Presidente da República
Fernando Henrique Cardoso

Ministro do Planejamento, Orçamento e Gestão
Martus Antônio Rodrigues Tavares

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Sérgio Besserman Vianna

Diretor-Executivo
Nuno Duarte da Costa Bittencourt

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Maria Martha Malard Mayer

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Paulo Roberto Ribeiro da Cunha

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Kaizô Iwakami Beltrão

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 60 número 213 janeiro/junho 1999

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 60, n. 213, p. 1-118, jan./jun. 1999.

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

Pedro Luís do Nascimento Silva
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.
Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2000

Capa
Renato J. Aguiar – Divisão de Criação – DIVIC/CDDI

Ilustração da Capa
Marcos Balster – Divisão de Criação – DIVIC/CDDI

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan/mar.1940), - Rio de Janeiro:IBGE, 1940- v.
Trimestral (1940-1986), semestral (1987-
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989).
Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN 0034-7175 = Revista brasileira de estatística.
I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.
IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO
Impresso no Brasil/Printed in Brazil

Sumário

NOTA DO EDITOR

Apresentamos mais um número da RBEs, com cinco artigos de diferentes especialidades e tratando de aplicações em áreas distintas. Abrindo o número, Bianchini e Albieri fazem "Uma revisão dos principais aspectos dos Planos Amostrais das Pesquisas Domiciliares realizadas pelo IBGE", trazendo informações atualizadas sobre as principais pesquisas domiciliares do País. Em seguida, dois artigos empregam simulação estocástica para tratar de problemas distintos: Fígoli avalia a incerteza de projeções populacionais no artigo "Intervalo de confiança para projeção de população baseado no método de Monte Carlo: Projeção de beneficiários urbanos da Previdência Social"; Mattos e Migon fazem uma "Análise da confiabilidade de itens submetidos a testes acelerados via simulação estocástica: O efeito da ortogonalização de parâmetros" na qual a simulação é empregada para permitir estimar parâmetros do modelo. Assunção, Reis, Marchesini e Sawyer elaboram "Mapas de malária em Rondônia usando o estimador Bayesiano empírico para dados binários", uma interessante aplicação de métodos modernos da estatística espacial. Fechando o número, Soares, César e Fonseca avaliam o "Desempenho das Escolas de Ensino Médio de Belo Horizonte no Vestibular da UFMG".

Ao sair este número, foi realizado mais um SINAPE. Encorajamos todos os autores de trabalhos apresentados no SINAPE a considerarem a submissão de versões completas de tais trabalhos para publicação na RBEs, levando em conta a política editorial em vigor.

Saudações,

Pedro Luis do Nascimento Silva

Editor Responsável

Uma Revisão dos Principais Aspectos dos Planos Amostrais das Pesquisas Domiciliares Realizadas pelo IBGE [1]

Zélia Magalhães Bianchini (IBGE)*

Sonia Albieri (IBGE)**

RESUMO

Desde 1960, o IBGE vem usando amostragem probabilística na realização de suas principais pesquisas domiciliares. Este artigo revisa alguns aspectos metodológicos das principais pesquisas domiciliares por amostra realizadas pelo IBGE, a saber: a Pesquisa Nacional por Amostra de Domicílios – PNAD –, iniciada em 1967; a Pesquisa Mensal de Emprego – PME –, desde 1980; a Pesquisa de Orçamentos Familiares – POF – , realizada duas vezes, em 1987/1988 e em 1995/1996; a Pesquisa sobre Padrões de Vida – PPV –, realizada em 1996/1997; a Pesquisa de Economia Informal Urbana – ECINF –, realizada em 1997 e a amostra para a coleta do questionário detalhado do Censo Demográfico desde 1960. O artigo aborda as semelhanças e as diferenças principais entre as pesquisas, no que se refere a objetivos, população-alvo, abrangência geográfica, plano amostral, incluindo aspectos tais como: cadastros, estratificação, conglomeração, número de estágios, tamanho da amostra, seleção da amostra, taxas de não-resposta e tratamentos adotados, estimação, avaliação de erros amostrais.

Palavras-chave: pesquisa por amostra de domicílios; plano amostral; estimação; taxa de resposta.

1. Introdução

O IBGE realiza várias pesquisas domiciliares com periodicidades diferentes (mensais, anuais, ou não-definidas), com graus variados de abrangência geográfica (nacional, regional, regiões metropolitanas, apenas áreas urbanas, alguns municípios de capitais de estado) e complexidade, de acordo não só com os recursos disponíveis, mas também com a área temática e

---

[1] Uma versão resumida deste artigo em inglês foi apresentada na Joint IASS/IAOS Conference, Statistics for Economic and Social Development, em Setembro de 1998, em Aguascalientes, México.

* Endereço para correspondência: IBGE – Diretoria de Pesquisas, Av. República do Chile 500, 10º andar – Rio de Janeiro – RJ – 20031-170 – Brasil – E-mail: zbianchi@ibge.gov.br .

** Endereço para correspondência: Deptº de Metodologia, Av. República do Chile 500, 10º andar – Rio de Janeiro – RJ – 20031-170 – Brasil - E-mail: salbieri@ibge.gov.br .

objetivos de cada pesquisa. Este artigo revisa alguns aspectos metodológicos das principais pesquisas domiciliares por amostra, realizadas pelo IBGE, a saber: a Pesquisa Nacional por Amostra de Domicílios – PNAD –, iniciada em 1967; a Pesquisa Mensal de Emprego – PME –, desde 1980; a Pesquisa de Orçamentos Familiares – POF –, realizada duas vezes, em 1987/1988 e em 1995/1996; a Pesquisa sobre Padrões de Vida – PPV –, realizada em 1996/1997; e a Pesquisa de Economia Informal Urbana – ECINF –, realizada em 1997.

Todas essas pesquisas adotam amostras probabilísticas de domicílios. Os desenhos amostrais das pesquisas listadas acima possuem várias semelhanças, as quais incluem amostragem de conglomerados em dois estágios (setores censitários + domicílios) ou mesmo três estágios (municípios + setores censitários + domicílios) com estratificação das unidades primárias de amostragem (UPAs). As UPAs (municípios ou setores censitários) são selecionadas com probabilidade proporcional a uma medida de tamanho – ppt (dentro de cada estrato). Os dados dos Censos Demográficos são freqüentemente usados para construir medidas de tamanho. Os setores censitários possuem em média 300 domicílios na área urbana e 200 domicílios na área rural. Para cada setor selecionado para a amostra de cada pesquisa é preparada uma listagem de todos os seus domicílios com o objetivo de preparar um cadastro atualizado para a seleção dos domicílios no último estágio de seleção. A precisão das estimativas é medida através dos coeficientes de variação (CVs) calculados para um conjunto de estimativas. O método do *ultimate cluster* de Hansen et al. (1953) é o que vem sendo usado na maioria das pesquisas para estimar a variância das estimativas de interesse.

As seções seguintes apresentam aspectos gerais de cada pesquisa aqui considerada, bem como uma breve descrição das principais diferenças entre as pesquisas, com relação a objetivos, população alvo, abrangência geográfica e plano amostral, incluindo aspectos tais como: cadastro, estratificação, conglomeração, número de estágios de seleção, tamanho da amostra, procedimento de seleção utilizado, taxas de não-resposta e tratamentos adotados, método de estimação e avaliação da precisão das estimativas.

Além dessas pesquisas, o sistema de pesquisas domiciliares inclui o Censo Demográfico, realizado a cada dez anos, que usa amostragem para obter dados sobre características selecionadas de pessoas, famílias e domicílios. O plano amostral e a metodologia de estimação adotados na pesquisa amostral do Censo são particularmente diferentes daqueles utilizados nas pesquisas realizadas no período intercensitário. Na seção 7, é apresentada uma breve descrição da amostra usada para a coleta do questionário de amostra dos últimos Censos Demográficos.

Vale ressaltar que este artigo não tem por objetivo descrever com profundidade os planos amostrais das pesquisas. Os detalhes metodológicos em geral e específicos de amostragem de cada pesquisa podem ser encontrados nas referências relacionadas no final do artigo.

2. Pesquisa Nacional por Amostra de Domicílios - PNAD

A PNAD foi implantada no País, gradativamente, a partir de 1967. O plano amostral da pesquisa foi inspirado no Plano Atlântida do US Bureau of the Census. Desde então, essa pesquisa tornou-se a pesquisa domiciliar anual mais importante no Brasil. É uma pesquisa com múltiplos propósitos que investiga características econômicas e sociais, principalmente aquelas relativas à situação da força de trabalho. Possui abrangência nacional, com exceção da área rural da Região Norte. Além disso, ocasionalmente são aplicados questionários suplementares na mesma amostra, e ao mesmo tempo em que se coleta a pesquisa tradicional, com o objetivo de investigar tópicos específicos, tais como: saúde, educação, trabalho, seguridade social, fertilidade, acesso aos serviços de saúde, contracepção e participação política.

A pesquisa possui um plano amostral autoponderado[2] com três estágios de seleção, estratificado, popularizado na década de 60 (Kish, 1965), para uso em pesquisas domiciliares. As unidades primárias de seleção são os municípios, que são estratificados por tamanho (população), e selecionados sistematicamente com probabilidade proporcional ao tamanho. No segundo estágio, os setores censitários são selecionados também de forma sistemática e com ppt, sendo que nesse caso o tamanho é medido pelo número de domicílios. Uma amostra sistemática simples de domicílios é então selecionada no terceiro estágio.

O plano amostral adotado para a PNAD considera uma estratificação das unidades primárias (municípios), definida separadamente em cada Unidade da Federação, com seleção de duas unidades por estrato. Municípios pertencentes à mesma microrregião geográfica foram agrupados em estratos com, aproximadamente, o mesmo tamanho. Os dados de população provenientes do Censo Demográfico foram as medidas de tamanho usadas para os procedimentos de estratificação e de seleção dos municípios. Os municípios grandes em termos populacionais e aqueles pertencentes às regiões metropolitanas foram tratados cada um como um estrato e portanto incluídos na amostra com certeza, e denominados auto-representativos. Os demais municípios selecionados em cada estrato são denominados não-auto-representativos e em cada um foram selecionados cinco setores.

O plano amostral é caracterizado por fração amostral fixa para cada região metropolitana, e para o restante da unidade da federação. Os municípios e os setores selecionados são mantidos na amostra até que estejam disponíveis os novos dados do Censo Demográfico, quando então são selecionadas novas unidades para a amostra. No momento em que foi feita a seleção de setores, o número de domicílios por setor para a amostra foi fixado e constante para todos os municípios. Quando a seleção da amostra foi atualizada com as novas definições de setores e com as medidas de tamanho baseadas nos dados do Censo Demográfico 1991, o número de domicílios por setor na amostra foi fixado em 13.

---

[2] Em um plano amostral autoponderado a probabilidade de uma unidade elementar qualquer pertencer à amostra é constante e igual à fração de amostragem geral, para um determinado nível geográfico.

A cada ano, em cada setor selecionado para a amostra, é preparada (ou atualizada) no campo uma listagem de domicílios, produzindo um cadastro atualizado para a seleção dos domicílios. Uma característica importante dessa operação de listagem refere-se ao Cadastro de Novas Construções, que é preparado de forma a conter os projetos responsáveis por alterações sérias nos tamanhos dos setores. O levantamento das novas construções é feito nos municípios da amostra, tanto nos setores da amostra como naqueles não selecionados para a amostra. Uma área de novas construções é excluída da área do setor original, e é tratada em separado no momento da seleção de domicílios que, nesse caso, é feita de acordo com a fração amostral da área.

Como a seleção de domicílios em cada setor selecionado para a amostra é feita sistematicamente, para garantir a autoponderação da amostra, o intervalo de seleção de domicílios permanece fixo de ano para ano. Esse procedimento acarreta um aumento anual no número de domicílios na amostra. Em um dado setor da amostra, o número de domicílios selecionados depende do tamanho atualizado do setor, em número de domicílios, atualização essa dada pela operação de listagem. Para ilustrar, na pesquisa de 1997, foram selecionados 109541 domicílios sendo 16,4 domicílios em média por setor (Tabela 2), enquanto que no dimensionamento da amostra com os dados de 1991, esse número médio foi fixado em 13.

O procedimento de estimação adotado na PNAD é baseado em estimação de razão com ajuste de população. Inicialmente, os pesos associados aos domicílios são obtidos considerando o plano amostral adotado sem tratamento para não-resposta. Então, esses pesos são ajustados (multiplicados) por um fator que é calculado independentemente para cada região metropolitana e para o restante de cada Unidade da Federação. Esse fator é a razão entre a estimativa independente de população para o período de referência da pesquisa e a estimativa de população proveniente da amostra. Os pesos associados às pessoas são aqueles calculados para os domicílios onde moram.

A precisão das estimativas provenientes da amostra é medida pelo coeficiente de variação (CV) calculado para um conjunto selecionado de estimativas de totais de variáveis categóricas. A variância do estimador de razão adotado é calculada pelo método denominado *ultimate cluster* (Hansen et al., 1953). As estimativas dos coeficientes de variação para variáveis categóricas são divulgadas através do ajuste de um modelo de regressão do tipo *Generalized variance functions* (Wolter, 1985). Ou seja, para cada domínio de publicação dos resultados, são ajustados modelos de regressão para explicar os CVs das estimativas de total como função das próprias estimativas. Os coeficientes dos modelos de regressão ajustados são publicados juntamente com os resultados da pesquisa a cada ano. Desde 1984, os resultados são publicados nos níveis nacional, Grandes Regiões, Unidades da Federação e regiões metropolitanas.

3. Pesquisa Mensal de Emprego - PME

A Pesquisa Mensal de Emprego foi iniciada em 1980. É uma pesquisa por amostra domiciliar realizada para fornecer estimativas do nível e das variações no emprego, desemprego e de outras características da força de trabalho. É realizada em seis Regiões Metropolitanas, a saber: Rio de Janeiro, São Paulo, Belo Horizonte, Recife, Salvador e Porto Alegre. Os resultados são produzidos para cada uma dessas regiões, bem como para o agregado das seis regiões, agregado esse que em 1996 representava cerca de 25% da população brasileira. Para cada uma dessas regiões, a amostra foi selecionada em dois estágios de forma independente. Em cada região metropolitana, os municípios constituem estratos de setores, os quais por sua vez são as unidades primárias de seleção dessa pesquisa. Os setores são selecionados sistematicamente com probabilidade proporcional a uma medida de tamanho (número de domicílios). Em cada região metropolitana, foi usada alocação proporcional dos setores ao longo dos estratos (municípios). Tal como na PNAD, após a atualização da listagem de domicílios em cada setor da amostra, o número de domicílios selecionados por setor aumenta de forma a manter a autoponderação. O Cadastro de Novas Construções preparado para a PNAD, também, é utilizado na PME. O tamanho da amostra referente ao mês de abril de 1998 é apresentado na Tabela 2, e o número médio de domicílios por setor é de 26,5.

Como a pesquisa é repetida a cada mês, foi estabelecido um esquema de rotação para a amostra a fim de evitar problemas de falta de cooperação dos entrevistados, que normalmente ocorrem em pesquisas por painéis fixos de domicílios. O procedimento de rotação foi definido de tal forma que cada domicílio permanece na amostra por quatro meses consecutivos, é retirado da amostra por oito meses, retorna para mais quatro meses de pesquisa e, então, é retirado definitivamente da amostra. Esse procedimento resulta em uma superposição de 75% da amostra a cada mês. Há também um processo de substituição dos setores que, por força do sistema de rotação da amostra, tiveram todos os seus domicílios incluídos na amostra.

O procedimento de estimação é semelhante ao da PNAD. É baseado no estimador de razão com ajuste para a população estimada para a data de referência. Para todas as estimativas de indicadores ou de totais calculados e divulgados a cada mês, são levadas em consideração apenas as informações da amostra daquele mês, ou seja, o aspecto longitudinal da amostra não é aproveitado na definição do estimador usado. Os coeficientes de variação são calculados para um conjunto selecionado de estimativas, usando o método do *ultimate cluster.*

4. Pesquisa de Orçamentos Familiares - POF

O objetivo principal da pesquisa é obter informações sobre os padrões de gastos das famílias para a construção dos novos pesos para os itens que compõem os diversos índices de preços que o IBGE calcula a cada mês. A primeira pesquisa sobre despesas familiares e renda no

Brasil foi o ENDEF - Estudo Nacional da Despesa Familiar, realizada em 1974/1975, que investigou também aspectos relativos à nutrição e antropometria, com plano amostral nos moldes do da PNAD. A Pesquisa de Orçamentos Familiares realizada em 1995/1996 foi planejada para representar a população urbana de nove regiões metropolitanas, do Distrito Federal e do Município de Goiânia. Essa pesquisa foi basicamente uma repetição da pesquisa realizada em 1987/1988, com alguma atualização metodológica.

Foram usados cinco tipos de questionários a saber: *questionário do domicílio*, incluindo características do domicílio e características demográficas dos moradores do domicílio; *questionário de despesa coletiva*, para as despesas com o domicílio e com bens duráveis; *caderneta de despesas coletivas diárias*, para despesas com alimentos, material de limpeza e higiene pessoal; *questionário de despesas individuais,* para despesas pessoais com alimentação fora do domicílio, transporte, fumo, livros, cuidados com a saúde, educação, roupas e outras; e *questionário de renda individual,* para todos os tipos de renda, incluindo salário, aluguel, rendimentos de aplicações financeiras e outras, além de taxas e deduções, tais como, imposto de renda. A coleta dos dados foi realizada no período de outubro/1995 a setembro/1996 de forma a captar os padrões sazonais de renda e despesas. As despesas foram obtidas através de entrevistas, usando diversos períodos de referência tais como: semana, mês, trimestre e semestre. Porém, as despesas com itens menos prováveis de serem lembrados exatamente pelos entrevistados (pequenas compras, despesas com alimentação e outras despesas de uso coletivo) foram obtidas através do registro diário em uma caderneta, feito preferencialmente pelo próprio informante, durante sete dias, no caso da POF1995/1996 e de 14 dias no caso da POF 1987/1988.

Cabe destacar vários aspectos inovadores que foram incorporados no planejamento da amostra da POF, já na sua primeira aplicação. A respeito ver IBGE (1992), sobre a metodologia da POF 1987/1988, e Bianchini e Vieira (1998), sobre a metodologia da POF 1995/1996.

Em cada área da pesquisa, o plano amostral considera uma amostra em dois estágios de seleção com estratificação da unidade primária. A unidade primária de seleção, o setor censitário, foi estratificado em duas etapas: estratos geográficos (núcleo e periferia) e renda média do chefe do domicílio no setor. Os setores foram selecionados sistematicamente com probabilidade proporcional ao tamanho (medido em número de domicílios particulares ocupados). Em cada setor da amostra, os domicílios foram selecionados através de amostragem aleatória simples sem reposição. Pela primeira vez, em pesquisas do IBGE, a seleção de domicílios a pesquisar em cada setor foi feita sem empregar amostragem sistemática, com o sorteio aleatório efetuado por computador.

Para a POF 1995/1996, o tamanho da amostra em cada área foi determinado a partir de uma precisão especificada para estimar a renda total do chefe (CV=0,05), com o número de domicílios a serem selecionados em cada setor da amostra fixado em dez, usando as informações investigadas no questionário básico do Censo Demográfico 1991. A amostra de setores foi dividida

em quatro subamostras, uma para cada trimestre de coleta. A alocação dos setores nas subamostras foi aleatória, preservando a estratificação adotada, de tal forma que todos os estratos estão representados em todos os trimestres.

Antes do início da coleta, em cada setor selecionado para a amostra, foi preparada uma listagem completa dos domicílios de forma a construir um cadastro atualizado para a seleção das unidades de segundo estágio. Entretanto, o número de domicílios selecionados em cada setor foi aumentado para 13 para compensar a seleção de domicílios vagos, fechados e possíveis recusas. Para setores com taxas de crescimento acima de certos níveis, o número de domicílios selecionados foi novamente aumentado, de acordo com patamares de crescimento, de forma a reduzir a variância dos pesos, mas o número máximo de domicílios selecionados por setor foi de 28. Uma novidade em relação à tradição das demais pesquisas domiciliares foi a eliminação do requisito da autoponderação. A amostra da POF 1996/1997 ficou com 19.816 domicílios selecionados (Tabela 2).

O procedimento de estimação é baseado no estimador de razão com calibração[3] na população residente em domicílios particulares urbanos dada pela Contagem de População de 1996. Os pesos associados a cada domicílio da amostra foram obtidos usando o estimador natural derivado do plano amostral, com tratamento para não-resposta, multiplicado por um fator que foi calculado independentemente para cada área da pesquisa. Esse fator é a razão entre a população residente em domicílios particulares permanentes urbanos dada pela Contagem de População 1996 e a estimativa da população correspondente proveniente da amostra. A data de referência da Contagem de População foi 01.08.96, que é próxima da data de referência definida para a POF, 15.09.96. A data de referência da POF foi usada para a correção da inflação, tornando todos os valores das despesas e receitas a preços constantes. Para o município da capital de cada Região Metropolitana, exceto Belém, foi calculado um outro ajuste para calibrar a população do município, gerando um conjunto de pesos à parte para ser usado apenas na obtenção de estimativas no nível do município.

A precisão das estimativas amostrais foi medida pelo coeficiente de variação calculado para um conjunto de estimativas de total, usando o método do *ultimate cluster*. A publicação dos coeficientes de variação para variáveis categóricas foi feita usando a função generalizada de variância (*generalized variance functions*, Wolter, 1985). Ou seja, para cada domínio de publicação, foi ajustado um modelo de regressão para explicar os CVs das estimativas de total como função do valor das próprias estimativas. Os coeficientes do modelo de regressão ajustados são publicados juntamente com as estimativas da pesquisa, bem como os próprios CVs para uma seleção de estimativas relativas a variáveis contínuas relacionadas com valores de despesas e de rendimentos.

---

[3] Calibração é o processo pelo qual os fatores de expansão para a amostra são determinados de forma a buscar a consistência das estimadas a partir da amostra com os totais conhecidos da população.

5. Pesquisa sobre Padrões de Vida - PPV

A Pesquisa sobre Padrões de Vida realizada em 1996/1997 foi uma pesquisa piloto baseada no *Living Standard Measurement Study* - LSMS estabelecido pelo Banco Mundial, em 1980, para desenvolver métodos para a coleta e análise de dados sobre padrões de vida em países em desenvolvimento. As pesquisas do tipo LSMS foram realizadas em vários países, com vários objetivos analíticos, tais como medir a distribuição do bem estar e o nível de pobreza dos domicílios, para entender como os domicílios reagem a programas econômicos e ambientais de governo, e permitir análises complexas da relação entre vários aspectos do bem estar do domicílio. (Grosh e Muñoz, 1996 e Caillaux, 1998).

Essa pesquisa caracteriza-se por: "inclusão de temas socioeconômicos, estudados de forma integrada em um mesmo domicílio, um ano de permanência no campo (março de 1996 a março de 1997) de forma a captar fenômenos sazonais e, manutenção de um controle rigoroso tanto na aplicação do questionário como na entrada de dados e no processo de crítica das informações" (Caillaux, 1998).

Para atingir esses objetivos, foram considerados vários aspectos inovadores no planejamento da pesquisa. Os domicílios foram entrevistados duas vezes, com intervalo de duas semanas entre as entrevistas. O questionário foi quase todo pré-codificado de forma que os dados puderam ser digitados diretamente em microcomputadores imediatamente após a entrevista, de forma descentralizada no campo. O programa de entrada de dados realiza críticas de consistência e de validade das respostas. Os dados inconsistentes ou errados eram marcados pelo programa de tal forma que as respostas de certas questões puderam ser verificadas durante a segunda entrevista. Os informantes foram instruídos para registrarem suas despesas durante as duas semanas que antecederam a segunda visita.

A pesquisa foi planejada para investigar uma diversidade de temas sociais e econômicos, a saber: características do domicílio, características básicas demográficas, migração, saúde, mão-de-obra, fecundidade, rendimentos, investimentos e créditos, despesas com bens duráveis, despesas com alimentação, empreendimentos domiciliares, agricultura, avaliação do padrão de vida e antropometria.

O plano amostral inclui estratificação dos setores censitários (as UPAs - unidades primárias de amostragem) em dez estratos geográficos: de Fortaleza, Recife, Salvador, o restante da área urbana da Região Nordeste, o restante da área rural da Região Nordeste, as Regiões Metropolitanas de São Paulo, Rio de Janeiro, Belo Horizonte, o restante da área urbana da Região Sudeste, o restante da área rural da Região Sudeste. Em cada estrato geográfico, foram definidos outros três estratos com base no rendimento médio domiciliar mensal do chefe do domicílio, usando os dados do Censo Demográfico 1991. O tamanho da amostra em número de setores em cada estrato foi determinado por alocação proporcional. Os setores foram selecionados com

probabilidade proporcional ao número de domicílios, com reposição, em cada um dos trinta estratos definidos. A amostra de setores foi dividida em quatro subamostras, uma para cada trimestre de coleta. A alocação da amostra nas subamostras foi aleatória e de tal forma a preservar a estratificação, ou seja, todos os estratos estão representados em todos os trimestres. Em cada setor selecionado para a amostra, foi realizada uma operação denominada listagem de domicílios, com o objetivo de obter um cadastro atualizado para a seleção das unidades de segundo estágio. A partir dessa listagem, os domicílios foram selecionados através de amostragem aleatória simples sem reposição. A operação de listagem foi realizada por partes, o mais próximo possível do início da coleta de cada trimestre. Nos setores pertencentes aos estratos de região metropolitana ou de área urbana, foram selecionados oito domicílios por setor. Nos estratos de setores rurais esse número foi fixado em 16.

O tamanho da amostra foi de 554 setores, 278 na Região Nordeste e 276 na Região Sudeste, correspondendo a um tamanho total de amostra de 4 944 domicílios. Com o objetivo de compensar a não-resposta por motivos de recusa, domicílios vagos ou fechados, foi selecionada uma segunda amostra de domicílios nos mesmos moldes da primeira. Os domicílios com não-resposta foram então substituídos por domicílios provenientes dessa segunda amostra. Esse procedimento foi adotado em função do reduzido tamanho amostra, que tornava indesejável perder informação. Apesar disso, foram perdidos quatro questionários do primeiro trimestre de pesquisa. Essa perda foi compensada por reponderação dos domicílios informantes pelo inverso da taxa de resposta. Com o objetivo de evitar um aumento na taxa de não entrevistas devidas à recusa dos informantes, no processo de seleção da amostra de setores foram adotados procedimentos para evitar a seleção de setores já selecionados para a PNAD, a PME e a POF, uma vez que essas pesquisas foram realizadas na mesma época da PPV e possuem planos amostrais semelhantes.

A estimação de totais e dos erros amostrais associados foi realizada usando o estimador natural derivado do plano amostral adotado. A precisão das estimativas foi medida pelo coeficiente de variação calculado para todos os indicadores usados para a análise e disseminação dos resultados, usando o método do *ultimate cluster* (Albieri e Bianchini, 1997).

6. Pesquisa de Economia Informal Urbana - ECINF

Em 1997, o IBGE realizou sua primeira pesquisa domiciliar por amostragem em nível nacional para identificar atividades econômicas desenvolvidas nos domicílios ou em pequenas unidades produtivas, de forma a medir o papel e a dimensão dessas atividades na economia brasileira, através da identificação dos proprietários de negócios informais e da investigação das características de funcionamento das unidades produtivas (Jorge, 1995).

A população-objetivo incluí as pessoas residentes na área urbana que trabalhavam por conta própria ou como empregadores com até cinco empregados, em pelo menos uma situação

de trabalho de atividades não-agrícolas. Os trabalhadores domésticos foram excluídos da população-objetivo.

O objetivo da pesquisa foi produzir estimativas para cada um dos 26 estados, o Distrito Federal, cada uma das dez regiões metropolitanas e o Município de Goiânia.

Foram usados dois tipos de questionários na coleta dos dados: o primeiro para obter informação sobre as características dos domicílios e das pessoas moradoras, com o objetivo de identificar as pessoas engajadas em unidades produtivas do setor informal, através das características do trabalho; o segundo questionário foi usado para investigar as características das unidades produtivas do setor informal e seus proprietários.

O plano amostral considera dois estágios de seleção e estratificação das unidades de seleção. As unidades primárias de amostragem (UPAs) foram os setores urbanos estratificados primeiramente pela localização geográfica. Em cada estado, foram definidos dois ou três estratos geográficos dependendo se o estado possui ou não região metropolitana, o município da capital do estado, os demais municípios da região metropolitana, o restante do estado. Foi definido um segundo nível de estratificação das UPAs, dentro de cada estrato geográfico, de acordo com a renda média domiciliar do setor, obtida dos dados do Censo Demográfico 1991. A seleção das unidades primárias em cada estrato foi feita através de amostragem sistemática com probabilidade proporcional ao tamanho (medida em número de domicílios ocupados). As unidades de segundo estágio foram os domicílios com moradores classificados como conta própria ou empregador com até cinco empregados. Os domicílios foram posteriormente classificados de acordo com o grupo de atividades e selecionados sistematicamente a partir da lista atualizada de domicílios realizada nos setores selecionados para a amostra.

Com o objetivo de evitar um aumento na taxa de não entrevistas devidas à recusa, no processo de seleção dos setores foi adotado um procedimento para eliminar as coincidências com os setores selecionados para a PNAD e PME, uma vez que essas duas pesquisas seriam realizadas na mesma época que a de economia informal.

O tamanho da amostra em cada área foi determinado a partir de um coeficiente de variação especificado para estimar o número de proprietários (conta próprias e empregadores) na economia informal (CV=5%, excepcionalmente, por motivos de custo, CV=6% para cada área da Região Norte). O número de domicílios a serem selecionados por setor foi fixado em 16. O tamanho da amostra é apresentado na Tabela 2. (Almeida e Bianchini, 1998).

No que se refere a aspectos de planejamento amostral, a pesquisa de economia informal difere das pesquisas domiciliares tradicionais do IBGE. Isto porque é preciso lidar com uma população rara, heterogênea e mais difícil de ser detectada. Todos esses fatores contribuem para aumentar a complexidade do plano amostral, da seleção da amostra, dos procedimentos de estimação e principalmente da preparação do cadastro de unidades amostrais de interesse, ou seja, da listagem de domicílios, a qual requer a realização de uma pesquisa em todos os

domicílios do setor para identificar as atividades desenvolvidas pelos moradores de cada domicílio. (Kalton e Anderson, 1986).

A listagem dos domicílios da ECINF foi uma operação com custo elevado, pois além de produzir uma lista completa de endereços das unidades domiciliares, envolveu a realização de entrevista para obter as informações necessárias para identificar a população-objetivo, para obter as informações para a segunda estratificação por grupo de atividades que constituíram o objetivo da pesquisa. Os oito grupos de atividades considerados foram: (1) indústria da transformação e extrativa mineral; (2) indústria da construção; (3) comércio de mercadorias; (4) serviços de alojamento e alimentação; (5) serviços de transporte; (6) serviços de reparação, pessoais, domiciliares e de diversões; (7) serviços técnicos e auxiliares; (8) outros serviços.

Em cada setor da amostra, a alocação dos 16 domicílios foi feita proporcionalmente ao número de domicílios existentes em cada grupo de atividades no setor. Além disso, foram feitos alguns ajustes que ocasionaram um aumento médio de 30% no número de domicílios a serem selecionados por setor.

A estimação foi feita com base no estimador natural que deriva do plano amostral adotado, com tratamento para não-resposta. Os domicílios listados como tendo proprietários do setor informal e que na entrevista não tinham mais essa característica foram excluídos na estimação de totais, mas receberam o valor zero para cada variável de interesse da investigação. Além dos pesos para estimação das características de proprietários de unidades produtivas do setor informal, foi também associado um peso à unidade produtiva do setor informal, que leva em conta o inverso do número de sócios da unidade produtiva. A precisão das estimativas foi medida pelos coeficientes de variação calculados para um conjunto selecionado de estimativas de total. O método do *ultimate cluster* foi usado para estimar a variância de cada variável.

7. A Amostra para a coleta de dados do questionário detalhado do Censo Demográfico

O Censo Demográfico 1960 foi o primeiro a utilizar amostragem na coleta de dados relativos a um conjunto selecionado de características de pessoas, famílias e domicílios. Foram usados dois tipos de questionários: um questionário pequeno aplicado a todos os domicílios e seus moradores, não-selecionados para a amostra (chamado questionário básico); e um questionário longo (chamado questionário da amostra) aplicado a todos os domicílios selecionados para a amostra, bem como seus moradores. Nos Censos de 1960, 1970 e 1980, foi utilizada uma única fração amostral de 25% dos domicílios. Em 1991, uma revisão amostral com grande impacto foi o emprego de duas frações amostrais diferentes de acordo com o tamanho do município, medido em função da projeção de população para a data de referência do Censo: 20% para os municípios com até 15 000 habitantes e 10% para os demais municípios.

O Território Nacional foi dividido em partições geográficas denominadas setores censitários, de tal forma que seus limites respeitam as divisões internas do município em zonas urbanas e rurais, e em distritos e subdistritos, caso existam. O setor censitário foi planejado de forma a que um entrevistador consiga realizar a operação de coleta no período de realização do Censo. Em 1991, foram definidos 163.266 setores censitários (que, no plano amostral, foram considerados como estratos). Os domicílios particulares foram selecionados por amostragem sistemática em cada setor censitário. As famílias ou pessoas sós moradoras em domicílios coletivos (alojamentos estudantis, quartéis, prisões, hospitais, orfanatos, conventos, etc.) foram selecionadas, também de forma sistemática, independentemente da seleção de domicílios particulares, usando a mesma fração amostral definida para o setor a que pertence cada domicílio coletivo.

O procedimento de estimação de totais foi aplicado em cada área de ponderação separadamente. Uma área de ponderação é um conjunto de setores censitários que constituem a menor área geográfica para a qual as estimativas provenientes da amostra são avaliadas em termos de precisão. Em 1991, o método de Mínimos Quadrados Generalizados em duas etapas *(Generalized Least Squares Estimation Procedure - GLSEP*, Bankier, Rathwell e Majkowski, 1992), denominado no IBGE por MQG2, foi usado na determinação dos pesos e as variáveis auxiliares utilizadas foram definidas dentre aquelas investigadas para 100% da população, no próprio Censo Demográfico. Esse procedimento de estimação de regressão atribui um único peso fracionário a cada domicílio e a cada um de seus moradores, sendo importante destacar essas duas situações novas em relação aos censos anteriores: o peso fracionário e único para domicílios, famílias e pessoas.

8. Resumo e conclusões

As tabelas a seguir apresentam de forma resumida algumas características gerais dos planos amostrais, os tamanhos das amostras, as distribuições da amostra por tipo de entrevista e as taxas de resposta para as várias pesquisas domiciliares, excetuando o Censo Demográfico.

Tabela 1 - Características gerais dos planos amostrais das várias pesquisas

(continua)

Pesquisa	*Abrangência Geográfica*	*Estágios de seleção*	*Tipo de estratificação das UPAs*
PNAD	Nacional (exceto o Norte rural)	3 (municípios, setores e domicílios)	Geográfica
PME	6 regiões metropolitanas	2 (setores e domicílios)	Geográfica
POF	11 áreas urbanas	2 (setores e domicílios)	Geográfica e classes de renda do chefe
PPV	Regiões Nordeste e Sudeste	2 (setores e domicílios)	Geográfica e classes de renda do chefe
ECINF	Nacional, somente áreas urbanas	2 (setores e domicílios)	Geográfica e classes de renda domiciliar

Tabela 1 - Características gerais dos planos amostrais das várias pesquisas

(continuação)

Pesquisa	*Auto-ponde-ração*	*Seleção de setores*	*Variável usada como medida de tamanho*	*Cadastro de seleção de domicílios*	*Seleção de domicílios*
PNAD	sim	sistemática com ppt	número de domicílios	Listagem + Novas construções	sistemática simples
PME	sim	sistemática com ppt	número de domicílios	Listagem + Novas construções	sistemática simples
POF	não	sistemática com ppt	número de domicílios particulares ocupados	Listagem	aleatória simples sem reposição
PPV	não	ppt com reposição	número de domicílios particulares	Listagem	aleatória simples sem reposição
ECINF	não	sistemática com ppt	número de domicílios ocupados	Listagem com entrevista para identificar população objetivo	seleção sistemática

Tabela 1 - Características gerais dos planos amostrais das várias pesquisas

(conclusão)

Pesquisa	*Periodicidade*	*Período de coleta*	*Aspectos longitudinais*
PNAD	anual (desde 1967, exceto em anos de Censo Demográfico)	3 meses	nova seleção da amostra de domicílios (3º estágio)
PME	mensal (desde 1980)	um mês	rotação da amostra de setores e domicílios
POF	1974/1975 (ENDEF) 1987/1988 e 1995/1996	um ano	nova seleção da amostra a cada execução da pesquisa
PPV	1996/1997 (pesquisa piloto)	um ano	—
ECINF	1997 (primeira pesquisa)	3 meses	—

Tabela 2 – Tamanho da amostra e distribuição por tipo de entrevista das várias pesquisas

Pesquisa	*Número de setores selecionados*	*Número de domicílios na amostra*			
		Selecionados	*Média por setor*	*Eleitos* (**)	*Entrevistados*
PNAD 1997	6 678	109 541	16,4	91 811	90 006
PME Abril/1998	1 510	40 090	26,5	32 549	30 951
POF 1995/1996	1 456	19 816	13,6	17 628	16 014
PPV 1996/1997 (*)	554	4 944	8 (urbanos) e 16 (rurais)	4 944	4 940
ECINF 1997	2 340	48 934	20,9	38 099	37 010

(*) **Foi adotado o procedimento de substituição de domicílios durante a coleta.**

(**) **Os domicílios eleitos foram definidos de acordo com a população alvo de cada pesquisa.**

Tabela 3 - Taxas de resposta das várias pesquisas

Pesquisa	*Taxas*		
	Entrevistados / selecionados (%)	*Eleitos / selecionados (%)*	*Entrevistados / eleitos (%)*
PNAD 1997	82,2	83,8	98,0
PME abril/1998	77,2	81,2	95,1
POF 1995/1996	80,8	89,0	90,8
PPV 1996/1997 (*)	99,9	100,0	99,9
ECINF 1997	75,6	77,9	97,1

(*) **Foi adotado o procedimento de substituição de domicílios durante a coleta.**

Tabela 4 - Procedimentos de estimação das pesquisas por amostra

Pesquisa	*Estimador*	*Variável de calibração*	*Tratamento de não-resposta*	*Publicação de erros amostrais*
PNAD 1997	razão	Projeção de População	não	ajuste de modelo de regressão para variáveis categóricas
PME abril/1998	razão	Projeção de População	não	calculados para um conjunto de estimativas
POF 1995/1996	razão	Contagem de População 1996	sim	seleção de variáveis contínuas e ajuste para variáveis categóricas
PPV 1996/1997	natural do desenho	não	sim	calculados para todos os indicadores divulgados
ECINF 1997	natural do desenho	não	sim	seleção de variáveis

As taxas de resposta representam a qualidade das pesquisas, pelo menos sob a ótica da aceitação por parte dos informantes, e são consideradas satisfatórias, em função da complexidade de cada pesquisa.

O IBGE vem realizando várias pesquisas domiciliares com diferentes níveis de complexidade em função das diferenças nos objetivos, na população-alvo, na abrangência geográfica, no nível de precisão e na periodicidade definidas para cada uma. Apesar das muitas semelhanças entre os planos amostrais das pesquisas citadas, as diferenças apontadas, também revelam mudanças e aperfeiçoamentos metodológicos que foram possíveis de serem implantados ao tempo em que cada pesquisa foi planejada ou realizada. Essas modificações (aperfeiçoamentos) incluem a incorporação de estratificação de acordo com o nível de renda do setor, a previsão de perda de unidades da amostra por não-resposta, o tratamento da não-resposta, a redução do número de domicílios selecionados por setor, a redução do número de estágios de seleção e a eliminação do requisito de autoponderação.

Essa revisão serviu de base para as discussões que vêm sendo realizadas com vistas à reformulação da pesquisa mensal de emprego. A reformulação dos planos amostrais das pesquisas de forma a reduzir custos através de amostras menores conjugada com novos e aperfeiçoados procedimentos de estimação colocam-se como um desafio para o IBGE.

Como aperfeiçoamentos possíveis, vale destacar: a incorporação de efeitos espaciais ou de vizinhança na estratificação dos setores; a consolidação de tratamentos para a não-resposta total; a introdução de estimadores que se beneficiem da estrutura longitudinal da amostra, no caso da

PME, por exemplo; o uso de sistemas genéricos para estimação de dados de pesquisas com planos amostrais complexos, que proporcionam economia de tempo e qualidade, com facilidades para o cálculo de erros amostrais também.

Referências Bibliográficas

ALBIERI, S. E BIANCHINI, Z.M. (1997). Aspectos de amostragem relativos à pesquisa domiciliar sobre padrões de vida. Rio de Janeiro: IBGE, 14p. mimeo.

ALMEIDA, R.A.P. E BIANCHINI, Z.M. (1998). Sampling aspects of the 1997 Brazilian survey of the urban informal sector. Proceedings of the Joint IASS/IAOS Conference.

ALMEIDA, R.A.P. E BIANCHINI, Z.M. (1998). Aspectos de amostragem da Pesquisa de Economia Informal Urbana 97. Rio de Janeiro: IBGE, 32p. (Texto para Discussão, nº 89).

BANKIER, M.D., RATHWELL, S. E MAJKOWSKI, M. (1992). Two step generalized least squares estimations in 1991 Canadian Census. Ottawa: Statistics Canada, Methodology Branch Working Paper.

BIANCHINI, Z.M. E VIEIRA, M. (1998). Aspectos de amostragem da Pesquisa de Orçamentos Familiares 1995-1996. Rio de Janeiro: IBGE, 106p. (Texto para Discussão, nº 93).

CAILLAUX, E.L. (1998). Living standard survey 1996-1997. Rio de Janeiro: IBGE. 10p. [Apresentado no Meeting of the Expert Group on Poverty Statistics (Rio Group), 13-15 Maio, 1998].

COCHRAN, W.G. (1977). Sampling Techniques (3rd ed.). New York : Wiley.

GROSH, M. E. E MUÑOZ, J. (1996). Manual for planning and implementing the LSMS Survey. Poverty e Human Resources Division, Policy Research Department, The World Bank.

HANSEN, M.H., HURWITZ, W.N. E MADOW, W.G. (1953). Sample Survey Methods and Theory, Vol. I e II. New York: Wiley.

IBGE (1983). Metodologia da Pesquisa Nacional por Amostra de Domicílios na década de 80. Rio de Janeiro: IBGE (Série Relatórios Metodológicos, vol.1).

IBGE (1983). Metodologia da Pesquisa Mensal de Emprego 1980. Rio de Janeiro: IBGE (Série Relatórios Metodológicos, vol.2).

IBGE (1992). Pesquisa Orçamentos Familiares. Volume 3. Aspectos de Amostragem. Rio de Janeiro: IBGE, 218p. (Série Relatórios Metodológicos, vol. 10).

JORGE, A. (1995). The survey of the urban informal economy in Brazil. Proceedings of the International Seminar on Informal Sector Employment Statistics, pp 239-256.

JORGE, A.F. (1996). Pesquisa de Economia Informal Urbana. Rio de Janeiro: IBGE, 17p. [Artigo apresentado no Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais].

KALTON, G. E ANDERSON, D.W. (1986). Sampling rare populations. The Journal of the Royal Statistical Society A, 149, part 1, pp 65-82.

KISH, L. (1965). Survey sampling. New York: Wiley.

SILVA, P.L.N. (1996). Planejamento, estimação e análise de dados em pesquisas por amostragem: desvendando a realidade brasileira com o "telescópio da estatística". Rio de Janeiro: IBGE, 28p. [Artigo apresentado no Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais].

WOLTER, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

ABSTRACT

Since 1960 IBGE (Instituto Brasileiro de Geografia e Estatística), the Brazilian Central Statistical Office, has used probabilistic sampling to conduct its major household surveys. This paper reviews some design aspects for each of the major household sample surveys: the national household sample survey (PNAD, started in 1967); the monthly employment survey (PME, since 1980); the income and expenditure survey (POF, taken twice in 1987/88 and 1995/96); the living standards survey (PPV, conducted in 1996/97); the survey of the urban informal sector (ECINF, conducted in 1997); and the sample collection of the decennial population census long form (since 1960). The paper will focus on both similarities and main differences among the surveys listed regarding their objectives, target population, geographical coverage, sampling design, including aspects such as: frames, stratification, clustering, number of stages, sample size, sample selection, nonresponse rates and treatments, estimation, sampling error evaluation.

Intervalo de Confiança para Projeção de População Baseado no Método de Monte Carlo: Projeção de Beneficiários Urbanos da Previdência Social

Moema Gonçalves Bueno Fígoli*

RESUMO

Neste estudo projetamos os beneficiários urbanos da Previdência Social relativos aos grupos de espécie: aposentadoria por tempo de serviço; por idade; e especial, de 1990 a 2040. O método de projeção utilizado foi o de simulação, onde as diversas simulações das possíveis trajetórias da população foram realizadas utilizando o modelo de coortes-componentes, mas as variáveis envolvidas: novos beneficiários e mortalidade; foram consideradas como aleatórias, geradas pelo método de Monte Carlo a partir de uma distribuição uniforme. O pressuposto de aleatoriedade tornou possível a construção de intervalo de confiança para a população futura.

A aplicação do modelo resultou no número de beneficiários em 2040 entre 19 956 755 e 15 348 687, com nível de confiança de 95%.

Palavras-chaves: Projeção aleatória, Simulação de Monte Carlo e Previdência Social.

1. Introdução

As projeções de população são ferramentas fundamentais para planejadores governamentais e privados, utilizadas para muitos propósitos. O interesse principal pode ser o de conhecer o total da população, uma componente do total, o número de pessoas em uma determinada faixa etária, ou alguma função da distribuição etária, como, por exemplo, a razão de dependência. Um exemplo atual da importância das projeções para o estabelecimento de políticas governamentais é a discussão

* Endereço para correspondência: UFMG - Rua: Curitiba, 832 – 8º andar Cep. 30170-120 – Belo Horizonte – MG.

acerca do impacto da dinâmica demográfica sobre o Sistema de Previdência Social em vigor no País. Neste contexto, o conhecimento da massa de beneficiários desta instituição a longo prazo pode influenciar a determinação e o planejamento do comportamento da taxa de contribuição previdenciária, e mesmo a alteração dos planos de benefícios, ou seja, o aumento ou a redução da idade de aposentadoria; ou, a alteração de algum benefício.

Apesar da sua importância, geralmente as projeções de populações têm resultado em prognósticos sobre as gerações futuras que diferem da realidade. No entanto, como afirma Keyfitz (1981):

> *Demographers can no more be held responsible for inaccuracy in forecasting population 20 years ahead than geologists, meteorologists, or economists when they fail to announce earthquakes, cold winters, or depressions 20 years ahead.*
>
> *What we can be held responsible for is warning one another and our public what the error of our estimates is likely to be. Statistics started to approach scientific maturity with the calculation of probable errors (for instance in astronomical observations), and then it went on to confidence intervals and tests of significance.*

Um dos procedimentos mais usados para projetar populações é o *método de coortes-componentes* que consiste basicamente em projetar anualmente, ou em intervalos de cinco anos, o número futuro de nascimentos, mortos e migrantes, adicionando-os para formar um novo vetor de população. Esse cálculo é repetido para cada ano de projeção (Shryock, Siegel, 1976). Tal metodologia, muito usada até hoje, foi introduzida por P. K. Welpton com uma seqüência de artigos iniciada em 1928. Não se trata de um método detalhado de projeção. Na verdade ele fornece uma estrutura geral, dentro da qual as taxas específicas das estatísticas vitais podem ser projetadas conforme o desejado. No entanto, uma de suas principais virtudes é permitir a decomposição do problema de projeção da população em uma questão de projeção das três componentes de mudança da população: fecundidade, mortalidade e migração.

Para considerar as incertezas com relação às componentes demográficas, são determinados *cenários*: normalmente um alto, um médio e um baixo, baseados em hipóteses sobre o comportamento futuro das variáveis, a partir dos quais são geradas projeções independentes utilizando-se esse método. Muitos elementos são levados em conta para a determinação da trajetória média: pesquisas de intenções, opiniões de *experts*, modelos estatísticos, fatores econômicos e sociais, etc.; em contrapartida, a construção das trajetórias alta e baixa é em geral menos sistemática. Apesar de usualmente não ser dada nenhuma interpretação de probabilidade ao resultado dessa projeção, o usuário é levado a pensar que o intervalo entre as séries contém o valor futuro.

Na verdade, esta prática oferece uma série de resultados, sem apontar, no entanto, aquele que seria o *mais provável de ocorrer*, como diz Keyfitz (1987):

The logic of this approach is that the demographer presents a few main possibilities in respect of the componentes of population growth, shows what population will result in twenty or more years later, and leaves the selection to the user. It is up to the user to study the assumptions on which the components were projected forward, choose the set of assumptions that seems right to him, and then accept only the demographers' arithmetic to read out the resulting future population.....

Assim, o demógrafo normalmente se limita a apresentar o resultado de algumas projeções, mas não quantifica a incerteza, digamos, na projeção média, que é a considerada mais provável de se realizar. Ele também não dá a sua opinião sobre a probabilidade de ocorrência das projeções alta e a baixa.

Com efeito, para Keyfitz (1981), *Without some probability statement, high and low estimates are useless to indicate in what degree one can rely on the medium figure, or when one ought to use the low or the high.* Desse modo, a questão que se coloca é: *como podemos determinar os erros nas projeções* de forma a estabelecer o intervalo dentro do qual deverá estar compreendida a população futura, com algum nível de confiança.

Neste sentido, alguns autores realizaram estudos com vistas a determinar se a *projeção alta e baixa da abordagem tradicional representam um intervalo de confiança para a projeção média.* Alho e Spencer (1985, 1991) argumentam que a conseqüência do método determinístico e não estatístico é a discrepante probabilidade de a população futura cair dentro do intervalo alto/baixo para projeções em diferentes pontos no tempo e para diferentes subgrupos. Essa conclusão baseou-se na comparação entre a razão dada pelas projeções alta e média da população dos EUA, realizada pelo *US Bureau of Censo*, e a razão dada pelo limite superior do intervalo de confiança de 90% em relação à projeção média, da previsão realizada por eles, utilizando modelo estatístico. Alho e Spencer, ao analisá-la, concluíram que o nível do intervalo de confiança da projeção, a partir de 1980, dos sobreviventes de grupos em idade entre 15 e 30 anos, era inicialmente menor que 90%, passando a ser superior a 90% após cinco anos. Para os nascimentos, no entanto, o intervalo de confiança encontrado apresentou nível de confiança inferior a 67%. Quando combinados a projeção dos nascimentos e a dos sobreviventes nas outras idades, o intervalo de confiança encontrado apresentou nível de confiança variável.

Por sua vez, Lee (1992) argumenta que a projeção por *métodos tradicionais não permitiria estabelecer limites de confiança para a previsão.* A fim de esclarecer a sua posição, ele realiza uma série de análises de várias componentes da projeção, dentre as quais, a do cálculo da razão de dependência: divide-se a população em idade igual ou superior a 65 anos, pela população entre 20-64. Segundo ele, a projeção baseada em séries altas combina denominador alto, resultante de alta fecundidade, com numerador alto, resultante de baixa mortalidade já a projeção baseada em séries

baixas combina baixo denominador com baixo numerador. Com base nesta análise, Lee conclui que o intervalo gerado dessa forma será mais estreito do que um que permitisse flutuações.

Considerando o exposto até aqui, de que não seria possível indicar os erros ou incertezas na projeção da população somente através de informações resultantes das projeções alta, média e baixa, *novas alternativas* para a projeção de população vêm sendo propostas e aplicadas por diversos autores. As soluções encontradas para a estimação do erro e a construção do intervalo de confiança têm sido: uso de métodos de séries temporais para projetar a taxa de crescimento da população (Cohen,1986; Pflaumer, 1992); simulação aleatória do crescimento da população com estatísticas vitais variando aleatoriamente (Pflaumer, 1988); análise dos erros de projeções passadas (Keyfitz, 1981; Stoto, 1983); desenvolvimento de modelos aleatórios para as estatísticas vitais, que são usadas em matrizes aleatórias de Leslie para gerar distribuições de probabilidade para a população futura (Lee, 1974; Saboia, 1974; McDonald,1979 and 1981; Alho and Spencer, 1985 and 1991; Carter and Lee, 1986; McNown and Rogers, 1990; Alho, 1990 e outros); e, uso de modelos demográficos e de séries temporais para a projeção das estatísticas vitais, e do produto de matrizes aleatórias, para a projeção da população com estatísticas vitais aleatórias (Lee, Tuljapurkar, 1994).

Considerando as vantagens relativas à *interpretação* dos resultados e de determinação antecipada do *grau de precisão*, que vêm sendo apontadas para as projeções de população que integram metodologias *demográficas e estatísticas*, discutidas acima, pretendemos com este trabalho utilizar *o método de simulação* desenvolvido por Pflaumer (1988), a fim de projetar beneficiários urbanos da Previdência Social no período de 1990 a 2040. Tal procedimento combina metodologias estatísticas e demográficas, uma vez que as diversas simulações de trajetórias da população são realizadas utilizando-se o modelo de coortes-componentes, mas as variáveis envolvidas na projeção são aleatórias. Com base nestas simulações, será determinado o erro na projeção e construído o intervalo de confiança.

O *método de simulação*, desenvolvido por Pflaumer em 1988, consiste em várias simulações das possíveis trajetórias da população. Para tanto, as estatísticas vitais são geradas pelo método de simulação de Monte Carlo, conforme uma distribuição de probabilidade assumida. Suas vantagens são: a de ser conceitualmente simples, a de fornecer os resultados em nível da distribuição por idade da população e a de permitir o estabelecimento do intervalo de confiança para todos os resultados da projeção.

Do elenco de benefícios oferecidos pela Previdência Social aos seus segurados, os *grupos para cujos beneficiários futuros pretendemos realizar a projeção* são: aposentadoria por tempo de serviço urbana; aposentadoria por idade urbana; e aposentadoria especial urbana. A opção de projetar os beneficiários futuros com direito a esses grupos de benefícios foi tomada, considerando que estes são previsíveis, assim, é possível aos filiados à Previdência, programarem a sua data de aposentadoria de forma a obterem maiores vantagens. Dessa forma, tais benefícios estão sujeitos a maiores flutuações.

2. Modelo de projeção

O método que utilizaremos para a projeção dos *novos beneficiários urbanos* que aderem às seguintes espécies de benefício: 1) aposentadoria por tempo de serviço; 2) aposentadoria por idade; e 3) aposentadoria especial , a cada qüinqüênio; e para a projeção da *massa de beneficiários urbanos* entre 1990 e 2040 nas mesmas espécies de benefícios, será o método demográfico de *coortes-componentes*. No entanto, diferentemente da abordagem tradicional deste método, onde as variáveis usadas são tratadas de forma determinística, neste trabalho aquelas nele envolvidas serão consideradas *aleatórias*. Este pressuposto de aleatoriedade implica que o tamanho da população em cada momento do tempo seja também uma variável aleatória. Com base nessa formulação, nos é possível deduzir a distribuição da população projetada e seu intervalo de confiança, tanto por métodos teóricos, como os utilizados por Sykes (1969), Cohen (1986), Lee e Tuljapurkar (1994), e outros, quanto por métodos de simulação, como o de simulação de Monte Carlo, utilizado por Pflaumer (1988).

Na projeção desenvolvida aqui, utilizaremos a versão da simulação do modelo de coortes-componentes, com os elementos envolvidos se comportando como variáveis aleatórias com uma distribuição específica. Assim sendo, faz-se necessário que definamos: 1) o intervalo dessas variáveis; 2) a distribuição a ser utilizada; e 3) o método de amostragem dessas variáveis aleatórias a partir da sua distribuição.

As variáveis que serão utilizadas na projeção serão o número de novos beneficiários[1] e a esperança de vida. Para definirmos os *limites de variação da variável novos beneficiários* no início de cada período, entre 1990 e 2040, nos baseamos no número e distribuição dos benefícios concedidos por espécie, idade e sexo, verificados no período de 1980 a 1993, descritos nas Tabelas 1 e 2, e consideramos as alterações advindas das tendências de crescimento futuro da população e cenários de comportamento das taxas de atividade, o que gerou as taxas de crescimento por período especificadas na Tabela 3[2]. Para obtermos os *limites de variação da variável esperança de vida,* no mesmo período, adotamos o método de projeção das taxas de mortalidade, aleatório com intervalo de confiança associado, desenvolvido por Lee e Carter (1992), e construímos as tábuas de vidas associadas à taxa de mortalidade projetada e os limites superior e inferior do intervalo de confiança de confiança de 95% desta variável.[3] Estes limites podem ser observados na Tabela 4.

Para estabelecermos a *forma da distribuição* dessas variáveis, poderíamos analisar a distribuição empírica dos seus valores passados. A adoção deste procedimento, no entanto, exige

[1] Por novos beneficiários entendemos aquela parcela da população vinculada à Previdência social, que passa da situação de segurado ativo para a situação de inativo.

[2] Para maiores detalhes sobre o estabelecimento destes limites veja Fígoli, M.G.B. (1997) págs. 42 a 65.

[3] Para maiores detalhes sobre a aplicação do modelo e elementos das tábuas de vida veja, Fígoli, M.B.G. (1998).

que disponhamos de uma longa e consistente série de dados e que admitamos a repetição do comportamento dessas variáveis no futuro. Tais exigências não podem ser atendidas. Primeiro, não dispomos de uma série extensa o suficiente. Segundo, não acreditamos que o seu comportamento passado se verifique no futuro. Assim, qualquer distribuição que viéssemos a estabelecer seria incerta, mesmo baseada no comportamento passado. Então, fizemos uma escolha subjetiva. Dentre as distribuições disponíveis, optamos por estabelecer que as variáveis aleatórias teriam uma *distribuição uniforme ou retangular* entre os limites *[a,m,b]*, dada a simplicidade e facilidade de operacionalização. Ou seja, estamos pressupondo que qualquer subintervalo de *[a,m,b]*, de mesmo comprimento, tem a mesma probabilidade de conter a variável aleatória.

Um campo de aplicação do método de Monte Carlo é o de *amostragem de variáveis aleatórias a partir de uma distribuição de probabilidade*. Assim, essa será a metodologia adotada para a obtenção dos valores das variáveis-número de novos beneficiários e esperança de vida, distribuídas uniformemente.

Este item será dividido em três subitens. No primeiro, descreveremos o *método de projeção de população*, o método das componentes, com as adaptações necessárias para a operacionalização dessa projeção. No segundo, apresentaremos *o método de Monte Carlo* e o procedimento de geração das variáveis aleatórias. No terceiro, *o processo de simulação da projeção* de novos beneficiários e da massa de beneficiários, e a metodologia de *construção de intervalo de confiança* para os resultados da projeção.

2.1 - O Método de projeção

O método aqui utilizado é o de projeções por *coortes-componentes*, apresentado pela primeira vez por Whelpton (1928), conforme o qual, partindo-se de uma população distribuída por grupos etários de cinco anos, os sobreviventes são computados ao longo de linhas de coortes e os nascimentos, menos os mortos na primeira infância, são adicionados a cada ciclo da projeção. Mais tarde este método de projeção foi formalizado por Bernardelli (1941), Lewis (1942), e Leslie (1945,1948). Este último o fez com tal detalhe que o processo passou a ser conhecido pelo seu nome.

De fato, Leslie observou que, se a população inicial estiver representada por um vetor N_0, distribuída em grupos etários, digamos, de cinco anos, e se fosse construída uma matriz X composta de razões de sobrevivência na diagonal principal ${}_5L_5 / {}_5L_0$, ${}_5L_{10} / {}_5L_5$, etc. , quando multiplicarmos esta matriz pelo vetor de população, estaremos levando a população ao longo de linhas de coortes sucessivas. A contribuição dos nascimentos, nesta formulação, é dado pela primeira linha da matriz. Assim, os sobreviventes depois de cinco anos seriam dados por:

$$N = X N_0$$

onde: N_0 é o vetor de população inicial;

N é o vetor de sobreviventes ao final de 5 anos; e

X é a matriz composta por razões de sobrevivência.

e após t anos teríamos:

$$N_t = X^t N_0$$

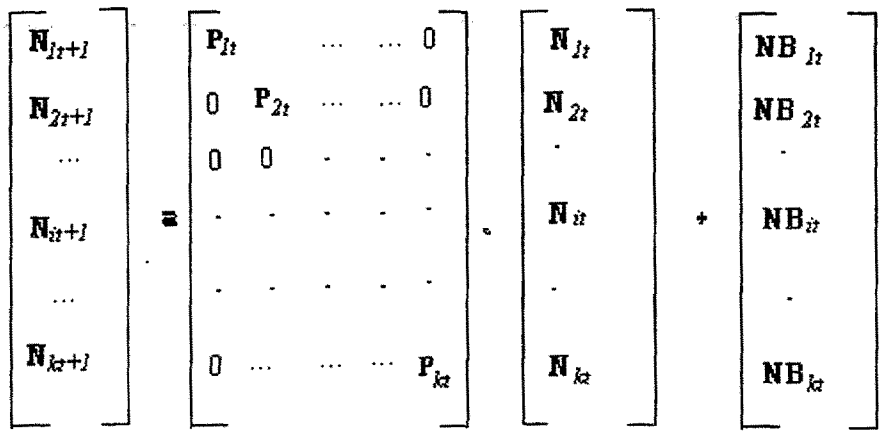
onde: N_t é o número de sobreviventes após t anos.

Se considerarmos que as variáveis que compõem a matriz de projeção mudam a cada período, a cada período teremos uma matriz de projeção; e, se ainda considerarmos que a população não é fechada e que temos um número de imigrantes (I_t) a ser adicionado a cada período, teremos após t anos:

$$N_t = X_{t-1} \ldots\ldots X_2 X_1 N_0 + I_t$$

onde: $X_{t-1} \ldots\ldots X_2 X_1$, são as matrizes de razões de sobrevivência nos diversos períodos.

Considerando que a população que pretendemos projetar é *adulta*, constituída pela massa de beneficiários da Previdência Social em dezembro de 1990, e que as incorporações ao sistema previdenciário não se dão nem via nascimento e nem via imigração, como no modelo de Leslie, e sim através da *transferência da atividade para a inatividade*, a matriz de transição será constituída unicamente de razões de sobrevivência na diagonal principal e zeros nas demais posições; e o vetor de imigrantes corresponderá aos novos beneficiários que entram em cada uma das aposentadorias a cada qüinqüênio. Dadas estas alterações, o modelo de projeção na forma matricial seria:



Onde:

N_{it+1} representa o número de pessoas do grupo de idade i no tempo t+1;

P_{it} representa as razões de sobrevivência ($_5L_i / _5L_{i-1}$) do grupo de idade i no tempo t;

N_{it} representa o número de pessoas do grupo de idade i no tempo t; e

NB_{it} representa o número de novos beneficiários do grupo de idade i no tempo t.

Para a operacionalização do modelo, é necessário que definamos as *variáveis que o compõem*, ou seja, a *mortalidade*, na forma de razões de sobrevivência, e o *número de novos beneficiários*. Como dissemos anteriormente, as variáveis que utilizaremos aqui serão aleatórias, obtidas pelo método de simulação de Monte Carlo, a partir de uma distribuição uniforme no intervalo *[a, m, b]*. Assim sendo, estabelecemos que: 1) a variável aleatória gerada e da qual obteremos posteriormente as razões de sobrevivência aleatórias, será a *esperança de vida;* e 2) a variável aleatória gerada, que representará as incorporações ao sistema previdenciário, será o *total de novos beneficiários* em cada período.

2.2 - Método de simulação de Monte Carlo

O método de *simulação* nos permite obter soluções numéricas, geralmente aproximadas, dos modelos matemáticos, como resultado da substituição das variáveis e parâmetros do modelo, por valores numéricos.

No sentido amplo, a simulação é uma técnica que nos permite realizar experimentos amostrais sobre um modelo. Por *simulação aleatória*, no entanto, entende-se um experimento que pressupõe amostragem de variáveis aleatórias a partir de uma distribuição de probabilidade. Como a amostragem a partir de uma distribuição em particular envolve o uso de números aleatórios, a

simulação aleatória é usualmente chamada de *simulação de Monte Carlo*, já que, historicamente, o método de Monte Carlo é considerado uma técnica que usa números aleatórios ou pseudo-aleatórios, para a solução de um modelo.

O método de Monte Carlo pode ser usado tanto para solução de problemas aleatórios como também para a solução de problemas determinísticos, desde que estes últimos tenham a mesma expressão formal que algum processo aleatório. Atualmente, o método vem sendo aplicado no cálculo de integrais complexas, na resolução de certas equações em física que não admitem solução analítica, como também na *amostragem de variáveis aleatórias a partir de distribuição de probabilidade*, como é o caso deste trabalho.

2.2.1 - Geração de variável aleatória a partir de uma distribuição

Neste ponto estamos interessados em empregar o método de Monte Carlo para gerar as variáveis aleatórias *número de novos beneficiários e esperança de vida*, a partir da *distribuição uniforme*, no intervalo *[a,m,b]*. Normalmente isto é feito tomando-se números de uma distribuição uniforme *[0,1]* e transformando-os, de alguma maneira. Para tanto, existem diversos procedimentos que podemos utilizar baseados nos três seguintes métodos: método transformação inversa; método composição; e método rejeição-aceitação[3].

O método que nos pareceu mais apropriado foi o método da *transformação inversa*. Segundo ele, para se gerar um valor x de uma variável aleatória X, com uma determinada distribuição, teremos que obter um valor y, de uma variável aleatória uniformemente distribuída no intervalo *[0,1]*; computar a inversa da função de distribuição acumulada de X no ponto y, $F_X^{-1}(y)$. O valor x, que desejamos obter, será igual a essa inversa.

Sabemos que se X é uma variável aleatória com função de distribuição acumulada $F_X(x)$, desde que $F_X(x)$ seja não decrescente, a função inversa $F_X^{-1}(y)$ pode ser definida para qualquer valor de y entre 0 e 1 como: $F_X^{-1}(y)$ é o menor x que satisfaz $F_X(x) \geq y$, isto é,

$$F_X^{-1}(y) = menor\{x : F_X(x) \geq y\}, \quad 0 \leq y \leq 1.$$

Se o valor y é obtido de variável aleatória *U* uniformemente distribuída no intervalo (0,1), então

$$x = F_x^{-1}(y)$$

tem função de distribuição acumulada $F_X(x)$.

---

[3]Neste sentido veja Hammersley, Handscomb (1992), páginas 36 a 40.

Neste caso, estamos pressupondo que as duas variáveis aleatórias, digamos *Y*, são distribuídas uniformemente no intervalo *[a,m,b]*. Essa distribuição é contínua e não decrescente. Logo a inversa pode ser calculada conforme prevê o método da transformação inversa. Assim, a *função de densidade* de cada uma destas variáveis é dada por:

$$f_y(y) = \begin{cases} 1/2\,(m-a), & a \le y < m \\ 1/2\,(b-m), & m \le y < b \\ 0 & \text{em outro caso} \end{cases}$$

O valor esperado e a variância são dados por:

$$E(Y) = (a+2m+b)/4,$$

$$Var(Y) = 1/6\,[(a^2 + a\ m + m^2) + (b^2 + b\,m + m^2)] - [(a+2m+b)/4]^2$$

A função de distribuição acumulada é dada por:

$$F_Y(y) = \begin{cases} 0, & y < a \\ (y-a)/(m-a), & a \le y \le m \\ (y-m)/(b-m), & m < y \le b \\ 1, & y > b \end{cases}$$

e a *inversa da função de distribuição acumulada* no ponto u, obtido da distribuição uniforme *[0,1]*, que nos fornecerá o valor da variável aleatória, digamos y, da variável aleatória Y, distribuída uniformemente no intervalo $[a, m, b]$, será:

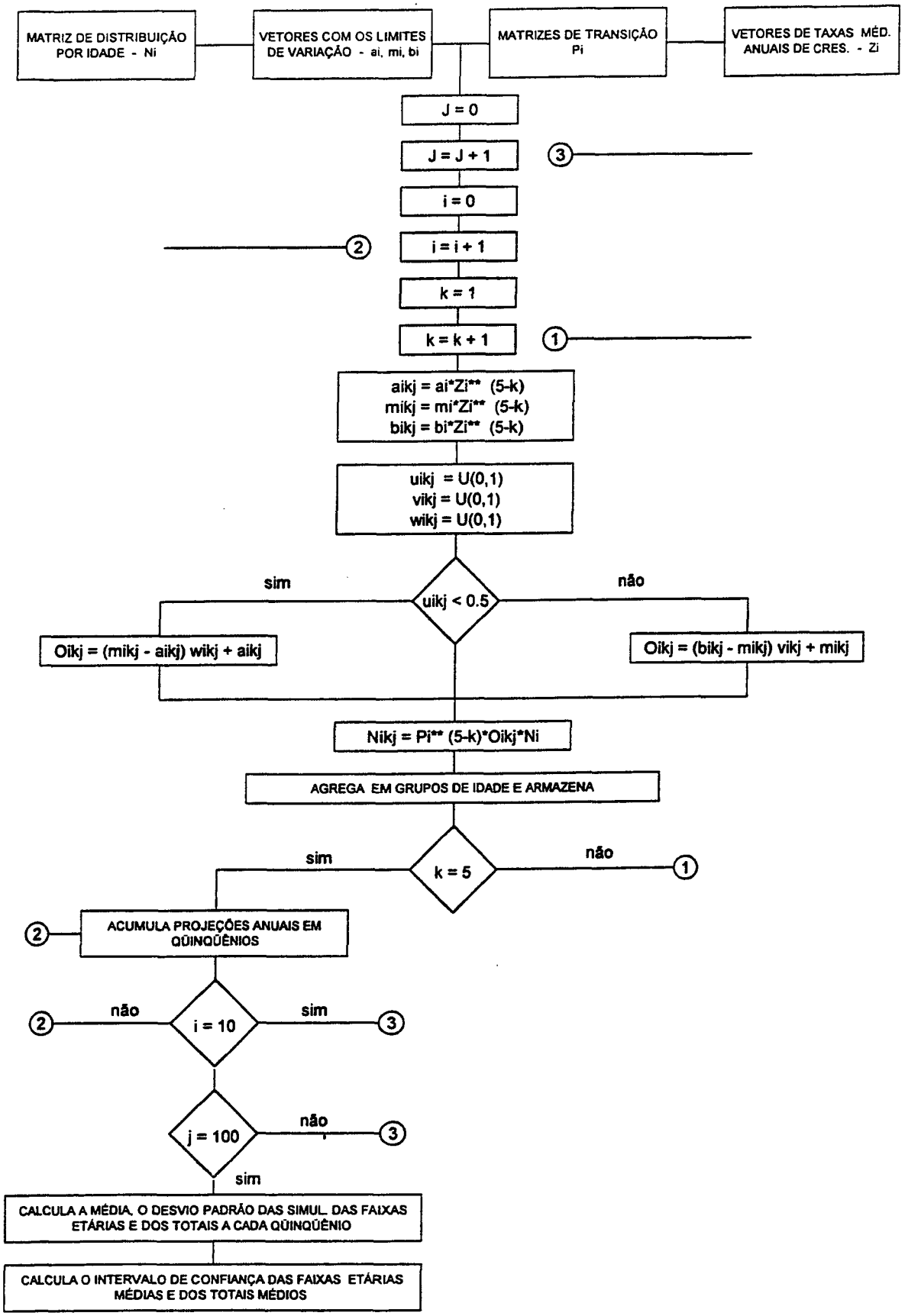
$$y = F_Y^{-1}(u) = \begin{cases} a + (m-a)\,u & para \quad a \le y \le m \\ m + (b-m)\ u & para \quad m < y \le b \end{cases}$$

2.3 - A simulação da projeção

Procederemos à projeção em duas etapas: na primeira projetaremos o *número de novos beneficiários* que aderem ao sistema previdenciário a cada ano para o final do qüinqüênio; em seguida projetaremos a *massa inicial de beneficiários*, agregando a ela, ao final de cada qüinqüênio, a nova massa de beneficiários. Para isso, apresentaremos a seguir o fluxograma do processo de simulação de cada uma das projeções, lembrando que este processo será realizado em separado para cada uma das espécies de benefício: 1) aposentadoria por tempo de serviço; 2) aposentadoria por idade; e 3) aposentadoria especial e, dentro de cada espécie, para o sexo feminino e masculino.

2.3.1 - A Simulação de novos beneficiários

Figura 1



onde: Ni - é uma matriz constituída de 10 vetores, cada um correspondendo à distribuição percentual dos novos beneficiários em idade simples, num dos período *i* e projeção;

a_i, m_i, b_i - são os limites inferior, médio e superior, respectivamente, do intervalo de variação da variável novos beneficiários, no início de cada período de projeção *i*[5] ;

Zi - vetor constituído por taxas médias de crescimento anual a cada período *i* dos limites iniciais de novos beneficiários[6];

Pi - representa as matrizes de transição, compostas na diagonal principal pelas razões de sobrevivência extraídas das tábuas de vida projetadas para cada período *i*;

j - corresponde ao número de simulações. Serão realizadas 100 simulações;

i - corresponde ao número de períodos. Como queremos projetar de 1990 a 2040, *i* variará de 1 até 10;

k - corresponde ao ano dentro de cada período. *k* variará de 1 até 5;

$U_{ikj}, V_{ikj}, W_{ikj}$ - correspondem a valores de variáveis aleatórias distribuídas uniformemente no intervalo *[0,1]*;

O_{ikj} - corresponde ao valor da variável aleatória número de novos beneficiários do ano *k*, do período *i*, da simulação *j*, obtido pelo método de simulação de Monte Carlo, a partir de uma distribuição uniforme no intervalo $[a_{ikj}, m_{ikj}, b_{ikj}]$; e

N_{ikj} - corresponde ao vetor de novos beneficiários por idade, que entram em aposentadoria no ano *k*, projetados para o final do período *i*, da simulação *j*.

Conforme o fluxograma acima, o *número de novos beneficiários* que adere a algum dos três tipos de benefício em estudo, a cada ano, gerado aleatoriamente pelo método de simulação de Monte Carlo, é projetado para o final de cada qüinqüênio. Os limites do intervalo da variável aleatória a_i, m_i e b_i são estabelecidos para cada início de período e corrigidos ano a ano por uma taxa média calculada. Uma vez projetados para o final do qüinqüênio, os novos beneficiários distribuídos em idade simples são agregados por faixas etárias. Ao final de cinco anos, os novos beneficiários projetados anualmente são agregados, de modo a representar o número de novos beneficiários que entraram em aposentadoria no período. O processo é operacionalizado para cada um dos dez períodos. A projeção da mesma população de novos beneficiários é repetida 100 vezes, gerando 100 diferentes trajetórias[7]. Com base nelas, calculamos o vetor que representa a trajetória média, o desvio padrão e o intervalo de confiança de cada elemento do vetor.

---

[5] Veja como estes limites foram estabelecidos no item 2 – MODELO DE PROJEÇÃO, terceiro parágrafo.
[6] Veja como o crescimento foi estabelecido no item 2 – MODELO DE PROJEÇÃO, terceiro parágrafo.
[7] O processo de simulação foi repetido 100 vezes porque não foi observada nenhuma diferença significativa nos parâmetros estimados ao aumentarmos o número de simulações.

São também calculados o total do vetor de cada trajetória, a média, o desvio padrão e o intervalo de confiança do total .

2.3.1.1 – Os estimadores para média e variância e cálculo do intervalo de confiança

Ao procedermos à projeção de novos beneficiários conforme o processo de simulação descrito acima, obteremos uma amostra aleatória constituída de 100 possíveis valores (ou vetores) para o número de novos beneficiários $(X_1, X_2, \ldots X_{100})$ ao final de cada período. Contudo, estamos interessados em determinar o valor médio da população constituída por todos os valores possíveis da variável. O processo de inferência estatística, por sua vez, permite estabelecer propriedades da população com base em resultados observados na amostra. Assim, podemos estimar o valor médio e a variância da população através de estimadores não tendenciosos, obtidos a partir dos dados da amostra. Então, a média da população será dada por:

$$X = (X_1 + X_2 + \ldots + X_{100})/n$$

onde n é o tamanho da amostra, e a variância da população por:

$$S^2 = 1/(n-1)\sum_{i=1}^{n}(X_i X)^2$$

Sabemos, em conseqüência do teorema do Limite Central, que a distribuição da média é assintoticamente normal, desde que tenhamos uma amostra grande (normalmente para n > 30). Assim sendo, podemos estabelecer os limites do intervalo de confiança para a média da população como:

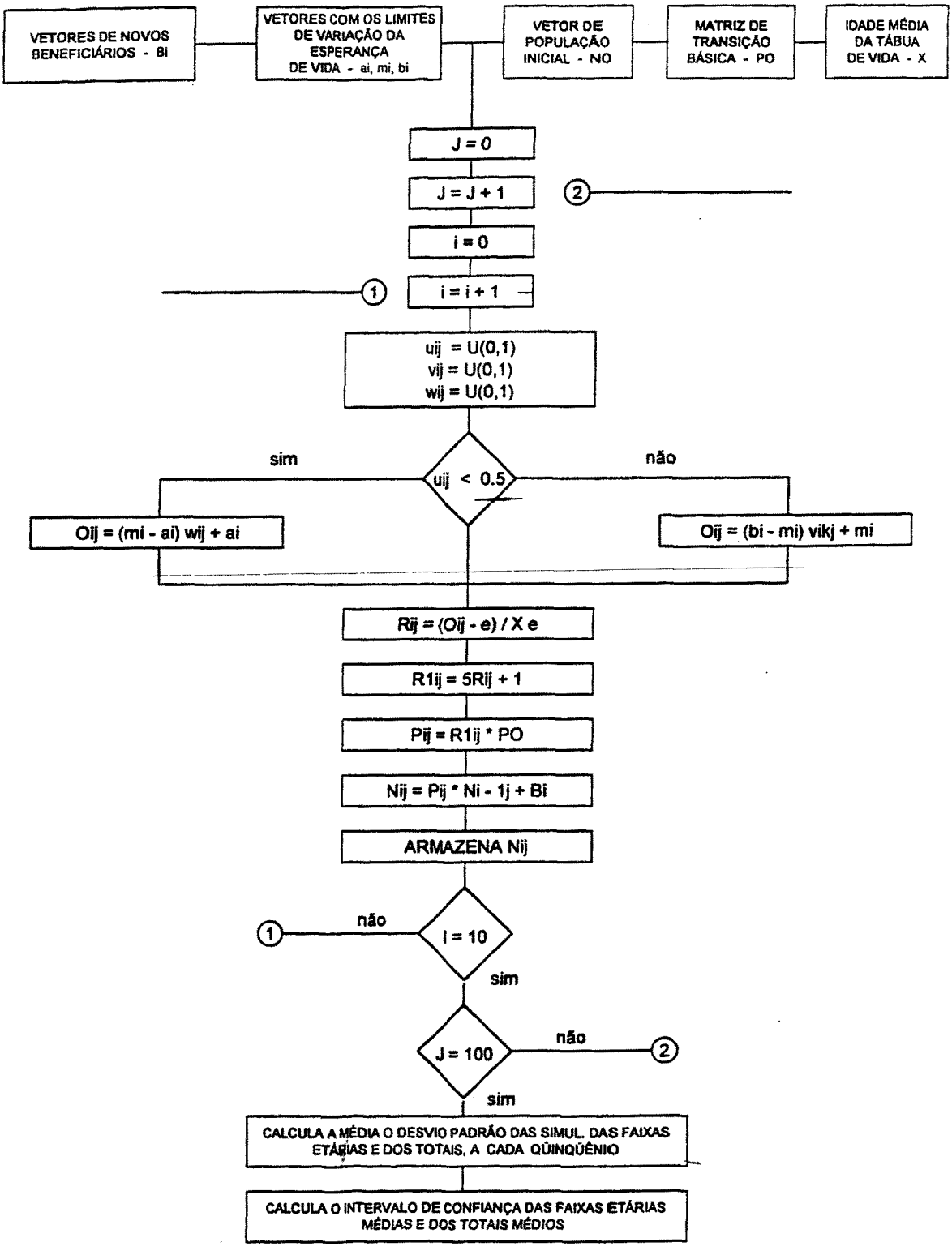
$$X \pm Z_c \sigma / \sqrt{n}$$

onde : Z_c são valores críticos de uma distribuição normal correspondente ao nível de confiança

$\sigma / \sqrt{n}$ é o desvio padrão da amostra e σ foi estimado por S^2.

2.3.2 - Simulação da massa de beneficiários

Figura 2



onde: Bi - corresponde a cada um dos vetores médios de novos beneficiários encontrados na projeção anterior, para cada um dos períodos *i*;

N_0 - é o vetor inicial de beneficiários, neste caso, aqueles que estavam aposentados em dezembro de 1990;

P_0 - matriz de transição básica, composta por razões de sobrevivência a serem extraídas da tábua de mortalidade, a partir do limite superior da taxa central de mortalidade para o período de 1990-1995. Estas razões de sobrevivência estão retratadas na Tabela 5;

a_i, m_i, b_i - são os limites inferior, médio e superior, respectivamente, do intervalo da variável esperança de vida, no início de cada período de projeção *i*;

X - idade média da tábua de vida básica;

j - número de simulações. Serão realizadas 100 simulações;

i - número de períodos. Como pretendemos projetar de 1990 a 2040, i varia de 1 até 10;

$[U_{ij}, V_{ij}, W_{ij}]$ - valores de variáveis aleatórias distribuídas uniformemente no intervalo *[0,1]*;

O_{ij} - valor da variável aleatória esperança de vida do período *i*, da simulação *j*, obtido pelo método de simulação de Monte Carlo, a partir de uma distribuição uniforme no intervalo $[a_{ij}, m_{ij}, b_{ij}]$;

R_{ij} - mudança na taxa central de mortalidade em decorrência do crescimento na variável aleatória esperança de vida. O valor *e* representa a esperança de vida da tábua de mortalidade base;

R_{lij} [8] - taxa de crescimento do período *i*, da simulação *j*, a ser aplicada às razões de sobrevivência da matriz de transição básica;

P_{ij} - matriz de transição do período *i* da simulação *j*; e

N_{ij} - corresponde ao vetor de massa de beneficiários projetado para o final do período *i*, da simulação *j*.

Neste último fluxograma, apresentamos o processo de *simulação da massa de beneficiários*. A massa de beneficiários em 1990 foi projetada para os diversos períodos, usando razões de sobrevivência aleatórias, e em cada período foram adicionados os novos beneficiários, projetados aleatoriamente. A variável aleatória esperança de vida foi gerada pelo método de Monte Carlo, a partir da distribuição uniforme no intervalo $[a_{ij}, m_{ij}, b_{ij}]$. Sua variação, com relação à esperança de vida da tábua-base, foi calculada e, levando-se em conta esta variação, estabelecida a taxa de

crescimento nas razões de sobrevivência da matriz de transição base. O processo de simulação foi repetido 100 vezes para cada período, considerando-se a natureza aleatória das variáveis mortalidade e novos beneficiários, o que gerou 100 diferentes trajetórias. Com base nelas foi calculado o vetor que representa a trajetória média, o desvio padrão e o intervalo de confiança de cada elemento do vetor. Foram também calculados o total do vetor de cada trajetória, a média, o desvio padrão e o intervalo de confiança do total .

2.3.2.1 - Estimadores para média e variância e cálculo do intervalo de confiança

Obteremos como resultado do processo de simulação acima, a cada período, uma amostra aleatória constituída de 100 trajetórias diferentes da massa de beneficiários. Podemos então estimar a *trajetória média* e a *variância da população* de trajetórias da mesma forma que fizemos para os novos beneficiários. Lembramos, porém, que, ao final de cada período, adicionamos à massa de beneficiários os novos beneficiários, estimados de forma independente, que aderiram ao sistema previdenciário durante o período anterior. Isto posto, a massa média composta pela massa dada pela trajetória média, estimada neste subitem, é somada ao número médio de novos beneficiários estimados no subitem anterior. Então, para construirmos o intervalo de confiança, teremos que estabelecer distribuição amostral da soma de médias da amostra.

Nesse sentido, sabemos que a variável padronizada associada às médias $(X_1 + X_2)$, é dada por:

$$Z = ((X_1 + X_2) - (\mu_1 + \mu_2)) / V(\sigma^2{}_1 / n_1 + \sigma^2{}_2 / n_2)$$

onde: X_1, X_2 representam a média de cada uma das amostras;

μ_1, μ_2 representam a média de cada uma das populações;

σ_1^2, σ_2^2 representam o desvio padrão de cada uma das populações;

n_1, n_2 representam o tamanho de cada uma das amostras; e

tem distribuição assintoticamente normal, se n_1, n_2 são grandes.

Então, concluímos que a *soma das médias das amostras tem distribuição aproximadamente normal* e podemos estabelecer os *limites do intervalo de confiança para a soma das médias da populações* como:

$$X_1 + X_2 \pm Z_c \, V(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

onde: Z_c são valores críticos de uma distribuição normal correspondente ao nível de confiança.

---

[8] Esta taxa de crescimento será calculada considerando que a mudança na taxa central de mortalidade será constante em todas as idades, conforme formulação de Keyfitz. (1985: 62).

3. Análise dos resultados

As Tabelas 6 a 8 apresentam a síntese dos resultados no que se refere aos: *totais do número médio de novos beneficiários* projetados para cada espécie de benefício e sexo, ao final de cada qüinqüênio; os limites inferiores - LI e os limites superiores - LS do intervalo de confiança de 95% destes totais; as percentagens destes totais que a metade da amplitude do intervalo de confiança representa; e as taxas de crescimento qüinqüenal dos totais, para cada espécie de benefício e sexo.

As *taxas de crescimento dos totais da massa média de beneficiários* apresentadas nas tabelas são diferenciadas entre as espécies, sexo e períodos. O comportamento destas taxas é fortemente determinado pelo peso relativo da massa de novos beneficiários de cada sexo, que aderem a cada espécie de benefício a cada qüinqüênio, sobre o total de beneficiários em aposentadoria. Nos primeiros anos da projeção, o número de novos beneficiários é elevado, proporcionalmente ao total em benefício, gerando altas taxas de crescimento entre os períodos. No decorrer do tempo, a relação entre novos beneficiários e total em aposentadoria reduz-se gradualmente, levando à redução das taxas de crescimento. Há que se observar, no entanto, que, como as taxas de crescimentos dos novos beneficiários em aposentadoria por idade são bastante elevadas e flutuam durante o período; o mesmo comportamento se dá em relação ao total da massa de beneficiários.

As *amplitude do intervalo de confiança dos totais da massa média de beneficiários,* em cada espécie de benefício e sexo, no entanto, depende do intervalo de variação estabelecido para as variáveis mortalidade e novos beneficiários. Ao estabelecermos os limites de variação da variável mortalidade, consideramos que eles seriam diferenciados unicamente por sexo, sendo a variação da esperança de vida em 2 040 das mulheres entre 71,21 a 83,48 anos, e a dos homens, de 64,63 a 78,07 anos, ou seja, uma variação de 12,27 anos para o sexo masculino e 13,44 anos para o sexo. Para a variável novos beneficiários, os limites de variação foram diferenciados por sexo e espécie. As maiores amplitudes do intervalo de confiança são relativas à projeção média da massa de beneficiários feminina, nas espécies tempo de serviço e especial.

4. Conclusões

Nesse estudo aplicamos um *método de simulação aleatória* para a estimativa futura dos beneficiários urbanos da Previdência Social. Nele integramos o método tradicional de coortes-componentes a métodos estatísticos, ao considerarmos as componentes como variáveis aleatórias com uma distribuição uniforme. Foi possível ainda determinar o erro ou desvio padrão do número de beneficiários projetado por grupos etário e total, em diversos períodos, e, consequentemente, calcular o intervalo de confiança para a projeção.

Uma das vantagens dessa metodologia é que através dela obtemos uma medida simples de ser interpretada, o intervalo de confiança, de um número grande de possíveis trajetórias da

população, no caso 100. No presente estudo, mesmo se operacionalizássemos somente as combinações das duas componentes, mortalidade e novos beneficiários, cada uma delas com três pressupostos de comportamento futuros alternativos (um alto, um médio e um baixo), teríamos nove séries diferentes de população. A apresentação das muitas trajetórias confunde o usuário e um simples intervalo de confiança é mais fácil de ser interpretado.

A comunicação entre os profissionais também é facilitada com esta abordagem, na medida em que tornamos explícito nosso ponto de vista a respeito da modelagem das incertezas inerente aos parâmetros, ao especificarmos uma determinada distribuição de probabilidade para eles. Assim, este procedimento permite a outros demógrafos observar como estamos considerando as fontes de incertezas do fenômeno e com ela concordar ou não.

Entre os muitos métodos que poderíamos utilizar para a construção do intervalo de confiança nessa projeção, acreditamos que esse seja preferível dada a sua simplicidade, flexibilidade e a não exigência de profundos conhecimentos de matemática e estatística. No que diz respeito ao *resultado da projeção*, podemos dizer que a massa de beneficiários da Previdência Social, relativa aos três grupos de benefícios contemplados neste trabalho, deverá crescer entre 1990 e 2040 aproximadamente 679,66 %, situando-se em 2040 entre 19 956 755 e 15 348 687 com nível de confiança de 95%, caso não haja alteração no plano de benefício da Previdência Social. Grande parte desse aumento pode ser imputado às *altas taxas de crescimento da população brasileira nas faixas etárias mais velhas*, previstas para as próximas décadas, e determinadas pela sua tendência de envelhecimento. No entanto, esse não é o único fator responsável pelo acentuado crescimento da massa de aposentados. Nesse sentido, à tendência consistente de aumento da *participação da mulher no mercado de trabalho* devemos o aumento proporcionalmente maior da massa feminina de aposentados. Outro fator que também participa na sua elevação é a *queda projetada da mortalidade*. Com efeito, à medida que a mortalidade cai e a esperança de vida aumenta, as pessoas tendem a uma maior sobrevida como aposentados. No entanto, a queda da mortalidade deverá ter um peso maior no crescimento da massa feminina, uma vez que a mortalidade dos homens é mais alta nas idades de aposentadoria. Isso deve gerar relativamente mais mulheres que homens em idade avançada.

Outro resultado a ser destacado diz respeito à *mudança na composição por sexo* da massa de aposentados, ocorrida principalmente em decorrência do aumento previsto da participação da mulher no mercado de trabalho. Enquanto em 1990 a participação das mulheres no total desses três grupos de benefícios era de 29,3%, em 2040, essa percentagem se eleva para 45,31%. Essa mudança na composição deverá ter como conseqüência a elevação do número de outros grupos de benefícios da previdência, como o de pensão, e poderá gerar ainda situações em que os filhos recebam simultaneamente duas pensões, em conseqüência da morte de ambos progenitores. Também poderá ocorrer com maior freqüência a acumulação de benefício de pensão e aposentadoria, pelo marido ou

esposa. Outra conseqüência que já pode ser antecipada é o aumento das pensões decorrentes da morte de aposentados.

Tabela 1 - Limites de Variação dos Novos Beneficiários por Período, segundo Espécie de Benefício, Sexo Feminino

(continua)

LIMITES	PERÍODO					
	1990	1995	2000	2005	2010	2015
TEMPO DE SERVIÇO						
a	10 891	13 840	17 694	22 777	29 019	34 117
m	16 821	21 374	27 327	35 177	44 818	52 691
b	39 403	50 069	64 013	82 403	104 986	123 430
ESPECIAL						
a	1 923	2 442	3 130	4 032	5 006	5 754
m	3 169	4 025	5 159	6 646	8 251	9 484
b	5 279	6 704	8 593	11 070	13 744	15 797
IDADE						
a	45 554	60 422	79 831	100 465	125 972	166 016
m	54 303	72 027	95 164	119 761	150 168	197 903
b	84 220	111 708	147 591	185 738	232 897	306 929

(conclusão)

LIMITES	PERÍODO				
	2020	2025	2030	2035	2040
TEMPO DE SERVIÇO					
a	39 615	43 702	47 557	51 444	54 877
m	61 182	67 494	73 447	79 450	84 752
b	143 320	158 106	172 050	186 112	198 532
ESPECIAL					
a	6 521	7 090	7 648	8 156	8 620
m	10 749	11 687	12 606	13 443	14 208
b	17 905	19 466	20 998	22 392	23 666
IDADE					
a	215 047	269 755	297 592	322 812	370 210
m	256 350	321 566	354 749	384 813	441 315
b	397 577	498 721	550 184	596 811	684 440

Tabela 2 - Limites de variação dos novos beneficiários por período, segundo espécie de benefício, sexo feminino

(Continua)

LIMITES	PERÍODO					
	1990	1995	2000	2005	2010	2015
a	69 536	78 685	91 758	109 380	131 617	151 027
m	78 503	88 831	103 589	123 484	148 588	170 501
b	121 788	137 811	160 707	191 571	230 517	264 513
ESPECIAL						
a	23 788	27 026	31 950	38 313	45 505	50 978
m	31 568	35 865	42 399	50 843	60 388	67 651
b	46 029	52 295	61 822	74 134	88 051	98 641
IDADE						
a	32 516	38 338	41 711	50 371	55 031	67 624
m	38 431	45 312	49 298	59 534	65 041	79 925
b	59 406	70 042	76 204	92 026	100 540	123 547

(conclusão)

LIMITES	PERÍODO				
	2020	2025	2030	2035	2040
a	173 557	191 904	205 982	216 605	225 165
m	195 936	216 649	232 543	244 536	254 199
b	303 972	336 106	360 763	379 369	394 361
ESPECIAL					
a	57 187	62 036	65 900	68 569	70 788
m	75 890	82 325	87 453	90 995	93 939
b	110 655	120 038	127 515	132 680	136 973
IDADE					
a	83 102	108 460	133 522	146 926	156 368
m	98 218	128 188	157 809	173 650	184 811
b	151 824	198 152	243 939	268 426	285 678

Tabela 3 - Taxa de crescimento médio anual de cada período, por período, segundo a espécie do benefício

(continua)

ESPÉCIE DO BENEFÍCIO	PERÍODO				
	1990-1995	1995-2000	2000-2005	2005-2010	2010-2015
T.S	4,91	5,04	5,18	4,96	3,29
ESPECIAL	4,90	5,09	5,20	4,42	2,82
IDADE	5,81	5,73	4,71	4,63	5,68
T.S	2,50	3,12	3,58	3,77	2,79
ESPECIAL	2,59	3,40	3,70	3,50	2,30
IDADE	3,35	1,70	3,85	1,79	4,21

(conclusão)

ESPÉCIE DO BENEFÍCIO	PERÍODO				
	2015-2020	2020-2025	2025-2030	2030-2035	2035-2040
T.S	3,03	1,98	1,70	1,58	1,30
ESPECIAL	2,54	1,69	1,53	1,29	1,11
IDADE	5,31	4,64	1,98	1,64	2,78
T.S	2,82	2,03	1,43	1,01	0,78
ESPECIAL	2,33	1,64	1,22	0,80	0,64
IDADE	4,21	5,47	4,25	1,93	1,25

Tabela 4 - Taxa de crescimento médio anual de cada período, por período, segundo a espécie de benefício e sexo

LIMITES	PERÍODO									
	1990	1995	2000	2005	2010	2015	2020	2025	2030	2035
HOMENS										
a	61,02	61,16	61,29	61,70	62,16	62,64	63,13	63,63	64,13	64,63
m	62,07	63,31	64,48	65,61	66,72	67,80	68,86	69,89	70,90	71,88
b	63,10	65,24	67,47	69,22	70,87	72,43	73,93	75,36	76,74	78,07
MULHERES										
a	67,73	67,75	67,83	68,26	68,73	69,22	69,72	70,22	70,72	71,21
m	68,99	70,29	71,52	72,67	73,76	74,78	75,75	76,67	77,55	78,38
b	70,19	72,33	74,62	76,26	77,71	79,04	80,26	81,40	82,47	83,48

Tabela 5 - Razões de sobrevivência que compõem a matriz de transição básica (PO) , por sexo

GRUPOS DE IDADE	SEXO	
	MASCULINO	FEMININO
0	0,97320	0,98575
5	0,99245	0,99585
10	0,99288	0,99588
15	0,98938	0,99375
20	0,98553	0,99124
25	0,98188	0,98849
30	0,97798	0,98512
35	0,97301	0,98063
40	0,96615	0,97444
45	0,95632	0,96567
50	0,94231	0,95329
55	0,92205	0,93530
60	0,88904	0,90568
65	0,83732	0,85800
70	0,76331	0,78689
P(75,W)	0,52680	0,54773

Tabela 6 - Totais da projeção média dos beneficiários de aposentadoria por tempo de serviço, limite Superior – LS – e limite inferior – LI – do intervalo de confiança de 95%, percentagem da metade da amplitude dos limites e crescimento entre os anos, segundo o ano – posição em dezembro de cada ano – e sexo

Ano	P. Média	LS	LI	((LS-LI)/2)/ P. Médio	Taxa de Crescimento P. Médio (%)
HOMENS					
1990	970 769				
1995	1 288 560	1 461 445	1 115 675	13,42	32,74
2000	1 613 319	1 834 901	1 391 737	13,73	25,20
2005	1 947 507	2 189 724	1 705 290	12,44	20,71
2010	2 335 418	2 663 657	2 007 179	14,05	19,92
2015	2 810 008	3 195 145	2 424 871	13,71	20,32
2020	3 323 095	3 760 894	2 885 296	13,17	18,26
2025	3 825 217	4 321 645	3 328 789	12,98	15,11
2030	4 379 445	4 957 968	3 800 922	13,21	14,49
2035	4 842 565	5 391 180	4 293 950	11,33	10,57
2040	5 229 736	5 811 027	4 648 445	11,12	8,00
MULHERES					
1990	156 151				
1995	259 545	352 425	166 666	35,79	66,21
2000	380 245	508 786	251 703	33,81	46,50
2005	515 807	663 878	367 736	28,71	35,65
2010	686 687	902 184	471 189	31,38	33,13
2015	912 346	1 173 565	651 128	28,63	32,86
2020	1 152 127	1 454 259	849 995	26,22	26,28
2025	1 357 016	1 697 590	1 016 442	25,10	17,78
2030	1 586 238	1 967 534	1 204 942	24,04	16,89
2035	1 748 270	2 097 066	1 399 474	19,95	10,21
2040	1 863 645	2 226 687	1 500 603	19,48	6,60
TOTAL					
1990	1 126 920				
1995	1 548 105	1 813 870	1 282 341	17,17	37,37
2000	1 993 564	2 343 687	1 643 440	17,56	28,77
2005	2 463 314	2 853 602	2 073 026	15,84	23,56
2010	3 022 105	3 565 841	2 478 368	17,99	22,68
2015	3 722 354	4 368 710	3 075 999	17,36	23,17
2020	4 475 222	5 215 152	3 735 292	16,53	20,23
2025	5 182 233	6 019 235	4 345 231	16,15	15,80
2030	5 965 683	6 925 502	5 005 864	16,09	15,12
2035	6 590 835	7 488 247	5 693 423	13,62	10,48
2040	7 093 381	8 037 714	6 149 048	13,31	7,62

Tabela 7 - Totais de projeção média dos beneficiários de aposentadoria especial, limite superior – LS – e limite – LI – do intervalo de confiança de 95%, percentagem da metade da amplitude dos limites e crescimento entre os anos , segundo o ano – posição em dezembro de cada ano – e sexo

Ano	P. Média	LS	LI	((LS-LI)/2)/ P. Médio	Taxa de Crescimento P. Médio (%)
HOMENS					
1990	276 675				
1995	427 432	495 312	359 552	15,88	54,49
2000	583 923	674 195	493 652	15,46	36,61
2005	750 219	852 385	648 052	13,62	28,48
2010	936 520	1 075 180	797 859	14,81	24,83
2015	1 151 317	1 308 789	993 845	13,68	22,94
2020	1 366 891	1 545 842	1 187 940	13,09	18,72
2025	1 563 323	1 768 790	1 357 856	13,14	14,37
2030	1 777 097	2 002 121	1 552 073	12,66	13,67
2035	1 953 413	2 175 120	1 731 706	11,35	9,92
2040	2 096 253	2 334 710	1 857 796	11,38	7,31
MULHERES					
1990	17 680				
1995	35 785	45 953	25 618	28,41	102,41
2000	57 068	71 714	42 422	25,66	59,47
2005	82 003	99 616	64 391	21,48	43,69
2010	112 172	137 492	86 852	22,57	36,79
2015	149 242	178 534	119 950	19,63	33,05
2020	187 336	221 333	153 339	18,15	25,52
2025	220 906	260 161	181 651	17,77	17,92
2030	256 338	297 892	214 784	16,21	16,04
2035	283 643	323 806	243 480	14,16	10,65
2040	304 055	347 096	261 014	14,16	7,20
TOTAL					
1990	294 355				
1995	463 218	541 265	385 170	16,85	57,37
2000	640 991	745 908	536 074	16,37	38,38
2005	832 222	952 002	712 442	14,39	29,83
2010	1 048 692	1 212 672	884 712	15,64	26,01
2015	1 300 559	1 487 323	1 113 795	14,36	24,02
2020	1 554 227	1 767 175	1 341 279	13,70	19,50
2025	1 784 229	2 028 951	1 539 507	13,72	14,80
2030	2 033 435	2 300 013	1 766 857	13,11	13,97
2035	2 237 056	2 498 926	1 975 186	11,71	10,01
2040	2 400 308	2 681 805	2 118 811	11,73	7,30

Tabela 8 - Totais de projeção média dos beneficiários de aposentadoria por idade, limite superior – LS – e limite – LI – do intervalo de confiança de 95%, percentagem da metade da amplitude dos limites e crescimento entre os anos, segundo o ano – posição em dezembro de cada ano – e sexo

Ano	P. Média	LS	LI	((LS-LI)/2)/ P. Médio	Taxa de Crescimento P. Médio (%)
HOMENS					
1990	353 489				
1995	479 201	562 236	396 166	17,33	35,56
2000	598 840	706 353	491 328	17,95	24,97
2005	680 180	783 695	576 664	15,22	13,58
2010	799 188	943 544	654 831	18,06	17,50
2015	914 240	1 064 541	763 939	16,44	14,40
2020	1 092 228	1 279 751	904 705	17,17	19,47
2025	1 297 337	1 525 734	1 068 940	17,61	18,78
2030	1 660 049	1 974 727	1 345 371	18,96	27,96
2035	2 037 313	2 363 890	1 710 736	16,03	22,73
2040	2 327 775	2 676 366	1 979 184	14,98	14,26
MULHERES					
1990	489 384				
1995	747 876	876 253	619 498	17,17	52,82
2000	1 049 329	1 236 258	862 400	17,81	40,31
2005	1 386 883	1 605 287	1 168 479	15,75	32,17
2010	1 781 841	2 092 743	1 470 939	17,45	28,48
2015	2 318 377	2 706 558	1 930 196	16,74	30,11
2020	3 013 255	3 510 405	2 516 105	16,50	29,97
2025	3 787 209	4 409 062	3 165 356	16,42	25,68
2030	4 662 324	5 401 902	3 922 746	15,86	23,11
2035	5 318 116	6 000 997	4 635 235	12,84	14,07
2040	5 831 257	6560869,2	5101644,8	12,51	9,65
TOTAL					
1990	842 873				
1995	1 227 077	1 438 490	1 015 664	17,23	45,58
2000	1 648 169	1 942 611	1 353 728	17,86	34,32
2005	2 067 063	2 388 983	1 745 143	15,57	25,42
2010	2 581 029	3 036 288	2 125 770	17,64	24,86
2015	3 232 617	3 771 100	2 694 135	16,66	25,25
2020	4 105 483	4 790 156	3 420 810	16,68	27,00
2025	5 084 546	5 934 796	4 234 296	16,72	23,85
2030	6 322 373	7 376 629	5 268 117	16,68	24,34
2035	7 355 429	8 364 887	6 345 971	13,72	16,34
2040	8 159 032	9 237 235	7 080 829	13,21	10,93

Referências Bibliográficas

ALHO, J. M., AND SPENCER, B. D. (1991), "A Population Forecast as a Database: Implementing the Stochastic Propagation of Error." *Jornal of official Statistics*, 7: 295-310.

ALHO, J. (1990), "Stochastic Methods in Population Forecasting." *International Journal of forecasting* 6: 521-530.

BERNARDELLI, H. (1941), "Population Waves." *Journal of Burma Research Society*, 31: 1-18.

CARVALHO, J. A. M. (1993). *Crescimento populacional e estrutura demográfica no Brasil.* Belo Horizonte. CEDEPLAR/UFMG.

CARVALHO, J. A. M. E, WONG, L. (1995). "A window of opportunity: some demographic and socioeconomic implication of the rapid fertility decline in Brazil". *"Texto para discussão no. 91".* Belo Horizonte. CEDEPLAR/UFMG.

COHEN, J.E. (1986), "Population Forecasts and Confidence Intervals for Sweden: A comparison of model model-based and empirical approches." *Demography*, 23: 105-126.

DAPONTE, B. O; KADANE, J. B.; WOLFSON, L. J. (1995), "Bayesian Demography: Projecting the Iraqi Kurdish Population, 1977-1990." paper presented in the meeting of the Population Association of American, San Francisco.

FÍGOLI, M.B.G. (1997), Intervalo de Confiança para Projeção de População: Projeção dos Beneficiários Urbanos da Previdência Social (1990-2040). Tese de doutorado. DEDEPLAR/UFMG

FÍGOLI, M.B.G. (1998), " Modelando e projetando a mortalidade no Brasil*" Revista Brasileira de Estudos da População* v. 15 n.1 Jan/Jun

HAMMERSLEY, J. M. AND HANDSCOMB, D. C. (1964). *Monte Carlo Methods*, Chapman & Hall. New York.

HYPPÖLÄ,J., TUNKELO, A., AND TÖRNQVIST. L. (1949), "Calculations on the Population of Finland, Its Renewal, and its Future Development." Tilastollisia Tiedonantoja 38. Helsinki: Central Statistical Office of Finland. (In Finnish)

KEYFITZ , N. (1987), "The Social and Political Context of Population Forecasting."*Reading in Population Methodology*, vol. 5, cap. 17, D.Bogue, E. Arriaga and D.Anderson, eds. Chicago, Social Development Center.

KEYFITZ, N. (1984), *Demography Through Problems.*Spring-Verlag. New York.

KEYFITZ, N. (1982), "Choice of function for mortality analysis: Effective forecasting depends on a minimum parameter representation." *Theoretical Population Biology*, 21: 329-352.

KEYFITZ, N. (1981), "The Limits of Population Forecasting." *Population and Development Review*, 7: 579-593.

KEYFITZ, N. (1977), *Applied Mathematical Demography.* Spring -Verlag, New York.

LAND, K. C. (1986), "Methods for National Population Forecasts: A Review." *Jornal of the American Statistical Association*, 81 (396): 888-901.

LEE, R. D. AND CARTER L. R. (1992), "Modeling and Forecasting the time Series of U. S. Mortality." *Journal of the American Association*, 87: 659-671.

LEE, R. D. (1992), "Stochastic Demograph Forecasting." *International Journal of Forecasting*, 8: 315-327.

LEE, R. D. (1974), "Forecasting Births in Post-Transitional Population: Sochastic Renewal With Serially Correlated Fertility ." *Journal of Statistical Association*, 69: 607-617.

LEE, R. D. AND TULJAPURKAR, S. (1994), "Stochastic Population Forecasts for the United States: Beyond High, Medium and low." *Journal of the American Statistic Association*, 89: 1175-2289.

LESLIE, P. H. (1945), "On the use of matrices in certain population mathematics." *Biometrika*, 33: 183-212.

LESLIE, P. H. (1948), " Some further notes on the use of matrices in certain population mathematics." *Biometrika*, 35: 213-245.

LEWIS, E. G. (1942), "On the generation and growth of a population ". *Sankhya, Indian Journal of Statistics*, 6: 93-96.

MACHADO, C. CAETANO (1993), Projeções Multiregionais de População: o caso brasileiro. Tese de doutorado. CEDEPLAR/UFMG

MCDONALD , J. (1981), "Modeling Demographic Relationships: An Analysis of Forecast Functions for Australian Births." *Jornal of American Statistical Association*, 76: 782-792

NAYLOR, T. J., BALINTFY, J.L., BURDICK, D.S. AND CHU, K. (1966), *Computer Simulation Techniques*. Wiley, New York.

PFLAMER, P. (1986a), "Stochastische Bevolkerungsmodelle zur Analyse der Auswirkungen Demographischer Prozesse auf die Systeme der Sozialen Sicherung." *Allgemeines Statistisches Archiv* 70: 52-74.

PFLAMER, P. (1986b), *Bevolkerung, Haushalte und Konsun: Statistische Analyse und Prognose.*Campus Verlag, Frankfurt-New York.

PFLAMER, P. (1988), " Confidencial Intervals for Population Projections Based on Monte Carlo Methods." *International Journal of Forecasting*, 4: 135-142

PFLAMER, P. (1992), "Forecasting US Population Totals With the Box-Jenkins Approch." *International Journal of Forecasting*, 8: 329-338.

POLLARD, A. H. (1990), *Demography Techniques*. Third Edition. Pergamon Press.

RIOS-NETO, E. L. G. E WAJNMAN, S. (1994), "Participação Feminina no Mercado de Trabalho no Brasil: Elementos para Projeção de Níveis e Tendências ". *Anais do Encontro de Estudos Populacionais* , Vol 3. Caxambu.

SABOIA, J.L.M., (1974), "Modeling and Forecasting Population by Time Series: The Swedish case." *Demography*, 2: 483-492.

SHRYOCK,H.S.,J.S. SIEGEL ET AL., CONDENSED EDITION BY E. G. STOCKWELL (1976), *The Methods an Materials of Demography*, Academic Press, New York.

SYKES, Z.M. (1969), "Some Stochastic Versions of the Matrix Model for Population Dynamics." *Jornal of American Statistical Association*, 44: 111-130.

TULJAPURKAR, SHRIPAD . (1990), *Population Dynamics in Variable Environments*. Lecture Notes In Biomathematics. Spring-Verlag.

TULJAPURKAR, S. (1992), "Stochastic Population Forecasts and their Uses." *International Journal of Forecasting* , 8: 385-392

WHELPTON, P. K., (1928), "Population of the United States, 1925 to 1975." *American Journal of Sociology*, 343: 253-270.

ABSTRACT

This paper applies Monte Carlo's simulation method for forecasting urban beneficiary of Social Security in Brazil between 1990 e 2040. The benefits in study are: time of work , age and special. This Technique attempts to incorporate the uncertainty involved in projecting population by letting the number of new beneficiary and mortality rates vary as a random variable with uniform distribution. This assumptions make it possible to construct confidence intervals for future populations.

Considering the results of the model for Brazil in 2040, for example, it is shown that the beneficiary will number between 19.956.755 and 15.348.687, with a probability of 95%.

Análise da Confiabilidade de Itens Submetidos a Testes Acelerados Via Simulação Estocástica: O Efeito da Ortogonalização de Parâmetros

Néli Maria Costa Mattos*

Hélio S. Migon**

RESUMO

Para se analisar a performance de itens cujo tempo para falhar, sob condições normais, é longo, faz-se uso de testes acelerados. O estudo da confiabilidade desses itens tem sido amplamente discutido na literatura, considerando-se um modelo estresse-resposta e especialmente a distribuição exponencial para os tempos de vida. Quando o enfoque é Bayesiano, surgem dificuldades com integrais que não possuem solução analítica e precisam ser resolvidas numericamente. Para contornar esse problema, propomos o uso do amostrador de Gibbs para analisar a confiabilidade de itens submetidos a testes acelerados com censura tipo II. Sugerimos uma transformação para ortogonalizar os parâmetros, objetivando a aceleração da convergência das cadeias. Uma aplicação numérica é apresentada e os resultados discutidos.

1. Introdução

Atualmente, a qualidade de produtos e serviços tem sido motivo de preocupação e, conseqüentemente, alvo de pesquisa. Para que se possa avaliar a qualidade é necessário quantificá-la; isto é feito pela obtenção de medidas de confiabilidade, manutenibilidade, disponibilidade, etc., através de testes em amostras aleatórias de itens produzidos.

Quando itens têm tempo de vida longo, faz-se uso de testes acelerados e coleta de dados sob esquema de censura (Nelson (1990)). Nesses testes, os itens são submetidos a um estresse mais

* Endereço para correspondência: IME/RJ & COPPE/UFRJ - e-mail: neli@leblon.ime.eb.br.

** Endereço para correspondência: IM & COPPE/UFRJ - e-mail: migon@pep.ufrj.br.

severo de operação do que o especificado para o seu funcionamento em condições normais; a censura determina o término do teste antes de serem observadas as falhas de todos os itens; a do tipo I fixa o término por tempo, e a do tipo II, por um número de falhas observadas.

Entre as publicações abordando testes acelerados sob o enfoque Bayesiano, citamos a de Louzada-Neto e Achcar (1993), na qual utilizaram a aproximação de Laplace para estimar as densidades *a posteriori* dos parâmetros de interesse.

Neste trabalho, temos como objetivo conduzir a análise Bayesiana da confiabilidade utilizando o amostrador de Gibbs, Gilks et al. (1996), que consiste de um método de simulação estocástica via cadeias de Markov, e explorar o efeito da ortogonalização para acelerar a convergência das cadeias geradas.

Entre as aplicações precursoras do uso do amostrador de Gibbs na análise da confiabilidade, podemos citar Dellaportas & Smith (1993).

Para ilustrar o procedimento proposto, apresentamos umas aplicações com os dados gerados por Louzada-Neto e Achcar (1993), segundo um modelo para testes acelerados.

Convém ressaltar que para a análise de sobrevivência e para a análise da confiabilidade os procedimentos são análogos e a modelagem similar. Desta forma, a teoria apresentada neste artigo pode ser adaptada para a análise de sobrevivência com aplicações na área social, por exemplo, em mercado de trabalho (Frechet et al., 1992 e Arrow, 1996). Além disso, as técnicas de ortogonalização são úteis para modelos de regressão em geral, com aplicação em qualquer área.

Este trabalho é organizado na forma a seguir: na seção 2, a distribuição exponencial é descrita, assumindo um modelo estresse-resposta com dados coletados sob esquema de censura tipo II. A estimação utilizando os parâmetros originais e uma transformação para encontrar parâmetros ortogonais é apresentada na seção 3. Algumas considerações sobre a implementação de métodos de MCMC ("Markov Chain Monte Carlo") são discutidas na seção 4. Na seção 5, uma aplicação numérica é apresentada e são comentados os resultados. Conclusões são encontradas na seção 6.

2. Modelagem

Suponha que se tenha n_i itens com tempos de falha T_i a serem testados sob o nível de estresse $V_i, i = 1,2,..,k$. Seja T_i uma variável aleatória com densidade exponencial dada por:

$$f(t|\lambda_i) = \lambda_i \exp[-(\lambda_i t)], \quad t > 0, \quad \lambda_i > 0, \quad i = 1,2..,k, \tag{2.1}$$

onde k representa o número de distintos níveis de estresse e λ_i a taxa de falha constante sob o nível de estresse V_i.

Utilizando-se o esquema de censura tipo II (o teste termina quando r_i itens do grupo i falham), tem-se como dados, os tempos de falha observados e ordenados: t_{ij}, $i=1,..,k$, $j=1,..,r_i$ e os tempos de falha censurados: t_{ij}, $i=1,..,k$, $j=r_{i+1},..,n_i$, maiores que t_{ir_i}.

Dessa forma, a função de verossimilhança para λ_i é dada por:

$$L(\lambda_i)=\prod_{i=1}^{k}\left\{\lambda_i^{r_i}\exp\left(-\lambda_i\left[\sum_{j=1}^{r_i}t_{ij}+(n_i-r_i)t_{ir_i}\right]\right)\right\}=\prod_{i=l}^{k}\left(\lambda_i^{r_i}\exp^{-\lambda_i A_i}\right),$$

onde $A_i=\sum_{j=1}^{r_i}t_{ij}+(n_i-r_i)t_{ir_i}$.

Supõe-se também um modelo estresse-resposta que relacione a taxa de falha falha λ_i e a variável de estresse V_i, através de:

$$\lambda_i=\exp^{\{-(z_i+\beta_0+\beta_1 X_i)\}},\quad i=1,2,\ldots,k, \tag{2.2}$$

onde $X_i=h(V_i)$, $Z_i=g(X_i)$ e $-\infty<\beta_0,\beta_1<\infty$.

Dessa forma, tem-se a função de verossimilhança para β_0 e β_1:

$$L(\beta_0,\beta_1)\,\alpha\exp\left(-\beta_0 r-\beta_1 a-e^{-\beta_0}\sum_{i=1}^{k}A_i e^{-Z_i-\beta_1 X_i}\right), \tag{2.3}$$

onde $r=\sum_{i=1}^{k}r_i$ e $a=\sum_{i=1}^{k}r_i X_i$.

Casos particulares do modelo estresse-resposta incluem:

1) modelo de potência:

$$X_i=-\log V_i,\quad Z_i=0. \tag{2.4}$$

2) modelo de Arrhenius:

$$X_i=-1/V_i,\quad Z_i=0. \tag{2.5}$$

3) modelo de Eyring:

$$X_i = -1/V_i, \quad Z_i = -\log V_i. \tag{2.6}$$

3. Estimação dos parâmetros

Para conduzir a análise Bayesiana é necessário fazer uso de aproximações numéricas. Uma alternativa muito utilizada é a aproximação de Laplace (Louzada-Neto e Achcar (1993)). Nesse artigo, os autores reparametrizaram o modelo definido por (2.1) e (2.2), substituindo β_0 por uma função de $\theta_s = 1/\lambda_s$, tempo médio de falha dos itens submetidos a um particular nível de estresse V_s. Para encontrar o estimador de máxima verossimilhança de β_1, que nesse caso é estimado na presença do parâmetro de distúrbio θ_s, necessitaram transformar (θ_s, β_1) em (ψ_s, β_1), parâmetros ortogonais. Assumindo *a priori* de Jeffreys para (ψ_s, β_1) e utilizando a aproximação de Laplace chegaram a formas fechadas para as densidades *a posteriori* de interesse.

A seguir, apresentamos a análise Bayesiana da confiabilidade para o modelo (2.1) e (2.2), via amostrador de Gibbs.

3.1 - O amostrador de Gibbs

Este método consiste em um procedimento iterativo de simulação de Monte Carlo via cadeia de Markov e é aplicável a situações onde não se consegue gerar amostras diretamente da densidade conjunta *a posteriori*. Essa densidade é construída a partir das densidades *a priori* dos parâmetros e da função de verossimilhança. A partir da densidade conjunta, as densidades condicionais completas requeridas pelo amostrador de Gibbs para a geração de amostras podem ser facilmente obtidas sem a constante de proporcionalidade. Como será mostrado na seção 4, no modelo exponencial (2.1) e (2.2), as densidades condicionais completas são log-côncavas e um método de amostragem de rejeição adaptativa (Gilks e Wild (1992)), pode ser utilizado.

Nos métodos de Monte Carlo via cadeias de Markov, as densidades *a posteriori* de qualquer função dos parâmetros podem ser facilmente obtidas a partir das amostras geradas. Uma questão crítica nesses métodos consiste em determinar o ponto de parada das iterações, o número de iterações a serem descartadas *burn-in* e o intervalo entre iterações a serem aproveitadas (lag) para a indicação da convergência. Entre os pesquisadores não existe um consenso nesse sentido. Por esta razão ferramentas adicionais devem ser utilizadas para monitorar a convergência e se possível acelerá-la. Na ausência de técnicas gerais para se determinar esses números *a priori*, análises estatísticas devem ser realizadas *a posteriori*. Estes procedimentos são denominados diagnósticos de convergência.

Para a utilização do amostrador de Gibbs, a ortogonalidade dos parâmetros não é necessária, porém, se existe alta correlação entre os mesmos, a cadeia caminha lentamente, exigindo um grande

número de iterações para a indicação de convergência. Nesse caso, uma reparametrização para ortogonalizar os parâmetros é sugerida. Com esse procedimento, a cadeia se move livremente através de todo o espaço paramétrico e o número de iterações exigidas para a indicação da convergência se reduz sensivelmente (Tabela 3 na aplicação numérica).

3.2 - O amostrador de Gibbs na parametrização original

Sejam β_0 e β_1 os parâmetros originais do modelo. Denotando as densidades *a posteriori* por π_1, segue-se:

$$\pi_1(\beta_0,\beta_1) \propto \pi_0(\beta_0)\pi_0(\beta_1)L(\beta_0,\beta_1),$$

onde L é a função de verossimilhança e $\pi_0(\beta_0)$ e $\pi_0(\beta_1)$ são as densidades *a priori* para β_0 e β_1, assumidas normais de média zero e variâncias σ_1^2 e σ_2^2, respectivamente.

Dessa forma, tem-se para a verossimilhança (2.3) a densidade conjunta *a posteriori*:

$$\pi_1(\beta_0,\beta_1) \propto \exp\left[-\beta_0 r - \beta_1 a - e^{-\beta_0}\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)} - \frac{\beta_0^2}{2\sigma_1^2} - \frac{\beta_1^2}{2\sigma_2^2}\right]. \tag{3.1}$$

A partir de (3.1), obtêm-se as densidades condicionais completas para β_0 e β_1:

$$\pi_1(\beta_0,\beta_1) \propto \exp\left(-\beta_0 r - \frac{\beta_0^2}{2\sigma_1^2}\right)\exp\left[-e^{-\beta_0}\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}\right], \tag{3.2}$$

$$\pi_1(\beta_1|\beta_0) \propto \exp\left(-\beta_1 a - \frac{\beta_0^2}{2\sigma_2^2}\right)\exp\left[-e^{-\beta_0}\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}\right]. \tag{3.3}$$

onde $r = \sum_{i=1}^{k} r_i, \quad a = \sum_{i=1}^{k} r_i X_i$.

Na próxima seção, apresentaremos procedimentos para transformar (β_0,β_1) em (ϕ,β_1), parâmetros ortogonais, e a utilização do amostrador de Gibbs no modelo (2.3) reparametrizado.

3.3 - O amostrador de Gibbs com parâmetros ortogonais

Para o modelo dado por (2.3), seja I a matriz de informação de Fisher com respeito a β_0 e β_1:

$$I(\beta_0, \beta_1) = \begin{bmatrix} r & a \\ a & a_1 \end{bmatrix} \tag{3.4}$$

onde $r = \sum_{i=1}^{k} r_i$, $a = \sum_{i=1}^{k} r_i X_i$ e $a_1 = \sum_{i=1}^{k} r_i X_i^2$.

O parâmetro ϕ ortogonal a β_1 pode ser obtido, (Cox e Reid (1987)) resolvendo-se a equação diferencial:

$$-I_{\beta_0\beta_1} = I_{\beta_0\beta_0} \frac{\partial \beta_0}{\partial \beta_1}. \tag{3.5}$$

De (3.4) e (3.5), tem-se:

$$-a = r \frac{\partial \beta_0}{\partial \beta_1}. \tag{3.6}$$

Uma solução de (3.6) é dada por:

$$\beta_0 = -\frac{a}{r}\beta + c(\phi),$$

onde $c(\phi)$ é uma função arbitrária de ϕ. Considerando $c(\phi) = \phi$, tem-se:

$$\beta_0 = -\frac{a}{r}\beta_1 + \phi. \tag{3.7}$$

Substituindo-se (3.7) na verossimilhança (2.3), tem-se:

$$L(\phi, \beta_1) \propto \exp\left\{-r\phi - \exp\left(\frac{a\beta_1}{r} - \phi\right)\sum_{i=1}^{k} A_i \exp[-(Z_i + \beta_1 X_i)]\right\}. \tag{3.8}$$

Associando aos parâmetros de interesse ϕ e β_1 distribuições *a priori* normais de média zero e variâncias σ_3^2 e σ_4^2, respectivamente, tem-se a densidade conjunta *a posteriori* para ϕ e β_1:

$$\pi_1(\phi,\beta_1) \propto \exp\left[-r\phi - \exp\left(\frac{a\beta_1}{r} - \phi\right)\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)} - \frac{\phi}{2\sigma_3^2} - \frac{\beta_1}{2\sigma_4^2}\right]. \tag{3.9}$$

A partir de (3.9), têm-se as densidades condicionais completas para ϕ e β_1:

$$\pi_1(\phi|\beta_1) \propto \exp\left[-r\phi - \frac{\phi}{2\sigma_3^2} - \exp\left(\frac{a\beta_1}{r} - \phi\right)\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}\right], \tag{3.10}$$

$$\pi_1(\beta_1|\phi) \propto \exp\left[-\frac{\beta_1}{2\sigma_4^2} - \exp\left(\frac{a\beta_1}{r} - \phi\right)\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}\right], \tag{3.11}$$

4. Implementação

Caso as densidades condicionais completas não pertençam a nenhuma família paramétrica conhecida, uma alternativa para a geração de amostras é utilizar um algoritmo de rejeição adaptativa (Gilks e Wild (1992)), sendo necessário, então, verificar a log-concavidade dessas densidades.

Uma função positiva f definida em um conjunto aberto C em R^n é log-côncava se *log* f é uma função real duas vezes continuamente diferenciável em C, com matriz Hessiana:

$$H = \left(\frac{\partial^2 \log f(x_1,\ldots x_n}{\partial x_i \partial x_j}\right), \tag{4.1}$$

semidefinida negativa para todo $x \in C$. Se a matriz Hessiana é definida negativa, a função f é dita estritamente log-côncava, (Dellaportas e Smith (1992)). Como as condicionais completas dadas por (3.2) e (3.3) são unidimensionais, as Hessianas da definição acima se reduzem a:

$$\frac{\partial^2 \log}{\partial \beta_0^2}\pi_1(\beta_0|\beta_1) = -\frac{1}{\sigma_1^2} - e^{-\beta_0}\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}, \tag{4.2}$$

$$\frac{\partial^2 \log}{\partial \beta_1^2}\pi_1(\beta_1|\beta_0) = -\frac{1}{\sigma_2^2} - e^{-\beta_0}\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}(-X_i)^2. \tag{4.3}$$

Avaliando os termos das expressões (4.2) e (4.3), verifica-se que essas expressões são negativas, caracterizando a log-concavidade.

Analogamente, para ϕ e β_1 temos :

$$\frac{\partial^2 \log}{\partial \phi^2}\pi_1(\beta_\phi|\beta_1) = \left[-\exp\left(\frac{a}{r}\beta_1 - \phi\right)\right]\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}, \qquad (4.4)$$

$$\frac{\partial^2 \log}{\partial \beta_1^2}\pi_1(\beta_1|\phi) = \left[-\exp\left(\frac{a}{r}\beta_1 - \phi\right)\right]\sum_{i=1}^{k} A_i e^{-(Z_i+\beta_1 X_i)}\left(X_i - \frac{a}{r}\right)^2, \qquad (4.5)$$

que também sendo negativas caracterizam a log-concavidade.

5. Aplicação e resultados

Nesta seção, apresentaremos os resultados obtidos pela análise Bayesiana da confiabilidade de itens submetidos a testes acelerados com censura tipo II via amostrador de Gibbs, implementados no software BUGS - *Bayesian Inference Using Gibbs Sampling*, Spiegehalter et al. (1997).

Este software permite monitorar qualquer função dos parâmetros estimados; se não reconhece conjugação nas condicionais completas, utiliza um algoritmo de rejeição adaptativa para a geração das amostras. Por essa razão, a log-concavidade dessas densidades é uma característica exigida pelo BUGS e foi verificada acima para as densidades dadas pelas expressões (4.2), (4.3), (4.4) e (4.5).

Os dados (tempos de falha) utilizados nesta aplicação foram gerados por Louzada-Neto e Achcar (1993), segundo uma distribuição exponencial, considerando um modelo estresse-resposta de potência dado por (2.2) e (2.4), com β_0 = 5.7366 e β_1 = 0.6. Estes dados são apresentados na Tabela 1.

São considerados $k = 3$ níveis de estresse: $V_1 = 4$ (nível normal), $V_2 = 10$ e $V_3 = 20$; para cada nível, foram testados $n_1 = 10$ itens, $i = 1,2,3$.

A censura foi feita no r_i-ésimo item a falhar em cada nível: $r_1 = 4$, $r_2 = 6$ e $r_3 = 8$.

Tabela 1 – Dados gerados

i	n_1	r_i	V_i	TEMPOS DE FALHA
1	10	4	04	18, 31, 47, 61, 61, 61, 61, 61, 61, 61
2	10	6	10	12,19, 34, 43, 52, 60, 60, 60, 60, 60
3	10	8	20	15, 24, 36, 37, 42, 43, 46, 57, 57, 57

Obs.: Os dados sublinhados são dados censurados.

Numa primeira etapa, foram considerados os parâmetros originais β_0 e β_1, com densidades *a priori* vagas normais com $\mu = 0$ e $\sigma^2 = 10^4$ com valores iniciais iguais a zero. Para uma corrida de 4000 iterações, foi constatada uma alta correlação entre β_0 e β_1 igual a 0.968. Esta alta correlação sugeriu que, numa segunda etapa, fizéssemos uma transformação de (β_0, β_1) para (ϕ, β_1) parâmetros ortogonais, e implementássemos o amostrador de Gibbs utilizando esses novos parâmetros.

Em ambos os casos, foram ainda monitoradas e representadas graficamente as seguintes distribuições *a posteriori* para o i-ésimo nível de estresse, $i = 1,2,3$:

i) tempo médio para falhar:

$$\theta_i = \frac{1}{\lambda_i}, \qquad i = 1,2,3;$$

ii) densidade preditiva (tempo para a próxima falha):

$$t[i,11] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_i(t_i|\beta)\pi_1(\beta_0,\beta_1)\partial\beta_0\partial\beta_1, \qquad i = 1,2,3; \text{ e}$$

iii)confiabilidade para os tempos w:

$$S[i,w] = \exp(-\lambda_i w), \qquad i = 1,2,3, \qquad w = 30,60,90.$$

Os resultados obtidos são apresentados na Tabela 2 (colunas III e IV) e nas Figuras de 1 a 9, e podem ser comparados com os valores utilizados para a geração dos dados (coluna I) e com as estimativas obtidas por Louzada-Neto e Achcar (1993) (coluna II), que não explicitaram a estimativa de β_0. Houve diferença A entre as estimativas.

Tabela 2 - Resultados

	I	II	III	IV
β_0	5.74	-	5.936 (4.215, 7.883)	5.876 (4.108, 7.914)
β_1	0.60	0.567	0.610 (-0.103, 1.362)	0.587 (-0.140, 1.373)
θ_1	134.93	124.45	181.3 (71.90, 430.9)	176.8 (69.13, 429.1)

Nas colunas III e IV são apresentadas as estimativas dos parâmetros (média) e intervalos de 95% de máxima densidade *a posteriori* obtidas neste trabalho, considerando-se a parametrização original (44 iterações de *burn-in* e 43 048 aproveitadas com "lag"$=1$ para β_0 e "lag"$=4$ para β_1) e a ortogonalização dos parâmetros (cinco iterações de *burn-in* e 5 801 todas aproveitadas), respectivamente.

Praticamente não houve diferenças entre as densidades estimadas pelos dois procedimentos (Figuras de 1 a 5), porém o esforço empreendido para ortogonalizar os parâmetros se justifica pela melhoria alcançada na aceleração da convergência.

Figura 1 – Densidade *a posteriori* para β_0 implementando o Gibbs sampler e utilizando a parametrização original (44 iterações de *burn-in* e 43 408 iterações aproveitadas) e o comportamento da cadeia ao longo das 6 000 primeiras iterações

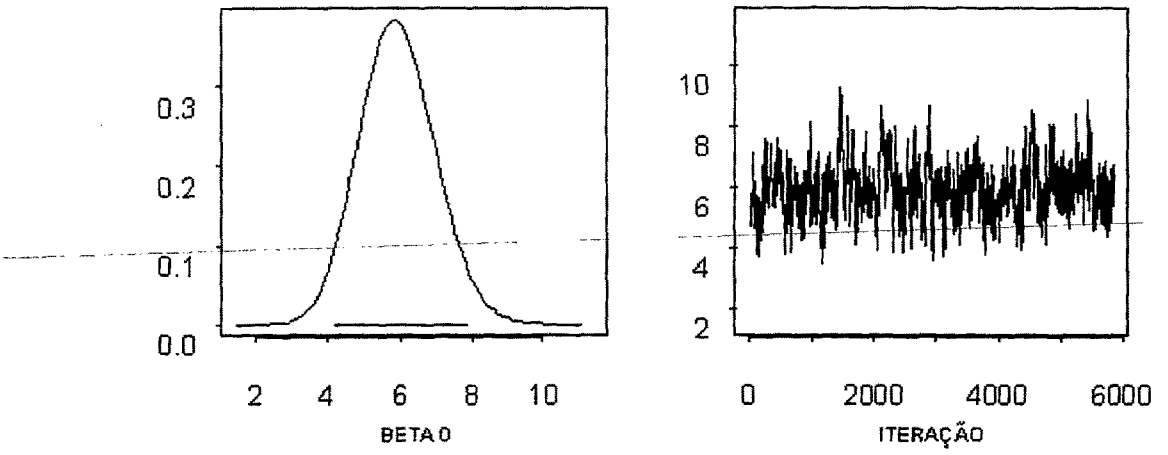


Figura 2 - Densidade *a posteriori* para β_0 e comportamento da cadeia, ortogonilizando os parâmetros (5 801 iterações aproveitadas e cinco de *burn-in*)

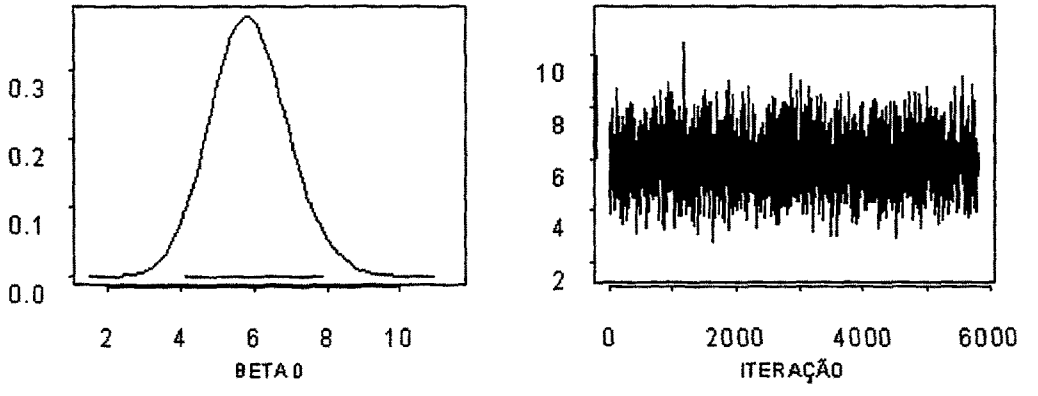


Figura 3 - Densidade *a posteriori* para β_1 implementado o Gibbs sampler e utilizando a parametrização original (44 iterações de *burn-in* e 43 408 iterações aproveitadas com *lag* = 4) e o comportamento da cadeia ao longo das 6 000 primeiras iterações aproveitadas

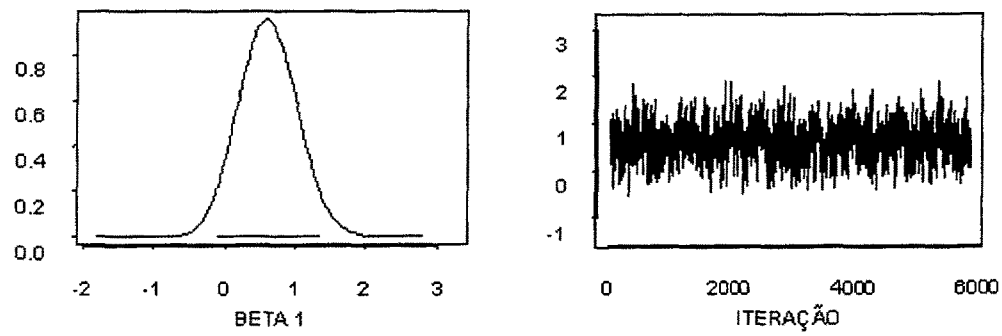


Figura 4 - Densidade *a posteriori* para β_1 e comportamento da cadeia, ortogonalizando os parâmetros (5 801 iterações aproveitadas e cinco de *burn-in*)

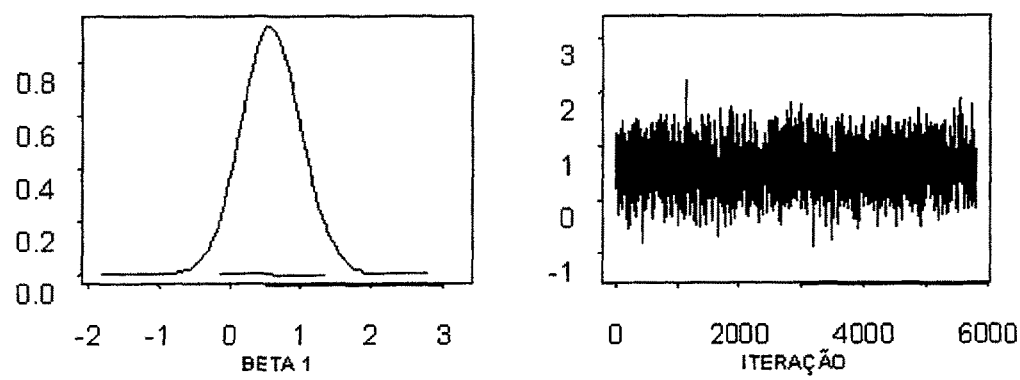
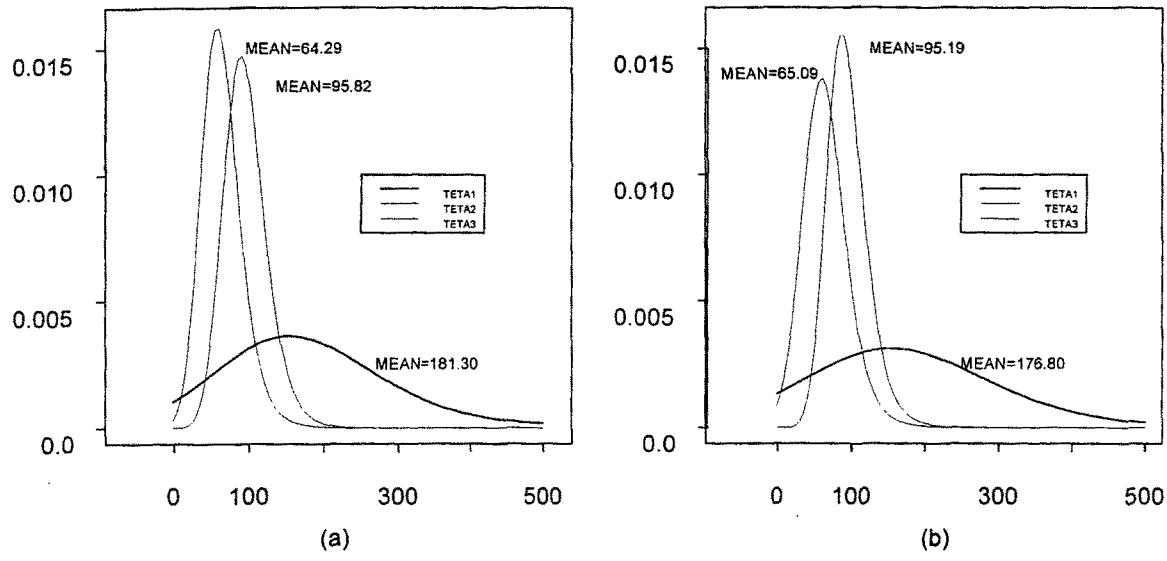


Figura 5 - Densidades *a posteriori* estimadas para os tempos médios de falha θ_1, θ_2 e θ_3, de itens submetidos aos três níveis de estresse, implementando o amostrador de Gibbs para a parametrização original (a) e ortogonalizando os parâmetros (b)



Como era esperado, houve diferenças significativas entre os tempos médios de falha estimados (Figura 5) para os diferentes níveis de estresse, nos dois casos (a) e (b), caracterizando uma taxa de falha crescente com o estresse.

Essa diferença observada entre os três níveis de estresse também se faz notar na função de confiabilidade (Figura 6) e na densidade preditiva (Figura 7).

Figura 6 - Boxplot da confiabilidade estimada para os tempos 30, 60, 90, de itens submetidos aos três níveis de estresse

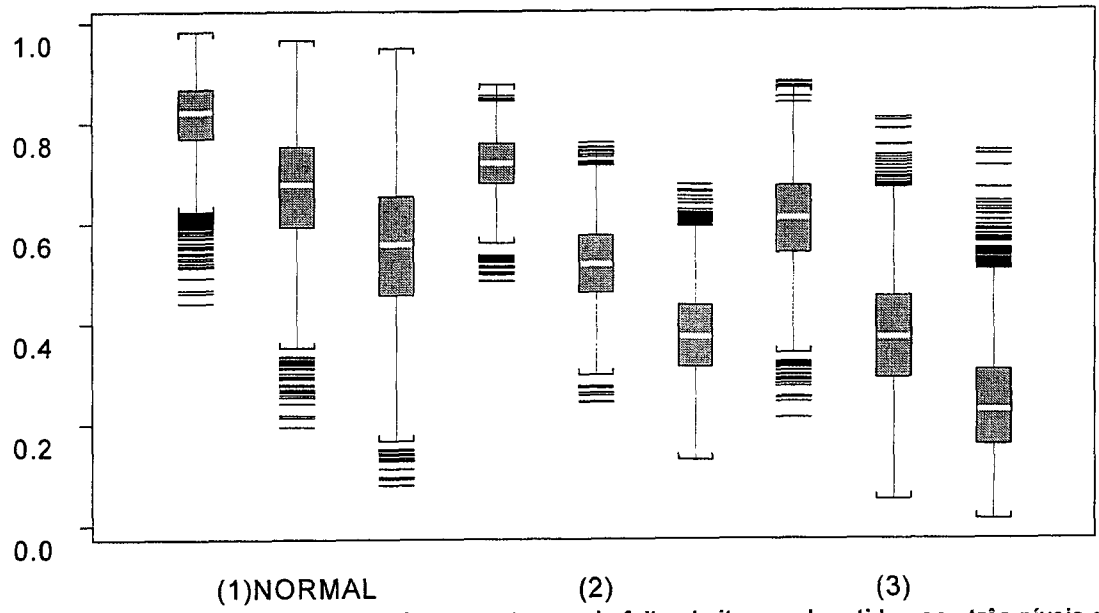
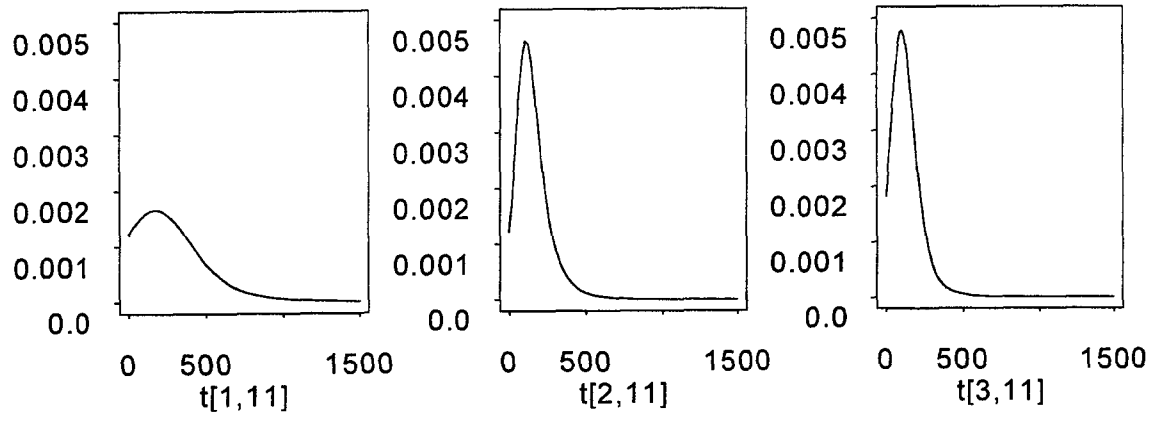


Figura 7 - Densidade preditiva estimada para o tempo de falha de itens submetidos aos três níveis de estresse (5 801 iterações aproveitadas e cinco de *burn-in*)



Entre os parâmetros β_0 e β_1 foi constatada uma alta correlação igual a 0,968 como mencionado anteriormente, o que nos levou a fazer uma nova análise com parâmetros ortogonalizados. Em ambos os casos, pelo diagnóstico de Geweke implementado no BUGS, em 4 000 iterações não houve evidências para descartar a existência de convergência para cada variável, ou seja, o diagnóstico não sugere que as primeiras 25% iterações da amostra sejam provenientes de distribuição diferente das 50% últimas.

Os diagnósticos de Raftery & Lewis, Heidelberger & Welch e também o de Geweke (10% primeiras iterações e 50% últimas), implementados no CODA - *Convergence Diagnosis and Output*

Analysis Software for Gibbs Sampling Output, Best et al. (1997), apresentaram valores diferentes entre as estatísticas de teste dos dois casos. Para a obtenção da convergência, também sugeriram números diferentes para o total de iterações, para o *burn-in* e para os *lags* (Tabela 3).

Tabela 3 – Iterações exigidas pelo diagnóstico de Raftery & Lewis
(utilizando os parâmetros originais / utilizando parâmetros ortogonais)

	BURN-IN	TOTAL	LAG	PARA A AMOSTRA
β_0	21 / 5	21629 / 5443	1 / 1	21629 / 5443
β_1	44 / 5	43048 / 5801	4 / 1	10762 / 5801

A eliminação da autocorrelação dentro de cada cadeia também foi acelerada com a ortogonalização dos parâmetros, como pode ser visto nas Figuras 8 e 9.

Figura 8 - Autocorrelação dentro de cada cadeia, para diferentes *lags*, utilizando a parametrização original (6 000 iterações)

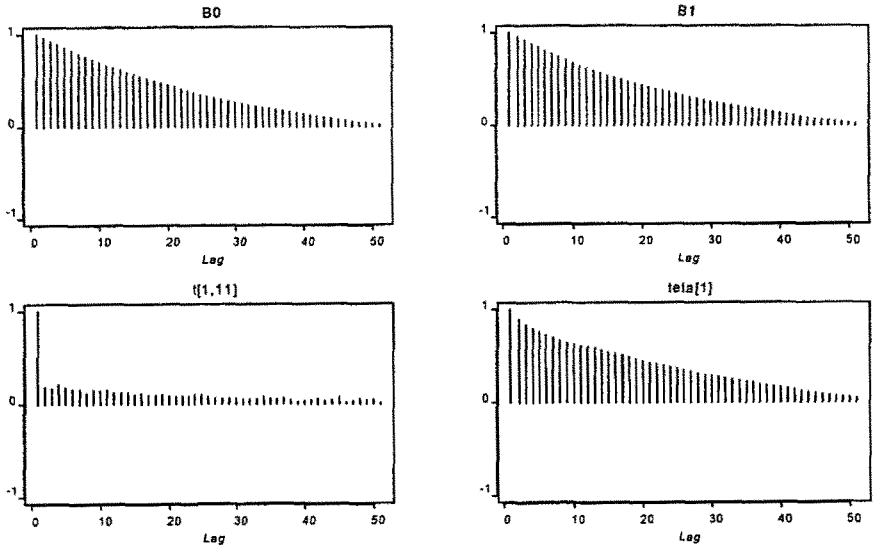
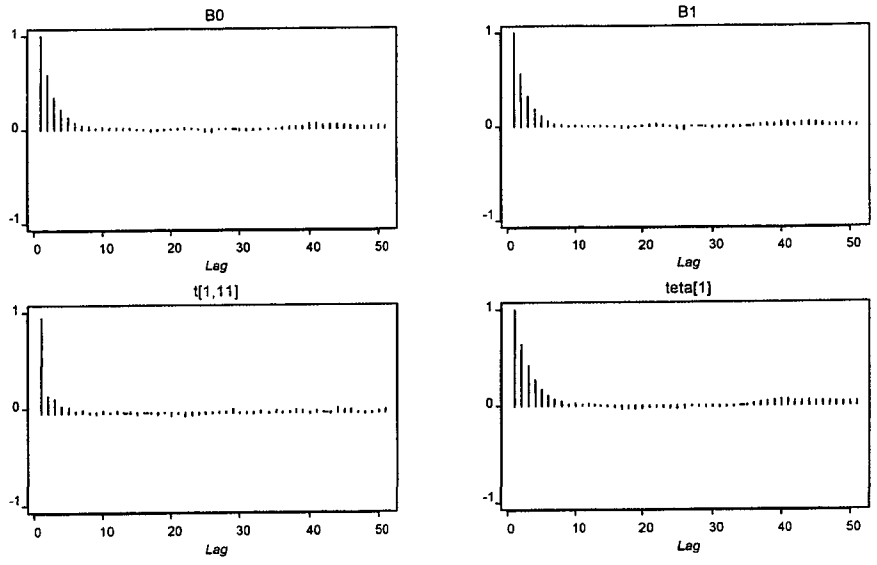


Figura 9 - Autocorrelação dentro de cada cadeia, para diferentes *lags*, utilizando a parametrização original (6 000 iterações)



6. Conclusões

Apesar de serem observadas pequenas diferenças nas estimativas pontuais, todas as estimativas obtidas por Laplace e os valores utilizados para a geração dos dados se encontram dentro dos intervalos de máxima densidade *a posteriori* gerados via amostrador de Gibbs ; esse fato e o pouco tempo de processamento sugerem que esses procedimentos podem ser considerados uma alternativa eficiente para a análise da confiabilidade ou análise de sobrevivência em problemas envolvendo modelos de regressão.

Não houve diferenças significativas entre as estimativas obtidas pelos dois procedimentos (utilizando a parametrização original e utilizando parâmetros ortogonais), o que torna o esforço empreendido com a reparametrização justificado pela aceleração da convergência e conseqüente redução do tempo requerido para o processamento (de três para meio minuto).

A redução de tempo é expressiva sobretudo levando-se em conta que no mundo real trabalha-se com diversos conjuntos de dados em muitos experimentos dessa natureza.

Referências Bibliográficas

ARROW, J. (1996) "Estimating the Influence of Health as a Risk factor on Unemployment: a Survival Analysis of Employment Durations for Workers Surveyed in the Germa Socio-Economic Panel (1984-1990)", *Social Science & Medicine* 42, 12 pp. 1651-1659

BEST, N., COWLES, M. K. & VINES, K. (1997) CODA - *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output* - versão 0.4. Technical Report, MRC Biostatistics Unit, University of Cambridge, England.

COX, D. R. e REID, N. (1987) "Orthogonal Parameters and Approximated Conditional Inference" (with discussion). Journal Royal Statistical Society, B, 49,1, pp. 1-39.

DELLAPORTAS, P. e SMITH, A. M.. F. (1993) "Bayesian Inference for Generalised Linear and Proportional Hazards Models Via Gibbs Sampling". *Applied Statistics*, 42, pp. 443-459.

FRECHET, G., LANGLOIS, S. e BERNIER, M. (1992) "Transition in the Labor-Market a Longitudinal Perspective". *Relations Industrielles-Industrial Relations*, 47, 1, pp. 79-99

GILKS, W. R., RICHARDSON, S. e SPIEGELHALTER, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

GILKS, W. R. e WILD, P. (1992) "Adaptive Rejection Sampling for Gibbs Sampling". *Applied Statistics*, 41, 337-348.

LOUZADA-NETO, F. e ACHCAR, J. A. (1993). "Uso de Dados Acelerados no Controle da Qualidade de Produtos Industriais Assumindo Uma Distribuição Exponencial e Um Modelo Estresse-resposta Geral" *Estadistica*, 45, 144, 145, pp. 81-106.

NELSON, W. (1990). "*Accelerated Testing-Statistical Model, Tests Plans and Data Analyses*". Wiley & Sons, New York.

SPIEGELHALTER, D., THOMAS, A., BEST, N. e GILKS, W. (1997) BUGS - *Bayesian Inference Using Gibbs Sampling* - version 0.6. Technical Report, Cambridge, England.

ABSTRACT

Accelerated and censored life tests are often used to analyze the performance of units whose time failure under normal operating conditions are very large or expensive. The Bayesian approach is widely applied in reliability, considering a stress-response model and specially the exponential distribution for the lifetimes. The implementation of the Bayesian paradigm involves many integrals which have no exact analytic solution, then need to be numerically solved. To circumvent this problem, in this paper is proposed the use Gibbs sampler to perform Bayesian inference to evaluate the reliability of units with exponential lifetime, submitted to accelerated and censored life test. For improve the convergence of the chains, a reparametrization is suggest. To illustrate the methodology, an application using artificially generated data are presented and its results are commented.

Mapas de Malária em Rondônia Usando o Estimador Bayesiano Empírico para Dados Binários

Renato Martins Assunção*

Edna Afonso Reis*

Paola B. Marchesini e Diana O. Sawyer*

RESUMO

Consideramos a estimação do risco de malária em pequenos lotes de assentamento rural na região de Machadinho, em Rondônia. Propomos um estimador bayesiano empírico adaptado à distribuição estocástica proposta para os dados de malária, a Binomial. Este estimador considera a informação de todos os lotes da região durante a estimação do risco em um determinado lote, podendo ser adaptado para utilizar a dependência espacial dos riscos entre os lotes para melhorar a estimação. Através de estudos de simulação, concluímos que o mapeamento do risco assim estimado reduz o erro de estimação global, permite visualizar a distribuição espacial da doença com menor efeito de flutuações aleatórias das quantidades observadas, e reproduz o arranjo espacial da doença de forma mais satisfatória que o estimador de máxima verossimilhança. Os mapas de malária em Machadinho com riscos estimados pelo estimador bayesiano empírico mostram que, ao longo do tempo, ainda permanece uma forte influência do componente ambiental na distribuição espacial da doença na região.

1. Introdução

A malária é a doença parasitária mais importante do mundo atual. Atinge de 200 a 400 milhões de pessoas, mata a cada ano mais de um milhão, na maior parte crianças, e incapacita muitos habitantes das áreas endêmicas (Bailey, 1982). Em Gana, por exemplo, a malária está em primeiro lugar entre as causas de invalidez temporária da população local, com uma média de 32,6 dias perdidos de trabalho por pessoa por ano (Nussenzweig e Nussenzweig, 1985).

* Endereço para correspondência: Deptº de Estatística – UFMG – C.Postal 702 – 30161-970 - Belo Horizonte – MG – e-mail: assuncao@est.ufmg.br.

No Brasil, grandes progressos foram feitos nos últimos 70 anos, e hoje sua incidência está praticamente restrita à Região Amazônica. Na Amazônia, porém, a incidência não só permanece alta como registra, desde 1970, um aumento no número de casos a cada ano (Tauil, 1984). A região apresenta uma conjunção de fatores que favorecem a transmissão da malária e outros que dificultam seu controle. São fatores de natureza física (acúmulo de grande quantidade de água), biológica (grande dispersão do principal mosquito vetor) e, de outro lado, de natureza socioeconômica, principalmente as condições precárias de habitação, saúde e educação.

No fim da década de 60, iniciou-se o processo de ocupação intensiva da Amazônia. O governo ofereceu incentivos fiscais a empresários de outras regiões do País para a implantação de projetos agropecuários. A abertura ou pavimentação de rodovias como a Transamazônica facilitaram o acesso e colonização de regiões desabitadas. A construção de hidrelétricas e a abertura de vários garimpos também contribuíram para atrair milhares de pessoas em busca de melhores condições de vida. O fluxo de pessoas não imunes para a área e o desmatamento para criação de áreas de plantio foram seguido por uma explosão na prevalência da malária na região. Enquanto em 1970 havia 52469 casos registrados de malária no Brasil, este número alcançou 568958 em 1991. Durante esse período, a malária foi erradicada na maior parte do País, ficando concentrada na Região Norte. Atualmente, 90% dos casos relatados ocorrem na Amazônia, com uma taxa de crescimento anual de 16%, em contraste com o aumento da população de 4.9% ao ano (Castilla and Sawyer, 1993).

A alta incidência de malária tornou-se um dos principais obstáculos ao processo de colonização da Amazônia. O agricultor atacado pela malária tem que interromper seu trabalho para ficar em repouso, muitas vezes por vários dias. Estando exposto ao alto risco de contaminação da doença na região, seu trabalho pode ser interrompido muitas vezes ao longo do ano, podendo coincidir com época de plantio ou colheita. Desse modo, torna-se economicamente inviável para muitas famílias permanecerem na região, o que gera a alta mobilidade populacional observada nas regiões de assentamento.

Pesquisadores do Centro de Desenvolvimento e Planejamento Regional da Universidade Federal de Minas Gerais - CEDEPLAR-UFMG - conduzem, desde 1984, um estudo sobre a malária em duas regiões de assentamento rural da Amazônia: Nortão, em Mato Grosso, e Machadinho, em Rondônia. O principal objetivo desse estudo é analisar os fatores socioeconômicos associados a diferenças na prevalência da malária dentro dessas regiões. Assim, existe interesse em saber por que, dentro de uma população exposta ao alto risco de malária (devido às condições em que vivem), alguns grupos experimentam mais malária e outros menos. Adicionalmente, deseja-se observar como essas relações mudam à medida que avança o processo de assentamento da população em cada região. Além dos fatores socioeconômicos, outros estudos são conduzidos paralelamente avaliando as características ambientais e geográficas e da região, envolvendo análise de imagens de satélite e de mapas da região e estudos entomológicos. Essas análises são apresentadas em Sawyer e Sawyer (1987 e 1995).

Este artigo refere-se a um tópico específico dentro da análise estatística da prevalência da malária na região estudada em Rondônia. O objetivo é construir mapas do risco da doença nos lotes de assentamento em que a região está dividida usando-se um estimador dos riscos mais estável que o usual estimador de máxima verossimilhança. Propomos um estimador bayesiano empírico que, além de apresentar menor erro de estimação, considera a dependência espacial do risco da doença. Justificamos nossa escolha em estudos de simulação que comprovam a superioridade desse estimador em relação ao estimador de máxima verossimilhança.

2. Malária em Machadinho

O Município de Machadinho localiza-se no extremo noroeste de Rondônia, próximo à fronteira com Mato Grosso. A área de estudo na região de Machadinho tem aproximadamente 2500 km^2, sendo dividida pelo rio Machadinho em duas sub-regiões chamadas Gleba 1 e Gleba 2. Essas, por sua vez, foram divididas pelo Instituto Nacional de Colonização e Reforma Agrária em 1755 lotes de assentamento rural. A Figura 6 mostra as duas glebas, onde as divisões menores são os lotes, e o rio Machadinho, que corre de sudoeste a nordeste.

O processo de assentamento rural em Machadinho começou por volta de 1984, sendo quase totalmente dependente da agricultura. Embora alguns poucos garimpos tenham sido criados, essa pode ser considerada uma atividade marginal.

O estudo sobre a incidência de malária nas Glebas 1 e 2 de Machadinho foi conduzido nos anos de 1986, 1987 e 1995. Uma vez em cada um desses três anos, os entrevistadores visitaram todas as casas da região preenchendo um questionário com cerca de 700 itens, coletando informações individuais dos membros da casa sobre características demográficas e socioeconômicas e suas experiências passadas com episódios de malária, diagnóstico e tratamento. Com relação à residência, foram coletadas informações sobre a percepção da malária, proteção contra o vetor, mobilidade e características da casa. Os detalhes da coleta de dados podem ser encontrados em Sawyer e Sawyer (1987).

Nem todos os lotes da região estavam ocupados na época do estudo. Os lotes ocupados tinham, em sua grande maioria, apenas uma família residindo nele. A Tabela 1 apresenta as informações relativas ao número de lotes pesquisados e ao número de moradores das duas glebas.

Tabela 1 - Estatísticas de ocupação dos lotes na área de estudo em Machadinho

Variável	Gleba 1 (606 lotes)			Gleba 2 (1147 lotes)		
	1986	1987	1995	1986	1987	1995
Nº de lotes ocupados	115	190	314	411	546	616
Nº total de pessoas	529	869	1748	2095	2884	3490
Nº médio de pessoas/lote	4.6	4.6	5.6	5.1	5.2	5.7

A ocorrência de malária durante o período do estudo foi monitorada para praticamente todos os moradores dos lotes. Cada morador preencheu um calendário onde marcava, a cada mês de um determinado ano de pesquisa, se esteve no lote naquele mês (e portanto, exposto à malária) e se teve malária durante aquele mês. Com essas informações, podemos ter uma idéia da intensidade da malária em cada lote da região e, assim, estimar o risco de se ter malária em cada lote. O estimador usado inicialmente para estimar o risco de malária em um lote em determinado ano foi:

$$\text{Proporção de malária no lote} = \frac{\text{Número de pessoas-meses de malária no lote}}{\text{Número de pessoas-meses de exposição no lote}} \qquad (1)$$

que assume valores entre 0, indicando baixo risco, e 1 indicando um risco muito alto, quando todas as pessoas do lote tiveram malária durante todo o tempo que ficaram no lote no ano em questão. A Tabela 2 mostra algumas estatísticas de exposição e infecção por malária. Note que o nível de malária esteve muito alto nesse período. Por exemplo, em 1986, cada pessoa da Gleba 1 ficava, em média, 39% do seu tempo incapacitada pela doença. Note também que o nível de malária na Gleba 2 foi menor durante todo o período.

Tabela 2 - Estatísticas de exposição e infecção por malária na área de estudo em Machadinho

Variável (per capita)	Gleba 1			Gleba 2		
	1986	1987	1995	1986	1987	1995
Nº médio de meses de exposição	8.5	9.5	11	9.4	10.4	11
Nº médio de meses de malária	3.6	3.4	1.2	2.8	2.1	0.4
Proporção de malária (média)	0.39	0.36	0.12	0.26	0.20	0.06

A proporção de malária pode ser calculada para vários subgrupos de indivíduos, em categorias como idade e sexo, por exemplo. Sawyer e Sawyer (1987) mostram que a malária afeta diferencialmente em relação à idade e sexo dos moradores de Machadinho, estando os homens e adultos expostos a um risco mais alto que mulheres e crianças. Esse fato se deve às diferentes atividades de cada grupo, criando diferentes graus de exposição. Por essa razão, restringimos nossa análise ao grupo mais atingido pela doença, os homens com idade maior ou igual a 15 anos, tornando mais homogêneo o grupo sob análise. O número de lotes analisados nos anos de 1986, 1987 e 1995, respectivamente, passa a ser 115, 188 e 314 na Gleba 1 e 407, 516 e 602 na Gleba 2.

Na construção de mapas do risco de malária em Machadinho, os lotes serão usados como unidade espacial de referência. A Figura 1 mostra o histograma da proporção de malária nos lotes

ocupados por pelo menos um homem com idade igual ou superior a 15 anos em cada ano pesquisado. Podemos ver que a incidência de malária medida em cada lote dessa região diminuiu de 1986 a 1995.

2.1. Mapas do risco de malária em Machadinho

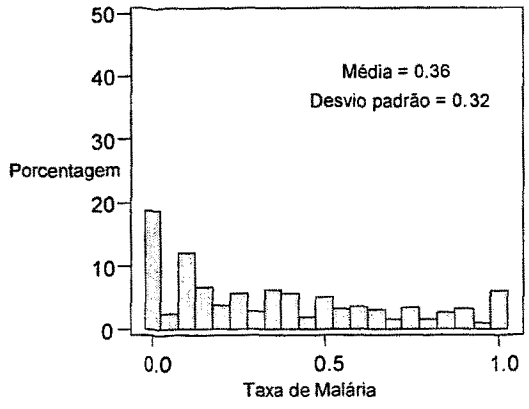
A estimação da distribuição geográfica de uma doença é uma aplicação crescente em Epidemiologia (Walter e Bernie, 1991; White, 1995). Estudos descritivos incluem a análise de mapas do risco da doença com o objetivo de visualizar uma possível presença de padrão espacial. Esses mapas são também instrumentos valiosos para apontar associações entre fontes potenciais de contaminação e áreas de risco elevado, podendo sugerir determinantes locais de doenças e fatores etiológicos desconhecidos (Donnely, 1995; Glass et al., 1995).

Essas finalidades são atingidas quando os mapas possuem resolução geográfica adequada. Portanto, para serem úteis, os mapas devem utilizar pequenas regiões geográficas como unidades de análise. Várias dessas pequenas áreas possuirão pequenas populações sob risco, o que acarretará taxas de incidência com muita instabilidade. Isto é, o acréscimo ou decréscimo de um caso nessas áreas poderá causar mudanças drásticas nas taxas. Em termos estatísticos, as taxas das diversas áreas que compõem a região de estudo não são comparáveis, já que possuem variâncias muito diferentes.

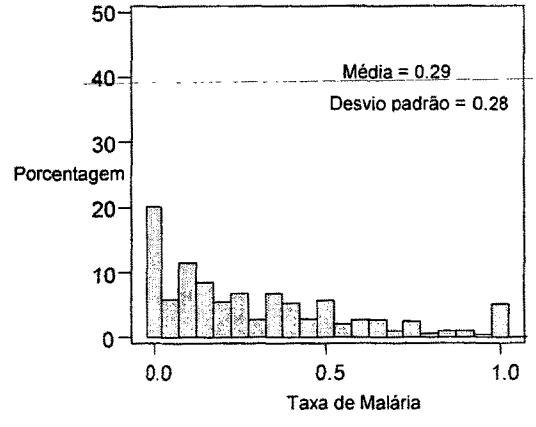
Métodos bayesianos e bayesianos empíricos de estimação dos riscos em pequenas áreas têm sido propostos na literatura para superar essas e outras dificuldades (Assunção et al., 1998). Esses métodos, ao estimar o risco de uma pequena área, têm como idéia central o uso de informações das outras áreas que compõem a região de estudo. A principal conseqüência é a diminuição do erro quadrático médio total das estimativas (Efron e Morris, 1973). Isto é, o conjunto dos riscos relativos de todas as áreas é estimado de forma mais precisa pelos métodos bayesianos ou bayesianos empíricos do que pelas taxas usuais (estimadores de máxima verossimilhança). Métodos não-bayesianos sofisticados e que levam em conta a autocorrelação espacial poderiam ser formulados, mas são de difícil formulação.

Figura 1 - Histogramas da proporção de malária (expressão 1) nos lotes ocupados por pelo menos um homem com idade igual ou superior a 15 anos na área de estudo em Machadinho

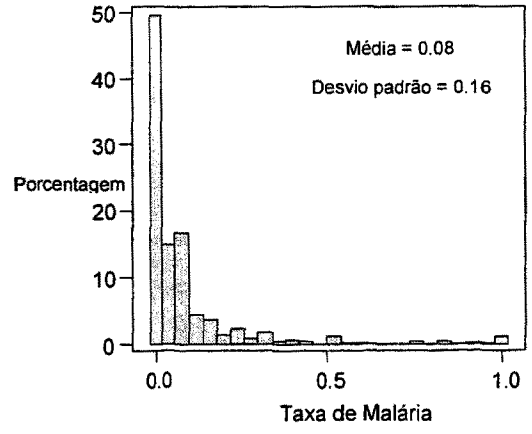
Ano: 1986



Ano: 1987



Ano: 1995



Na estimação de fenômenos espaciais, os métodos Bayesianos ou Bayesianos empíricos têm a vantagem adicional de possibilitarem a incorporação das similaridades espaciais nos riscos entre áreas contíguas. Essas similaridades são devidas à usual variação suave do risco sobre a região sendo mapeada. Incorporar essa informação na estimação dos riscos pode levar a mapas com estimativas mais estáveis e a uma diferenciação mais precisa entre o que é de fato risco elevado (ou muito baixo) e o que é flutuação aleatória causada por pequenas populações. Além disso, espera-se que as estimativas reproduzam o padrão espacial dos riscos reais.

Na literatura de mapas de doenças raras, usando-se a distribuição de Poisson para as contagens de casos da doença, não têm sido encontradas diferenças substanciais entre os métodos Bayesianos empíricos e aqueles inteiramente Bayesianos para estimar o risco pontualmente (Bernadinelli e Montonolli, 1992). Os métodos bayesianos empíricos têm sido preferidos, em geral, devido ao seu menor custo computacional (Devine et al., 1994).

Marshall (1991) propõe um estimador bayesiano empírico para o risco de doenças raras, quando pode-se aproximar a distribuição do número de casos pela distribuição de Poisson, com parâmetros estimados pelo método dos momentos. Ele mostra que este estimador possui erro médio de estimação menor que o usual estimador de máxima verossimilhança e é comparável a outros estimadores bayesianos empíricos obtidos de forma mais difícil tal como, por exemplo, pelo algoritmo EM.

Os mapas bayesianos de doenças usuais, com grandes regiões como unidade de observação, possuem um grande número de pessoas-anos e assim a única fonte de instabilidade possível das taxas de morbidade é o seu numerador. Neste estudo, lidamos com uma situação onde ambos, tanto o numerador quanto o denominador da proporção de malária calculada por lote na expressão (1) são bastante pequenos. Isso significa que uma pequena alteração aleatória no numerador (ou no denominador) pode levar a uma alteração substancial na proporção de malária. Além disso, esta situação inviabiliza a adoção, como nos mapas usuais, de uma distribuição de Poisson para as contagens no numerador.

Assim, para estudar a distribuição espacial da malária nos lotes de Machadinho sem correr o risco de basear nossas conclusões em variações aleatórias da proporção de malária da equação (1), dissociadas do risco real, propomos usar um estimador bayesiano empírico adaptado à distribuição binomial assumida para esses dados. O desenvolvimento desse estimador é mostrado na próxima seção, bem como os resultados dos experimentos simulados, que mostram a sua superioridade sobre a proporção de malária por lote dada na equação (1), o estimador de máxima verossimilhança para o risco de malária em cada lote. Com base nos resultados favoráveis do estimador bayesiano empírico, mostramos, na seção 4, sua aplicação na estimação do risco de malária nos lotes de Machadinho.

3. O Estimador Bayesiano Empírico para a distribuição Binomial

Considere uma região dividida em n lotes indexados por i, $i = 1, 2, ..., n$. Em um determinado ano, seja Y_i o número total de meses de malária em N_i meses de exposição para todos os indivíduos do lote i, em uma determinada classe de sexo e idade. No caso de Machadinho, em cada lote, Y_i e N_i são contados apenas para os homens com idade igual ou superior a 15 anos. Condicionalmente no valor do risco de malária no lote i, assuma que Y_i, $i=1,...,n$, são independentes e têm distribuição Binomial com parâmetros N_i e P_i. Assim, P_i representa o risco de se ter malária no lote i em um dado mês, a quantidade que se deseja estimar.

O estimador de máxima verossimilhança de P_i, que chamaremos de *taxa bruta*, é $x_i = y_i/N_i$, onde y_i é o valor observado para Y_i. A taxa bruta é, assim, a proporção de malária dada pela equação (1) calculada para lotes. O uso da palavra "taxa" neste artigo é um abuso de terminologia já que não estamos descontando do denominador os meses em que os indivíduos estão doentes e, portanto, não estão expostos ao risco. Sabe-se que $E(x_i|\ P_i) = P_i$ e que $Var(x_i|\ P_i) = P_i(1-P_i)/N_i$.

Suponha que os riscos $\{P_i\}$ tenham uma densidade de probabilidade (qualquer) *a priori* com média M e variância A. Assim, não condicionado a P_i, a taxa bruta x_i tem média $E_x(x_i) = E_P[E(x_i|\ P_i)] = M$ e variância $Var_x(x_i) = Var_P[E(x_i|\ P_i)] + E_P[Var(x_i|\ P_i)] = A + (M-A-M^2)/N_i$.

Como critério de comparação entre possíveis estimadores dos riscos $\{P_i\}$, considere a função perda dada pelo erro quadrático total $\sum_i (\theta_i - P_i)^2$, onde θ é um estimador dos riscos. Dados M e A, o estimador de Bayes linear ótimo para P_i é o estimador de contração:

$$B_i = M + C_i(x_i - M) \qquad (2)$$

onde $C_i = Var_P(P_i)/Var_x(x_i) = A/(A + (M-A-M^2)/N_i)$ é um valor entre 0 e 1 (Efron e Morris, 1973).

A expressão (2) mostra que o efeito do estimador de Bayes (B_i) sobre a taxa bruta (x_i) é torná-la mais próxima da média M, reduzindo ou aumentando seu valor quando este for muito alto ou muito baixo. A intensidade da contração em torno da média M depende da quantidade $E_P[Var(x_i|\ P_i)]$. Um valor grande indica que, em média, x_i é instável e, em conseqüência, o estimador terá C_i próximo de 0, efetuando uma grande contração. Se o valor de $E_P[Var(x_i|\ P_i)]$ for pequeno, então, em média, x_i tem variância pequena e $C_i \approx 1$, resultando em uma contração pequena. Escrevendo $B_i = C_i x_i + (1-C_i)M$, podemos ver que a estimativa de Bayes do risco no lote é uma média ponderada entre a taxa bruta naquele lote e a média a priori dos riscos, onde o peso da taxa bruta é a constante C_i.

Na abordagem bayesiana empírica, os valores observados de x_i são tratados como uma amostra da sua distribuição marginal e utilizados na estimação de M e A, que pode ser feita pelo método dos momentos. Como $E_x(x_i) = M$, qualquer média ponderada das taxas brutas é um estimador

não viciado para M. Fazendo o peso de cada lote igual ao número total de meses de exposição no lote, um estimador de M é:

$$m = \frac{\sum_{i=1}^{n} N_i x_i}{\sum_{i=1}^{n} N_i} = \frac{\sum_{i=1}^{n} y_i}{N} \tag{3}$$

onde $N = \sum_i N_i$. Note que m é a taxa bruta de malária em toda a região, isto é, desconsiderando sua divisão em lotes.

Para estimar A, vamos considerar o estimador de momentos proposto por Marshall (1991), no seu caso de dados Poisson, adaptado para dados binomiais por Martuzzi e Elliott (1996), dado por:

$$a = \frac{s^2 - (m - m^2)n/N}{1 - n/N} \qquad \text{onde} \qquad s^2 = \frac{\sum_{i=1}^{n} N_i (x_i - m)^2}{N} \tag{4}$$

Substituindo M e A respectivamente por m e a na equação (2), o estimador bayesiano empírico (EBE) de P_i torna-se:

$$b_i = m + c_i(x_i - m) \qquad \text{onde} \qquad c_i = \frac{a}{a + \frac{m - a - m^2}{N_i}} \tag{5}$$

Em alguns casos, pode ocorrer um valor de c_i que torne b_i negativo, maior que um ou indefinido quando o seu denominador é igual a zero. Nesses casos, uma correção é fazer $b_i = m$. A correção pode ser necessária quando o número de lotes usados no cálculo é muito pequeno.

3.1. Estimador Bayesiano Empírico global e local

A expressão (5) mostra que a estimativa bayesiana empírica do risco em cada lote é baseada na informação sobre malária específica para aquele lote (em x_i) e a informação fornecida por *todos* os outros lotes (em m e c_i). Esse estimador bayesiano empírico é chamado *global*.

Em doenças como a malária, é razoável pensar que o risco seja similar em lotes próximos, em virtude da similaridade espacial em aspectos ambientais relacionados à exposição ou infecção. A distribuição espacial do risco na região pode se mostrar na forma de uma curva suave ou como uma mistura de áreas bem definidas de baixo ou alto risco de malária.

Uma maneira simples de considerar essa configuração espacial na estimativa do risco é calcular a estimativa bayesiana empírica no lote i usando-se apenas os dados dos lotes pertencentes a uma *região de vizinhança* do lote i, ao invés de usar todos os lotes da área considerada. Esse estimador bayesiano empírico é chamado *local*.

Assim como o próprio valor do risco, a sua configuração espacial na região é desconhecida. Acreditamos que o estimador bayesiano empírico local, além de reduzir o erro de estimação da taxa bruta, tenha a propriedade adicional de revelar esse padrão espacial. A definição das regiões de vizinhança dos lotes para o cálculo das estimativas locais envolve o conhecimento dos mecanismos de espalhamento da doença e inclui ainda a decisão sobre o número de lotes usados no cálculo da estimativa do risco.

3.2. Estudo de simulação

O objetivo dessas simulações é comparar o desempenho, em termos do erro global de estimação, do estimador bayesiano empírico para dados binários em relação à taxa bruta. Pretende-se também comparar estimadores bayesianos empíricos locais com o estimador global.

Tentamos aproximar ao máximo o cenário das simulações com a situação real dos dados de malária em Machadinho. Resultados positivos das simulações em relação à eficiência do estimador bayesiano empírico nos dão segurança para utilizá-lo no problema da malária, substituindo as estimativas fornecidas pela taxa bruta de malária apresentada anteriormente.

Decidimos não usar a própria região de Machadinho como base geográfica devido a seu grande número de lotes. Desse modo, seguindo o esquema das simulações de Marshall (1991), vamos considerar uma região hipotética formada por 100 lotes localizados em uma área quadrada com coordenadas (X,Y), $X, Y = 1,2,...,10$.

Na construção dos estimadores bayesianos empíricos locais, definimos a região de vizinhança do lote i como sendo o círculo de raio r centrado nas suas coordenadas (X,Y) e, assim, todos os lotes cujas coordenadas estão dentro desse círculo são considerados vizinhos do lote i. Definimos seis estimadores bayesianos empíricos locais, correspondentes aos raios r = 2, 2.3, 3, 3.2, 4 ou 4.3, o que gera um número de lotes vizinhos (mínimo, máximo) = {(3,8), (7,20), (8,24), (12,36), (14,44), (17,52)}, respectivamente. Definiu-se ainda uma região de vizinhança formada por todos os lotes, correspondendo ao estimador bayesiano empírico global.

O número total de meses de malária em cada lote, Y_i, é gerado da distribuição binomial com tamanho N_i (número total de meses de exposição) e probabilidade de sucesso P_i (risco real). O número total de meses de exposição N_i dentro de cada lote foi gerado de uma distribuição uniforme nos inteiros entre 1 e 12. Consideramos três situações de intensidade da doença na região, com os riscos reais P_i de cada lote assumindo valores em intervalos correspondendo a doenças muito freqüentes $(0.40 \leq P_i \leq 0.80)$, freqüentes $(0.10 \leq P_i \leq 0.40)$ e pouco freqüentes $(0.02 \leq P_i \leq 0.10)$. Em cada

uma dessas classes de intensidade da doença, a configuração espacial dos riscos na região foi caracterizada por dez arranjos espaciais, mostrados na Tabela 3 e na Figura 2.

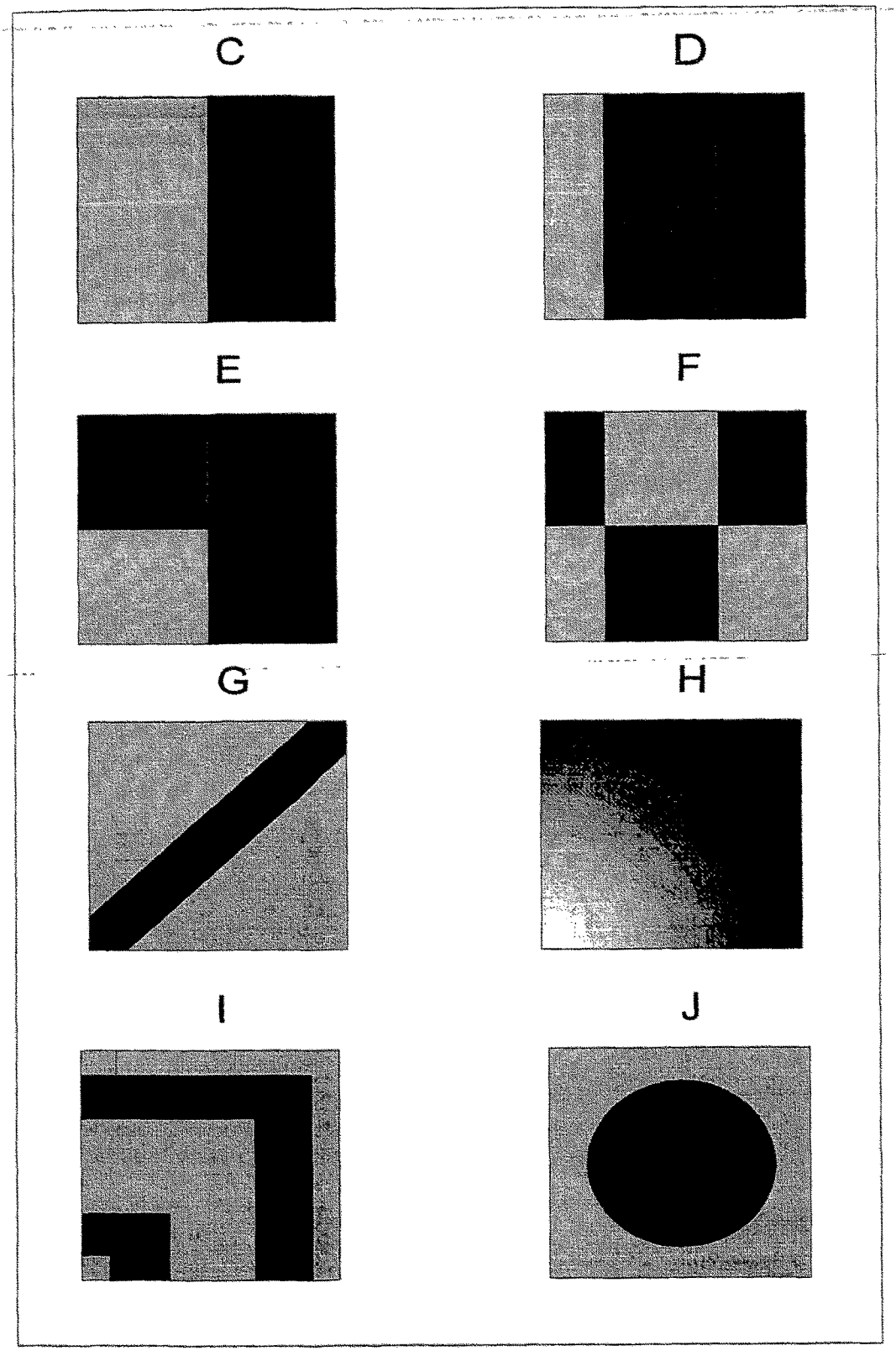
Tabela 3 - Arranjos espaciais dos riscos $\{P_i\}$ usados nos experimentos simulados

Descrição de P_i	Coordenadas dos lotes[1]	Valor de P_i por classe de intensidade da doença		
		Muito freqüente	Freqüente	Pouco freqüente
A) Constante	$\forall$ (X,Y)	0.50	0.25	0.10
B) Aleatório[2]	$\forall$ (X,Y)	U[0.4;0.8]	U[0.2;0.4]	U[0.02; 0.1]
C) Dois setores	$X \leq 5$	0.50	0.10	0.05
	$X > 5$	0.70	0.40	0.10
D) Três setores	$X \leq 3$	0.40	0.10	0.02
	$3 < X < 7$	0.60	0.25	0.08
	$X \geq 7$	0.80	0.40	0.10
E) Quatro setores	$X \leq 5$ e $Y \leq 5$	0.40	0.10	0.02
	$X \leq 5$ e $Y > 5$	0.50	0.20	0.05
	$X > 5$ e $Y \leq 5$	0.70	0.30	0.07
	$X > 5$ e $Y > 5$	0.80	0.40	0.10
F) Seis setores	$X \leq 3$ e $Y \leq 5$	0.40	0.10	0.02
	$X \leq 3$ e $Y > 5$	0.70	0.30	0.08
	$3 < X < 7$ e $Y \leq 5$	0.75	0.35	0.09
	$3 < X < 7$ e $Y > 5$	0.45	0.15	0.03
	$X \geq 7$ e $Y \leq 5$	0.50	0.20	0.04
	$X \geq 7$ e $Y > 5$	0.80	0.40	0.10
G) Diagonal	$-4 < X-Y < 4$	0.80	0.40	0.10
	caso contrário	0.40	0.10	0.05
H) Linear crescente	$\forall$ (X,Y)	0.4 + 0.02(X+Y)	0.1 + 0.015(X+Y)	0.02+0.004(X+Y)
I) Forma de L	$7 \leq X,Y \leq 9$	0.60	0.25	0.06
	$2 \leq X,Y \leq 4$	0.80	0.40	0.10
	caso contrário	0.40	0.10	0.02
J) Circular	$(X-5.5)^2+(Y-5.5)^2 < 4$	0.80	0.40	0.10
	$4 \leq (X-5.5)^2+(Y-5.5)^2 < 9$	0.70	0.30	0.07
	$9 \leq (X-5.5)^2+(Y-5.5)^2 \leq 16$	0.50	0.20	0.05
	caso contrário	0.40	0.10	0.02

Nota: Ver na Figura 2.

(1) $X, Y = 1,2, \ldots 10.$, (2) U: distribuição uniforme.

Figura 2 - Arranjos espaciais dos riscos (P_i) usados nos experimentos simulados



Nota: Quanto mais escura a imagem maior é o valor do risco no local (Tabela 3).

A comparação entre os estimadores taxa bruta (EMV), EBE global e EBE locais será feita através de uma medida de erro relativo de suas estimativas ao valor real do risco em cada lote. Considera-se o melhor estimador aquele que apresenta o menor erro relativo para o conjunto de n lotes (erro global de estimação), ou seja, aquele que apresentar o menor erro relativo médio. O erro relativo médio de um estimador **z** dos riscos $\{P_i\}$ é definido como:

$$ERM(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{|z_i - P_i|}{P_i}. \tag{6}$$

Nas simulações, fixada uma classe de intensidade da doença, para cada um dos dez arranjos espaciais dos riscos $\{P_i\}$ da Figura 2, foram gerados os valores de Y_i (número total de meses de malária) e calculados os estimadores da taxa bruta, bayesianos empíricos global e locais com os seis raios de vizinhança definidos anteriormente. Esse experimento foi repetido 100 vezes, em cada uma das simulações sendo calculado o erro relativo médio dos oito estimadores. Ao final das 100 simulações, calculou-se uma medida de comparação R entre as estimativas bayesianas empíricas b e a taxa bruta (x), definida por:

$$R = \frac{\sum_{j=1}^{100} ERM_j(b)}{\sum_{j=1}^{100} ERM_j(x)}. \tag{7}$$

Um estimador bayesiano empírico b é considerado mais satisfatório que a taxa bruta se $R < 1$, ou seja, se seu erro global de estimação é, em média, menor que o erro da taxa bruta. A comparação entre dois estimadores empíricos, b_1 e b_2, também é feita em termos do valor de R, isto é, o melhor estimador dos riscos será aquele que apresentar menor R.

3.3 Resultados e discussão

A Tabela 4 mostra os valores de R dos seis estimadores bayesianos empíricos locais e do estimador bayesiano empírico global, para as três classes de intensidade e aos dez arranjos espaciais da doença definidas anteriormente. Estes valores de R são bastante precisos, pois em 100 valores adicionais de R calculados para o arranjo espacial H, com $0.10 \leq P_i \leq 0.40$, o desvio padrão ficou entre 0.005 e 0.009 (coeficiente de variação entre 1.2% e 2.3%).

Em todas as situações sob simulação, o EBE apresentou melhores resultados que a taxa bruta, ou seja, $R<1$ em todos os casos. De fato, o menor e o maior valor de R foram, respectivamente, 0.14 e 0.82, indicando que o EBE reduz o erro global de estimação da taxa bruta em, no mínimo, 18%, chegando a reduzir em até 86%.

Tabela 4 - Valores de R (equação 8) para três classes de intensidade e dez arranjos espaciais dos riscos reais em 100 simulações do estimador bayesiano empírico global e seis estimadores locais com raio de vizinhança igual a r

Raio (r)	Arranjo Espacial									
	A	B	C	D	E	F	G	H	I	J
	Pouco freqüente: $0.02 \leq P_i \leq 0.10$									
2	0.54	0.72	0.61	0.67	0.64	0.67	0.57	0.58	0.78	0.75
2.3	0.38	0.52	0.41	0.49	0.43	0.51	0.42	0.40	0.62	0.57
3	0.35	0.49	0.39	0.49	0.42	0.50	0.40	0.37	0.61	0.57
3.2	0.30	0.46	0.34	0.44	0.37	0.50	0.35	0.32	0.61	0.56
4	0.28	0.44	0.33	0.45	0.37	0.49	0.35	0.31	0.62	0.59
4.3	0.25	0.42	0.30	0.45	0.36	0.50	0.34	0.27	0.58	0.60
Global	0.16	0.35	0.33	0.68	0.49	0.48	0.33	0.25	0.49	0.54
	Freqüente: $0.10 \leq P_i \leq 0.40$									
2	0.54	0.72	0.59	0.58	0.57	0.60	0.64	0.54	0.67	0.60
2.3	0.38	0.62	0.46	0.45	0.44	0.51	0.57	0.38	0.64	0.50
3	0.36	0.60	0.47	0.44	0.43	0.51	0.55	0.36	0.67	0.51
3.2	0.30	0.57	0.45	0.41	0.40	0.53	0.59	0.30	0.68	0.53
4	0.28	0.56	0.45	0.42	0.40	0.54	0.61	0.29	0.71	0.57
4.3	0.24	0.54	0.47	0.43	0.41	0.57	0.68	0.27	0.66	0.60
Global	0.14	0.52	0.76	0.72	0.65	0.59	0.82	0.37	0.59	0.58
	Muito freqüente: $0.40 \leq P_i \leq 0.80$									
2	0.55	0.73	0.56	0.57	0.58	0.62	0.63	0.55	0.64	0.59
2.3	0.39	0.63	0.41	0.45	0.44	0.55	0.56	0.38	0.60	0.52
3	0.37	0.61	0.40	0.45	0.44	0.56	0.55	0.36	0.61	0.54
3.2	0.31	0.59	0.35	0.41	0.41	0.59	0.58	0.31	0.61	0.56
4	0.29	0.57	0.35	0.43	0.41	0.61	0.59	0.30	0.63	0.59
4.3	0.25	0.56	0.34	0.44	0.42	0.64	0.63	0.28	0.61	0.61
Global	0.15	0.54	0.50	0.66	0.64	0.64	0.71	0.40	0.57	0.60

Consideremos agora a comparação entre os EBEs locais (espaciais) e o EBE global (não-espacial). A ausência de correlação espacial dos riscos nos arranjos A e B tornou os resultados do estimador global melhores do que os resultados dos estimadores locais. Esse resultado era esperado, pois, na ausência de dependência espacial, a vantagem adicional do estimador local sobre o global de incorporar esta informação adicional torna-se inútil. Como o estimador global usa mais lotes no seu cálculo, ele torna-se mais estável que o estimador local neste caso.

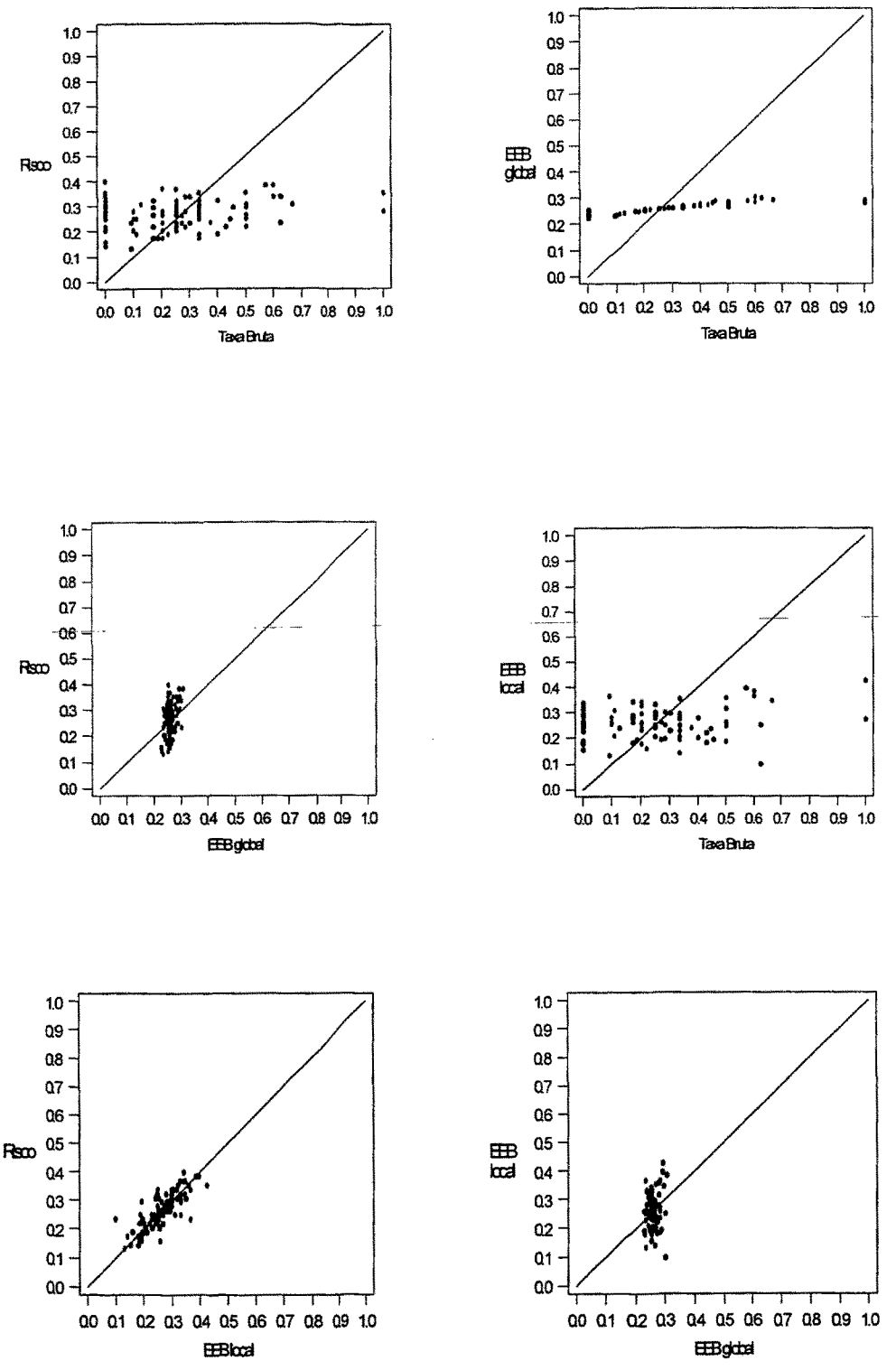
Nas classes de intensidade dos riscos "freqüente" e "muito freqüente", pelo menos um dentre os seis estimadores locais apresentou-se melhor que o estimador global, com exceção do arranjo espacial *I*. Na classe "pouco freqüente", os estimadores locais apresentaram resultados menos ou tão satisfatórios quanto os resultados do estimador global. Isso pode ser explicado pelo fato de que, quando $0.02 \leq P_i \leq 0.10$, o número total de meses da doença gerado (Y_i) é freqüentemente igual a zero, o que torna a correlação espacial fraca. Os resultados melhores do estimador global em relação aos estimadores locais no arranjo *I* podem ser explicados, possivelmente, pelo fato de nenhum raio testado ter-se adaptado às mudanças do nível do risco que ocorrem em curtos intervalos nesse arranjo.

Na maior parte das situações, o estimador local com raio igual a 2 apresentou resultado menos satisfatório que os outros estimadores locais. Esse fato é explicado pelo pequeno número de vizinhos que formam a vizinhança nesse caso, tornando maior a variância do estimador. Dentre os estimadores locais (raios > 2), é difícil concluir qual apresenta melhor resultado nos arranjos *C, D, E e I*. Notamos, entretanto, que nos arranjos onde há mudanças mais bruscas do risco *(F, G e J)* os estimadores locais com raio menor apresentam melhores resultados. O contrário ocorre no arranjo *H,* onde a mudança no nível do risco ocorre suavemente e os estimadores locais com raios maiores mostram-se melhores.

Para efeito ilustrativo, vamos considerar uma das 100 simulações do arranjo espacial *H,* com riscos $0.10 \leq P_i \leq 0.40$. A Figura 3 compara os riscos reais $\{P_i\}$ com as estimativas fornecidas pela taxa bruta, pelo EBE global e pelo EBE local (raio = 4). Analisando os gráficos à esquerda da figura, podemos observar que a dispersão em torno da linha de igualdade com o risco é muito menor para os estimadores bayesianos empíricos. De fato, o erro relativo médio da taxa bruta nesse caso é 0.59, enquanto os ERMs para os estimadores bayesianos empírico global e local são 0.19 e 0.11, respectivamente. Note também que grande parte (43%) das estimativas da taxa bruta estão fora do intervalo *[0.10;0.40]* definido para o risco real. Esses valores muito altos ou muito baixos da taxa bruta foram corretamente suavizados pelos estimadores bayesianos empíricos (gráficos à direita na Figura 3). Entretanto, o EBE global suavizou excessivamente, já que suas estimativas se restringiram ao intervalo *[0.22;0.31]*, enquanto as estimativas do EBE local estão no intervalo *[0.09;0.43]*.

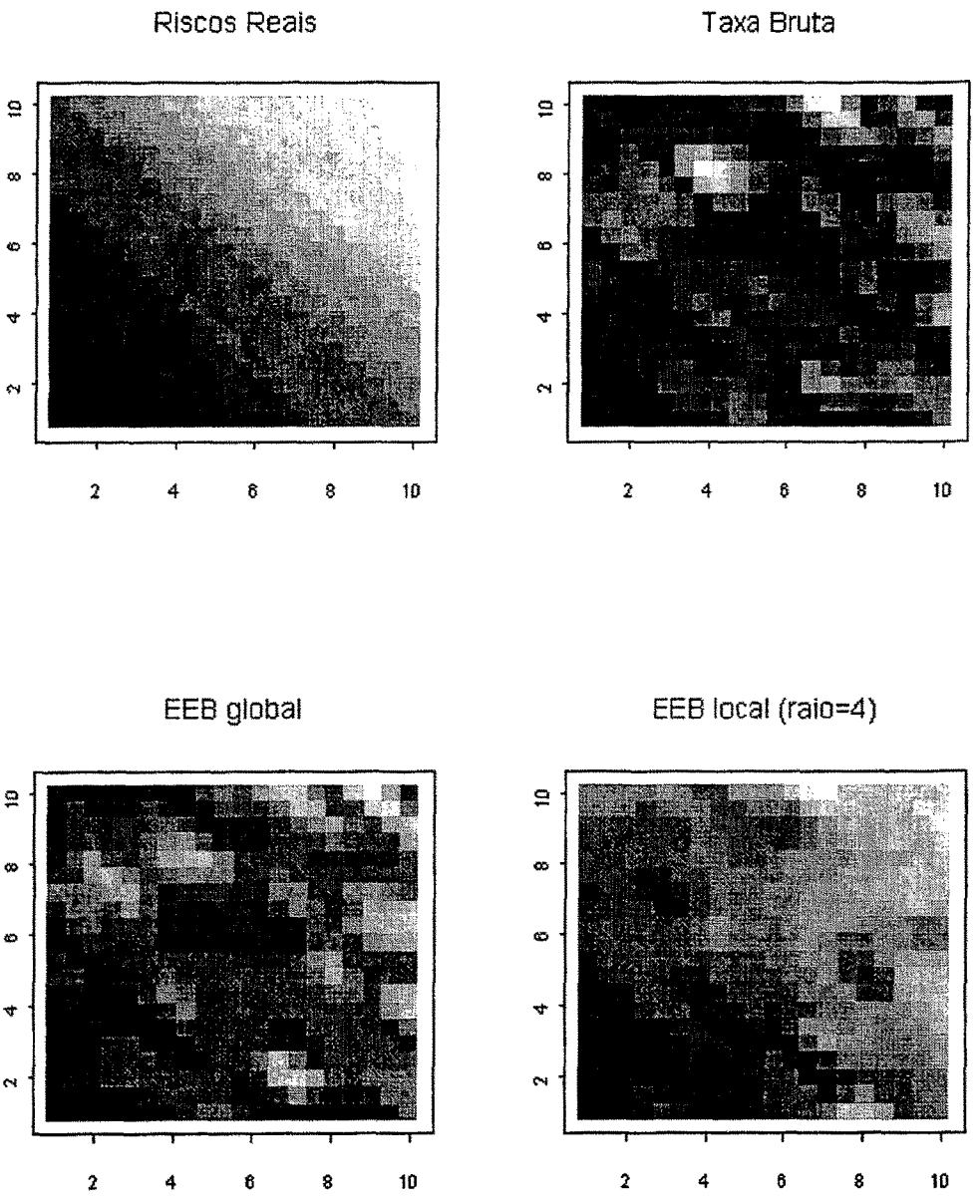
Na Figura 4, mostramos uma imagem dessas estimativas na região hipotética. Podemos ver claramente que apenas o EBE local conseguiu recuperar, ao menos parcialmente, o arranjo espacial dos riscos reais. Isto sugere que, embora em uma medida de erro de estimação global os estimadores global e local possam produzir resultados similares, o estimador local possui a vantagem adicional de reproduzir a estrutura espacial de forma mais nítida, caso ela esteja presente.

Figura 3 - Estimativas do risco, segundo a taxa bruta e os Estimadores Bayesianos Empíricos



Nota: Global e Local (EBE global e EBE local, respectivamente) para uma das simulações do arranjo espacial H, com $0.10 \leq Pi \leq 0.40$

Figura 4 - Mapa das estimativas do risco, segundo a taxa bruta e o Estimador Bayesiano Empírico Global (EEB global) e Local (EEB local) com raio=4 para uma das simulações do arranjo espacial H, com $0.10 \leq P_i \leq 0.40$



Nota: Quanto mais clara a imagem maior é o valor do risco ou de sua estimativa no local.

4. Estimação do risco de malária via Estimador Bayesiano Empírico

Considerando os resultados favoráveis do estimador bayesiano empírico local na redução do erro global de estimação dos riscos e na recuperação do padrão espacial dos riscos, obtidos nos estudos de simulação anteriores, decidimos estimar o risco de malária nos lotes de Machadinho aplicando esse estimador aos dados de malária para homens com idade igual ou superior a 15 anos.

Para o cálculo do estimador bayesiano empírico local, definimos a região de vizinhança de cada lote como sendo formada por todos os lotes cujos centróides se encontram dentro de um círculo de raio igual a 5 km centrado no lote para o qual se quer calcular a estimativa. O valor 5 km foi escolhido por se tratar de um raio de influência do mosquito razoavelmente aceito pelos pesquisadores do grupo. Com raio de vizinhança igual a 5 km, o número médio de vizinhos de cada lote no ano de 1995 é 63.

A Figura 5 mostra o efeito suavizador do estimador bayesiano empírico local sobre as taxas brutas de malária nos lotes, reduzindo os valores muito altos e aumentando as taxas muito baixas. Como vimos nos estudos de simulação anteriores, essa correção reduz o erro global de estimação da taxa bruta, pois os valores extremos fornecidos pela taxa bruta para alguns lotes podem não corresponder a riscos extremos, mas serem resultado da alta variabilidade da taxa bruta.

Como nem todos os lotes da região estavam ocupados na época do estudo, a visualização da distribuição espacial da malária através de mapas das estimativas dos riscos é prejudicada pela presença de lotes sem estimativas atribuídas a eles. Para melhorar a interpretação dos mapas, atribuímos estimativas do risco aos lotes vagos através da interpolação das estimativas dadas pelo EBE local aos lotes ocupados. Primeiramente, fizemos uma interpolação linear para 10 000 pontos dispostos em uma grade regular na região de estudo usando os valores do EBE local dos lotes ocupados. Posteriormente, atribuímos a cada lote vago o valor da estimativa do risco no ponto dessa grade mais próximo de seu centróide. Esse procedimento é justificado pelo fato dos lotes vagos estarem uniformemente espalhados entre os lotes ocupados nos três anos do estudo e pelo fato do estimador bayesiano empírico fornecer estimativas suavizadas dos riscos reais e com menor variabilidade. Usar interpolação com as taxas brutas seria arriscado, dada a grande flutuação de seus valores.

Os mapas obtidos com esse procedimento para os anos de 1986, 1987 e 1995 são mostrados nas Figuras 6, 7 e 8. Inicialmente, podemos ver que o nível do risco de malária decresceu ao longo dos três anos, especialmente de 1987 para 1995. De fato, os valores médios das estimativas dadas pelo EBE local (incluindo as interpolações) para os anos de 1986, 1987 e 1995 são, respectivamente, 0.40, 0.32 e 0.08 (desvios padrões iguais a 0.24, 0.20 e 0.10).

Figura 5 - Estimativas da taxa bruta e do Estimador Bayesiano Empírico Local (EBE local) para risco de malária nos lotes com pelo menos um homem com idade igual ou superior a 15 anos

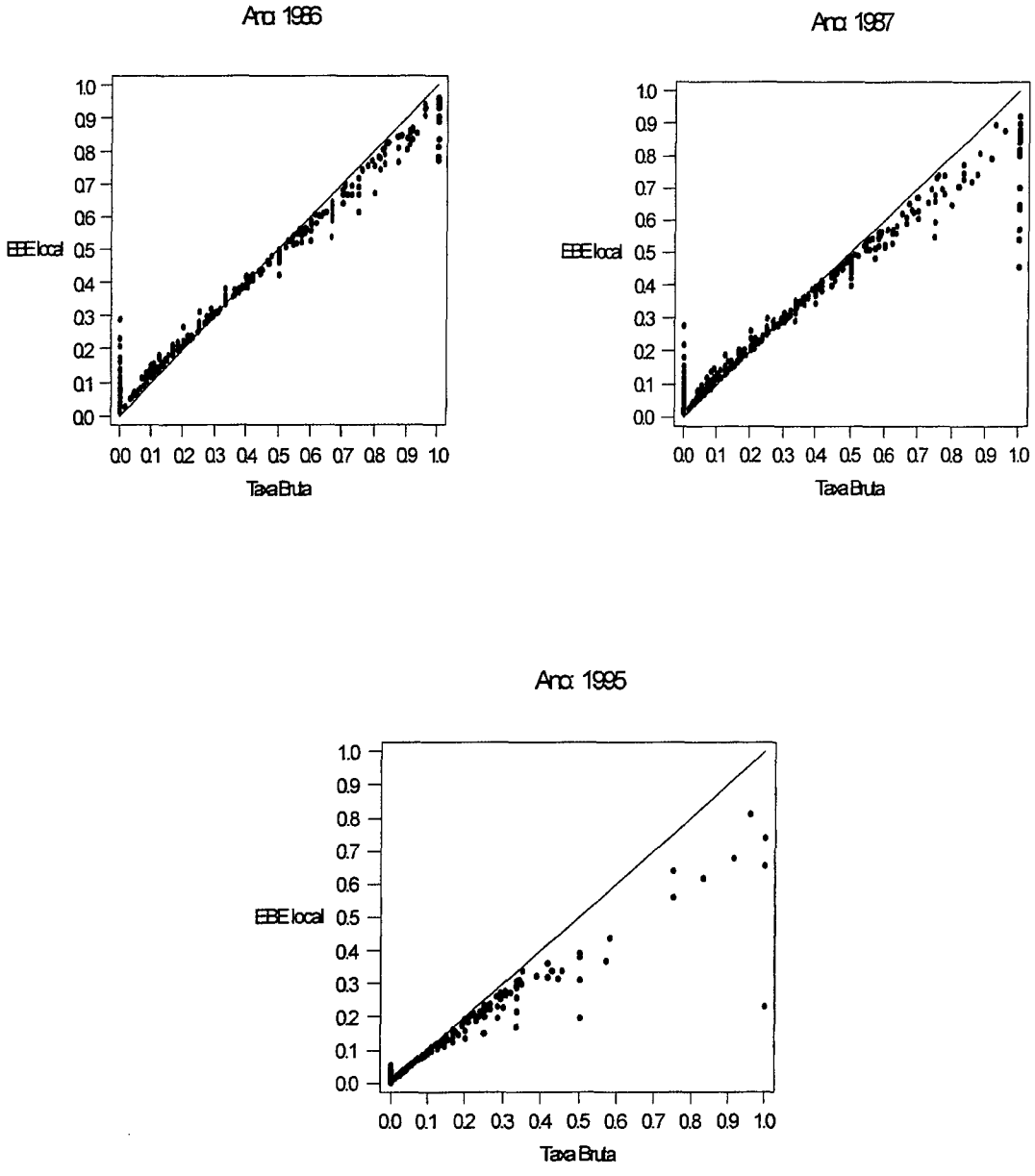


Figura 6 - Mapa das estimativas do EBE local para o risco de malária nos lotes com pelo menos um homem com idade igual ou superior a 15 anos em 1986

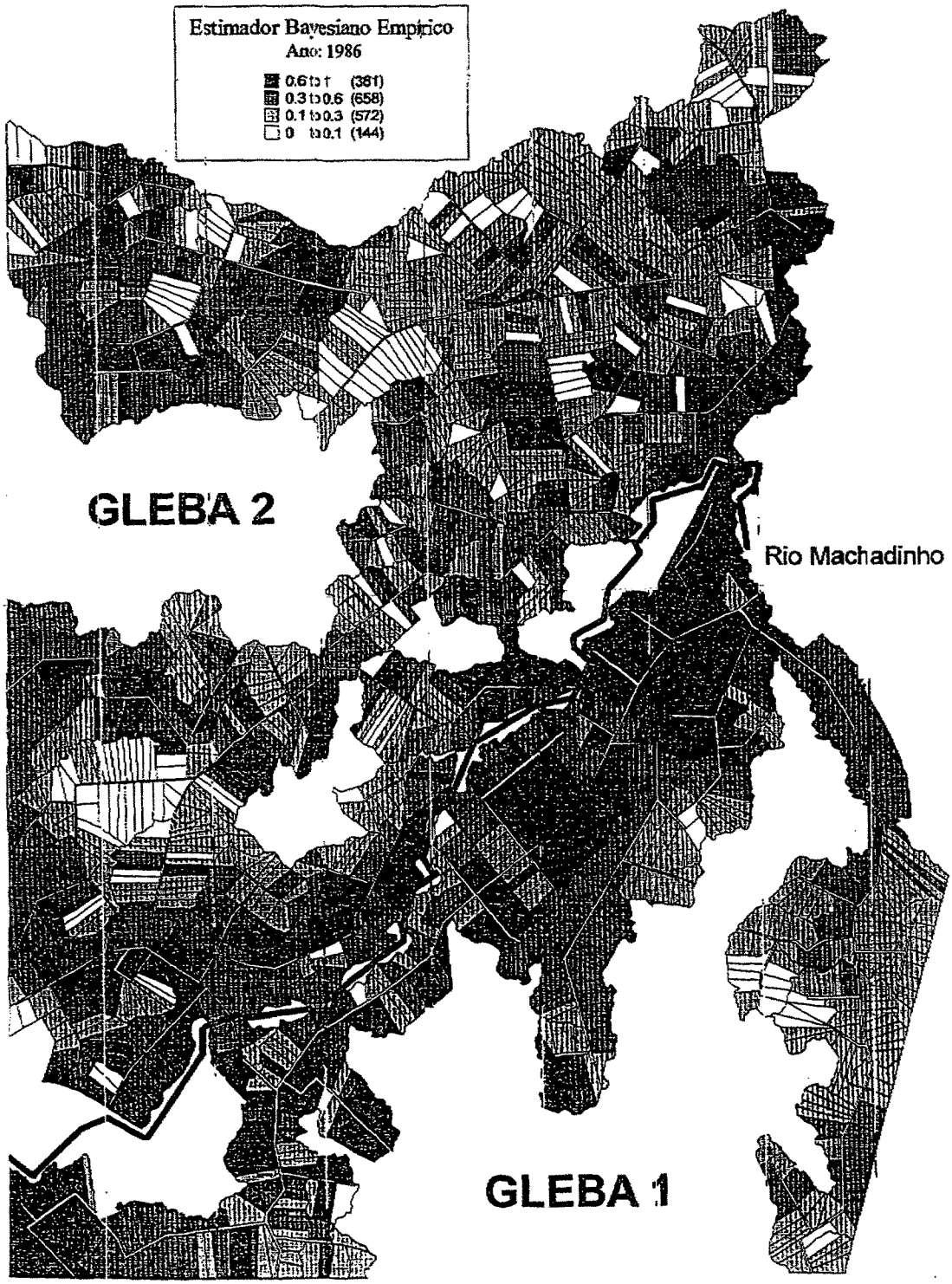
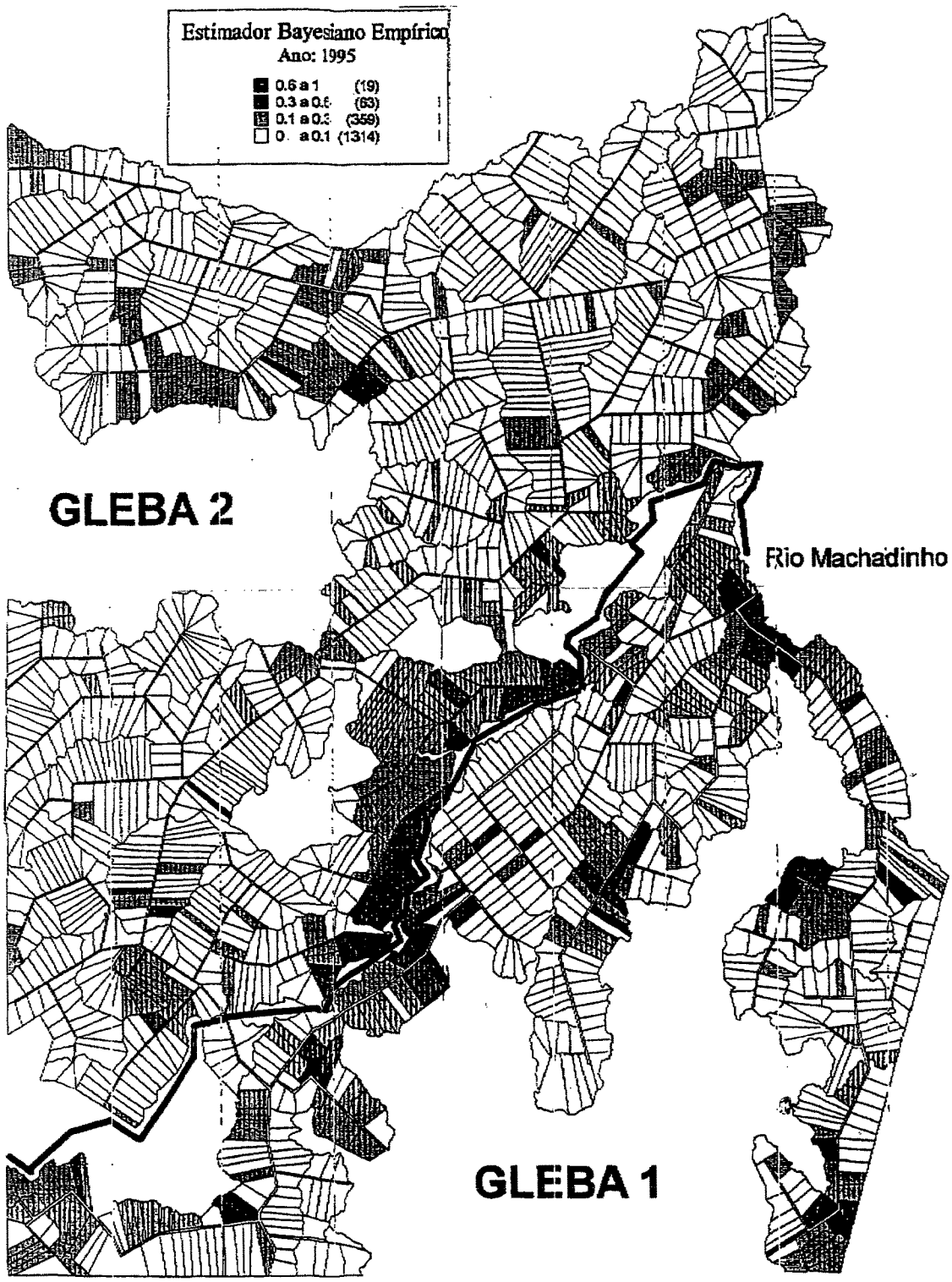


Figura 7 - Mapa das estimativas do EBE local para o risco de malária nos lotes com pelo menos um homem com idade igual ou superior a 15 anos em 1987

Figura 8 - Mapa das estimativas do EBE local para o risco de malária nos lotes com pelo menos um homem com idade igual ou superior a 15 anos em 1995



Notamos também que a Gleba 2 foi menos atingida que a Gleba 1. Em 1986, o valor médio do EBE local para a Gleba 1 foi 0.48 (dp=0.25), enquanto a Gleba 2 teve média igual a 0.35 (dp=0.21). Em 1987, o valor médio do EBE local para as Glebas 1 e 2 foi igual a 0.38 (dp=0.19) e 0.28 (dp=0.20), respectivamente. Em 1995, as médias foram 0.11 para a Gleba 1 (dp=0.12) e 0.07 para a Gleba 2 (dp=0.09).

Comparando os três mapas, podemos ver que, ao longo do tempo, os altos valores do risco tenderam a se dispersar do centro das glebas em direção às bordas próximas às reservas florestais e ao rio Machadinho.

As doenças infecciosas não se distribuem aleatoriamente ou homogeneamente na população. As causas das doenças são sempre múltiplas, um processo que envolve o agente infeccioso, o hospedeiro e o ambiente. A malária é uma doença claramente relacionada ao ambiente natural, no qual a presença de água e floresta, e os níveis de umidade e temperatura são fatores importantes. Para se reproduzir, o mosquito vetor precisa de água limpa, parcialmente sob sombra e com pouca ou nenhuma correnteza. A Região Amazônica tem todas essas condições ambientais favoráveis ao desenvolvimento do mosquito vetor. Assim, o ambiente natural representa um importante fator determinante dos altos níveis de transmissão de malária. Superposto a esse ambiente natural, está o intenso processo de ocupação humana da área, incluindo a chegada de imigrantes freqüentemente não imunes à doença, que se ocupam de atividades econômicas que os colocam em contato direto com a floresta, hábitat natural do mosquito vetor.

A estimação do risco de malária, utilizando o estimador bayesiano empírico local, usa a informação sobre malária em cada lote individual, mas também considera a similaridade espacial dos riscos dentro de uma vizinhança de cada lote. A interpretação dos mapas é uma análise visual que envolve apenas as hipóteses acerca de mudanças ocorridas no período de estudo que estão relacionadas ao meio ambiente, desde que é mais provável que fatores ambientais, mais que socioeconômicos, sejam similares entre lotes vizinhos. Um estudo mais profundo deveria levar em conta outras variáveis, tais como: o tempo de moradia no lote, condição socioeconômica, características da casa, dentre outras.

Dadas essas considerações, podemos fazer alguns comentários acerca dos aspectos ambientais relacionados à exposição e infecção de malária baseados nos mapas para o período de 1986 a 1995. Em 1986, os moradores tinham acabado de chegar e, por causa da necessidade de limpar os lotes para construir suas casas e plantar, tiveram intenso contato com o mosquito vetor. Isso provavelmente explica o alto risco de malária espalhado praticamente em toda a região. Em 1987, alguns desses moradores já tinham limpado a área próxima a suas casas e tiveram menos contato com o hábitat do mosquito, embora o macroambiente ainda seja uma fonte infindável de mosquitos. Em 1995, o nível de malária cai bastante e os focos com os valores mais altos estão basicamente concentrados nas bordas próximas às florestas e ao rio Machadinho.

Esses resultados estão em acordo com aqueles obtidos no estudo de Marchesine et al. (1996) sobre a incidência de malária nas Glebas 1 e 2 no ano de 1994. Nesse estudo, a distribuição espacial

da malária é estudada no nível de 133 localidades, que englobam um número variável de lotes. Para cada localidade, dispõe-se da informação do número de exames de sangue com resultados positivos para malária no período considerado. A variável mapeada é o Índice Parasitológico Anual – IPA – que é calculado dividindo-se o número total de resultados positivos dos 12 meses do ano pelo tamanho da população, ambos referentes à localidade da infecção. Analisando visualmente o mapa dos valores de IPA divididos em cinco categorias, os pesquisadores descobriram que, de maneira geral, as localidades situadas às margens do rio Machadinho e/ou às margens das reservas florestais foram aquelas que apresentaram os maiores valores de IPA. Além disso, observaram que os valores de IPA não parecem se distribuir aleatoriamente na região, sendo que apenas algumas localidades isoladas apresentam valores de IPA que se destoam das localidades vizinhas.

Os resultados do estudo apresentado neste artigo sugerem, de forma mais definitiva, que há uma forte influência do componente ambiental na distribuição espacial da malária em Machadinho ao considerar a análise espacial em uma escala micro como é o lote. Além disso, trabalhamos com informações que permitem calcular taxas de incidência ao invés do índice IPA.

5. Conclusões

Neste trabalho consideramos a construção de mapas do risco de malária nos lotes de assentamento rural da região de Machadinho, em Rondônia. O objetivo é buscar um estimador para os riscos nos lotes mais estável que o usual estimador de máxima verossimilhança (EMV). Propomos o uso da metodologia bayesiana empírica apresentada por Martuzzi e Elliott (1996) a partir do artigo de Marshall (1991). O estimador proposto é o estimador bayesiano empírico (EBE) adaptado para a distribuição binomial assumida para os dados de malária na forma em que foram coletados.

O estimador bayesiano empírico considera a informação de outros lotes da região durante a estimação do risco em um lote. Consideramos duas abordagens para esse estimador: o EBE global, que não considera a similaridade espacial dos riscos da doença entre os lotes e EBEs locais, que incorporam essa informação espacial na estimação dos riscos.

Nos estudos de simulação apresentados neste trabalho, o EBE mostrou-se superior ao EMV na estimação do risco da doença nos lotes, apresentando, em média, erro global de estimação entre 14% e 82% do erro do EMV. Assim, o EBE sofre menos efeito de flutuações aleatórias das quantidades observadas, fornecendo estimativas mais precisas para a construção de mapas de doenças com o objetivo de visualizar sua distribuição espacial. A escolha entre o EBE global ou local depende da estrutura espacial do risco, pois nenhuma das duas formas mostrou-se superior à outra em todas as situações. Entretanto, recomendamos o uso dos estimadores locais quando há razões para acreditar que os riscos são espacialmente correlacionados, pois o EBE local mostrou-se mais eficaz na recuperação desse padrão espacial dos riscos. Na ausência de estrutura espacial, seja porque o risco é constante em toda área ou porque a doença espalha-se de forma completamente aleatória, EBE global apresenta um erro global de estimação menor que os EBEs locais.

Os mapas das estimativas bayesianas empíricas locais dos riscos de malária nos lotes de Machadinho, referentes aos anos de 1986, 1987 e 1995, revelam uma queda no nível de malária ao longo desse período e a tendência de concentração de alta incidência nos lotes próximos às margens do rio que corta a região e/ou às margens das reservas florestais. Esses resultados concordam com aqueles obtidos em um estudo similar feito por Marchesine et al. (1996) sobre a distribuição de malária na mesma região, dividida, no entanto, em áreas muito maiores e sugerem que há, de fato, uma forte influência do componente ambiental na distribuição espacial da malária em Machadinho.

Referências Bibliográficas

ASSUNÇÃO, R. M., BARRETO, S.M., GUERRA, H. L., SAKURAI, E. (1998), Mapas de taxas epidemiológicas: a abordagem bayesiana. Cadernos de Saúde Pública, 1998.

BAILEY, M. T. J. (1982) *The biomathematics of malaria*. Charles Griffin & Company Ltd: High Wycombe.

BERNADINELLI, L., MONTONOLLI, C. (1992), Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, v. 11, p. 983-1007.

CASTILLA, R. E. F., SAWYER, D.O. (1993), Malaria rates and fate: a socioeconomic study of malaria in Brazil. *Social Science Medicine*, v. 37, n. 9, p. 1137-1145.

DEVINE, O. J., LOUIS, T. A., HALLORAN, E. (1994), Empirical Bayes Methods for stabilizing incidence rates before mapping. *Epidemiology*, v. 5, p. 622-630.

DONNELY, C. A. (1995), The spatial analysis of covariates in a study of environmental epidemiology. *Statistics in Medicine*, v. 14, p. 2393-2409.

EFRON, B., MORRIS, C. (1973), Stein's estimaton rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association*, v. 68, p. 117-130.

GRIFFIN, B.S., KRUTCHKOFF, R.G. (1971), Optimal linear estimators: an empirical Bayes version with application to the binomial distribution. *Biometrika*, v. 58, n. 1, p. 195-201.

GLASS, G. E., SCHWARTZ, B. S, MORGAN, J. M., JOHNSON, D. T., NOY, P. M., ISRAEL, E. (1995), Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health*, v. 85, p. 944-948.

MARSHALL, R. J. (1991), Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*, v. 40, n. 2, p. 283-294.

MARCHESINI, P. B., SPENCER, B., LIMA, M. C. (1996), Distribuição espacial da malária no município de Machadinho/RO, 1994. *Anais do* X Encontro Nacional de Estudos Populacionais. Belo Horizonte: ABEP, v. 4, p. 2427.

MARTUZZI, M., ELLIOTT, P. (1996), Empirical Bayes estimation of small area prevalence of non-rare conditions. *Statistics in Medicine*, v. 15, p. 1867-1873.

NUSSENZWEIG, V., NUSSENZWEIG, R. (1985), Malária: a vacina possível. *Ciência Hoje*, v. 3, n. 16, p. 26-35.

SAWYER, D.R., SAWYER, D.R.T.O. (1987.), Malaria on the Amazon frontier: economic and social aspects of transmission and control. Relatório de Pesquisa, CEDEPLAR/UFMG, Belo Horizonte.

SAWYER, D.R., SAWYER, D.R.T.O. (1995), A research and education initiative on human health and effective utilization of tropical forests. Relatório de Pesquisa, CEDEPLAR/UFMG, Belo Horizonte.

SAWYER, D.R.T.O. (1995), O papel da malária na mortalidade das áreas endêmicas no Brasil. Tese para concurso de Professor Titular, CEDEPLAR/UFMG, Belo Horizonte.

TAUIL, P. L. Malária. (1984), Agrava-se o quadro da doença no Brasil. *Ciência Hoje*, v. 2, n. 12, p. 58-64.

WALTER, S. D., BIRNIE, S. E. (1991), Mapping mortality and morbidity patterns: an international comparison. *International Journal of Epidemiology*, v. 20, p. 678-689.

WHITE, A. A. (1995), Mapping and geographic display of data. *Statistics in Medicine*, v. 14, p. 697-699.

ABSTRACT

We consider the malaria risk estimation in small settlement lots in a rural area (Machadinho) in Northern Brazil. We propose an empirical Bayes estimator adapted to the binomial characteristic of our data. This estimator takes in account the information of all other lots to estimate the risk in a specific lot. It can be adapted to take advantage of the spatial dependence among the lot's risks in order to improve estimation. Through simulation, we find our proposed estimator produces smaller total estimation error and it is more stable to recover the spatial distribution of the disease than the usual maximum likelihood estimator. The Machadinho malaria maps with risks estimated by the empirical Bayes proposal show that a strong environmental components in the incidence of the disease is persistent through time.

Desempenho das Escolas de Ensino Médio de Belo Horizonte no Vestibular da UFMG

José Francisco Soares*

Cibele Comini César*

José Aguinaldo Fonseca*

RESUMO

Nos últimos anos, diversas universidades brasileiras publicaram os escores médios alcançados em seus exames de admissão (vestibulares) para estudantes que vieram das mesmas escolas secundárias. O emprego desta metodologia para comparar escolas secundárias apresenta duas grandes limitações. A metodologia não considera as diferenças socioeconômicas entre os estudantes das diferentes escolas, e ignora a correlação entre as observações. Este artigo, baseado em resultados da primeira etapa do vestibular da Universidade Federal de Minas Gerais em 1997, trata destes problemas mediante incorporação da estrutura hierárquica dos dados num modelo de regressão. O modelo hierárquico fornece estimativas corretas dos parâmetros do modelo, uma medida do coeficiente de correlação entre escores de estudantes da mesma escola e, mais importante, uma medida do efeito de cada escola secundária. O modelo ajustado aos dados revela que a variação nos escores das escolas pode ser melhor explicada por diferenças socioeconômicas entre os estudantes de cada escola do que pelas práticas e políticas da escola. O modelo nulo estima que 35,9% da variabilidade dos escores dos estudantes é devida às escolas, proporção que cai para 19,6 quando os resultados são controlados pelas covariáveis dos estudantes.

1. Introdução

Nos últimos anos várias universidades brasileiras, entre elas a UFMG e a UFRJ, divulgaram

*Endereço para correspondência: Departamento de Estatística – UFMG.

sínteses do desempenho no vestibular das escolas de segundo grau. Usaram para isto ou a nota média dos vestibulandos, classificados pela escola de ensino médio freqüentada ou a porcentagem de alunos aprovados no concurso de cada uma das escolas consideradas. A forma de divulgação, com as escolas ordenadas pela nota média obtida por seus ex-alunos ou pela proporção de aprovados, revelava uma vontade subjacente de classificar as escolas. Rapidamente a imprensa publicou reportagens salientando as qualidades das escolas privadas, principalmente as confessionais, e criticando a qualidade das escolas públicas.

A classificação de escolas através da porcentagem de aprovados é opção metodológica falha, já que o vestibular é classificatório e há competitividade diferente nos diversos cursos. A classificação pela nota média dos ex-alunos retira o efeito do curso escolhido pelo aluno, mas assume que a clientela das diversas escolas é semelhante, em relação a todas as características relevantes para o processo educacional. Sabe-se, entretanto, que o escore final do aluno é influenciado não somente pela escola mas também por características do aluno e do ambiente no qual ele está inserido e que estas características não são igualmente distribuídas nas diferentes escolas. Em particular são grandes as influências das condições socioeconômicas e na habilidade natural e motivação dos alunos das diversas escolas. Portanto a comparação pelas médias atribui, indevidamente, qualquer diferença observada no resultado final do aluno à escola.

A discussão sobre o melhor método de comparar escolas realizou-se de forma acirrada na Inglaterra nos últimos anos, depois que o governo conservador estabeleceu como preceito legal a divulgação dos resultados dos alunos através de médias. Estas são publicadas em forma de tabelas - as *league tables*, veiculadas pela imprensa nacional e local. Estas tabelas apresentavam as escolas secundárias da Inglaterra e País de Gales, ordenadas pelo rendimento médio de seus alunos em exame nacional. A principal justificativa oficial para esta política era informar os pais sobre as melhores escolas, de forma que eles tivessem subsídios para escolher a escola para seus filhos (Goldstein e Thomas, 1996).

A partir da necessidade de se analisar criticamente as *league tables*, foram desenvolvidos modelos e metodologias alternativos, que, na medida que os dados comportam, são utilizados neste trabalho. A modelagem que se mostrou adequada está baseada na Teoria do Valor Agregado (Thomas e Mortimore, 1996; Goldstein e Thomas, 1996), através da utilização dos Modelos Hierárquicos (Bryk e Raudenbush, 1992).

Ao se considerar o nível do aluno no ingresso, retira-se a influência de processos anteriores e independentes da escola. O que deve ser observado na determinação da efetividade da escola é o que foi acrescentado ao aluno durante o período em que ele esteve naquele estabelecimento de ensino, isto é, entre a sua entrada e sua saída, e não o patamar alcançado ao final do processo. Ou seja, o foco da análise deve ser o progresso do estudante, dentro da instituição de ensino. É a este acréscimo de conhecimento observado no aluno que, muitas vezes, dá-se o nome de "Valor Agregado", embora este termo não tido completa aceitação entre todos os estudiosos do assunto (Goldstein a Spiegelhalter, 1996).

A consideração das características individuais e das variáveis de contexto tem por objetivo tornar comparáveis as clientelas das diversas escolas. A não obtenção da comparabilidade leva a resultados errôneos: existem escolas que, apesar de acrescentarem muito a seus alunos, devido ao perfil de sua clientela, não conseguem se destacar no quadro das médias de desempenho. Em suma, uma forma mais justa de comparar as escolas é julgá-las pelo que acrescenta ao aluno padrão e não pelo resultado final médio do seu grupo de alunos.

A utilização de Modelos Hierárquicos incorpora à análise a estrutura hierárquica dos dados. Na análise de desempenho escolar, isto se dá pela incorporação no modelo do reconhecimento que existe variabilidade relacionada às escolas, uma vez que se reconhece que estas diferem em eficiência e, desta forma, alunos de uma mesma escola tendem a ser mais semelhantes entre si que alunos de escolas distintas. Além disso, existe variabilidade dentro da escola, uma vez que os alunos não têm todos o mesmo rendimento. Em outras palavras, os modelos hierárquicos permitem a incorporação ao modelo de análise de mais de um termo de erro. Ao incorporar ao modelo a variabilidade proveniente dos dois níveis é possível quantificar a contribuição de cada um deles para a variação total. Mais ainda, através da introdução de covariáveis específicas a cada um dos níveis é possível explicar grande parte da variação presente nos dados e definir políticas facilitadoras do processo ensino-aprendizagem. Ao se incluir no modelo uma covariável escolar que tem efeito positivo no desempenho dos alunos, diminui-se a variabilidade entre escolas e detecta-se um fator potencialmente causal do bom rendimento do aluno. Desta forma, é possível definir políticas administrativas e pedagógicas que visem a uma maior eficácia da rede de ensino.

A utilização dos Modelos Hierárquicos na análise do dados ingleses mostraram que diferentes modelos levam a diferentes classificações das escolas (Goldstein e Thomas, 1996). Mostraram ainda que a introdução de variáveis de desempenho do aluno em testes anteriores (modelo baseado na Teoria do Valor Agregado) diminuía drasticamente a variabilidade dentro e entre as escolas, quando comparado com um modelo que não incluía esta informação (Goldestein e Thomas, 1996; Thomas e Mortimore, 1996). Conclui-se, então, pela imprecisão e inexatidão da tabela construída a partir dos dados não ajustados.

Para o Brasil ainda não é possível utilizar, na sua totalidade, a Teoria do Valor Agregado na comparação das escolas, por não existir uma informação básica: uma medida do conhecimento do aluno ao ingressar no nível no qual se deseja fazer a comparação. Assim, se desejamos comparar o desempenho das escolas no vestibular, que seria uma medida da eficiência da escola no nível médio, necessitaríamos de uma medida do aluno na entrada neste nível, ou seja, ao final do ensino fundamental.

É preciso enfatizar que o objetivo deste estudo é simplesmente produzir uma comparação das escolas da Grande Belo Horizonte que forneceram candidatos para o vestibular de 1997, através do desempenho destes alunos no concurso. Estudos muito mais interessantes, como o efeito das desigualdades sociais sobre o processo de ensino-aprendizagem no Brasil ou uma análise da estrutura das escolas sobre o desempenho de seus alunos, não são viáveis com os dados

disponíveis. No entanto, este artigo pretende chamar atenção para os modelos estatísticos necessários para este tipo de estudo.

2. Dados utilizados

O vestibular da UFMG é realizado em duas etapas. A primeira etapa consiste de provas de múltipla escolha, que verificam o conhecimento do conteúdo básico do ensino fundamental e médio, e é comum a todos os candidatos. A segunda etapa consiste de provas discursivas e é específica por curso. Somente uma parte dos candidatos, selecionada a partir dos resultados da primeira etapa, faz as provas da segunda etapa. Neste trabalho, somente o resultado na primeira etapa foi utilizado.

No vestibular da UFMG, em 1997, 45.994 candidatos concorreram a um dos 45 cursos que a universidade oferece. Deste total 3 365 (7,3%) não compareceram a nenhum dia das provas. Como objetivo deste trabalho é a crítica a uma metodologia de comparação de escolas, apenas os alunos que terminaram o segundo grau em 1996 e que compareceram a pelo menos um dia de prova foram considerados. As escolas representadas por quatro ou menos alunos não foram analisadas. Desta orma somente 163 escolas e um total 10.476 alunos foram incluídos na análise.

A limitação do estudo aos alunos que concluíram o segundo grau em 1996 tem duas razões principais. A primeira seria evitar o "efeito coorte" que poderia existir ao se considerar alunos que cursaram o nível médio em épocas distintas. Além disto, grande parte dos candidatos que concluíram o 2º grau em anos anteriores, se matricularam em cursinhos pré-vestibular. Sendo a ação deste sobre o aluno mais próxima temporalmente do exame vestibular, existe o risco de exercer maior impacto sobre o desempenho do aluno que a escola de conclusão do 2º grau. Certamente que muitos candidatos recém-egressos do 2º grau freqüentaram cursinho pré-vestibular e o controle para isto é eito através de sua inclusão em um dos fatores considerados como covariáveis no modelo.

Para se inscrever ao vestibular da UFMG, o candidato preenche junto com o requerimento de matrícula um questionário com informações demográficas e socioeconômico e culturais. Uma descrição detalhada do questionário foi feita em Soares e Fonseca (1998). As informações deste questionário foram utilizadas para fazer uma categorização socioeconômica dos alunos e para a construção de alguns fatores explicativos do desempenho. A seleção dos itens a serem incluídos em cada fator foi feita através de um estudo descritivo de sua associação com o desempenho do candidato (Soares e Fonseca, 1998). Os fatores considerados procuram medir quatro dimensões e foram identificados como habilidade acadêmica do aluno (***habilidade***), qualidade do ambiente escolar freqüentado (***escolar***), preparação para o vestibular (***preparo***) e condições socioeconômicas da família (***posição***). Os fatores habilidade, escolar e preparo foram construídos somando-se variáveis indicadoras, onde o valor 1 indica a presença de condições favoráveis a um melhor desempenho no vestibular. O fator posição foi construído a partir de variáveis contendo informações sobre escolaridade do pai, ocupação do pai e renda familiar. Todas as três variáveis foram codificadas em seis níveis, aos quais foram atribuídos postos de 1 a 6. O fator posição é a soma dos postos relativos

às três variáveis. O Quadro 1 apresenta as descrições das categorias das variáveis, segundo os postos atribuídos.

Quadro 1 - Categorização das variáveis Renda familiar, Escolaridade do pai e Ocupação do pai, segundo os postos atribuídos

Posto	Renda familiar	Escolaridade do pai	Ocupação do pai
1	Até 2 SM	Nenhuma	Do lar
2	De 2 a 5 SM	1º grau incompleto	Muito Baixo
3	De 5 a 15 SM	1º grau completo	Baixo
4	De 15 a 20 SM	2º grau completo	Médio
5	De 20 a 40 SM	Superior incompleto	Alto
6	40 SM e MAIS	Superior completo	Muito Alto

As tabelas que se seguem apresentam a nota média dos candidatos para cada nível dos fatores e justificam seu uso em modelos explicativos do desempenho. Para a confecção destas tabelas foram utilizados somente os candidatos analisados neste trabalho e não todos aqueles que se submeteram ao vestibular. Na leitura das tabelas deve-se ter em mente que a porcentagem de aprovados é calculada com relação ao total de inscritos pertencentes àquela categoria enquanto que a porcentagem de inscritos é calculada em relação ao total de candidatos na amostra.. Maiores detalhes sobre a construção destes fatores e os resultados obtidos com todos os candidatos podem ser encontrados em Soares e Fonseca (1998).

2.1 Fator *preparo*

O fator *preparo* foi construído somando os valores assumidos pelas variáveis indicadoras relativas à freqüência, a cursinho pré-vestibular e exercício de atividade remunerada, tomado como indicação de tempo disponível para estudo.

Tabela 1 -Desempenho médio dos candidatos, segundo os níveis do fator *preparo*

FATOR preparo	Aprovados (%)	Nota Média	Inscritos (%)
0 característica	4,4	40,3	15,5
1 característica	8,4	47,4	51,3
2 características	13,9	53,8	33,2
Total	1 010(9,6%)	48,4	10 476 (100%)

Como se pode ver na Tabela 1 os candidatos com índice zero obtiveram, além de notas mais baixas, também um baixo índice de aprovação, comparados com os candidatos na situação oposta.

Muitos candidatos têm as duas características que indicam boas condições para se preparar para o vestibular.

2.2 Fator *escolar*

O fator *escolar* foi construído somando os valores assumidos pelas variáveis indicadoras relativas a: ter cursado o científico; ter concluído o 2º grau, em Belo Horizonte ou Exterior; ter cursado a maior parte do 2º grau em turno diurno. Todos os alunos da amostra apresentavam pelo menos uma das características, não tendo sido observado, portanto, o valor 0 para o fator *escolar*.

Tabela 2 - Desempenho médio dos candidatos, segundo os níveis do fator *escolar*

FATOR escolar	Aprovados (%)	Nota Média	Inscritos (%)
1 característica	2,0	37,5	7,4
2 características	5,0	41,0	22,9
3 características	12,0	52,0	69,7
Total	1 010 (9,6%)	48,4	10 476 (100%)

Candidatos que freqüentaram ambientes escolares onde é proporcionado melhor qualidade no ensino (três características), obtiveram notas na 1ª etapa e taxas de aprovação muito superiores, quando comparados tanto com o total dos candidatos quanto com o subgrupo formado pelos candidatos que apresentaram uma ou duas características. Considerando ainda o baixo percentual de candidatos com apenas uma característica, optou-se, nas análises subseqüentes, por considerar apenas duas categorias para o fator *escolar,* agrupando os estudantes com até duas características no nível 0 e os que apresentavam três características receberam o valor 1.

2.3 Fator *habilidade*

A mensuração correta do fator que retrata a habilidade acadêmica prévia do candidato exige o conhecimento de resultados de testes anteriores ao vestibular. No Brasil, esse dado não existe. Para este trabalho utilizamos itens do questionário socioeconômico-cultural para medir indiretamente este fator, que foi construído somando as variáveis indicadoras das seguintes situações: ter concluído o 2º grau com até 18 anos; não ter tido reprovações no 2º grau; ter domínio de alguma língua estrangeira; ter prestado vestibular anteriormente com aprovação ou sem ter concluído o 2º grau.

Tabela 3 - Desempenho médio dos candidatos, segundo os níveis do fator *habilidade*

FATOR habilidade	Aprovados (%)	Nota Média	Inscritos (%)
0 característica	1,9	38,9	8,0
1 características	3,2	40,0	17,7
2 características	5,5	44,8	31,3
3 características	13,7	53,4	27,7
4 características	22,1	61,4	15,3
Total	1 010 (9,6%)	48,4	10 476 (100%)

O fator habilidade pode ter no máximo quatro pontos e é sintetizado na Tabela 3. As notas crescem à medida que aumenta a "habilidade" acadêmica do candidato, medida pelo número de características consideradas no fator. Os candidatos de baixa "habilidade" obtiveram, como era de se esperar, taxas de aprovação muito baixas e notas abaixo da nota média global. É importante ressaltar a enorme diferença nas taxas de aprovação e notas observadas entre candidatos de baixa e alta habilidade. Isto, de certa forma, mostra que essa variável, apesar de sua deficiência estrutural, consegue discriminar bem os candidatos.

2.4 Fator *posição*

Este fator, embora discreto, será, nas análises subseqüentes, considerado contínuo, assumindo valores no intervalo [4 ; 18], com média igual a 12 e desvio padrão igual a 3,6 pontos.

Tabela 4 - Desempenho médio dos candidatos, segundo os níveis do fator *posição*

FATOR posição	Aprovados (%)	Nota Média	Inscritos (%)
Até 9 pontos	4,2	40,3	27,2
De 9 até 12 pontos	7,4	46,3	22,9
De 12 até 15 pontos	10,6	51,3	26,6
Acima de 15 pontos	17,1	56,7	23,2
Total	1 010 (9,6%)	48,4	10 476 (100%)

As notas crescem linearmente com o aumento da posição socioeconômica dos candidatos. Candidatos com alto valor para o fator posição (acima de 15 pontos – em torno de 23% dos candidatos) obtiveram nota média e taxa de aprovação bem acima dos valores totais.

3. Modelos de análise

Um modelo para análise do desempenho dos alunos deve considerar a estrutura hierárquica presente nos dados. Esta estrutura advém do fato que os alunos estão agrupados em escolas e, existindo efeito de escola, espera-se que alunos provenientes de uma mesma escola tenham desempenhos dependentes. Conseqüentemente, o modelo necessita incorporar explicitamente as diferenças entre escolas.

Com o objetivo de diminuir o efeito de composição, decorrente do fato que diferentes escolas têm perfis de alunos distintos, e tornar comparáveis os resultados das diversas escolas, são introduzidas covariáveis específicas do aluno. A variável resposta - total de pontos na primeira etapa do vestibular - é modelada como uma combinação linear destas covariáveis e do nível da escola, representado por uma constante (intercepto) própria para cada escola. Em suma, o modelo utilizado é um modelo hierárquico em dois níveis, sendo o nível inferior relativo ao aluno e o nível superior à

escola, no qual supõe-se que as escolas apresentam médias distintas, representadas pelos interceptos, mas que os efeitos das covariáveis são fixos.

Temos, então,

$$Y_{ij} = \beta_{0j} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_q X_{qij} + \varepsilon_{ij} \qquad (1) \quad \text{e} \quad \beta_{0j} = \beta_0 + u_j \qquad (2)$$

Substituindo (2) em (1):

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_q X_{qij} + u_j + \varepsilon_{ij}$$

onde,

Y_{ij} – total de pontos obtidos pelo i - ésimo aluno e j - ésima escola,

$i = 1, 2, \cdots, n_j \qquad j = 1, 2, \cdots, k, \qquad$ onde $k =$ número de escolas

X_{lij} – valor da covariável l para o aluno i da escola j,

$l = 1, 2, \cdots, q$

β_l – efeito da covariável X_l sobre o total de pontos

β_{0j} – intercepto relativo à escola j

β_0 – valor esperado de β_{0j}, ou seja $E(\beta_{0j})$

Além disto, as componentes de erro u_j e ε_{ij} são supostas independentes, com distribuição. Normal com média 0 e variâncias σ_u^2 e σ_ε^2.

A comparação entre as escolas é feita através dos valores $\hat{u}_j$, que representam os valores estimados do efeito da escola, ajustados pelas covariáveis presentes no modelo. A princípio, u_j poderia ser considerado um efeito fixo a ser estimado - o efeito da escola na nota do aluno. Desta forma, estimaríamos k parâmetros, representativos dos efeitos de escola na nota dos alunos. Entretanto, duas razões levam a considerá-lo como um componente aleatório. Do ponto de vista da Estatística, número de escolas presentes no estudo exigiria um grande número de parâmetros a serem estimados com a conseqüente ineficiência do procedimento de estimação dos parâmetros. Do ponto de vista da análise substantiva do problema, é interessante considerar o efeito de escola como aleatório, uma vez que esta opção nos propicia a oportunidade de avaliar a extensão da variabilidade entre escolas que é explicada pelos fatores incorporados no modelo.

Vários modelos podem ser ajustados aos dados, dependendo de quais variáveis explicativas são incorporadas. No entanto, primeiramente, é importante ajustar um modelo que incorpora apenas a escola onde o aluno concluiu o segundo grau. Comparando a variabilidade entre as escolas obtida neste modelo que não considera as características dos alunos com aquela obtida do modelo contendo as covariáveis tem-se o indicativo mais explícito da inadequação da comparação de médias na avaliação do desempenho das escolas.

O Modelo 1, que incorpora apenas a escola é escrito da forma que segue:

$$Y_{ij} = \beta_{oj} + \varepsilon_{ij} \qquad \beta_{oj} = \beta_o + u_j$$

ou,

$$Y_{ij} = \beta_o + u_j + \varepsilon_{ij}$$

Considerando as suposições sobre as componentes de erro, tem-se:

$$Y_{ij} \sim N(\beta_0; \sigma_u^2 + \sigma_\varepsilon^2)$$

A síntese do ajuste deste modelo é apresentada na Tabela 5.

Tabela 5 - Resultado do Ajuste do Modelo, contendo apenas o indicador de escola

	Estimativa (Erro Padrão)	Valor-p
Parte fixa		
Intercepto (β_0)	**42,947 (0.657)**	**0,000**
Parte aleatória		
Variação total	**180,417**	
Entre escolas (σ_u^2)	**64,511**	
Entre alunos (σ_ε^2)	**115,907**	
Correlação intra-escolas	**0,359**	

Observa-se que a variância total fica divida em duas partes- uma devida à variação entre os alunos e outra devida à escola. O coeficiente de correlação intra-escolas fornece a fração da variação total observada nas notas que é explicada pela variação entre as médias das escolas. Desta forma, no primeiro modelo ajustado temos, aproximadamente, 36% da variabilidade presente na amostra que pode ser explicada pelas escolas.

Em um segundo momento, incluímos covariáveis específicas para os alunos, com o objetivo que alcançar comparabilidade. As covariáveis incluídas foram aquelas já discutidas anteriormente - ***habilidade***, ***preparo***, ***escolar*** e ***posição*** - acrescidas da covariável ***sexo***. Considerou-se mais adequado tratar as três primeiras covariáveis - ***habilidade***, ***preparo*** e ***escolar*** - como categóricas e introduzi-las no modelo através de variáveis indicadoras. Como as covariáveis têm, respectivamente, cinco, três e dois níveis, foram criadas as variáveis indicadoras ***habil1***, ***habil2***, ***habil3*** e ***habil4***; ***preparo1*** e ***preparo2*** e ***escolar***. A covariável ***sexo*** é, por sua própria natureza, indicadora. Atribuiu-se o valor 1 para o sexo feminino. A variável ***posição*** foi considerada contínua e incluída no modelo na forma de escore padronizado. A opção por esta forma de inclusão tem duas razões. Uma vez que ***posição***, nível socioeconômico do aluno, não assume o valor 0, para dar significado claro ao intercepto β_0, é interessante que seja feita uma mudança de origem na covariável. Optou-se por centralizá-la em sua média geral, 12,44. Além disto, para facilidade de interpretação, dividiu-se o desvio em relação à média pelo desvio padrão da covariável, 3,55. Assim, β_0 representa a nota média do aluno que tem nível socioeconômico igual à media do grupo e às demais covariáveis iguais a zero.

Embora fosse razoável considerar que a dificuldade de classificação no curso de escolha do candidato pudesse influenciar o seu desempenho, a inclusão no modelo do número mínimo de pontos para classificação à segunda etapa, observado no ano anterior, como uma covariável de aluno não apresentou coeficiente significativamente diferente de zero e tampouco produziu modificações no modelo anterior. Por estas razões a covariável não foi mantida no modelo.

Diante destas considerações, o Modelo 2 se escreve como:

$$Y_{ij} = \beta_0 + \beta_1 sexo_{ij} + \beta_2 posição_{ij} + \beta_3 preparo1_{ij} + \beta_4 preparo2_{ij} + \beta_5 habil1_{ij} + \beta_6 habil2_{ij} + \beta_7 habil3_{ij} + \beta_8 habil4_{ij} + \beta_9 escolar_{ij} + u_j + \varepsilon_{ij}$$

Uma observação importante relativa ao modelo diz respeito às covariáveis incluídas no modelo. Ainda que possa haver discordância com relação à classificação do fator escolar, uma vez que ele engloba características que podem ser consideradas da escola, estão presentes no modelo basicamente covariáveis medidas a nível do aluno. Não foram incluídas a características institucionais - sejam estas em nível de composição de seus alunos, de recursos educacionais disponíveis ou de propostas pedagógicas. Assim sendo, não é possível através desta análise avaliar políticas institucionais internas às escolas. O modelo permite apenas comparar o desempenho das escolas, ajustado pelo perfil de seus alunos. Em outras palavras, o modelo analisado serviria para guiar pais e alunos na escolha da escola. Não seria, entretanto, adequado para avaliação do sistema educacional com vistas à definição de políticas de intervenção no sistema.

A Tabela 6 apresenta o resultado do ajuste do Modelo 2.

Tabela 6 - Resultado do ajuste do Modelo 2, contendo o indicador de escola e as covariáveis específicas por aluno

	Estimativa (Erro Padrão)	*Valor p*
Parte fixa		
intercepto	**39.770 (0.598)**	**0.000**
sexo	**-4.125 (0.207)**	**0.000**
posição	**1.030 (0.138)**	**0.000**
preparo1	**0.718 (0.315)**	**0.023**
preparo2	**3.303 (0.357)**	**0.000**
habil1	**0.524 (0.426)**	**0.220**
habil2	**2.591 (0.416)**	**0.000**
habil3	**6.078 (0.447)**	**0.000**
habil4	**9.847 (0.514)**	**0.000**
escolar	**3.572 (0.330)**	**0.000**
Parte aleatória		
Variação total	**126.114**	
Entre escolas	**24.708**	
Entre alunos	**10.106**	
Correlação intra-escolas	**0.196**	

Comparando as Tabelas 5 e 6 podemos observar que a inclusão das covariáveis diminuiu consideravelmente a variação total não explicada pelo modelo que passou de 180,417 para 126,114 - uma redução em torno de 30%. Esta redução, embora tenha ocorrido nos dois níveis, foi muito mais acentuada a nível da escola, tendo passado de 64,511 para 24,708 – uma redução de 62%, enquanto em nível do aluno a redução observada, de 115,906 para 101,406, foi da ordem de 13%. Como resultado da ação assimétrica das covariáveis sobre as duas fontes de variação, com maior impacto sobre o nível escola, o coeficiente de correlação intra-escolas passou de 0,358 para 0,196, ou seja, aproximadamente 55% do valor original. Conclui-se, então, que embora exista variação entre as escolas, esta não é tão expressiva como a comparação das médias não ajustadas parece indicar.

O sinal negativo e altamente significativo da covariável sexo contraria os resultados de outros estudos empíricos semelhantes a este. Uma possível explicação é o vício de seleção a que estes dados certamente estão submetidos. Este vício seria produzido pela ausência no vestibular de candidatos que, avaliando suas poucas chances de aprovação não comparecem, enquanto as candidatas de mesmo perfil, por outro lado, são incentivadas, por motivos sociais variados, a se inscreverem no vestibular. Assim sendo, estaria observando-se no vestibular um excesso de candidatas com desempenho baixo.

Uma questão que é freqüentemente colocada na discussão da eficiência das escolas é o desempenho da escola pública. Sendo esta, diferentemente da privada, a princípio acessível a todos os membros da sociedade, podendo assim desempenhar o papel de provedora de "igualdade de oportunidades" e sendo, além disto, mantida pela sociedade, torna-se alvo de especulações e críticas. Existe, atualmente, uma grande discussão sobre a real qualidade da escola pública, cujo desempenho é confrontado com o da escola particular.

Uma vez que é possível fazer a classificação das escolas segundo o caráter público/privado, foi ajustado o Modelo 3 que, além de incluir as covariáveis individuais presentes no Modelo 2, contém também a covariável indicadora ***particular***, que assume o valor 1 para as escolas particulares e 0 para as escolas públicas. A Tabela 7 apresenta os resultados obtidos deste ajuste.

Tabela 7 - Resultado do ajuste do Modelo 3, contendo o indicador de escola, as covariáveis específicas por aluno e o indicador particular

	Estimativa (Erro Padrão)	Valor-p
Parte fixa		
intercepto	**38.075 (0.746)**	**0.000**
sexo	**-4.111 (0.207)**	**0.000**
posição	**0.977 (0.139)**	**0.000**
preparo1	**0.681 (0.316)**	**0.031**
preparo2	**3.247 (0.358)**	**0.000**
habil1	**0.536 (0.426)**	**0.209**
habil2	**2.627 (0.416)**	**0.000**
habil3	**6.086 (0.447)**	**0.000**
habil4	**9.841 (0.514)**	**0.000**
escolar	**3.490 (0.330)**	**0.000**
particular	**3.132 (0.847)**	**0.000**
Parte aleatória		
Variação total	**124.448**	
Entre escolas	**23.072**	
Entre alunos	**10. 376**	
Correlação intra-escolas	**0.185**	

Embora a inclusão da covariável ***particular*** praticamente não tenha afetado os valores relativos à variabilidade não explicada pelo modelo e os coeficientes estimados para as demais covariáveis no modelo, seu coeficiente foi significativo, o que justifica a sua manutenção no modelo.

Argumenta-se na literatura educacional (Willms, 1992) que não somente as características do aluno exercem efeito sobre o seu desempenho, mas também o contexto em que este aluno está inserido. Com o objetivo de contextualizar a análise foi ajustado o Modelo 4 que incorpora ao Modelo 3 a covariável , **posição média**, que representa o nível socioeconômico médio da escola.

O Modelo 4 se escreve como:

$$Y_{ij} = \beta_0 + \beta_1 sexo_{ij} + \beta_2 posição_{ij} + \beta_3 preparo1_{ij} + \beta_4 preparo2_{ij} + \beta_5 habil1_{ij} + \beta_6 habil2_{ij} + \beta_7 habil3_{ij} + \beta_8 habil4_{ij} + \beta_9 escolar_{ij} + \beta_{10} particular_j + \beta_{11} posição\_média_j + u_j + \varepsilon_{ij}$$

A Tabela 8 apresenta o resultado do ajuste do Modelo 4.

A introdução da covariável posição média, produziu uma redução de aproximadamente 25% na variação entre escolas, não alterando a variação entre aluno. Os coeficientes relativos às covariáveis medidas em nível do aluno não sofreram mudanças, a menos da covariável indicadora da posição socioeconômica do aluno que teve seu efeito um pouco reduzido, sem entretanto perder a significância. A covariável particular, por sua vez, sofreu profunda alteração e seu efeito anteriormente positivo passou a ser negativo e ainda significativo. A evidência captada dos dados mostra que o efeito do sistema privado de ensino está associado ao tipo de clientela que atende e ao fato de que os alunos com características favoráveis a melhor desempenho freqüentam as mesmas escolas. Com isto há uma sinergia de fatores com resultados muito expressivos.

Tabela 8 - Resultado do ajuste do Modelo 4, contendo as covariáveis específicas do aluno e as covariáveis de escola

	Estimativa (Erro Padrão)	*Valor-p*
Parte fixa		
intercepto	**40.810 (0.820)**	**0.000**
sexo	**-4.107 (0.207)**	**0.000**
posição	**0.874 (0.140)**	**0.000**
preparo1	**0.648 (0.316)**	**0.040**
preparo2	**3.190 (0.358)**	**0.000**
habil1	**0.531 (0.426)**	**0.213**
habil2	**2.614 (0.416)**	**0.000**
habil3	**6.036 (0.446)**	**0.000**
habil4	**9.748 (0.514)**	**0.000**
escolar	**3. 272 (0.331)**	**0.000**
particular	**-2.064(1.112)**	**0.065**
Posição média	**1.439 (0.228)**	**0.000**
Parte aleatória		
Variação total	**119.396**	
Entre escolas	**18.047**	
Entre alunos	**101.349**	
Correlação intra-escolas	**0.151**	

4. Resultados

A Tabela do Apêndice, apresenta os efeitos de cada escola segundo os modelos 1, 3 e 4. Considerando a média não ajustada – Modelo 1 - o efeito de escola é dado pela diferença entre a média observada para os alunos daquela escola e a média geral. Para os modelos 2 a 4, o efeito de escola é dado pelo resíduo de nível 2, $\hat{u}_j$, como explicado anteriormente. As escolas encontram-se ordenadas pelo efeito estimado através da média não ajustada e o posto 1 é dado à escola de menor valor agregado.

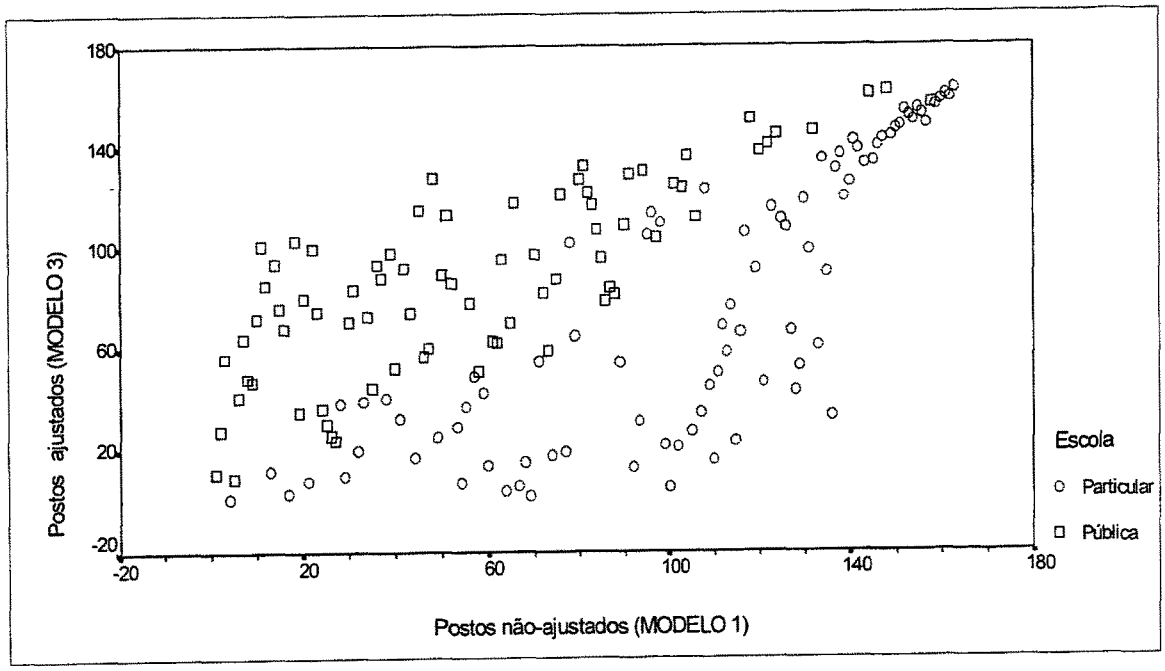
Serão considerados nesta discussão os modelos 1, 3 e 4 pelas razões explicitadas a seguir. O Modelo 1 é aquele subjacente à comparação das médias, prática mais utilizada para comparação de escolas. Os modelos 3 e 4 ajustam pelas covariáveis de aluno e por covariáveis de escola, sendo que o Modelo 3 inclui apenas a covariável indicadora do sistema administrativo a que a escola está inserida e o Modelo 4 considera também o contexto socioeconômico da escola. Estes dois modelos são contemplados na discussão, uma vez que a comparação entre eles torna mais clara a discussão sobre o papel da escola particular.

Observa-se primeiramente que, quando se controla pelas características dos alunos e pelo sistema administrativo, o efeito de escola é reduzido. O intervalo de variação que era $[-11{,}2\ ;\ 29{,}2]$ passa para $[-10{,}7\ ;\ 19{,}1]$ no modelo ajustado pelo indicador de escola particular, com uma clara redução das diferenças. A inclusão da covariável de contexto diminui ainda mais o comprimento do intervalo que passa a ser $[-10{,}4\ ;\ 15{,}3]$.

Outro resultado importante é a mudança de posição das escolas. Comparando os postos obtidos para os modelos 1 e 3, verifica-se que escolas que trabalham com clientela de posição social mais baixa, mas que fazem um bom trabalho, tem o seu valor agregado aumentado. O oposto vale para escolas que trabalham com alunos de posição social mais alta. A tabela com as escolas e seus respectivos postos e efeitos estão em anexo no final do trabalho.

Para melhor visualização dos resultados da tabela em anexo, o Gráfico 1 apresenta os postos atribuídos às escolas pelos dois modelos (Modelo 1 e 3). Se as classificações fossem exatamente equivalentes, todos os pontos estariam na diagonal principal do gráfico. Quanto maior o desvio em relação à diagonal, maior a discrepância no resultado obtido pelas duas análises. Observa-se que somente no canto superior direito da figura há concordância entre as duas classificações. No restante da figura, observa-se discrepância entre as classificações.

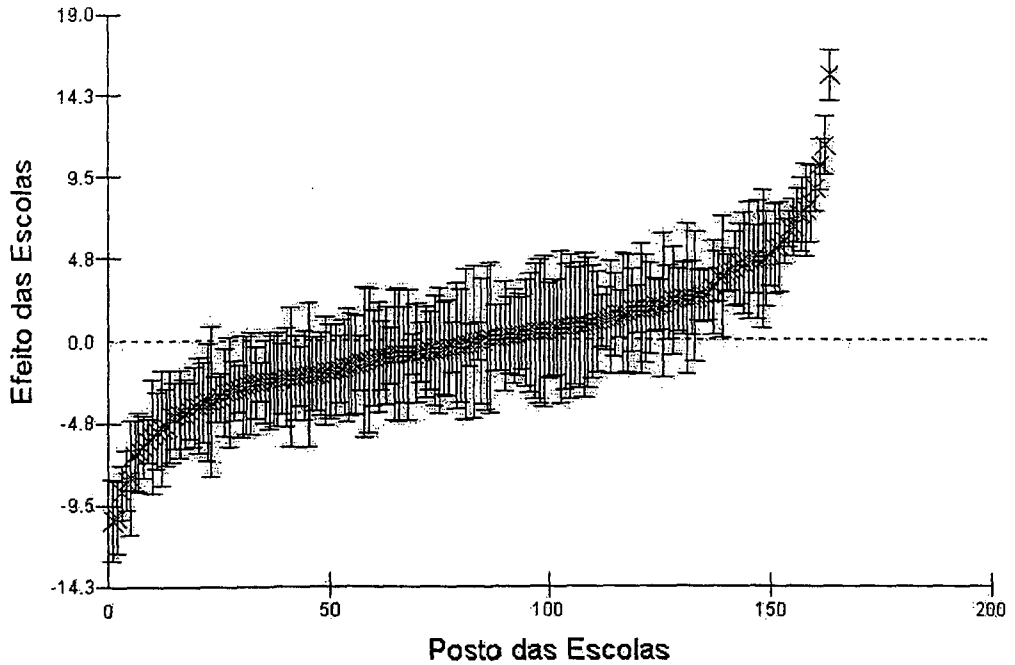
Gráfico 1 - Postos atribuídos às escolas pelos Modelos 1 e 3



Devido à incerteza relacionada ao efeito da escola, a comparação entre instituições não pode se basear em valores pontuais. A imprecisão da medida deve ser levada em conta e, portanto, devem ser construídos intervalos de confiança.

O Gráfico 2 apresenta os intervalos de confiança para os efeitos de escola, segundo o Modelo 4. A grande maioria destes intervalos apresenta intercessão e, portanto, não se pode considerar que as escolas diferem quanto à eficiência. Embora seja possível definir grupos de escolas com valores agregados distintos, a maioria delas é classificada em um grande grupo, com eficácia inferior às demais.

Gráfico 2 - Intervalos de confiança para os efeitos de escola - Modelo 4



5. Conclusões

Os resultados dos modelos ajustados mostram que grande parte da variabilidade observada nos resultados dos alunos de diferentes escolas pode ser explicada pelos fatores introduzidos no modelo que se referem primordialmente ao aluno. Assim, a ordenação das escolas aqui obtida se mostra mais justa que aquela baseada simplesmente nas média dos resultados, já que é feito o controle do efeito de fatores que, embora influenciando o desempenho, não podem ser modificados por políticas internas à escola.

Fica claro também que os dados obtidos quando da realização do vestibular não permitem estudo completo do desempenho das escolas de segundo grau. O acesso dos alunos às diferentes escolas é determinado, primeiramente, pela questão socioeconômica. As escolas que mostraram melhor desempenho são aquelas, sabidamente, inacessíveis à grande maioria da população. Como a condição socioeconômica determina também o ambiente no qual o aluno está inserido e este, por sua vez, tem forte influência no resultado acadêmico do aluno, há uma confusão de efeitos dificilmente desemaranháveis. Além disto, nem todos os alunos de todas as escolas se candidatam ao vestibular e não há nenhuma informação sobre as respectivas proporções de inscritos no vestibular da UFMG. Para uma comparação completamente não viciada seria necessário obter, concomitantemente, dados de desempenho de amostras aleatórias de alunos de todas as escolas de interesse.

Uma limitação importante dos dados é a inexistência de informação de entrada. Há ampla evidência na literatura (Willms, 1992) que a variável com maior poder explicativo do desempenho de alunos é a sua habilidade medida antes do nível escolar que se quer avaliar. Neste caso deveria ser medida através de testes de desempenho anos antes do vestibular. Como não existe no Brasil uma avaliação da escola de ensino médio, estes dados não existem. Por outro lado, muitas das escolas de Belo Horizonte selecionam seus alunos através de resultados de testes, quer no setor público quer no privado. Desta forma, na entrada, os alunos já são diferenciados. Aquelas escolas conhecidas ou tidas como eficientes, determinam o perfil do aluno que desejam ter e através de procedimento de seleção elaborado por cada uma delas. Com esta estratégia atingem seu objetivo.

Mesmo com boa informação de entrada e modelos de análise corretos, a idéia de comparação de escolas tem limitações estruturais, como explicitado por Goldestein e Thomas (1996. Uma melhor utilização da Teoria do Valor Agregado é como instrumento de triagem isto é para identificação de escolas que seriam objeto de estudos de caso. Esta combinação de metodologia quantitativa e qualitativa forneceria informações preciosas para a melhoria do sistema. As boas escolas poderiam ser imitadas.

A instituição na qual o aluno termina o ensino médio não é, necessariamente, a instituição na qual ele cursou todo este nível. Ou seja, mesmo existindo a informação necessária para utilização da Teoria do Valor Agregado, existiria ainda o efeito da migração entre escolas, não disponível nos dados do questionário socioeconômico do vestibular.

É importante mais uma vez ressaltar que a grande utilidade dos modelos hierárquicos não pode ser apreciada neste trabalho, pela limitação dos dados. A mera identificação das escolas que, após ajuste por fatores socioeconômicos, apresentam melhor desempenho não deve ser objetivo final do trabalho. Deve-se procurar identificar as características das escolas que apresentam resultado melhor que o esperado, dada as características de seus alunos. Estes resultados poderiam subsidiar a proposição de políticas públicas adequadas.

Referências Bibliográficas

GOLDSTEIN, H. AND THOMAS, S. 1996: ***Using Examination Results as Indicators of School and College Performance.*** **J. R. Statistic. Soc. A, 159, 149-163.**

THOMAS, S. AND MORTIMORE, P. 1996: ***Comparison of value-added models for secondary-school effectiveness.*** **Research Papers in Education 11(1), 5-23.**

BRYK, A. S. AND RAUDENBUSH, S. W. 1992: ***Hierarchical Linear Models.*** **Newbury Park: Sage.**

GOLDSTEIN, H. AND SPIEGELHALTER, D. J. 1996: ***League Tables and Their Limitations: Statistical Issues in Comparissons of Institutional Performace.*** **J. R. Statistic. Soc. A, 159, Part 3, 385-443.**

SOARES, J. F. E FONSECA, J. A. 1998: ***Fatores Socioeconômicos e o Desempenho no Vestibular da UFMG-97*** **(Relatório), UFMG ,Belo Horizonte.**

WILLMS, J. DOUGLAS, 1992: ***Monitoring School Performance, Washington DC: The Falmer Press.***

APÊNDICE

Tabela A - Efeitos de cada escola segundo, os Modelos 1, 3 e 4.

Escolas	Postos			Efeito de escola		
	Mod.1	Mod.3	Mod.4	Mod.1	Mod.3	Mod.4
COLÉGIO STº ANTÔNIO	163	163	163	29,20	19,05	15,34
COLTEC DO CENTRO PEDAG. DA UFMG	148	162	162	13,50	15,17	11,28
CEFET - CENTRO FED. ED. TECNOLOGICA	144	161	161	10,90	13,41	10,10
COLÉGIO LOGOS GONZ PECOTCHE I	161	160	158	22,60	12,06	7,51
COLÉGIO LOYOLA	162	159	156	23,00	11,82	7,16
COLÉGIO STª DOROTEIA	160	158	159	20,30	11,40	7,85
COLÉGIO STº AGOSTINHO	158	157	160	18,70	10,71	8,74
COLÉGIO MARISTA DOM SILVERIO	159	156	153	19,00	9,40	5,66
COLÉGIO PITAGORAS TIMBIRAS	155	155	154	16,60	9,11	6,34
COLÉGIO STª MARIA I	152	154	157	15,00	8,28	7,22
COLÉGIO SÃO PAULO	156	153	147	16,70	7,74	4,64
COLÉGIO MAGNUM AGOSTINIANO	153	152	150	15,50	7,24	4,84
EM. GOV. CARLOS LACERDA	118	151	155	3,00	6,70	6,45
COLÉGIO PROMOVE SAVASSI	154	150	141	16,40	6,65	3,97
COLÉGIO PROMOVE MANGABEIRAS	157	149	135	17,20	6,30	2,54
COLÉGIO SÃO BENTO	151	148	144	14,90	5,99	4,38
COLÉGIO SAGRADO CORAÇÃO DE JESUS	150	147	143	13,80	5,90	4,30
COLÉGIO MUNICIPAL MARCONI	132	146	120	4,80	5,27	1,70
EE. AUGUSTO DE LIMA	124	145	137	3,40	4,79	2,98
COLÉGIO PROMOVE PAMPULHA	149	144	136	13,70	4,32	2,57
COLÉGIO PITÁGORAS PAMPULHA	147	143	112	12,50	4,20	1,05
COLÉGIO ARQUIDIOCESANO DE BH	141	142	149	8,60	4,20	4,66
EM. SANTOS DUMONT	122	141	142	3,30	4,18	4,05
INST. METODISTA IZABELA HENDRIX	146	140	134	11,10	4,07	2,53
COLÉGIO BATISTA MINEIRO	142	139	140	9,00	4,06	3,75
EM. PROF. LOURENÇO DE OLIVEIRA	120	138	130	3,20	4,03	2,36
ESC. NOSSA SENHORA DA PIEDADE	138	137	148	7,30	3,57	4,66
COLÉGIO MUNICIPAL DE B. HORIZONTE	104	136	125	0,00	2,83	1,99
COLÉGIO PADRE EUSTÁQUIO	134	135	138	5,50	2,71	3,58
COLÉGIO SAGRADO CORAÇÃO DE MARIA	145	134	102	11,10	2,65	0,69
COLÉGIO IMACULADA CONCEIÇÃO	143	133	94	10,80	2,47	0,35
COLÉGIO IMACO	81	132	124	-2,90	2,35	1,97
INST. PADRE MACHADO	137	131	116	7,20	2,15	1,41
EE. ENG. FRANCISCO BICALHO	94	130	132	-1,80	1,86	2,39
EE. OLEGÁRIO MACIEL	91	129	117	-2,10	1,78	1,48
EM. PROF. PEDRO GUERRA	48	128	139	-5,00	1,77	3,65
EE. DES. RODRIGUES CAMPOS	80	127	118	-3,10	1,64	1,50

Escolas	Postos			Efeito de escola		
	Mod.1	Mod.3	Mod.4	Mod.1	Mod.3	Mod.4
COLÉGIO STª MARCELINA	140	126	89	8,00	1,59	0,10
EE. GOV. MILTON CAMPOS	101	125	84	-0,40	1,51	-0,06
COLÉGIO TIRADENTES PMMG	103	124	69	0,00	1,38	-0,80
COLÉGIO TÉCNICO DE CONTAGEM	108	123	152	0,90	1,37	5,29
EE. PADRE JOÃO BOSCO PENIDO BURNIER	82	122	126	-2,90	1,27	2,02
EM. PROF. TABAJARA PEDROSO	76	121	131	-3,30	1,22	2,36
CENTRO EDUC. MINEIRO STª D'ÁVILA	139	120	71	7,90	1,21	-0,74
COLÉGIO SÃO MIGUEL ARCANJO	130	119	119	4,50	1,11	1,68
EM. PRES. TANCREDO NEVES	66	118	129	-3,80	1,05	2,23
EM. GERALDO TEIXEIRA DA COSTA	83	117	115	-2,70	1,05	1,41
INST. SAGRADA FAMÍLIA	123	116	127	3,30	0,92	2,10
EE. SÃO RAFAEL	45	115	109	-5,10	0,89	0,95
CESU - MARIA VIEIRA BARBOSA	96	114	151	-1,10	0,88	5,22
EE. CÂNDIDO PORTINARI	51	113	121	-4,90	0,60	1,75
EM. ARTHUR VERSIANI VELLOSO	106	112	39	0,70	0,56	-2,22
COLÉGIO SÃO FRANCISCO DE ASSIS	125	111	101	3,80	0,39	0,64
CURSO REG. SUPLÊNCIA PALOMAR	98	110	133	-0,80	0,31	2,46
EE. PROF. HILTON ROCHA	90	109	107	-2,10	0,31	0,79
COLÉGIO FREI ORLANDO CARLOS PRATES	126	108	113	3,90	0,30	1,17
EE. PROF. CAETANO AZEREDO	84	107	100	-2,70	0,28	0,47
COLÉGIO SÃO PAULO DA CRUZ	117	106	122	2,60	0,14	1,76
SUPLETIVO VISÃO	95	105	145	-1,60	0,13	4,56
EE. SAGRADA FAMÍLIA	97	104	62	-1,10	0,02	-0,94
EE. CARLOS DRUMOND DE ANDRADE	18	103	123	-7,00	-0,07	1,82
INST. DE EDUCAÇÃO DE MINAS GERAIS	78	102	146	-3,20	-0,13	4,57
EM. JOAQUIM DOS SANTOS	11	101	128	-7,60	-0,16	2,22
EM. HUGO PINHEIRO SOARES	22	100	105	-6,60	-0,17	0,73
COLÉGIO ARNALDO	131	99	78	4,70	-0,20	-0,46
EE. PROFª MARIA AMÉLIA GUIMARÃES	39	98	86	-5,50	-0,27	0,06
INST. ELIZABETH KALIL	70	97	32	-3,60	-0,35	-2,49
EM. SALGADO FILHO	85	96	66	-2,60	-0,43	-0,84
EE. ODILON BEHRENS	63	95	92	-4,00	-0,48	0,21
EE. PROF. RICARDO DE SOUZA CRUZ	14	94	108	-7,20	-0,49	0,80
EE. PROF. LEON RENAULT	36	93	40	-5,70	-0,56	-2,15
EE. KENNEDY	42	92	61	-5,10	-0,56	-1,02
CENTRO EDUC. MINEIRO	119	91	110	3,00	-0,57	0,98
COLÉGIO SOMA	135	90	25	6,30	-0,58	-3,26
FUNEC-UNIDADE CENTEC	50	89	82	-5,00	-0,58	-0,28
EE. PAULO DAS GRAÇAS DA SILVA	37	88	87	-5,70	-0,58	0,09
EE. DEP. ILACIR PEREIRA LIMA	75	87	76	-3,40	-0,72	-0,51

Escolas	Postos			Efeito de escola		
	Mod.1	Mod.3	Mod.4	Mod.1	Mod.3	Mod.4
EE. MARIA LUIZA MIRANDA BASTOS	52	86	93	-4,80	-0,73	0,29
EE. PROF. CLÁUDIO BRANDÃO	12	85	85	-7,60	-0,74	-0,05
EM. PAULO MENDES CAMPOS	87	84	52	-2,40	-0,78	-1,65
EM. CAIO LÍBANO SOARES	31	83	49	-6,00	-0,83	-1,78
EE. DR. LUCAS MONTEIRO MACHADO	72	82	104	-3,60	-0,87	0,72
EE. MAESTRO VILLA LOBOS	88	81	20	-2,30	-0,91	-3,75
EE. JUSCELINO K. DE OLIVEIRA	20	80	99	-6,70	-0,91	0,47
EE. PROF. LEOPOLDO DE MIRANDA	86	79	21	-2,50	-0,92	-3,56
EM. OSWALDO CRUZ	56	78	81	-4,60	-0,94	-0,30
COLÉGIO PAMPULHA	114	77	60	2,00	-0,96	-1,07
EE.LUIZ DE BESSA	15	76	98	-7,20	-1,01	0,46
EE. ASSIS CHATEAUBRIAND	23	75	68	-6,50	-1,05	-0,84
EE. TÉC. IND. PROF. FONTES	43	74	30	-5,10	-1,07	-2,64
EE. PROF. ALISSON PEREIRA GUIMARÃES	34	73	38	-5,70	-1,08	-2,24
EE. GETÚLIO VARGAS	10	72	79	-7,70	-1,18	-0,36
EM. SEC. HUMBERTO ALMEIDA	30	71	77	-6,00	-1,19	-0,50
EE. PEDRO II	65	70	64	-3,90	-1,27	-0,90
INST. ITAPOÃ	112	69	90	-1,40	-1,39	0,17
EE. PASCHOAL COMANDUCCI	16	68	103	-7,10	-1,40	0,69
COLÉGIO PIO XII	127	67	29	3,90	-1,50	-2,79
COLÉGIO N. Srª DO MONTE CALVÁRIO	116	66	46	2,50	-1,51	-1,97
COLÉGIO SUPLETIVO PEDRO II	79	65	111	-3,10	-1,52	1,00
EE.CEL. MANOEL SOARES DO COUTO	7	64	97	-8,20	-1,53	0,43
EM. ANTÔNIO SALES BARBOSA	61	63	80	-4,30	-1,55	-0,35
EE. NOSSA Srª DO CARMO	62	62	47	-4,20	-1,63	-1,95
COLÉGIO MÓDULO	133	61	13	5,40	-1,63	-4,98
EE. PROF. PEDRO ALEIXO	47	60	59	-5,10	-1,65	-1,15
EM. MESTRE ATAÍDE	73	59	63	-3,40	-1,66	-0,91
COLÉGIO N. Srª DAS DORES	113	58	73	1,70	-1,70	-0,65
EE. CELSO MACHADO	46	57	56	-5,10	-1,83	-1,38
EE. PROF. FRANCISCO BRANT	3	56	54	-8,70	-1,83	-1,43
COLÉGIO ABGAR RENAULTI	71	55	96	-3,60	-1,83	0,37
CESPRO - CENTRO ENSINO SUP. PROMOVE	89	54	34	-2,20	-1,88	-2,40
COLÉGIO SISTEMA	129	53	24	4,20	-1,90	-3,38
EE. AMÉLIA SANTANA BARBOSA	40	52	45	-5,40	-2,07	-2,01
EE. MADRE CARMELITA	58	51	37	-4,50	-2,07	-2,32
COLÉGIO ADVENTISTA DE B. HORIZONTE	111	50	75	1,10	-2,08	-0,59
COLÉGIO TITO NOVAIS	57	49	67	-4,50	-2,08	-0,84
EM. MILTON CAMPOS	8	48	72	-8,20	-2,12	-0,73
EE.NA DE CARVALHO SILVEIRA	9	47	65	-8,10	-2,17	-0,86

Escolas	Postos			Efeito de escola		
	Mod.1	Mod.3	Mod.4	Mod.1	Mod.3	Mod.4
COLÉGIO MÉTODO	121	46	43	3,20	-2,20	-2,04
COLÉGIO DOM CABRAL	109	45	55	1,00	-2,23	-1,40
EE. ORDEM E PROGRESSO	35	44	28	-5,70	-2,27	-2,87
ESC. ARNALDINUM SÃO JOSÉ	128	43	22	4,20	-2,36	-3,56
COLÉGIO ÔMEGA	59	42	41	-4,40	-2,38	-2,14
EE. CARMO GIFFONI	6	41	58	-8,20	-2,58	-1,26
COLÉGIO RUI BARBOSA	38	40	114	-5,60	-2,63	1,23
COLÉGIO DA AEC	33	39	106	-5,80	-2,66	0,79
COLÉGIO MINAS GERAIS	28	38	91	-6,10	-2,67	0,21
SISTEMA ENS BH - COLÉGIO PH7	55	37	95	-4,60	-2,73	0,37
EE. PROF. MORAIS	24	36	26	-6,50	-2,74	-2,96
EE. SANTOS DUMONT	19	35	51	-6,80	-2,75	-1,69
COLÉGIO SÃO JOSÉ	107	34	36	0,70	-2,75	-2,32
INST. ZILAH FROTA	136	33	6	6,70	-2,83	-6,59
CRESA - CURSO REG. SUPLÊNCIA APROVA	41	32	88	-5,40	-2,89	0,09
COLÉGIO MONTE LÍBANO	93	31	42	-1,90	-2,92	-2,13
EE. TRÊS PODERES	25	30	15	-6,40	-2,96	-4,31
COLÉGIO VISCONDE DE CAIRU	53	29	74	-4,70	-3,05	-0,62
EE. DONATO WERNECK DE FREITAS	2	28	70	-11,00	-3,06	-0,76
ESC. SALESIANA	105	27	33	0,10	-3,07	-2,42
EM. HILDA RABELLO MATTA	26	26	48	-6,30	-3,16	-1,94
CENTRO EDUC. TEC. DE ARTES PROFIS.	49	25	83	-5,00	-3,32	-0,13
EM. DOM ORIONE	27	24	14	-6,20	-3,53	-4,34
COLÉGIO MODELO	115	23	9	2,20	-3,54	-5,73
INST. MARIA MONTESSORI	99	22	50	-0,80	-3,60	-1,77
COLÉGIO MAXIMUS	102	21	35	-0,10	-3,85	-2,38
COLÉGIO CARRIER	32	20	57	-5,90	-4,16	-1,30
ESC. TÉC. POLIMIG PLATINA	77	19	31	-3,30	-4,21	-2,56
ESC. DA COM. DOMICIANO VIEIRA	74	18	53	-3,40	-4,27	-1,53
CURSO SUPLETIVO ROMA	44	17	17	-5,10	-4,34	-4,21
COLÉGIO ANCHIETA	110	16	11	1,00	-4,47	-5,22
COLÉGIO TÉC. DE ELET. DE MG-COTEMIG	68	15	18	-3,70	-4,62	-4,04
COLÉGIO PEDRO II	60	14	44	-4,40	-4,78	-2,04
COLÉGIO ATENAS	92	13	8	-1,90	-4,80	-6,08
COLÉGIO PRES. WENCESLAU BRÁS	13	12	23	-7,30	-5,05	-3,55
EE. PEDRO AMÉRICO	1	11	10	-11,20	-5,09	-5,55
COLÉGIO SÃO LUÍS GONZAGA	29	10	19	-6,00	-5,29	-3,81
EM. LUIZ GATTI	5	9	16	-8,30	-5,35	-4,22
COLÉGIO BRASILEIRO	21	8	27	-6,60	-5,60	-2,96
COLÉGIO CLEMENTE FARIA	54	7	12	-4,70	-5,67	-5,04

Escolas	Postos			Efeito de escola		
	Mod.1	Mod.3	Mod.4	Mod.1	Mod.3	Mod.4
SUPLETIVO MODELO	67	6	3	-3,70	-6,10	-8,74
COLÉGIO ROMA	100	5	4	-0,70	-6,24	-8,05
COLÉGIO JOSÉ DE ALENCAR ROGEDO	64	4	5	-4,00	-6,69	-7,97
COLÉGIO PRISMA	17	3	7	-7,10	-7,00	-6,42
COLÉGIO MILTON CAMPOS	69	2	2	-3,70	-7,31	-10,22
COLÉGIO PADRE LEBRET	4	1	1	-8,60	-10,67	-10,39

ABSTRACT

In the last years, several Brazilian universities published the mean score in the their entrance examination for the students that attend the same high school. The use of this methodology for comparing high schools presents two big limitations. It doesn't consider the socioeconomic differences among the students of different schools and ignores the correlation between the observations. This paper, based in the results from the first stage of the entrance examination of the Federal University of Minas Gerais in 1997, deals with this shortcomings by incorporating the hierarchical structure of the data in a regression model. The hierarchical models provides correct estimates for the model's parameters, a measure of the correlation coefficient between the scores of students of the same school and, more importantly, a measure of the effect for each high school. The model fitted to the data show that the variation in school's scores can be explained more by the socioeconomic differences among the students of each school than by the school's policies and practices. The null model estimate that 35,9% of the student scores variability is due to the schools, percentage that falls to 19,6 when the results are controlled by the student's covariables.

Política Editorial

A Revista Brasileira de Estatística - RBEs - objetiva promover a estatística relevante para aplicação em questões sociais, interpretadas amplamente para incluir questões educacionais, de saúde, demográficas, econômicas, legais, de políticas públicas e de estatísticas oficiais, entre outras. A revista apresenta artigos num formato que permite fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações devem ter conteúdo estatístico substancial. As análises deverão ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

Artigos contendo exposição de métodos são aceitáveis, desde que estes sejam relevantes para as áreas cobertas pela revista, auxiliem na compreensão do problema e contenham interpretação clara das expressões matemáticas apresentadas. A apresentação de aplicações ilustrativas envolvendo dados adequados é requerida. Tratamentos algébricos extensos devem ser evitados.

A revista tem periodicidade semestral e publica também artigos escritos à convite e resenha de livros, bem como artigos abordando os diversos aspectos de metodologias relevantes para órgãos produtores de estatísticas, incluindo:

a) planejamento de pesquisas;

b) avaliação e mensuração de erros em pesquisas;

c) uso e combinação de fontes alternativas de informações e integração de dados;

d) novos desenvolvimentos em metodologia de pesquisa;

e) crítica e imputação de dados;

f) amostragem e estimação;

g) disseminação e confiabilidade de dados;

h) análise de dados;

i) análise de séries temporais;

j) modelos e métodos demográficos; e

k) modelos e métodos econométricos.

Todos os artigos submetidos são avaliados quanto à qualidade e relevância por dois especialistas indicados pelo Comitê Editorial da Revista Brasileira de Estatística. Os artigos submetidos devem ser inéditos e não ter sido, simultaneamente, submetidos a qualquer outro periódico nacional. O processo de avaliação é do tipo duplo cego, isto é, os artigos são avaliados sem identificação da autoria, e os comentários dos avaliadores também são repassados aos autores sem identificação.

Instruções para submissão de artigos

Os artigos submetidos para publicação devem ser remetidos em três vias, que não serão devolvidas, para:

Pedro Luis do Nascimento Silva
Editor responsável
Revista Brasileira de Estatística
Av. República do Chile, 500 10º andar
Rio de Janeiro - RJ - 20031-170
Tel.: xx 55-21-515 0470
Fax: xx 55-21-514 4785
E-mail: pedrosilva@ibge.gov.br

Para cada artigo publicado, são fornecidas gratuitamente 20 separatas.

Instruções para Preparo de Originais

Os originais entregues para publicação devem obedecer às seguintes normas:

1 - O texto deve ser editado, preferencialmente, em Word, sem formatação (*default*), configurado em A4.

2 - As páginas do original devem ser numeradas seqüencialmente.

3 - A primeira página do original (folha de rosto) deve conter o título do artigo, nome completo do(s) autor(es), com indicação das instituições a que está(ão) vinculado(s) e endereço para correspondência. Agradecimentos a colaboradores e instituições e auxílios recebidos devem figurar também nessa página.

4 - A segunda página do original deve conter um resumo informativo de no máximo 150 palavras, em português e inglês (Abstract), destacando os pontos relevantes do artigo. Deve seguir o mesmo padrão do texto, em um único parágrafo, sem inclusão de fórmulas. Ver a respeito a norma da Associação Brasileira de Normas Técnicas - ABNT -, *Resumos*: NBR 6028, de julho de 1988.

5 - As notas explicativas devem ser numeradas numa seqüência única, listadas no pé-de-página onde elas se encontram.

6 - As tabelas e gráficos devem ser precedidos de títulos que permitam a identificação do conteúdo. Devem ser numerados seqüencialmente (Tabela 1, Figura 3, etc.) e com ordem de indicação de entrada no texto.

Toda tabela e gráfico deve ter fonte. Recomenda-se a indicação dos documentos publicados que foram utilizados na sua elaboração, identificados por referências bibliográficas completas, com as páginas ou volumes específicos de onde foram extraídas as informações. No caso de publicação que contenha tabelas com dados numéricos resultantes de uma única fonte, já identificada na própria publicação, é dispensável a apresentação da fonte em cada uma das tabelas.

No caso de tabelas e demonstrações extensas ou outros elementos de suporte, podem ser incluídos anexos, que devem ter título e numeração.

7 - As citações bibliográficas no texto devem ser feitas de acordo com a norma da ABNT, *Apresentação de citações em documentos*: NB 896, de maio de 1990.

8 - As referências bibliográficas devem ser redigidas segundo a norma da International Organization for Standardization - ISO -, *Referência bibliográfica e documentação*: n. 690, de 1987, contendo os elementos necessários à identificação da publicação. Devem ser organizadas em ordem alfabética.

Se o assunto é Brasil,
procure o IBGE

http://www.ibge.gov.br
http://www.ibge.net

atendimento
0800 21 81 81