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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 70 número 233 julho/dezembro 2009

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 70, n. 233, p. 1-111, jul./dez. 2009

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

Francisco Louzada-Neto
Editor responsável - RBEs - IBGE.
Rua André Cavalcanti, 106 - Santa Teresa
20231-050 - Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Sandra Cavalcanti de Barros - ENCE/IBGE
Raquel Rodrigues Levy Barbosa - ENCE/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2009.

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

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

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

Trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).
Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN 0034-7175 = Revista brasileira de estatística.

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

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (rev.2009) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor

Este segundo volume da RBEs do ano de 2009 é composto por quatro artigos. O primeiro artigo, de autoria de Karina Pretto e Rinaldo Artes, discute algumas propostas de mensuração da influência de atributos ausentes na avaliação de respondentes quando estímulos incompletos são apresentados aos mesmos no contexto de análise de preferência conjunta. O segundo artigo, de autoria de Rita de Cássia de Lima Idalino, Pâmela Sabrina de Oliveira e Paulo Sérgio Lucio, apresenta modelos que descrevem os impactos de eventos raros na sociedade e ecossistemas para prevenção e propostas de mitigação via teoria de valores extremos. O terceiro artigo, de autoria de Fernando A. Moala, apresenta uma implementação prática do método bayesiano de elicitação de distribuição de *priori* proposto por Oakley e O'Hagan (2007) para construção de uma distribuição a *priori* do risco de fratura em pacientes que sofrem de osteoporose sob um determinado tratamento. O quarto artigo, de autoria de Rafael Perez Ribas e Sergei Suarez Dillon Soares, apresenta um algoritmo de emparelhamento avançado capaz de reduzir a perda de pessoas no painel da Pesquisa Mensal de Emprego do IBGE.

Aproveito a oportunidade para agradecer a Associação Brasileira de Estatística-ABE pela confiança, a colaboração de Ismenia Blavastsky (Editora-Executiva) e a todos os Editores Associados, revisores do periódico, autores e a equipe do IBGE.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Análise de Preferência Conjunta: um estudo sobre a omissão de atributos

Karina Pretto [1]
Rinaldo Artes [2]

Resumo

A Análise de Preferência Conjunta - APC, ou *Conjoint Analysis*, é uma metodologia estatística, amplamente utilizada em pesquisas de *marketing*, que se preocupa em identificar características (atributos) de um produto que mais se associam com a preferência do consumidor. Quando se tem uma quantidade excessiva de atributos, o número de possíveis estímulos pode ser muito elevado. Uma alternativa, nestes casos, é a apresentação de estímulos incompletos. Embora o ganho de tempo e a redução no número de configurações apresentadas possam ser significativos, cabe avaliar se a apresentação de um estímulo incompleto interfere no processo de julgamento, ou seja, deve-se investigar até que ponto um atributo, mesmo ausente, interfere na avaliação do respondente. Neste trabalho, são discutidas algumas propostas de mensuração da influência de atributos ausentes na avaliação quando estímulos incompletos são apresentados aos respondentes. É apresentada uma aplicação a dados reais da área de recursos humanos.

[1] IBMEC (.), São Paulo - SP, Brasil.
[2] USP (Instituto de Matemática e Estatística), São Paulo - SP, Brasil.

1. Introdução

A Era Industrial, iniciada no Século XVII, foi marcada pelo aumento na eficiência dos métodos de produção, o que além de baratear os preços estimulou o consumo. Nesse período, a padronização e produção em massa eram as chaves do sucesso. Já em meados do Século XX, o aumento da concorrência e a rápida disseminação de informações através da Internet fizeram com que muitas empresas reestruturassem seus processos de produção em busca da satisfação e fidelização de seu cliente. Atualmente, a produção em massa cedeu lugar à produção personalizada, que utiliza informações sobre hábitos de consumo para oferecer produtos e serviços que atendam tanto às necessidades quanto ao desejo do consumidor. Basta observar a grande expansão da TV a cabo, que trouxe a opção de informação e entretenimento que se adapta a cada estilo de vida de seus assinantes. Ou ainda, a diversidade de produtos eletroeletrônicos disponíveis no comércio, que deixam a critério do consumidor a difícil missão da escolha do produto de sua preferência.

Ao oferecer um produto personalizado, ou seja, que tenha a “cara do consumidor”, as empresas procuram aproximar-se mais do seu cliente, tornando-se mais atrativas e competitivas frente à concorrência. Surge assim o jogo de sedução entre vendedores e compradores que se inicia muito antes do lançamento de um novo produto no mercado, pois para identificar qual a “cara” que o produto precisa ter para diferenciar-se dos demais é preciso conhecer o cliente e entender seus hábitos de consumos, financeiro e cotidiano.

A utilização de pesquisas científicas sobre o comportamento do mercado consumidor talvez seja uma das práticas mais frequentes e antigas utilizadas para mensurar a preferência. Desde o início do Século XX, a aplicação de questionários e utilização de medidas estatísticas básicas auxiliam no entendimento do perfil do consumidor e do mercado.

Com o passar dos anos e com o desenvolvimento de novas metodologias estatísticas de análise de dados, Modelos de Regressão Múltipla, Análise de Correlação, Modelos de Planejamento de Experimentos, além das Técnicas Multivariadas, foram incorporados a estes estudos enriquecendo ainda mais este processo. Segundo Friedmann (19998), a tendência atual da pesquisa em *marketing* aponta para estudos sobre a análise de preferência do consumidor e o processo de decisão presente no momento da compra de um bem ou contratação de um serviço.

Nestes casos, a intenção é identificar fatores que apresentam uma maior influência na compra, ou seja, fatores capazes de, por exemplo, diferenciar um produto dos seus concorrentes a fim de torná-lo mais competitivo. Outros estudos verificam se há a presença de alguma regra de decisão implícita utilizada pelo consumidor no momento da compra, e de que forma o conhecimento *a priori* sobre o produto pode influenciar na sua aquisição. Seus resultados podem fornecer informações sobre a aceitabilidade do produto, suas vantagens e desvantagens ante os concorrentes e ainda dar suporte a estudos de simulação de penetração do produto antes mesmo da sua fabricação (*Green* e *Savitz*, 1994).

Uma técnica estatística multivariada muito utilizada neste tipo de estudo é a Análise de Preferência Conjunta - APC, também conhecida por *Conjoint Analysis*. A APC se preocupa em estimar modelos que expressem as preferências individuais de um consumidor como função de um conjunto de características prédefinidas. É possível ainda mensurar o impacto de cada uma dessas características sobre a preferência do consumidor, dado que um dos objetivos em uma análise de preferência é estimar a preferência em relação a diferentes configurações de um mesmo produto. Este produto pode ser qualquer objeto tal como um aparelho eletroeletrônico, automóvel, alimentos, produtos de beleza, ou, então, pode representar um serviço, como, por exemplo, a composição de um plano de saúde, forma de atendimento em uma agência bancária ou estrutura de um curso de pós-graduação.

As características que descrevem o produto em estudo são conhecidas como atributos. Assim, se o produto alvo for um automóvel, a potência do motor, a presença de direção hidráulica e o número de portas são possíveis atributos presentes na caracterização do mesmo. As variações de um atributo, por exemplo, o número de portas (duas ou quatro portas) são denominadas níveis de um atributo. Quando diferentes níveis de diferentes atributos são combinados, temos uma possível configuração do produto denominada estímulo ou perfil.

Para cada estímulo, o respondente deve atribuir um valor que quantifica sua preferência pela configuração apresentada. Este valor é conhecido como valor de preferência (utilidade) e é resultado da combinação das preferências individuais ou utilidades de cada atributo. Logo, o valor de preferência por um estímulo pode ser modelado como combinação das utilidades parciais de todos os atributos.

A quantidade excessiva de características ou atributos é um dos pontos mais críticos neste tipo de estudo, pois quanto mais detalhes de um produto forem avaliados simultaneamente, um maior número de estímulos pode ser gerado, o que aumenta significativamente o tempo de resposta do questionário e a fadiga dos respondentes. Para contornar esse problema, *Green et alli*.(1978) propõem o uso de delineamentos experimentais fatoriais fracionários (*Montgomery*, 1996). Há situações, no entanto, em que mesmo esse artifício não consegue reduzir satisfatoriamente o número de produtos hipotéticos, sendo necessária a utilização de planejamentos de experimentos fracionários combinados com a omissão de um ou mais atributos na descrição do estímulo.

Também é comum assumir que atributos ausentes em um estímulo não interferem na quantificação da sua preferência, contudo vários estudos questionam tal posicionamento. Quando um atributo muito importante é omitido na descrição de um estímulo, a percepção de um indivíduo sobre o produto pode ser afetada, e consequentemente seu julgamento sobre a preferência também pode sofrer alterações. Por exemplo, Broniarczyk e Alba (1994) discutem experimentos em que os respondentes atribuíram valores de preferências distintos ao classificar um mesmo estímulo apresentado de forma completa ou incompleta, o que leva os autores a crerem na influência dos demais fatores na quantificação da preferência de um atributo ausente. Feldman e Lycnh (1988) acreditam que os respondentes usem de crenças, atitudes e pré-conceitos ao avaliar um estímulo. Assim, ao se deparar com a omissão de um atributo, o respondente pode associar o nível ausente aos níveis dos atributos presentes.

Resta avaliar até que ponto o valor de preferência atribuído a um estímulo é afetado pela ausência de um atributo e quais são os artifícios utilizados pelos respondentes para incorporar tal informação no processo de julgamento.

O procedimento adotado para esta verificação consiste em definir uma regra de imputação da característica omitida na descrição do perfil, aplicá-la na matriz de delineamento e posteriormente estimar o valor de preferência para cada um dos perfis avaliados por um mesmo indivíduo, usando inferência bayesiana e simulação de Monte Carlo.

Dentre as diferentes propostas passíveis de aplicação estão o método do valor padrão (ou valor de referência da característica avaliada), imputação do valor mais recente visualizado pelo individuo e o método de aprendizado, proposto por Bradlow, Hu e Ho (2004).

Embora seja pouco usual, a imputação de valores combinada com a estimação bayesiana em aplicações de APC se justifica devido ao fato de que a independência entre as avaliações de diferentes perfis feitas por um mesmo indivíduo é bastante questionável.

Através da execução de um experimento real, pretende-se comparar alternativas de análise que incorporam o efeito da ausência de atributos na classificação de estímulos em uma aplicação sobre o efeito de características do emprego na satisfação e motivação do funcionário. Embora o produto em estudo aqui não seja um bem de consumo, a função exercida pode muito bem ser avaliada como sendo um produto, cujo fabricante é a própria organização para a qual o empregado trabalha.

2. Antecedentes

Os primeiros trabalhos sobre APC e efeitos conjuntos de atributos foram publicados a partir da década de 1960. Luce e Tukey (1964) apresentaram um estudo sobre o efeito conjunto de dois ou mais atributos na ordenação de estímulos. Em 1971, Green e Rao publicaram o primeiro artigo sobre medida conjunta aplicado a *marketing*, enfatizando a estimação de um modelo de preferência. Primeiro em 1978 e depois em 1990, Green e Srinivasan descreveram as etapas fundamentais para a realização de um estudo de APC enfatizando aspectos teórico e prático importantes a serem considerados neste tipo de pesquisa. Louviere, Eagle e Cohen (2005) e Rao (2007) trazem bons resumos sobre técnicas de APC.

São apresentados, na década de 1980, estudos com modelos com estruturas híbridas (*Green*, 1984), que buscam minimizar o número de estímulos respondidos pelos participantes quando um grande número de atributos são avaliados simultaneamente.

Há duas maneiras de se construir estímulos: na primeira, todos os atributos são incluídos na descrição, trata-se do método perfil completo. Uma maneira alternativa é construir estímulos que utilizam apenas subconjuntos dos atributos; esse é o método perfil incompleto. Quando o número de atributos é muito grande, além do uso de planejamentos fracionários, pode ser vantajosa a utilização de perfis incompletos, para tornar o estímulo mais simples. No entanto, é necessário saber se há algum efeito da omissão de um atributo

no estímulo sobre a sua avaliação. Embora Green (1984) assuma que a preferência por um estímulo não sofra alteração diante da falta de alguns atributos. Alguns autores, tais como Lynch e Srull (1982), Yamagishi e Hill (1981) e Bradlow, Hu e Ho (2004), entre outros, acreditam que os respondentes ponderam suas decisões levando em consideração a falta de informação na descrição dos perfis, ou seja, que eles assumam algum valor para atributos que estejam faltando. Por exemplo, ao avaliar um automóvel, quando se fornece a potência do motor, é possível que o respondente imagine um preço razoável para o automóvel e utilize essa informação na avaliação. Tal ponderação pode ser interpretada como um fenômeno de imputação feita mentalmente pelos indivíduos. Logo, descrever matematicamente um modelo que capte este processo de imputação mostra-se uma tarefa bastante árdua e rica em detalhes, dado que cada sujeito pode utilizar um ou mais mecanismos diferentes de ponderação ao longo de sua participação da pesquisa.

Lynch e Srull (1982) acreditam que os respondentes tendem a utilizar o último nível visto de um atributo quando este estiver ausente, ou seja, assumem que quando nada for mencionado sobre um atributo, o seu nível não tenha sofrido alteração em relação ao perfil anterior. Já Yamagishi e Hill (1981) supõem que todos os níveis já apresentados ao respondente podem influenciar igualmente em sua decisão, assumindo então que a imputação é feita pela “média” dos níveis do atributo ausente.

Segundo Huber e McCann (1982) é possível ainda que o respondente associe dois ou mais atributos por meio de uma estrutura arbitrária de correlação, por exemplo, associar preço à qualidade ou horas de trabalho e remuneração. Logo, quando qualquer um destes atributos estiver ausente, a informação seria imputada com base na informação do atributo presente.

Combinando conceitos de associação de perfis (*Camerer* e *Ho*, 1998) e a forma de aprendizado e memória adquiridos ao longo do processo de classificação (*Ho* e *Chong*, 2003). Bradlow, Hu e Ho (2004) propõem um novo modelo de imputação, denominado de modelo de aprendizado. Os autores acreditam que haja um processo de aprendizagem do respondente à medida que mais informação sobre o produto lhe é apresentada e propõem modelos que visam descrever e incorporar esse processo na análise dos dados. Convém ressaltar que os modelos descritos neste artigo e a notação utilizada são similares aos de Bradlow, Hu e Ho (2004).

3. Caracterização de um modelo de APC

Como exemplo ilustrativo, considere uma APC cujo delineamento seja descrito pela Tabela 1. Supondo que o objeto em estudo seja a função desempenhada em uma organização e que os três atributos sejam a Jornada de Trabalho (xi1), Remuneração (xi2) e Frequência de exposição a novos desafios (xi3). Embora o produto seja caracterizado por três atributos com dois níveis cada, o delineamento apresenta a formação de perfis incompletos, pois em cada estímulo omite-se a informação de um dos três atributos.

Tabela 1 - Delineamento experimental de uma APC com três atributos.

Estímulo (t)	Valor de preferência (yij(t))	Atributos		
		Jornada de Trabalho	Remuneração	Desafios
1	$y_i(1)$	10h/dia		constantes
2	$y_i(2)$		sem aumento	raros
3	$y_i(3)$		sem aumento	constantes
4	$y_i(4)$	8h/dia		constantes
5	$y_i(5)$	10h/dia		raros
6	$y_i(6)$	8h/dia	sem aumento	
7	$y_i(7)$	10h/dia	sem aumento	
8	$y_i(8)$		aumento de 20%	raros

O primeiro estímulo caracteriza uma função que exige trabalho de 10 horas diárias em que o funcionário é desafiado constantemente. Nada é dito sobre a remuneração. A suposição de que o atributo ausente, no caso a remuneração, não interfere na avaliação é bastante questionável, pois é possível que um respondente, ao avaliar o primeiro estímulo, pondere sua decisão sobre o valor de preferência segundo suas expectativas salariais, caso este atributo seja importante para ele. Ou seja, a influência da omissão de um atributo na descrição de um estímulo pode levar a algum viés na estimativa o grau de importância dos demais atributos.

Quando um atributo muito importante for omitido na descrição de um estímulo, espera-se que os respondentes compensem sua ausência acrescentando ao valor de preferência uma parcela subjetiva (imputação) referente ao nível do atributo ausente. Este processo de imputação presente durante a avaliação de um perfil incompleto é feito mentalmente pelo respondente, sendo difícil de ser captado via modelagem, pois a cada nova avaliação uma

diferente regra de decisão pode ser adotada. A complexidade pode tornar-se ainda maior se a heterogeneidade dos participantes da pesquisa for levada em consideração.

Um dos atrativos da APC é a possibilidade de construir um modelo de preferência para cada respondente, assim é possível estimar a importância (utilidade) atribuída por um indivíduo para cada nível do atributo. Como premissa, um modelo de APC assume que a avaliação dos consumidores combina diferentes parcelas de utilidades, ou seja, o valor de preferência é formado pela combinação linear dos atributos considerados na descrição de um estímulo.

A contribuição de cada atributo na composição da preferência é também conhecida por utilidade parcial. Assumindo que o valor de preferência seja formado pela combinação linear das utilidades parciais (valores de preferências) dos atributos considerados na descrição de um estímulo, e que estejamos trabalhando com atributos quantitativos ou que possuam apenas dois níveis (níveis 0 e 1), é possível descrever um modelo de preferência que sob forma de uma equação linear, dada pela expressão (1)

$$y_i(t) = \alpha_i + \sum_{j=1}^{J} \left[\beta_{ij} X_{ij}(t)\right] + \varepsilon_i(t)\,, \qquad (1)$$

sendo $y_i(t)$ o valor de preferência atribuído pelo i-ésimo indivíduo ao perfil t, β_{ij} o parâmetro que representa a utilidade parcial referente ao atributo *j* e $X_{ij}(t)$ é o valor numérico que representa o nível do atributo *j* avaliado no instante *t*. Aqui, α_i e $\varepsilon_i(t)$ representam, respectivamente, um parâmetro de escala e o erro aleatório que representa a influência de todos os demais atributos que não foram considerados no estudo e que possam interferir na preferência do respondente. Usualmente, assume-se que o erro $\varepsilon_i(t)$ seja normalmente distribuído ($\varepsilon_i(t) \sim N(0,\sigma_i^2)$).

Neste caso, *J* atributos descrevem um estímulo, identificado pelo índice *t*. O índice *i* identifica a qual indivíduo este modelo faz referência, isso porque a percepção sobre um determinado produto pode variar de indivíduo a indivíduo. Assim, $y_i(t)$ pode assumir tanto valores métricos, por exemplo, uma nota variando de 0 a 10, quanto valores não métricos, quando a preferência for quantificada sob a forma de uma escala de importância.

Os níveis dos atributos avaliados em cada estímulo são representados por $X_{ij}(t)$. Por simplificação, assume-se neste trabalho que cada atributo pode ser bem descrito por

apenas dois níveis, logo $X_{ij}(t)$ pode assumir dois valores 0 e 1. O primeiro representa o nível padrão do atributo avaliado e o segundo diz respeito ao nível alternativo em avaliação. A utilidade parcial do nível do atributo *j* para o indivíduo i é dada por β_{ij} .

4. Formas de imputação

À medida que mais estudos sobre a utilização de estímulos ou perfis incompletos como alternativa para diminuir o número de perfis avaliados foram sendo realizados, o número de pesquisadores que passaram a adotar este tipo de método de formação de perfis também aumentou. Isso porque diferentes estudos sugerem que as estimativas das utilidades parciais mostram-se muito similares àquelas encontradas quando perfis completos são utilizados, como visto por Bradlow, Hu e Ho (2004) em um estudo de preferências aplicado a máquinas fotográficas digitais. No entanto, para que isso aconteça, é preciso assumir que:

a) Não há grandes perdas de informação ao se utilizar perfis incompletos em vez de perfis completos, ou seja, a omissão dos atributos não afeta o valor de preferência nem as utilidades parciais:

b) Que o atributo ausente não influencia na preferência atribuída ao perfil: e

c) Há independência entre as respostas das preferências atribuídas para dois perfis diferentes, sendo descritos ou não pelos mesmos atributos.

Tais suposições são questionáveis, uma vez que é natural que um indivíduo, ao classificar um perfil, considere a existência do atributo faltante e acabe por “imputar” algum nível coerente com os demais atributos presentes no perfil. Pode-se pensar que a preferência do respondente é influenciada pela ausência de um atributo e que o maior desafio seja conhecer quais são os processos que são considerados para imputar esta informação na avaliação do perfil.

Assim, surge o interesse em estudar se a ausência de um atributo em um perfil realmente afeta o valor da utilidade a ele atribuída, se a ordem de apresentação dos estímulos interfere nos valores das suas utilidades e de que maneira o respondente imputa

a informação do atributo ausente. Cabe lembrar que o processo de imputação é realizado mentalmente pelo respondente, o que torna o desafio de expressar matematicamente um modelo que capte este processo, tão particular, ainda maior.

Uma alternativa para mensurar a influência da presença/ausência de um determinado atributo é acrescentar mais um termo ao modelo (1) que indique a presença de cada atributo no estímulo sob avaliação. Fazendo isso, o modelo passa a ser escrito como:

$$y_i(t) = \alpha_i + \sum_{j=1}^{J} \left[\beta_{ij} X'_{ij}(t) + \beta'_{ij}\, r_{ij}(t)\right] + \varepsilon_i(t) \qquad (2)$$

sendo $y_i(t)$ o valor de preferência atribuído pelo i-ésimo indivíduo ao perfil *t*, α_i o parâmetro de escala, β_{ij} o parâmetro que representa a utilidade parcial referente ao atributo *j* e $X_{ij}(t)$ é o valor numérico que representa o nível do atributo *j* avaliado no instante *t*.

A variável indicadora da presença do atributo *j* no perfil *t* respondido pelo indivíduo *i* é denotada por *rij(t)* e o parâmetro β'_{ij} é o parâmetro que mensura o incremento no valor de preferência quando o atributo j está descrito no estímulo *t* do indivíduo *i*. O erro aleatório do modelo é representado por $\varepsilon_i(t)$.

A seguir, são descritas várias maneiras de modelar o processo de imputação de um atributo ausente.

4.1 Imputação utilizando um valor padrão

Ao se deparar com a uma informação ausente na descrição de um estímulo, um respondente pode ainda adotar um padrão de resposta tal como admitir que o nível do atributo ausente seja igual a um nível padrão. Por exemplo, o segundo estímulo da Tabela 1 descreve uma função cuja remuneração não se altera em relação à atual situação, com raros desafios apresentados ao funcionário. Nada é mencionado sobre a jornada de trabalho. Dado que o padrão assumido na maioria das empresas é de 8 horas/dia,

espera-se que um respondente assuma que este deve ser o valor padrão adotado ao julgar este estímulo.

4.2 Imputação pela média

É possível ainda que um respondente considere que o nível ausente seja dado pela média dos possíveis níveis (no caso de atributos quantitativos). Por exemplo, considere que o atributo salário anual oferecido por uma empresa é descrito quantitativamente e que por conhecimento prévio do respondente, ou os valores descritos nos estímulos anteriores descreverem um salário entre R$ 20.000,00 e R$ 24.000,00, nada impede do candidato assumir que, quando nada for informado sobre o salário uma média de R$ 22.000,00 pode estar sendo considerada.

4.3 Imputação pela recência

Uma forte similaridade entre o estímulo atual e o último avaliado, associada ao fato de que ambos sejam distintos dos demais, pode levar a um padrão de imputação baseado na recência. Neste caso, o respondente assume que o nível atual seja idêntico ao apresentado no estímulo que o antecede, ou seja, é um processo de memória curta. Ou, ainda, pode-se assumir que os demais estímulos podem também contribuir no processo de imputação, basta supor que o tempo que os distanciam do momento presente seja um fator ponderador. Neste caso, um modelo de médias ponderadas pela recência da ocorrência do nível mostra-se bastante adequado. Considerando novamente o segundo estímulo da Tabela 1, o nível adotado para o atributo jornada de trabalho seria igual a 10 horas/dia, caso o respondente venha a utilizar um procedimento de imputação via recência ao quantificar sua preferência.

4.4 Imputação por associação

Embora exista uma forte suposição sobre a similaridade dos perfis no modelo de recência, espera-se que alguns perfis mostrem-se mais influentes, especialmente quando os atributos possuírem alguma correlação. Por exemplo, Huber e McCann (1982) mostram que os indivíduos utilizam suas crenças sobre a correlação entre preço e qualidade quando ao menos um destes atributos estiver ausente. Diferente dos processos descritos anteriormente, a imputação por associação concentra-se na informação dos níveis dos atributos presentes no estímulo para definir o nível do atributo ausente. Os padrões de associações podem variar de acordo com a interpretação feita pelo respondente, sendo ainda mais subjetivo que os procedimentos anteriores.

4.5 Imputação pelo Aprendizado

À medida que um respondente avalia mais estímulos, toda a informação recebida anteriormente tende a apresentar alguma influência na quantificação de sua preferência, isto porque sua avaliação torna-se mais crítica, dada sua experiência, ou aprendizado anterior. Alba e Cooke (2004) veem cada estímulo como sendo uma amostra do mundo real que vai sendo descoberto pelo respondente a cada nova avaliação. Aliada a essas descobertas, todo o conhecimento *a priori* e informações disponíveis vão compor o valor final da preferência.

Com o intuito de modelar este processo de aprendizagem, Bradlow, Hu e Ho (2004) propõem um modelo, denominado de modelo de aprendizado, que procura captar a forma como o respondente infere os níveis ausentes de um atributo combinando elementos de similaridade dos perfis, níveis de conhecimento *a priori* das características do produto e diferentes conceitos de métodos de imputação.

Uma vantagem desta técnica de imputação é que ela capta a influência não só dos níveis *a priori* de um mesmo atributo, mas também leva em consideração os níveis dos outros atributos presentes no estudo, útil quando existem atributos naturalmente correlacionados.

4.6 Modelagem estatística do processo de imputação

Descreveremos a modelagem do processo de imputação sugerida por Bradlow, Ho e Ho (2004), denominada de modelo de aprendizado.

Para descrever a forma como o i-ésimo indivíduo imputa a informação segundo o modelo de aprendizado, utilizam-se conceitos de associação, ou similaridade entre o perfil a ser avaliado e perfis anteriores. Tais conceitos bem definidos por Hock, Bradlow e Wansink (1999) e Camerer e Ho (1998). Com esses dois elementos, as probabilidades de escolha (imputação) de cada um dos níveis são derivadas, tornando-se possível descrever o processo de recência de memória envolvido durante a classificação da preferência do perfil e como ocorre o enfraquecimento (ou decaimento) da influência dos níveis dos atributos à medida que as comparações com os perfis anteriores são realizadas.

4.7 Similaridades dos perfis

A similaridade entre dois ou mais perfis é medida através da comparação par a par dos atributos, métrica de *Hamming*, própria para casos em que atributos binários são utilizados. Esta medida de similaridade associa o valor 1 quando dois estímulos possuem o mesmo nível em um determinado atributo e 0 caso contrário. Algebricamente, a métrica de *Hamming* pode ser traduzida como uma variável indicadora, que será igual a 1 sempre que o atributo ou os atributos observados possuírem níveis idênticos em dois estímulos distintos.

Considere um delineamento composto por somente três estímulos e quatro atributos, conforme descrito pela Tabela 2. Note que os estímulos não são descritos por todos os atributos simultaneamente, ou seja, são perfis incompletos. No primeiro deles, por exemplo, o atributo 3 é omitido na descrição. No instante seguinte, o quarto atributo é omitido e por fim é a vez do atributo 2 ser omitido na descrição do terceiro perfil.

Suponha também que o interesse seja imputar um nível para o segundo atributo do terceiro estímulo via método de aprendizado e que todos os atributos considerados assumem somente dois níveis (0 ou 1).

Tabela 2 - Exemplo ilustrativo de um delineamento com 4 atributos e 3 perfis incompletos.

Perfil	Atributo1	Atributo2	Atributo3	Atributo4
1	1	0		1
2	0	1	1	
3	1		1	0

Fonte: Bradlow, Hu e Ho (2004).

Como primeiro passo do processo de imputação via aprendizado, deve-se contabilizar as semelhanças encontradas entre os níveis dos atributos do terceiro estímulo e os anteriores. Para tal, deve-se simular todas as possíveis configurações dos perfis e assim compará-las. Tome, por exemplo, o atributo 2, assumindo que ele possui somente dois níveis, logo duas novas configurações para o terceiro estímulo são possíveis. Por exemplo, tem-se que xi2(3) representa o nível que o atributo 2 assume no terceiro estímulo e xi2(1) o nível do mesmo atributo para o primeiro estímulo. Caso seja assumido que o valor imputado inicialmente seja o valor padrão (nível 0) temos xi2(3) = xi2(1) = 0. Assim, a métrica de Hamming, dada por I[xi2(3) = xi2(1)], é igual a 1, indicando igualdade entre os níveis observados. Caso assumamos que o valor imputado seja 1, 1=xi2(3)≠ xi2(1) = 0, o que faz com que métrica de Hamming seja 0 (I[xi2(3) ≠ xi2(1)] =0).

4.8 Cálculo de coincidências

Admite-se que além da similaridade, a distância no tempo entre os estímulos avaliados também influencia no nível imputado e que tal influência decresce exponencialmente à medida que um maior afastamento no tempo seja observado. Tal efeito representa um efeito de memória no processo de imputação, que será modelado através de parâmetros λ ($\lambda \in [0; 1]$). Dessa forma, é possível combinar os conceitos de similaridade de perfis, aqui representado pela métrica de *Hamming*, com o efeito de memória para que seja possível quantificar o número de coincidências entre estímulos. As comparações devem ser feitas entre o estímulo sob avaliação e todos os estímulos anteriormente apresentados, bem como entre o atributo ausente e os demais atributos considerados no estudo.

A notação utilizada, a seguir, segue a sugestão dos autores Bradlow, Hu e Ho (2004). Considere o terceiro estímulo da Tabela 2, temos duas possibilidades: ou o entrevistado imputa o valor 1 ou o valor 0 o nível do atributo ausente. Assuma, inicialmente, que o entrevistado i imputou o valor 1 ao atributo ausente 2. Como os valores dos níveis do atributo 2 diferem entre o primeiro e o terceiro estímulos, temos que o primeiro estímulo não influenciou na imputação e sua contribuição na medida de influência será nula. Já ao comparar os estímulos 2 e 3, temos igualdade entre os níveis do atributo 2, portanto, iremos admitir que este estímulo tenha alguma influência na avaliação do estímulo 3. Ao se comparar os estímulos 2 e 3, note que há ainda coincidência no nível do atributo 3. A medida da influência dos estímulos 1 e 2 sobre o estímulo 3, ao imputar o valor 1 no atributo 2, será igual à influência do estímulo 2 sobre o estímulo 3 dada pela expressão (3)

$$N_{i2}(3; 1) = \lambda_2 + \lambda_{23}, \qquad (3)$$

sendo Ni2(3; 1) a experiência anterior do indivíduo i ao imputar o nível l do atributo 2 no 3 estímulo. Esta medida quantifica a coincidência encontrada entre os estímulos quando o nível 1 é imputado nos atributos ausentes.

$\lambda 2$ a intensidade com que a coincidência de níveis atributo 2 entre o segundo e terceiro estímulos influencia na imputação.

$\lambda 23$ a intensidade com que a coincidência entre níveis dos atributos 2 e 3 para o segundo e terceiro estímulos influencia na imputação.

Repetindo o exercício para o caso do valor imputado ser igual a zero encontra-se coincidência apenas no estímulo 1. Ao se comparar os estímulos 1 e 3 temos coincidência entre os atributos 1 e 2 e o número de coincidências passa a ser dado por:

$$N_{i2}(3; 0) = \lambda_2^2 + \lambda_{21}^2 \qquad (4)$$

O expoente 2 é uma penalização, uma vez que o estímulo 3 foi o segundo atributo apresentado ao entrevistado após o atributo 1. Caso estivéssemos comparando os estímulos 2 e 5, todos os parâmetros deveriam ser elevados à terceira potência. Como esses parâmetros (λ_{ij}) estão entre zero e 1, temos que quanto mais próximos forem os estímulos, maior será a influência na imputação.

4.9 Probabilidade de Imputação

Definidas as contagens das experiências anteriores (Ni2(3; 1) = λ2 + λ23 e Ni2(3; 0) = $\lambda_2^2 + \lambda_{21}^2$), o próximo passo é calcular a probabilidade de imputação de cada um dos níveis. Intuitivamente, aquele nível que apresentar uma maior similaridade entre estímulos e atributos será o imputado. A probabilidade de imputação de cada um dos níveis é calculada seguindo conceitos de probabilidades condicionais. Por exemplo, a probabilidade de que o atributo 2 do terceiro estímulo tenha um valor imputado igual a 1 é dada por:

$$P[x_{i0}(3)=1] = \frac{N_{i2}(3;1)}{N_{i2}(3;1)+N_{i2}(3;0)} = \frac{\lambda_2+\lambda_{23}}{\lambda_2^2+\lambda_{21}^2+(\lambda_2+\lambda_{23})} \tag{5}$$

4.10 Modelo Geral

De uma maneira geral, denota-se por *Nij(t,lj)* a experiência anterior do indivíduo *i* ao avaliar o nível l do atributo *j* no t-ésimo perfil. Para o caso *t=0*, *Nij(0,lj)* representa o conhecimento *a priori* do indivíduo *i* a respeito do atributo *l* enquanto que para os demais casos um processo de imputação de recência similar ao apresentado anteriormente é considerado. A expressão mais geral de *Nij(t,lj)*, mostrada em (6) é formada pela a soma do conhecimento *a priori* de cada indivíduo com as experiências anteriormente avaliadas pelos demais atributos, estando eles ausentes ou presentes.

$$N_{ij}(t,l_j) = \underbrace{N_{ij}(0,l_j)}_{priori} + \sum_{t'=1}^{t-1} \underbrace{(\lambda_{i(j,j)}^{t-t'} \times I[x_{ij}^{'}(t), x_{ij}(t')]}_{efeito\ do\ atributo\ ausente} + \underbrace{\sum_{j'=1, j'\neq j}^{J} \{(\lambda_{i(j,j)}^{t-t'} \times I[x_{ij'}(t), x_{ij'}(t')] \times I[x_{ij}^{'}(t), x_{ij}(t'),]\})}_{efeito\ dos\ atributos\ presentes} \tag{6}$$

É possível calcular a probabilidade de imputação de um determinado nível *lj* no t-ésimo perfil utilizando a expressão (7).

$$P[x_{ij}^{'}(t)=l_j] = \begin{cases} 1 \ \ se\ \ r_{ij}(t)=1 \ \ e\ \ x_{ij}^{'}(t)=l_j \\ 0 \ \ se\ \ r_{ij}(t)=1 \ \ e\ \ x_{ij}^{'}(t)\neq l_j \\ \dfrac{N_{ij}(t,l_j)}{N_{ij}(t,0)+N_{ij}(t,1)} \ \ se\ \ r_{ij}(t)=0 \end{cases} \tag{7}$$

O valor imputado será aquele que maximizar (7). Levando em conta uma eventual dependência nas avaliações, Bradow, Hu e Ho (2004) sugerem que se considere um efeito auto-regressivo para o erro aleatório $\varepsilon_i(t)$.

$$\varepsilon_i(t) = \gamma_i \varepsilon_i(t-1) + u_i(t) \quad (8)$$

O modelo estatístico utilizado, em todos estes casos, é o descrito pela expressão (2) e sua estimação será feita através da utilização de procedimentos bayesianos. O uso deste método de estimação vem crescendo muito nos últimos anos em pesquisas de *marketing*, principalmente depois da década de 1980, quando os primeiros trabalhos utilizando métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC) foram publicados (*Allemby et al*., 1995). Em 2004 Allemby, Bakken e Rossi exploram os avanços nas pesquisas em *Marketing* quanto à utilização de técnicas bayesianas, enfatizando seus benefícios.

Para a estimação dos parâmetros do modelo, utilizou-se os *softwares WinBugs* e *R* que dentre as suas vantagens constroem internamente a expressão para a distribuição *a posteriori* a partir das distribuições *prioris* apresentadas pela Tabela 3.

Tabela 3 - Distribuições a Priori e Hiperprioris utilizadas na estimação dos modelos.

Parâmetro	Priori	Hiperpriori
$u_i(t)/\sigma^2$	$N(0;\sigma^2)$	$\sigma^2 \sim \Gamma - Inv(a,b)$
γ_i	$U[m;n]$	
λ_{im}	$Beta(a_m, b_m$	$a_m e\ b_m \sim U[k;l]$
β_{ij}	$N(\overline{\beta_j};\sigma^2_{j\beta})$	$\overline{\beta_j} \sim N(r,s)\ e\ \sigma^2_{j\beta} \sim \Gamma - inv(c,d$
β'_{ij}	$N(\overline{\beta'_j};\sigma^2_{j\beta'})$	$\overline{\beta'_j} \sim N(u,v)\ e\ \sigma^2_{j\beta'} \sim \Gamma - inv(e,$
α_{ij}	$N(\overline{\alpha_j};\sigma^2_{j\alpha})$	$\overline{\alpha_j} \sim N(z,w)\ e\ \sigma^2_{j\alpha} \sim \Gamma - inv(g,h$

Fonte: Bradlow, Hu e Ho (2004).

sendo $N(\mu;\sigma^2)$ uma distribuição normal com média μ e variância σ^2; $\Gamma - Inv(a,b)$ uma distribuição *Gama* Inversa com parâmetros *a* e *b*; *U[k;l]* uma distribuição Uniforme no intervalo *[k;l]* e *Beta[a;b]* uma distribuição *Beta* com parâmetros *a* e *b*.

5. Aplicação: preferências no Ambiente de Trabalho

Pretende-se avaliar a performance do modelo de imputação via aprendizado, proposto por Bradlow, Hu e Ho (2004), através de sua aplicação em estudo sobre motivação no trabalho.

Nas últimas décadas, a preocupação com a satisfação dos funcionários aumentou, isso porque a produtividade de um indivíduo está atrelada a sua capacitação e a sua motivação. Muitas empresas desenvolvem campanhas de motivação, pois acreditam que uma equipe motivada possui maior propensão a desenvolver novos projetos e inovar tecnologicamente, o que gera vantagem competitiva da companhia em relação aos concorrentes.

Em conjunto com um profissional da área de Recursos Humanos, e tendo como apoio a obra de Robbins (2002), definiram-se os principais atributos que afetariam a preferência dos indivíduos no ambiente de trabalho. A escolha dos níveis e a definição de qual seria o nível padrão baseou-se nas características de trabalho de empresas do ramo financeiro, uma vez que funcionários deste segmento da economia foram os participantes deste estudo. Dentre os atributos selecionados temos: Remuneração, Reconhecimento, Exposição a níveis superiores de hierarquia, Jornada de trabalho, Cima na área, Forma de gestão e Novos Desafios.

Estes sete atributos, com dois níveis cada, selecionados, geraram 128 (27=128) estímulos compostos por diferentes combinações de atributos e níveis. Dado o elevado número de estímulos, torna-se inviável a avaliação do delineamento completo por parte dos indivíduos (*Bradlow*, *Hu* e *Ho*, 2004). A estratégia adotada para a formação dos estímulos contou com utilização de delineamentos fatoriais fracionários (*Plackett* e *Burman*, 1946). Com isso, 20 estímulos foram selecionados: 16 para estimação das utilidades e 4 para validação (*holdouts*), todos eles ortogonais, o que garante a estimabilidade dos efeitos principais sem que seja necessário testar todas as combinações dos níveis dos fatores.

Para comparar as estimativas das utilidades parciais sob diferentes métodos de estimação, e verificar a influência da omissão de um ou mais atributos na preferência dos respondentes, optou-se por realizar a pesquisa em duas etapas, respeitando procedimentos de um estudo *cross-over* (*Agresti*, 1990). Assim, na primeira etapa da pesquisa, um grupo

de respondentes avaliou estímulos completos, ou seja, estímulos descritos por todos os atributos considerados no estudo, enquanto isso, o segundo e o terceiro grupos avaliaram atributos, formados pela mesma combinação de níveis de atributos, exceto pela omissão de um ou mais atributos em cada um dos estímulos. Na etapa seguinte, aqueles respondentes que receberam estímulos completos passaram a avaliar estímulos incompletos (com um ou dois atributos ausentes) e vice-versa. Optou-se ainda pela não omissão dos atributos Remuneração e Reconhecimento, isso porque eles são frequentemente confundidos pelos indivíduos ao questionar sobre quais fatores são importantes em um ambiente de trabalho. O primeiro refere-se mais ao aspecto financeiro e o segundo ao reconhecimento ou elogio recebido por uma atividade bem desempenhada.

Foram pesquisados 74 funcionários de distintas instituições financeiras e áreas de negócio. Na primeira fase da pesquisa, além da classificação das preferências, coletaram-se algumas variáveis demográficas, tais como sexo, idade, tempo de empresa e escolaridade, a fim de verificar se a maturidade de um funcionário pode alterar substancialmente suas preferências.

Na fase de classificação e quantificação de suas preferências, os respondentes foram convidados a atribuir uma nota de 0 a 100 para cada estímulo apresentado, tomando o cuidado de não repetir a mesma nota para estímulos distintos.

A apresentação dos estímulos foi antecedida por uma breve descrição dos atributos e níveis avaliados, bem como uma avaliação de um estímulo composto por todos os níveis alternativos. Para a quantificação da preferência, foram utilizados os 16 primeiros estímulos para a estimação das utilidades parciais e os quatro restantes para validação (estímulos *holdout*).

As duas etapas de classificação propostas pelo delineamento *cross-over* foram espaçadas por 30 dias, o que pode minimizar o efeito de memória de um questionário para o seguinte.

6. Resultados

Modelos distintos quanto à regra de imputação e à forma funcional foram considerados[3] na estimação das utilidades parciais com a finalidade de comparar os resultados das estimativas dos parâmetros com as obtidas no modelo de perfis completos. No geral, todos os modelos apresentaram um bom ajuste. Quanto à convergência, cerca de 5% dos casos de estimação, segundo o modelo de aprendizado, apresentaram algum problema, principalmente quando os valores de preferência eram muito similares. Já os demais modelos não apresentaram problemas de convergência. As distribuições *a priori* utilizadas neste estudo são não informativas cujos parâmetros são apresentados no Anexo I.

Além do modelo de perfis completos e da imputação através do modelo de aprendizado, a imputação via valor padrão e a imputação via recência também foram testadas. Nos dois últimos casos, houve ainda um teste quanto à forma funcional do modelo com a inclusão do parâmetro que captura o efeito da omissão do atributo na descrição de um estímulo (β'), conforme apresentado na equação (2).

As estimativas das utilidades parciais médias exibidas pela Tabelas 4, no geral, apresentam similaridades no que diz respeito ao sinal dos coeficientes dos atributos Reconhecimento e Jornada de Trabalho, ambos negativos. O sinal negativo representa, respectivamente, que a preferência dos funcionários é menor quando o reconhecimento vem dos colegas e/ou quando a jornada de trabalho proposta supera 50 horas semanais.

No caso do modelo de perfis completos, dos 48 respondentes, 70% apontam a Remuneração como principal fator de preferência por um emprego, sendo que os demais atributos não foram os mais importantes em mais de 10% dos casos. Para os modelos de perfis incompletos com um atributo omitido, embora a remuneração tenha sido encontrada mais frequentemente como o atributo mais importante, de 37% a 54% dos casos, outros atributos como a Forma de Gestão, Clima na área e Desafios também se destacam.

[3] As macros utilizadas na estimação desses modelos encontram-se em Pretto (2007).

Tabela 4 - Comparação entre utilidades parciais estimadas para os diferentes modelos.

Modelo		Remuneração	Reconhecimento	Exposição	Jornada de trabalho	Gestão	Clima na área	Desafios	Intercepto
Completo	Média	22,76	-2,44	5,9	-4,47	11,37	2,79	9,29	28,46
	d.p.	16,13	13,13	6,62	13,15	10,47	12,97	4,7	18,6
Padrão	Média	24,02	-1,32	9,06	-5,58	4,65	13,86	10,98	28,36
	d.p.	15,34	5,57	10,36	12,64	9,09	7,42	6,91	14,08
Padrão + β'	Média	24,17	-2,47	5,99	-10,37	7,69	8,93	9,67	8,72
	d.p.	11,23	1,21	4,61	2,52	1,57	2,12	5,44	7,51
Recência	Média	22,56	0,07	8,76	-3,81	11	10,92	12,19	21,69
	d.p.	15,95	6,37	9,97	9,98	8,38	9,65	7,86	13,09
Recência + β'	Média	23,18	-2,1	7,19	-5,94	10,61	8,44	7,11	7,06
	d.p.	9,89	1,55	3,64	1,57	2,54	4,36	2,83	4,15
Aprendizado	Média	20,92	-0,23	7,48	-6,02	8,5	10,47	11,1	8,72
	d.p.	17,62	3,78	7,71	9,66	7,6	8,73	7,93	7,51

Avaliando a importância relativa de cada atributo avaliado sobre a preferência total de cada indivíduo, percebe-se uma boa concordância entre as estimativas encontradas dos modelos de aprendizado com o modelo de estímulos completos. A correlação de Spearman entre as importâncias relativas do modelo completo e do modelo de aprendizado é de 71%. Correlações altas também foram encontradas entre o modelo completo e os modelos com imputação através do método de recência, com e sem o indicador do efeito da omissão do atributo, respectivamente, 75% e 86%.

Outros resultados de ajuste do modelo podem ser encontrados no Anexo I.

7. Considerações finais

Neste trabalho, foram apresentados diferentes métodos de estimação das utilidades parciais para o caso em que perfis incompletos são utilizados em estudos de APC. Em especial, avaliou-se o desempenho do método de imputação via aprendizado, proposto por Bradlow, Hu e Ho (2004), que mostrou bons resultados quando comparado aos demais métodos mais conhecidos na literatura.

O desafio de aplicar a metodologia de APC, em um estudo sobre preferências no trabalho, tem por objetivo ampliar a utilização desta metodologia de pesquisa em outras áreas de conhecimento, uma vez que já possui boa aceitação em áreas de pesquisa de mercado e desenvolvimento de produtos.

No que se refere à estimação dos modelos, a utilização de perfis incompletos na estimação das utilidades parciais parece ser uma boa saída nos casos onde muitos atributos ou níveis são testados simultaneamente. Mesmo que ao avaliar um estímulo com atributos omitidos (perfil incompleto) um respondente tenda a compensar a ausência de informação utilizando algum procedimento de imputação, as estimativas das utilidades parciais encontradas não se mostram muito diferentes daquelas encontradas quando perfis completos são avaliados.

Neste estudo sobre preferências no ambiente de trabalho, o modelo de Aprendizado, embora tenha apresentado bons resultados, demandou um tempo de processamento muito superior aos demais métodos, o que o deixa em desvantagem frente ao modelo de imputação via recência, por exemplo. O modelo de recência, além de resultados muito próximos ao modelo de perfis completos, possui formulação relativamente simples, além de apresentar um baixo tempo de processamento, mostrando-se assim um bom método de imputação para estes dados.

Como primeira impressão, embora a implementação e a estimação do modelo exijam um tempo relativamente maior do que os métodos clássicos, a estimação por métodos bayesianos mostrou-se uma ferramenta muito promissora em estudos de APC, uma vez que é menos restritiva do que os métodos clássicos, principalmente em estudos onde a heterogeneidade entre respondentes é grande, sendo também robusta, mesmo quando poucos estímulos são avaliados.

Por fim, embora a aplicação proposta contivesse somente dois níveis para cada atributo, o modelo proposto suporta que um número maior de níveis seja avaliado não necessariamente sendo o mesmo entre todos os atributos. Sendo assim, estudos futuros podem ser realizados aumentando o número de níveis dos atributos ou, ainda, aumentando o número de omissões dos estímulos. Outra sugestão seria a utilização simultânea de diferentes procedimentos de imputação em um único modelo, tornando o modelo mais flexível e abrangente do que o atual.

Anexo

As distribuições *a priori* utilizadas neste estudo são *prioris* não informativas, similares àquelas utilizadas no estudo original dos autores Bradlow, Hu e Ho (2004).

Tabela 5 - Distribuições *Priori* não informativas

Parâmetro	Priori	Hiperpriori
$u_i(t)$ / σ^2	$N(0;\sigma^2)$	$\sigma^2 \sim \Gamma - Inv(0.1,0.1)$
γ_i	$U[-1;1]$	
λ_{im}	$Beta(a_m, b_m$	$a_m e\ b_m \sim U[1;100]$
β_{ij}	$N(\overline{\beta_j};\sigma^2_{j\beta})$	$\overline{\beta_j} \sim N(0,0.01)\ e\ \sigma^2_{j\beta} \sim \Gamma - inv(0.1,0.$
β'_{ij}	$N(\overline{\beta'_j};\sigma^2_{j\beta'})$	$\overline{\beta'_j} \sim N(0,0.01)\ e\ \sigma^2_{j\beta'} \sim \Gamma - inv(0.1,0$
α_{ij}	$N(\overline{\alpha_j};\sigma^2_{j\alpha})$	$\overline{\alpha_j} \sim N(0,0.01)\ e\ \sigma^2_{j\alpha} \sim \Gamma - inv(0.1,0.$

A Tabela 6, abaixo, traz as medidas de diagnóstico dos modelos de APC avaliados. Comparando-se os resultados ao modelo completo, observam-se os bons resultados dos modelos de Recência e Padrão para o critério de AIC e o bom desempenho do modelo de Aprendizado nas medidas de erros de estimação, tais como: Erro Quadrático Médio - EQM e Erro Absoluto Médio - EAM, bem como na correlação entre valores estimados e observados.

Tabela 6 - Medidas de diagnóstico dos modelos de APC

Medidas de ajuste		Completo	Padrão	Padrão (β')	Recência	Recência (β')	Aprendizado
	lnL	-2,91	-2,97	-2,96	-2,93	-3,01	-2,26
	AIC	186,11	190,21	331,86	187,52	337,57	307,50
EAM	Estimação	7,69	9,9	10,36	8,47	10,78	6,58
	(%)		28,7%	34,7%	10,1%	40,2%	-14,4%
	Validação	15,55	18,36	13,35	15,97	14,87	15,19
	(%)		18,1%	-14,1%	2,7%	-4,4%	-2,3%
EQM	Estimação	116,59	204,83	204,80	136,65	217,92	140,52
	(%)		75,7%	75,7%	17,2%	86,9%	20,5%
	Validação	439,21	500,61	322,59	381,76	335,45	384,02
	(%)		14,0%	-26,6%	-13,1%	-23,6%	-12,6%
Corr	(y e ^y)	76,7%	78,6%	73,9%	81,3%	77,8%	89,9%
	γ	0,092	0,192	0,18	0,177	0,177	0,1

Referências Bibliográficas

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley and Sons.

Alba, J. W. e Cooke, A. D. J. (2004). When Absence Begets Inference in Conjoint Analysis. Journal of Marketing Research, (41): 382-387.

Allemby, G.M., Arora, N., Ginter, J.L. (1995). Incorporating Prior knowledge into the analysis of conjoint studies. Journal of Marketing Research, (XXXII): 152-162 .

Allenby, G. M., Bakken, D. G., Rossi, P.E. (2004). The HB Revolution: How Bayesian Methods Have Changed the Face of Marketing Research. Journal of Marketing Research, XVI(2): 20-25.

Bradlow, E.T, Hu. Y. e Ho, T.(2004). A Learning-Based Model for Imputin Missing Levels in Partial Conjoint Profiles. Journal of Marketing Research, (XLI): 369-381.

Broniarczyk, S. M. e Alba, J. A. (1994). The Role of Consumers'Intuitions in Inference Making. Journal of Consumer Research, (XXI): 393-407.

Camerer, C. e Ho, T.H. (1998). EWA Learning in Condination Games: Probability Rules, heterogeneity, and Time Variation. Journal of Mathematical Psycology, (42): 305-326.

Feldman, J.M. e Lynch, J. G. (1988). Self-generated Validity and Other E□ects of Measuerem on Belief, Attitude, Intention and Behavior. Journal of Applied Psychology, (73): 421-435.

Friedmann, L. S. (1998). Análise de Preferência. Dissertação de Mestrado. Escola de Administração de Empresas da Fundação Getúlio Vargas (FGV -EAESP). São Paulo.

Green, P.E. (1974). On the Design of Choice Experiments Involving Multifactor Alternatives. Journal of Consumer Research, (1): 61-68.

Green, P. E.(1984). Hybrid Model for Conjoint Analysis: An Expository Review. Journal of Marketing Research, (XXI): 155-169.

Green, P.E., Carrol, J.D. e Carmone, F.J. (1978). Some New Types of Fractional Factorial Designs for Marketing Experiments. Research in Marketing, (1): 99-122.

Green, P. e Rao, V. (1971). Conjoint Measurement for Quantifying Judgmental Data. Journal of Marketing Research, (8): 355-363 .

Green, P. E. e Savitz, J.(1994). Applying Conjoint analysis to Product Assortment and Pricing in Retailing Research. Pricing Strategy and Pratice Journal , (2): 4-19.

Green, P. E. e Srinivasan, V. (1978). Conjoint Analysis in Consumer Research"Issues and Outlook. Journal of Consumer Research, (5): p 103-123.

Green, P. E. e Srinivasan,V. (1990). Conjoint Analysis in Marketing: New Developments with implications for Research and Practice. Journal of Marketing, (54): 3-19.

Ho, T.H. e Chong, J.K (2003). A Parsimonious Model of SKU Choice. Journal of Marketing Research, (40): 351-365.

Hoch, S.J., Bradlow, E. T. e Wansink, B (1999). The variety of an Assortment. Marketing Science, 18(4): 527-546.

Huber, J. e McCann, J. W. (1982). The Impact of Inferential Baliefs on Product Evaluations. Journal of Marketing Research, (XIX): 324-333.

Louviere, J.L., Eagle, T.C e Cohen, S.H. (2005). Conjoint Analysis: Methods, Myths and Much More. Centre for the Study of Choice of the University of Technology of Sidney Working Paper Series. http://www.business.uts.edu.au/censoc/papers/wp05001.pdf, acessado em 07/08/2008.

Luce, R. D. e Tukey, J.W. (1964). Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. Journal of Mathematical Psycology, (1): 1-27.

Lynch, J. G., Jr e Srull, T.K.(1982). Memory and Attentional Factors in Consumer Choice: Concepts and Research Methods. Journal of Consumer Research, (9): p18-37.

Montgomery, D. C. (1996). Design and Analysis of Experiments. (4nd ed.). John&Wiley Sons.

Plackett, L. R. e Burman, J.P. (1946). The Design of Optimum Multifactorial Experiments. Biometrika, (XXXIII): 305-325.

Pretto, K. Modelando o Efeito da Omissão de Atributos em um Estudo de Análise de Preferência Conjunta. Dissertação de Mestrado (IME-USP).

Rao, V.R. (2007). Developments in Conjoint Analysis. Working paper. Cornell University. http://forum.johnson.cornell.edu/faculty/rao/Developments%20in%20Conjoint%20Analysis%20July%202007.pdf. Acessado em 07/08/2008.

Robbins, S.P.(2002). Comportamento organizacional I. (9 ed.) S˜ao Paulo: Prentice Hall.

Yamagishi, T. e Hill, C.T.(1981). Adding Versus Averaging Models Revisited: A test of a Path-Analytic Integration Model. Journal of Personality ans Social Psycology, 41(1): 13-25.

Abstract

Conjoint Analysis is a statistical technique widely applied in marketing research. This methodology is applied to identify attributes which are associated with customer's preferences. When a large number of attributes are considered in a conjoint study, the final number of all possible profiles increases dramatically. In the case of excessive possible profiles, fractional designs and incomplete profiles can be used instead. In this study, the missing information effects and the magnitude of the impact of dealing with partial profiles will be tested. Different calculation methods are discussed and a real case study is presented at the end of this paper.

MODELAGEM DE EXTREMOS METEOROLÓGICOS VIA GEV E GPD – UMA ANÁLISE COMPARATIVA DE ALGUMAS CAPITAIS BRASILEIRAS

Rita de Cássia de Lima Idalino[4]
Pâmela Sabrina de Oliveira[4]
Paulo Sérgio Lucio[4]

Resumo

Motivados pela tendência em extremos climáticos em conexão com as supostas mudanças climáticas, o objetivo deste estudo é apresentar, por meio da teoria de valores extremos, modelos que mostrem os impactos de eventos raros na sociedade e ecossistemas para prevenção e propostas de mitigação. Assim, medidas estruturais e não estruturais podem ser tomadas a fim de reduzir os impactos. Os dados desse estudo correspondem a precipitações diárias dos anos de 1951 a 2005 de cinco capitais brasileiras. Dada a ocorrência de eventos extremos, buscou-se verificar a adequação da distribuição generalizada de valores extremos e distribuição generalizada de Pareto para estimar a precipitação que ocorrerá em um dado período. Através dos ajustes via GEV e GPD, pode-se concluir que os níveis de retorno de 100 mm são mais frequentes em Recife. A comparação dos resultados para os dois métodos utilizados abre uma discussão sobre a confiabilidade das estimativas fornecidas.

Palavras-chave: POT, Nível de Retorno, Período de Retorno e Precipitação.

[4] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, 59072-970. Natal, RN - Brasil.

1. introdução

Ocorrências meteorológicas recentes, como grandes catástrofes, chamaram a atenção pela severidade das consequentes perdas. Preocupada com as mudanças climáticas ocorridas em todo o planeta, a Organizações das Nações Unidas - ONU cria em 1988 o Painel Intergovernamental para Mudanças Climáticas Intergovernmental Panel on Climate Change – IPCC, o qual afirmou que há evidências de que eventos extremos como secas, enchentes, ondas de calor e de frio, furacões e tempestades têm afetado diferentes partes do planeta, produzindo assim enormes perdas econômicas e de vidas (IPCC 2001b). A agência meteorológica da ONU divulgou que no início de 2007 o mundo registrou uma série de eventos climáticos extremos, como as enchentes na Ásia, as ondas de calor na Europa e a precipitação de neve na África do Sul. No Brasil, as chuvas e as secas intensas já estão aumentando, afirma o Instituto de Pesquisas Espaciais - INPE. O furacão Catarina, ocorrido em 2004, o aumento da incidência de tornados, a seca da Amazônia ocorrida em 2005 e as secas já observadas no Nordeste brasileiro (NEb) estão cada vez mais frequentes e intensas. Nos meses de novembro e dezembro de 2008, vários municípios de Santa Catarina, do Rio de Janeiro e de Minas Gerais decretaram estado de calamidade devido às fortes chuvas, as quais causaram destruição total de casas, deixando milhares de pessoas desabrigadas e centenas de mortos.

Eventos deste porte, por serem de baixa frequência e grande impacto, são de difíceis previsões por parte de qualquer especialista que se proponha a prover proteção ou a reter este risco. Assim, torna-se importante a identificação e análise estatística de extremos severos. Os eventos severos podem ser caracterizados tanto pela intensidade de manifestação de um parâmetro meteorológico, tais como: chuvas e ventos intensos, ou pela duração prolongada de um dado fenômeno, a exemplo do que ocorre em regiões serranas, onde uma chuva leve e constante pode levar a sérios deslizamentos de terra.

É de suma importância que os riscos meteorológicos que estamos expostos sejam devidamente avaliados e corretamente dimensionados. Com intuito de prover informações a respeito do período de retorno dos eventos severos e avaliar os riscos de situações extraordinárias, relativas à precipitação de algumas capitais brasileiras, este trabalho faz uso da teoria de valores extremos, uma ferramenta poderosa para inferir sob as caudas das distribuições de probabilidades destes eventos.

2. Dados observacionais

Foram analisadas as séries históricas de precipitações diárias correspondentes ao período de janeiro de 1961 a dezembro de 2005 das seguintes cidades: Belo Horizonte, Curitiba, Goiânia, Manaus e Recife. As observações foram registradas nas estações meteorológicas do Instituto Nacional de Meteorologia - INMET. A base de dados não está completa, havendo algumas observações faltantes por motivos não esclarecidos. Esses dados foram tratados sob a ótica de valores extremos, metodologia detalhada mais adiante. As séries temporais desses dados encontram-se na Figura 1.

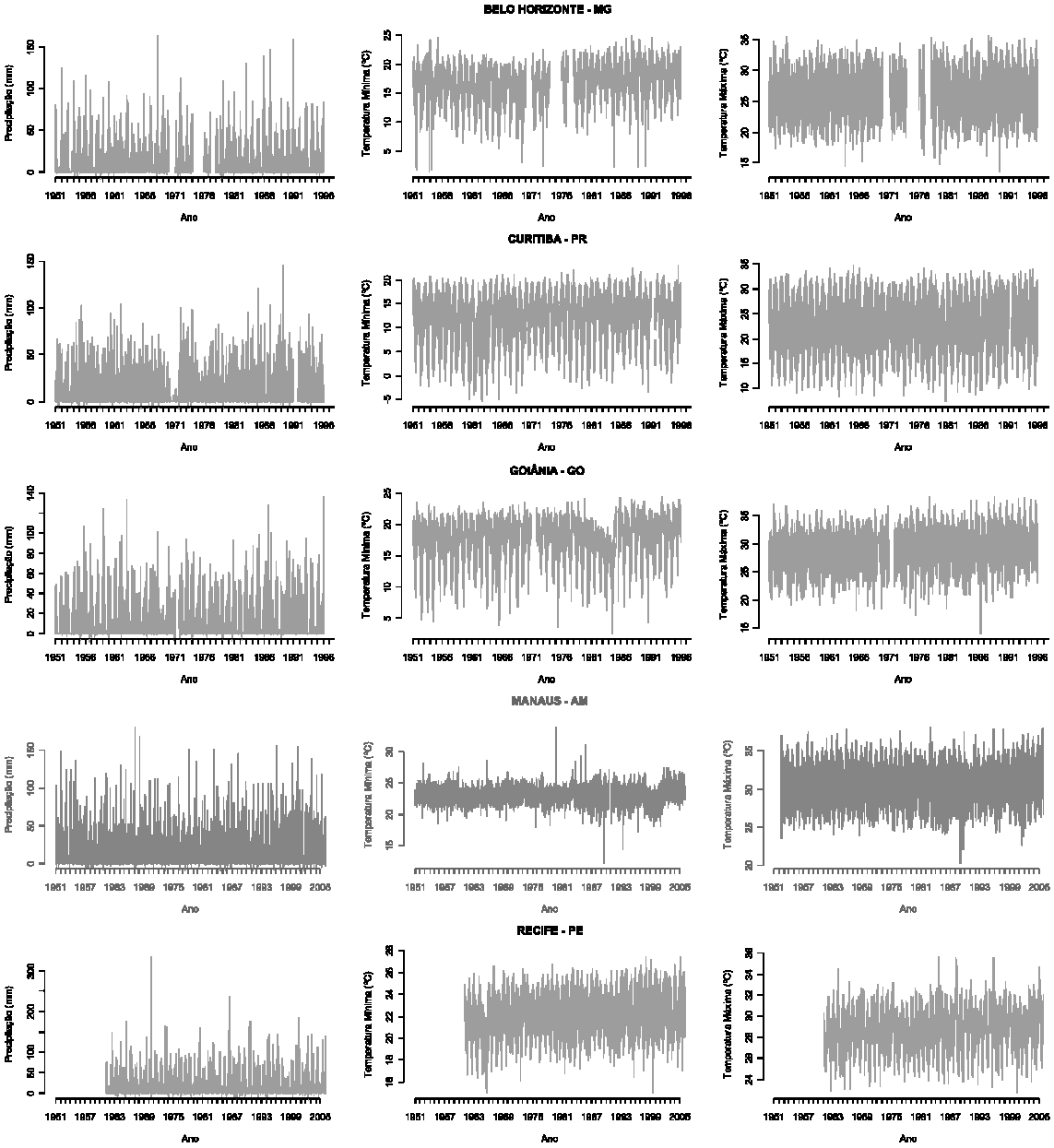


Figura 1: Série temporal das precipitações e temperaturas mínimas e máximas diárias (1961-2005) Dados oriundos do INMET

3. Valores extremos

Eventos extremos são aqueles que ocorrem nas caudas das distribuições, sendo esses de baixa frequência. Não há um consenso em relação à definição exata de valores extremos. Algumas literaturas afirmam que extremos são definidos como sendo os valores máximos em blocos (*Block Maxima*), enquanto outras definem como sendo os dados excedentes de um limiar (*POT - Peaks Over Threshold*). A Figura 2 ilustra essas duas situações.

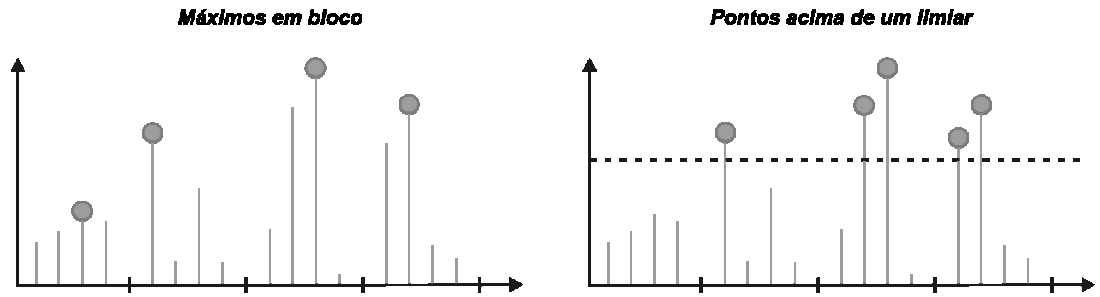


Figura 2: Ilustração das definições de extremo.

Os primeiros trabalhos teóricos em extremos foram baseados no máximo (empírico) de um conjunto de observações consideradas como variáveis aleatórias (máximo do bloco/conjunto). A distribuição estatística para eventos extremos a partir do método máximos em blocos foi determinada por Fisher e Tippet (1928) e demonstrada teoricamente por Gnedenko (1943) através do teorema dos três tipos. Este teorema estabelece que a distribuição de uma amostra aleatória de valores extremos somente pode, no limite, tender a três tipos de distribuições: Gumbel; Frèchet; e Weibull. Jenkinson (1955) demonstrou e unificou as três distribuições em uma única equação (Equação 1), denominada de GEV (COLES, 2001).

Precisamente, dadas n variáveis aleatórias independente e identicamente distribuídas (i.i.d.'s) $X_1, X_2, \ldots, X_n$, a teoria de valores extremos busca (dentre outras coisas) as leis de probabilidade de $Y_n = \max(X_1, X_2, \ldots, X_n)$. O Teorema dos Três Tipos garante a existência de uma distribuição limite $F(X)$ de Y_n quando $n \to \infty$ (n é grande). As três possíveis distribuições de probabilidade são:

(Gumbel ou EV-1) $P(Y_n \leq x) = F(x) = \exp\left[-\exp(-x)\right]$, $\forall x \in \Re$;

(Fréchet ou EV-2) $P(Y_n \leq x) = F(x) = \exp\left[(-x^{-\xi})\right]$, $\forall x > 0$;

(Weilbull ou EV-3) $P(Y_n \leq x) = F(x) = \exp\left[-|x|^{\xi}\right]$, $\forall x < 0$.

Estas três distribuições podem ser estendidas através da distribuição de Valores Extremos Generalizados - GEV:

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{(x-\mu)}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad (3.1)$$

para $1+\xi\left(\frac{(x-\mu)}{\sigma}\right) > 0.$ Nesta equação μ é um parâmetro de locação, σ é um parâmetro de escala, e ξ é um parâmetro de forma. As três distribuições individuais previamente apresentadas podem ser desdobradas a partir de uma GEV da seguinte forma:

(Gumbel ou GEV-1), quando $\xi \to 0$;

(Fréchet ou GEV-2), quando $\xi > 0$;

(Weilbull ou GEV-3), quando $\xi < 0$.

Basicamente, a Teoria de Valores Extremos trabalha ou mesmo maneja a cauda da distribuição principal (ou parental). Na prática, isto pode ser feito buscando uma distribuição empírica que se "ajuste" à distribuição da cauda da distribuição principal, usando os dados disponíveis. Dados os poucos registos de muitas séries temporais observadas, uma transigência é necessária de ser acordada quando se aplica a teoria de valores extremos no "mundo real". Poucas estratégias são geralmente aplicadas:

Considerar máximo ou mínimo de referência no período sob estudo (informações mensais ou anuais), pois nesta etapa muitos dados observados são desconsiderados;

Considerar pontos acima/abaixo de um dado patamar (fixo ou aleatório): pontos acima/abaixo do limiar. Este esquema envolve a teoria de processos pontuais e nos conduz à Distribuição Generalizada de Pareto - GPD: $G(x) = P(X > u + x | X > u) = (1+\xi x/\sigma)^{-1/\xi}$;

Ordenar (hierarquizar) os dados e analisar sob o enfoque estatístico as k maiores.

Uma vez definida a estratégia de análise, os parâmetros da distribuição resultante podem ser estimados via uma das seguintes metodologias:

Técnicas Gráficas – gráficos de probabilidade ou gráficos do tipo quantil-quantil (Q-Q Plot);

Método dos Momentos;
Método baseado nos Momentos – L;
Método da Máxima Verosimilhança;
Métodos Bayesianos.

É importante observar que o estudo da teoria dos valores extremos nos permite:

modelar eventos extremos com o propósito de predição, em geral, relação entre mudanças na média e/ou variância da distribuição e a intensidade e frequência de eventos extremos;

calcular o nível de retorno z_p, correspondente ao período de retorno $1/p$ ($p << 1$). Este é o $(1-p)-ésimo$ quantil, x_{1-p}, da distribuição de valores extremos F dada por: $z_p = x_{1-p} = F^{-1}(1-p)$. O $p-ésimo$ quantil, x_p, satisfaz à relação $F(x_p) = p$ ou analogamente $x_p = F^{-1}(p)$. O nível de retorno é uma quantidade muito útil em Planejamentos Estrutural e Territorial;

estudar o acesso a impactos de eventos extremos na sociedade e ecossistemas para prevenção e propostas de mitigação;

analisar a dependência temporal de eventos extremos, que é muito útil no acesso a tendências em extremos climáticos em conexão com as supostas mudanças climáticas.

A GPD (Equação 2) representa três distribuições: a Exponencial, a Pareto, e a Beta. Assim como a GEV é a distribuição limite para os máximos em blocos, a do tipo GPD é a forma paramétrica para distribuição dos extremos a partir de um limiar (Teorema de r *Balkema-de Haan*). O limiar considerado neste trabalho foi aquele que selecionava 2% dos dados localizados na cauda superior da distribuição.

Para ambas as abordagens, o parâmetro das distribuições mais importante é o de forma (ξ), pois este define para qual distribuição convergirá.

$$F(x;\mu,\sigma,\xi)=1-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\begin{array}{l}\xi=0,\ \text{Exponencial}\\ \xi>0,\ \text{Pareto}\\ \xi<0,\ \text{Beta}\end{array} \qquad (3.2)$$

Os parâmetros das distribuições foram estimados através do método da máxima verossimilhança. Para verificar a adequabilidade dos dados para as distribuições GEV e GPD foi realizado o teste não paramétrico de Kolmogorov-Smirnov, o qual verifica se uma das distribuições de probabilidade subjacentes difere da distribuição em hipótese.

4. Estimação dos parâmetros da distribuição GEV

O método mais geral e flexível para a estimação de um parâmetro desconhecido θ pertencente a uma família F é a máxima verossimilhança. Cada valor de θ potencia diferentes valores de probabilidades dos dados observados. A probabilidade de um dado observado como função de θ é chamada função verossimilhança. Funções de θ que tenham alta verossimilhança correspondem a modelos que nos dão uma alta probabilidade para os dados observados.

Considerando que $x_1, x_2, ..., x_n$ são uma série de realizações aleatórias independentes e identicamente distribuídas e ordenadas, com função densidade de probabilidade da GVE, a função de verossimilhança é

$$L(\theta)=L(\mu,\sigma,\xi)=\prod_{i=1}^{n} f(x_i;\theta)=\frac{1}{\sigma^n}\prod_{i=1}^{n}\left\{\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)}\right\}\exp\left\{\sum_{i=1}^{n}\left\{-\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\right\} \qquad (4.1)$$

que para $\xi<0$, assume valores diferentes de zero, se todos os valores de x_i $(i=1,2,...,n)$ forem menores do que $\mu-\sigma/\xi$, ou seja, se $\mu-\sigma/\xi>x_n$, sendo x_n o maior valor da série de observações, e para $\xi>0$, se todos os valores de x_i $(i=1,2,...,n)$ forem maiores que $\mu-\sigma/\xi$, ou seja, $\mu-\sigma/\xi<x_1$ o menor valor da série de observações. Caso contrário $L(\theta)=0$.

É mais conveniente tomar o logaritmo da função verossimilhança, que é dado por:

$$l(\mu,\sigma,\xi)=\ln[L(\mu,\sigma,\xi)]=-n\ln\sigma-\left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{n}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

$$=\sum_{i=1}^{n}\left\{-\ln\sigma-\left(\frac{1+\xi}{\xi}\right)\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (4.2)$$

para $\mu-\sigma/\xi>x_n$ e $\xi<0$ ou $\mu-\sigma/\xi<x_1$ se $\xi>0$. Caso contrário $l(\mu,\sigma,\xi)$ não existe.

Os estimadores de máxima verossimilhança de μ, σ e ξ são obtidos maximizando o logaritmo da função verossimilhança $l(\mu,\sigma,\xi)$ em relação a cada parâmetro e a raiz obtida, a sua solução. Assim:

$$\frac{\partial}{\partial\mu}l(\mu,\sigma,\xi)_{\mu=\mu_0}=0$$

$$\frac{\partial}{\partial\sigma}l(\mu,\sigma,\xi)_{\sigma=\sigma_0}=0$$

$$\frac{\partial}{\partial\xi}l(\mu,\sigma,\xi)_{\xi=\xi_0}=0$$

ou, seja:

$$\frac{1}{\hat{\sigma}}\sum_{i=1}^{n}\left(\frac{1+\hat{\xi}-w_i^{-\frac{1}{\hat{\xi}}}}{\hat{\sigma}}\right)=0$$

$$-\frac{n}{\hat{\sigma}}+\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\left\{\frac{\left(x_i-\hat{\mu}\right)\left[\left(1+\hat{\xi}\right)-w_i^{-\frac{1}{\hat{\xi}}}\right]}{w_i}\right\}=0 \quad (4.3)$$

$$\sum_{i=1}^{n}\left\{\left(1-w_i^{-\frac{1}{\hat{\xi}}}\right)\left[\frac{1}{\hat{\xi}^2}\ln(w_i)-\frac{\left(x_i-\hat{\mu}\right)}{\hat{\xi}\,\hat{\sigma}\,w_i}\right]-\frac{\left(x_i-\hat{\mu}\right)}{\hat{\sigma}\,w_i}\right\}=0,$$

sendo $w_i=1+\xi\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)$,

Como o sistema de equações (4.3) não possui solução analítica, utilizaram-se procedimentos iterativos para obter as estimativas dos parâmetros de máxima verossimilhança, usando a informação da matriz informação esperança M. A formula iterativa é, para $j \geq 0$:

$$\theta^{(j+1)} = \theta^{(j)} + M\left(\hat{\theta}\right)^{-1} grad\ l\left(\theta^{j}\right)$$

onde $\theta = (\mu, \sigma, \xi)$.com :

$$-\ grad\ l(\theta) = \left(-\frac{\partial l}{\partial \mu}, -\frac{\partial l}{\partial \sigma}, -\frac{\partial l}{\partial \xi}\right)$$

$$M\left(\hat{\theta}\right) = \begin{bmatrix} -E\left(\frac{\partial^2 l}{\partial \sigma^2}\right) & -E\left(\frac{\partial^2 l}{\partial \mu \partial \sigma}\right) & -E\left(\frac{\partial^2 l}{\partial \xi \partial \sigma}\right) \\ -E\left(\frac{\partial^2 l}{\partial \mu \partial \sigma}\right) & -E\left(\frac{\partial^2 l}{\partial \mu^2}\right) & -E\left(\frac{\partial^2 l}{\partial \xi \partial \mu}\right) \\ -E\left(\frac{\partial^2 l}{\partial \xi \partial \sigma}\right) & -E\left(\frac{\partial^2 l}{\partial \xi \partial \mu}\right) & -E\left(\frac{\partial^2 l}{\partial \xi^2}\right) \end{bmatrix}$$

os elementos de M podem ser expressos em termos da função gamma $\Gamma(r) = \int e^{-x}\, x^{r-1}\, dx$ e $\psi(r) = d \log \Gamma(r)/dr$ como:

$$E\left(-\frac{\partial^2 l}{\partial \sigma^2}\right) = \frac{n}{\sigma^2 \xi^2}\left[1 - 2\Gamma(2-\xi) + p\right]$$

$$E\left(-\frac{\partial^2 l}{\partial \sigma \partial \mu}\right) = \frac{n}{\sigma^2 \xi}\left[p - \Gamma(2-\xi)\right]$$

$$E\left(-\frac{\partial^2 l}{\partial \sigma \partial \xi}\right) = \frac{n}{\sigma \xi^2}\left[1 - \gamma - \frac{\{1 - \Gamma(2-\xi)\}}{\xi} - q - \frac{p}{\xi}\right]$$

$$E\left(-\frac{\partial^2 l}{\partial \mu^2}\right) = \frac{n}{\sigma^2}\, p$$

$$E\left(-\frac{\partial^2 l}{\partial \mu \partial \xi}\right) = \frac{n}{\sigma \xi}\left[q + \frac{p}{\xi}\right]$$

$$E\left(-\frac{\partial^2 l}{\partial \xi^2}\right)=\frac{n}{\xi^2}\left[\frac{\pi^2}{6}+\left(1-\gamma-\frac{1}{\xi}\right)^2+\frac{2q}{\xi}+\frac{p}{\xi^2}\right],$$

sendo

$$p=(1-\xi)^2\Gamma(1-2\xi)$$

$$q=\Gamma(2-\xi)\left\{\psi(1-\xi)-\frac{(1-\xi)}{\xi}\right\}$$

e $\gamma=0.5772157$ a constante de Euler.

No procedimento iterativo, fixa-se um valor inicial arbitrário ξ_0 para ξ, e sugere-se como valores iniciais μ_0 e σ_0 para μ e σ, valores tais que $E(X)=\bar{x}$ e $Var(X)=s^2$, sendo $\bar{x}$ a média e s^2 a variância da série de observações. Considerando-se a função densidade de probabilidade dada por (3.1), obtém-se:

$$E(X)=\mu+\frac{\sigma}{\xi}[\Gamma(1-\xi)-1],\ \text{se}\ \ \xi<1,\ \text{e}$$

$$Var(X)=\frac{\sigma^2}{\xi^2}[\Gamma(1-2\xi)-\Gamma^2(1-\xi)],\ \text{se}\ \ \xi<\frac{1}{2}\ ,$$

sendo as seguintes expressões para os valores iniciais:

$$\sigma_0=s\sqrt{\frac{\xi_0^2}{\Gamma(1-2\xi_0)-\Gamma^2(1-\xi_0)}} \tag{4.4}$$

$$\mu_0=\bar{x}-\frac{\Gamma(1-\xi_0)-1}{\xi_0}\sigma_0=\bar{x}-\frac{s}{\xi_0}[\Gamma(1-\xi_0)-1]\sqrt{\frac{\xi_0^2}{\Gamma(1-2\xi_0)-\Gamma^2(1-\xi_0)}} \tag{4.5}$$

Jenkinson sugeriu que se devia usar a matriz informação de Fisher (ou esperança) para amostras completas, mas para amostras censuradas estas esperanças não existem no sentido usual, e foi observado num numero de estudos simulados, que a convergência para

θ é consideravelmente mais rápida, usando a matriz $V\left(\hat{\theta}\right)$ do que a matriz $M\left(\hat{\theta}\right)$. Assim é usual aproximar a matriz $M\left(\hat{\theta}\right)$ por esta nova matriz $V\left(\hat{\theta}\right)$, descrita por:

$$V\left(\hat{\theta}\right)=\begin{bmatrix} -\left(\frac{\partial^2 l}{\partial\sigma^2}\right) & -\left(\frac{\partial^2 l}{\partial\mu\,\partial\sigma}\right) & -\left(\frac{\partial^2 l}{\partial\xi\,\partial\sigma}\right) \\ -\left(\frac{\partial^2 l}{\partial\mu\,\partial\sigma}\right) & -\left(\frac{\partial^2 l}{\partial\mu^2}\right) & -\left(\frac{\partial^2 l}{\partial\xi\,\partial\mu}\right) \\ -\left(\frac{\partial^2 l}{\partial\xi\,\partial\sigma}\right) & -\left(\frac{\partial^2 l}{\partial\xi\,\partial\mu}\right) & -\left(\frac{\partial^2 l}{\partial\xi^2}\right) \end{bmatrix}.$$

Com esta nova matriz, o calculo iterativo de θ envolve rapidez computacional e converge para $|grad\, l| < 10^{-3}$ em menos de 5 iterações.

Para o caso particular da distribuição generalizada de valores extremos com $\xi \to 0$, uma distribuição Gumbel, o logaritmos da função verossimilhança é dados por:

$$l(\mu,\sigma)=\sum_{i=1}^{n}\left\{-\ln\sigma-\left(\frac{x_i-\mu}{\sigma}\right)-\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right\}, \qquad (4.6)$$

e os estimadores de máxima verossimilhança de μ e σ são obtidos pela solução de:

$$\frac{\partial}{\partial\mu}l(\mu,\sigma)_{\mu=\mu_0}=0$$

$$\frac{\partial}{\partial\sigma}l(\mu,\sigma)_{\sigma=\sigma_0}=0,$$

ou seja:

$$-\frac{1}{\hat{\sigma}}\left\{\left[\sum_{i=1}^{n}\exp\left(-\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]-n\right\}=0$$

$$-\frac{1}{\hat{\sigma}}\left\{\left\{\sum_{i=1}^{n}\left[\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)-\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\exp\left(-\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]\right\}-n\right\}=0$$

Mais uma vez, este sistema não possui solução analítica e usou-se o mesmo método iterativo descrito acima para a obtenção da solução numérica, tomando como valores iniciais μ_0 e σ_0 para μ e σ a soluções obtidas através do calculo dos momentos. Para este caso tem-se:

$E(X)=\mu+\gamma\sigma$

$Var(X)=\frac{\pi^2\sigma^2}{6}$, com $\gamma=0.5772157$ a constante de Euler, logo

$\mu_0=\bar{x}-\gamma\frac{\sqrt{6}}{\pi}s\cong\bar{x}-0.45005\ s$

$\sigma_0=\frac{\sqrt{6}}{\pi}s\cong 0.77970\ s$,

que correspondem aos limites de (4.4) e (4.5) quando $\xi_0\to 0$.

Foram aplicados neste trabalho três testes de ajuste a distribuições de valores extremos. O teste de Anderson Darling, o de Kolmogorov Smirnov e o da razão de verossimilhança.

O teste de Anderson Darling é baseado numa função de distribuição empírica A^2,

$$A^2=-\left[\sum_i(2i-1)\{\log z_i+\log(1-z_{n+1-i})\}\right]\Big/n-n$$

com $z=F(x_i)$, em que $F(x_i)$ é a função distribuição cumulativa de uma das distribuições de valores extremos que são:

$$F(x_i)=\exp\left\{-\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

$x_i : 1+\xi(x_i-\mu)/\sigma>0$

$-\infty<\mu<+\infty$

$-\infty<\xi<+\infty \quad e \quad \sigma>0$ (Weibull e Fréchet)

$$F(x_i)=\exp\left[-\exp\left\{-\left(\frac{x_i-\mu}{\sigma}\right)\right\}\right],\qquad -\infty<x_i<+\infty$$

(Gumbel)

com os parâmetros μ, σ e ξ das distribuições, estimados de amostras aleatórias de forma que $F(x_i)$ esteja completamente especificada, e a amostra que deve ser arranjada em ordem ascendente. É um teste que mede a discrepância entre a função empírica de uma amostra e a distribuição teórica. O resultado é comparado com um valor crítico a um determinado nível de confiança. Para valores abaixo do valor crítico, considera-se a hipótese de que a amostra corresponde à distribuição específica.

O teste de Kolmogorov Smirnov é baseado numa função distribuição empírica D:

$$D = \max_{1 \le i \le N} \left| F(x_i) - \frac{i}{N} \right|$$

com $F(x_i)$, uma qualquer distribuição cumulativa expressa em cima. O teste é efetuado da mesma forma que foi descrita em cima e mais uma vez os parâmetros μ, σ e ξ devem ser conhecidos. Os resultados do teste são comparados da mesma forma que em cima. Este teste é mais sensível na parte central da distribuição do que na cauda como é o caso do teste de Anderson Darling.

O teste de taxa de verossimilhança, testa se as observações seguem uma distribuição de valores extremos tipo I, II ou III, supondo que uma distribuição M_0 (no nosso caso a distribuição Gumbel), com parâmetro localização μ e escala σ, é um submodelo de M_1 (distribuições Weibull ou Fréchet), com parâmetro localização μ, escala σ e de forma ξ, sob o constrangimento de que $\xi = 0$. Seja $l_0(M_0)$ e $l_1(M_1)$ o valor maximizado dos logaritmos da verossimilhança das distribuições M_0 e M_1, respectivamente, o teste valida a distribuição M_0, relativamente a M_1, ao nível de significância α, ou seja, rejeita-se M_0 em favor de M_1 se:

$$D = 2\{l_1(M_1) - l_0(M_0)\} > c_\alpha$$

onde c_α é o quantis $(1-\alpha)$ da distribuição assintótica χ^2_k com 1 grau de liberdade. Este teste foi aplicado a todas as distribuições cujo o parâmetro de forma era aproximadamente igual a zero para testar a que distribuição realmente pertencia a amostra.

5. Estimação dos parâmetros da GPD

Para a determinação do limiar, recorre-se à análise gráfica da linearidade de n_u observações que excedem os vários limiares u determinados da própria amostra. Assim, o gráfico de vida residual média usado para a determinação visual de u é construído da seguinte forma:

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_i - u)\right) : u < x_{\max}\right\},$$

em que $x_1, x_2, \ldots, x_{n_u}$ consistem nas observações que excedem u e $x_{\max}$ é o valor mais elevado das observações.

A escolha do limiar implica um balanço entre a bias e a variância, e devido a isto deve-se escolher um limiar que não seja muito elevado, nem muito baixo e para tal existem duas técnicas disponíveis, uma técnica exploratória e outra de contribuição para estabilidade dos parâmetros estimados, baseados no modelo de ajuste para o alcance dos diferentes limiares, descrito em cima. O primeiro método é baseado na média da distribuição generalizada de Pareto. Se Y função densidade de probabilidade de Pareto com parâmetros σ e ξ, então:

$$E(Y) = \frac{\sigma}{1-\xi}, \quad \text{com} \quad \xi < 1.$$

Quando $\xi \geq 1$ a média é infinita.

$$Var(Y) = \frac{\sigma^2}{1-2\xi}, \quad \text{com} \quad \xi < \frac{1}{2}.$$

Seja u_0 o limiar mais baixo de uma série $X_1, X_2, \ldots, X_n$ arbitrária, então:

$$E(Y) = E(X - u_0 \mid X > u_0) = \frac{\sigma_{u_0}}{1-\xi} \quad \text{com} \quad \xi < 1,$$

em que σ_{u_0} é o parâmetro escala correspondente aos excessos do limiar u_0. Mas se a distribuição de Pareto é valida para os excessos de u_0, também é igualmente valida para os excessos do limiar $u > u_0$, sujeito à apropriada variação no parâmetro escala para σ_u. Então, para $u > u_0$,

$$E(X-u \mid X>u)=\frac{\sigma_u}{1-\xi}=\frac{\sigma_{u_0}+\xi u}{1-\xi}$$.

A distribuição generalizada de Pareto é um modelo razoável para os excessos acima do limiar u_0, assim como para um limiar mais elevado u. Os parâmetros de forma das duas distribuições são idênticos. No entanto, o valor do parâmetro de escala para o limiar $u > u_0$ é:

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$,

que varia com u a menos que $\xi = 0$. Esta dificuldade pode ser remediada pela reparametrização do parâmetro de escala, como:

$\sigma^* = \sigma_u - \xi u$ e $\sigma_u = \bar{x}(1-\xi)$, com $\bar{x}$ a média dos excessos para de cada limiar u, e ξ determinado da média e do desvio padrão dos excessos de cada limiar u, e consequentemente as estimativas de ambos σ^* e ξ serão constantes acima de u_0, se u_0 é um limiar válido para os excessos que seguem uma distribuição generalizada de Pareto. Assim, são desenhados os gráficos de σ^* e ξ *versus* u, juntamente com os intervalos de confiança que são obtidos pela matriz variância-covariância V para ξ e para σ^* pelo método delta, usando:

$$Var(\sigma^*) \approx \nabla\sigma^{*T} V \nabla\sigma^*$$, com $$\nabla\sigma^{*T} = \left[\frac{\partial\sigma^*}{\partial\sigma_u}, \frac{\partial\sigma^*}{\partial\xi}\right] = [1, -u]$$.

Determinado o limiar, os parâmetros da distribuição generalizada de Pareto podem ser estimados por máxima verosimilhança. Suponha-se que $y_1, y_2, ..., y_n$ são n excessos de um limiar u. Então, para $\xi \neq 0$ o logaritmo da função verossimilhança é:

$$l(\sigma, \xi) = -k \log\sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right),$$

com $1 + \frac{\xi y_i}{\sigma} > 0$ para $i = 1,2,...,k$, de outra forma $l(\sigma, \xi) = -\infty$. No caso em que $\xi \to 0$, o logaritmo da função verossimilhança é:

$$l(\sigma) = -k \log\sigma - \left(\frac{1}{\sigma}\right) \sum_{i=1}^{k} y_i$$

Procedendo-se da mesma forma, como para a GEV e considerando como valores iniciais os valores determinados em cima pela análise exploratória, determinam-se os

parâmetros da distribuição generalizada de Pareto, assim como o valor de máxima verossimilhança.

Todo procedimento computacional para determinação dos parâmetros foi realizado na plataforma R 2.8.1, *Software* livre disponível em http://www.R-project.org.

6. Resultados e discussão

Para realização dos ajustes a partir da GEV, foram considerados como extremos os valores máximos encontrados em cada trimestre anual. Os ajustes para as capitais em estudo foram bastante satisfatórios, o que pode ser verificado através dos resultados do teste de Kolmogorov-Smirnov, analisado ao nível de significância $\alpha = 0,05$, (Tabela 1) e dos gráficos de Probabilidade, QQ-*plot* e densidade (Figura 2).

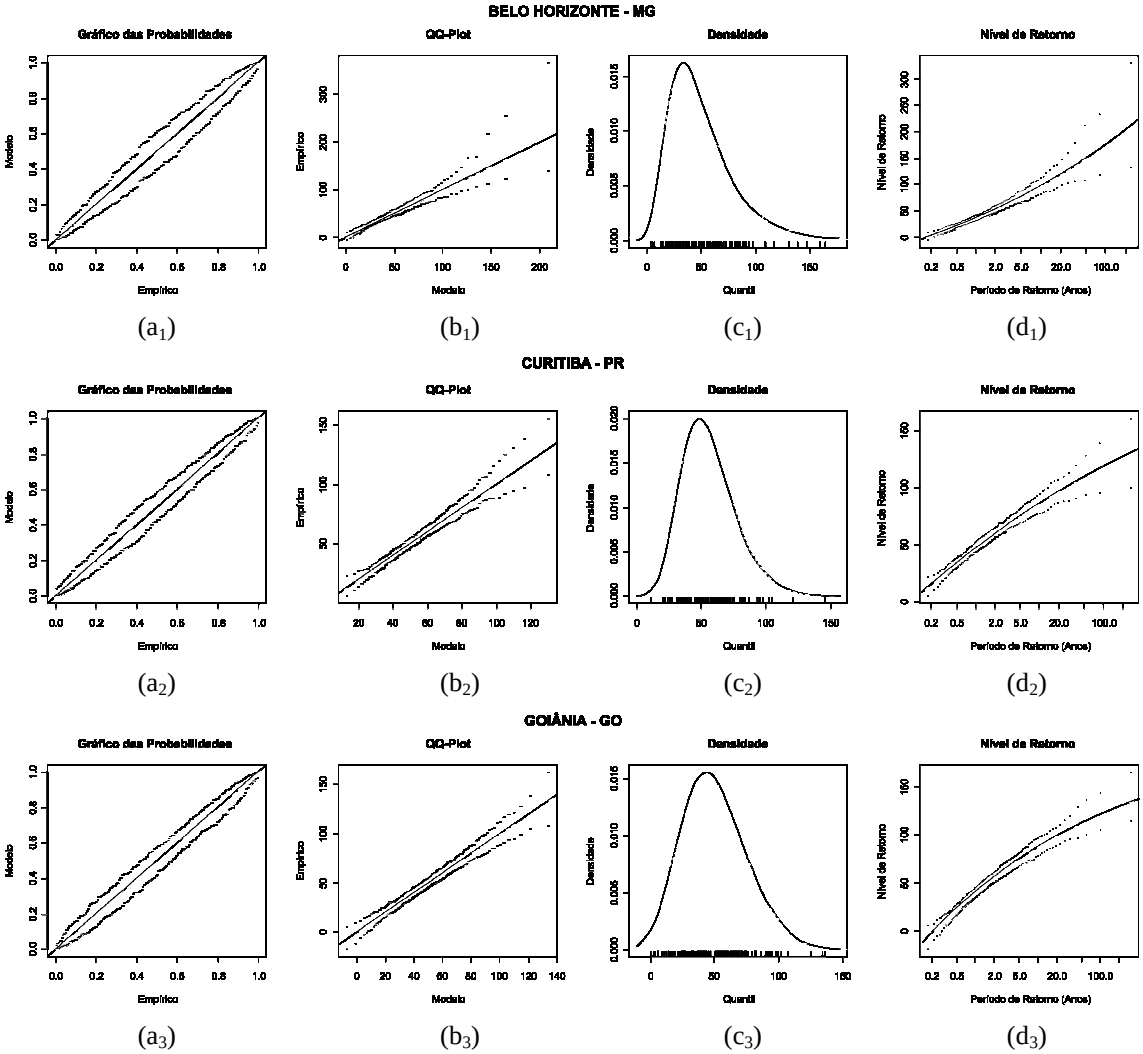
Tabela 1 - Teste de Kolmogorov-Smirnov e estimativa dos parâmetros para o ajuste da GEV.

Capital	Valor-p*	Estimativas dos Parâmetros			Intervalo de Confiança de 90% para ξ
		Locação (μ)	Escala (σ)	Forma (ξ)	
Belo Horizonte – MG	0,4820	36,00	22,8202	0,1146	[-0,0201; 0,2492]
Curitiba – PR	0,4478	46,97	18,4882	-0,0770	[-0,1620; 0,0080]
Goiânia – GO	0,5393	39,91	23,9193	-0,1380	[-0,2193;-0,0567]
Manaus – AM	0,8519	56,86	27,8097	-0,0499	[-0,1466; 0,0467]
Recife – PE	0,3271	53,41	33,5958	0,0444	[-0,0524; 0,1412]

Dados oriundos do INMET
*** Valor-p referente ao teste de Kolmogorov Smirnov.**

Em Belo Horizonte-MG, no mês de janeiro de 2008, o INMET registrou em um único dia uma precipitação de 94,2 mm, o que causou alagamentos e queda de árvores. Situações como essas, de acordo com o modelo ajustado na Figura 2.d1, podem ocorrer a cada 11 anos, aproximadamente. Observa-se que em um período correspondente a cada 47 anos, espera-se que ocorra uma precipitação de pelo menos 147 mm em Belo Horizonte. No ano de 1973, na capital do Paraná, em 24h choveu cerca de 122 mm, segundo o Centro de Previsão de Tempo e Estudos Climáticos - CPTEC. Valores como esse pode ocorrer a cada 25 anos, aproximadamente (Figura 2.d2), o que pôde ser comprovado no ano de 1999, onde

houve uma precipitação superior a esse, cerca de 140 mm. O CPTEC registrou, em março de 2008, uma precipitação de 89 mm em 24h ocorrida em Goiânia-GO. Essa chuva, com duração de duas horas, deixou bairros alagados e cerca de 50.000 pessoas sem energia elétrica. De acordo com o modelo ajustado através da GEV (Figura 2d3), precipitações como essa é comum na cidade, ocorrendo a cada oito anos. Situações muito extremas, como a ocorrida em 1967, onde choveu cerca de 180 mm em um único dia, podem voltar a ocorrer em Manaus-PA a cada 240 anos, aproximadamente. Situações mais frequentes são as chuvas com 33.6 mm, as quais, segundo o modelo GEV, ocorrem anualmente. Em Recife-PE, precipitações acima de 100 mm podem ocorrer a cada 15 anos.



(a_1) (b_1) (c_1) (d_1)

(a_2) (b_2) (c_2) (d_2)

(a_3) (b_3) (c_3) (d_3)

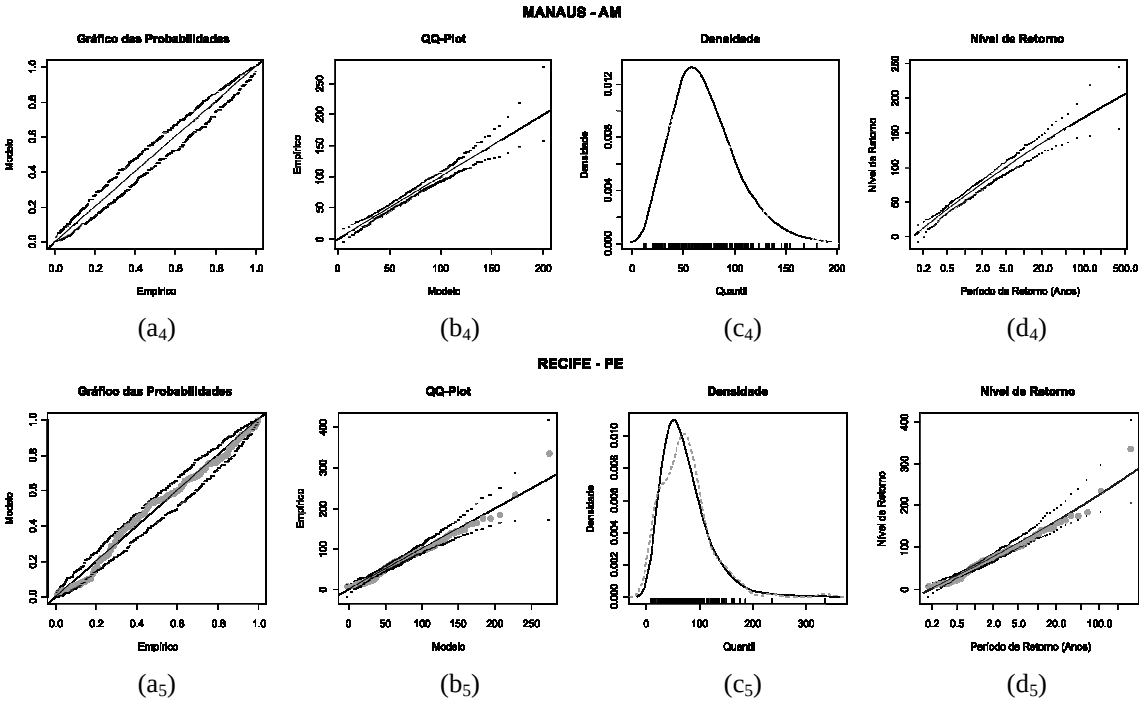


Figura 2: Gráficos de diagnósticos através da GEV para precipitação (1951-2005), Dados oriundos do INMET.

Os ajustes realizados para GPD foram todos satisfatórios. O teste de Kolmogorov-Smirnov não nos leva a rejeitar, ao nível de significância $\alpha = 0,05$, a hipótese de nulidade, a qual afirma que os dados são provenientes de uma GPD. Os gráficos da Figura 3.a, Figura 3.b e Figura 3.c são indicadores da qualidade do ajuste. A Tabela 2 informa o valor-*p* calculado para o teste de Kolmogorov-Smirnov e as estimativas, calculadas a partir do método da máxima verossimilhança, dos parâmetros associados à GPD.

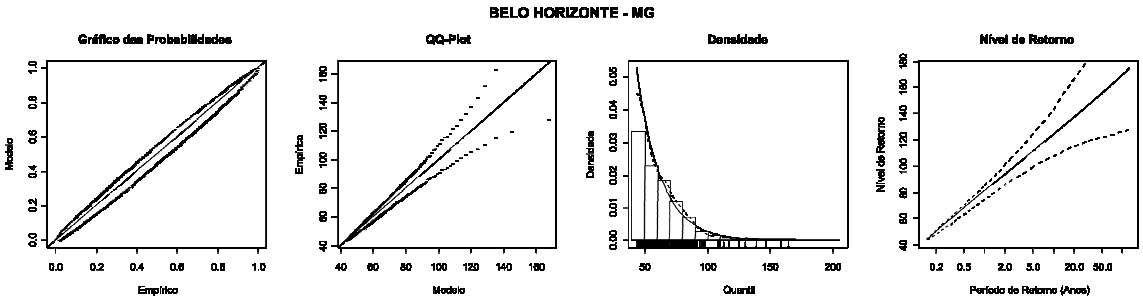
Tabela 2: Teste de *Kolmogorov-Smirnov* e estimativa dos parâmetros para o ajuste da GPD.

Capital	**Valor-p***	**Estimativas dos Parâmetros**			**Intervalo de Confiança de 90% para ξ**
		Locação (μ)	Escala (σ)	Forma (ξ)	
Belo Horizonte – MG	0,8940	43,30	18,9464	0,0090	[-0,0866; 0,1046]
Curitiba – PR	0,3628	37,60	16,8125	-0,0260	[-0,1214; 0,0694]
Goiânia – GO	0,8735	40,88	15,9356	-0,0107	[-0,0974; 0,0759]
Manaus – AM	0,5659	48,50	21,2983	0,0451	[-0,0533; 0,1435]
Recife – PE	0,4459	54,90	23,4727	0,0972	[0,0044; 0,1900]

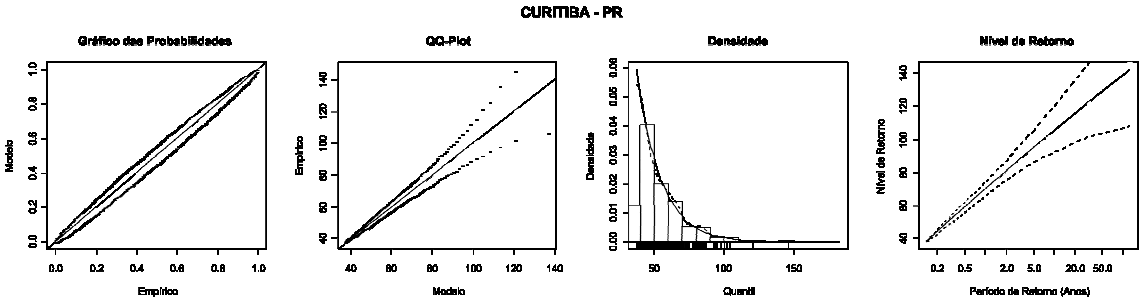
Dados oriundos do INMET

*** Valor-p referente ao teste de *Kolmogorov Smirnov*.**

Pode-se verificar através da Figura 3d1 que um nível de precipitação próximo a 94,2, como o ocorrido em janeiro de 2008, poderá acontecer novamente na cidade de Belo Horizonte em, aproximadamente, 18 anos de acordo com o modelo ajustado com a GPD. Através desse modelo, prever que precipitações acima de 110 mm podem ocorrer na capital a cada 35 anos. Chuvas como a de 1973 em Curitiba, através do modelo ajustado pela GPD, podem ocorrer a cada 33 anos, aproximadamente. Em janeiro de 2007, houve uma precipitação semelhante, de 107 mm, a qual deixou centenas de casas destelhadas e destruídas. Em Curitiba, precipitações acima de 90 mm podem ocorrer a cada quatro anos. Precipitações acima de 89 mm podem ocorrer em Goiânia-GO a cada três anos, aproximadamente. O que se comprova através da série histórica e fatos recentes, onde houve uma precipitação superior a esta em 2005 e tornou-se a repetir em 2008, onde a chuva deixou bairros alagados. Em Manaus, precipitações acima de 100 mm podem ocorrer a cada um ano e meio. Enquanto precipitações acima de 180 mm preveem que aconteça a cada 154 anos, aproximadamente. Chuvas acima de 51 mm já causam alagamentos e transtornos no trânsito na cidade do Recife. Precipitações como esta podem voltar a acontecer a cada dois meses, aproximadamente, situação esta comprovada no ano de 2007. Chuvas acima de 150 mm ocorrem, aproximadamente, a cada seis anos.



(a_1) (b_1) (c_1) (d_1)



(a_2) (b_2) (c_2) (d_2)

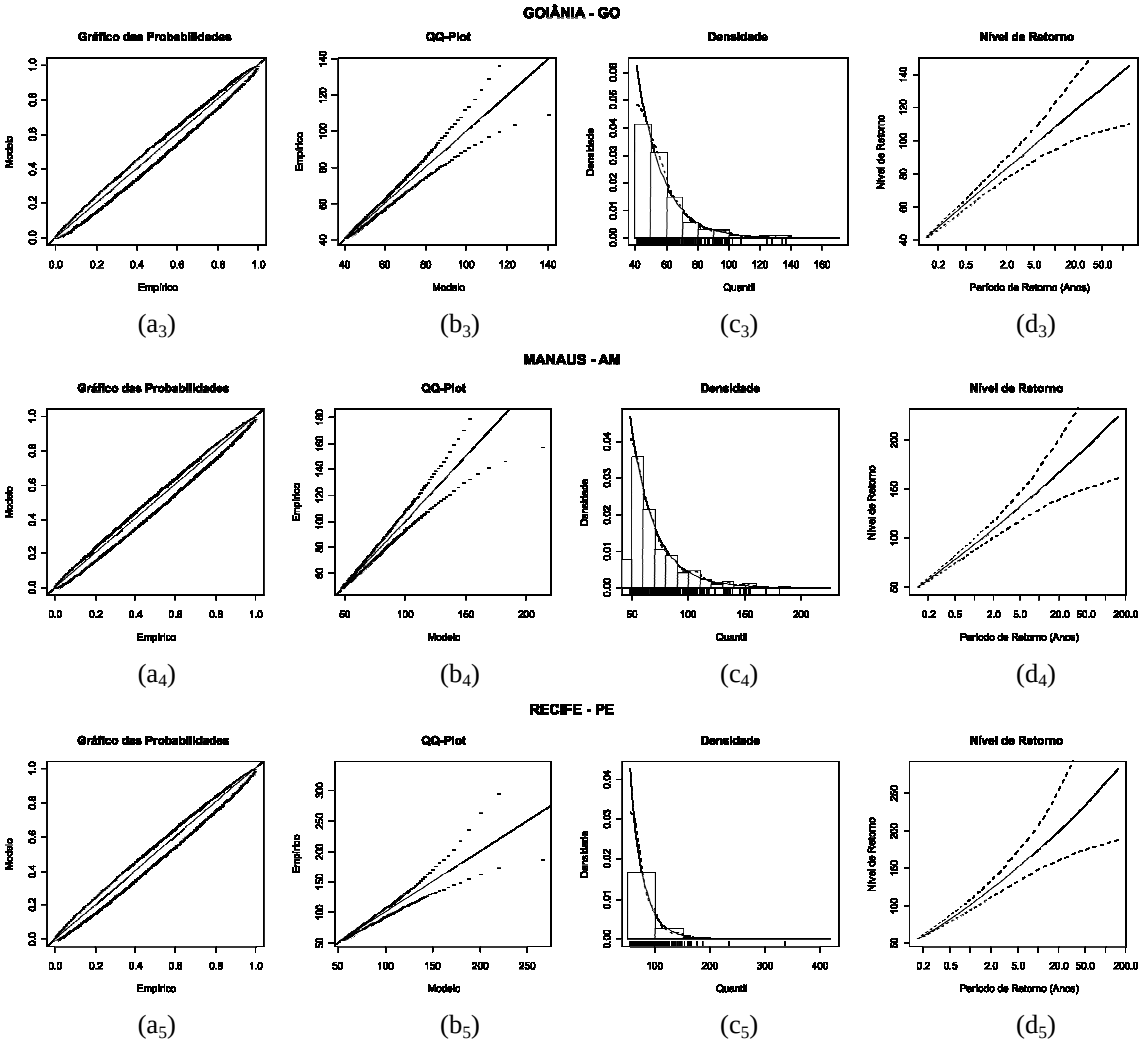


Figura 3: Gráficos de diagnósticos através da GPD para precipitação (1951-2005) Dados oriundos do INMET.

7. Considerações finais

As estimativas alcançadas pela GEV e GPD foram satisfeitas, porém o grande problema de se trabalhar com a GEV é o fato de ela não diferir situações em que ocorrem extremos de máximo em um período, uma vez que o máximo em um período pode não ser, necessariamente, um evento extremo.

A qualidade do ajuste da GEV e a da GPD foram avaliadas graficamente através dos gráficos de probabilidade da distribuição empírica (PP-Plot) e quantil-qualtil (QQ-Plot), e pela aplicação dos testes de Kolmogorov-Smirnov. Os resultados obtidos indicam que o ajuste obtido no ajuste GPD se mostrou visualmente mais adequado.

Através do intervalo de confiança definido para o parâmetro de forma, pode-se observar que as capitais de Belo Horizonte, Curitiba, Manaus e Recife convergem para a distribuição *Gumbel* e Goiânia para a *Weibull*, quando modeladas através da GEV. As cidades de Belo Horizonte, Curitiba, Goiânia e Manaus convergem para a exponencial e Recife para a Pareto, sendo estas modeladas pela GPD.

8. Trabalhos futuros

1 - A partir da distribuição generalizada de valores extremos, pretende-se estimar como, e com qual precisão, as durações de eventos máximos e mínimos acontecem.

2 - Realizar inferência sobre os parâmetros da distribuição generalizada de valores extremos através do Método dos Momentos e comparar com as estimativas obtidas pelo Método de Máxima Verossimilhança.

3 - Explorar a relação multiparamétrica existente entre as distribuições separadas do máximo e mínimo.

Referências bibliográficas

BALKEMA, A. A. e de HAAN, L. (1974). Residual life time at great age. Annals of Probability, 2, 792-804.

Climatologia de precipitações e temperaturas. Disponível em: http://www.cptec.inpe.br/products/climanalise/cliesp10a/chuesp.html.

COLES, S. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer Series in Statistics. Springers Series in Statistics, London.

GALLO, N,F., (2007). Análise comparativa de incertezas em métodos para estimação de freqüências de vazões máximas diárias com incorporação de variação climática em bacias do médio rio Uruguai. Porto Alegre:UFRS, Dissertação.

HOSKING, J. e WALLIS, J. (1987). Parameter and quantile estimation for the generalised Pareto distribution. Technometrics. 29(3), 339–349.

JENKINSON, A. F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. Quarterly Journal of the Royal Meteorological Society, London, 81, 158-171, 1955.

LUCIO, P. S. (2004a). Geostatistical Assessment of HadCM3 Simulations via NCEP Reanalyses over Europe. Atmospheric Science Letters, 5, 118-133.

LUCIO, P. S. (2004b). Assessing HadCM3 Simulations from NCEP Reanalyses over Europe: Diagnostics of block-seasonal extreme temperature's regimes. Global and Planetary Change, 44 (1-4), 39-57.

FISHER, R. A., e TIPPETT, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest members of a sample. Proc. Cambridge Philos. Soc., 24, 180–190.

GNEDENKO, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Annals of Mathematics, 44, 423-453.

GUMBEL, E. J., (1958). Statistics of Extremes. Columbia. University Press, 375 pp.

PICKANDS III, J. (1975). Statistical inference using extreme order statistics. Annals of Statistics. 3, 119-131.

RIBATET, M. (2007). POT: Modelling Peaks over a Threshold. The Newsletter of the R Project. 7-1: 34-36.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

WILKS, D. S. (1995). Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, 467 pp.

Agradecimentos

Os autores agradecem ao Instituto Nacional de Meteorologia - INMET por ter, gentilmente, cedido os dados utilizados nesta pesquisa.

Abstract

Motivated by the trend in extreme weather - in connection with the alleged climate change - the purpose of this study is to present, through the theory of extreme values, models that show the impact of rare events in society and ecosystems for prevention and mitigation proposals. Thus, structural and non-structural measures can be taken to reduce the impacts. The data of this study are the daily rainfall of the years 1951 to 2005 of five Brazilian capitals. Given the occurrence of extreme events, we tried to verify the appropriateness of the widespread distribution of extreme values and generalized Pareto distribution to estimate the precipitation will occur in a given period. Through adjustments via GEV and GPD, we can conclude that the levels of return of 100 mm are more frequent in Recife. The comparison of results for both methods opens a discussion about the reliability of the estimates provided.

Key-Words: POT, Return Level, Return Period, Precipitation.

Elicitação da distribuição *a priori* para o risco de fratura em pacientes com osteoporose

Fernando A. Moala[5]

Resumo

No contexto de análise estatística, elicitação é o processo de se extrair o conhecimento de um especialista sobre alguma quantidade desconhecida na forma de uma distribuição de probabilidade. Essa distribuição de probabilidade é frequentemente usada como distribuição a *priori* em uma análise bayesiana e então a informação elicitada do especialista pode ser usada para complementar a informação dos dados observados.

Oakley e O'Hagan (2007) desenvolveram um procedimento bayesiano não-paramétrico para elicitação da distribuição *a priori* considerando uma única variável de interesse. Em vez de assumir uma forma particular paramétrica para a função densidade, ela é tratada como uma função desconhecida, e a inferência é feita sobre esta função baseada apenas nas especificações probabilísticas fornecidas pelo especialista.

O principal objetivo desse artigo é apresentar uma implementação prática do método bayesiano de elicitação proposto por Oakley e O'Hagan (2007) para construção de uma distribuição *a priori* f(θ) do risco de fratura em pacientes que sofrem de osteoporose sob um determinado tratamento. Este artigo também tem o propósito de divulgar e provocar interesse pela Elicitação, de forma que mais estatísticos possam aplicá-la em muitos problemas

[5] Dep. de Matemática - FCT - UNESP - 19060-900 - Presidente Prudente - SP - BR. E-mail: femoala@fct.unesp.br

práticos. A teoria do método de elicitação utilizado neste artigo é também apresentada de forma sintética e acessível, a fim de facilitar a sua compreensão e implementação.

Palavras-chave: Elicitação, especialista, processo gaussiano, distribuição *a priori*, *a posteriori*, método bayesiano, osteoporose.

1. Introdução

A capacidade para incorporar a informação *a priori* em uma pesquisa científica é um importante benefício da inferência bayesiana. Ela permite o uso de toda a informação disponível além dos dados, e consequentemente chegar-se a conclusões mais realísticas. A Elicitação de *prioris* é um tópico da inferência bayesiana e tem sido utilizada em muitas áreas aplicadas do conhecimento, principalmente em situações nas quais os dados experimentais não são tão numerosos devido à dificuldade ou custo para obtê-los. Isto realmente ocorre na apresentação de um caso de custo-eficácia de novas drogas, em que necessitamos de vários anos de experimentação para se obter mais dados, de mais pacientes, acarretando uma demora na aprovação e produção do medicamento. Há, porém, uma informação *a priori* sobre a droga proveniente da indústria farmacêutica que a desenvolve e que poderia ser utilizada na análise estatística. Através da Elicitação, o estatístico pode incorporar formalmente essa informação do especialista na forma de uma distribuição de probabilidade.

Freedman e Spiegelhalter (1983), Carlin *et al*. (1993), Chaloner *et al.* (1993) e Spiegelhalter *et al*. (1994) discutem a importância da quantificação da opinião *a priori* e apresentam regras e métodos para a utilização da elicitação numa pesquisa clínica.

Há vários métodos diferentes de se elicitar distribuições *a priori* propostos na literatura; a maioria deles requer a elicitação de momentos de segunda ordem e/ou a elicitação de hiperparâmetros o que torna difícil sua aplicação prática.

Oakley e O'Hagan (2007) desenvolveram um procedimento bayesiano não-paramétrico para elicitação do conhecimento do especialista na forma de uma distribuição *a priori* para uma única variável. A ideia é pensar em elicitar uma distribuição *a priori* como a

estimação de parâmetro na inferência bayesiana. O estatístico deseja fazer inferências sobre uma função desconhecida $f(\theta)$, a densidade *a priori* do especialista. Ele formula primeiramente suas próprias convicções *a priori* sobre $f(\theta)$ e pede então ao especialista que forneça as probabilidades sobre θ às quais são pensadas como dados sobre $f(\theta)$. Atualiza-se então suas convicções sobre $f(\theta)$ levando em conta estes dados.

Neste trabalho, nosso interesse será aplicar o procedimento de elicitação proposto por Oakley e O'Hagan (2007) para estimação da distribuição do risco de fraturas em pacientes que sofrem de osteoporose. Para um paciente recebendo um particular tratamento, há risco incerto de fratura do osso e então nosso interesse será a construção de uma distribuição *a priori* para este risco sob um determinado tratamento. A informação sobre o risco para qualquer tratamento virá das informações do clínico. Este trabalho é então baseado em nossa experiência prática, obtida ao aplicarmos nosso procedimento de elicitação em uma situação real e expor os desafios envolvidos para implementá-la.

Este artigo é estruturado da seguinte forma. Na próxima seção, é apresentado o enfoque teórico detalhado do procedimento de elicitação sob estudo nesse artigo. A seção 3 descreve os resultados do processo de elicitação do especialista e os dados elicitados. Na seção 4, estimamos a densidade esperada do estatístico, e então sumarizamos o que aprendemos do processo de elicitação, os seus resultados e alguns dos problemas observados na implementação deste processo. A seção 5 traz um *feedback* do especialista, para verificar se ele concorda com as estimações propostas pelo estatístico. Finalmente, a seção 6 contém uma discussão dos resultados da elicitação e as implicações de sua implementação.

2. O Método de Elicitação

Nesta seção, apresentamos uma descrição do método de elicitação proposto por Oakley e O'Hagan (2007). Denotando a densidade *a priori* do especialista por $f(\theta)$, precisamos considerar primeiramente as convicções *a priori* do estatístico sobre $f(\theta)$.

O método pressupõe uma função densidade *f*(θ) que seja lisa, infinitamente diferenciável, cuja forma paramétrica é desconhecida.

2.1 Uma distribuição *a priori* para *f(θ)*

É assumido que as convicções *a priori* do estatístico sobre *f(θ)* podem ser representadas por um processo gaussiano, em que para qualquer coleção de *n* valores $\theta_1, \theta_2, \cdots, \theta_n$ tem-se que os valores correspondentes da função densidade $f(\theta_1), f(\theta_2), \cdots, f(\theta_n)$ seguem uma distribuição normal multivariada. Uma representação apropriada para *a priori* de *f(θ)* é então dada por

$$f(\cdot)|\alpha \sim GP\big(g(\cdot), C(\cdot,\cdot)\big), \tag{1}$$

cujo α é um vetor de hiperparâmetros.

Processos gaussianos *a priori* para funções foram propostos em várias diferentes situações, inclusive regressão (O'Hagan (1978) e Neal (1999)), classificação (Neal, 1999) e análise numérica (O'Hagan, 1992).

Para o processo gaussiano, é preciso especificar uma forma paramétrica para a esperança *a priori* de *f(θ)* e a covariância *a priori* entre *f(θ)* e *f*(ϕ).

Uma vez que sabemos que *f(θ)* é uma função densidade, é natural escolher *a priori* uma função densidade paramétrica *g(θ)* para *E{f(θ)|α}* de forma que podemos então pensar em modelar a razão $h(\theta) = \dfrac{f(\theta)}{g(\theta)}$ como um processo gaussiano com uma média constante. 1. Consideraremos o caso em que o estatístico acredita que uma função densidade normal é uma escolha apropriada para a função média *g(θ).* Não será necessário especificar a média e a variância em *g(θ),* denotados por *m* e *v*; estes serão tratados como desconhecidos. Portanto, tem-se:

$$E\{f(\theta)|\alpha\} = g(\theta) = \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2}\left(\frac{\theta - m}{v}\right)^2\right\}. \tag{2}$$

A correlação entre quaisquer dois pontos *h(θ)* e *h*(ϕ) é dada por alguma função *c*(θ , ϕ), implicando

$$Cov\{h(\theta), h(\phi) | \alpha\} = \sigma^2 c(\theta, \phi) . \quad (3)$$

O hiperparâmetro *σ2* especifica quão próximo a verdadeira função de densidade estará de sua média *a priori*, e assim governa quão bem ela se aproxima da função média *g(θ).*

Em geral, a função *c*(θ, ϕ) deve assegurar que a matriz de variância e covariância *a priori* de qualquer conjunto de observações de *f(θ)* (ou funcionais de *f(θ))* seja semi-definida positiva. Aqui é escolhida a função

$$c(\theta, \phi) = \exp\left\{-\frac{1}{2b}(\theta - \phi)^2\right\} . \quad (4)$$

Isto será visto também como uma escolha matematicamente conveniente, e implica que f(θ) seja infinitamente diferenciável com probabilidade 1. O hiperparâmetro *b* controla a suavidade da verdadeira densidade. Se *b* for grande, então dois pontos *f(θ)* e *f*(ϕ) estarão altamente correlacionados, mesmo se *θ* e ϕ estiverem muito distantes um do outro.

A função de covariância entre *f(θ)* e *f*(ϕ) pode então ser escrita como:

$$Cov\{f(\theta), f(\phi) | \alpha\} = C(\theta, \phi) = \sigma^2 g(\theta) g(\phi) c(\theta, \phi) . \quad (5)$$

Esta formulação foi dada por Kennedy e O'Hagan (1996), que estavam interessados na quadratura para funções densidades computacionalmente complexas de se avaliar.

O estatístico descreve suas convicções *a priori* sobre a densidade desconhecida *f(θ)* por um modelo que a permita desviar de uma densidade paramétrica conhecida *g(θ),* porém, o modelo é não-paramétrico e permite a verdadeira *f(θ)* ter qualquer forma.

Os hiperparâmetros deste modelo são então $\alpha = (m, v, b, \sigma^2)$ e eles serão considerados valores desconhecidos.

2.2 Dados elicitados do especialista

Claramente, não é razoável esperar que o especialista seja capaz de estabelecer valores da sua densidade *a priori f(θ)* para vários valores de *θ*. Momentos de segunda ordem ou mais de uma distribuição (excluindo possivelmente a média) também não deveriam ser pedidos ao especialista (*Kadane* e *Wolfson*, 1998). Seria mais razoável pedir probabilidades como $P\{\theta \le x\}$, ou os percentis. Assim, *n* valores $x_1, x_2, \ldots, x_n$ são escolhidos, e o especialista fornece suas probabilidades $P_{x_1}, P_{x_2}, \ldots, P_{x_n}$ para obter os dados $D = \{P_{x_1}, P_{x_2}, \ldots, P_{x_n}\}$ com

$$P_{x_1} = P\{\theta \le x_1\} = \int_{-\infty}^{x_1} f(\theta) d\theta \quad . \tag{6}$$

É assumido implicitamente que a observação $\int f(\theta) d\theta = 1$ é dada.

Uma vez que *f(θ)* é distribuído sob um processo gaussiano, a distribuição de qualquer funcional linear de *f(θ)* (que inclui momentos e outras quantidades como as probabilidades que definem determinados quantis) é uma normal.

Condicional em $\alpha = (m, v, b, \sigma^2)$, o vetor $D = \{P_{x_1}, P_{x_2}, \ldots, P_{x_n}\}$ é normalmente distribuído (por definição de processo gaussiano), isto é,

$$D|\alpha \sim N_n(H, \sigma^2 A) \tag{7}$$

com vetor de médias *H* tendo elementos $E(P_x) = \Phi\left(\frac{x-m}{\sqrt{v}}\right)$ cujo Φ representa a acumulada da normal padrão, e matriz de variância-covariância $\sigma^2 A$ com elementos dados por

$$\operatorname{cov}\left(P_x, P_y | \alpha\right) = \sigma^2 \sqrt{\frac{b}{b+2}} \int_{P_x} \int_{P_y} N_2\left((\theta, \phi) | (m, m), S\right) d\theta d\phi \tag{8}$$

e

$$S^{-1} = \frac{1}{v} \begin{bmatrix} \frac{1+b}{b} & -\frac{1}{b} \\ -\frac{1}{b} & \frac{1+b}{b} \end{bmatrix}. \tag{9}$$

Daí, a função de verossimilhança *L(α | D)* é dada por

$$L(\alpha \mid D) = \frac{|A|^{-\frac{1}{2}}}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(D-H)^t A^{-1}(D-H)\right\}. \quad (10)$$

2.3 *Posteriori* como atualização da *priori*

Como visto na seção 2.2, os dados entrarão na forma de quantis da distribuição e momentos simples. Condicionais em α, a distribuição *a posteriori* de *f(θ)* pode ser derivada analiticamente, usando as propriedades da distribuição normal multivariada.

Primeiramente, precisamos determinar a covariância *a priori* entre *f(θ)* e *D*, isto é,

$$\operatorname{cov}\left(f(\theta), P_x|\alpha\right) = \operatorname{cov}\left(f(\theta), \int_{-\infty}^{x} f(\phi)d\phi|\alpha\right) = \int_{-\infty}^{x} \operatorname{cov}\left(f(\theta), f(\phi)|\alpha\right)d\phi =$$

$$= \sigma^2 g(\theta)\sqrt{\frac{b}{b+1}} \exp\left\{\frac{(\theta-m)^2}{v(b+1)}\right\} \Phi\left[\left(x - \frac{\theta+mb}{b+1}\right)\sqrt{\frac{1+b}{vb}}\right]. \quad (11)$$

Agora, como *f(θ)* e todos os elementos em *D* que ainda vamos observar são normalmente distribuídos então, a distribuição conjunta de *f(θ)* e *D* é dada por

$$\begin{pmatrix} D \\ f(\theta) \end{pmatrix} \sim N\left[\begin{pmatrix} H \\ g(\theta) \end{pmatrix}, \begin{pmatrix} \sigma^2 A & \sigma^2 t(\theta) \\ \sigma^2 t(\theta)^T & C(\theta,\theta) \end{pmatrix}\right]. \quad (12)$$

e *t(θ)* é o vetor de covariância *a priori* entre *f(θ)* e os elementos de *D*.

Segue, então, imediatamente das fórmulas usuais para média e variância das distribuições condicionais da distribuição normal multivariada que a distribuição *a posteriori* $f(\cdot)|D,\alpha$ é um processo gaussiano com:

$$E\left\{f(\theta)|D,\alpha\right\} = g(\theta) + t(\theta)^t A^{-1}(D-H) \quad (13)$$

e

$$\operatorname{cov}\left(f(\theta), f(\phi)|D,\alpha\right) = \sigma^2\left[g(\theta)g(\phi)c(\theta,\phi) - t(\theta)^t A^{-1} t(\phi)\right]. \quad (14)$$

2.4 Estimadores *a posteriori* para os hiperparâmetros

Considerando que estamos usando um procedimento Bayesiano, precisamos completar o modelo hierárquico especificando uma distribuição *a priori* π(α) que refletirá as convicções *a priori* do analista sobre os hiperparâmetros desconhecidos α = (*m,v, b, σ2*) envolvidos no modelo. *Prioris* não-informativas são usadas para m, v e σ^2 na forma

$$\pi(m, v, \sigma^2) \propto \frac{1}{v\sigma^2}, \tag{15}$$

porém, uma *priori* imprópria para *b* levaria a uma *posteriori* imprópria. Assim, a incerteza sobre *log(b)* é representada por uma distribuição normal N(0,65; 0,252). Oakley e O'Hagan (2006) especificam essa normal através de gráficos das gerações da distribuição *a priori* da razão *f(θ)/g(θ)*, para diferentes valores de *b*. Uma vez que acreditamos que *f(θ)* seja uma função lisa, não esperamos que esta razão flutue muito. Além disso, acreditamos que *f(θ)* pode diferir substancialmente de uma densidade normal, e assim não devemos por muito peso *a priori* para valores de *b* que signifiquem uma razão quase constante.

A distribuição *a priori* de $\alpha = (m, v, b, \sigma^2)$ tem a forma:

$$\pi(\alpha) \propto \frac{1}{v\sigma^2}\pi(b). \tag{16}$$

em que π(b) é *a priori* informativa para b.

A distribuição a *posteriori* é, então, usando (10) e (16), dada por:

$$p(\alpha \mid D) \propto \pi(b)\frac{|A|^{-\frac{1}{2}}}{v\sigma^{n+2}}\exp\left\{-\frac{1}{2\sigma^2}(D-H)^t A^{-1}(D-H)\right\}. \tag{17}$$

O condicionamento em σ^2 pode ser removido como

$$p(m, v, b \mid D) \propto \pi(b)\frac{|A|^{-\frac{1}{2}}}{v}\int_0^\infty \frac{1}{\sigma^{n+2}}\exp\left\{-\frac{1}{2\sigma^2}(D-H)^t A^{-1}(D-H)\right\}d\sigma^2 \tag{18}$$

e notando que $\int_0^\infty x^{-(p+1)}e^{-ax^{-q}}dx = \frac{1}{q}a^{-p/q}\Gamma\left(\frac{p}{q}\right)$ obtemos

$$p(m, v, b \mid D) \propto \frac{1}{v}(\hat{\sigma}^2)^{-n/2}|A|^{-\frac{1}{2}}\pi(b) \tag{19}$$

onde

$$\hat{\sigma}^2 = \frac{1}{n-2}(D-H)^t A^{-1}(D-H) \quad (20)$$

.

Como a distribuição *a posteriori* conjunta é complicada e com isso não podemos obter analiticamente as distribuições marginais *a posteriori*, precisaremos empregar métodos numéricos de aproximação para obtê-las. Optamos aqui pelo uso do método de Monte Carlo via cadeias de Markov (MCMC) e implementamos o algoritmo de Metrópole-Hastings (M-H) como descrito a seguir. As seguintes distribuições foram escolhidas para o algoritmo:

$$\begin{aligned} m_t \,|\, m_{t-1} &\sim N(m_{t-1}; 0{,}01) \\ \log v_t \,|\, m_{t-1}, m_t, v_{t-1} &\sim N(\log v_{t-1}; 0{,}1(1+|\, m_t - m_{t-1} \,|\, / 0{,}2)) \\ \log b_t \,|\, b_{t-1} &\sim N(\log b_{t-1}; 0{,}01) \end{aligned} \quad (21)$$

.

Estimamos a moda de *m, v e b* para, então, começar o procedimento iterativo com valores a uma distância razoável da moda.

Finalmente, para cada valor de *m, v e b* gerado da *posteriori* $p(m,v,b \mid D)$ via MCMC e $\hat{\sigma}^2$ como função destes podemos amostrar uma função densidade $f(\cdot)$ da *posteriori* $p(f(\cdot)|m,v,b,D)$, para um número finito de valores de θ.

Após a convergência do algoritmo, temos uma amostra das funções da distribuição *a posteriori* $p(f(\cdot)|D)$, e, portanto, estimativas e limites para $f(\cdot)$ podem ser obtidas.

Usando o estimador para $f(\cdot)$ dado na equação (13), podemos plotar a função média e os percentis amostrais da distribuição da função densidade.

3. Quantificando a opinião do especialista sobre a eficácia do tratamento de osteoporose

Discutimos agora uma aplicação prática do procedimento de elicitação apresentado neste artigo com a estimação da distribuição do risco de fraturas em pacientes que sofrem de osteoporose. Nesta seção, descrevemos como o processo de elicitação para quantificar a opinião do especialista foi conduzido.

A informação sobre o risco de osteoporose sob um dado tratamento vem da opinião de um pesquisador em modelagem de tratamentos para osteoporose na *School of Health and Related Research (*ScHARR*)* da Universidade de Sheffield. O processo de elicitação foi realizado entre agosto-2005 e janeiro-2006 através de vários encontros entre o especialista e o estatístico.

Embora, em muitos casos, seja necessária uma fase de treinamento para permitir ao especialista se familiarizar com o formato do processo de elicitação que será implementado, isto não foi necessário aqui, pois este já havia trabalhado com um processo de elicitação semelhante anteriormente. Houve, contudo, uma reunião inicial para dar-lhe uma exposição do propósito da pesquisa, detalhes sobre as tarefas de avaliação que lhe seriam pedidas que executasse como avaliações sobre seu conhecimento do tratamento de osteoporose em termos probabilísticos, além de uma breve revisão do significado dos percentis de uma distribuição de probabilidade.

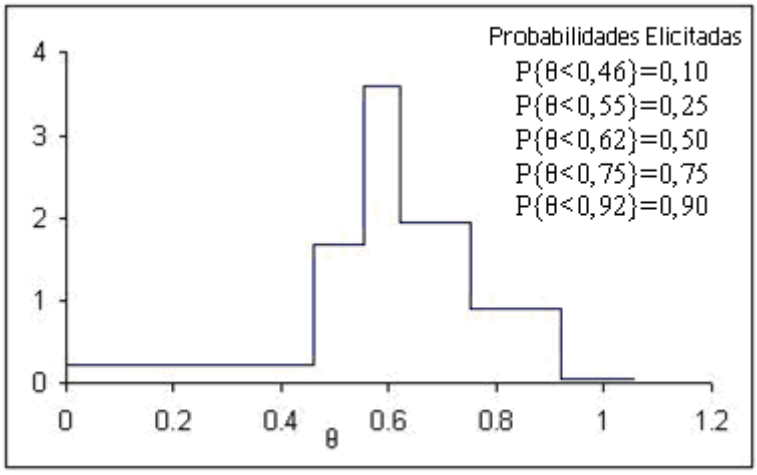
As pesquisas têm mostrado não ser fácil responder perguntas como “qual é a probabilidade da variável assumir um determinado valor”, o mais conveniente é perguntar “qual o valor que a variável poderia assumir para uma determinada probabilidade” (Qing, 1998). Assim, no processo de elicitação é pedido ao especialista um domínio de valores dentro do qual ele acredita θ pertencer com uma probabilidade especificada, efetivamente pedir um intervalo de cobertura especificada, por exemplo, se a probabilidade dada é 0,25, então, o especialista dá o valor "x" para qual ele sente haver uma chance de 25% que a variável seja menor que "x". A escolha da probabilidade de cobertura para se pedir é, então, uma tarefa importante. Escolhemos simplesmente os 10^o, 25^o, 50^o, 75^o. e 90^o. percentis para elicitar, pois os quartis são bem espaçados sobre o domínio da variável e mais fáceis de se

elicitar, enquanto os 10º. e 90º. percentis podem ajudar o analista a obter mais informação sobre as espessura das caudas da densidade a ser estimada.

Pedimos, então, que o especialista fornecesse os percentis acima para o risco relativo θ de fraturas para pacientes recebendo o medicamento "alendronate". As avaliações do especialista e o histograma do risco relativo de fratura θ são dadas na Figura 1. Este dispositivo gráfico é a primeira e mais simples ferramenta para ajudar o analista a entender a forma da distribuição $f(\cdot)$.

Vale ressaltar que, durante todo o processo, foi exigido do analista revisar periodicamente se as probabilidades eram coerentes, isto é, se as probabilidades numéricas obtidas eram consistentes com a teoria de probabilidade.

Figura 1: Informações do especialista e histograma do risco relativos θ



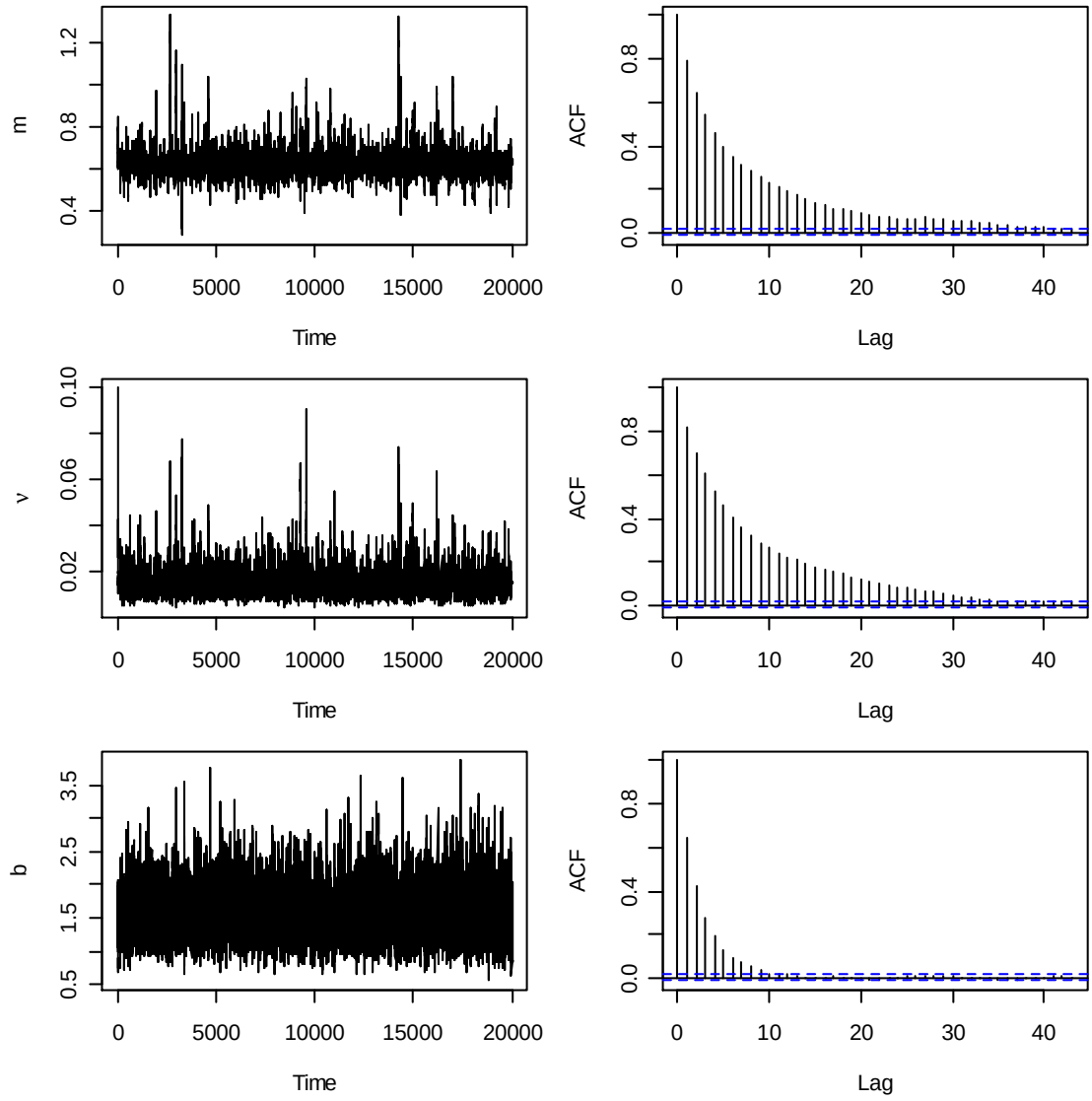
4. Determinando *a priori f*(·) do especialista

Com a informação do conhecimento do especialista transformada em percentis elicitados, o estatístico trata-os como um conjunto de dados *D* com o qual estimará os hiperparâmetros *{m, v, b, σ2}.*

Cadeias do MCMC são, então, geradas com 20 000 iterações para *{m, v, b}.* Uma avaliação visual das simulações na Figura 2 sugere que todos os hiperparâmetros

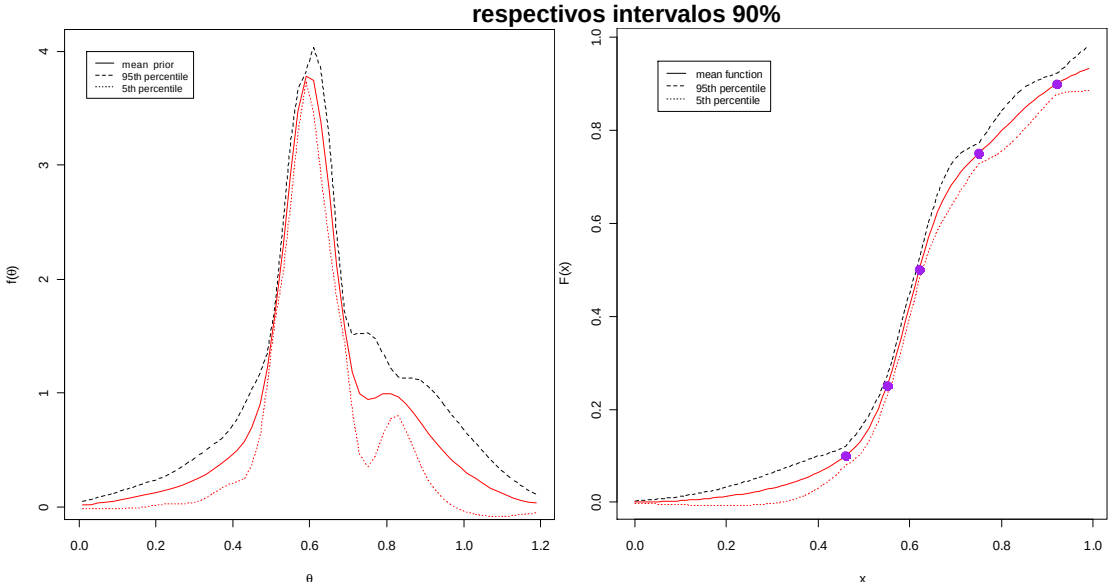
convergem e o M-H mostrou uma taxa de aceitação em torno de 35-40%. Os *plots* de autocorrelação gerados sugerem convergência também.

Figura 2: Cadeias geradas pelo MCMC para os hiperparâmetros {m, v, b} e correspondentes correlogramas



Depois das amostras de MCMC terem sido geradas ajustamos a densidade $f(\cdot)$ como a média a posteriori do processo gaussiano usando os últimos 2000 valores da cadeia. O gráfico da densidade média, 5^o e 95^o. percentis *a posteriori* de $f(\cdot)$ são mostrados na Figura 3. Os intervalos de credibilidade ilustram o grau de incerteza em torno da densidade. Na Figura 3, também mostramos a função acumulada média, 5^o. e 95^o. percentis *a posteriori*.

Figura 3: Gráficos da densidade *f*(·) e função de distribuição *F*(·) esperadas *a posteriori* do estatístico e respectivos intervalos 90%



Observamos que a densidade elicitada mostrada na Figura 3 concentra quase toda sua densidade em torno de seu centro, implicando uma pequena variância e, consequentemente, mostrando que o especialista é bastante "confiante". A Literatura tem mostrado que os especialistas, quando avaliam subjetivamente probabilidades, exibem frequentemente muita confiança até adquirirem um pouco de experiência com o processo de elicitação. Há, porém, estudos mostrando que o treinamento para elicitação do especialista pode ajudar a reduzir esta superconfiança, mas não eliminá-la. Isto pode explicar a superconfiança do especialista em nosso processo de elicitação aqui, apesar de sua experiência anterior no processo de elicitação.

As funções limites (pontilhados) do intervalo 90% mostram que restou pouquíssima incerteza sobre a densidade estimada. A maior região de incerteza aparece sobre a segunda moda e nas caudas. Possivelmente esta incerteza nas caudas é causada pela superconfiança do especialista e devido aos 10º e 90º percentis elicitados. O'Hagan (1998) afirma que pedir intervalos com alta (baixa) probabilidade de cobertura, por exemplo, 90% ou mais (10% ou menos), produz respostas não confiáveis do especialista que dão uma indicação muito pobre da variância. Além disso, é sabido que nas caudas de uma distribuição, qualquer má avaliação das probabilidades correspondentes a regiões de pequena área da cauda da distribuição pode ter um efeito drástico no intervalo avaliado (Hora *et al.*, 1992). Neste caso onde a incerteza nas caudas da distribuição estimada é alta,

poderíamos considerar uma reestruturação dos intervalos elicitados para propiciar uma cobertura adequada para suas caudas. Isto é discutido em Gosling *et al.* (2007).

Apesar das convicções *a priori* do analista de que a densidade do especialista pudesse ter a forma de uma densidade normal, isto é, ser aproximadamente simétrica e unimodal, e apesar de o histograma ser também unimodal, a Figura 3 mostra que os dados do especialista proporcionaram uma densidade bimodal. Um resultado semelhante foi observado em O'Hagan e Oakley (2007).

Note, também, na Figura 3 que praticamente não há incerteza alguma sobre a distribuição acumulada nas regiões contendo pontos elicitados devido à interpolação de $F(\cdot)$, enquanto nas caudas onde não há nenhum dado elicitado, claramente a incerteza é mais alta.

A análise dos gráficos sugere então que os quantis elicitados fornecem uma estimação razoável da densidade do especialista, contudo se a segunda moda for importante para o estudo, então quantis adicionais deveriam ser elicitados.

Ao término desta fase de estimação, no processo de elicitação, é necessário assegurar que o especialista aceite completamente e concorde que suas convicções estão realmente representadas pela distribuição *a priori* estimada e são consistentes.

5. Verificando a densidade estimada do especialista

Para verificar se o especialista propiciou uma verdadeira representação das suas convicções, foi apresentado a ele dois gráficos: a densidade $f(\cdot)$ e a função de distribuição acumulada $F(\cdot)$ como foram mostradas na Figura 3. Esta avaliação visual da distribuição esperada do estatístico ajuda o especialista a avaliar sua concordância na forma da distribuição. Às vezes, resultados inesperados do conjunto de dados elicitados podem levar à nova reflexão do especialista sobre as respostas dadas e quantis elicitados. O especialista deveria simplesmente estar satisfeito que a forma da distribuição e as caudas da densidade refletem suas incertezas, embora com exceção de características como número de modas e assimetria, um especialista pode não ser capaz de avaliar se temos uma representação boa ou não das suas convicções (Oakley e O'Hagan, 2007).

Através da função de distribuição acumulada, o especialista pode ver também se valores de $P\{\theta \leq x\}$ para novos valores de θ sugeridos pela *posteriori* do estatístico concordam com as suas convicções. A Tabela 1 fornece as probabilidades absolutas estimadas e elicitadas para o risco relativo θ. Os valores de outros quantis poderiam, então, ser checados para confirmação do especialista.

Tabela 1: Probabilidades estimadas da função de distribuição esperada do estatístico

Valores do risco θ	Probabilidades justadas $P\{\theta \leq x\}$	Desvio Padrão $DP(P\{\theta \leq x\})$
0,10	**0,003**	**1,02e-05**
0,15	**0,007**	**6,26e-05**
0,20	**0,013**	**3,12e-04**
0,25	**0,020**	**0,00116**
0,30	**0,030**	**0,00315**
0,35	**0,044**	**0,00579**
0,40	**0,064**	**0,00616**
0,45	**0,093**	**0,00137**
0,46*	**0,100**	**0,0**
0,50	**0,143**	**0,00350**
0,55*	**0,250**	**0,0**
0,60	**0,425**	**0,00239**
0,62*	**0,500**	**0,0**
0,65	**0,597**	**0,00552**
0,70	**0,697**	**0,00954**
0,75*	**0,750**	**0,0**
0,80	**0,799**	**0,00950**
0,85	**0,847**	**0,00860**
0,90	**0,887**	**0,00252**
0,92*	**0,900**	**0,0**
0,95	**0,916**	**0,00285**
0,98	**0,929**	**0,00454**

*** valores elicitados do especialista**

Ao final da aplicação do processo de elicitação o especialista concordou que a densidade estimada era uma representação razoável de suas convicções sobre o risco relativo θ.

6. Conclusões

Informações *a priori* podem proporcionar conclusões mais realísticas, particularmente onde o tamanho da amostra é relativamente pequeno como é frequentemente o caso em análises de custo-eficácia e análise de risco. Com o uso da informação *a priori* bem estruturada, podemos produzir também conclusões defensáveis e significantes.

As formas de algumas distribuições *a priori* elicitidas propostas na literatura parecem ser bastante irreais, difíceis de especificar e há poucas instruções de como elicitá-las em situações práticas. Garthwaite, Kadane e O'Hagan (2005), O'Hagan *et al*. (2006) e Wolfson (1995) proporcionaram uma boa revisão dos conceitos de elicitação e dos métodos propostos na literatura. O método apresentado neste trabalho fornece uma diretriz para desenvolver um processo de elicitação como também construir uma distribuição *a priori* válida.

O processo de elicitação ilustrado aqui é muito simples para implementar e requer elicitar apenas algumas probabilidades do especialista. Além disso, esse método nos permite medir a incerteza sobre $f(\cdot)$. Moala (2006) estendeu o método proposto por Oakley e O'Hagan para o caso multivariado.

Referências Bibliográficas

Carlin, B.P. ; Chaloner, K. ; Church, T. ; Louis, T.A. ; Matts, J.P. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. Statistician. 42; 355-367.

Chaloner, K. ; Church, T. ; Matts, J.P. ; Louis, T.A. (1993). Graphical elicitation of a prior distribution for an AIDS clinical trial. Statistician; 42; 341- 353.

Freedman, L.S. ; Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. Statistician. 32; 153-160.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. Journal of the American Statistical Association, 100, 680-701.

Gosling, J.P., Oakley, J.E. and O'Hagan, A. (2007). Nonparametric elicitation for heavy-tailed prior distributions. Bayesian Analysis 2, 693-718.

Hora, S. C., Hora, J. A. and Dodd, N. G. (1992) Assessment of probability distributions for continuous random variables: a comparison of the bisection and fixed-value methods. Organznl Behav. Hum. Decis. Process., 51,133-155.

Kadane, J. B. ; Wolfson, L. J. (1998) Experiences in elicitation. Statistician, 47,3-19.

Kennedy, M.C. ; O'Hagan, A. (1996). Iterative rescaling for Bayesian quadrature in Bayesian Statistics 5, edited by Bernardo, J.M. ; Berger, J.O. ; Dawid, A.P. ; Smith, A.F. M., pp 639-645. Oxford University Press.

Moala, F.A. (2006). Elicitation of Multivariate Prior Distributions. PhD Thesis. Department of Statistics. University of Sheffield.

Neal, R. (1999). Regression and classification using gaussian process priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, Bayesian Statistics 6, pp 69–95. Oxford: University Press.

Oakley, J. E. and O'Hagan, A. (2007). Uncertainty in prior elicitations: a nonparametric approach. Biometrika 94, 427-441.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). J.Roy. Statist. Soc. Ser. B, 40:1–42.

O'Hagan, A. (1992). Some Bayesian numerical analysis. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, Bayesian Statistics 4, pp 345–363. Oxford: University Press.

O'Hagan, A. (1998). Eliciting Expert Beliefs in Substantial Practical Applications, Statistician 47: (1), pp. 21-35 (with discussion, pp 55-68).

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. E., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). Uncertain Judgements: Eliciting Expert Probabilities. Wiley.

Qing, X. (1998). Structured Subjective Judgment on Quantifying Uncertainty of the Real Estate Project. Paper for the 1998 Conference of American Real Estate Society.

Spiegelhalter, D.J. ; Freedman, L.S. ; Parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion). Journal of the Royal Statistical Society, Series A; 157; 357-416.

Wolfson, L.J. (1995). Elicitation of priors and utilities for Bayesian analysis. PhD Thesis. Department of Statistics. Carnegie Mellon University, Pittsburgh.

Agradecimento

Gostaríamos de agradecer Matt Stephenson pela sua participação na aplicação do processo de elicitação fornecendo-nos as informações do risco de osteoporose.

Abstract

In the context of statistical analysis, elicitation is the process of gathering an expert's knowledge about some unknown quantity of interest in the form of a probability distribution. This probability distribution is generally used as prior distribution in the Bayesian analysis and hence the elicited expert's information can be incorporated in a statistics analysis to complement the information from the observed data.

Oakley and O'Hagan (2007) developed a nonparametric Bayesian approach to elicit the prior distribution for just one variable of interest. The density function is considered as an unknown function, instead of assuming a particular parametric form, and the inference is carried out about this function based just in the probabilistic specifications provided by the expert.

The main objective of this paper is to present a practical implementation of the Bayesian method proposed by Oakley and O'Hagan (2007) for elicitation of a prior distribution f(θ) for the fracture risk in patients suffering from osteoporosis under treatment. This paper also has the purpose of to divulge and to provoke interest for Elicitation, such that more statisticians can apply it in many practical problems. The theory of the elicitation approach used in this paper is also presented in a summarized and accessible way, in order to make easy its understanding and implementation.

Key words: Elicitation, expert, gaussian process, prior distribution, posterior, Bayesian approach, osteoporosis.

Sobre o Painel da Pesquisa Mensal de Emprego - PME do IBGE: problemas e soluções para o emparelhamento usando microdados

Rafael Perez Ribas[6]
Sergei Suarez Dillon Soares[7]

Resumo

O objetivo deste texto é fornecer as informações necessárias para que pesquisadores possam utilizar a Pesquisa Mensal de Emprego - PME e, em particular, o painel de domicílios/indivíduos. O painel da PME é um instrumento extremamente útil de análise que, no entanto, tem sido subutilizado devido a uma série de dificuldades práticas, principalmente na sua (re)construção. Uma das dificuldades é a ausência de uma chave identificadora da mesma pessoa ao longo das entrevistas. Assim, o emparelhamento de pessoas, feito de maneira indireta, está sujeito aos erros na informação reportada, o que reduz consideravelmente a amostra do painel. Para minimizar esse problema, fornecemos alguns procedimentos para aumentar a taxa de identificação no painel de pessoas. Esses procedimentos tomam em consideração possíveis erros na informação reportada pelos entrevistados, como na data de nascimento e na escolaridade. Com o uso de nosso algoritmo, é possível reduzir em quase 50% a taxa de atrito nos primeiros meses em relação aos algoritmos convencionais.

[6] Pesquisador do Centro Internacional de Pobreza (IPC/UNDP).
Endereço: SBS 1, Bl. J, Ed. BNDES, 10º andar. Brasília, DF, 70076-900.
rafael.ribas@undp-povertycentre.org
[7] Técnico de Planejamento e Pesquisa do IPEA.
Endereço: SBS 1, Bl. J, Ed. BNDES, 16º andar. Brasília, DF, 70076-900.
sergei.soares@ipea.gov.br

1. Introdução

A Pesquisa Mensal de Emprego - PME do IBGE é uma pesquisa domiciliar de periodicidade mensal, que apresenta um esquema de amostragem igual ao do *US Current Population Survey* (CPS). Suas informações são obtidas de uma amostra probabilística de aproximadamente 40 000 domicílios situados nas Regiões Metropolitanas do Rio de Janeiro, de São Paulo, de Porto Alegre, de Belo Horizonte, de Recife e de Salvador. O tema básico da pesquisa é o trabalho, constando algumas características demográficas e educacionais com o objetivo de possibilitar melhor entendimento sobre este tema.

Os microdados da PME podem ser trabalhados tanto na forma empilhada como na forma de série de tempo mensal. Contudo, esta base de dados é muito mais rica, visto que possibilita a estimação de transições, pois contém um painel que acompanha as unidades amostrais por até oito entrevistas. Dados em painel possuem uma série de vantagens em relação a dados *cross-section* e séries de tempo, como: inferências mais precisas sobre parâmetros de interesse (Hsiao *et al.*, 1995); maior possibilidade de modelagem do comportamento humano, como em avaliações de programas sociais (Heckman *et al.*, 1998); controle de variáveis omitidas (MaCurdy, 1981; Hsiao, 1986); e estimativa de relações dinâmicas (Nerlove, 2002).

Pesquisas em painel vêm se tornando cada vez mais comuns ao redor do mundo. Os dois principais exemplos são a *National Longitudinal Survey of Labor Market Experience* (NLS) e a *Panel Study of Income Dynamics* (PSID), ambas realizadas nos Estados Unidos (Juster, 2000). Na Europa, muitos países possuem uma pesquisa periódica em painel como o *Netherlands Socio-Economic Panel* (SEP), o *German Social Economics Panel* (GSOEP), o *Luxembourg Social Panel* (PSELL), e a *British Household Panel Survey* (BHPS). Além disso, o *European Community Household Panel* (ECHP) representa um esforço de coordenar pesquisas existentes para o levantamento de um painel representativo da Comunidade Europeia como um todo. Nos países em desenvolvimento, a disponibilidade de pesquisas ainda é pequena. Contudo, existe o apoio de Organizações Não-Governamentais - ONGs e de Organismos Internacionais interessados em informações para a avaliação e monitoramento de políticas, para que o número de pesquisas em painel cresça (Hsiao, 2006).

Todas estas pesquisas têm por característica o acompanhamento do mesmo indivíduo ao longo do tempo, independente de sua mobilidade geográfica dentro de uma mesma área de cobertura. Para isso, métodos como envio de correspondências e entrevista por telefone são frequentemente usados para evitar a perda de pessoas no painel. A PME, por outro lado, faz o acompanhamento do mesmo domicílio (residência), e não necessariamente da mesma família. Se os indivíduos que residem neste domicílio não mudarem, é possível também acompanhá-los longitudinalmente. Contudo, não é possível observar todos os indivíduos da amostra em todas as oito entrevistas, o que faz da PME um painel incompleto, ou desequilibrado (*unbalanced panel*), do ponto de vista individual.

Basicamente, são duas as principais causas para o chamado desgaste (ou atrito) em painéis: a mobilidade geográfica das pessoas na amostra e a recusa de entrevista (Peracchi e Welch, 1995). Neste artigo, contudo, levantamos ainda a hipótese de uma terceira fonte de desgaste no painel da PME. Essa fonte estaria relacionada à imprecisão na declaração das informações utilizadas como critério de emparelhamento na reconstituição do painel, por meio dos microdados. A PME não reporta um código que possibilite identificar com certeza a mesma pessoa em períodos distintos. Para contornar esse problema, é comum utilizar algumas características individuais reportadas na pesquisa para identificar a mesma pessoa em duas ou mais entrevistas. Contudo, se alguma destas características for inconsistente entre os períodos, a pessoa nunca será encontrada. Dessa forma, um "falso atrito" de observações pode ser gerado, sobreestimando o verdadeiro desgaste do painel.

O objetivo deste artigo é levantar alguns pontos sobre como construir (ou reconstituir) o painel da PME e testar qual a implicação do "falso atrito" sobre alguns resultados. Para isso, descrevemos como é o esquema de rotação da amostra na pesquisa, como identificar o mesmo domicílio em diferentes entrevistas e, por fim, como lidar com o problema do emparelhamento de pessoas. Para este emparelhamento, propomos um uso de um novo algoritmo, mais avançado que o convencional, que leva em consideração possíveis erros na informação reportada pelos entrevistados. Esse algoritmo é aplicado tanto na antiga PME, que foi realizada até dezembro de 2002, como na nova PME, que vem sendo realizada desde março de 2002.[8]

[8] Os principais objetivos da mudança na pesquisa foram: implementação de algumas mudanças conceituais no tema trabalho; ampliação da investigação para se ter melhor conhecimento da População Economicamente Ativa - PEA e da População em Idade Ativa - PIA; e melhor operacionalização dos quesitos para captação das informações de forma a aprimorar a mensuração dos fenômenos (IBGE, 2002).

Nossos resultados apontam que, com o uso de um algoritmo de emparelhamento mais avançado do que o convencionalmente usado, foi possível reduzir em quase 50% a taxa de atrito nos primeiros intervalos de meses do painel. Nos intervalos maiores, onde o desgaste é maior, a recuperação de casos representa algo próximo de seis pontos percentuais na antiga pesquisa e de 20 pontos percentuais na nova pesquisa. Além disso, a utilização deste algoritmo pode vir a aumentar a precisão das estimativas devido ao maior número de observações, principalmente quando o pesquisador trabalhar com uma amostra muito restrita.

2. A Pesquisa Mensal de Emprego

A Pesquisa Mensal de Emprego - PME é a pesquisa domiciliar com maior série ininterrupta no Brasil. A primeira PME foi a campo em janeiro de 1980, nas Regiões Metropolitanas (RMs) do Rio de Janeiro e de São Paulo, com 12 quesitos de identificação, quatro perguntas sociodemográficas e 17 variáveis sobre emprego, desemprego e rendimento do trabalho. As RMs de Porto Alegre e de Belo Horizonte foram incorporadas à pesquisa em abril de 1980, enquanto as RMs de Recife e de Salvador ingressaram em junho do mesmo ano.

Do início da pesquisa a janeiro de 1982, a PME era parte integrante da Pesquisa Nacional por Amostra de Domicílios - PNAD, utilizando-se do mesmo desenho amostral. Em fevereiro de 1982, deu início à implantação gradativa de reformulações na amostra, baseadas nas informações do Censo Demográfico 1980[9]. As reformulações foram concluídas em maio daquele mesmo ano, quando se introduziu também um novo e mais amplo questionário. Este questionário possuía 16 quesitos de identificação, seis questões sociodemográficas, cinco sobre educação e 28 relacionadas ao emprego e desemprego. O questionário permaneceu inalterado até dezembro de 2002, o que gera uma série de 248 meses com a mesma metodologia de pesquisa.

[9] Em agosto de 1988, efetuou-se nova alteração no desenho da amostra, que resultou numa redução do número de unidades selecionadas, diminuindo os custos operacionais, mas mantendo a precisão dos resultados finais em níveis aceitáveis. Em outubro de 1993, a PME começou a implantar gradativamente uma nova amostra com base no Censo Demográfico 1991, concluindo essa implantação em janeiro de 1994.

Ao se pretender resultados para cada RM abrangida, separadamente, a PME incorporou aspectos de um plano amostral autoponderado. Dentro de cada RM, a probabilidade de um domicílio ou de uma pessoa qualquer pertencer à amostra é constante e igual à fração amostral. A Tabela 1 reporta as frações amostrais das RMs e a respectiva quantidade de domicílios selecionados para a amostra, para a série anterior a 2002. Note que, para cerca de um quarto dos domicílios selecionados para a amostra, a entrevista não era efetivamente realizada.

Tabela 1 – Distribuição regional da amostra e das entrevistas nos domicílios da antiga PME

Regiões Metropolitanas	Número de setores	Fração	Número de domicílios	Entrevistas realizadas*	%
Recife	196	1/170	5 022	3 965	78,95
Salvador	169	1/170	5 100	3 912	76,71
Belo Horizonte	244	1/170	7 019	5 643	80,40
Rio de Janeiro	315	1/430	7 826	5 792	74,01
São Paulo	332	1/600	8 366	6 001	71,73
Porto Alegre	254	1/170	6 762	5 310	78,53
Total	1 510	-	40 095	30 623	76,38

Fonte: IBGE, 1998.

Nota: * Valores referentes ao mês de junho de 1998.

De acordo com IBGE (2002), as recentes mudanças na estrutura produtiva, na alocação da mão-de-obra e nas relações de trabalho, juntamente com as novas recomendações da Organização Internacional do Trabalho - OIT para a investigação da força de trabalho, levaram o instituto a realizar uma revisão da PME em todos os seus aspectos metodológico e processual. Assim, em março de 2002, uma nova PME, com um questionário ainda mais amplo, foi a campo com 14 quesitos de identificação e seis questões sociodemográficas[10], 14 sobre educação, 67 sobre emprego e desemprego. Os principais objetivos da revisão foram: implementação de algumas mudanças conceituais no tema trabalho; ampliação da investigação para se ter melhor conhecimento da População Economicamente Ativa - PEA e da População em Idade Ativa - PIA; e melhor

[10] Ao contrário da antiga pesquisa, que só entrevistava pessoas com 10 anos ou mais de idade, as novas questões sociodemográficas são perguntadas a todos os indivíduos da unidade domiciliar. Na prática, isso possibilita uma reconstituição mais precisa da composição demográfica domiciliar.

operacionalização dos quesitos para captação das informações de forma a aprimorar a mensuração dos fenômenos.

Note que ambas as pesquisas, com a antiga e a nova metodologia, foram a campo concomitantemente por dez meses, o que, em princípio, permite comparar as mudanças ocorridas. A nova PME deverá permanecer inalterada, pelo menos, até dezembro de 2010, quando o IBGE a substituirá pela mais completa Pesquisa Domiciliar Contínua. Isto levará a uma série de 106 observações no tempo.

Em relação ao desenho amostral, na Tabela 2, aumentou-se em quase 500 o número de setores selecionados e calculou-se uma nova fração amostral, que diminuiu a probabilidade de um domicílio qualquer pertencer à amostra. Por consequência, a nova pesquisa reduziu em quase 3 000 casos o número de unidades domiciliares selecionadas. Contudo, o percentual de entrevistas realizadas manteve-se igual. O IBGE (2002) recomenda que alguns cuidados devem ser tomados ao compararmos os resultados da nova pesquisa com os da antiga. Um deles diz respeito justamente à composição da amostra por RM.

Tabela 2 – Distribuição regional da amostra e das entrevistas nos domicílios da nova PME

Regiões Metropolitanas	Número de setores	Fração	Número de domicílios	Entrevistas realizadas*	%
Recife	261	1/200	4 715	3 714	78,77
Salvador	243	1/200	4 684	3 260	69,60
Belo Horizonte	359	1/200	6 644	5 253	79,06
Rio de Janeiro	406	1/500	7 576	5 339	70,47
São Paulo	431	1/700	7 820	6 276	80,26
Porto Alegre	329	1/200	5 773	4 470	77,43
Total	2 029	-	37 212	28 312	76,08

Fonte: IBGE, 2002.

Nota: * Valores referentes ao mês de junho de 2002.

Por fim, uma outra alteração ocorrida na nova pesquisa foi na forma de leitura dos microdados. Os microdados da PME são disponibilizados pelo IBGE em arquivos de texto no formato ASCII (*American Standard Code for Information Interchange*). Para a antiga

PME, os arquivos ASCII se apresentam com “pessoas na linha do domicílio”. Esta forma de apresentação consiste em ordenar as informações de pessoas depois das informações do domicílio em que habitam. Conforme o Diagrama 1, por exemplo, o primeiro domicílio, identificado pela variável I_1 e descrito pelas variáveis D_1^k, possui duas pessoas residentes, com características $P_{1,1}^k$ e $P_{1,2}^k$. O segundo domicílio, identificado pela variável I_2 e descrito pelas variáveis D_2^k, possui três residentes, descritos por $P_{2,1}^k$, $P_{2,2}^k$ e $P_{2,3}^k$.

Diagrama 1 – Informações de domicílio na linha das pessoas

I_1	D_1^1	D_1^2	D_1^3	$P_{1,1}^1$	$P_{1,1}^2$	$P_{1,1}^3$	$P_{1,2}^1$	$P_{1,2}^2$	$P_{1,2}^3$			
I_2	D_2^1	D_2^2	D_2^3	$P_{2,1}^1$	$P_{2,1}^2$	$P_{2,1}^3$	$P_{2,2}^1$	$P_{2,2}^2$	$P_{2,2}^3$	$P_{2,3}^1$	$P_{2,3}^2$	$P_{2,3}^3$

Na nova PME, os arquivos ASCII se apresentam com o “domicílio na linha das pessoas”. Este formato consiste em repetir, para cada pessoa, todas as informações do seu domicílio, conforme mostra o Diagrama 2.

Diagrama 2 – Informações de domicílio na linha das pessoas

I_1	D_1^1	D_1^2	D_1^3	$P_{1,1}^1$	$P_{1,1}^2$	$P_{1,1}^3$
I_1	D_1^1	D_1^2	D_1^3	$P_{1,2}^1$	$P_{1,2}^2$	$P_{1,2}^3$
I_2	D_2^1	D_2^2	D_2^3	$P_{2,1}^1$	$P_{2,1}^2$	$P_{2,1}^3$
I_2	D_2^1	D_2^2	D_2^3	$P_{2,2}^1$	$P_{2,2}^2$	$P_{2,2}^3$
I_2	D_2^1	D_2^2	D_2^3	$P_{2,3}^1$	$P_{2,3}^2$	$P_{2,3}^3$

A leitura dos microdados, particularmente na antiga PME, pode ser penosa para os não-iniciados. Felizmente, a partir de 2001, o IBGE passou a disponibilizar os microdados da antiga PME desde 1991, em um formato tal como o do Diagrama 2.

3. Painel da PME

Na PME, de modo a se ter maior segurança nas comparações mensais dos resultados sobre o mercado de trabalho, os mesmos informantes deveriam ser entrevistados ao longo do tempo. Somente dessa forma haveria a garantia de que as variações verificadas não estariam sendo provocadas pela troca de informantes. Contudo, para minimizar o cansaço imposto aos informantes de serem entrevistados por repetidas vezes, a pesquisa adota um esquema de rotação de painéis (*rotating panel*). Isso quer dizer que os domicílios não são entrevistados durante todos os meses de pesquisa, mas eles entram e saem da amostra de acordo com um padrão pré-definido.

A reconstituição deste painel torna a pesquisa ainda mais rica em termos de possibilidades de estudos. No entanto, alguns cuidados devem ser tomados no processo de reconstituição. Detalhes sobre a estrutura e a identificação de pessoas e domicílios no painel são colocados a seguir.

Na antiga PME, um painel equivale a um conjunto de domicílios selecionados que é dividido em quatro grupos rotacionais correspondentes cada qual a uma remessa de setores entrevistados em uma semana específica do mês. Os grupos rotacionais são indicados por uma letra, que identifica o painel, acompanhada de um subscrito correspondente à semana do mês. Se no mês *t* for aplicado o painel *B* (B1, B2, B3, B4), no mês $t+1$ será aplicado apenas 75% do seu todo (B1, B2, B3), entrando um quarto do painel seguinte *C* (C4), e assim sucessivamente. Assim, há a garantia de que 75% dos domicílios são comuns em dois meses consecutivos.

O esquema de rotação, chamado 4-8-4, determinava que, de outubro de um ano ímpar a setembro do ano seguinte, todo mês um grupo de domicílios entra na pesquisa e é entrevistado por quatro meses consecutivos. Do quinto ao décimo segundo mês, este grupo sai da amostra, retornando no décimo terceiro mês e sendo entrevistado por mais quatro vezes. Os domicílios saem da amostra definitivamente 16 meses depois de sua entrada. Importante salientar que, de outubro de um ano par a setembro do ano seguinte, nenhum grupo novo de domicílios entra na amostra. Neste período, a cada mês, apenas retornam os grupos entrevistados que estavam no intervalo de oito meses sem entrevista. Por consequência, a cada par de anos, 100% da amostra se repetem. O Quadro 1, feito para os anos de 1996 a 1999, facilita a compreensão do esquema de rotação.

Quadro 1 – Painel rotativo da antiga PME

Mês/ano	1996				1997				1998				1999			
Janeiro	D1	D2	D3	D4	D1	D2	D3	D4	G1	G2	G3	G4	G1	G2	G3	G4
Fevereiro	D1	D2	D3	E4	D1	D2	D3	E4	G1	G2	G3	H4	G1	G2	G3	H4
Março	D1	D2	E3	E4	D1	D2	E3	E4	G1	G2	H3	H4	G1	G2	H3	H4
Abril	D1	E2	E3	E4	D1	E2	E3	E4	G1	H2	H3	H4	G1	H2	H3	H4
Maio	E1	E2	E3	E4	E1	E2	E3	E4	H1	H2	H3	H4	H1	H2	H3	H4
Junho	E1	E2	E3	F4	E1	E2	E3	F4	H1	H2	H3	I4	H1	H2	H3	I4
Julho	E1	E2	F3	F4	E1	E2	F3	F4	H1	H2	I3	I4	H1	H2	I3	I4
Agosto	E1	F2	F3	F4	E1	F2	F3	F4	H1	I2	I3	I4	H1	I2	I3	I4
Setembro	F1	F2	F3	F4	F1	F2	F3	F4	I1	I2	I3	I4	I1	I2	I3	I4
Outubro	F1	F2	F3	D4	F1	F2	F3	G4	I1	I2	I3	G4	I1	I2	I3	J4
Novembro	F1	F2	D3	D4	F1	F2	G3	G4	I1	I2	G3	G4	I1	I2	J3	J4
Dezembro	F1	D2	D3	D4	F1	G2	G3	G4	I1	G2	G3	G4	I1	J2	J3	J4

Fonte: Lopes (2002).

Por exemplo, o grupo rotacional D1 entrou em janeiro de 1996, permaneceu até abril, ficou de maio até dezembro fora, voltou em janeiro de 1997 e, após abril de 1997, saiu definitivamente da pesquisa. Cabe notar que, de anos pares para anos ímpares, de janeiro a setembro, exatamente os mesmos domicílios compõem a amostra. Isto pode ser visto comparando, por exemplo, junho de 1996 com o mesmo mês em 1997. Nos meses de outubro a dezembro, alguns grupos rotacionais mudam. De um ano ímpar para um ano par, os grupos de domicílios são todos diferentes, o que pode ser verificado de junho 1997 para junho de 1998. Mais uma vez, a exceção são os grupos de domicílios que entram na amostra entre os meses de outubro e dezembro. Conhecer este esquema de rotação evita que o pesquisador gaste muito tempo procurando grupos rotacionais que já saíram da amostra.

No caso da nova PME, o padrão 4-8-4 foi mantido, mas houve um ajustamento no processo de rotação para dar mais condições de acompanhamento longitudinal dos resultados. Os grupos rotacionais foram sincronizados de tal forma que não aconteça o chamado *blackout* a cada dois anos, quando toda a amostra era trocada. Dessa forma, aumentou-se o número de grupos rotacionais de quatro para oito, rodando dois grupos por mês ao invés de apenas um, como era anteriormente. O Quadro 2 ilustra o novo sistema.

Quadro 2 – Painel rotativo da nova PME

Mês/ano	2003								2004								2005							
Janeiro	C1	C2	C3	C4	D5	D6	D7	D8	F1	F2	F3	F4	D5	D6	D7	D8	F1	F2	F3	F4	G5	G6	G7	G8
Fevereiro	E1	C2	C3	C4	C5	D6	D7	D8	E1	F2	F3	F4	F5	D6	D7	D8	H1	F2	F3	F4	F5	G6	G7	G8
Março	E1	E2	C3	C4	C5	C6	D7	D8	E1	E2	F3	F4	F5	F6	D7	D8	H1	H2	F3	F4	F5	F6	G7	G8
Abril	E1	E2	E3	C4	C5	C6	C7	D8	E1	E2	E3	F4	F5	F6	F7	D8	H1	H2	H3	F4	F5	F6	F7	G8
Maio	E1	E2	E3	E4	C5	C6	C7	C8	E1	E2	E3	E4	F5	F6	F7	F8	H1	H2	H3	H4	F5	F6	F7	F8
Junho	D1	E2	E3	E4	E5	C6	C7	C8	G1	E2	E3	E4	E5	F6	F7	F8	G1	H2	H3	H4	H5	F6	F7	F8
Julho	D1	D2	E3	E4	E5	E6	C7	C8	G1	G2	E3	E4	E5	E6	F7	F8	G1	G2	H3	H4	H5	H6	F7	F8
Agosto	D1	D2	D3	E4	E5	E6	E7	C8	G1	G2	G3	E4	E5	E6	E7	F8	G1	G2	G3	H4	H5	H6	H7	F8
Setembro	D1	D2	D3	D4	E5	E6	E7	E8	G1	G2	G3	G4	E5	E6	E7	E8	G1	G2	G3	G4	H5	H6	H7	H8
Outubro	F1	D2	D3	D4	D5	E6	E7	E8	F1	G2	G3	G4	G5	E6	E7	E8	I1	G2	G3	G4	G5	H6	H7	H8
Novembro	F1	F2	D3	D4	D5	D6	E7	E8	F1	F2	G3	G4	G5	G6	E7	E8	I1	I2	G3	G4	G5	G6	H7	H8
Dezembro	F1	F2	F3	D4	D5	D6	D7	E8	F1	F2	F3	G4	G5	G6	G7	E8	I1	I2	I3	G4	G5	G6	G7	H8

Fonte: IBGE (2002).

Na nova metodologia, o grupo E1 em vermelho, por exemplo, é entrevistado de fevereiro a maio de 2003 (quatro meses) e novamente de fevereiro a maio de 2004. A principal mudança é que a cada mês dois grupos, cada um equivalente a um oitavo da amostra, saem. Isto fica claro de janeiro a fevereiro de 2003, período quando os grupos C1 e D5 são trocados pelo E1 e C5. Ou seja, a sobreposição de 75% da amostra de um mês para o outro foi mantida, mas com a rotação de dois grupos de um oitavo cada e não de um único grupo de um quarto. Como resultado, a cada 12 meses, metade da amostra é sempre comum. Repare que os meses de junho em 2003 e em 2004 compartilham os grupos E2, E3, E4 e E5. Da mesma forma, os meses de junho em 2004 e em 2005 compartilham os grupos G1, F6, F7 e F8. Não há dúvida que o esquema atual de rotação amplia a possibilidade de se investigar certos fenômenos social e econômico que ocorrem entre um ano e outro.

Conhecidos os esquemas de rotação das PMEs, o primeiro passo para reconstituir o painel é saber como identificar o mesmo domicílio em dois períodos. As variáveis que identificam o domicílio na pesquisa são as seguintes:

Antiga	Descrição	Nova	Descrição
v010	Unidade da Federação	v035	Região Metropolitana
v101	Número no 202/203	v040	Número de controle
v102	Número de controle	v050	Número de série
v103	Número de série	v060	Painel
v106	Remessa	v063	Grupo rotacional

Para identificar a entrevista de cada domicílio univocamente, basta acrescentar ainda a variável de 'mês da pesquisa' (v070 na nova e v105 na antiga) ou a variável de 'número da pesquisa no domicílio' (v072 na nova). A variável de 'número da pesquisa no domicílio', assim como a variável 'painel' (que na nova PME é representada por v060), não consta nas bases de dados da antiga PME. Estas variáveis são muito úteis na reconstituição do painel e podem ser construídas a partir das demais variáveis de identificação de acordo com a necessidade.

A Tabela 3 apresenta a sobreposição da amostra nas duas PMEs. A primeira coluna representa a sobreposição máxima que o desenho da pesquisa permite entre os meses em termos de domicílios; a segunda coluna apresenta a sobreposição observada de domicílios; e a terceira coluna apresenta a sobreposição observada de pessoas. A razão entre a segunda coluna e a primeira coluna fornece a taxa de atrito de domicílios, enquanto a razão entre a terceira coluna e a primeira coluna fornece a taxa de atrito de indivíduos na pesquisa.

Para os domicílios, essa sobreposição é calculada a partir das variáveis descritas acima. Para os indivíduos, o emparelhamento é mais complicado. A PME não reporta um código que possibilite identificar com certeza a mesma pessoa em períodos distintos[11]. Além disso, a pesquisa não coleta informações sobre números de registros (NIS, RG, CPF, Título de Eleitor, etc.) e nem divulga os nomes das pessoas entrevistadas, por razões óbvias de confidencialidade.

De acordo com Lopes (2002), as informações disponíveis para o emparelhamento de pessoas são dia, mês e ano de nascimento e sexo. Denominamos esta combinação de variáveis de "emparelhamento básico". Somada a essas características, alguns autores utilizam ainda a escolaridade, a condição no domicílio e/ou o número de ordem da pessoa como critérios de emparelhamento.[12] Porém, essas características são sensíveis a mudanças no tempo, já que as pessoas nascem, morrem ou se mudam e, por conseqüência, alteram a composição domiciliar. Além disso, elas não acrescentam muito

[11] Apesar de alguns pensarem que o número de ordem da pessoa no domicílio não muda entre as entrevistas, esta variável não possui o propósito de identificador no painel. Este número é atribuído independentemente a cada mês de acordo com a condição das pessoas no domicílio naquele instante. Basta uma pessoa no domicílio alterar a sua condição, que o número de ordem dela e das demais não será o mesmo entre duas entrevistas.
[12] Além de Lopes (2002), ver Neri *et al*. (1997), Corseuil e Carneiro (2001), Lemos (2002), Woltermann (2002), Penido e Machado (2003), Gonzaga e Reis (2005), e Machado *et al*. (2007).

rigor ao emparelhamento, pois, com exceção dos gêmeos, é muito difícil encontrar, em um mesmo domicílio, pessoas com a mesma data de nascimento.

Cabe ressaltar que, se durante o período em que o domicílio permanece na amostra, a família mudar de endereço e outra família passar a ocupar a unidade domiciliar, a informação será obtida com a nova família pelo período restante. Portanto, é possível que a sobreposição de domicílio no painel não implique em sobreposição de famílias ou pessoas.

Tabela 3 – Sobreposição da amostra na PME entre pares de meses seguindo os critérios básicos de emparelhamento

Intervalo de meses	Antiga PME			Nova PME		
	Máximo	Domicílios	Indivíduos	Máximo	Domicílios	Indivíduos
1	75,0%	71,6%	65,1%	75,0%	72,0%	65,2%
2	50,0%	47,0%	42,0%	50,0%	47,3%	41,0%
3	25,0%	23,1%	20,2%	25,0%	23,2%	19,2%
4-8	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
9	25,0%*	22,8%	19,8%	12,5%	11,1%	5,6%
10	50,0%*	45,5%	38,9%	25,0%	22,4%	11,4%
11	75,0%*	67,9%	57,1%	37,5%	33,6%	17,0%
12	**100%***	**89,7%**	**74,1%**	**50,0%**	**44,5%**	**22,3%**
13	75,0%*	67,0%	54,5%	37,5%	30,5%	15,2%
14	50,0%*	44,3%	35,4%	25,0%	18,3%	9,0%
15	25,0%*	21,8%	17,1%	12,5%	8,1%	3,9%
16 ou mais	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Fonte: Elaboração própria a partir da PME 1998, 1999, 2003 e 2004.

Nota: * De ano par para ímpar. Valores referentes à média das amostras presentes nos meses de janeiro a setembro de 1998, para a antiga PME, e nos meses de janeiro a dezembro de 2003, para a nova PME.

Em decorrência dos domicílios que, em períodos subsequentes, passam a ser inexistentes, estão fechados ou cujos habitantes se recusaram a responder à pesquisa, o emparelhamento dos domicílios não é perfeito. Como a identificação do domicílio é precisa, nada se pode fazer com relação a esta forma de desgaste no painel. Se o domicílio não foi entrevistado, nenhum procedimento estatístico pode revelar sua informação. A segunda forma de desgaste – pessoas não emparelhadas porque seu domicílio não foi entrevistado ou porque ela não fazia parte do domicílio no momento da entrevista – é obviamente maior. Contudo, à informação utilizada para o emparelhamento de indivíduos pode não ser tão precisa quanto a informação utilizada no emparelhamento de domicílios e, portanto, a sobreposição de pessoas pode estar subestimada, desde que haja algum erro na informação individual reportada em algum dos meses.

A Tabela 4 mostra a taxa de atrito da amostra, que é a proporção do desgaste ocorrido entre os meses sobre a perda máxima de domicílios imposta pelo esquema de rotação. Em ambas as PMEs, as taxas de atrito no intervalo de um mês são em torno de 4% para domicílios e de 13% para indivíduos, sendo crescentes com o aumento no intervalo dos meses. Até o intervalo de 12 meses, a perda proporcional de domicílio continua semelhante entre as PMEs, chegando perto de 10%. Porém, as perdas relativas de indivíduos são muito mais altas na PME nova a partir dos nove meses de intervalo, com mais de 50% de perda, chegando quase a 70% no 16º mês. Na antiga PME, a taxa de atrito de indivíduos varia de 20% a 30% a partir dos nove meses de intervalo.

Tabela 4 – Perdas entre pares de meses como proporção do limite rotacional da PME, seguindo os critérios básicos de emparelhamento

Intervalo de meses	Antiga PME			Nova PME		
	Domicílios	Indivíduos		Domicílios	Indivíduos	
		total	em domicílios emparelhados		total	em domicílios emparelhados
1	4,5%	13,2%	9,4%	4,0%	13,0%	9,6%
2	6,0%	16,1%	11,0%	5,5%	18,0%	13,1%
3	7,7%	19,3%	12,7%	7,3%	23,1%	16,4%
9	8,6%	20,8%	13,0%	11,1%	54,9%	45,1%
10	9,0%	22,3%	14,4%	10,5%	54,5%	45,4%
11	9,5%	23,9%	15,6%	10,4%	54,7%	45,6%
12	**10,3%**	**25,9%**	**16,9%**	**11,0%**	**55,5%**	**45,8%**
13	10,7%	27,3%	18,1%	18,7%	59,6%	42,3%
14	11,5%	29,3%	19,3%	26,9%	64,1%	38,6%
15	13,0%	31,6%	20,4%	35,4%	68,9%	34,8%

Fonte: Elaboração própria a partir da PME 1998, 1999, 2003 e 2004.
Nota: Valores referentes à média das amostras presentes nos meses de janeiro a setembro de 1998, para a antiga PME, e nos meses de janeiro a dezembro de 2003, para a nova PME.

Uma diferença clara entre o painel da antiga e da nova PME é a razão entre as taxas de atrito de domicílios e indivíduos. Na antiga, o desgaste de indivíduos é cerca de duas vezes e meia maior que o desgaste de domicílios, enquanto na nova, o desgaste de indivíduos é de três a cinco vezes maior que o desgaste de domicílios. O aumento relativo no desgaste de indivíduos pode ser tanto relacionado à mudança na metodologia (tamanho e composição da amostra) e/ou à queda na qualidade da informação utilizada como critério para o emparelhamento quanto a mudanças conjunturais e nos arranjos domiciliares. Pela Tabela 4, constatamos que entre 65% e 70% dos indivíduos atritados estão em domicílios

emparelhados na antiga pesquisa, enquanto, na nova pesquisa, essa proporção pode ser maior que 80%, dependendo do intervalo de meses.

4. Os problemas no emparelhamento de pessoas e como solucioná-los

Como já foi colocado, se a informação utilizada para o emparelhamento de indivíduos não for precisa, a sobreposição de pessoas na amostra pode ser subestimada. Apresentamos a seguir alguns destes exemplos, encontrados aleatoriamente nas PMEs de 2002 e 2003. No primeiro caso, a única pessoa residente no domicílio identificado por '26000012 1 D 1' (v040 + v050 + v060 + v063) apresentava-se na sua primeira entrevista, em junho de 2002, da seguinte forma:

v203	v204	v214	v224	v234	v208	vdae1
2	99	99	9999	65	1	4

Ela era uma mulher (v203 = 2), branca (v208 = 1), de 65 anos (v234 = 65) e com escolaridade entre 8 e 10 anos de estudo completos (vdae1 = 4). Note que, o dia (v204), o mês (v214) e o ano de nascimento (v224) não foram informados. No mês seguinte, possivelmente a mesma mulher se apresenta da seguinte forma:

v203	v204	v214	v224	v234	v208	vdae1
2	20	5	1934	68	1	5

Desta vez, ela declarou a sua data de nascimento. Pelo critério básico de emparelhamento, diríamos que elas não são a mesma pessoa, justamente porque uma das datas de nascimento não foi declarada. Além disso, se fosse utilizada outra variável como critério, como idade presumida (v234)[13] ou escolaridade (vdae1), o emparelhamento também não iria funcionar. Mais um mês à frente, a encontramos novamente, porém com um novo nível de escolaridade:

v203	v204	v214	v224	v234	v208	vdae1
2	20	5	1934	68	1	3

Ou seja, em um mês, essa mulher passou de um nível de escolaridade com mais de 11 anos de estudo completos para menos de 7 anos de estudo completos.

O segundo caso, referente ao domicílio identificado por '26000012 3 D 1', é um exemplo de mudança na composição da família combinada com algumas inconsistências nas informações declaradas. Em junho de 2002, o domicílio se apresentava da seguinte forma:

v201	v203	v205	v206	v207	v208	v204	v214	v224	v234
1	2	1	1	1	1	24	12	1970	31
2	1	3	3	1	4	7	2	1990	12
3	1	3	3	1	4	99	9	1992	9
4	2	5	1	2	4	12	6	1980	21
5	1	4	3	2	4	2	4	1998	4
6	1	4	3	2	4	99	99	1999	2

Eram duas mulheres, provavelmente irmãs, com dois filhos homens cada uma. Em julho daquele ano, uma nova pessoa apareceu no domicílio, provavelmente o irmão das mulheres de acordo com a relação de parentesco (v205):

[13] A idade presumida é uma variável calculada a partir da diferença entre a data de nascimento da pessoa e a data de entrevista. Porém, caso o informante não saiba o ano ou o mês de nascimento da pessoa, o entrevistador pergunta qual a idade que o informante presume que esta pessoa tenha. Por se tratar de uma variável que exige um menor rigor em sua precisão, a probabilidade de erro na idade presumida é menor do que na data de nascimento.

v201	v203	v205	v206	v207	v208	v204	v214	v224	v234
1	2	1	1	1	4	24	12	1970	31
2	1	3	3	1	4	7	1	1990	12
3	2	3	3	1	4	14	7	1992	9
4	2	4	1	2	4	12	6	1980	22
5	1	4	3	2	4	2	4	1998	4
6	1	4	3	2	4	5	5	1999	3
7	1	4	4	1	4	13	12	1968	33

Note que a chefe (v201 = 1) passou de branca (v208 = 1) a parda (v208 = 4), um de seus filhos (v201 = 2) mudou o mês de nascimento (v214) e o outro (v201 = 3) mudou de sexo (v203), além de passar a responder o dia e o mês de nascimento. O critério básico de emparelhamento, neste caso, levaria à conclusão que a chefe perdeu um de seus filhos homens de um mês para o outro, mas ganhou uma filha mulher. A entrada do novo membro no domicílio fez com que a ordem das pessoas entrevistadas (v201) mudasse no mês seguinte:

v201	v203	v205	v206	v207	v208	v204	v214	v224	v234
1	2	1	1	1	4	24	12	1970	31
2	1	3	3	1	4	7	1	1990	12
3	2	4	1	2	4	12	6	1980	22
4	1	4	4	1	4	13	12	1968	33
5	1	4	3	2	4	2	4	1998	4
6	1	4	3	2	4	5	5	1999	3
7	1	3	3	1	4	16	8	1992	9

A pessoa v201 = 3 passou para v201 = 7, a v201 = 4 passou para v201 = 3 e a v201 = 7 passou para v201 = 4. Além disso, a criança que mudou de sexo no mês anterior, trocou o sexo e a data de nascimento novamente. Esta família foi observada neste domicílio até a quarta entrevista. Na quinta entrevista, depois dos oito meses de intervalo, ninguém foi encontrado no domicílio. Na sexta entrevista, havia apenas um homem de 22 anos residindo no local e, na oitava e última entrevista, já havia uma outra família de três pessoas.

O último exemplo é de uma família que foi entrevista oito vezes, residente no domicílio ‘26000012 4 D 1’. Essa família é formada por um casal com dois filhos, sendo que um deles possui uma esposa e um filho. Note como as datas de nascimento (v204, v214 e v224) e os níveis de escolaridade (vdae1) do chefe e do cônjuge neste domicílio mudam de um mês para o outro:

v070	v072	v075	v203	v205	v204	v214	v224	v234	vdae1
6	1	2002	2	1	5	11	1959	42	3
6	1	2002	1	2	99	99	9999	65	5
7	2	2002	2	1	5	12	1956	45	3
7	2	2002	1	2	1	1	1939	63	4
8	3	2002	2	1	5	12	1956	45	3
8	3	2002	1	2	1	1	1939	63	5
9	4	2002	2	1	5	12	1956	45	3
9	4	2002	1	2	1	1	1939	63	2
6	5	2003	2	1	5	2	1956	47	3
6	5	2003	1	2	1	1	1939	64	4

As variáveis v070, v072 e v075 indicam, respectivamente, o mês, o número e o ano da entrevista.

A partir da observação de casos como os apresentados acima, sentimos a necessidade de montar um algoritmo mais avançado de emparelhamento. A ideia é eliminar ao máximo o “falso atrito” gerado na amostra decorrente de erros nas informações declaradas, tendo o cuidado, porém, de não emparelhar pessoas distintas. Este algoritmo, com sua sintaxe para *Stata* apresentada no Apêndice[14], utiliza critérios não só de exatidão nas variáveis de identificação, mas também de proximidade nas respostas.

Para o emparelhamento, construiu-se uma nova variável, denominada p201, que identifica o mesmo indivíduo em todo o painel. Essa variável é igual ao número de ordem da pessoa (v201) no domicílio, caso ela seja observada pela primeira vez no painel durante a primeira entrevista em seu domicílio. Se a pessoa aparece pela primeira vez no painel durante a segunda entrevista no domicílio, p201 será igual a v201 somado 100; se ela aparecer pela primeira vez durante a terceira entrevista, p201 será igual a v201 somado 200; e assim sucessivamente. Além dessa variável, outras duas foram criadas, uma indicando se a mesma pessoa foi identificada na entrevista seguinte (denominada *forw*) e outra indicando se a mesma pessoa foi identificada na entrevista anterior (denominada *back*).

O algoritmo de emparelhamento nada mais é do que um processo de busca da mesma pessoa em uma posição anterior na base de dados, seguindo um critério de ordenação crescente em relação ao número da entrevista no domicílio e que otimize essa

busca de acordo com as variáveis que identificam esta pessoa. Para construção da variável p201, em uma primeira rodada, emparelhamos os indivíduos na segunda entrevista com aqueles da primeira entrevista. Assim, o número de ordem da pessoa na primeira entrevista é atribuído à variável p201 dela na segunda entrevista. Na segunda rodada, emparelhamos os indivíduos na terceira entrevista com aqueles na segunda entrevista, atribuindo o valor de p201 de um ao outro. Este processo é repetido sucessivamente em um *loop* até emparelhar as pessoas na oitava entrevista com as na sétima entrevista. Ao final, para reconstituir os casos onde o indivíduo sai da amostra, porém, retorna meses depois, utiliza-se um outro loop semelhante ao anterior, porém retrospectivo. Esse *loop* verifica se as pessoas que apareceram pela primeira vez na oitava entrevista não são observadas em entrevistas anteriores à sétima; em seguida, verifica se as pessoas que apareceram pela primeira vez na sétima entrevista não são observadas em entrevistas anteriores à sexta; e assim sucessivamente.

No primeiro *loop*, os critérios de emparelhamento seguem uma hierarquia do mais ao menos rigoroso e do menos ao mais preciso. Assim, só participam do processo menos rigoroso aqueles indivíduos que não emparelharam no processo mais rigoroso. Além disso, a manipulação das variáveis *forw* e *back* garante que os indivíduos não sejam emparelhados com duas pessoas ao mesmo tempo.

As ideias de rigor e precisão geram, na realidade, um *trade-off*. Um critério rigoroso (com baixa precisão) é aquele em que as pessoas emparelhadas possuem uma alta probabilidade de serem as mesmas. Contudo, as pessoas não-emparelhadas por este critério não possuem necessariamente baixa probabilidade de serem as mesmas. Um exemplo de critério muito rigoroso e com baixa precisão é a data de nascimento. Por outro lado, um critério preciso (com baixo rigor) é aquele em que as pessoas não-emparelhadas possuem uma baixa probabilidade de serem as mesmas, porém as pessoas emparelhadas não possuem necessariamente alta probabilidade de serem as mesmas. Exemplos de critérios pouco rigorosos com alta precisão são sexo e escolaridade.

Por causa dessa diferença na classificação dos critérios, há uma hierarquia em sua utilização, tal que os mais precisos e menos rigorosos só são adotados quando já há poucas pessoas não-emparelhadas nos domicílios. Isso, consequentemente, aumenta a

[14] Para rodar a sintaxe em *Stata*, sugerimos que o usuário remonte as bases de dados da PME por painel (variável v060), incluindo todas as informações da primeira à oitava entrevista. Assim, não seria exigido um uso excessivo de memória.

probabilidade das pessoas emparelhadas serem as mesmas quando utilizado um critério pouco rigoroso.

O Quadro 3 apresenta os critérios do emparelhamento avançado na ordem em que são aplicados. Cabe notar que, após a utilização dos critérios mais rigorosos, alguns filtros amostrais são utilizados. A justificativa para o primeiro filtro ("chefes, cônjuges e filhos com 25 anos ou mais") é que alguns critérios mais precisos só possuem certo rigor quando aplicados a pessoas adultas que pertencem ao núcleo familiar. Já o segundo filtro ("indivíduos em domicílio onde alguém já emparelhou") é aplicado para que famílias distintas, que residiram no mesmo domicílio, não sejam emparelhadas ao se adotar os critérios menos rigorosos.

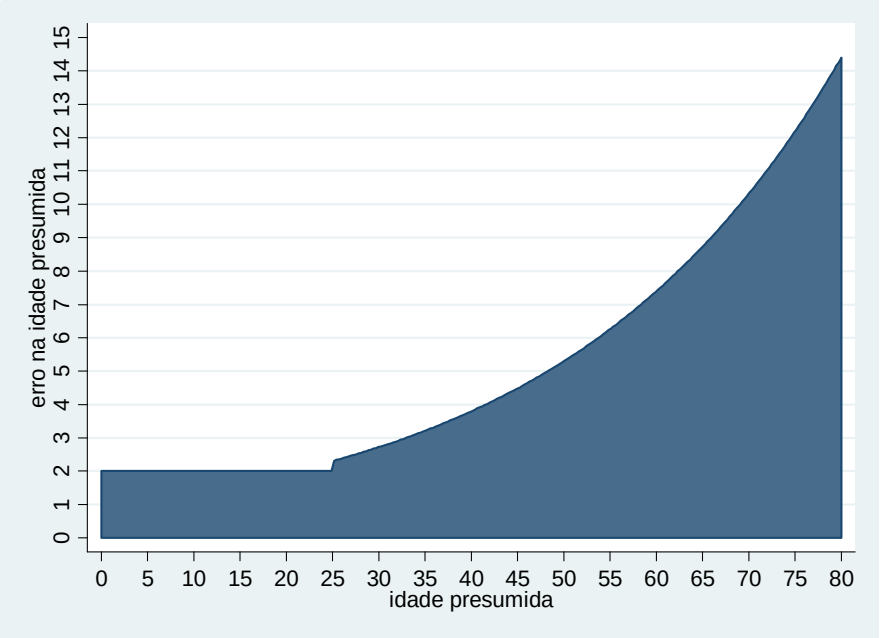
Quadro 3 – Combinação dos critérios de emparelhamento avançado

	Ordem de combinação dos critérios											
	básico	avançado										
			somente para chefes, cônjuges e filhos com 25 anos ou mais				somente para indivíduos em domicílio onde alguém já emparelhou					
Critérios de emparelhamento	1	2	3	3 [ii]	3 [iii]	3 [iv]	4	4 [v]	5	6	7	8
mesmo sexo	x		x	x	x	x	x	x	x	x	x	x
mesmo dia de nascimento	x	x										
até 4 dias de diferença no dia de nascimento			x		x							
mesmo mês de nascimento	x	x										
até 2 meses de diferença no mês de nascimento			x	x								
mesmo ano de nascimento	x											
mesma idade presumida							x					
diferença na idade presumida igual a 1									x			
diferença na idade presumida igual a 2										x		
diferença na idade presumida [i] $\leq \begin{cases} 2 & \text{se idade} < 25 \\ \exp(\text{idade}/30) & \text{se idade} \geq 25 \end{cases}$			x	x	x	x					x	
função acima multiplicada por dois (sse idade ≥ 25)												x
mesmo número de ordem		x										
mesma condição no domicílio								x				
mesmo nível de escolaridade								x				
até 1 ciclo de diferença no nível de escolaridade				x	x	x						

Nota: [i] ver Figura 1 abaixo. [ii] Se dia de nascimento for não-observado. [iii] Se mês de nascimento for não-observado. [iv] Se dia e mês de nascimento forem não-observados. [v] Se idade presumida for não-observada.

No Quadro 3, é possível notar que alguns critérios, como dia e mês de nascimento e idade, perdem rigor à medida que os emparelhamentos são executados. Por exemplo, depois de emparelhadas as pessoas que possuem a mesma data de nascimento, procuram-se aquelas que possuem datas de nascimento próximas, porém não exatamente iguais. Outro exemplo é o de erro na idade presumida em função da própria idade da pessoa. A idéia de utilizar essa função surgiu depois de constatarmos que pessoas mais velhas possuem maior dificuldade em reportar a sua idade corretamente. Portanto, definimos uma função que impõe um limite máximo de diferença na idade presumida entre dois períodos. O erro aceitável para a idade presumida está demarcado na Figura 1.

Figura 1 – Área correspondente ao erro aceitável para a idade presumida



A Tabela 5 mostra as taxas de atrito após a utilização do algoritmo de emparelhamento avançado, considerando todos os critérios mencionados a cima. Em ambas as pesquisas, é possível recuperar entre 7 e 10 pontos percentuais da amostra nos primeiros meses de intervalo. Isso significa que a taxa de atrito cai quase que pela metade ao passarmos do emparelhamento básico para o avançado. Em domicílios emparelhados, essa recuperação de indivíduos representa uma redução de mais de 60% na taxa de atrito. Nos intervalos maiores, recuperaram-se entre 5 e 6 pontos percentuais da amostra na antiga PME e entre 15 e 25 pontos percentuais da amostra na nova PME. Em domicílios

emparelhados, isso representa reduções na taxa de atrito entre 25% e 40% e entre 40% e 55%, respectivamente.

Tabela 5 – Perdas entre pares de meses como proporção do limite rotacional da PME, seguindo os critérios avançados de emparelhamento

Intervalo de meses	Antiga PME			Nova PME		
	Percentual de indivíduos			Percentual de indivíduos		
	atritados (não recuperados)		recuperados	atritados (não recuperados)		recuperados
	total	em domicílios emparelhados		total	em domicílios emparelhados	
1	6,0%	2,1%	7,3%	6,1%	2,7%	6,9%
2	8,4%	3,3%	7,7%	9,3%	4,5%	8,6%
3	11,2%	4,6%	8,1%	12,7%	6,1%	10,3%
9	15,3%	7,6%	5,4%	30,2%	20,4%	24,7%
10	16,6%	8,7%	5,7%	30,3%	21,1%	24,3%
11	18,0%	9,7%	5,9%	30,8%	21,7%	23,9%
12	**19,6%**	**10,7%**	**6,2%**	**31,9%**	**22,2%**	**23,6%**
13	21,1%	11,9%	6,2%	41,2%	23,9%	18,4%
14	24,7%	14,7%	4,6%	45,8%	20,2%	18,4%
15	26,0%	14,7%	5,7%	54,6%	20,5%	14,3%

Fonte: elaboração própria a partir da PME 1998, 1999, 2003 e 2004.
Nota: Valores referentes à média das amostras presentes nos meses de janeiro a setembro de 1998, para a antiga PME, e nos meses de janeiro a dezembro de 2003, para a nova PME.

Após o uso do algoritmo de emparelhamento avançado, constatamos que a taxa de atrito nos intervalos maiores continua mais elevada na nova PME que na antiga. No entanto, a recuperação de observações no painel da nova pesquisa é consideravelmente mais elevada. Isso ressalta a hipótese que o aumento relativo no desgaste de indivíduos de uma pesquisa para a outra se deve, entre outras coisas, à queda na qualidade da informação utilizada como critério para o emparelhamento.

Para corrigir os problemas informacionais e facilitar a reconstituição do painel não só da PME, mas de outras pesquisas que venham a ser realizadas neste formato, sugerimos que o IBGE adote um processo de crítica dos dados que garanta uma consistência nas informações entre as entrevistas. Por exemplo, a data de nascimento de uma pessoa entrevistada em janeiro teria que ser consistente com a data declarada em fevereiro. Além disso, seria interessante o uso de um código único por pessoa que a identifique da primeira à oitava entrevista. Com isso, seria possível perguntar inclusive a razão pela ausência ou inclusão da pessoa no domicílio. Por fim, acreditamos que estas sugestões não são de difícil

operacionalização, principalmente porque o IBGE vem adotando o uso de *Personal Digital Assistants* - PDAs na coleta de informações desde 2002.

5. Implicação de mudanças no emparelhamento sobre alguns resultados

Se o erro na informação que impossibilita o emparelhamento de pessoas no painel ocorre de maneira aleatória, podemos dizer que a utilização do algoritmo de emparelhamento avançado possui apenas a vantagem de aumentar o tamanho da amostra. Caso contrário, além da redução no tamanho amostral, a utilização apenas do emparelhamento básico resultaria em análises viesadas.

Para investigar a ocorrência deste viés, as Tabelas 6 e 7 apresentam matrizes de transições de posição na ocupação, mensais e anuais, respectivamente, calculadas com base em dois painéis construídos. Ambos os painéis são de adultos, entre 18 e 60 anos, que estavam inicialmente ocupados no mercado de trabalho. Contudo, o primeiro painel foi construído com base no algoritmo de emparelhamento básico, enquanto o segundo painel foi construído com base no algoritmo de emparelhamento avançado.

Nas transições mensais (Tabela 6), apesar de a amostra aumentar 11% quando se utiliza o emparelhamento avançado, nenhuma diferença considerável entre os resultados é identificada. Portanto, neste caso em específico, o chamado ‘falso atrito’ entre os meses parece ser ortogonal aos resultados de interesse.

Tabela 6 - Transições mensais de posição na ocupação

Resultado do emparelhamento básico

mês anterior	posição na ocupação							
	inativo	desempregado	empregado SC	empregado CC	conta-própria	empregador	não-remunerado	total
empregado SC	5,46	3,27	73,20	10,77	6,10	0,90	0,31	100
empregado CC	1,96	0,99	4,19	91,60	0,93	0,31	0,02	100
conta-própria	5,98	2,09	7,87	2,70	76,69	4,25	0,42	100
empregador	1,97	0,51	4,27	3,17	15,83	73,62	0,63	100
não-remunerado	14,26	2,96	11,35	2,41	11,76	4,72	52,54	100
total	3,64	1,73	21,14	50,47	17,49	4,96	0,57	100
n. observações	1.700.806							

Resultado do emparelhamento avançado

mês anterior	posição na ocupação							
	inativo	desempregado	empregado SC	empregado CC	conta-própria	empregador	não-remunerado	total
empregado SC	5,68	3,36	72,52	11,03	6,19	0,90	0,31	100
empregado CC	2,12	1,09	4,40	91,03	1,01	0,33	0,02	100
conta-própria	6,22	2,18	8,06	2,94	75,85	4,32	0,43	100
empregador	2,10	0,52	4,42	3,35	16,33	72,63	0,65	100
não-remunerado	14,52	3,04	11,43	2,47	11,99	4,82	51,73	100
total	3,83	1,82	21,34	50,23	17,33	4,86	0,57	100
n. observações	1.886.293							

Fonte: elaboração própria a partir da PME 2002-2007.
Nota: SC = sem carteira; CC = com carteira.

Nas transições anuais (Tabela 7), podemos notar que as taxas de permanência na mesma posição estão significativamente sobreestimadas quando se utiliza a amostra com emparelhamento básico. Em contrapartida, as taxas de transição para a posição de empregado com ou sem carteira assinada e, principalmente, para a inatividade estão subestimadas. Além disso, a amostra derivada do emparelhamento avançado 55% maior que a amostra derivada do emparelhamento básico. Independente do viés, esse aumento na amostra pode significar um incremento considerável na eficiência de estimativas, principalmente quando o pesquisador trabalhar com uma amostra muito restrita, em geral mulheres, entre 20 e 30 anos, casadas, com filhos.

Tabela 7 - Transições anuais de posição na ocupação

Resultado do emparelhamento básico

ano anterior	posição na ocupação							
	inativo	desempregado	empregado SC	empregado CC	conta-própria	empregador	não-remunerado	total
empregado SC	10,79	6,45	52,12	19,46	9,24	1,55	0,38	100
empregado CC	4,70	3,49	5,45	83,55	2,13	0,64	0,05	100
conta-própria	10,64	2,81	10,62	5,83	63,61	5,98	0,50	100
empregador	3,98	1,05	5,90	5,27	20,69	62,29	0,82	100
não-remunerado	19,55	4,82	14,17	6,37	17,92	6,80	30,37	100
total	7,30	3,88	16,88	48,89	17,14	5,43	0,48	100
n. observações	422.547							

Resultado do emparelhamento avançado

ano anterior	posição na ocupação							
	inativo	desempregado	empregado SC	empregado CC	conta-própria	empregador	não-remunerado	total
empregado SC	11,48	6,50	51,07	19,55	9,49	1,53	0,37	100
empregado CC	5,15	3,64	6,18	81,90	2,40	0,68	0,06	100
conta-própria	11,51	3,17	11,29	6,65	60,72	6,07	0,58	100
empregador	4,68	1,10	6,77	6,02	22,09	58,48	0,86	100
não-remunerado	21,79	4,87	14,09	6,03	18,20	7,28	27,74	100
total	7,95	4,08	17,63	48,29	16,53	5,03	0,48	100
n. observações	656.929							

Fonte: Elaboração própria a partir da PME 2002-2007.
Nota: SC = sem carteira; CC = com carteira.

6. Conclusão

A finalidade deste artigo é de apresentar alguns aspectos sobre como lidar com o painel da PME, particularmente em relação ao seu desgaste. Na reconstituição do painel através dos microdados da pesquisa, mostramos que, além das diversas razões que fazem as pessoas não serem encontradas nas entrevistas seguintes, há ainda outra fonte de desgaste ou atrito. Essa fonte trata-se da imprecisão nas informações declaradas pelas pessoas entrevistadas.

A PME não reporta um código que possibilite identificar com certeza a mesma pessoa em períodos distintos. Portanto, é comum utilizarmos algumas características individuais, reportadas com certa imprecisão, para identificar a mesma pessoa em duas ou mais entrevistas. De fato, quanto maior a imprecisão nas informações que servem como uma chave de reconstituição, maior é o "falso atrito" gerado, o que sobreestima o verdadeiro desgaste do painel.

De acordo com o processo de emparelhamento convencional dos microdados, a perda de pessoas no painel no intervalo de um mês é em torno de 13%, sendo crescente com o aumento no intervalo dos meses. A partir dos nove meses de intervalo, a taxa de atrito de indivíduos varia entre 20% e 30%, na antiga PME, e chega a mais de 50% na nova pesquisa. Com o uso de um algoritmo de emparelhamento avançado, foi possível recuperar entre 7 e 10 pontos percentuais das amostras nos primeiros meses de intervalo, o que representa uma queda de quase 50% na taxa de atrito. Nos intervalos maiores, recuperamos entre 5 e 6 pontos percentuais da amostra na antiga PME e entre 15 e 25 pontos percentuais da amostra na nova PME. Portanto, parte do aumento relativo no desgaste de indivíduos de uma pesquisa para a outra se deve, de fato, à queda na qualidade da informação utilizada como critério para o emparelhamento. Contudo, mesmo após o emparelhamento avançado, a taxa de atrito nos intervalos maiores continua mais elevada na nova PME que na antiga, sugerindo que há também um efeito de mudanças no desenho e na composição amostral sobre esta taxa.

Além de reduzir consideravelmente o tamanho amostral, o erro de informação, que dificulta a reconstituição do painel, pode causar também mudanças nos resultados de algumas análises, principalmente quando se trata da investigação de mudanças anuais. Portanto, um melhor emparelhamento de pessoas, que recupere os casos com imprecisão nas informações, implica uma maior precisão nas estimativas. Resta testar, por meio de simulações, se o algoritmo que propusemos é realmente efetivo nessa recuperação.

Como sugestão para facilitar a reconstituição do painel não só da PME, mas também de outras pesquisas que venham a ser realizadas neste formato, recomendamos que o IBGE adote um processo de crítica dos dados que garanta uma consistência nas informações entre as entrevistas, além do uso de um código único por pessoa que a identifique da primeira à oitava entrevista. Com isso, seria possível perguntar ainda a razão

pela ausência ou inclusão da pessoa no domicílio, o que possibilitaria avanços em pesquisas relacionadas a mudanças nas composições domiciliares.

Referências Bibliográficas

Corseuil, C. H. and F. G. Carneiro, 2001. "Os Impactos do Salário Mínimo sobre Emprego e Salários no Brasil: Evidências a partir de Dados Longitudinais e Séries Temporais," Texto para Discussão n. 849, IPEA, Rio de Janeiro.

Gonzaga, G. and M. C. Reis, 2005. "Os Efeitos Trabalhador Adicional e Desalento no Brasil," Anais do XXXIII Encontro Nacional de Economia, ANPEC, João Pessoa.

Heckman, J.J., H. Ichimura, J. Smith and P. Todd, 1998. "Characterizing Selection Bias Using Experimental Data", Econometrica 66: 1017-1098.

Hsiao, C., 1986. Analysis of Panel Data, Econometric Society monographs No. 11, New York: Cambridge University Press.

Hsiao, C., 2006. "Longitudinal Data Analysis," The New Palgrave Dictionary of Economics, MacMillan.

Hsiao et al., D.C. Mountain and K. Ho-Illman, 1995. "Bayesian Integration of End-Use Metering and Conditional Demand Analysis", Journal of Business and Economic Statistics 13: 315-326.

IBGE, Departamento de Emprego e Rendimento. Para Compreender a PME: (um texto simplificado), 4ª ed. Rio de Janeiro, 1998.

IBGE, Departamento de Emprego e Rendimento. Pesquisa Mensal de Emprego. Rio de Janeiro, 2002. (Relatório Metodológico v. 23).

IBGE, Departamento de Estatísticas de População e Sociais. Manual do Entrevistador, Pesquisa Mensal de Emprego – 1980. Rio de Janeiro, 1980.

Lemos, S., 2002. "The Effects of the Minimum Wage on Wages and Employment in Brazil: a Menu of Minimum Wage Variables," Discussion Paper 02-02, Department of Economics, University College London.

Lopes, M. D., 2002. "Avaliação de Desgaste de Painéis em Estudos Longitudinais: Uma Aplicação na Pesquisa Mensal de Emprego (PME/IBGE)," ENCE, Rio de Janeiro (Dissertação de Mestrado, Orientadora: Denise Britz do Nascimento Silva).

Machado, A. F., R. P. Ribas and M. Penido, 2007. "Mobilidade entre estados de pobreza e inserção no mercado de trabalho: uma análise para o Brasil Metropolitano em 2004," Economia Aplicada 11 (2): 253-279.

MaCurdy, T.E., 1981. "An Empirical Model of Labor Supply in a Life Cycle Setting", Journal of Political Economy 89: 1059-85.

Nerlove, M., 2002. Essays in Panel Data Econometrics, Cambridge: Cambridge University Press.

Neri, M., D. Coelho, M. Ancora, A. Pinto, 1997. "Aspectos Dinâmicos do Desemprego e da Posição na Ocupação," Estudos Econômicos 27, n. especial: 137-159.

Penido, M. and A. F. Machado, 2003. "Duração do desemprego no Brasil Metropolitano" In: Wajnman, S. and A. F. Machado (eds.) Mercado de Trabalho: Uma análise a partir das pesquisas domiciliares no Brasil, Editora UFMG, pp. 203-218.

Peracchi, F. and F. Welch, 1995. "How representative are matched cross-sections? Evidence from the Current Population Survey," Journal of Econometrics 68 (1): 153-179.

Woltermann, S., 2002. "Job-Search Methods and Labor Market Transitions in a Segmented Economy: Some empirical evidence from Brazil," Discussion Paper 88, Ibero-America Institute for Economic Research (IAI), Georg-August-Universität Göttingen.

Agradecimentos

Os autores agradecem os comentários de Carlos Henrique Corseuil e Maurício Cortez Reis. Os erros remanescentes são de responsabilidade dos autores. Todas as sintaxes usadas neste trabalho encontram-se à disposição para todos que os queiram usar. Enviar pedidos para sergei.soares@ipea.gov.br ou rpribas.rs@gmail.com.

Abstract

The objective of this paper is to provide the necessary information so that others researches can use more properly the panel of the Brazilian Monthly Employment Survey (PME). The PME's panel is a quite useful instrument of analysis, but it has been underutilized due to several difficulties in putting it together. One of these difficulties is the absence of a key code which identifies the same person in different interviews. Then the matching, which must be done indirectly, is subject to measurement errors. As a consequence, the panel sample is considerably smaller than it could be. To minimize such a problem we provide an algorithm that improves the identification of people in the panel. This algorithm takes into account measurement errors in variables, such as birth date and education, and allows us to reduce 50% the attrition rate with respect to other conventional algorithms.

Apêndice – Sintaxe para reconstituição do painel da PME

```
*************************************************************
* Nota
*************************************************************

/* Este algoritmo pode ser aplicado tanto à Nova quanto à Antiga
   PME. As variáveis utilizadas a seguir são as da Nova pesquisa.
   Para utilizar o algoritmo com a Antiga PME basta substituir
   as seguintes variáveis:

   na Nova                                 na Antiga
   v035 = Região Metropolitana             = v010
   v040 = Número de controle               = v102
   v050 = Número de série                  = v103
   v060 = Painel                           = deve ser construída
   v063 = Grupo rotacional                 = v106
   v070 = Mês da pesquisa                  = v105
   v075 = Ano da pesquisa                  = deve ser construída
   v072 = Número da pesquisa no domicílio  = deve ser construída
   v201 = Número de ordem                  = v201
   v203 = Sexo                             = v202
   v204 = Dia de nascimento                = v206
   v214 = Mês de nascimento                = v236
   v224 = Ano de nascimento                = v246
   v234 = Idade calculada                  = v256
   v205 = Condição no domicílio            = v203
   vdae1= Anos de estudo I                 = deve ser construída

   Recomenda-se rodá-lo com arquivos pequenos. Para PME, a sugestão
   é de um arquivo por painel (variável v060) */


*************************************************************
* Variáveis do painel
*************************************************************

* Variável de identificação da pessoa no painel
g p201 = v201 if v072 == 1 /* definido com base na 1a entrevista */

* Variáveis que identificam o emparelhamento
g back = . /* com uma entrevista anterior */
g forw = . /* com uma entrevista posterior */


*************************************************************
* Emparelhamento - 1a loop
*************************************************************

* Emparelhamento para cada par de entrevista por vez
forvalues i = 1/7 {


   ***************************************************************
   * Emparelhamento padrão - se a data de nascimento está correta
   ***************************************************************

   * Ordenando cada indivíduo pelo mês de entrevista
   sort v035 v040 v050 v060 v063 v203 v204 v214 v224 v075 v070 v201

   * Loop para procurar a mesma pessoa em uma posição anterior
   loc j = 1    /* j determina a posição anterior na base */
   loc stop = 0 /* se stop=1, a loop para */
   loc count = 0
   while `stop' == 0 {
```

```
      loc lastcount = `count'
      count if p201 == . & v072 == `i'+1 /* observações não
                                            emparelhadas */
      loc count = r(N)
      if `count' == `lastcount' {

         * Parar caso a loop não esteja emparelhando mais
         loc stop = 1
      }
      else {
         if r(N) != 0 {

            * Captando a identificação p201 da observação anterior
            replace p201 = p201[_n - `j'] if /*
                              Identificação do domicílio
                           */ v035 == v035[_n - `j'] & ///
                              v040 == v040[_n - `j'] & ///
                              v050 == v050[_n - `j'] & ///
                              v060 == v060[_n - `j'] & ///
                              v063 == v063[_n - `j'] & /*
diferença entre períodos */   v072 == `i'+1 & v072[_n - `j'] == `i' /*
excluir emparelhados */       & p201 ==. & forw[_n - `j'] != 1 & /*
                              Características individuais
sexo */                       v203 == v203[_n - `j'] & /*
dia de nascimento */          v204 == v204[_n - `j'] & /*
mês de nascimento */          v214 == v214[_n - `j'] & /*
ano de nascimento */          v224 == v224[_n - `j'] & /*
informação observada */       v204!=99 & v214!=99 & v224!=9999

            * Identificação de emparelhamento para quem está a frente
            replace forw = 1 if v035 == v035[_n + `j'] & ///
                                v040 == v040[_n + `j'] & ///
                                v050 == v050[_n + `j'] & ///
                                v060 == v060[_n + `j'] & ///
                                v063 == v063[_n + `j'] & ///
                                p201 == p201[_n + `j'] & ///
                                v072 == `i' & v072[_n + `j']==`i'+1 ///
                                & forw != 1

            loc j = `j' + 1 /* passando para a próxima observação */
         }
         else {

            * Parar se não há observações para emparelhar
            loc stop = 1
         }
      }
   }

   * Recodificar variáveis de identificação do emparelhamento
   replace back = p201 !=. if v072 == `i'+1
   replace forw = 0 if forw != 1 & v072 == `i'


   ****************************************************************
   * Emparelhamento avançado
   ****************************************************************
   * Se sexo e ano de nascimento não estiverem corretos

   * Isolando observações já emparelhadas
   tempvar aux
   g `aux' = (forw==1 & (v072==1 | back==1)) | (back==1 & v072==8)

   * Ordenando cada indivíduo pelo mês de entrevista
   sort `aux' v035 v040 v050 v060 v063 v204 v214 v201 v075 v070

   * Loop para procurar a mesma pessoa em uma posição anterior
   loc j = 1    /* j determina a posição anterior na base */
```

```
   loc stop = 0 /* se stop=1, a loop para */
   loc count = 0
   while `stop' == 0 {
      loc lastcount = `count'
      count if p201 == . & v072 == `i'+1 /* observações não
                                            emparelhadas */
      loc count = r(N)
      if `count' == `lastcount' {

         * Parar caso a loop não esteja emparelhando mais
         loc stop = 1
      }
      else {
         if r(N) != 0 {

            * Captando a identificação p201 da observação anterior
            replace p201 = p201[_n - `j'] if /*
                              Identificação do domicílio
                           */ v035 == v035[_n - `j'] & ///
                              v040 == v040[_n - `j'] & ///
                              v050 == v050[_n - `j'] & ///
                              v060 == v060[_n - `j'] & ///
                              v063 == v063[_n - `j'] & /*
diferença entre períodos */   v072 == `i'+1 & v072[_n - `j'] == `i' /*
excluir emparelhados */       & p201 ==. & forw[_n - `j'] != 1 & /*
                              Características individuais
dia de nascimento */          v204 == v204[_n - `j'] & /*
mês de nascimento */          v214 == v214[_n - `j'] & /*
mesmo número de ordem */      v201 == v201[_n - `j'] & /*
informação observada */       v204!=99 & v214!=99

            * Identificação de emparelhamento para quem está à frente
            replace forw = 1 if v035 == v035[_n + `j'] & ///
                                v040 == v040[_n + `j'] & ///
                                v050 == v050[_n + `j'] & ///
                                v060 == v060[_n + `j'] & ///
                                v063 == v063[_n + `j'] & ///
                                p201 == p201[_n + `j'] & ///
                                v072 == `i' & v072[_n + `j']==`i'+1 ///
                                & forw != 1

            loc j = `j' + 1 /* passando para a próxima observação */
         }
         else {

            * Parar se não há observações para emparelhar
            loc stop = 1
         }
      }
   }


   *************************************************************
   * Emparelhamento avançado
   *************************************************************
   * Somente para chefes, cônjuges e filhos adultas

   tempvar ager aux

   * Função de erro na idade presumida
   g `ager' = cond(v234>=25 & v234<999, exp(v234/30), 2)

   * Isolando observações já emparelhadas
   g `aux' = (forw==1 & (v072==1 | back==1)) | (back==1 & v072==8)

   * Ordenando cada família pelo mês de entrevista
   sort `aux' v035 v040 v050 v060 v063 v203 v075 v070 v234 vdae1 v201
```

```
   * Loop para procurar a mesma pessoa em uma posição anterior
   loc j = 1
   loc stop = 0
   loc count = 0
   while `stop' == 0 {
      loc lastcount = `count'
      count if p201==. & v072==`i'+1 & ///
               (v205<=2 | (v205==3 & v234>=25 ///
               & v234<999))     /* observações não emparelhadas */
      loc count = r(N)
      if `count' == `lastcount' {
         loc stop = 1
      }
      else {
         if r(N) != 0 {
            replace p201 = p201[_n - `j'] if /*
                              Identificação do domicílio
                           */ v035 == v035[_n - `j'] & ///
                              v040 == v040[_n - `j'] & ///
                              v050 == v050[_n - `j'] & ///
                              v060 == v060[_n - `j'] & ///
                              v063 == v063[_n - `j'] & /*
diferença entre períodos */   v072 == `i'+1 & v072[_n - `j'] == `i' /*
excluir emparelhados */       & p201 ==. & forw[_n - `j'] != 1 & /*
                              Características individuais
sexo */                       v203 == v203[_n - `j'] & /*
diferença na idade */         abs(v234 - v234[_n - `j'])<=`ager' & /*
idade observada */            v234!=999 & /*
se chefe ou cônjuge */        ((v205<=2 & v205[_n - `j']<=2) | /*
ou filho com mais de 25 */    (v234>=25 & v234[_n - `j']>=25 & ///
                              v205==3 & v205[_n - `j']==3)) & /*
até 4 dias de erro na data */ ((abs(v204 - v204[_n - `j'])<=4 & /*
até 2 meses de erro na data*/ abs(v214 - v214[_n - `j'])<=2 & /*
informação observada */       v204!=99 & v214!=99) /*
ou */                         | /*
1 ciclo de erro na educação*/ (abs(vdae1 - vdae1[_n - `j'])<=1 /*
e */                          & /*
até 2 meses de erro na data*/ ((abs(v214 - v214[_n - `j'])<=2 & /*
informação observada */       v214!=99 & /*
informação não observada */   (v204==99 | v204[_n-`j']==99)) /*
ou */                         | /*
até 4 dias de erro na data */ (abs(v204 - v204[_n - `j'])<=4 & /*
informação observada */       v204!=99 & /*
informação não observada */   (v214==99 | v214[_n - `j']==99)) /*
ou */                         | /*
informações não-observadas */ ((v204==99 | v204[_n - `j']==99) & ///
                              (v214==99 | v214[_n - `j']==99)))))

            replace forw = 1 if v035 == v035[_n + `j'] & ///
                                v040 == v040[_n + `j'] & ///
                                v050 == v050[_n + `j'] & ///
                                v060 == v060[_n + `j'] & ///
                                v063 == v063[_n + `j'] & ///
                                p201 == p201[_n + `j'] & ///
                                v072 == `i' & v072[_n + `j']==`i'+1 ///
                                & forw != 1
            loc j = `j' + 1
         }
         else {
            loc stop = 1
         }
      }
   }
   replace back = p201 !=. if v072 == `i'+1
   replace forw = 0 if forw != 1 & v072 == `i'


   ************************************************************
```

```
   * Emparelhamento avançado
   ************************************************************
   * Somente em domicílio onde alguém já emparelhou

   * Quantas pessoas emparelharam no domicílio
   tempvar dom
   bys v075 v070 v035 v040 v050 v060 v063: egen `dom' = sum(back)

   * Loop com os critérios de emparelhamento
   foreach w in /*mesma idade*/ "0" /*erro na idade = 1*/ "1" /*
      erro na idade = 2*/ "2" /*erro na idade = f(idade)*/ "`ager'" /*
      2xf(idade)*/ "2*`ager' & v234>=25" {


      * Isolando observações já emparelhadas
      tempvar aux
      g `aux' = (forw==1 & (v072==1 | back==1)) | ///
                  (back==1 & v072==8) | (`dom'==0 & v072==`i'+1)

      sort `aux' v035 v040 v050 v060 v063 v203 v075 v070 v234 ///
            vdae1 v201
      loc j = 1
      loc stop = 0
      loc count = 0
      while `stop' == 0 {
         loc lastcount = `count'
         count if p201 == . & v072 == `i'+1 & `dom'>0 & `dom'!=.
         loc count = r(N)
         if `count' == `lastcount' {
            loc stop = 1
         }
         else {
            if r(N) != 0 {
               replace p201 = p201[_n - `j'] if /*
                                  Identificação do domicílio
                                */ v035 == v035[_n - `j'] & ///
                                   v040 == v040[_n - `j'] & ///
                                   v050 == v050[_n - `j'] & ///
                                   v060 == v060[_n - `j'] & ///
                                   v063 == v063[_n - `j'] & /*
diferença entre períodos */        v072 == `i'+1 & v072[_n-`j'] == `i' /*
excluir emparelhados */            & p201 ==. & forw[_n - `j'] != 1 & /*
há emparelhados no domicilio*/     `dom' > 0 & `dom'!=. & /*
                                   Características individuais
sexo */                            v203 == v203[_n - `j'] & /*
critérios mudam com a loop */      ((abs(v234-v234[_n - `j'])<=`w' & /*
se a idade é observada */          v234!=999) /*
caso contrário */                  | /*
mesma escolaridade */              (vdae1==vdae1[_n - `j'] & /*
mesma condição no domicílio */     v205==v205[_n - `j'] & /*
idade não observada */             (v234==999 | v234[_n - `j']==999)))

               replace forw = 1 if v035 == v035[_n + `j'] & ///
                                   v040 == v040[_n + `j'] & ///
                                   v050 == v050[_n + `j'] & ///
                                   v060 == v060[_n + `j'] & ///
                                   v063 == v063[_n + `j'] & ///
                                   p201 == p201[_n + `j'] & ///
                                   v072 ==`i' & v072[_n+`j']==`i'+1 ///
                                   & forw != 1
               loc j = `j' + 1
            }
            else {
               loc stop = 1
            }
         }
      }
   }
```

```
   replace back = p201 !=. if v072 == `i'+1
   replace forw = 0 if forw != 1 & v072 == `i'

   * Identificação para quem estava ausente na última entrevista
   replace p201 = `i'00 + v201 if p201 == . & v072 == `i'+1
}


*************************************************************
* Recuperar quem saiu e retornou para o painel - 2a loop
*************************************************************

* Variável temporária identificando o emparelhamento à frente
tempvar fill
g `fill' = forw

* Loop retrospectivo por entrevista
foreach i in 7 6 5 4 3 2 1 {
   tempvar ncode1 ncode2 aux max ager

   * Função de erro na idade presumida
   g `ager' = cond(v234>=25 & v234<999, exp(v234/30), 2)

   * Variável que preserva o antigo número
   bys v035 v040 v050 v060 v063 p201: g `ncode1' = p201

   * Isolando observações emparelhadas
   g `aux' = ((`fill'==1 & (v072==1 | back==1)) | (back==1 & v072==8))

   * Variável identificando a última entrevista
   bys v035 v040 v050 v060 v063 p201: egen `max' = max(v072)

   sort `aux' v035 v040 v050 v060 v063 v203 v072 v201 p201
   loc j = 1
   loc stop = 0
   loc count = 0
   while `stop' == 0 {
      loc lastcount = `count'
      count if p201>`i'00 & p201<`i'99 & back==0
      loc count = r(N)
      if `count' == `lastcount' {
         loc stop = 1
      }
      else {
         if r(N) != 0 {
            replace p201 = p201[_n - `j'] if /*
                           Identificação do domicílio
                        */ v035 == v035[_n - `j'] & ///
                           v040 == v040[_n - `j'] & ///
                           v050 == v050[_n - `j'] & ///
                           v060 == v060[_n - `j'] & ///
                           v063 == v063[_n - `j'] & /*
quem entrou na entrevista i*/ p201>`i'00 & p201<`i'99 & /*
não emparelhado */            back==0 & `fill'[_n - `j']!=1 & /*
uma entrev. de diferença */   `max'[_n - `j']<`i' & ///
                              p201[_n - `j']<`i'00-100 & /*
sexo */                       v203 == v203[_n - `j'] & /*
diferença na idade */         ((abs(v234 - v234[_n - `j'])<=`ager' & /*
idade observada */            v234!=999 & /*
até 4 dias de erro na data */ ((abs(v204 - v204[_n - `j'])<=4 & /*
até 2 meses de erro na data*/ abs(v214 - v214[_n - `j'])<=2 & /*
informação observada */       v204!=99 & v214!=99) /*
ou */                         | /*
1 ciclo de erro na educação*/ (abs(vdae1 - vdae1[_n - `j'])<=1 /*
e */                          & /*
até 2 meses de erro na data*/ ((abs(v214 - v214[_n - `j'])<=2 & /*
informação observada */       v214!=99 & /*
informação não observada */   (v204==99 | v204[_n - `j']==99)) /*
```

```
ou */                          | /*
até 4 dias de erro na data */ (abs(v204 - v204[_n - `j'])<=4 & /*
informação observada */       v204!=99 & /*
informação não observada */   (v214==99 | v214[_n - `j']==99)) /*
ou */                          | /*
nada é observado */            ((v204==99 | v204[_n - `j']==99) & ///
                               (v214==99 | v214[_n - `j']==99)))))) /*
ou */                          | /*
mesma escolaridade */          (vdae1==vdae1[_n - `j'] & /*
e número de ordem */           v205==v205[_n - `j'] /*
se idade não é observada */    & (v234==999 | v234[_n - `j']==999)))

            * Identificação de emparelhamento para quem está à frente
            replace `fill' = 1 if v035 == v035[_n + `j'] & ///
                                  v040 == v040[_n + `j'] & ///
                                  v050 == v050[_n + `j'] & ///
                                  v060 == v060[_n + `j'] & ///
                                  v063 == v063[_n + `j'] & ///
                                  p201 == p201[_n + `j'] & ///
                                  `fill' == 0 & `max'<`i' & ///
                                  (v072[_n + `j'] - v072)>=2
            loc j = `j' + 1
         }
         else {
            loc stop = 1
         }
      }
   }

   * Igualando o número de quem era igual
   bys v035 v040 v050 v060 v063 `ncode1': egen `ncode2' = min(p201)
   replace p201 = `ncode2'
}

*************************************************************
* Fim do arquivo
*************************************************************
```

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

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

Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

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

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

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

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

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

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

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

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

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

Revista Brasileira de Estatística – RBEs
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - IBGE
Rua André Cavalcanti, 106, sala 111
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tels.: 55 21 2142-4682 (Sandra Cavalcanti Barros – Secretária)
55 21 2142-4686 (Ismenia Blavatsky – Editor–Executivo)
Fax: 55 21 2142-0501

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Podem ser submetidos originais processados pelo editor de texto *Word for Windows* ou originais processados em LaTeX (ou equivalente) desde que estes últimos sejam encaminhados e acompanhados de versões em pdf, conforme descrito no item 3, a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, se for o caso de constarem no documento, também devem figurar nesta página;

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

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

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

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

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

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

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

10. Não serão permitidos itens que identifiquem os autores do artigo dentro do texto, tais como: número de projetos de órgãos de fomento, endereço, *e-mail*, etc. Caso ocorra, a responsabilidade será inteiramente dos autores; e

11. No caso de o artigo ser aceito para a publicação após a avaliação dos pareceristas, serão encaminhadas as sugestões/comentários aos autores sem a sua identificação. Uma vez nesta condição, é de responsabilidade única dos autores fazer o *download* da formatação padrão da revista (em doc ou em LaTeX) para o envio da versão corrigida.