Documentos para Disseminação

Memória Institucional

22

40 Anos da Unidade de Métodos Estatísticos do IBGE

Alguns Passos

IBGE
Instituto Brasileiro de Geografia e Estatística

Presidente da República
Michel Miguel Elias Temer Lulia

Ministro do Planejamento, Orçamento e Gestão
Dyogo Henrique de Oliveira

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Roberto Luís Olinto Ramos

Diretor-Executivo
Fernando J. Abrantes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Claudio Dutra Crespo

Diretoria de Geociências
Wadih João Scandar Neto

Diretoria de Informática
José Sant`Anna Bevilaqua

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Maysa do Sacramento de Magalhães

UNIDADE RESPONSÁVEL

Diretoria de Pesquisas

Coordenação de Métodos e Qualidade
Sonia Albieri

Ministério do Planejamento, Desenvolvimento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE
Diretoria de Pesquisas
Coordenação de Métodos e Qualidade

Documentos para Disseminação

Memória Institucional 22

40 Anos da Unidade de Métodos Estatísticos do IBGE

Alguns passos

Rio de Janeiro
2017

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

ISSN 0103-6335 **Documentos para Disseminação**
ISSN 0103-6459 **Memória institucional**
Divulga textos sobre aspectos históricos do IBGE e/ou de seus estudos e pesquisas bem como sobre personalidades que contribuíram para a história do Instituto.

ISBN 978-85-240-4430-4

Esse volume foi organizado por Sonia Albieri e Antonio José Ribeiro Dias, servidores do IBGE lotados na Coordenação de Métodos e Qualidade da Diretoria de Pesquisas - DPE.

Produção do *e-book*
Roberto Cavararo

Capa
Marcos Balster Fiore
Ilustração
Aldo Victorio Filho e Fabio Muniz de Moura
Coordenação de *Marketing* e Gerência de Editoração/ Centro de Documentação e Disseminação de Informações -CDDI

40 anos da unidade de métodos estatísticos do IBGE : alguns passos / [organizado por Sonia Albieri e Antonio José Ribeiro Dias]. - Rio de Janeiro : IBGE, Coordenação de Métodos e Qualidade ; 2017.
216p. : il - (Documentos para disseminação. Memória institucional, ISSN 0103-6459 ; 22)

Inclui bibliografia.
ISBN 978-85-240-4430-4

1. Estatística – Metodologia - História. 2. Controle de qualidade – Métodos estatísticos. 3. IBGE - História. 4. Pesquisa histórica. 5. Memória. I. Albieri, Sonia . II. Dias, Antonio José Ribeiro. III. IBGE. Coordenação de Métodos e Qualidade. IV. Título: Quarenta anos da unidade de métodos estatísticos do IBGE : alguns passos. V. Série.

Gerência de Biblioteca e Acervos Especiais CDU311(093)
RJ/IBGE/2017-21 IBGE

Impresso no Brasil/*Printed in Brazil*

Sumário

Apresentação

Com o presente volume, a série Memória Institucional, divulgada pelo Instituto Brasileiro de Geografia e Estatística - IBGE, comemora os 40 anos de existência da unidade de metodologia estatística do Instituto, cuja atuação, ao longo desse período, trouxe significativos aprimoramentos e avanços, tanto no que se refere aos métodos estatísticos e sistemas computacionais utilizados para a produção, análise e disseminação, como no que diz respeito à qualidade estatística dos dados.

A área de métodos estatísticos, cabe ressaltar, é tão ampla que delimitar uma relação de temas de interesse para este livro comemorativo seria por demais ambicioso. Tradicionalmente, e de uma forma geral, a unidade de metodologia estatística do IBGE vem trabalhando com métodos e sistemas voltados para amostragem, crítica, imputação e outros métodos de tratamento de não resposta, análise de dados, avaliação da qualidade, sigilo de informações estatísticas, análise de séries temporais, modelos estatísticos e análise espacial. Nessa perspectiva, foram convidados, então, pesquisadores da área que, aceitando participar da empreitada comemorativa desses 40 anos, escreveram especialmente para o presente livro, mesmo que o assunto já tenha sido objeto de divulgação em outro formato. Vale destacar que todos os convidados foram ou são servidores da Coordenação de Métodos e Qualidade, participam ou participaram das atividades da unidade desde sua criação, e que os capítulos por eles elaborados abordam as temáticas da metodologia de produção, análise e disseminação de informações estatísticas, no âmbito das atividades do IBGE.

Cláudio Dutra Crespo

Diretor de Pesquisas

Prefácio

As contínuas e diversificadas atividades, exigidas pelo sistema estatístico, requerem constantes aprimoramentos metodológicos.

A tecnologia em geral, as possibilidades de sofisticados tratamentos de dados, e os avanços da teoria e dos métodos estatísticos — da amostragem aos modelos afetos à inteligência artificial — determinam e condicionam o fazer das pesquisas, do empolgante esforço censitário a um simples inquérito administrativo.

A dimensão e complexidade do Brasil, as necessidades cada vez mais prementes de informação, seja por parte dos gestores públicos, seja por parte das comunidades e cidadãos, não permitem que o ciclo anual destas atividades se conduza alheio à dinâmica e preceitos da boa metodologia.

Tenho o orgulho de haver iniciado, de forma moderna e sistemática, a área de métodos, sem medo de incorporar as inovações então existentes, mudando procedimentos, critérios e planos amostrais, sempre guiado pela busca de mais rigor e melhor qualidade, ao lado de ganhos em custo e tempo de obtenção dos resultados finais.

Tenho a emoção de prefaciar esse livro, décadas após, testemunho inconteste de que o esforço vingou. Identifico entre os vários autores, muitos que a mim se aliaram; em quem acreditei, contratei, estimulei e comigo trabalharam pelas necessárias mudanças.

O zelo pela metodologia, a fé no seu poder transformador e a mentalidade de contínuo aperfeiçoamento técnico estão implícitos em todos os capítulos.

Isso é o mais importante dentre tudo o que arduamente plantei.

Poderia também dizer que, para mim, é surpreendente que em nosso país, em uma instituição federal, tal espírito de excelência profissional, de visão de longo prazo, de preocupação com a pesquisa e o desenvolvimento tenham perdurado e florescido durante todos esses anos.

Recuso a visão negativista, onde tudo o que é bom é exceção. O mérito desse inegável sucesso é de todos, todos os que começaram comigo, os que depois se agregaram, os que colaboraram lateralmente, das sucessivas gerações de funcionários que acreditaram e acreditam na importância da metodologia.

O mérito é da Instituição, nosso IBGE, tão querido e por vezes tão maltratado, que sobreviveu a cortes orçamentários, crises as mais diversas,

administrações mais e outras menos esclarecidas, à incompetência e ignorância não de muitos, mas de uns poucos que, por vezes, ocuparam posições onde dano causaram.

A Instituição, entretanto, segue. O livro é símbolo da imensa maioria, desde jovens técnicos com poucos anos de casa a analistas calejados, que defende, preserva e exerce as boas práticas.

É isso que faz um país avançar, é isso que ajuda a desenvolver a economia e a sociedade, ao torná-las mais conhecidas e analisadas; bem mensuradas, enfim.

Encerro não com uma alegria juvenil do "dever cumprido", embora com parte da alma em paz, ao ver os progressos da obra.

Encerro, mais do que tudo, com palavras de ânimo e alerta. O trabalho não tem fim, responsabilidade não é uma brisa passageira.

Cabe aos que estão nesse livro, e a todos os outros que com eles cooperam, ou na área labutam, o dever de manter vivo esse espírito.

De lutar sempre por ser um profissional exemplar, mais informado dos avanços da estatística, da inteligência artificial e dos métodos de pesquisa em geral. Ouvir a voz dos usuários e dos formuladores do conteúdo da investigação. Dialogar, refletir, planejar e, sobretudo, jamais ter medo de inovar.

Mais do que uma peça de especialistas, esse livro é um registro do que funcionários públicos brasileiros podem realizar e, ao mesmo tempo, um marco de alerta às gerações futuras, de que mais e melhor há que ser feito.

Pelo IBGE, pelo povo brasileiro.

Renato Galvão Flôres Junior

Breves considerações

Em maio de 1977, foi criada a primeira unidade formal na estrutura organizacional do IBGE destinada aos estudos de métodos. Era um departamento, o Departamento de Coordenação de Métodos, subordinado à Superintendência de Estatísticas Primárias, da Diretoria Técnica, responsável, à época, pela implantação e gerência das diversas pesquisas para coleta de informações estatísticas.

O responsável pela idealização, criação e implantação do Departamento de Coordenação de Métodos foi Renato Galvão Flôres Junior, que nos deu a honra de prefaciar este livro. Após algum tempo à frente do Departamento, Renato assumiu a Superintendência de Estatísticas Primárias e, em 1981, convidou José Paulo Quinhões Carneiro, Professor de Matemática da Escola Nacional de Ciências Estatísticas - ENCE e da Universidade Federal do Rio de Janeiro - UFRJ, para assumir a chefia do Departamento. A parceria desses dois pioneiros foi responsável pela sedimentação da área de metodologia estatística do IBGE nos cinco anos em que José Paulo, que também nos brinda com um depoimento (apresentado a seguir) sobre esse período, esteve à frente da unidade.

Ao longo dos 40 anos de existência, a unidade de metodologia do IBGE teve diferentes denominações[1] e subordinações administrativas, essas últimas aqui entendidas como as localizações na estrutura organizacional da Instituição, além de distintas atribuições, definidas no momento de sua criação e dependentes da subordinação administrativa. Uma alteração importante foi feita em 2007, ampliando as áreas de atuação da unidade, e, consequentemente, mudando o seu nome para Coordenação de Métodos e Qualidade, subordinando-a à Diretoria de Pesquisas.

Este livro foi planejado para comemorar os 40 anos dessa unidade de metodologia completados em 2017, por meio da publicação de uma coletânea de capítulos relacionados com os temas que foram e vêm sendo objeto dos trabalhos realizados na área. O livro tem, também, o objetivo de documentar, não exaustivamente, e promover a divulgação de trabalhos realizados e avanços ocorridos no IBGE no que se refere a métodos estatísticos e sistemas

[1] Departamento de Coordenação de Métodos, Coordenadoria de Avaliação e Metodologia, Coordenadoria de Metodologia, Núcleo de Metodologia, Divisão de Metodologia, Departamento de Metodologia, Coordenação de Métodos e Qualidade.

computacionais para a produção, análise e disseminação de dados obtidos por meio de pesquisas amostrais, censos e registros administrativos, bem como aos avanços recentes na área da qualidade estatística. Dada a diversidade de temas metodológicos afetos à produção de informações estatísticas, vale destacar que seria muito ambicioso tentar abranger todas as áreas de estudo em que os técnicos da unidade estiveram envolvidos nesse período.

Finalmente, gostaríamos de, mais uma vez, ressaltar a importância de Renato Flôres e José Paulo Carneiro na criação e implantação da área de metodologia estatística do IBGE e agradecer-lhes por suas iniciativas. Este livro é, sobretudo, uma homenagem a eles e a muitos outros técnicos e pesquisadores que passaram pelo então Departamento de Coordenação de Métodos e suas outras denominações até chegar à atual Coordenação de Métodos e Qualidade, ou colaboraram de alguma forma, nesses 40 anos, para o avanço metodológico em nossa Instituição.

Sonia Albieri e Antonio José Ribeiro Dias

Depoimento de José Paulo Quinhões Carneiro

A colaboração de José Paulo Quinhões Carneiro à elaboração do presente registro comemorativo veio por meio da resposta a apenas uma pergunta sobre sua fundamental participação na implantação da área de metodologia estatística do IBGE. Reproduzimos, a seguir, a pergunta e a respectiva resposta.

Pergunta: O Decme [Departamento de Coordenação de Métodos] foi criado em 1977 e você assumiu a chefia do Departamento em 1981. Como foi assumir essa área nova ligada à produção de estatísticas primárias? Que pontos você destaca no seu trabalho frente à área de metodologia nesses 5 anos?

Resposta: Assumi a área com muito entusiasmo, muita curiosidade diante do desafio colocado, muito estímulo e apoio do Renato Flôres, o criador dessa área no IBGE e, finalmente, contei com uma equipe muito competente e motivada.

Um dos pontos que mais destaco na atuação do meu grupo foi, em primeiro lugar, a valorização da transparência, da qual cito alguns exemplos:

1. Na amostra do Censo Demográfico de 1980, pela primeira vez o IBGE publicou estimativas dos erros amostrais. A partir daí, tal procedimento tornou-se um hábito nosso em todas as pesquisas por amostragem do IBGE. Quero mesmo crer que, neste ponto, o IBGE influenciou toda a comunidade – como convém ao órgão máximo do Sistema Estatístico. Até as pesquisas eleitorais, que jamais sequer mencionavam erros amostrais, passaram a fazê-lo e hoje em dia esta é uma saudável rotina difundida nos meios de comunicação.

2. Com a coleção Relatórios Metodológicos, cujo primeiro volume foi sobre a Pesquisa Nacional por Amostra de Domicílios - PNAD [na década de 70], iniciamos o procedimento de publicar as metodologias das pesquisas do IBGE, desvendando cada vez mais um certo mistério que as cercava.

3. Os esboços geográficos dos setores demográficos deixaram de ser considerados "segredos militares" e passaram a ser acessíveis a todos os pesquisadores.

Um segundo ponto, não menos importante, foi o gradativo, mas consistente, aumento do número de pesquisas por amostragem, iniciando pela Pesquisa Industrial Anual - PIA, em 1981.

Finalmente, chamo a atenção para a Pesquisa de Orçamentos Familiares - POF, pressuposto básico para as ponderações dos índices de preços. Por motivos alheios à vontade dos técnicos do IBGE, uma pesquisa desse tipo só tinha sido feita em 1974, o Estudo Nacional da Despesa Familiar - ENDEF. No início dos anos 80, os planejadores governamentais ficavam inconformados de não poderem incluir no Índice de Preços, por exemplo, o álcool combustível, cuja importância então era clara para todo mundo. Mas incluí-lo com que ponderação? Só em 1986 o IBGE conseguiu a verba necessária para fazer uma segunda POF, cuja amostra foi planejada e coordenada pela nossa área de Metodologia. A partir de então, uma POF mais frequente passou a ser um sonho de consumo do IBGE, só não realizado quando não há recursos. Note-se também que já em 1986, desenvolvemos estudos para uma POF contínua, que já existia em outros países. Tenho certeza de que o IBGE vai chegar logo lá.

Os métodos de pesquisa no IBGE: um breve resumo dos últimos 50 anos

Maurício Teixeira Leite de Vasconcellos

Introdução

Foi com imenso prazer que aceitei o convite para produzir um capítulo para o livro comemorativo dos 40 anos da unidade de metodologia do Instituto Brasileiro de Geografia e Estatística - IBGE, apesar das dificuldades antevistas para a tarefa.

Lembro-me de uma conversa com Nelson de Castro Senra sobre as dificuldades que ele apontava para escrever sobre a história recente do IBGE, história que viveu no dia a dia da Instituição e da qual foi um dos atores. De fato, não é fácil fazer uma análise histórica imparcial e isenta de um período do qual participamos. Neste sentido, deixo claro desde as primeiras linhas deste texto que não tenho qualquer pretensão de ser totalmente isento nessa descrição, mas que descreverei o que vivi e o que era discutido sobre o assunto desde que entrei no IBGE no início dos anos 1970.

Outra dificuldade reside na vastidão do tema. Métodos de pesquisa é um tema que abrange muito mais do que os métodos de seleção de amostras de uma população de pesquisa.

Método de pesquisa abrange métodos e técnicas que começam pela apropriação de conhecimento do quadro referencial teórico de cada área a pesquisar (o que define o que levantar e como fazê-lo); pelos métodos de coleta (entrevista direta, por telefone, por troca de arquivos magnéticos, por consulta a sistemas via Internet, etc.); pela forma de perguntar e pela ordenação das perguntas ou consultas; pelos métodos operacionais e administrativos, pelos métodos de avaliação e controle da informação obtida; pelos métodos de crítica de consistência dos dados levantados; pelos métodos de imputação e tratamento de não respostas; pelos métodos de seleção de amostras e geração dos pesos amostrais necessários à produção das estimativas; pelos métodos de construção de variáveis derivadas e indicadores; pelos métodos de análise da informação obtida; dentre outros.

Nesse sentido, os métodos estão distribuídos em todas as unidades do IBGE e, como bem escreveu Sonia Albieri, a unidade de metodologia do IBGE está intimamente ligada à evolução do uso de amostragem na Instituição (ALBIERI, 2003). No entanto, essa área não se limita a desenhar e expandir amostras. Tem contribuições em diferentes aspectos do método de pesquisa

estatística quantitativa, como crítica e imputação de dados; sistemas computacionais para produção de pesquisas; disseminação de métodos das pesquisas do IBGE e formas de disseminar dados; capacitação profissional (como professores da Escola Nacional de Ciências Estatísticas - ENCE, em cursos de treinamento e no Curso de Desenvolvimento de Habilidades em Pesquisa - CDHP), dentre outros. O texto de Albieri (2003) apresenta essas ideias de forma mais pormenorizada.

Assim, vou abordar uma parte vivida da discussão de métodos de pesquisa no IBGE das décadas de 1960-1970, a criação da área de métodos da então Diretoria Técnica do IBGE e finalizar trazendo o momento atual dessas questões no IBGE. Propositalmente, não tratarei dos métodos mais estatísticos, que serão abordados em outros capítulos deste livro.

A discussão de métodos nos anos 1960-1970

Entrei no IBGE para participar de um grupo ligado à presidência que tinha como objetivo principal definir e planejar uma pesquisa de consumo alimentar e orçamentos familiares (que depois recebeu o nome de Estudo Nacional da Despesa Familiar - ENDEF por sugestão da empresa que foi contratada para fazer a propaganda da pesquisa). Nesse período, a Instituição, presidida pelo Professor Isaac Kerstenetzky, desde março de 1970, passava por grandes transformações que visavam a transformar a antiga autarquia IBGE na nova fundação, criada pelo Decreto-Lei n. 161, de 13.02.1967.

Como indicou com extrema propriedade Senra (2006), em capítulo do livro *Isaac Kerstenetzky: legado e perfil*, publicado pelo IBGE, o Professor Isaac traz sua visão do método que considerava adequado para nortear um sistema estatístico. Seu método era baseado nas estatísticas derivadas com seu quadro referencial de conhecimentos, e sua elaboração demandava a produção de estatísticas primárias, sendo, portanto, a norteadora da estrutura da produção de estatísticas primárias. E, ao assumir a presidência do IBGE colocou em prática esse seu pensamento, que foi apresentado na I Conferência Nacional de Estatística - CONFEST, em 1968, já então mais amadurecido. Como descreveu Maristela Afonso de André Sant'Anna, em outro capítulo intitulado *Professor Isaac:*

> Em 1970, como presidente do IBGE, o professor Isaac me propôs compor sua equipe, na realização de um ambicioso sonho dele: o projeto de disponibilizar ao País modelos macroeconômicos de desenvolvimento econômico-social, integradores das dimensões global e setorial, da nacional e regional, da econômica e da social. O passo inicial consistia na reconstrução do Sistema de Contabilidade Social do País, ao qual o professor havia dedicado anos preciosos de sua vida profissional, na Fundação Getúlio Vargas, e que seria integrado com sistemas de indicadores e de estudos sociais, a serem criados.
>
> Segundo a sua concepção, o projeto desencadeador, que daria consistência a esse novo Sistema de Contabilidade Social (o *bench-mark*), consistia na construção quinquenal de uma matriz nacional de relações intersetoriais de produção de bens e serviços, então conhecida como Matriz de *Input-Output* da economia brasileira, sendo a primeira referente ao ano censitário de 1970. A partir dela, caberia redefinir a Conta de Produção do Sistema de Contas Nacionais e orientar a reconstrução das demais Contas, fossem de periodicidade anual como trimestral, no plano nacional como no regional. Seriam desenvolvidas pesquisas econômicas e sociais para ampliar o conhecimento da apropriação da renda pelas famílias, pelo setor público, na formação de capital dos vários agentes econômicos. O arcabouço inicial viria a dar estímulo a uma profunda renovação nos indicadores, os de produção industrial, serviços, agropecuária, emprego, salários, preços e outros. O sonho do professor se estendia à criação de amplo projeto de indicadores sociais de qualidade de vida, de distribuição de renda, de estrutura de consumo das famílias, de oportunidades de ascensão social, de condições de vida dos vários estratos da população, nas várias regiões do País.
>
> O Sistema Nacional de Estatísticas Primárias viu-se submetido a uma demanda hercúlea, em extensão, em qualidade, em consistência conceitual e metodológica. Emergiu demanda irresistível por ampliar o sistema e integrar as estatísticas primárias do IBGE, econômicas, sociais, demográficas e geográficas, por se integrar com outras fontes estatísticas, em especial com registros administrativos, como os de comércio exterior, construção civil e outros, gerados em órgãos governamentais e privados.

> Foram criados novos levantamentos no âmbito do IBGE, em âmbitos como o de órgãos do governo, serviços, construção civil, agropecuária. Foi absorvido o sistema de índices de preços do Ministério do Trabalho e reconstruído, cada área metropolitana passou a ter os seus índices, com conceituação, metodologia e confiabilidade até então desconhecida no País, pautada em experiências bem sucedidas em países avançados nessa área (SANT'ANNA, 2006, p. 180)

No mesmo livro, Jane Souto de Oliveira descreve sua experiência daqueles tempos:

> Assim, ao mesmo em tempo em que adequava o IBGE às novas tecnologias de informação, criando o Instituto de Informática e agilizando o processo de coleta, sistematização e disseminação das informações estatísticas, Prof. Isaac nos desafiava, a todos, com textos que indagavam: Is it time to stop counting? Sua abertura intelectual favoreceu abordagens interdisciplinares, impulsionou a combinação de métodos quantitativos e qualitativos na rotina de trabalho institucional, seu compromisso ético levou a que temas como pobreza e desigualdade fossem priorizados no âmbito dos estudos e levantamentos do órgão.
>
> Difícil expressar o significado dessa inovação transformadora, produzida por Prof. Isaac ao longo dos dez anos em que esteve à frente do IBGE. Ainda que a mim tenha cabido focalizar sua contribuição no que se refere mais propriamente ao social, penso que ele seria o primeiro a questionar uma demarcação rígida de fronteiras entre o social, o econômico e o geográfico. E isto porque foi fundamentalmente uma concepção integrada de sistema de informação, o que sempre pontuou o pensamento e a prática de Prof. Isaac como mestre, pesquisador e condutor de instituições.
>
> Dessa visão integrada e integradora, testemunha a criação, praticamente em simultâneo, do projeto da Matriz de Relações Intersetoriais, embrião do projeto de Contas Nacionais, do projeto de Indicadores Sociais, bem como a implementação do Estudo Nacional da Despesa Familiar - ENDEF (OLIVEIRA, 2006, p. 183-184).

De acordo com a tese de doutorado de Maurício Teixeira Leite de Vasconcellos:

> [...] a sua [Isaac Kerstenetzky] opção de constituir dois grupos para desenvolver os projetos [da Matriz e dos Indicadores Sociais] e de ligá-los diretamente à Presidência do IBGE tinha objetivos bem delineados: (a) integrar as informações econômicas com as sociais, gerando um sistema de contabilidade social que recuperasse o pensamento social dos economistas clássicos; (b) utilizar esse sistema de contabilidade social como norteador e definidor do âmbito e abrangência das pesquisas realizadas pelo Sistema Estatístico Nacional; e (c) utilizar o quadro referencial teórico do Sistema de Contas Nacionais para integrar as informações de consumidores (pesquisas domiciliares e censo demográfico) com as de produtores (censos e pesquisas agropecuárias, da indústria, do comércio e dos serviços) (VASCONCELLOS, 2001, p. 31).

Ainda segundo Oliveira (2006, p. 184), "Complementando e subsidiando estes dois projetos, um terceiro – o ENDEF – acoplava o estudo nutricional ao levantamento dos orçamentos familiares e das condições de reprodução social da população brasileira".

Essa visão de organização do Sistema Estatístico Nacional, usando as necessidades das estatísticas derivadas como definidoras do conjunto de pesquisas e levantamentos primários, foi implantada no IBGE por meio dos projetos da matriz (matriz de relações intersetoriais, desagregação da conta de produção das contas nacionais por setores de produção e consumo) e do sistema de indicadores sociais, que foram mantidos junto à presidência do IBGE até que estivessem estruturados, amadurecidos e com as pesquisas primárias que necessitavam para fornecer os dados para suas elaborações. Dessa forma, o presidente poderia acompanhar de perto o desenvolvimento desses projetos (o que, aliás, era uma característica pessoal do pesquisador que tinha que presidir a Instituição) e assegurar os meios (inclusive para as pesquisas e levantamentos primários) para sua consecução. Só depois foram criadas as áreas (ou departamentos) do IBGE responsáveis pelas contas nacionais e pelos indicadores sociais.

O ENDEF, concebido para prover informações para o projeto da matriz e dos indicadores sociais, foi exemplo da busca incansável do Professor Isaac por um método de pesquisa que pudesse atender aos objetivos dessas duas áreas. Antes de decidir a linha de pesquisa a adotar no ENDEF, ele estudou e avaliou três linhas de pesquisas de orçamentos familiares: (1) a norte-americana, nascida nas pesquisas de saúde e que incorporou os orçamentos familiares, que já conhecia por ter sido o diretor técnico da Fundação Getúlio Vargas - FGV, quando da realização das quatro pesquisas de orçamentos familiares da década de 1960 (FUNDAÇÃO GETÚLIO VARGAS, 1967); (2) a inglesa, de origem mais econômica e que foi introduzida no IBGE pelas Pesquisas de Orçamentos Familiares a partir da 1987-1988; e (3) a francesa, que nasceu das pesquisas

de consumo de alimentos realizadas na África por diferentes pesquisadores franceses a partir dos anos 1960, que incluiu os orçamentos familiares para explicar as decisões familiares relacionadas ao consumo de alimentos. A linha francesa foi escolhida por ser a única que permitia um maior aprofundamento nas questões sociais e no conhecimento das reais condições nutricionais da população brasileira à época.

E não foi simples realizar o Endef dentro do programa de pesquisas domiciliares do IBGE. Inúmeras eram as resistências por um questionário semiaberto, em vez do questionário norte-americano proposto pelos consultores que apoiavam o desenvolvimento da Pesquisa Nacional por Amostra de Domicílios - PNAD à época. E não era só o questionário, era preciso alterar o plano de amostragem para selecionar dois domicílios vizinhos e atender à necessidade de deslocamento rápido entre os domicílios selecionados para poder pesar os alimentos. Além disso, realizar entrevistas domiciliares por sete dias consecutivos, com pesagem de alimentos três a quatro vezes por dia, parecia impossível para muitos técnicos da Instituição: falava-se em 100% de recusa e ao final o Endef teve apenas 1,6% de recusa em nível nacional, sendo 15,3% delas no Distrito Federal.

Para eliminar resistências e fazer valer a decisão tomada, o Professor Isaac fez uma reunião com todos os envolvidos no Endef (os do grupo Endef, os do Grupo Executivo de Pesquisas Domiciliares, que mantinha a PNAD, todos os delegados de estatística dos estados, territórios e Distrito Federal e todos os envolvidos das demais áreas do IBGE) e mostrou que cerca de 70% da produção nacional à época era destinada ao consumo, uma variável estimada por resíduo na conta de produção do sistema de contas nacionais. Complementou que só isto mostrava a importância do Endef para as estatísticas nacionais e que contava com o apoio de todos, mas entendia que quem não colaborasse com o Endef estava pedindo sua demissão no IBGE e que, com tristeza, acataria a decisão de cada um. Para mim, acostumado com a gentileza do dia-a-dia do presidente, foi um choque ver como colocou de maneira absolutamente clara sua decisão. Em seguida, os delegados foram chamados, um a um na ordem geográfica das Unidades da Federação, para colocarem suas necessidades materiais para a realização do Endef. E todos tiveram suas demandas atendidas, como forma complementar de mostrar a importância do projeto para a Instituição.

Depois de realizada a pesquisa, com seus dados parcialmente divulgados, o grupo Endef saiu da presidência e transformou-se no Departamento de Estudos do Consumo (Resolução n. 06, de 24.04.1978, da Presidência do IBGE), da Superintendência de Estudos Geográficos e Socioeconômicos, à qual já estavam ligados o Departamento de Contas Nacionais e o Departamento de Indicadores Sociais, dentre outros.

Desnecessário dizer que a discussão de métodos para definição do quadro referencial das contas nacionais e dos indicadores sociais, assim como os métodos de avaliação nutricional a partir da ingestão de alimentos e da antropometria eram alvo de discussão nos três projetos ligados à presidência do IBGE e continuaram nos correspondentes departamentos da atual Diretoria de Pesquisas.

Além da discussão sobre o método a orientar o sistema de pesquisas e levantamentos primários do IBGE e da concepção dos quadros referenciais teóricos da contabilidade social e da avaliação nutricional, distintas outras discussões relacionadas eram mantidas nesse período. Uma era relacionada à melhor forma de organizar a estrutura da Diretoria Técnica (correspondente à atual Diretoria de Pesquisas, com as seguintes áreas hoje na Diretoria de Geociências: Geografia, Recursos Naturais e Base Operacional Geográfica): (1) se por meio de duas superintendências, uma responsável pelas pesquisas primárias e outra responsável pelas derivadas; ou (2) separando as áreas por tema de atuação e cada uma delas cuidando tanto das pesquisas primárias quanto das derivadas. A grande questão por trás dessa discussão era como motivar estudos e análises que cobrissem todas as áreas de atuação da fundação. Essa questão transpôs os anos 1970 nas diferentes estruturações organizacionais pelas quais o IBGE passou.

Outra questão intensamente discutida dizia respeito à classificação das atividades econômicas, nas pesquisas de empresas, e a classificação de ocupações e atividades nas pesquisas domiciliares, aí incluídos os Censos Demográficos. As estatísticas de pessoal

ocupado nas pesquisas econômicas eram muito distintas do número de pessoas por atividade nos Censos e pesquisas demográficas. Um dos pontos observados para as diferenças estava nas distintas classificações empregadas nos dois grupos de pesquisas.

A dificuldade de realizar pesquisas amostrais sobre estabelecimentos, tanto para conhecer seus dados gerais quanto para estimar a produção física de produtos, informação necessária às contas nacionais, estava intimamente ligada à falta de cadastros atualizados. Isto conduziu à estratégia de utilizar os Censos Econômicos para gerar cadastros de estabelecimentos, a partir de 1975, e adotar a periodicidade quinquenal para esses Censos. Dessa forma, a cada cinco anos o cadastro era atualizado e as informações para as matrizes intersetoriais quinquenais eram coletadas.

Na realidade, as discussões sobre classificações de atividades e cadastros eram intimamente ligadas e não somente porque os cadastros de estabelecimentos deveriam ter sua classificação para estratificação das amostras, mas, sobretudo porque o conjunto de métodos de pesquisa (que muitos chamam de metodologia, cuja acepção mais adequada é análise crítica dos métodos) deve ser integrado, coerente e complementar.

A definição de novas pesquisas para atender às demandas das estatísticas derivadas também trazia outro ponto de reflexão à direção do IBGE: como assegurar que métodos semelhantes fossem aplicados a pesquisas conduzidas por distintas áreas do IBGE. No âmbito da amostragem, por exemplo, há duas formas distintas de selecionar amostras: (1) a partir dos cadastros de áreas, nas pesquisas demográficas; e (2) a partir de cadastros de empresas (ou estabelecimentos), nas pesquisas econômicas. No âmbito do tratamento estatístico das informações, em particular nos sistemas de crítica e imputação, também eram observados métodos distintos entre áreas do IBGE.

Essa discussão sobre a necessidade de padronizar métodos e, também, de como assegurar a qualidade das informações deu origem à criação do Departamento de Coordenação de Métodos, o ponto inicial da contagem dos 40 anos que ora se comemora.

A criação da área de métodos da Diretoria Técnica

O Departamento de Coordenação de Métodos foi criado pela Resolução n. 04, de 20.05.1977, da Presidência, com três áreas de atuação: (1) amostragem; (2) coordenação e avaliação; e (3) técnicas de levantamento. A Divisão de Amostragem tinha dois serviços de amostragem de estatísticas, um para as demográficas e sociais (as amostras de áreas) e outro para as econômicas (amostra de cadastro de empresas, entidades, estabelecimentos, etc.). A Divisão de Coordenação e Avaliação tinha o serviço de avaliação e análise e o de coordenação metodológica. A Divisão de Técnicas de Levantamentos tinha o serviço de bases operacionais (na prática só a geográfica) e o de estudos metodológicos.

Observa-se nessa organização, uma intenção clara de fazer um serviço de amostragem com suas bases cadastrais (apesar de apenas a base operacional geográfica estar no Departamento de Coordenação de Métodos) e ter uma área de avaliação e coordenação dos métodos aplicados nas distintas pesquisas primárias do IBGE.

Na concepção de Amaro da Costa Monteiro, diretor técnico do IBGE à época, o Departamento de Coordenação de Métodos atenderia às principais questões relacionadas à padronização dos métodos entre diferentes áreas do IBGE através da sua ação de coordenação e avaliação e das amostras que produziria. No entanto, não foi bem assim que as coisas aconteceram.

Em 1981, deixei o Departamento de Estudos do Consumo, já com a base de dados do ENDEF elaborada, e fui para o Departamento de Coordenação de Métodos, em particular para a sua Divisão de Amostragem. Foi feito um esforço para documentar os métodos de diferentes pesquisas (Série de Relatórios Metodológicos, que de fato eram relatórios de métodos); expansão da amostra do Censo Demográfico 1980 e calibração

de seus pesos; a Pesquisa Especial da Indústria 1981 e outros trabalhos demandados, mas não havia coordenação de métodos.

Em 1984, o Departamento de Coordenação de Métodos foi extinto e, pela Resolução n. 27, de 13.07.1984, da Presidência, foi criada a Coordenadoria de Avaliação e Metodologia, ligada à Presidência do IBGE. Nessa época, fui transferido da Coordenadoria de Avaliação e Metodologia para a Superintendência de Estatísticas Industriais, Comerciais e dos Serviços, com orientação direta do Presidente do IBGE, Professor Jessé de Souza Montello, de colaborar com a área de métodos na implantação de amostragem probabilística nas pesquisas de indústria, comércio e serviços sob a responsabilidade da Superintendência de Estatísticas Industriais, Comerciais e dos Serviços. Não que houvesse resistências à implantação de amostragem nas pesquisas dessa área. A preocupação maior era com a mudança de cultura das equipes de crítica e correção de dados e de sistemas computacionais, decorrente do uso de amostragem. O operador de crítica e correção deveria ter em mente que um erro, na apuração da pesquisa censitária, seria, de fato, "peso amostral" vezes erros na pesquisa por amostra e que imputações determinísticas pela média, por exemplo, afetariam as medidas de variabilidade. Já os sistemas de crítica de valores extremos, por exemplo, passariam a depender do peso amostral para estimar a média e o desvio padrão das distribuições de valores.

No entanto, esperava-se que a nova posição institucional da área de métodos junto à Presidência permitisse ampliar o escopo de sua atuação, envolvendo as distintas diretorias do IBGE e promovendo uma coordenação de métodos mais efetiva no IBGE, o que não ocorreu: a Coordenadoria de Avaliação e Metodologia durou pouco mais de um ano.

Em 1985, a Coordenadoria de Avaliação e Metodologia foi transformada na Coordenadoria de Metodologia, subordinada ao Centro de Ensino e Desenvolvimento Metodológico, pela Resolução n. 40, de 16.08.1985, da Presidência, unidade que também tinha a ENCE sob sua responsabilidade. Sua colocação fora da área de produção das estatísticas (primárias ou derivadas), é um indicador da falta de percepção da gestão do IBGE à época em relação à importância dos métodos de pesquisa como definidores do grau de confiança que se pode atribuir aos achados e resultados de uma pesquisa científica.

Em 1987, a área foi transferida para a Diretoria de Pesquisas, na qual permanece até os dias de hoje, com distintos nomes e pequenas alterações em suas atribuições até sua transformação na Coordenação de Métodos e Qualidade, em junho de 2003. Sonia Albieri (2003) fez um histórico pormenorizado dessas transformações, desde a criação do Departamento de Coordenação de Métodos até a constituição da atual Coordenação de Métodos e Qualidade.

Classificação, cadastros e métodos no IBGE

A questão das classificações, observada desde os anos 1970 pela discrepância entre os resultados de pesquisas de fontes consumidoras e fontes produtoras, deu origem, em 1994, a uma área responsável pela manutenção da Classificação Nacional de Atividades Econômicas - CNAE. O cadastro de estabelecimentos que derivava dos Censos Econômicos (sendo o último realizado em 1985) foi substituído pelo Cadastro Central de Empresas - Cempre (ZACHARIAS, 1993).

A partir de 1994, foi implantado um novo modelo de produção das estatísticas econômicas, suspendendo a realização dos Censos Econômicos e tendo como elemento central de ordenamento do sistema o Cempre (usado nas amostras das pesquisas econômicas anuais) e baseado na integração com gestores dos principais registros administrativos, que resultou na adoção, por parte de vários órgãos da esfera federal, a partir de 1995, da CNAE. Esta, por sua vez, passou a ser o resultado de um trabalho conjunto entre diversas entidades

produtoras e usuárias de estatísticas econômicas e de registros administrativos de âmbito nacional, sob a coordenação do IBGE, que teve por objetivo construir uma classificação-padrão, compatibilizada com a revisão 3 da Clasificación Industrial Internacional Uniforme de Todas las Actividades Económicas - CIIU (International Standard Industrial Classification of all Economic Activities - ISIC), das Nações Unidas.

Para atualizar o Cadastro de Empresas com a nova classificação de atividades, foi realizado o Censo Cadastro 1995 (ALBIERI, S.; BIANCHINI, Z. M.; VASCONCELLOS, 1996), com o objetivo de construir um cadastro com cobertura censitária para as empresas consideradas importantes (ou de grande porte) e investigação por amostragem para o restante da população, atribuindo às empresas o novo código da CNAE.

Atualmente, tanto o Cempre quanto a CNAE estão sob a responsabilidade da Coordenação das Estatísticas de Empresas, Cadastro e Classificações, da Diretoria de Pesquisas, que também atua como Secretaria Executiva da Comissão Nacional de Classificação - Concla.

A base operacional geográfica, cadastro de áreas (setores) para seleção de amostras das pesquisas domiciliares, foi transferida para a atual Coordenação de Estruturas Territoriais, da Diretoria de Geociências, quando o Departamento de Coordenação de Métodos foi extinto e é atualizada antes da realização dos Censos Demográficos. Após cada Censo Demográfico, o IBGE produz uma nova versão do arquivo agregado por setores censitários dos resultados do universo, disponível para download em seu sítio, para quem precisa selecionar amostras por setor. Atualmente, existe um projeto de base operacional que trabalha com atualização dos setores de forma contínua, congelando o conjunto de setores em datas definidas para a realização dos Censos Agropecuário e Demográfico.

Os demais métodos de pesquisa do IBGE estão distribuídos nos distintos departamentos da Diretoria de Pesquisas, Diretoria de Geociências, Diretoria de Informática e Centro de Documentação e Disseminação de Informações. Os departamentos dessas unidades são os responsáveis pelos estudos para atualização dos quadros referenciais de suas pesquisas, pelas novas técnicas de informática e de disseminação de informações.

Nesse sentido, o papel de coordenação dos métodos do antigo Departamento de Coordenação de Métodos foi transferido para o nível hierárquico superior (Diretoria de Pesquisas, Diretoria de Geociências, Diretoria de Informática e Centro de Documentação e Disseminação de Informações) e, no caso de métodos que transcendam esse nível, para o Conselho Diretor do IBGE.

Referências

ALBIERI, S. *A unidade de metodologia e a evolução do uso de amostragem no IBGE*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 42 p. (Textos para discussão n. 12). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv2282.pdf>. Acesso em: set. 2017

ALBIERI, S.; BIANCHINI, Z. M.; VASCONCELLOS, M. T. L. de. *Aspectos de amostragem relativos ao censo cadastro de 1995*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1996. 47 p. (Textos para discussão, n. 80). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv25810.pdf>. Acesso em: set. 2017.

BRASIL. Decreto-Lei n. 161, de 13 de fevereiro de 1967. Autoriza o Poder Executivo a instituir a "Fundação Instituto Brasileiro de Geografia e Estatística" e dá outras providências. *Diário Oficial [da] República do Brasil*, Brasília, DF, ano 105, n. 130, 14 fev. 1967. Seção 1, p. 1. Disponível em: < http://www.planalto.gov.br/ccivil_03/decreto-lei/1965-1988/Del0161.htm>. Acesso em: set. 2017.

FUNDAÇÃO GETÚLIO VARGAS. *Food consumption in Brazil*: family budget surveys in the early 1960's. Jerusalém: U.S. Department of Agriculture, Economic Research Service, 1970. 283 p. (Israel Program for Scientific Translations).

_____. *Projections of supply and demand for agricultural products of Brazil through 1975*. Rio de Janeiro: Center for Agricultural Studies, 1967.

IBGE. Resolução do Presidente n. 4, de 20 de maio de 1977. Dispõe sobre a estrutura, competência e atribuições dos órgãos de Assessoramento Superior, das Diretorias e Unidades Regionais do IBGE, e dá outras providências. *Boletim de Serviço*, Rio de Janeiro, n. 1296, p. 1, 10 jun. 1977.

_____. Resolução do Presidente n. 6, de 24 de abril de 1978. Cria a Divisão de Índices de Preços, a Divisão de Estatísticas do Setor Público e o Departamento de Estudos do Consumo, e dá outras providências. *Boletim de Serviço*, Rio de Janeiro, n. 1342, p. 1, 28 abr. 1978.

_____. Resolução do Presidente n. 27, de 13 de julho de 1984. Criar a Coordenadoria de Avaliação de Metodologia (CAM), órgão de Assessoramento Superior de que trata o capítulo III, Seção I, do Estatuto. *Boletim de Serviço*, Rio de Janeiro, n. 1649, p. 1, 23 jul. 1984.

_____. Resolução do Presidente n. 40, de 16 de agosto de 1985. Cria o Centro de Ensino e Desenvolvimento Metodológico e dá outras providências. *Boletim de Serviço*, Rio de Janeiro, n. 1703, p. 2, 26 ago. 1985.

ISAAC Kerstenetzky: legado e perfil. Rio de Janeiro: IBGE, Centro de Documentação e Disseminação de Informações, 2006. 210 p. Disponível em: <http://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232590>. Acesso em: set. 2017.

OLIVEIRA, J. S. de. Isaac Kerstenetzky: um tributo ao mestre. In: ISAAC Kerstenetzky: legado e perfil. Rio de Janeiro: IBGE, 2006. p. 182-187. Disponível em: <http://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232590>. Acesso em: set. 2017.

SANT'ANNA, M. A. de A. Professor Isaac. In: ISAAC Kerstenetzky: legado e perfil. Rio de Janeiro: IBGE, 2006. p. 179-181. Disponível em: <http://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232590>. Acesso em: set. 2017.

SENRA, N. de C. A fundação é refundada na administração Kerstenetzky. In: ISAAC Kerstenetzky: legado e perfil. Rio de Janeiro: IBGE, 2006. p. 160-178. Disponível em: <http://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232590>. Acesso em: set. 2017.

VASCONCELLOS, M. T. L. de *Análise crítica dos métodos de avaliação nutricional de populações, a partir de dados de consumo familiar de energia*. 2001. 231 p. Tese (Doutorado)- Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rio de Janeiro, 2001. Disponível em: <https://www.arca.fiocruz.br/bitstream/icict/4542/2/71.pdf>. Acesso em: set. 2017.

ZACHARIAS, M. L. B. *Cadastros estatísticos de empresas construídos a partir de registros administrativos*. Santiago do Chile: Comissão Econômica para a América Latina e o Caribe - Cepal, 2003. Trabalho apresentado na segunda reunião da Conferência de Estatística das Américas da Cepal, realizada em Santiago do Chile, 2003. Disponível em: <https://www.cepal.org/deype/ceacepal/documentos/lcl1892p.pdf>. Acesso em: set. 2017.

Sigilo das informações individualizadas no IBGE

Zélia Magalhães Bianchini

Introdução

O objetivo do Instituto Brasileiro de Geografia e Estatística - IBGE sempre foi e continua sendo o de fornecer informação relevante e adequada à sociedade. Ao mesmo tempo, os avanços tecnológicos oferecem oportunidades para armazenar, processar, acessar, combinar e analisar grandes conjuntos de dados de forma mais eficiente, reforçando a exigência dos usuários por informações cada vez mais detalhadas. Porém, aumentar a riqueza e os detalhes das informações implica o risco de revelação de dados identificáveis e sigilosos.

Nesse contexto, existe um dilema sobre até onde vão os benefícios e riscos de disponibilizar informações. A questão recai sobre o que pode ou não ser feito para assegurar que os compromissos de sigilo das informações individualizadas sejam mantidos e prover acesso aos dados.

A Lei n. 5.534, de 14.11.1968, que confere ao IBGE o seu mandato para a produção de estatísticas é explícita no sentido de dispor sobre a obrigatoriedade de prestação de informações ao IBGE, mas ao mesmo tempo impõe como contrapartida, conforme o parágrafo único do Art. 1º: as informações prestadas terão caráter sigiloso, serão usadas exclusivamente para fins estatísticos e não poderão ser objeto de certidão, nem, em hipótese alguma, servirão de prova em processo administrativo, fiscal ou judicial [...].

Uma das condições essenciais para a manutenção da integridade e a credibilidade de uma instituição produtora de estatísticas como o IBGE (e como todos os seus similares no mundo) é a preservação intransigente do sigilo das informações individuais ou identificadas que utiliza como parte de seu processo de produção de estatísticas. O IBGE depende de maneira crítica da confiança pública para obter os dados individualizados de que necessita como matéria prima básica para a produção estatística. O dever de sigilo imposto ao IBGE pela legislação presta-se, justamente, a gerar a necessária confiança daqueles que prestam as informações, garantindo, inclusive, a fidedignidade dos dados coletados. Portanto, para honrar o contrato estabelecido com os informantes é fundamental manter o sigilo prometido quando as informações foram coletadas, através da anonimidade e do requerimento de que o risco de revelação das informações individualizadas seja aceitavelmente baixo.

Uma importante área da pesquisa estatística desenvolve métodos que permitem divulgar o máximo possível dos dados, respeitando o aspecto legal para proteger a confidencialidade prometida aos informantes. Vários institutos nacionais de estatística têm desenvolvido procedimentos e investido em soluções a fim de encontrar o equilíbrio certo para enfrentar esse desafio.

Este capítulo apresenta os referenciais institucionais e princípios norteadores, bem como uma descrição de métodos e práticas no âmbito das atividades relacionadas com a preservação do sigilo das informações individualizadas, incluindo acesso ao acervo institucional de dados e a segurança. São abordados métodos e procedimentos para proteção no que se refere: aos dados agregados / tabulados, aos microdados para uso público, ao acesso a microdados não disseminados para uso público, aos microdados com identificadores (cadastros) e aos dados em grades.

Os desafios e perspectivas para o futuro identificam diversas ações e linhas de trabalho a serem implementadas no IBGE, com ênfase na atualização com as experiências dos institutos nacionais de estatística mais avançados.

Referenciais e princípios

Neste tópico são abordados referenciais institucionais e princípios norteadores na produção de informações estatísticas, que fornecem diretrizes, valores e boas práticas para a produção e organização de estatísticas oficiais, com destaque para as questões relacionadas com o sigilo das informações individualizadas.

Referenciais das Nações Unidas

A Comissão de Estatística das Nações Unidas (United Nations Statistical Commission) adotou os Princípios Fundamentais das Estatísticas Oficiais (Fundamental Principles of National Official Statistics), na 28ª sessão, realizada em 1994, que foram endossados pela Assembleia Geral das Nações Unidas em 2014 (UNITED NATIONS, 2014). Trata-se de um conjunto de 10 princípios para as atividades estatísticas, que refletem os valores com os quais devem ser analisadas as estatísticas.

O Princípio número seis é dedicado à confidencialidade e estabelece o seguinte: "Dados individuais coletados por órgãos de estatística para produção de informação estatística, sejam referentes à pessoa física ou jurídica, devem ser estritamente confidenciais e usados exclusivamente para fins estatísticos." (UNITED NATIONS, 2014, p. 2, tradução nossa).

Em 2003, foi publicada pela Divisão de Estatística do Departamento de Assuntos Econômicos e Sociais (Department of Economic and Social Affairs - DESA) das Nações Unidas a terceira edição do manual de organização estatística, *Handbook of statistical organization*, que trata da operação e organização de uma agência de estatística e define os fundamentos do sistema nacional de estatísticas oficiais. O manual está estruturado nas seguintes áreas: princípios gerais, coleta de dados e políticas dos informantes, necessidades dos usuários, princípios de organização e gerenciamento e diretrizes de disseminação. Há um capítulo específico sobre o respeito à privacidade e preservação do sigilo: honrar o contrato com o informante (UNITED NATIONS, 2003).

Códigos de boas práticas e de ética do IBGE

Em 2013, o IBGE publicou seu Código de Boas Práticas das Estatísticas com base no Código Regional de Boas Práticas das Estatísticas para a América Latina e o Caribe (Código Regional de Buenas Prácticas en Estadísticas para América Latina **y** el Caribe). O Código é constituído por um conjunto de diretrizes – recomendações e orientações – estruturadas em três seções (ambiente institucional e coordenação, processos estatísticos e produtos estatísticos), 17 princípios e 80 indicadores de boas práticas (IBGE, 2013).

O princípio quatro do Código trata da confidencialidade estatística: "O IBGE deve garantir a proteção e a confidencialidade das informações individualizadas com as quais são produzidas as estatísticas oficiais" (IBGE, 2013, p. 19).

Para o referido princípio foram estabelecidos os seguintes indicadores de boas práticas:

> 4.1 A legislação vigente deve prever a confidencialidade dos dados individualizados e a garantia de que são usados, exclusivamente para fins estatísticos, e que não podem ser usados para fins comerciais, de tributação fiscal, de investigação judicial e outros.
>
> 4.2 Deve haver uma declaração assinada pelas pessoas que têm acesso a informações de caráter individual ou confidencial, especificando o compromisso com a confidencialidade e as penalidades em caso de não cumprimento.
>
> 4.3 Deve haver normas e compromissos jurídicos de confidencialidade de informações estabelecidos para o pessoal envolvido com a geração e a análise das estatísticas oficiais, que estipulem penalidades em caso de não cumprimento.
>
> 4.4 Deve haver protocolos que estabeleçam diretrizes sobre segurança e integridade dos processos e das bases de dados estatísticos do Instituto.
>
> 4.5 Devem ser explicitados os principais usos e limitações de acesso que se aplicam às informações obtidas pelo Instituto junto aos informantes.
>
> 4.6 O acesso aos microdados não desidentificados deve estar sujeito a protocolos de confidencialidade, estabelecidos para usuários externos que têm acesso com a finalidade de análise e pesquisa estatística.
>
> 4.7 O arquivamento das informações pelo Instituto deve ser feito de acordo com os protocolos de segurança e confidencialidade estabelecidos e com as normas vigentes. (IBGE, 2013, p. 19-20).

O IBGE possui também o Código de Ética Profissional do Servidor Público do IBGE, instituído em 2014, que estabelece os preceitos sobre a conduta ético-moral de seus servidores. Consta do Código de Ética a vedação ao servidor do IBGE de disponibilizar informações de caráter sigiloso e confidencial sobre pessoas físicas ou jurídicas (IBGE, 2014).

Cabe registrar a existência da Comissão de Ética do IBGE, encarregada de orientar e aconselhar sobre a ética profissional dos servidores da Casa, no tratamento com as pessoas e com o patrimônio público.

Comitês relacionados com o sigilo

O IBGE conta com um Comitê de Sigilo, instância criada pela Resolução n. 2, de 19.02.2001, do Conselho Diretor do IBGE, que é o locus da reflexão para as questões relacionadas ao sigilo, com as seguintes atribuições:

> a) propor soluções para questões relacionadas ao sigilo das informações de natureza estatística e geocientífica, individuais ou identificadas, coletadas, produzidas, armazenadas e disseminadas pelo IBGE;
>
> b) apreciar as solicitações de acesso às informações confidenciais, dando parecer sobre autorização e condições de uso; e
>
> c) avaliar rotineiramente o cumprimento da promessa de assegurar a manutenção do sigilo das informações confidenciais. (IBGE, 2001, p. 1).

Com o objetivo de atender a demanda de pesquisadores em projetos de relevante interesse público e/ou acadêmico e que necessitam de acesso à base de dados que não está disponível para uso público, o IBGE oferece, desde 2003, o Serviço de acesso a microdados não desidentificados, pelo qual se assegura o sigilo das informações individuais, em sala de acesso a dados restritos. A decisão sobre tal acesso se dá no Comitê de Avaliação de Acesso a Dados não Desidentificados, criado através da Resolução n. 7, de 04.06.2003, do Conselho Diretor do IBGE.

Cabe ainda registrar a existência do Comitê de Segurança da Informação e Comunicações do IBGE criado pela Resolução n. 26, de 21.10.2015, do Conselho Diretor do IBGE, com o objetivo de reunir competências para elaboração e gestão permanente da Segurança da Informação e Comunicações do IBGE.

Métodos para limitar revelação

Existem diferentes abordagens de proteção de dados para garantir a confidencialidade das informações individualizadas, dentre as quais se destacam: uso da criptografia, divulgar /compartilhar com restrição dos dados e compartilhar dados através de acesso restrito ou da combinação desses dois últimos tipos de abordagem.

No caso de restrição dos dados, os métodos de limitação de revelação estatística podem ser aplicados aos microdados e/ou a dados tabulados antes do lançamento. As abordagens tradicionais estão relacionadas à agregação (divulgação apenas de resumos tabulares); arredondamento (de estimativas em tabelas); supressão parcial (de celas em dados tabulados); recodificação (em tabelas e microdados); anonimização (em microdados); amostragem (em microdados); modificação (mascaramento, perturbação – adição de ruído aleatório a determinados campos em microdados); permutação de dados (em microdados).

A restrição do acesso abrange o controle do acesso, as limitações do acesso e as condições sobre quem pode acessar os dados, para qual propósito, quais variáveis podem ser acessadas, as restrições às ligações, o que pode ser publicado, etc.

A revelação da identidade ocorre quando terceiros podem identificar um respondente a partir dos dados divulgados. Em microdados, a identificação geralmente é considerada como revelação. Por esse motivo, os métodos para limitação da revelação, aplicáveis a microdados, limitam ou modificam as informações que poderiam ser usadas para identificar respondentes específicos.

Práticas de proteção para dados tabulados

A seleção de uma regra do sigilo estatístico para os dados apresentados nas tabelas (dados tabulados) depende se os dados são frequências ou outras quantidades de interesse, que devem medir algo mais que a contagem de unidades na cela da tabela.

No caso de tabelas de frequências, os principais métodos para proteger o sigilo são os seguintes:

- Amostragem e estimação das frequências nas celas.
- Regras especiais - impõem restrições sobre o nível de detalhe que pode ser dado a uma tabela, por exemplo, proíbem tabulações nas quais uma determinada cela é igual ao total marginal.
- Regra do patamar - uma cela da tabela de frequências é definida como de risco se o número de respondentes é menor que algum número específico. Para reduzir o risco de revelação, as tabelas podem ser reestruturadas e as categorias combinadas ou podem ser usados métodos tais como supressão de celas, arredondamento aleatório, arredondamento controlado ou edição de confidencialidade:
  - Supressão de celas - numa linha ou coluna com uma cela de risco suprimida, pelo menos uma cela adicional deve ser suprimida (supressão complementar);
  - Arredondamento aleatório - em vez de usar o arredondamento padrão, uma decisão aleatória é tomada indicando se o valor será arredondado para cima ou para baixo;
  - Arredondamento controlado - é uma forma de arredondamento aleatório, mas com a restrição de ter a soma das entradas publicadas em cada linha e coluna igual ao total marginal publicado correspondente;
  - Edição de confidencialidade - técnica de limitação de revelação estatística que pode ser aplicada aos arquivos de microdados antes que estes sejam usados na preparação das tabelas.

As tabelas que contêm dados quantitativos têm um problema único de revelação. As distribuições destes valores são usualmente assimétricas, com poucas unidades tendo valores muito grandes. A limitação da revelação, neste caso, concentra-se em

evitar que os dados publicados sejam usados para deduzir os valores das maiores unidades. Protegendo os maiores valores, todos os valores estarão protegidos.

Existem regras desenvolvidas para tabelas de dados quantitativos, chamadas regras de supressão primária ou medidas de risco linear, que determinam se uma dada cela da tabela poderia revelar informações individuais do respondente. Tais celas são chamadas de celas de risco e não podem ser publicadas. Uma vez identificadas as celas de risco, há duas opções: reestruturar a tabela e juntar celas até que acabem as celas de risco ou suprimir celas.

Várias agências utilizam uma forma administrativa para evitar a supressão de celas. Elas obtêm uma permissão por escrito dos respondentes que contribuem para as celas de risco, que autorizam a publicação das mesmas. Essa permissão é conhecida como "*waiver*" à promessa de proteger o sigilo das informações individuais.

A prática corrente no IBGE de tratamento para desidentificação de dados tabulados para divulgação ocorre, tradicionalmente, nas pesquisas da área econômica (indústria, comércio e serviços), no Censo Agropecuário e nas pesquisas de estabelecimentos da área social e da agropecuária. O método utilizado é o da regra do patamar, onde são exigidos pelo menos três respondentes por cela. A opção preferencial se dá pelo procedimento de supressão de celas, dentro do objetivo de menor supressão de informações.

A título de exemplo, nas pesquisas da Indústria a agregação de celas com menos de três informantes se dá por dentro dos agrupamentos da Classificação Nacional de Atividades Econômicas - CNAE, seguindo o critério de escolha da classe/grupo/divisão em função do menor Valor de Transformação Industrial - VTI. Já na Pesquisa Anual do Comércio - PAC, a agregação por dentro dos agrupamentos da CNAE obedece a critério de homogeneidade entre as atividades, pois a classificação do comércio tem menos abertura (de dois e três dígitos) do que a da indústria.

Os procedimentos relacionados com o sigilo das informações individuais coletadas no Censo Demográfico 2010 estão descritos na segunda edição da publicação *Metodologia do censo demográfico 2010* (METODOLOGIA..., 2016).

Práticas de proteção para microdados disseminados

Os tipos de dados presentes nos arquivos de microdados podem ser classificados como identificadores, atributos de domínio público (classificadores) e dados confidenciais, sendo diferenciados de acordo com o tipo de unidade pesquisada: domicílio e/ou pessoa, unidade econômica (empresa, estabelecimento, etc.), produto e preço.

Alguns microdados incluem identificadores explícitos, tais como nome e endereço. A remoção de tais identificadores obviamente é o primeiro passo na preparação para a liberação de um arquivo cuja confidencialidade da informação individualizada deve ser protegida.

Além disso, há outros fatores que contribuem para aumentar os riscos de revelação. Um primeiro fator é a existência de informações com alta visibilidade. Alguns registros podem representar respondentes com uma característica única no local, tais como determinadas ocupações em certos municípios (juiz, padre, médico, etc.), ou com rendimentos muito altos. O segundo fator de risco é a possibilidade de batimento do arquivo de microdados com outros arquivos com informações mais detalhadas.

Há uma grande dificuldade em proteger um arquivo de microdados da revelação devido à possibilidade de associação com fontes de dados externas. Além disso, não há medidas aceitáveis do risco de revelação para um arquivo de microdados, então não há um padrão que possa ser aplicado para garantir que a proteção é adequada.

Os métodos para reduzir o potencial de revelação limitam ou modificam as informações. Em geral, os arquivos de microdados de uso público: incluem apenas os dados de uma amostra da população; não incluem identificadores óbvios; limitam o detalhe geográfico; e limitam o número de variáveis no arquivo.

Os métodos adicionais usados para disfarçar variáveis mais evidentes incluem: recodificação por intervalos, agregação ou arredondamentos; adição ou multiplicação por números aleatórios (ruídos); permutação; seleção de registros aleatoriamente, atribuindo valor branco para algumas variáveis selecionadas e imputando-as novamente; agregação de pequenos grupos de respondentes, substituindo o valor informado de cada indivíduo pela média do grupo.

Desde o início da década de 1990, o IBGE divulga arquivos de microdados de uso público das pesquisas domiciliares por amostragem probabilística. Para proteger a confidencialidade desses microdados, adota-se a retirada dos identificadores, geografia pouco detalhada e ordenação aleatória dos registros. Por outro lado, não há liberação de arquivos de microdados das pesquisas econômicas e do Censo Agropecuário, nem das informações investigadas para toda a população do Censo Demográfico, que corresponde ao conteúdo das informações do questionário básico do referido Censo.

Práticas de proteção no acesso a microdados não disseminados

Apesar da ampla divulgação dos dados do IBGE por meio de publicações impressas e na Internet, a impossibilidade de liberar os arquivos com dados mais detalhados limita a utilidade desses dados para fins de pesquisa e de políticas públicas.

Nas pesquisas econômicas, a prática dos principais órgãos de estatística internacionais é não liberar arquivo de microdados. Os riscos de revelação em microdados de pesquisas da área econômica são muito mais altos que aqueles para microdados de pesquisas domiciliares. As informações da área econômica são na maioria dados quantitativos, que possuem geralmente distribuição assimétrica e possibilitam que os respondentes sejam facilmente identificados a partir de outras informações disponibilizadas publicamente, e, além disso, podem conter informações estratégicas. Por essas razões não são disseminados microdados da área econômica para uso público.

A crescente demanda por microdados detalhados, o avanço da tecnologia e o aumento da preocupação com questões de privacidade levaram o IBGE, a partir de 2003, a propiciar acesso pelos pesquisadores a arquivos de dados que não são liberados para o público em geral, permitindo análises mais aprofundadas do que aquelas possíveis com dados tabulados ou agregados. Em resposta ao interesse da comunidade de pesquisa, o IBGE criou uma sala de acesso especial a dados restritos onde os pesquisadores podem acessar arquivos criptografados de dados detalhados em um ambiente seguro, sem comprometer a confidencialidade dos informantes.

Conforme descrito por Koeller, Vilhena e Zacharias (2013), foi realizado um amplo estudo sobre os procedimentos e práticas existentes e desenvolvidos por outros institutos nacionais de estatística, para servir de base para as melhorias implementadas no acesso a dados restritos. Os procedimentos para obtenção de acesso a arquivos de microdados para fins científicos estão descritos por Zacharias, Bianchini e Albieri (2013), juntamente com as mudanças e melhorias que estão sendo implementadas.

> (a) Submissão do projeto de pesquisa
>
> Para poder acessar dados restritos na sala destinada a este fim, os pesquisadores devem, inicialmente, apresentar uma proposta de projeto de pesquisa, [..] obedecendo às instruções a respeito das informações que devem ser incluídas no pedido. [...] um formulário padronizado está sendo desenvolvido, contendo campos específicos a serem preenchidos com as informações solicitadas para serem analisadas para a concessão do acesso. Este formulário estará disponível na página do IBGE na internet, permitindo que a proposta seja submetida eletronicamente [...].
>
> [...] um guia do usuário está sendo elaborado [...] para que os pesquisadores fiquem cientes de todos os procedimentos e regras a serem seguidos para obter autorização de uso da sala de acesso a dados restritos. [...](ZACHARIAS; BIANCHINI; ALBIERI, 2013, p. 2-3).

Um sistema eletrônico para gerenciar todo o processo está em desenvolvimento para substituir o atual sistema que ainda é feito manualmente. Os procedimentos continuam ser descritos pelas autoras:

> (b) Avaliação do projeto
>
> O Comitê de Avaliação de Acesso a Dados Restritos é responsável por avaliar o projeto, com base em informações fornecidas pela área técnica responsável pela produção dos dados da pesquisa. O Comitê autoriza (ou não) o uso da sala de acesso a dados restritos, assegurando que não haverá risco de quebra de confidencialidade. [...]
>
> A incorporação de bases de dados externas aos arquivos de microdados solicitados geralmente é permitida, a menos que se perceba a existência de algum risco de revelação da identidade dos informantes. [...] essa avaliação é feita pelo analista temático [...].
>
> Outra iniciativa para tornar esta incorporação de dados externos mais restritiva é exigir que os pesquisadores apresentem uma autorização formal, assinada pelos gestores das bases de dados externas, para garantir que houve consentimento da fonte externa para utilização daqueles dados no projeto de pesquisa.
>
> (c) Termos de compromisso para acesso aos dados
>
> Depois do projeto de pesquisa ter sido aprovado, um termo de compromisso é assinado entre o pesquisador e o instituto de estatística, especificando as condições e os valores das taxas de uso da sala de acesso. [...]
>
> (d) Acesso local
>
> Na sala de acesso, os arquivos com os microdados restritos são instalados em computadores especiais munidos de aspectos de segurança, tais como bloqueio à rede e à internet para transferência de dados, e desativação das unidades de disco externas e das portas paralelas e seriais. Os bancos de dados de pesquisas de empresas do IBGE ou de fontes externas têm a identificação das empresas criptografadas.
>
> Após a conclusão do trabalho, o pesquisador grava as saídas produzidas no disco rígido do computador e, em seguida, elabora um relatório documentando o que foi feito. [...] (ZACHARIAS; BIANCHINI; ALBIERI, 2013, p. 3-4).

Ainda de acordo com Zacharias, Bianchini e Albieri (2013, p. 4), "para aprimorar este procedimento, está em desenvolvimento o uso de uma rede interna de dados protegida para transmitir os arquivos de dados" para a análise dos especialistas da área temática responsável pela base de dados da pesquisa.

> Outra melhoria a ser implementada é que apenas as variáveis solicitadas, e que sejam realmente necessárias para a realização das análises propostas, serão incluídas no arquivo de microdados [...].
>
> (e) Verificação e liberação das saídas
>
> [...] A verificação das saídas é feita pela área técnica responsável pelos microdados da pesquisa. [..] depois que a área temática avalia que não há risco de revelação, outro termo de compromisso é assinado.
>
> Este termo estabelece as condições de uso dos dados gerados por esta modalidade especial de acesso, onde o usuário reconhece que os dados são de propriedade do IBGE, e se compromete a informar que os resultados e análises eventualmente divulgados envolvendo esses dados foram obtidos por meio desse acesso especial (ZACHARIAS; BIANCHINI; ALBIERI, 2013, p. 4).

Práticas de proteção para microdados com identificadores (cadastros)

A questão do sigilo em cadastros é bastante abrangente, em função dos diversos cadastros existentes: cadastro de empresas, de estabelecimentos agropecuários, de endereços, de produtos e preços. Nos cadastros o problema de revelação é da identificação, diferentemente dos microdados, quando se têm associados os atributos.

O IBGE vem fornecendo seletivamente o Cadastro Central de Empresas - Cempre, disponibilizando apenas nome, endereço, atividades e faixas de pessoal ocupado, aos usuários que apresentam justificativas de uso para fins estatísticos.

Dependendo da natureza da instituição, mesmo que para fins estatísticos, o Cempre só é disponibilizado para o atendimento através da sala de acesso a dados restritos, cumpridos os requisitos estabelecidos. Por exemplo, no caso de solicitação para fins de desenho e seleção de amostra, dá-se o acesso aos dados do cadastro necessários à amostragem, para que o usuário proceda a seleção, obtenha as informações necessárias

à determinação dos pesos de expansão e produza a relação de códigos das unidades selecionadas. Com base nessa relação é gerado o arquivo contendo as informações cadastrais necessárias à condução dos trabalhos de coleta prevista na pesquisa.

Em particular, para a Fundação Sistema Estadual de Análise de Dados - Seade, do Estado de São Paulo, o IBGE forneceu o Cempre com nome, endereço, atividade, pessoal ocupado e receita, com a finalidade de seleção da amostra da Pesquisa da Atividade Econômica Paulista - PAEP, mediante convênio estabelecido e termo de compromisso de resguardo do sigilo.

No caso do cadastro de estabelecimentos agropecuários, os procedimentos são semelhantes aos adotados para o Cempre.

O IBGE disponibiliza publicamente, na Internet, o Cadastro Nacional de Endereços para Fins Estatísticos - CNEFE com as informações sobre o nome e número do logradouro, a espécie associada ao endereço (domicílio particular, domicílio coletivo - hotéis, alojamentos, asilos, etc. -, estabelecimento agropecuário, estabelecimento de ensino, estabelecimento de saúde, estabelecimento de outras finalidades e edificação em construção) e as coordenadas geográficas na área rural, retirando o nome e todos os demais atributos.

No caso do cadastro de produtos e preços, o IBGE, tradicionalmente, fornece apenas preços médios para produtos genéricos, sem identificação de marca. Apesar da demanda pelo fornecimento de preços médios de determinados produtos, por marca, o IBGE vem respondendo negativamente com base na garantia do sigilo, na proteção ao informante e na preservação da imagem da instituição.

Práticas de proteção para dados em grades

A utilização de dados geoespaciais juntamente com tecnologias afins permite aos pesquisadores e tomadores de decisão entender melhor as relações dinâmicas entre os fatores críticos para a pesquisa em muitas áreas. Desenvolvimentos em sensoriamento remoto e tecnologia de computação têm melhorado a resolução dos dados geoespaciais e facilitado a integração destes dados com outros, oferecendo uma maior capacidade de análise das informações. No entanto, na medida em que os dados são espacialmente precisos, existe um aumento correspondente no risco de identificação das pessoas ou organizações para os quais os dados se aplicam.

Uma grade estatística é uma estrutura espacial hierárquica composta por células regulares e utilizada para disseminar dados estatísticos agregados. Embora as grades estatísticas apresentem vantagens significativas como escalabilidade, comparabilidade temporal, portabilidade, flexibilidade, acurácia e detalhamento, na mesma proporção intensificam a questão do sigilo no contexto espacial.

A Presidência do IBGE criou, por meio da Portaria n. 485, de 06.12.2013, o Grupo de Trabalho sobre Sigilo de Informações em Grades Estatísticas (2015), encarregado de desenvolver estudos e procedimentos que possibilitem manter o sigilo das informações individualizadas na disseminação de dados através de grades estatísticas.

A disseminação de dados censitários em grades regulares já é uma realidade em Institutos de Estatística de alguns países como, por exemplo, na Noruega, na Finlândia e no Japão. Praticamente todos os produtores de dados enfrentam restrições na disseminação de dados em grade, por oferecer um maior detalhamento, implicando no aumento das chances de quebra de sigilo. De um modo geral, o método mais utilizado diferencia dado mais e menos sensível e define, de acordo com esta sensibilidade, o limite de supressão e o tamanho mínimo da cela.

Um dos grandes problemas para disseminar os dados do Censo Demográfico 2010 em grades, para as diversas variáveis investigadas, está no risco de revelação que pode vir da identificação pela diferenciação geográfica envolvendo os dois recortes, grade e setor censitário, mesmo considerando que seriam aplicados tratamentos para proteger o sigilo também na divulgação por grade. Já tendo sido divulgada a base de informações por setor censitário, uma estratégia discutida no grupo de trabalho foi

manter ou aumentar os limites para a supressão das informações para as celas da grade em relação aos limites adotados para a base de informações por setor censitário[1].

A título de estudo, foi realizado um exercício de supressão de informações para uma grade para avaliar a quantidade de celas que teriam a supressão das informações em três hipóteses: celas com menos de cinco domicílios particulares permanentes; com menos de 10 domicílios particulares permanentes; e com menos de cinco domicílios particulares permanentes e menos de vinte pessoas. Foi utilizada a grade gerada em projeção geográfica, cuja dimensão nas áreas rurais é de 1 km e nas áreas urbanas de 200 m. As quantidades de celas que teriam as informações omitidas foram bastante expressivas (53 a 67%).

Utilizando os resultados do Censo Demográfico 2010 e a divisão em Grade Estatística, que divide o território em mais de 2,5 milhões de celas, de 200 x 200 m nas áreas urbanas e 1 x 1 km nas áreas rurais, o IBGE disponibilizou em 2016 os seguintes produtos: uma aplicação interativa para visualização e consulta, que permite que o usuário desenhe qualquer área na tela e obtenha os dados sobre população (total e por sexo) e domicílios; mapas interativos, com as informações por biomas, regiões hidrográficas, clima, altitude e relevo; o Atlas Digital Brasil 1 por 1, que apresenta uma visão detalhada e inédita das principais variáveis coletadas pelo Censo Demográfico 2010 e, por fim, os arquivos digitais para serem utilizados em programas de geoprocessamento.

Mais trabalho precisa, ainda, ser feito, no que diz respeito à divulgação futura de dados censitários: explorar os dados para avaliar a possibilidade de divulgação da grade considerando as demais variáveis do conjunto universo do Censo Demográfico 2010 e definir a metodologia para o tratamento da confidencialidade; além dos cuidados especiais advindos do risco de revelação pela diferenciação geográfica envolvendo grade e setor censitário.

Acesso aos dados e a segurança

O acesso ao acervo institucional de dados e a segurança estão ligados às várias etapas do fluxo das informações, desde a coleta, passando pela apuração e a disseminação, até o usuário.

A necessidade de proteger as informações coletadas, armazenadas, analisadas, produzidas e disseminadas pelo IBGE com o uso de recursos de Tecnologias da Informação e Comunicação - TIC é crescente, tendo em vista as ameaças e vulnerabilidades às quais as informações estão expostas, em virtude da grande conectividade e disponibilidade das informações na rede.

A Política de Segurança da Informação e Comunicações - Posic do IBGE é o documento corporativo que define os princípios, diretrizes, responsabilidades e competências para garantir a confidencialidade, integridade, autenticidade e disponibilidade das informações e que norteia a segurança de informação no IBGE, estabelecendo quais controles de segurança a serem aplicados e, ainda, as responsabilidades e competências na aplicação, gerenciamento e monitoramento dos controles definidos (IBGE, 2015).

A implementação da Posic é sustentada por planos e políticas, normas e ordens de serviço, alinhados às diretrizes estabelecidas na mesma. A título de exemplo cabe mencionar: Política de acesso à Internet; Política do uso do correio eletrônico; Ordem de Serviço de controle de acesso físico – que define os controles de acesso aos Centros de Processamento de Dados - CPD e as condições físicas a serem observadas; Ordem de Serviço de acesso lógico - que determina os diferentes tipos de credenciais, as formas de administração (concessão, revogação e revisão de privilégios de acesso) e as regras para uso das credenciais para acesso aos Ativos de Tecnologia da Informação; Ordem de Serviço de Senhas; Ordem de Serviço para Armazenamento de Dados.

---

[1] Nos setores com menos de cinco domicílios particulares permanentes foram omitidos os valores da maioria das variáveis, mantendo apenas as variáveis estruturais, tais como: a identificação das subdivisões geográficas, o número de domicílios e a população por sexo (METODOLOGIA..., 2016).

Outros procedimentos de segurança e integridade dos processos e das bases de dados estatísticos são: segurança física das instalações, defesa contra *hackers*, transmissão dos dados com criptografia, remoção de atributos de identificação e proteção contra adulteração dos arquivos de dados; realização de testes dos instrumentos e procedimentos de coleta; realização de testes dos sistemas de coleta e apuração; e constituição do Grupo de Trabalho sobre Segurança das Informações Estatísticas - GSIE.

Avaliação e linhas de trabalho no IBGE

Do que foi visto antes, constata-se que apesar das realizações envolvendo o tema sigilo das informações, ainda há muito por fazer no IBGE.

Numa avaliação da situação do IBGE, de acordo com Silva (2017), foram destacados os seguintes pontos críticos: arcabouço legal antiquado e desatualizado frente às novas tecnologias e fontes de informação; baixo nível de capacitação técnica no tema; práticas 'simples', sem padronização, baixo nível de automação de soluções; melhorias recentes na situação da 'segurança física' ainda insuficientes; riscos de revelação não avaliados de forma sistemática; níveis de utilidade aquém do desejável em várias áreas.

Ainda de acordo com Silva (2017) as prioridades para investimento no IBGE recaem em: aumentar capacitação técnica no tema; revisar e atualizar arcabouço legal; aprofundar melhorias na situação da 'segurança física'; avaliar de forma sistemática os riscos de revelação para cada 'linha de produtos'; adotar soluções padronizadas e automatizadas para proteção de confidencialidade; adotar novas práticas que permitam ampliar a utilidade dos dados e resultados, sem ampliar em demasia os riscos de revelação.

Trata-se de uma oportunidade para investir em estudos de métodos e implantação de sistemas automatizados para a proteção do sigilo no IBGE, com base nas experiências dos institutos nacionais de estatística mais avançados. Outro fator a ser considerado diz respeito à transparência para a sociedade, através da sistematização, documentação e disponibilização dos procedimentos adotados na proteção do sigilo em todas as etapas do processo de produção das informações e em cada linha de produtos. Para tanto, as ações implicam em mudanças na cultura institucional, requerendo atuação em todos os níveis funcionais, sem exceção, dos envolvidos no processo de produção das informações.

Esses desafios e perspectivas conduzem a passos importantes a serem conquistados no sentido de avançar e planejar os dados do futuro e abordagens de acesso. É possível identificar ações e linhas de trabalho a serem implementadas, a saber: pesquisar, estudar, conscientizar, diagnosticar, organizar diretrizes, documentar, preparar normas e procedimentos, automatizar procedimentos, recomendar, implementar, monitorar, dentre outras.

Por fim, cabe registrar a atenção a ser dada com a relação de compromisso da proteção do sigilo *versus* acesso, uso de registros administrativos, ligação de bases de dados e o georreferenciamento de informações.

Referências

BRASlL. Lei n. 5.534, de 14 de novembro de 1968. Dispõe sobre a obrigatoriedade de prestação de informações estatísticas e dá outras providências. *Diário Oficial [da] República Federativa do Brasil,* Brasília, DF, ano 106, n. 222, 18 nov. 1968. Seção 1, p. 9985. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L5534.htm>. Acesso em: set. 2017.

DOYLE, P. et al. (Ed.). *Confidentiality, disclosure and data access:* theory and practical applications for statistical agencies. Amsterdam: Elsevier Science, 2001. 452 p.

IBGE. *Código de boas práticas das estatísticas do IBGE*. Rio de Janeiro, 2013. 48 p. Disponível em: <ftp://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>. Acesso em: set. 2017.

_____. *Código de ética profissional do servidor público do IBGE*. Rio de Janeiro: IBGE, 2014. 18 p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv98031.pdf>. Acesso em: set. 2017.

_____. *Política de segurança da informação e comunicações do IBGE* - POSIC 2016. Rio de Janeiro, 2015, 35 p. Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/Politica_de_Seguranca_da_Informacao_e_Comunicacoes_2016.pdf>. Acesso em: set. 2017.

_____. Portaria n. 485, de 06 de dezembro de 2013. Cria o Grupo de Trabalho sobre Sigilo de Informações em Grades Estatísticas. *Boletim de Serviços,* Rio de Janeiro, n. 2690, p. 3, 6 dez. 2013.

_____. *Resolução do Conselho Diretor n. 2, de 19 de fevereiro de 2001*. Cria Comitê de Sigilo. Rio de Janeiro, 2001.

_____. *Resolução do Conselho Diretor, n. 26, de 21 de outubro de 2015*. Cria o Comitê de Segurança da Informação e Comunicações do IBGE - CSI. Rio de Janeiro, 2015.

_____. *Resolução do Conselho Diretor, n. 7, de 4 de junho de 2003*. Cria o Comitê de Avaliação de Acesso a Microdados Não-Desidentificados. Rio de Janeiro, 2003.

GRUPO DE TRABALHO SOBRE SIGILO DE INFORMAÇÕES EM GRADES ESTATÍSTICAS. *Relatório...* Rio de Janeiro: IBGE, 2015. 78 p.

KOELLER, P.; VILHENA, F.; ZACHARIAS, M. L. B. *Disponibilização de acesso a microdados em institutos nacionais de estatística:* experiência de países selecionados e Eurostat. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2013. 29 p. (Textos para discussão, n. 44). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv64305.pdf>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2010. 2. ed. Rio de Janeiro: IBGE, 2016. 720 p. (Série relatórios metodológicos, v. 41). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv95987.pdf>. Acesso em: set. 2017.

SILVA, P. L. do N. *Sigilo e disseminação de dados:* em busca do equilíbrio sustentável. Rio de Janeiro: IBGE, 2017. Trabalho apresentado no Seminário IBGE Especial: Sigilo Estatístico, realizado no Rio de Janeiro, 2017.

UNITED NATIONS. *Handbook of statistical organization:* the operation and organization of a statistical agency. 3rd. ed. New York: United Nations, Statisticts Division, 2003. 205 p. Disponível em: <https://unstats.un.org/unsd/publication/SeriesF/SeriesF_88E.pdf> Acesso em: set. 2017.

_____. General Assembly. *Fundamental principles of official statistics.* New York, 2014. 2 p. Adotado pela Resolução A/RES/68/261 da Assembléia Geral das Nações Unidas, Nova Iorque, em 29 de janeiro de 2014. Disponível em: <https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf>. Acesso em: set. 2017.

ZACHARIAS, M. L. B.; BIANCHINI, Z. M.; ALBIERI, S. *Aperfeiçoamentos no processo de acesso a microdados restritos no IBGE*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2013. 6 p. Disponível em: <ftp://ftp.ibge.gov.br/Artigos_e_Apresentacoes/CES_2013_MariaZacharias_et_ZeliaBianchini_et_SoaniaAlbieri_portugues.pdf>. Acesso em: set. 2017.

Comentários sobre alguns projetos de consultoria executados na Coordenação de Métodos e Qualidade do IBGE

Djalma Galvão Carneiro Pessoa

Introdução

Inicialmente agradeço o convite para participar da elaboração deste volume comemorativo dos 40 anos da área de metodologia do Instituto Brasileiro de Geografia e Estatística - IBGE.

Durante vários anos trabalhei como consultor da Coordenação de Métodos e Qualidade, podendo ver de perto os relevantes serviços prestados por esta Coordenação ao IBGE e à sociedade em geral.

Neste capítulo descrevo, em ordem cronológica de execução, seis projetos em que participei como consultor, selecionando os que tiveram, no meu entender, maior alcance. Todos os projetos descritos foram executados em colaboração com técnicos da Coordenação, que são devidamente mencionados no texto.

Inicio mencionando trabalho em conjunto com Pedro Silva na elaboração do livro *Análise de dados amostrais complexos* (PESSOA; SILVA, 1998a), que serviu de material auxiliar do minicurso sobre técnicas de análise de dados amostrais complexos, ministrado durante o Simpósio Nacional de Probabilidade e Estatística - Sinape de 1998. Esse material foi de grande utilidade para a divulgação de técnicas de análise de dados de pesquisas amostrais, entre os usuários das pesquisas produzidas IBGE, por ressaltar a importância de se considerar nas análises estatísticas as características do plano amostral usado na coleta dos dados da pesquisa. Nessa mesma linha, menciono ainda os artigos *Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões* (PESSOA; SILVA; DUARTE, 1997) e *Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral* (SILVA; PESSOA; LILA, 2002).

Outro projeto que destaco é a *Imputação de rendimentos no questionário da amostra do censo demográfico 2000* (PESSOA; MOREIRA; SANTOS, 2004), trabalho conjunto com Guilherme Moreira e Alexandre Santos. Sua implementação computacional foi em linguagem S e para sua execução foi desenhado, pela Diretoria de Informática do IBGE, um sistema no qual as rotinas de imputação, utilizando técnicas Não-Paramétricas de Árvores de Regressão, foram executadas em microcomputador no modo *batch* e as

tarefas de gerenciamento de banco de dados em computador de grande porte, sem haver necessidade de interferência de operador. Essa foi a primeira vez em que a imputação da variável numérica rendimento foi executada no Censo Demográfico do IBGE. Anteriormente, o IBGE havia utilizado o software DIA apenas para imputação de variáveis categóricas (BECKER; CHAMBERS; WILKS, 1988; GARCÍA RUBIO; VILLÁN CRIADO, 1988).

Pessoa (2007) faz um resumo da técnica de imputação de rendimentos apresentada no Satellite Meeting on Innovative Methodologies for Censuses in the New Millennium, na 56ª sessão da conferência organizada pelo International Statistical Institute - ISI, realizada em Lisboa, 2007.

Com a finalidade de programar, em linguagem R, rotinas de análise de dados contidas no software SUDAAN, descritas por Shah e outros (1995), foi construído, em colaboração com Guilherme Moreira, o pacote de funções Análise de Dados Amostrais Complexos - *adac*. À época ainda não havia disponível o pacote *survey* (LUMLEY, 2017). O pacote de funções *adac* foi um instrumento importante de divulgação do sistema R no IBGE, sistema que já era comumente utilizado na área acadêmica.

Por solicitação do então Ministério do Desenvolvimento Social e Agrário[1], para subsidiar informações ao Programa Bolsa Família, o IBGE elaborou o *Mapa de pobreza e desigualdade: municípios brasileiros 2003* (MAPA..., 2008), com o objetivo de estimar a taxa de pobreza de todos os municípios do país. A partir das pesquisas do IBGE, não é possível obter estimativas diretas dessas taxas com precisão adequada. Para isso, foi utilizada a técnica de estimação em pequenas áreas apresentada por Elbers, Lanjouw e Lanjouw (2003). Posteriormente, surgiu uma metodologia alternativa proposta por Molina e Rao (2010). Em trabalho conjunto com Debora Souza, Nícia Brendolin, Viviane Quintaes e Solange Onel (SOUZA et al., 2014), executamos um estudo de simulação comparando os métodos propostos por Elbers, Lanjouw e Lanjouw (2003) e Molina e Rao (2010).

Comumente, não é possível utilizar a técnica de linearização de Taylor para estimar a variância de índices de pobreza e de concentração de renda, por não serem satisfeitas as condições de regularidade exigidas para aplicação da técnica. Além disso, os tamanhos dos bancos de dados envolvidos tornam computacionalmente proibitiva a utilização de técnicas de replicação como *bootstrap* e *jackknife*. Uma alternativa que tem se mostrado frutífera é utilizar o método de linearização por meio de funções de influência, proposto por Deville (1999) para obtenção de estimativas de variância. Em trabalho conjunto com André Costa essa metodologia foi utilizada para estimar a variância do índice FGT de medidas de pobreza (COSTA; PESSOA, 2003). Desse projeto da Coordenação de Métodos e Qualidade, originou-se o pacote *convey* de funções do R (PESSOA; DAMICO; JACOB, 2017), que utiliza a técnica de linearização por funções de influência para estimar variâncias de um grande número de medidas de pobreza e de concentração de renda.

A cada Censo Demográfico, desde 1970, o IBGE realiza uma Pesquisa de Avaliação da Cobertura da Coleta do Censo - PA. Pela primeira vez, em 2010, foi adotado pelo IBGE um processo automatizado de pareamento de registros de domicílios e de pessoas no Censo 2010 e na PA. Anteriormente, o IBGE utilizava um procedimento de comparação visual de questionários e listagens. Em trabalho conjunto com Fábio Farias e Vinicius Xavier foram desenvolvidos *scripts* do R, descritos por Pessoa, Farias e Xavier (2012) para implementar o pareamento probabilístico do Censo e da PA, utilizando o pacote *RecordLinkage* (BORG; SARIYAR, 2016).

Finalmente, ressalto que os projetos mencionados nesse artigo representam uma pequena parcela dos trabalhos em que participei como consultor de Estatística na Coordenação de Métodos e Qualidade e cuja seleção reflete uma escolha pessoal de temas.

[1] Atual Ministério do Desenvolvimento Social.

Curso ministrado no 13º Simpósio Nacional de Probabilidades e Estatística - Sinape

Em trabalho conjunto com Pedro Silva, foi submetido à Comissão Organizadora do 13º Sinape o projeto de um curso sobre Análise de Dados Amostrais Complexos - *adac*. A proposta foi aceita e o material abordado no curso foi organizado em um texto, cujos objetivos foram descritos como a seguir:

> Este livro trata de problema de grande importância para os usuários de dados obtidos através de pesquisas amostrais por agências produtoras de informações estatísticas. Tais dados são comumente utilizados em análises descritivas envolvendo o cálculo de estimativas para totais, proporções, médias e razões, nas quais, em geral, são devidamente considerados os pesos distintos das observações e o planejamento da amostra que lhes deu origem.
>
> Outro uso desses dados, denominado secundário, é a construção de modelos, feita geralmente por analistas que trabalham fora das agências produtoras dos dados. Neste caso, o foco é, essencialmente, estabelecer a natureza de relações ou associações entre variáveis. Para isso, a estatística clássica conta com um arsenal de ferramentas de análise, já incorporado aos principais pacotes estatísticos disponíveis. O uso desses pacotes se faz, entretanto, sob condições que não refletem a complexidade usualmente envolvida nas pesquisas amostrais de populações finitas. Em geral, partem de hipóteses básicas que são válidas quando os dados são obtidos através de amostras aleatórias simples com reposição (AASC). Tais pacotes estatísticos não consideram os seguintes aspectos relevantes no caso de amostras complexas:
> i) probabilidades distintas de seleção das unidades;
> ii) conglomeração das unidades;
> iii) estratificação;
> iv) não-resposta e outros ajustes.
>
> As estimativas pontuais de parâmetros da população são influenciadas por pesos distintos das observações. Além disso, as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos. Ao ignorar esses aspectos, os pacotes tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais. (PESSOA; SILVA, 1998a, p. 9-10)

O livro teve três objetivos principais:

1. Ilustrar e analisar o impacto das simplificações feitas ao utilizar pacotes usuais de análise de dados quando esses dados são provenientes de pesquisas amostrais complexas;
2. apresentar uma coleção de métodos e recursos computacionais disponíveis para análise de dados amostrais complexos, equipando o analista para trabalhar com tais dados, reduzindo assim o risco de inferências incorretas; e
3. ilustrar o potencial analítico de muitas das pesquisas produzidas por agências de estatísticas oficiais para responder questões de interesse, mediante uso de ferramentas de análise estatística agora já bastante difundidas, aumentando assim o valor adicionado dessas pesquisas.

Para alcançar tais objetivos, adotou-se uma abordagem fortemente ancorada na apresentação de exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também recursos de sistemas especializados. A comparação dos resultados das análises feitas das duas formas permite avaliar o impacto de não se considerar os pontos I a IV citados anteriormente em Pessoa e Silva (1998a, p. 9-10).

O livro[2] constou dos seguintes capítulos: Introdução; Referencial para inferência; Estimação baseada no plano amostral; Efeitos do plano amostral; Ajuste de modelos paramétricos; Modelos de regressão; Testes de qualidade de ajuste; Testes em tabelas de duas entradas; Agregação vs. desagregação; e Pacotes para análise de dados amostrais.

Considerando a ampla utilização do livro no país, julgamos que os objetivos descritos foram plenamente atingidos.

---

[2] Uma versão atualizada do texto está disponível em: <https://djalmapessoa.github.io/adac/>, na Internet.

Imputação da variável rendimento no Censo Demográfico 2000

O objetivo desse projeto foi elaborar um procedimento de imputação de rendimento para o responsável do domicílio no universo do Censo Demográfico 2000 bem como para diferentes tipos de variáveis de rendimentos dos moradores dos domicílios na amostra.

Na coleta dos dados do Censo Demográfico 2000 foram usados dois tipos de questionário: longo e curto. O longo, mais detalhado, só foi aplicado em 10% dos domicílios de municipalidades com mais de 15 000 habitantes e em uma amostra de 20% dos domicílios para municipalidades com menos de 15 000 habitantes. O questionário longo continha todas as questões incluídas no questionário curto.

Entre as questões no questionário curto havia uma questão referente ao valor do rendimento bruto em Reais ($) originado de todas as fontes, recebido pela pessoa responsável pelo domicílio ou por indivíduos em domicílios coletivos, durante o mês de julho de 2000.

No questionário longo havia um grupo de sete perguntas sobre valores e fontes de rendimentos de moradores com idades iguais a ou maiores que 10 anos no mês de referência:

1. rendimento bruto do trabalho principal;
2. rendimento bruto de outros trabalhos;
3. aposentadoria e pensão;
4. rendimentos de aluguel;
5. doações de fontes privadas;
6. doações do governo; e
7. outros.

O rendimento total era a soma dos rendimentos de todas as fontes.

Quando há perda de informação causada por não-resposta, as estimativas baseadas apenas nos valores disponíveis podem ser tendenciosas. Uma forma de diminuir a tendenciosidade é através da imputação apropriada dos valores de rendimentos para os não-respondentes. No Censo Demográfico 2000, foi adotada uma metodologia que permitia a correção da tendência e, ao mesmo tempo, possibilitava um processamento rápido e automatizado da grande massa de dados produzida. O método de imputação explorou a relação existente entre as variáveis de rendimentos e variáveis auxiliares cujos valores eram conhecidos para todos os domicílios pesquisados, e a partir dessa relação foram imputados valores para os rendimentos faltantes. A metodologia usada é baseada na técnica de Árvores de Regressão (BREIMAN et al., 1984). A grande assimetria e a ocorrência de valores zero na distribuição das variáveis de rendimentos influenciaram na escolha dessa técnica não-paramétrica de regressão.

Análises baseadas em dados do Censo Demográfico 1991 mostraram que a não-resposta das variáveis de rendimentos não era completamente ao acaso, do tipo *Missing Completely at Random* - MCAR, segundo Little e Rubin (2002, p. 11-13). Por exemplo, a taxa de não-resposta tendia a aumentar com o nível de educação do responsável pelo domicílio. Geralmente, valores de variáveis auxiliares que indicavam rendimentos maiores eram acompanhadas por taxas maiores de não-resposta. Dessa forma, estimativas baseadas apenas nas observações disponíveis poderiam ser tendenciosas, indicando assim a necessidade de uma imputação apropriada. Estudos semelhantes foram também executados para as variáveis de rendimentos do questionário longo, e o mesmo tipo de não-resposta diferencial foi observado.

No caso de não-resposta de item, geralmente é preferível tratar a não-resposta por meio da imputação dos valores faltantes, devido a sua simplicidade no processamento subsequente dos dados, particularmente quando eles precisam ser publicados na forma de microdados.

Para a imputação de valores faltantes nas variáveis de rendimentos tanto para o questionário longo como para o curto, foi selecionado um conjunto diversificado de variáveis explanatórias para os rendimentos que poderia fornecer poder preditivo para todo o país. Para o questionário curto foi usado o seguinte conjunto de variáveis preditivas no processo de imputação:

- idade em anos completos do responsável pelo domicílio;
- anos de estudo do responsável pelo domicílio;
- sexo do responsável pelo domicílio;
- espécie de domicílio (particular permanente, particular improvisado ou coletivo);
- tipo de domicílio (casa, apartamento ou cômodo);
- número total de moradores no domicílio;
- número total de empregados domésticos vivendo no domicílio;
- condição do domicílio (próprio já pago, próprio ainda pagando, alugado, ...);
- quantidade de banheiros no domicílio;
- indicador de existência de banheiro no domicílio;
- tipo de abastecimento de água (rede geral, poço/nascente ou outra);
- tipo de canalização de água (canalizada em pelo menos um cômodo; canalizada só na propriedade ou terreno; não canalizada);
- tipo de escoadouro (rede geral de esgoto ou pluvial, fossa séptica, fossa rudimentar, ...);
- tipo de coleta de lixo (coletado por serviço de limpeza, colocado em caçamba de serviço de limpeza, ...); e
- tipo de setor censitário (rural, urbano).

Da mesma forma, entre as variáveis disponíveis no questionário longo, selecionamos algumas para explicar os vários tipos de rendimentos dos moradores. Infelizmente, foi apenas possível achar variáveis explanatórias com poder preditivo satisfatório para o rendimento bruto da atividade principal e para o total das variáveis de rendimentos. Para as outras categorias de rendimentos, foram usadas as mesmas classes de imputação obtidas para a variável de rendimento total.

A seguir apresentamos o conjunto de variáveis selecionadas no processo de imputação para as variáveis de rendimentos do questionário longo:

- condição do morador na atividade principal;
- atividade principal da instituição ou entidade de negócio, firma, empresa onde o morador trabalhava;
- uma combinação do número de banheiros e da existência de instalações sanitárias em um domicílio sem sanitários;
- disponibilidade de certos aparelhos no domicílio, tais como videocassete, máquina de lavar, microondas e computador;
- número de aparelhos de televisão;
- número de automóveis para uso privado no domicílio;
- número de aparelhos de ar condicionado;
- relação do morador com o responsável pelo domicílio;
- idade do morador;
- número de anos completos de escolarização;
- indicador de se o morador tinha salários provenientes do emprego principal; e
- indicador de se o morador tinha rendimentos de aposentadoria/pensão.

A metodologia de imputação usada tanto para o formulário longo quanto para o curto supôs que a perda de dados de rendimentos foi do tipo *Missing at Randon* - MAR,

conforme Little e Rubin (2002, p. 11-13), e baseou-se na aplicação do método não-paramétrico de árvores de regressão, *Classification and Regression Trees* - CART, apresentado por Breiman e outros (1984). Resumidamente, esse método gera uma árvore binária cujos nós terminais definem uma partição dos dados mais homogênea com relação aos valores da variável resposta na regressão. Os caminhos entre o nó raiz e o nó terminal são determinados por partições binárias repetidas, baseadas nas variáveis explicativas que podem ser numéricas ou categóricas. Os nós terminais gerados no ajuste das árvores foram tomados como classes de imputação. Dentro de cada classe, a imputação foi executada pelo método *hot deck*.

Para implementar todo o procedimento foram programadas funções em linguagem S e usada o pacote *tree* do S-Plus. Na aplicação da técnica, a totalidade de domicílios foi agrupada em 526 lotes para o formulário curto e 216 lotes para as variáveis do formulário longo. Os lotes foram os mesmos que haviam sido definidos na fase de crítica e imputação executada pelo software DIA. Para cada lote, uma árvore de regressão foi construída. Devido à impossibilidade de avaliar a curva de *deviance* para todas as árvores, seus tamanhos foram fixados em termos do número máximo de partições e do tamanho mínimo das classes. Baseado em experimentos feitos com base nos dados do Censo Demográfico 1991, o número máximo de partições foi fixado em 25, e o tamanho mínimo da classe foi 100.

As taxas de não-resposta nos lotes variaram de 0,54% a 6,24%, com valor da mediana 1,60%. A taxa global de não-resposta foi 1,75%. Entre as variáveis auxiliares usadas para definir as classes de imputação, aquelas diretamente relacionadas ao responsável pelo domicílio apresentaram maior poder de explicação para a variável rendimento. A variável número de banheiros no domicílio, diretamente relacionada ao domicílio, ficou entre os melhores preditores. As duas variáveis: nível de instrução e número de banheiros geraram a primeira partição em 87,00% e 94,70% das árvores, respectivamente. Por outro lado, as seguintes variáveis, mencionadas anteriormente no formulário curto, não foram usadas na construção das árvores em mais de 80,00% dos lotes: espécie de domicílio (particular permanente, particular improvisado ou coletivo); número total de moradores no domicílio; número total de empregados domésticos vivendo no domicílio; tipo de abastecimento de água (rede geral, poço/nascente ou outra); tipo de coleta de lixo (coletado por serviço de limpeza, colocado em caçamba de serviço de limpeza, ...); e tipo de setor censitário (rural, urbano).

A fim de controlar a implementação do processo, em cada lote foi aplicado um teste de Kolmogov-Sminorv para testar a hipótese de igualdade da distribuição do rendimento, antes e depois da imputação, em cada classe de imputação. O lote era aceito só quando não havia rejeição em nenhuma classe de imputação.

Outro aspecto no processo de imputação era relacionado ao tratamento de valores zero de rendimentos. Uma opção considerada foi construir uma variável de rendimento binária com níveis zero e diferente de zero, e então usar o sistema DIA para imputação. Contudo, foi observado que as variáveis auxiliares preditivas no questionário curto tinham baixo poder preditivo para essa variável binária. Dessa forma, foi decidido tratar a imputação de valores zero nas variáveis de rendimento como parte do procedimento de imputação de variáveis numéricas. Por outro lado, para as variáveis de rendimento derivadas no questionário longo, as variáveis auxiliares foram consideradas como tendo suficiente poder preditivo para predizer valores zero. Assim, a imputação de valores zero dessas variáveis foi feita por meio do sistema DIA. Quando foi observado valor faltante em mais de uma variável, os valores para imputação não foram selecionados do mesmo doador, pois havia grande probabilidade de não encontrar um doador tendo todas os valores faltantes diferentes de zero. Quando não havia doador para alguma categoria de variável na classe de imputação, um doador era selecionado no subgrupo de registros que originava a classe.

Para implementar o procedimento de imputação no Censo Demográfico 2000, foi projetado um sistema para baixar automaticamente os arquivos de dados do ambiente de grande porte para um microcomputador, onde o programa de aplicação em linguagem S estava instalado, e depois de executada a imputação os arquivos resultantes eram

enviados de volta para o sistema de grande porte. Finalmente, mencionamos alguns aspectos da metodologia de imputação usada que poderiam sofrer alterações. Por exemplo, o procedimento adotado para lidar com *outliers* foi considerado ingênuo e deveria ser aperfeiçoado.

Um procedimento alternativo ao utilizado foi proposto por Chambers, Hentges e Zhao (2004). Neste artigo, os autores sugerem algumas formas de estender as técnicas de árvore de regressão para o caso de resposta multivariada, o que poderia permitir o uso do mesmo doador para o caso de mais de uma variável faltante por pessoa.

Pacote *adac* de funções do R

Na análise de dados de pesquisas amostrais complexas é fundamental considerar os aspectos especificados por Pessoa e Silva (1998a, p. 9-10): "i) probabilidades distintas de seleção das unidades; ii) conglomeração das unidades; iii) estratificação; iv) não-resposta e outros ajustes". Como foi ressaltado, os softwares estatísticos são geralmente adequados para lidar com dados obtidos por Amostras Aleatórias Simples Com Reposição - AASC. Os dados amostrais produzidos por agências de estatísticas oficiais, como o IBGE, são obtidos por meio de amostras complexas. Já há algum tempo, vários programas especializados estão disponíveis para a análise desses dados, entre eles: SUDAAN, STATA, WesVarPc, PCCarp. Uma revisão desses softwares foi realizada por Pessoa e Silva (1998b).

O R é um software de código aberto que fornece um ambiente flexível para análise de dados e contém ferramentas para implementar de forma elegante muitas ideias estatísticas. Pode ser estendido de forma natural por procedimentos escritos pelos usuários, usando sua própria linguagem ou outras como C e FORTRAN. Ao longo do tempo, a comunidade de usuários vem contribuindo com uma profusão de funções para implementar procedimentos estatísticos de interesse. Atualmente é o software estatístico mais utilizado para a pesquisa estatística e para a distribuição de novos métodos. O R possui várias bibliotecas de funções especializadas para a análise de dados amostrais complexos, com destaque para o pacote *survey*.

O pacote de funções *adac* foi um esforço pioneiro desenvolvido na Coordenação de Métodos e Qualidade para a análise de dados amostrais complexos em linguagem R, contendo a quase totalidade dos procedimentos do SUDAAN descritos por Shah e outros (1995). No pacote *adac* foram incluídas funções para estimar: totais, razões e percentis, com as respectivas medidas de precisão, considerando vários planos amostrais. Foram ainda disponibilizadas funções para análise de tabelas de contingência, usando a estatística de Pearson com correções de primeira e segunda ordem de Rao-Scott e o teste de Wald para as hipóteses de homogeneidade e independência. Há também funções para o ajuste de modelos de regressão linear e logística com base em dados de pesquisas amostrais.

A seguir listamos as principais funções do pacote *adac* com suas finalidades:

- **cdf.samp** - função de distribuição acumulada;
- **chisq** - estatística de teste qui-quadrado de Pearson com correções de Rao-Scott de primeira e segunda ordem;
- **crosstab** - estatísticas de testes de independência e de homogeneidade para tabelas de contingência;
- **logist** - ajuste de regressão logística;
- **quant.samp** - quantis da distribuição;
- **ratio e ratiopos** - razão de dois totais com pós-estratificação dos pesos do plano amostral;
- **regress** - ajuste de regressão linear;
- **strwor.var** - variância de totais para plano amostral estratificado com conglomerados primários selecionados sem reposição;
- **strwr.var** - variância de totais para plano amostral estratificado com conglomerados primários selecionados com reposição

- **total.grp** - total por domínios;
- **total.pos** - total com pós-estratificação de pesos;
- **waldtest** - estatísticas de teste de Wald;
- **wor2.var** - matriz de covariância para o plano amostral com dois estágios, sendo o primeiro sem reposição; e
- **wr2.var** - matriz de covariância para o plano amostral com dois estágios, sendo o primeiro com reposição.

O pacote *adac* foi instrumental na divulgação do ambiente R no IBGE, e disponível antes de o pacote *survey* ser lançado, quando a análise de dados de pesquisas no IBGE era feita exclusivamente por meio do pacote SUDAAN. Atualmente, é comum no IBGE a utilização do pacote *survey* do R e de programas em linguagem R específicos para a execução de tarefas especiais.

Estudo de comparação de técnicas de estimação em pequenas áreas

Esse projeto foi elaborado em conjunto com Debora Souza, Nícia Brendolin, Viviane Quintaes e Solange Onel e teve como objetivo comparar o desempenho de três métodos de estimação em pequenas áreas para estimar índices de pobreza em municípios de Minas Gerais. Detalhes sobre esse projeto são apresentados nesta publicação, no Capítulo **Indicadores de pobreza nos municípios de Minas Gerais: comparação de métodos de estimação em pequenas áreas**.

Estimação de variância de indicadores de pobreza e concentração de renda

Várias medidas de pobreza e de concentração são definidas por funções não-diferenciáveis de variáveis de rendimentos. Sendo assim, não é possível utilizar o método de linearização de Taylor para estimar suas variâncias. Uma alternativa é utilizar funções de influência como descritas por Deville (1999) e Osier (2009).

Como exemplos dessas medidas, mencionamos o índice de Gini e a família FGT de indicadores para medir pobreza, proposta por Foster, Greer e Thorbecke (1984). Anualmente, o índice de Gini é calculado pelo IBGE com base nos microdados da Pesquisa Nacional por Amostra de Domicílios - PNAD. O índice FGT tem sido utilizado pelo IBGE na elaboração do *Mapa de pobreza e desigualdade* (2008). Além de estimar essas medidas é importante divulgar suas correspondentes medidas de precisão. Para isso, é possível utilizar-se técnicas de replicação como *bootstrap*, *jackknife*, etc., que, contudo, envolvem esforço computacional intensivo e grande tempo de processamento.

Costa e Pessoa (2003) implementaram uma função em linguagem R para estimar a variância do FGT utilizando a técnica de linearização do indicador por função de influência, conforme proposto por Deville (1999). Os mesmos autores produziram, para uso interno na Coordenação de Métodos e Qualidade, uma função do R para estimar a variância do índice de Gini por meio de função de influência.

Essas atividades da Coordenação motivaram a construção do pacote *convey* de funções do R, lançado recentemente, para estimar índices de pobreza, medidas de desigualdade e índices multivariados a partir de dados de pesquisas amostrais complexas (JACOB; DAMICO; PESSOA, 2016). O pacote *convey*[3] funciona em articulação com o pacote *survey*

[3] Mais detalhes estão disponíveis em: <https://CRAN.R-project.org/package=convey>, na Internet.

do R e produz estimativas de variâncias de indicadores tanto pelo método de linearização, usando funções de influência, como por replicação, usando o método *bootstrap*.

Pareamento de registros na Pesquisa de Avaliação da Cobertura da Coleta - PA do Censo Demográfico 2010

Para avaliar a cobertura da coleta do Censo Demográfico 2010, foi realizada a Pesquisa de Avaliação da Cobertura da Coleta - PA. A partir dos dados da PA, foram estimados indicadores de omissão de domicílios e de pessoas no Censo Demográfico 2010. No cálculo dessas estimativas, foi necessário identificar nos arquivos de domicílios e de pessoas os pares de registros coincidentes nos dois levantamentos. Em 2010, foi utilizado pela primeira vez um método automático para identificação desses pares (pareamento automático). A partir dos dados da PA e do Censo foi gerado o Quadro 1.

Quadro 1 - Comparação de registros do Censo e da PA

		Pesquisa de Avaliação (PA)		
		Dentro	Fora	Total
Censo	Dentro	N_{11}	N_{12}	N_{1+}
	Fora	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N_{++}

Fonte: Comparações do Autor com base nos dados do Censo Demográfico 2010 e Pesquisa de Avaliação da Cobertura da Coleta.

Utilizando as contagens N_{11}, N_{12} e N_{21}, foram estimados os indicadores de omissão desejados. As celas da tabela foram obtidas a partir do pareamento, realizado em três fases: 1) automática; 2) assistida e 3) de reconciliação.

Na fase automática, foi utilizado um método probabilístico de pareamento, onde se calculam escores que dependem das probabilidades de concordância e discordância de variáveis nos pares de registros.

Na implementação dessa fase, foi utilizado o pacote *RecordLinkage* do R, que contém funções para estimar os escores necessários, utilizando o algoritmo EM. A decisão de parear foi tomada quando o escore ultrapassava um patamar fixado. Pessoa, Farias e Xavier (2012) detalham o procedimento utilizado.

No pareamento com base num patamar para escores, um registro de um arquivo poderia vir a ser pareado com vários registros do outro arquivo (associação m:n). Inicialmente, a associação 1:1 de registros foi obtida por meio do algoritmo simplex. Essa abordagem foi descartada por envolver uso intensivo de memória. A associação 1:1 de registros pode ser reformulada como um problema de designação (*assignement*), e a solução pode ser obtida por meio da função *solve_LSAP* do pacote *clue* do R (por sugestão de José André Brito e Flávio Montenegro). Foi desenvolvido um *script* do R implementando todos os passos do pareamento pela metodologia de Fellegi-Sunter e posteriormente executando a associação 1:1 de registros.

As variáveis utilizadas no pareamento de domicílios foram: tipo de logradouro, título e nome concatenados; parte de valor do número; complemento; primeiro nome do responsável concatenado com o primeiro nome do cônjuge (ou na ordem inversa); último nome do responsável pelo domicílio; total de homens; e total de mulheres.

No pareamento de pessoas foram usadas: primeiro nome; último nome; idade.

Na comparação de variáveis de tipo *string* foi usada a medida de similaridade de Jaro-Winkler, tendo sido fixados patamares de valores para cada variável, usados para dicotomizar os valores da similaridade em 1 e 0.

O pareamento de pessoas foi feito dentro dos domicílios pareados e, diferentemente do pareamento de domicílios, os pesos de concordância foram pré-definidos, conforme Quadro 2.

Quadro 2 - Pesos de concordância

Concordância nas variáveis	Peso
Idade, primeiro e último nome	1
Primeiro nome e Idade	0,8
Primeiro nome e último nome	0,7

Fonte: O Autor com base nos dados do Censo Demográfico 2010 e Pesquisa de Avaliação da Cobertura da Coleta.

Depois da atribuição de escores aos pares de pessoas dentro do domicílio, foi feita uma redução 1:1. O Pareamento de todas as pessoas dentro do domicílio serviu de indicativo de que o pareamento de domicílio foi correto. Para os domicílios em que, pelo menos, uma pessoa não foi pareada, foi feita a verificação do pareamento de domicílio. A partir dessa verificação, obteve-se a estimativa da probabilidade de falsos positivos de 1,82% (SILVA; FREITAS; PESSOA, 2015).

Referências

BECKER, R. A.; CHAMBERS, J. M.; WILKS, A. R. *The new S language.* London: Chapman & Hall, 1988. 702 p.

BORG, A; SARIYAR, M. *RecordLinkage*: record linkage in R. R package version 0.4-10. [S.l.]: Comprehensive R Archive Network - CRAN, 2016. Disponível em: <https://CRAN.R-project.org/package=RecordLinkage>. Acesso em: set. 2017.

BREIMAN, L. et al. *Classification and regression trees.* Belmont, California: Wadsworth International Group, 1984. 368 p.

CHAMBERS, R.; HENTGES, A.; ZHAO, X. Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society.* Series A, Statistics in Society. London: Royal Statistical Society, v. 167, n. 2, p. 323-339, May 2004.

COSTA, A. W. N.; PESSOA, D. G. C. *Estimação de variância para o estimador de índice FGT de medidas de pobreza.* Rio de Janeiro: IBGE, 2003. Não publicado.

DEVILLE, J. C. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology,* Ottawa: Statistics Canada, v. 25, n. 2 p. 193-203, Dec. 1999. Disponível em: <http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X19990024882&ObjType=47&lang=en&limit=0>. Acesso em: set. 2017.

ELBERS, C.; LANJOUW, J. O.; LANJOUW, P. Micro-level estimation of poverty and inequality. *Econometrica,* New York: The Econometric Society, v. 71, n. 1, p. 355-364, Jan. 2003.

FOSTER, J.; GREER, J.; THORBECKE, E. A class of decomposable poverty measures. *Econometrica,* New York: The Econometric Society, v. 52, n. 3, p. 761-766, May 1984.

GARCÍA RUBIO, E.; VILLÁN CRIADO, I. *Sistema DIA:* sistema de detección e imputación automática de errores para datos cualitativos. Madrid: Instituto Nacional de Estadística - INE, 1988.

HORNICK, K. *Clue:* cluster ensembles. R package version 0.3-54. [S.l.]: Comprehensive R Archive Network - CRAN, 2017. Disponível em: <https://CRAN.R-project.org/package=clue>. Acesso em: set. 2017.

IHAKA R.; GENTLEMAN R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics,* Oxfordshire: Taylor & Francis, v. 5, n. 3, p. 299-314, May 1995.

JACOB, G.; DAMICO, A.; PESSOA, D. G. C. *Poverty and inequality with complex survey data.* 2016. Disponível em: <https://guilhermejacob.github.io/context/index.html#introduction>. Acesso em: set. 2017.

LITTLE, R. J. A; RUBIN, D. B. *Statistical analysis with missing data.* 2nd ed. Hoboken: Wiley, 2002. 408 p.

LUMLEY, T. *Survey:* analysis of complex survey samples. R package version 3.32-1. [S.l.]: Comprehensive R Archive Network - CRAN, 2017. Disponível em: <https://CRAN.R-project.org/package=survey>. Acesso em: set. 2017.

MAPA de pobreza e desigualdade: municípios brasileiros 2003. Rio de Janeiro: IBGE, 2008. 1 DVD.

MOLINA, I.; RAO, J. N. K. Small area estimation of poverty indicators. *The Canadian Journal of Statistics,* Ottawa: Statistical Society of Canada, v. 38, n. 3, p. 369-385, Sep. 2010.

OSIER, G. Variance estimation for complex indicators of poverty and inequality. *Survey Research Methods,* Southampton: European Survey Research Association, v. 3, n. 3, p. 167-195, 2009. Disponível em: <https://ojs.ub.uni-konstanz.de/srm/article/view/369>. Acesso em: set. 2017.

PESSOA, D. G. C. *Imputation for income variables in the brazilian demographic census.* 4 p. Trabalho apresentado na 56ª sessão do International Statistical Institute - ISI, realizada em Lisboa, 2007. Disponível em: <https://unstats.un.org/unsd/censuskb20/Attachments/2007BRA_ISI-GUIDf489fd67d4224cf2a243d2f0eb726e71.pdf>. Acesso em: set. 2017.

PESSOA, D. G. C.; DAMICO, A.; JACOB, G. *Convey:* income concentration analysis with complex survey samples. R package version 0.2.0. [S.l.]: Comprehensive R Archive Network - CRAN, [2017]. Disponível em: <https://CRAN.R-project.org/package=convey>. Acesso em: set. 2017.

PESSOA, D. G. C.; FARIAS, F. F. *Estimação de variância por linearização usando função de influência.* Trabalho apresentado na III Escola de Amostragem e Metodologia de Pesquisa - ESAMP, realizado em Juiz de Fora, 2011.

PESSOA, D. G. C.; FARIAS, F. F.; XAVIER, V. L. *Pareamento automático na pesquisa de avaliação da cobertura da coleta do censo demográfico de 2010.* Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2012. 50 p. (Textos para discussão, n. 41). Disponível em: <ftp://ftp.dpe.ibge.gov.br/gdi/texdisc/texdisc41-12.pdf>. Acesso em: set. 2017.

PESSOA, D. G. C.; MOREIRA, G. G.; SANTOS, A. R. *Imputação de rendimentos no questionário da amostra do censo demográfico 2000.* Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2004. 19 p.

PESSOA, D. G. C.; SILVA, P. L. do N. Análise de dados amostrais complexos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 13., 1998, Caxambu. *Anais...* São Paulo: Associação Brasileira de Estatística - ABE, 1998a. 170 p. Disponível em: <http://www.ie.ufrj.br/download/livro.pdf>. Acesso em: set. 2017.

_____. Pacotes para análise de dados amostrais. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 13., 1998, Caxambu. *Anais...* São Paulo: Associação Brasileira de Estatística - ABE, 1998b. Cap. 10, p. 155-170. Disponível em: <http://www.ie.ufrj.br/download/livro.pdf>. Acesso em: set. 2017.

PESSOA, D. G. C.; SILVA, P. L. do N.; DUARTE, R. P. N. Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. *Revista Brasileira de Estatística,* São Paulo: Associação Brasileira de Estatística, Rio de Janeiro: IBGE, v. 58, n. 210, p. 53-75, 1997. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

R: A language and environment for statistical computing. Version 3.4.2. Vienna, Austria: The R Foundation for Statistical Computing, 2017. Disponível em: <https://www.R-project.org/>. Acesso em: set. 2017.

SHAH, B. V. et al. *Statistical methods and mathematical algorithms used in SUDAAN.* Research Triangle Park, NC: Research Triangle Institute - RTI, 1995. 74 p.

SILVA, A. D.; FREITAS, M. P. S.; PESSOA, D. G. C. Assessing coverage of the 2010 Brazilian census. *Statistical Journal of the IAOS,* The Hague: International Association for Official Statistics, v. 31, p. 215-225, 2015. Disponível em: <http://content.iospress.com/download/statistical-journal-of-the-iaos/sji897?id=statistical-journal-of-the-iaos%2Fsji897>. Acesso em: set. 2017.

SILVA, P. L. do N.; PESSOA, D. G. C.; LILA, M. F. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. *Ciência & Saúde Coletiva,* Rio de Janeiro: Associação Brasileira de Saúde Coletiva - ABRASCO, v. 7, n. 4, p. 659-670, 2002. Disponível em: <http://www.scielo.br/pdf/csc/v7n4/14597.pdf>. Acesso em: set. 2017.

Planejamento, estimação e análise de dados em pesquisas por amostragem: desvendando a realidade brasileira com o "telescópio da estatística": trajetória até 1996

Pedro Luis do Nascimento Silva

Introdução

As pesquisas por amostragem desempenham na Estatística papel semelhante ao do telescópio na Astronomia. Sua invenção em meados do século XX e popularização na segunda metade do mesmo permitiram explorar, com rigor científico, vastas áreas do universo econômico e social, no Brasil e no resto do mundo, de forma antes não imaginada. Dados obtidos através de pesquisas por amostragem estão hoje na base de conhecimento de todas as ciências sociais.

Isso justifica estudar como o instrumento (pesquisas por amostragem) tem evoluído ao longo do tempo, e em particular como tem sido empregado no Instituto Brasileiro de Geografia e Estatística - IBGE. Este capítulo, escrito para a 4ª Conferência Nacional de Estatística - Confest, realizada no âmbito do 1º Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais, em 1996 mas nunca antes publicado, apresenta uma revisão abrangente, embora superficial, dos desenvolvimentos ocorridos no IBGE com relação à utilização de amostragem para obtenção e produção de informações e sua posterior análise até o ano de 1996. Ao mesmo tempo, procura rever progressos recentes da teoria de amostragem e indicar suas conexões e relevância para o IBGE e os usuários de suas informações. A amplitude do tema a ser coberto é grande, mas a ambição desse texto é modesta. Pretende-se aqui destacar os principais aspectos e ideias, e não fazer uma descrição ou revisão detalhada dos mesmos.

Pesquisas por amostragem envolvem três principais questões em termos de metodologia estatística: como selecionar amostras, como estimar as quantidades desconhecidas de interesse (parâmetros) e a precisão destas estimativas, e como formular e ajustar modelos com os dados das amostras observadas. Essas questões orientam a divisão deste capítulo. O tópico **Planejamento de pesquisas por amostragem** contém uma discussão da evolução dos aspectos ligados ao planejamento de pesquisas por amostragem. Após breve revisão histórica da introdução da amostragem no IBGE, as modificações introduzidas em várias pesquisas são revistas e os progressos alcançados indicados.

Os aspectos ligados à estimação de totais, médias e proporções nas pesquisas por amostragem do IBGE são discutidos no tópico **Estimação em pesquisas por amostragem**, destacando-se o emprego de estimadores que incorporam informações auxiliares na estimação e alternativas para tratamento de não-resposta.

A análise de dados provenientes de pesquisas por amostragem é a questão tratada no tópico **Análise de dados de pesquisas por amostragem**. São apontadas dificuldades causadas porque os dados são gerados por desenhos amostrais complexos. Argumenta-se que a utilização automática de métodos da estatística clássica, tais como análise de regressão e outros pode ser problemática. Algumas estratégias para lidar com o problema são apontadas bem como se argumenta que essa é uma área em que o IBGE e seus usuários, tem muito que aprender e que fazer para superar as dificuldades.

O tópico **Conclusões e desafios** destaca as conclusões do capítulo e indica desafios que o IBGE enfrentará nos anos vindouros para aproveitar plenamente o poderoso instrumento da amostragem em benefício do conhecimento preciso e detalhado da realidade brasileira.

Planejamento de pesquisas por amostragem

A utilização de amostragem nas pesquisas do IBGE já tem uma história considerável. Iniciando-se com a coleta de dados por amostra durante o Censo Demográfico 1960, o IBGE foi sendo progressivamente conquistado pelo uso de amostragem. A introdução da Pesquisa Nacional por Amostra de Domicílios - PNAD em 1967 é um dos marcos dessa história (METODOLOGIA..., 1981), seguindo-se o Censo Demográfico 1970 (METODOLOGIA..., 1983b).

Nos censos de população, a amostragem foi introduzida primordialmente para reduzir custos de coleta de dados mais detalhados (questionário longo), cuja obtenção para a população como um todo era muito cara. Em 1970 empregou-se amostragem também na pesquisa de avaliação da qualidade, para preparar tabulações avançadas e para gerar arquivos de uso público (pequena amostra com registros individuais desidentificados para servir ao público interessado em desenvolver suas próprias análises). A realização do Estudo Nacional da Despesa Familiar - ENDEF em 1974-1975 marcou mais uma etapa desse processo de consolidação no IBGE do emprego da amostragem como uma ferramenta indispensável ao processo de pesquisa e produção de informações (VASCONCELLOS, 1983).

Nessa primeira fase, o uso de amostragem pelo IBGE se restringiu a pesquisas em que o domicílio era a unidade de investigação. Nas pesquisas de instituições ou estabelecimentos, a norma era a realização de censos ou pesquisas exaustivas, tentando cobrir todas as unidades das respectivas populações de interesse. Uma exceção digna de nota foi a Pesquisa Industrial Mensal - Produção Física - PIM-PF iniciada em 1971, a qual pesquisa repetidamente uma amostra intencional das maiores indústrias (INDICADORES…, 1991). Entretanto essa pesquisa não adota esquema probabilístico de seleção da amostra como as demais. Outra característica desse primeiro período foi a ênfase no emprego de assessoria externa para implementação da amostragem nas pesquisas do IBGE, levando os desenhos amostrais então adotados a se parecerem bastante com aqueles usados nas pesquisas congêneres de outros países.

No fim dos anos 1970 e primeira metade dos anos 1980 várias razões levaram o IBGE a ampliar bastante o uso de amostragem em suas pesquisas. Entre essas destacaram-se a maior demanda por informações ágeis sobre vários setores da economia e da sociedade, aumentos significativos dos tamanhos das populações-alvo e a necessidade de pesquisar populações até então não cobertas.

Datam desse período a implantação do Sistema Nacional de Índices de Preços ao Consumidor - SNIPC (SISTEMA..., 1994), em 1979, e da Pesquisa Mensal de Emprego - PME (METODOLOGIA..., 1983a), em 1980, o uso de amostragem em pesquisas da indústria, começando com a Pesquisa Especial da Indústria - PEI (PESQUISA..., 1982), de 1981, prosseguindo com a Pesquisa Industrial Anual - PIA (PLANO..., 1982), desde 1981, o Índice de Preços ao Produtor - IPP de 1981 a 1985, e mais tarde com a reformulação

da Pesquisa Industrial Mensal de Dados Gerais - PIM-DG em 1985 (INDICADORES..., 1991). Vale mencionar ainda que no Censo Demográfico 1980 a amostragem foi empregada de forma ainda mais ampla que em 1970, com a introdução de esquemas de controle de qualidade baseados em inspeção por amostragem durante as etapas de coleta e processamento.

Essa fase de consolidação da amostragem como instrumento básico para a realização de pesquisas foi caracterizada por maior independência e autonomia técnica. O planejamento amostral passou a ser realizado por pessoal do próprio IBGE, incorporando inovações técnicas e refletindo o desenvolvimento de uma cultura própria. Data ainda dessa fase o reconhecimento da necessidade de documentar e disseminar entre os usuários as metodologias das pesquisas, em particular detalhes dos procedimentos amostrais adotados, tendo sido publicados relatórios descrevendo as metodologias adotadas nas principais pesquisas do IBGE[1].

Foram várias as inovações introduzidas na área de planejamento amostral. Nas pesquisas por estabelecimentos vale mencionar o emprego de desenhos amostrais com probabilidades desiguais de seleção. A PIA 1981 empregou o método de *Poisson sampling* (PLANO..., 1982) para selecionar respondentes produtores de uma lista de produtos pré-determinada a fim de investigar dados de produção física da indústria. Foi também empregada alocação ótima para distribuir a amostra entre os estratos na pesquisa de dados gerais. O IPP empregou amostragem sistemática com probabilidades desiguais para selecionar estabelecimentos nos quais seria observada a variação dos preços de produtos industriais.

Os desenhos amostrais das PIAs 1983 e 1984 foram modificados para refletir mudanças nos objetivos da pesquisa. Foram utilizados um novo cadastro base, nova estratificação e introduzidas amostras complementares de novos estabelecimentos captados através de registros administrativos (PLANO..., 1984; CABRAL, 1985). Na implantação do SNIPC empregou-se amostragem para montar o cadastro e selecionar a amostra da Pesquisa de Locais de Compra - PLC, os quais seriam visitados para a pesquisa de preços (PESQUISA..., 1994), e também na Pesquisa de Especificação de Produtos e Serviços - PEPS cujos preços seriam monitorados (SISTEMA..., 1994).

Outra tendência desse período foi o uso de computadores para o planejamento, seleção e expansão de amostras. Programas usando o sistema SAS foram desenvolvidos para implementar as estratégias amostrais de várias pesquisas (SILVA, 1989).

Essa fase de consolidação foi seguida por uma fase de aperfeiçoamento (após 1985), em que as estratégias amostrais de várias pesquisas foram revistas, buscando redução dos custos, maior eficiência, adequação a novos objetivos e incorporação de novos cadastros. Entre essas, vale mencionar a redução da amostra da PNAD desde 1986 em cerca de 44% (MENEZES et al., 1986), e da PME desde 1988 em 30% (SILVA; MOURA, 1988).

Também a PIA teve seu esquema amostral reformulado para o período 1986-1990. Foram introduzidas múltiplas unidades de amostragem (empresas e unidades locais) e amostragem de conglomerados (selecionando-se unidades locais e investigando-se todos os estabelecimentos nelas contidos).

Uma outra revisão amostral com grande impacto foi o emprego de duas frações amostrais no Censo Demográfico 1991, a saber 10% em municípios com população projetada de mais de 15.000 habitantes e 20% nos demais (SILVA; BIANCHINI,1990). Essa modificação reduziu a fração global de amostragem empregada no Censo Demográfico de 25% para cerca de 12,5%, resultando em substantiva economia de recursos durante a coleta e processamento dos dados, sem que fosse observada redução apreciável da precisão das estimativas derivadas da amostra.

Nesse período, novas pesquisas empregando desenhos amostrais considerados modernos foram iniciadas, tais como a Pesquisa Anual do Comércio - PAC, em 1988, (PESQUISA..., 1991) e a Pesquisa de Orçamentos Familiares - POF realizada em 1987-1988 (PESQUISA..., 1992). A PAC inovou pelo uso de dois cadastros distintos para

[1] Para mais informações, ver textos da Série Relatórios Metodológicos do IBGE, iniciada em 1981.

seleção da amostra (Censo 1985 e a Relação Anual de Informações Sociais - RAIS) para compensar problemas que cada um desses cadastros teria isoladamente. Foi também empregado esquema amostral com estratificação automática e otimização do tamanho amostral (HIDIROGLOU, 1986) e introduzido o uso de amostra reserva para compensar as elevadas perdas de coleta esperadas devido aos problemas cadastrais (SILVA et al 1999), além de se ter adotado a empresa como unidade de seleção e investigação.

Já a POF 1987-1988 inovou por ter uma amostra de domicílios desenhada sob medida, com um desenho em dois estágios apenas, incorporando estratificação de setores censitários pela renda além da localização geográfica, e com setores selecionados com reposição e probabilidades proporcionais ao tamanho no primeiro estágio e domicílios com equiprobabilidade no segundo estágio (PESQUISA..., 1992). Pela primeira vez, a seleção dos domicílios a pesquisar em cada setor foi feita sem empregar amostragem sistemática, mas sim amostragem aleatória simples, com o sorteio efetuado por computador. O dimensionamento da amostra também incorporou previsão de perda por não-resposta, outra novidade em relação à tradição das demais pesquisas domiciliares.

Também data desse período a introdução da amostragem em pesquisas da área agropecuária, com a Pesquisa Agropecuária do Paraná 1986-1987. Essa pesquisa empregou amostragem de áreas, com sofisticado esquema de estratificação multivariada incorporando restrições de contiguidade (MULLER; SILVA; VILLALOBOS, 1988). Essa pesquisa foi mais tarde estendida para três outros estados (São Paulo, Distrito Federal e Santa Catarina) e sua extensão para quatro outros (Mato Grosso do Sul, Goiás, Rio Grande do Sul e Minas Gerais) era planejada.

O início da década de 1990 marcou um período de retrocesso, provocado por aguda crise de financiamento das operações do IBGE. Tais problemas levaram ao cancelamento de pesquisas (PNAD em 1994) ou a drásticas reduções não desejadas nas amostras de pesquisas da área econômica (PAC e PIA desde 1992) etc.

Essa fase de crise foi superada e em 1996 estavam em curso várias iniciativas que representavam progresso renovado no uso de amostragem. Foi iniciado em 1995 um programa abrangente de reformulação das estatísticas econômicas, apoiado na ampliação do uso de amostragem. Pela primeira vez, a montagem de cadastros para a seleção de amostras das pesquisas econômicas periódicas foi baseada numa pesquisa cadastral também por amostragem (ASPECTOS..., 1995). No passado esses cadastros eram obtidos dos censos econômicos quinquenais que, entretanto, não foram realizados no início da década de 1990 como previsto. Já a pesquisa Censo Cadastro 1995 teve como base a Relação Anual de Informações Sociais - RAIS, procurando aproveitar os dados cadastrais desse registro administrativo para suprir a lacuna deixada pela falta dos censos. Numa segunda fase, as diversas pesquisas econômicas repetidas teriam suas amostras revistas, passando a selecionar unidades da amostra mestra do Censo Cadastro 1995.

Tal modificação implicou em novos problemas e dificuldades, causadas pela falta das bases cadastrais (*benchmarks*) antes geradas pelos censos. Mas ao mesmo tempo ensejou oportunidades para avanços e modernização dos desenhos amostrais das várias pesquisas econômicas. Uma vantagem do uso da RAIS é que novos dados são coletados anualmente. Isso não implica, entretanto, que tenha diminuído a importância dos censos como fornecedores de *benchmarks* para as pesquisas periódicas.

Outra iniciativa recente foi a introdução da Pesquisa Mensal do Comércio - PMC em 1995, a princípio apenas no Rio de Janeiro. Essa pesquisa inovou no emprego de métodos de estratificação ótima (FARIAS; BARBOSA, 1993-1996), por ser a primeira incursão do IBGE no setor de comércio com uma pesquisa de periodicidade mensal e pela parceria com setores da sociedade civil no planejamento e financiamento da pesquisa.

Foi ainda realizada uma pesquisa piloto sobre economia informal, Economia Informal Urbana - Ecinf 1994, limitada ao município do Rio de Janeiro, com desenho amostral similar ao da POF 1986-1988. Uma inovação nessa pesquisa foi que a operação de listagem dos setores selecionados compreendeu uma primeira etapa da pesquisa, chamada varredura ou *screening*, na qual foram identificadas as unidades pertencentes à população alvo das quais

foi selecionada a amostra efetivamente pesquisada. Essa pesquisa foi realizada em escala completa em 1997, cobrindo as áreas urbanas dos 200 maiores municípios brasileiros.

Também estava em curso desde outubro de 1995 a coleta de dados da POF 1995-1996, cujo desenho amostral é bastante similar ao adotado na POF 1986-1988, com pequenos ajustes decorrentes do uso de nova base cadastral obtida do Censo Demográfico 1991 (BIANCHINI; VIEIRA, 1996).

A Pesquisa Domiciliar sobre Padrões de Vida - PPV, cuja pesquisa piloto foi iniciada em março de 1996, também faz parte dessa série de iniciativas recentes em que a amostragem tem papel de destaque. Seu desenho amostral se assemelha àquele adotado na POF 1986-1988, embora com abrangência geográfica distinta (regiões Nordeste e Sudeste do país) e menor número de unidades selecionadas em cada estágio de amostragem (ALBIERI; BIANCHINI; CARDOSO, 1995).

Apesar da maturidade do IBGE em relação ao emprego de amostragem em suas pesquisas e à questão do planejamento amostral, ainda há vários desafios a serem superados. Primeiramente há que terminar a conquista da área de estatísticas agropecuárias, onde ainda se faz pouco uso de amostragem, em comparação com outras áreas. Um forte candidato seria o Censo Agropecuário, onde o emprego de amostragem durante a coleta poderia garantir economias significativas de tempo e recursos, nos moldes do que ocorre no Censo Demográfico.

Há também muito que fazer na direção de incorporar melhor a estrutura longitudinal de pesquisas repetidas como a PME, a PIM-DG e a PIM-PF ao rever os respectivos desenhos amostrais. A introdução de procedimentos de rotação nas amostras das pesquisas da área econômica, por exemplo, poderia proporcionar redução do ônus dos respondentes de menor porte, ao mesmo tempo que daria oportunidade para atualizações mais frequentes das mesmas. Já no caso da PME, há razões para crer que reduções ainda maiores no tamanho da amostra seriam possíveis sem grande perda de precisão, bastando para isso que a estrutura longitudinal da pesquisa seja mais plenamente aproveitada na hora da estimação, como se vai discutir no próximo tópico.

Um último desafio seria a montagem de sistemas mais estáveis de geração e atualização das bases cadastrais das pesquisas, menos dependentes de censos, como se esboça na área econômica com a introdução da pesquisa Censo Cadastro 1995. Tal objetivo ainda está fora de alcance para toda a área de pesquisas domiciliares pela falta de uma base cadastral digitalizada (incluindo mapas) e/ou de listas de endereços domiciliares em meio digital, tais como as existentes em outros países.

Estimação em pesquisas por amostragem

Este tópico discute a questão da estimação em pesquisas por amostragem. A revisão do uso de amostragem no IBGE apresentada no tópico **Planejamento de pesquisas por amostragem** serve como pano de fundo para ajudar a compreender o desenvolvimento também observado no uso de métodos de estimação.

Na primeira fase do uso de amostragem nas pesquisas do IBGE os estimadores empregados para totais e médias baseavam-se na ideia de ponderar as observações amostrais pelo inverso das respectivas probabilidades de seleção e tinham em comum o objetivo de simplicidade de cálculo. Os estimadores para totais eram da forma

$$\hat{Y} = \sum_{i \in s} w_i y_i \qquad (1)$$

onde $\hat{Y}$ é o estimador de Horvitz-Thompson do total populacional de uma variável *y* qualquer, y_i é o valor dessa variável para a i-ésima unidade da população, w_i é o peso dado à i-ésima unidade, calculado como $w_i = \pi_i^{-1}$, sendo π_i a probabilidade de inclusão na amostra da i-ésima unidade, e a soma é sobre o conjunto *s* de todas as unidades selecionadas para a amostra.

Devido ao emprego de desenhos amostrais autoponderados[2] (no Censo Demográfico, PNAD e PME, por exemplo), a simplicidade dos estimadores era assegurada pelo desenho amostral. Com as probabilidades de seleção todas iguais à fração amostral $f=n/N$, onde n é o tamanho da amostra e N o tamanho da população, os pesos w_i são todos iguais ao inverso dessa fração ($w_i = N/n \ \forall i$) e a expressão do estimador de total em (1) se reduz a $\hat{Y} = (N/n)\sum_{i \in s} y_i = N\,\bar{y}$, onde $\bar{y}$ é a média amostral da variável y. Nesse período a preocupação com a estimação e divulgação da precisão das estimativas era secundária.

Tal esquema se justificava pelo imperativo prático de efetuar os cálculos das estimativas sem suporte computacional sofisticado. A ocorrência de problemas práticos tais como não-resposta (possivelmente diferencial em relação a subgrupos da amostra) e desatualização da base cadastral à medida que o período de referência da pesquisa se afastava do último censo, dificultava o uso puro e simples de estimadores autoponderados. Na PNAD, por exemplo, o tratamento dispensado às novas construções foi motivado em parte pela aderência ao uso de estimadores autoponderados (MENEZES; ALMEIDA; BIANCHINI, 1991). Na PME, um dos problemas que levou à necessidade de redução da amostra em 1988 foi provocado pela tentativa de manter a autoponderação do desenho mediante uso de fração fixa de amostragem no segundo estágio (SILVA; MOURA, 1988).

A despeito desses esforços para manter os estimadores autoponderados, observou-se que tais estimadores apresentavam vício não desprezível. A saída foi adotar estimadores que calibrassem a pesquisa, de forma que as estimativas de totais da população em certas categorias coincidissem com totais obtidos de projeções demográficas independentes. Isso era feito mediante emprego de estimadores de razão (quando apenas o total global da população é usado) ou pós-estratificação (quando a calibração se dá para duas ou mais categorias da população), em comum com a prática adotada em vários outros países. Nesse caso, os pesos amostrais são dados pelo inverso da fração amostral observada em cada categoria, isto é, $w_i = N_h/n_h$ se a unidade i pertence ao pós-estrato h.

Estimadores de pós-estratificação são uma forma simples de incorporar informações auxiliares para melhorar as estimativas amostrais. Essa melhoria se dá pela redução de vício provocado por não-resposta ou desatualização cadastral, ou pela redução da variância das estimativas devido ao efeito da estratificação. Quando o número de categorias nas quais a amostra é pós-estratificada é pequeno, o emprego desses estimadores é relativamente simples. Foi desse tipo o esquema de estimação adotado nas amostras dos Censos Demográficos 1960 e 1970. Na PNAD e na PME eram empregados estimadores de razão simples, calibrando-se apenas para o total da população residente projetada. Estimadores de razão com totais de pessoal ocupado e valor da produção foram também adotados na expansão da PIA 1981 (PLANO..., 1981).

Uma vantagem óbvia da calibração é a concordância das estimativas da pesquisa com os totais de população projetados, sem prejuízo da simplicidade do processo de estimação. Entretanto, ela ajuda a mascarar problemas tais como a inexistência de tratamento adequado para a não-resposta (na PNAD e na PME), e as dificuldades causadas pela tentativa de aderência à autoponderação. Outra dificuldade é a qualidade das projeções de população usadas para calibrar a amostra (BIANCHINI, 1989). O Censo Demográfico 1991 mostrou que as projeções populacionais para o final da década de 1980 e início da década de 1990 estavam superestimadas, e esse efeito foi repassado para as estimativas de totais originadas da PNAD e PME.

A incorporação de informações populacionais auxiliares na etapa de estimação de pesquisas por amostragem é a ligação comum a todas essas aplicações. Nesse assunto o IBGE já adquiriu larga experiência e tem feito, em algumas áreas, uso efetivo dos desenvolvimentos recentes da teoria. Para corroborar essa afirmação é preciso examinar a evolução dessa teoria e de como ela tem sido aplicada no IBGE.

---

[2] Assim chamados porque todas as unidades populacionais têm idêntica probabilidade de inclusão na amostra.

O aproveitamento de informações populacionais auxiliares para estimação em pesquisas por amostragem é uma das partes da teoria de amostragem que mais progrediu desde os anos 1970. O livro de Cochran (1977) representava o estado da arte da amostragem até então e contemplava o uso de informações auxiliares através de estimadores de razão ou de regressão simples (ambos incorporando apenas uma variável auxiliar) ou de pós-estratificação. Entretanto essas técnicas eram apresentadas como ferramentas separadas, sem uma ligação comum.

O IBGE adotou na expansão da amostra do Censo Demográfico 1980 o Processo Iterativo de Estimação por Totais Marginais - PIETOM (METODOLOGIA..., 1983b) aplicado separadamente para cada uma das 4219 áreas de ponderação[3]. Esse método consistia em definir uma tabela (ou matriz) de pós-estratificação de dupla entrada, cujas linhas e colunas eram dadas por combinações de valores das variáveis auxiliares, as quais foram investigadas a 100% através do questionário básico. Eram, portanto, conhecidos os totais populacionais das celas, linhas e colunas dessa tabela. Os pesos amostrais para unidades em cada cela eram calculados por um processo iterativo de ajuste dos pesos simples iniciais, de forma tal que as estimativas amostrais eram sucessivamente calibradas nos totais das linhas e depois das colunas, até que fosse observada convergência dos pesos.

O uso desse método permitiu ampliar bastante o número de variáveis auxiliares consideradas para a calibração das estimativas amostrais: a tabela de pós-estratificação empregada em 1980 tinha 720 celas, em comparação com os 46 pós-estratos adotados no Censo Demográfico 1970. No entanto, tal método envolvia um algoritmo iterativo que podia não convergir, requerendo ajustes na pós-estratificação quando isso ocorria. Além disso, o método foi implementado em programas desenvolvidos sob medida, requeria cálculos complicados para estimação da precisão e não oferecia uma teoria para explicar a definição da pós-estratificação. Essas dificuldades levaram o IBGE a procurar um novo método para expandir a amostra do Censo Demográfico 1991, apesar do sucesso da aplicação do PIETOM em 1980.

Nesse período houve grande progresso na teoria. No início dos anos 1990, o livro de Särndal, Swensson e Wretman (1992) correspondia ao estado da arte da amostragem no início dos anos 1990, apresentando as técnicas de pós-estratificação, estimação de razão e de regressão como casos particulares do estimador de regressão generalizado, o qual fornece uma estrutura flexível e eficiente para incorporar informações auxiliares na etapa de estimação. Esse estimador pode ser escrito na forma (1), mas nesse caso os pesos são dados por

$$w_i = \pi_i^{-1} g_i \tag{2}$$

onde g_i é um fator de ajuste ou calibração que incorpora informações sobre um vetor $x_i = (x_{i1},\ldots,x_{iQ})'$ de variáveis auxiliares.

O estimador de regressão generalizado é motivado por um modelo linear relacionando a variável de interesse y com o vetor de variáveis auxiliares x tal como

$$y_i = \beta_0 + x_i'\beta + \varepsilon_i \tag{3}$$

onde β_0 e β são parâmetros desconhecidos, e se assume que os erros ε_i tem média 0, são não correlacionados e têm variância conhecida exceto pelo fator de escala σ_e.

No caso mais simples em que a amostra é aleatória simples sem reposição e os erros do modelo (3) são supostos homoscedásticos, o fator de ajuste g_i é dado por:

$$g_i = 1 + (\bar{X}-\bar{x})'\hat{S}_x^{-1}(x_i-\bar{x}) \tag{4}$$

onde $\bar{X}$ é o vetor de médias conhecido das variáveis auxiliares x na população, e $\bar{x}$ e $\hat{S}_x$ são respectivamente o vetor de médias e a matriz de covariâncias das variáveis auxiliares x na amostra.

[3] Área de ponderação é a menor área para a qual foram calculadas estimativas com base na amostra do Censo Demográfico, e na maior parte das vezes, coincidia com um município, podendo ser uma subdivisão de um município de maior população.

Com os fatores de ajuste dados por (4) é fácil mostrar que o estimador de regressão do total *Y* pode ser escrito como

$$\hat{Y}_R = N\left[\bar{y} + \left(\overline{\mathrm{X}} - \bar{\mathrm{x}}\right)'\mathrm{b}\right] \tag{5}$$

onde $\mathbf{b} = \hat{\mathbf{S}}_x^{-1}\hat{\mathbf{S}}_{xy}$ é estimador de mínimos quadrados ordinários do parâmetro β e $\hat{\mathbf{S}}_{xy}$ é a covariância amostral entre **x** e *y*.

O estimador de regressão é útil por, pelo menos, três motivos. Primeiro porque oferece calibração nas variáveis auxiliares, isto é, se aplicado a qualquer das variáveis do vetor **x**, replicará exatamente seu total conhecido na população. Segundo porque oferece ganhos de eficiência em relação ao estimador de Horvitz-Thompson sempre que a variância dos erros ε_i for menor que a dos y_i (SÄRNDAL; SWENSSON; WRETMAN, 1992). Terceiro porque tem grande flexibilidade, já que o vetor **x** de variáveis auxiliares pode incluir qualquer número de variáveis de tipo contínuo ou binário, ou ambos, e também devido ser facilmente generalizado para o caso de desenhos amostrais complexos.

Estimadores de regressão podem também ser justificados sob várias abordagens alternativas da teoria de amostragem. Särndal, Swensson e Wretman (1992) enfatizam uma abordagem *model assisted*, em que o modelo de regressão (3) é usado para motivar o estimador, mas em que as propriedades do mesmo são avaliadas com respeito à distribuição gerada por repetidas replicações do processo de seleção da amostra.

Já Royall (1970) e seus seguidores apresentam esse estimador como o estimador ótimo sob uma família de modelos lineares, ignorando o desenho amostral na inferência e avaliando as propriedades do estimador com respeito a diferentes realizações do modelo postulado.

Mais recentemente, Deville e Särndal (1992) identificaram o estimador de regressão como um dos membros de uma família de estimadores de calibração da forma (1), em que os pesos w_i são da forma (2) com fatores de ajuste g_i obtidos de forma a minimizar a distância:

$$\mathrm{D} = \mathrm{E_P}\left[\sum_{i\in s}\pi_i^{-1}(g_i - 1)^2 \ / \mathrm{v_i}\right] \tag{6}$$

sujeito à restrição de que $\sum_{i\in s}\pi_i^{-1}g_i\mathrm{x}_i = \sum_{i\in U}\mathrm{x}_i$, onde E_p representa uma esperança com relação à distribuição gerada pelo desenho amostral, e v_i são números positivos conhecidos, por exemplo, proporcionais às variâncias dos erros ε_i no modelo linear (3). Empregando-se outras funções de distância se gera uma ampla família de estimadores que inclui os *raking ratio estimators* do método Pietom, estimadores de regressão, de razão, de pós-estratificação e outros.

Estimadores de Mínimos Quadrados Generalizados em Dois Estágios - MQG2 dessa família foram aplicados para a expansão da amostra do Censo Demográfico 1991 (SILVA; BIANCHINI; ALBIERI, 1993; ALBIERI; DIAS, 1994). Eles substituíram o Pietom com algumas vantagens. Esses estimadores não dependem de processo iterativo de cálculo, são suportados por uma teoria sólida, simplificaram a estimação da precisão e forneceram recursos para automação do processo de seleção das variáveis auxiliares nas quais calibração pode ser imposta. Foi possível contar com programas mais ou menos gerais para implementação do método, preparados e cedidos ao IBGE pelo Statistics Canada, cuja adaptação para uso no censo brasileiro foi feita com custo muito menor que o do desenvolvimento sob medida anterior.

Um dos usos habituais da amostra do Censo Demográfico é para estimação de medidas de concentração da renda. Essas medidas podem geralmente ser derivadas da função de distribuição da renda na população. Estimadores de pós-estratificação

similares aos empregados na expansão da amostra do Censo Demográfico 1991 foram considerados em Silva e Skinner (1995) para estimar a função de distribuição populacional. Tais estimadores oferecem ganhos de precisão razoáveis em comparação com estimadores mais simples (Horvitz-Thompson, razão e diferença), sem necessidade de recorrer a estimadores não-lineares mais complexos disponíveis na literatura.

A ênfase dada aqui à discussão do método de estimação adotado no Censo Demográfico 1991 se justifica porque dela se derivam vários outros aspectos de interesse. Primeiramente vale notar que tem aumentado no IBGE o emprego de sistemas genéricos em substituição a sistemas desenvolvidos sob medida, em resposta às necessidades de redução de custos e aumento da confiabilidade e eficiência das etapas de processamento dos dados (SILVA; BIANCHINI, 1997). Na área de estimação em amostragem, em particular, há hoje em dia várias opções de sistemas genéricos: SUDAAN (SHAH et al., 1992), Generalized Estimation System - GES (ESTEVAO; HIDIROGLOU; SÄRNDAL, 1995), Bascula (GÖTTGENS et al., 1991), PC-CARP (FULLER et al., 1986), CALMAR (SAUTORY, 1993), CLAN (ANDERSSON; NORDBERG, 1994), WESVARPC (A USER'S..., 1995). Todos esses sistemas são capazes de calcular estimativas de totais e médias, e respectivas medidas de precisão (exceto CALMAR), para uma ampla gama de desenhos amostrais e tipos de estimadores.

Em particular, o sistema *GES* desenvolvido no Statistics Canada implementa a metodologia de estimadores de regressão generalizados tal como descrita no livro de Särndal, Swensson e Wretman (1992). Embora a disponibilidade de tais sistemas não fosse a mesma no passado, o que ajuda a entender a prática do desenvolvimento local sob medida, eles não podem mais ser ignorados. Por justiça, vale mencionar que a estimação do ENDEF foi feita empregando um sistema genérico denominado ARIEL, à época considerado o estado da arte. No futuro, o IBGE deverá basear a estimação de suas pesquisas no emprego de sistemas genéricos desse tipo.

O método adotado no Censo Demográfico 1991 incorpora tratamento de não-resposta por reponderação automaticamente. Apesar da existência de outras opções para o tratamento de não-resposta, é consensual no mundo das estatísticas oficiais que não-resposta em nível de unidades perdidas por completo (*unit nonresponse*) em pesquisas domiciliares deve ser tratada mediante reponderação. Portanto é um bônus contar com um método de estimação que tenha essa facilidade embutida.

Para não-resposta de itens, o tratamento mais comumente aplicado é a imputação, cuja discussão está fora do escopo desse artigo. As experiências e progressos do IBGE nesse assunto foram revistas em Silva e Bianchini (1997).

Nas pesquisas domiciliares tradicionais do IBGE (PNAD e PME) não se emprega um tratamento adequado para não-resposta de unidades perdidas por completo. Já na POF 1986-1988 foram empregadas correções baseadas no inverso da taxa de não-resposta de cada setor da amostra, ao mesmo tempo que estimadores de razão baseados na projeção da população residente (PESQUISA..., 1992). Nas pesquisas da área econômica, a prática adotada tem sido a de reponderar as unidades respondentes de cada estrato para compensar a perda por não-resposta, como é o caso da PIA (PESQUISA...,1983) e da PAC (PESQUISA..., 1991).

Apesar de se tratar de assunto no qual o IBGE já acumula alguma experiência, há grande desigualdade no reconhecimento do problema pelas suas diversas áreas. Há que investir mais no emprego de técnicas adequadas para monitoramento e tratamento da não-resposta, bem como na prática de reportar junto dos resultados os níveis de não-resposta observados nas várias pesquisas, para permitir avaliação da qualidade dos dados pelos usuários. Nessa questão a POF 1986-1988 também deu exemplo, reportando dados detalhados de não-resposta total e por tipo. Mas a prática ainda não é adotada amplamente.

A metodologia MQG2 adotada no Censo Demográfico 1991 permite incorporar grande número de variáveis auxiliares, mas não oferece uma teoria para a escolha ótima das mesmas. Esse é um dos aspectos do emprego de estimadores de regressão que tem merecido atenção da comunidade de pesquisa recentemente. Em particular, Silva e Skinner (1997) apresentam um método para seleção de variáveis auxiliares quando se

utiliza estimadores de regressão cuja eficiência para estimar a média de uma variável resposta especificada foi maior que a de vários competidores. Silva e Skinner (1997) apontam ainda para perda de precisão do estimador de regressão quando o número de variáveis auxiliares cresce demasiadamente, alertando para a necessidade de estabelecer um compromisso entre a calibração no maior número possível de variáveis auxiliares sem impor grande perda de eficiência no estimador.

O método MQG2 fornece estimativas de precisão de maneira mais simples que o Pietom. Essa foi uma das áreas em que o IBGE progrediu bastante. Desde meados dos anos 1980 passou a ser rotina o cálculo e disseminação de medidas da precisão das estimativas obtidas de pesquisas por amostragem, tanto as domiciliares como as da área econômica. A forma adotada para disseminar as estimativas de precisão tem variado no tempo e ainda representa um problema interessante.

Algumas vezes, como na PIA 1981, optou-se por duplicar o plano tabular publicando para cada tabela com estimativas de totais uma outra com estimativas dos respectivos coeficientes de variação (CVs). Essa opção dá estimativas de precisão em grande detalhe, mas sua implementação é cara. Mais recentemente o padrão tem sido publicar medidas agregadas da qualidade das estimativas, tais como tabelas resumindo as distribuições de frequência dos CVs das estimativas de total, como é o caso da PAC. Essa forma tem baixo custo, mas não permite conhecer o valor do CV de uma dada estimativa de total.

Um meio termo foi encontrado para disseminar a precisão das estimativas da amostra do Censo Demográfico 1980, no qual foram empregadas funções de variância generalizadas (*generalized variance functions*), Wolter (1985), para permitir que usuários dos dados calculassem eles mesmos a precisão aproximada de qualquer estimativa de total desejada. Essas funções foram obtidas mediante um ajuste de modelos de regressão relacionando a precisão (CV) com o valor da estimativa (METODOLOGIA..., 1983b). Essa abordagem foi adotada também na PNAD desde 1983. Já a PME não divulga estimativas de precisão, embora sejam calculados para uso interno os CVs estimados de algumas das taxas e totais divulgados com base nessa pesquisa. Uma solução definitiva para o problema ainda não está disponível, e esse é um tema que deve merecer atenção nos esforços de pesquisa metodológica do IBGE.

Um outro tema conectado com o método de estimação adotado no Censo Demográfico 1991 é a questão da estimação para pequenos domínios. Até o Censo Demográfico 1980, qualquer estimativa para domínios (áreas) menores que uma área de ponderação dependia de um processo de estimação caro e demorado, o que frustrava muitos usuários ou mesmo desencorajava um uso mais pleno dos dados do censo. O mesmo não pode ser dito do método adotado em 1991. Como a calibração é feita em dois níveis, um deles considerando totais populacionais por setor, é possível gerar estimativas para agregados de setores dentro de uma mesma área de ponderação utilizando os pesos calculados pela metodologia MQG2 adotando-se técnicas usuais de estimação em subpopulações. Se esses domínios tiverem tamanho mínimo da ordem de 5.000 habitantes, é provável que as respectivas estimativas tenham precisão aceitável.

Portanto não deverá mais ser necessário recorrer a métodos *ad hoc* para gerar estimativas para agregados de setores a partir da amostra do Censo Demográfico 1991. Isso não elimina a possibilidade de que essas estimativas possam ser melhoradas usando métodos mais sofisticados, capazes de incorporar outras informações auxiliares. A questão de como estimar para pequenos domínios em pesquisas por amostragem foi outra em que a teoria da amostragem avançou bastante desde os anos 1970. Na preparação de tabulações avançadas do Censo Demográfico 1980 foram empregados estimadores sintéticos (GARCIA, 1986) para gerar as estimativas em nível de microrregião (METODOLOGIA..., 1983b).

Mais recentemente, Moura (1994) estudou o emprego de modelos hierárquicos para obtenção de estimadores para pequenos domínios. Esses estimadores se revelaram superiores quando comparados com estimadores convencionais encontrados na literatura mediante estudo de simulação usando dados do Censo Demográfico Experimental de Limeira 1988, indicando que há ganhos quando se incorpora a

estrutura hierárquica dos dados na modelagem e estimação. Apesar desses esforços, ainda há pouca aplicação prática no IBGE das modernas técnicas disponíveis para obtenção de estimativas para pequenos domínios. Essa é uma área em que investimentos relativamente modestos permitiriam aumentar substancialmente o valor dos dados das pesquisas através da divulgação de estimativas para áreas geográficas mais detalhadas, por exemplo. Associada com a digitalização da base geográfica e o emprego de Sistemas de Informações Geográficas - SIG, tal iniciativa permitiria considerar a dimensão geográfica na análise dos resultados das pesquisas de forma bem mais detalhada, o que enriqueceria bastante as possibilidades analíticas.

Métodos baseados no estimador de regressão podem também ser usados para combinar resultados de pesquisas distintas, a fim de aumentar a precisão das estimativas. Por exemplo, as pesquisas mensais poderiam ter suas estimativas calibradas anualmente à luz dos resultados das pesquisas anuais, cujas amostras (em geral bem maiores) fornecem estimativas mais precisas. Ou também pode ser empregada amostragem em duas etapas, como deverá ocorrer nas novas pesquisas da área econômica, que deverão ser subamostras da pesquisa cadastro (ALBIERI; BIANCHINI; VASCONCELLOS, 1995).

Um último aspecto da metodologia de estimação do Censo Demográfico 1991 que se quer comentar aqui diz respeito aos pesos que ela gera para produção das estimativas. Essa metodologia, como também a do Censo Demográfico 1980 e de outras pesquisas como a POF 1986-1988, implica em que os pesos atribuídos aos registros individuais são complexos, e não derivados da mera inversão das respectivas probabilidades de seleção. Isso tem implicações para os usuários de arquivos de microdados divulgados pelo IBGE, como se verá com mais detalhes na próxima seção.

A utilização de métodos mais elaborados de estimação nas pesquisas da área econômica é ainda incipiente. Nessa área há muito para fazer em termos de aproveitar melhor as informações auxiliares disponíveis e também em relação ao emprego de métodos capazes de compensar as deficiências da base cadastral existente. Isso sem falar que várias questões tratadas no âmbito das pesquisas domiciliares também se aplicam para essas pesquisas.

Encerrando essa seção, discute-se a questão da estimação em pesquisas repetidas, na qual pouco foi feito no IBGE. A grande maioria das pesquisas repetidas que o IBGE realiza ignora a estrutura longitudinal dos dados na hora da estimação, tratando cada rodada da pesquisa como se fosse uma pesquisa transversal (*cross-section*) separada. Isso é muito ineficiente se comparado com estimadores que combinam informações amostrais da rodada atual com as séries históricas armazenadas, como mostrado em Silva (1992) com uma aplicação a dados da PIM-DG. Tem havido grande progresso na teoria, com inúmeros artigos publicados recentemente, conforme a revisão em Silva (1997).

Trata-se de técnicas com grande potencial em termos de oferecer estimadores de melhor precisão sem necessidade de ampliar tamanhos de amostra. Ao contrário, provavelmente seu emprego propiciaria margem para redução dos tamanhos de amostra atuais. Entre as candidatas mais óbvias para emprego dessas técnicas estão a PIM-DG, a PMC e a PME. Esta última poderia se beneficiar também de um redesenho que aproveitasse ideias como as empregadas na Pesquisa de Emprego e Desemprego - PED da Fundação Sistema Estadual de Análise de Dados - Seade em convênio com o Departamento Intersindical de Estatística e Estudos Socioeconômicos - Dieese, em particular a adoção de uma amostra trimestral com utilização de trimestres móveis para divulgação mensal de estimativas das taxas de interesse.

Análise de dados de pesquisas por amostragem

A análise (estatística) de dados e resultados de pesquisas por amostragem é uma das áreas em que o IBGE tem muito que fazer e avançar. Entre o início dos anos 1970 e meados da década de 1980 grande importância era dada à questão da análise dos resultados das pesquisas no IBGE, embora essa análise nem sempre fosse apoiada na metodologia estatística. Esse período foi seguido por outro em que as prioridades foram alteradas e o IBGE passou a se concentrar mais na produção das informações

básicas. Portanto não há hoje em dia uma tradição estabelecida de analisar em detalhe os resultados das pesquisas.

A prática de analisar os resultados das pesquisas enriquece as mesmas, aumenta o valor dos dados que elas produzem, fomenta o desenvolvimento técnico e contribui para o planejamento de pesquisas semelhantes no futuro. Em certa medida, também facilita um diálogo maior entre os usuários e o IBGE, pois este último passa a ter interlocutores mais qualificados e um entendimento melhor das necessidades dos primeiros. Por essas razões, argumenta-se que o IBGE deve investir no desenvolvimento da sua capacidade de analisar dados e resultados das pesquisas, processo no qual uma parceria com a comunidade acadêmica e de usuários é indispensável.

Um outro motivo que justifica a realização de análises é o fato de que o IBGE tem acesso aos dados individuais detalhados de cada pesquisa, enquanto os usuários frequentemente têm acesso apenas a dados agregados, por razões ligadas à preservação do sigilo das informações. Como se verá mais adiante, essa é uma limitação séria para o emprego de métodos modernos de análise estatística. Isso sem mencionar que nem todos os usuários estão capacitados a processar grandes volumes de dados como os habitualmente produzidos pelas pesquisas do IBGE.

Tendo como objetivo ampliar a capacidade e o papel analítico do IBGE, é importante discutir como isso pode ser feito e que recursos a Estatística tem hoje para oferecer em suporte a essa tarefa. Para dar uma ideia resumida da evolução nessa questão, é interessante comparar dois cenários.

Por volta de 1970, os resultados das pesquisas por amostragem (ou censos) eram essencialmente tabelas com totais das variáveis pesquisadas por categorias de interesse. Essas tabelas eram condensadas em pesados livros ou volumes de divulgação, provendo a única fonte de acesso aos dados por usuários externos. Um analista típico teria que compilar informações provenientes de vários desses volumes de tabelas e efetuar manualmente cálculos (necessariamente simples) para atingir conclusões. Como não se publicavam estimativas da precisão, as análises ficavam empobrecidas, pois não se tinha ideia da significância de efeitos ou diferenças encontradas nos dados. Comentários publicados junto das tabelas eram muitas vezes mera descrição das mesmas em palavras, pouco acrescentando em termos de uma compreensão dos resultados. Essa situação caracterizava essencialmente uma análise descritiva (manual) dos dados.

Hoje em dia ainda se publicam tabelas com totais das variáveis pesquisadas, mas também se coloca à disposição dos usuários arquivos com microdados (desidentificados) e mais recentemente até acesso eletrônico direto às principais coleções, por exemplo via Internet. Um analista típico dispõe de um computador com acesso aos dados em formato eletrônico (local ou remoto) e de sistemas com sofisticados recursos para manejo dos dados, produção de gráficos e análise estatística. Com esses sistemas é possível analisar diretamente os dados individuais e formular, ajustar, testar e validar modelos levando em conta a precisão das estimativas e o fato de que os dados foram obtidos por amostragem. Análises desse tipo contribuem de maneira efetiva para a evolução do conhecimento, merecendo a descrição de pesquisa no sentido acadêmico da palavra.

Que fatores diferenciam o segundo cenário do primeiro? Em primeiro lugar, a disponibilidade de computadores e meios de comunicação cada vez mais poderosos e de baixo custo, garantindo a possibilidade de acesso direto aos dados pelos analistas interessados. Em segundo lugar, a ampliação do acesso aos microdados, mediante preparação de produtos de pronta entrega e de facilidades para acesso eletrônico remoto às bases de dados. Ainda há muito que fazer nessa área, mas há que registrar o progresso alcançado. Hoje o IBGE divulga rotineiramente arquivos de microdados de suas principais pesquisas domiciliares.

Além desses fatores, há que reconhecer a evolução da teoria Estatística, em grande parte sustentada pela evolução do poder computacional. No primeiro cenário, um analista interessado em investigar o diferencial de salários recebidos segundo o sexo do indivíduo ficaria limitado a comparar médias de salários dos respectivos grupos classificados por mais uma ou duas variáveis de controle, tais como setor de atividade

ou idade. Tais comparações eram meramente descritivas, pois sem medidas da precisão das estimativas era impossível estabelecer a significância das diferenças encontradas.

Hoje em dia, dispondo dos dados individuais de uma pesquisa como a PNAD, por exemplo, um analista interessado na mesma questão pode formular modelos de regressão múltipla para explicar a variação dos salários em função de outras variáveis, e então testar a significância dos efeitos principais e interações. E é aqui que a evolução da teoria estatística se revela crucial. Se em meados dos anos 1970 era comum o uso de modelos de regressão e outros modelos estatísticos em várias áreas de ciências, só mais recentemente essas técnicas foram adaptadas para aplicação na análise de dados provenientes de pesquisas por amostragem.

Na modelagem estatística convencional se assume que as observações de uma amostra[4] são independentes e identicamente distribuídas - IID. Tal hipótese é inadequada para o caso das pesquisas por amostragem realizadas para produção de estatísticas oficiais e muitas outras. Por razões práticas, tais como inexistência de cadastros (de pessoas, por exemplo) ou de eficiência, essas pesquisas geralmente empregam desenhos amostrais complexos envolvendo estratificação, conglomeração e probabilidades desiguais de seleção. Embora adequados para gerar amostras cujas estimativas de totais são representativas da população de interesse, esses desenhos produzem amostras que não podem ser adequadamente modeladas como IID.

Só recentemente se teve acesso amplo a uma abordagem teórica coerente e amadurecida para lidar com essas situações: o livro de Skinner, Holt e Smith (1989) discute métodos para a análise de dados de pesquisas amostrais complexas. Uma das ideias ali enfatizadas é a de que é essencial considerar o desenho amostral ao fazer inferência estatística empregando dados provenientes de amostras complexas.

Para ilustrar esse ponto, reconsidere o analista interessado em testar se há diferença significativa entre as médias dos salários recebidos por pessoas do sexo masculino e feminino, usando dados de uma amostra complexa. Se esse analista tentasse aplicar um teste *t* de Student usando um pacote estatístico qualquer, baseando sua análise na amostra de dados individuais, sua inferência estaria incorreta. Sem exceção, os procedimentos automatizados nos pacotes estatísticos de uso geral assumem que a amostra é IID. Isso geralmente implica em estimativas de precisão (essenciais para se fazer testes de hipótese e inferência estatística em geral) que são viciadas, subestimando grosseiramente o erro padrão efetivamente alcançado com a amostra.

Skinner, Holt e Smith (1989) propõem o uso de fatores de correção, denominados de *misspecification effects* - MEFF, para corrigir as estimativas de precisão habituais em razão da especificação incorreta da distribuição da amostra. O IBGE publicou estimativas de fatores desse tipo para algumas variáveis com base na amostra do Censo Demográfico 1980 (SILVA; MOURA, 1990). Essa abordagem daria conta de resolver o problema para questões inferenciais relativamente simples, tais como comparações de médias por subgrupos (Análise de variância - Anova) e outras similares. Mas para a efetiva modelagem dos dados de pesquisas por amostragem essa abordagem não é suficiente, e várias seções do livro são dedicadas a estudar alternativas.

Há dois caminhos principais: análises agregadas ou análises desagregadas. Na análise desagregada, os modelos postulados envolvem a especificação de distribuições para as variáveis de interesse na população alvo e para a relação dessas com variáveis auxiliares consideradas no desenho amostral. Nesse caso, as inferências podem ser feitas ignorando o desenho amostral, de vez que as variáveis auxiliares do desenho já estão diretamente incorporadas no modelo. A análise agregada requer apenas a especificação de modelos para descrever as distribuições das variáveis de interesse na população alvo, mas nesse caso as inferências para os parâmetros desses modelos devem ser feitas levando-se em conta o desenho amostral empregado.

Para ilustrar e analisar melhor as abordagens alternativas, considere o problema de ajustar um modelo de regressão linear usando dados provenientes de uma pesquisa por amostragem. Sejam *y* a variável resposta e *z* o vetor das variáveis explicativas de

[4] Aqui o termo é empregado no sentido amplo de um conjunto de observações de um certo fenômeno cujas propriedades se quer investigar.

interesse no exercício de modelagem, e seja *x* o vetor de variáveis auxiliares consideradas no desenho amostral (por exemplo, variáveis indicadoras para os estratos de seleção ou para indicar a que conglomerado cada unidade pertence). O modelo de regressão linear relacionando *y* com *z* pode ser especificado alternativamente por (7) ou (8) a seguir:

$$y_i = \beta_0 + z_i'\beta + x_i'\gamma + \varepsilon_i \quad (7)$$

$$y_i = \beta_0 + z_i'\beta + \varepsilon_i \quad (8)$$

onde β_0, β e y são parâmetros desconhecidos e os erros ε_i são não correlacionados, com média 0 e variância conhecida a menos de um fator de escala σ_e.

O modelo desagregado (7) incorpora diretamente as variáveis do desenho amostral. Nathan e Holt (1980), considerando o caso em que há apenas uma variável explicativa *z* e uma variável de desenho *x*, afirmam que ' não há nada de novo na situação: estimadores *model-based* não viciados dos parâmetros desse modelo estão disponíveis e a função do desenho amostral baseado na variável *x* é melhorar as propriedades desses estimadores'. Esse seria o domínio da estatística clássica, pois a modelagem explícita da resposta como função das variáveis auxiliares torna o desenho amostral não informativo (SMITH, 1989).

Apesar de parecer simples, essa abordagem envolve questões complexas. Raramente a estrutura populacional considerada no planejamento amostral pode ser codificada numa única variável auxiliar. É mais comum haver dezenas de variáveis para representar os diferentes estratos e conglomerados aos quais pertencem as unidades. Isso motivou em parte o desenvolvimento recente no uso de modelos hierárquicos para analisar dados de pesquisas por amostragem (GOLDSTEIN; SILVER, 1989; BRYK; RAUDENBUSH, 1992), pois esses modelos permitem considerar a dependência entre unidades da amostra, incluindo efeitos de conglomeração.

Por outro lado, a inclusão de todas as variáveis auxiliares no modelo apresentaria desafios não triviais para o analista, em particular para a estimação adequada de todos os seus parâmetros. Além disso, a interpretabilidade e a validação do modelo se tornariam mais complexas (SKINNER; HOLT; SMITH, 1989, p. 9). Outra dificuldade é que essa abordagem requer conhecimento detalhado sobre as variáveis *x* empregadas no desenho amostral, algo que frequentemente não está disponível para o analista (secundário) devido a restrições de confidencialidade e outras razões práticas.

Modelos agregados como (8), nos quais as variáveis auxiliares não figuram diretamente, podem também ser preferidos por razões de substância: as variáveis de desenho *x* são em geral escolhidas por razões práticas e não porque fazem parte da explicação científica da resposta *y* pelas variáveis *z*. Apesar de não figurarem explicitamente no modelo (8), as variáveis auxiliares não podem ser ignoradas na estimação dos parâmetros do mesmo. Isso foi demonstrado de forma convincente por Holt, Smith e Winter (1980). Supondo que os vetores $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ têm distribuição normal multivariada, esses autores mostraram que o estimador de Máxima Verossimilhança do parâmetro β no modelo (8) depende da variável *x*.

A hipótese de normalidade não é válida em geral. Mas o que se verifica sob normalidade é que não se pode ignorar o efeito de seleção (SMITH, 1993) das unidades populacionais com base nas variáveis auxiliares $\mathbf{x}$ a menos que uma condição seja satisfeita: (y_i, z_i) é independente de x_i na população. Como essa condição é raramente satisfeita e de difícil verificação, é, portanto, essencial considerar as variáveis auxiliares e/ou o desenho amostral na realização de inferências sobre β.

Holt, Smith e Winter (1980) e Nathan e Holt (1980) recomendam uma abordagem *model-based*, que ignora o desenho amostral e se baseia no emprego de estimadores de Máxima Verossimilhança - MV no modelo normal multivariado com dados incompletos. Binder (1983) propôs uma abordagem denominada de Máxima Pseudo

Verossimilhança - MPV que se aplica a modelos da família exponencial, mas que é bastante simples no caso do modelo linear (8) se os erros são normalmente distribuídos. Suponha que um censo fosse realizado. Nesse caso, o estimador de MV *B* para β baseado em toda a população finita seria obtido resolvendo-se a equação de verossimilhança

$$\sum_{i\in U} \mathbf{z}_i(y_i - \mathbf{z}_i'\mathbf{B}) = \mathbf{0} \Leftrightarrow (\mathbf{Z}_U'\mathbf{Z}_U)\mathbf{B} = \mathbf{Z}_U'\mathbf{Y}_U \Rightarrow \mathbf{B} = (\mathbf{Z}_U'\mathbf{Z}_U)^{-1}\mathbf{Z}_U'\mathbf{Y}_U \quad (9)$$

onde $\mathbf{Z}_U$ e $\mathbf{Y}_U$ são a matriz e o vetor de dados populacionais sobre as variáveis **z** e *y* respectivamente, e se assume que a matriz $(\mathbf{Z}_U'\mathbf{Z}_U)$ é inversível.

Como apenas uma amostra da população é observada, a ideia é usar os pesos amostrais w_i para estimar o total dos escores na população, e então obter o estimador de MPV para β mediante a solução das equações de pseudo verossimilhança:

$$\begin{aligned} &\sum_{i\in s} w_i\mathbf{z}_i(y_i - \mathbf{z}_i'\mathbf{b}_{PV}) = \mathbf{0} \quad \Leftrightarrow \\ &(\mathbf{Z}_s'\mathbf{W}_s\mathbf{Z}_s)\mathbf{b}_{PV} = \mathbf{Z}_s'\mathbf{W}_s\mathbf{Y}_s \quad \Rightarrow \\ &\mathbf{b}_{PV} = (\mathbf{Z}_s'\mathbf{W}_s\mathbf{Z}_s)^{-1}\mathbf{Z}_s'\mathbf{W}_s\mathbf{Y}_s \end{aligned} \quad (10)$$

onde $\mathbf{Z}_s$ e $\mathbf{Y}_s$ são a matriz e o vetor de dados amostrais sobre as variáveis *z* e *y* respectivamente, $\mathbf{W}_s$ é a matriz diagonal com os pesos amostrais w_i e a matriz $(\mathbf{Z}_s'\mathbf{W}_s\mathbf{Z}_s)$ é suposta inversível.

Quando os pesos amostrais são todos iguais, o estimador em (10) coincide com o estimador de Mínimos Quadrados Ordinários - MQO dado por

$$\mathbf{b}_{MQO} = (\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{Z}_s'\mathbf{Y}_s \quad (11)$$

Já quando os pesos são os do estimador de Horvitz-Thompson, se obtém o estimador de Mínimos Quadrados Ponderados dado por:

$$\mathrm{b}_\pi = (\mathrm{Z}_s \prod_s^{-1} \mathrm{Z}_s)^{-1}\mathrm{Z}_s \prod_s^{-1} \mathrm{Y}_s \quad (12)$$

onde $\prod_s^{-1}$ é a matriz diagonal com as probabilidades de inclusão π_i.

O estimador MQO é viciado quando (y_i, z_i) é correlacionado com x_i e a seleção da amostra depende dessa variável (NATHAN; SMITH, 1989). Já o estimador de MPV é consistente (sob o desenho) para o parâmetro B definido em (9), embora haja perda de eficiência em relação ao estimador de MQO, que é ótimo sob o modelo. Essa troca entre vício e eficiência representa o leque de escolhas do analista na análise de dados de pesquisas por amostragem. Ignorar o desenho amostral pode implicar em vício considerável no estimador e respectivas estimativas de precisão. Para reduzir o vício considera-se o desenho amostral, mas há geralmente um preço a pagar em termos de eficiência. O debate sobre o papel que esses pesos amostrais devem desempenhar na inferência sobre modelos como (8) ainda não foi definitivamente resolvido (SMITH, 1988; PFEFFERMANN, 1993). Uma coisa é certa, entretanto: ignorar pura e simplesmente o fato de que os dados provêm de uma amostra complexa não é recomendável e os riscos dessa estratégia são grandes em termos da possibilidade de inferências incorretas.

Uma vantagem do método de MPV é sua relativa simplicidade e o fato de que depende de menos hipóteses que os métodos *model-based*. Isso permitiu a sua implementação em sistemas genéricos como SUDAAN e PC-CARP, o que facilitou bastante a modelagem de dados amostrais complexos.

Recentemente, Silva (1996) estudou o efeito de empregar nas equações de pseudo-verossimilhança os pesos do estimador de regressão dados por (2), ao invés dos pesos do estimador de Horvitz-Thompson. Esse estudo tem interesse prático porque muitas pesquisas divulgam dados individuais com pesos desse tipo, adotados para obter calibração ou para compensar não-resposta. Os resultados até agora apontam que esses pesos têm propriedades muito semelhantes aos de Horvitz-Thompson, embora haja cuidados a serem tomados em termos da estimação da precisão.

Essa discussão mostra a necessidade do IBGE prover seus usuários com informações suficientes para que possam fazer uso adequado e correto dos dados. Isso implica incluir pelo menos os pesos amostrais básicos (os inversos das probabilidades de inclusão) e variáveis indicadoras de estratificação e conglomeração junto dos microdados, ainda que por razões de sigilo estas variáveis talvez tenham que ser anonimizadas.

Pela mesma razão, se nota que o IBGE ocupa uma posição privilegiada em relação à possibilidade de efetuar análises dos dados. Provavelmente é o único capaz de experimentar com o uso de modelos desagregados como (7) devido às dificuldades causadas para analistas secundários pelas restrições de proteção do sigilo das informações. Isso é especialmente verdadeiro para estimação de precisão, pois esta requer informações mais detalhadas sobre o desenho amostral que aquelas necessárias para obtenção de estimativas pontuais.

A abordagem de MPV proposta por Binder (1983) é atrativa também porque se estende facilmente para outros modelos da família exponencial, tais como regressão logística, modelos log-lineares para análise de tabelas de contingência e outros modelos lineares generalizados. Alguns desses estão já implementados em sistemas genéricos como SUDAAN.

Essa facilidade e a maior disseminação de microdados levarão a um uso mais sofisticado dos mesmos, por sua vez gerando maior demanda por novos microdados. Atender satisfatoriamente essa demanda é mais um desafio para o IBGE, que precisa ao mesmo tempo cumprir sua responsabilidade de preservar o sigilo de informações individuais identificadas.

Outra área em que o progresso da teoria vai desafiar o IBGE é a de análise das pesquisas longitudinais. Já existe alguma tradição em pesquisas como a Pesquisa Industrial Mensal - Produção Física - PIM-PF e Pesquisa Industrial Mensal - Dados Gerais - PIM-DG de analisar as séries históricas agregadas e publicar séries sazonalmente ajustadas (INDICADORES..., 1991). Mas essa é uma análise bastante crua, e nem mesmo praticada em todas as pesquisas repetidas do IBGE. A PME, por exemplo, ainda não faz ajuste sazonal e só recentemente se tem investigado a obtenção de séries sazonalmente ajustadas dos índices de preços calculados pelo IBGE (BUZANOVSKY; PINTO; CRUZ, 1995). Há muito que fazer nessa área, e é encorajador que pesquisadores do IBGE tenham recentemente submetido ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq projeto de pesquisa sobre o tema para complementar o financiamento do trabalho (FEIJÓ; SILVA; CARVALHO, 1996).

Finalmente, vale ainda comentar que o valor dos dados produzidos pelas pesquisas pode ser bastante aumentado quando essas são combinadas entre si e com dados de outras fontes, mediante estruturas analíticas integradoras, tais como a das contas nacionais ou os relatórios integrados de indicadores sociais. Esses esquemas de análise permitem criticar e validar dados de forma que as pesquisas isoladas não contemplam, mas são ainda pouco explorados, diante de seu potencial.

Conclusões e desafios

Esse capítulo procurou rever a evolução observada no emprego de métodos para seleção de amostras, estimação e análise de dados amostrais complexos no IBGE, fazendo um paralelo com o desenvolvimento da teoria de amostragem em geral. O IBGE emerge dessa revisão como um usuário maduro de métodos para seleção e estimação de amostras, embora o mesmo não se possa dizer em termos da análise de resultados.

Há grandes desafios a superar para manter a tradição nas áreas em que o IBGE tem experiência adquirida, bem como para ampliar a utilização das modernas técnicas estatísticas disponíveis em benefício da redução de custos via amostras menores, aumento da eficiência e da precisão mediante revisão dos desenhos amostrais e estimadores empregados, e aumento da velocidade de obtenção dos resultados via emprego de sistemas genéricos de estimação tais como SUDAAN e GES. Há que ampliar o uso de técnicas estatísticas de análise de dados, tais como modelos hierárquicos, modelos de séries temporais, etc. implantando uma cultura de análise dos resultados que realimente o planejamento de pesquisas futuras e contribua de maneira mais efetiva no debate informado dos grandes temas de interesse nacional.

Há também desafios em termos de ampliar ainda mais o acesso aos microdados e outros resultados das pesquisas, o principal deles sendo a divulgação mais rápida das informações. Os recentes progressos no emprego de tecnologias modernas como CD-ROM e redes como a Internet para a disseminação das informações deverão provocar aumento da demanda e do número de usuários em contato com o IBGE, que precisará estar preparado para atender essa demanda de forma mais ágil e satisfatória.

Será cada vez mais importante que o IBGE divulgue junto ao público usuário as metodologias adotadas na produção dos dados, juntamente com informações sobre a qualidade dos mesmos, incluindo estimativas da precisão amostral, taxas de não-resposta e de imputação de itens, quando for o caso. Esse processo contribuirá para educar os usuários no entendimento mais pleno do processo de produção de informações empregando pesquisas por amostragem.

Vale mencionar ainda a necessidade de completar a conquista de grupos ainda não completamente adeptos do uso de amostragem, tanto dentro do IBGE como entre seus usuários menos crédulos.

Para vencer todos esses desafios é imprescindível que o IBGE recrute, forme e mantenha em seus quadros pessoal técnico qualificado, em particular nas áreas de Estatística e de amostragem. Sem competência no manejo do "telescópio da Estatística", ficará prejudicada a visão IBGEana da realidade brasileira.

Referências

A USER'S guide to WesVarPC: version 1.0. Rockville: Westat, 1995.

ALBIERI, S.; BIANCHINI, Z. M.; CARDOSO, R. L. *Pesquisa domiciliar sobre padrões de vida:* planejamento da amostra. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1995. 25 p.

ALBIERI, S.; BIANCHINI, Z. M.; VASCONCELLOS, M. T. L de. Aspectos de amostragem relativos ao Censo Cadastro de 1995. Rio de Janeiro: IBGE, 1996. 47 p. (Texto para discussão. Diretoria de Pesquisas; n. 80) Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv25810.pdf>. Acesso em: set. 2017.

ALBIERI, S.; DIAS, A. J. R. *Metodologia de expansão da amostra do censo demográfico de 1991:* uma descrição resumida. Rio de Janeiro: IBGE, 1994. Não publicado.

ANDERSSON, C.; NORDBERG, L. A method for variance estimation of non-linear functions of totals in surveys: theory and software implementation. *Journal of Official Statistics*, Stockholm: Statistics Sweden, v. 10, n. 4, p. 395-405, 1994.

ASPECTOS de amostragem relativos à pesquisa cadastro de 1995. Rio de Janeiro: IBGE, 1995. Não publicado.

BIANCHINI, Z. M. *Projeções de população e o ajuste das estimativas das pesquisas domiciliares:* solução ou problema?. Rio de Janeiro: IBGE, 1989. 18 p.

BIANCHINI, Z. M.; VIEIRA, M. *O planejamento da amostra da pesquisa de orçamentos familiares 95/96.* Rio de Janeiro: IBGE,1996. Não publicado.

BINDER, D. A. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review,* Medford: Wiley, v. 51, n. 3, p. 279-292, Dez. 1983.

BRYK, A. S.; RAUDENBUSH, S. W. *Hierarchical linear models:* applications and data analysis methods. Newbury Park: Sage Publications, 1992. 265 p. (Advanced quantitative techniques in the social sciences, 1).

BUZANOVSKY, A. M.; PINTO, L. M. C.; CRUZ, M. M. *Ajustamento sazonal nos índices de preços ao consumidor amplo.* Rio de Janeiro: IBGE, 1995. Não publicado.

CABRAL, M. D. B. *Plano de amostragem da PIA 1984.* Rio de Janeiro: IBGE, 1985. Não publicado.

COCHRAN, W. G. *Sampling techniques.* 3rd ed. New York: Wiley, 1977. 428 p. (Wiley series in probability and mathematical statistics). Acesso em: set. 2017.

CRÍTICA quantitativa, qualitativa e fórmulas de cálculo do IPP. Rio de Janeiro: IBGE, 1982. Documento preliminar do Projeto Índice de Preços ao Produtor.

DEVILLE, J. C.; SÄRNDAL, C. E. Calibration estimators in survey sampling. *Journal of the American Statistical Association,* Alexandria [Estados Unidos]: American Statistical Association - ASA, v. 87, n. 418, p. 376-382, June 1992. Acesso em: set. 2017

ESTEVAO, V.; HIDIROGLOU, M. A.; SÄRNDAL, C. E. Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics,* Stockholm: Statistics Sweden, v. 11, n. 2, p. 181-204, 1995.

FARIAS, A. M. L.; BARBOSA, M. T. S. Estudos para definição do desenho amostral da pesquisa mensal de comércio: região metropolitana do Rio de Janeiro. *Revista Brasileira de Estatística,* Rio de Janeiro: IBGE, v. 54-57, n. 201-208, p. 7-46, 1993-1996. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

FEIJÓ, C. A.; SILVA, D. B. N.; CARVALHO, P. G. M. *Análise longitudinal de dados industriais:* projeto integrado de pesquisa CNPq. Rio de Janeiro: Escola Nacional de Ciências Estatísticas - ENCE, 1996. Não publicado.

FULLER, W. A. et al. *PC-CARP.* Ames: Statistical Laboratory, Iowa State University,1986.

GARCIA, R. M. *Estimação em pequeno domínio.* Rio de Janeiro: Instituto de Matemática Pura e Aplicada - IMPA, 1986. (Informes de matemática. Série D, 20).

GOLDSTEIN, H.; SILVER, R. Multilevel and multivariate models in survey analysis. In: SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. (Ed.). *Analysis of complex surveys*. Chichester; Wiley, 1989. cap. 11. (Wiley series in probability and mathematical statistics).

GÖTTGENS, R. et al. *Bascula version 1.0*: a weighting package under MS-DOS: user manual. Voorburg: Statistics Netherlands, 1991.

HIDIROGLOU, M. A. The construction of a self-representing stratum of large units in survey design. *The American Statistician*, Alexandria [Estados Unidos]: American Statistical Association - ASA, v. 40, n. 1, p. 27-31, Feb. 1986.

HOLT, D.; SMITH, T. M. F.; WINTER, P. D. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*. Series A, Statistics in Society. London, v. 143, p. 474-487, 1980.

INDICADORES conjunturais da indústria: produção, emprego e salário. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1991. 154 p. (Série relatórios metodológicos, v. 11). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv22891.pdf>. Acesso em: set. 2017.

MENEZES, A. C. F. de; ALMEIDA, R. A. P.; BIANCHINI, Z. M. *Revisão da PNAD*: a questão amostral: módulo II do anteprojeto. Rio de Janeiro: IBGE, Diretoria de Pesquisas,

1991. 33 p. (Textos para discussão, n. 54). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv22895.pdf>. Acesso em: set. 2017.

MENEZES, A. C. F. de et al. *Pesquisa nacional por amostra de domicílios*: redução do número de domicílios na amostra e proposta de novos estimadores. Rio de Janeiro: IBGE, 1986. 26 p.

METODOLOGIA da pesquisa mensal de emprego 1980. Rio de Janeiro: IBGE, 1983a. 88 p. (Série relatórios metodológicos, v. 2). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv12497.pdf>. Acesso em: set. 2017.

METODOLOGIA da pesquisa nacional por amostra de domicílios na década de 70. Rio de Janeiro: IBGE, 1981. 559 p. (Série relatórios metodológicos, v. 1). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv9765.pdf>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico de 1980. Rio de Janeiro: IBGE, 1983b. 478 p. (Série relatórios metodológicos, v. 4). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv13083.pdf>. Acesso em: set. 2017.

MOURA, F. A. S. *Small area estimation using multilevel models.* 1994. Tese (P.h.D)-Department of Social Statistics, University of Southampton, Southampton.

MULLER, C. C.; SILVA, G.; VILLALOBOS, A. G. Pesquisa agropecuária do Paraná: safra 1986-87 (Programa de aperfeiçoamento das estatísticas agropecuárias). *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 49, n. 191, p. 55-84, jan./jun. 1988. Disponível em: <http://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

NATHAN, G.; HOLT, D. The effect of survey design on regression analysis. *Journal of the Royal Statistical Society. Series B, Methodological*, London, v. 42, p. 377-386, 1980.

NATHAN, G.; SMITH, T. M. F. The effect of selection on regression analysis. In: SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. (Ed.). *Analysis of complex surveys*. Chichester; Wiley, c1989. cap. 7. (Wiley series in probability and mathematical statistics).

PESQUISA anual de comércio PAC. Rio de Janeiro: IBGE, 1991. 61 p. (Série relatórios metodológicos, v. 12). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv22904.pdf>. Acesso em: set. 2017.

PESQUISA de locais de compra PLC/1988. Rio de Janeiro: IBGE, 1994. 109 p. Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv24725.pdf>. Acesso em: set. 2017.

PESQUISA de orçamentos familiares: aspectos de amostragem. Rio de Janeiro: IBGE, 1992. v. 3. (Série relatórios metodológicos, v. 10). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv81836_v3.pdf>. Acesso em: set. 2017.

PESQUISA especial da indústria: notas metodológicas. Rio de Janeiro: IBGE, 1982. 132 p.

PESQUISA industrial de 1981: tratamento das situações especiais durante a fase de expansão. Rio de Janeiro: IBGE, 1983. Não publicado.

PFEFFERMANN, D. The role of sampling weights when modeling survey data. *International Statistical Review*, Medford: Wiley, v. 61, p. 317-337, 1993.

PLANO amostral da PIA-81. Rio de Janeiro: IBGE, 1982. 8 p. Não publicado.

PLANO de amostragem da PIA-83. Rio de Janeiro: IBGE, 1984. Não publicado.

PLANO de expansão da amostra de dados gerais da PIA 1981. Rio de Janeiro: IBGE, 1981.

ROYALL, R. M. On finite population sampling theory under certain linear regression models. *Biometrika*, Oxford: Oxford University Press, v. 57, n. 2, p. 377-387, Aug. 1970. Disponível em: <https://academic.oup.com/biomet>. Acesso em: set. 2017.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling*. New York: Springer-Verlag, 1992. 694 p. (Springer series in statistics).

SAUTORY, O. La macro CALMAR: redressement d'un échantillon par calage sur marges. Paris: Institut National de la Statistique et dês Études Economiques - INSEE, 1993. 51 p. (Série dês documents de travail, n. F 9310). Disponível em: <https://www.insee.fr/fr/information/2021902>. Acesso em: set. 2017.

SHAH, B. V. et al. *SUDDAN user's manual*: professional software for survey data analysis for multi-stage sample designs: release 6.0. Research Triangle Park, NC: Research Triangle Institute - RTI, 1992.

SILVA, D. B. N. *Aplicação de modelos lineares dinâmicos bayesianos em pesquisas repetidas no tempo*. 1992. Dissertação (Mestrado)- Instituto de Matemática e Estatística, Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro, 1992.

_____. *Modelling compositional time series from repeated surveys*. 1997. 237 p. Tese (P.h.D.)-Faculty of Mathematical Studies, University of Southampton, Southampton.

SILVA, L. F.; BIANCHINI, Z. M. A redução da amostra e a utilização de duas frações amostrais no censo demográfico de 1990. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1990. 49 p. (Textos para discussão, n. 33).

SILVA, P. L. N. et al. Frame problems and survey design for the Brazilian annual retail and wholesale trade survey. *Estadística*, v. 49-51, n. 152-157, p. 211-230, 1999.

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

_____. *Utilizing auxiliary information in sample survey estimation and analysis*. 1996. Thesis (P.h.D.)-Department of Social Statistics, University of Southampton, Southampton, 1996. 247 p.

SILVA, P. L. N.; BIANCHINI, Z. M. Apuração de dados no IBGE: problemas e perspectivas. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 54-57, n. 201-208, p. 47-112, 1997. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

SILVA, P. L. N.; BIANCHINI, Z. M.; ALBIERI, S. *Uma proposta de metodologia para a expansão da amostra do censo demográfico de 1991*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1993. 106 p. (Textos para Discussão, n. 62).

SILVA, P. L. N.; MOURA, F. A. S. *Efeito de conglomeração da malha setorial do censo demográfico de 1980*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1990. 115 f. (Textos para Discussão, n. 32). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv66426.pdf>. Acesso em: set. 2017.

_____. Redução da amostra da pesquisa mensal de emprego: estratégia para reduzir o custo da pesquisa. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 49, n. 192, p. 65-95, 1988. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

SILVA, P. L. N.; SKINNER, C. J. Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics*, Stockholm: Statistics Sweden, v. 11, n. 3, p. 277-294, [Dec.] 1995.

_____. Variable selection for regression estimation in finite populations. *Survey Methodology*, Ottawa: Statistics Canada, v. 23, n. 1, p. 23-32, Jun. 1997. Disponível em: <http://www5.statcan.gc.ca/olc-cel/olc.action?objId=12-001-X19970013102&objType=47&lang=en&limit=0>. Acesso em: set. 2017.

SISTEMA nacional de índices de preços ao consumidor: métodos de cálculo. Rio de Janeiro: IBGE, 1994. 105 p. (Série relatórios metodológicos, v. 14). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv82218.pdf>. Acesso em: set. 2017.

SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. (Ed.). *Analysis of complex surveys*. Chichester; Wiley, c1989. 309 p. (Wiley series in probability and mathematical statistics).

SMITH, T. M. F. Introduction to part b. In: SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. (Ed.). *Analysis of complex surveys*. Chichester ; Wiley, 1989. p. 146-147. (Wiley series in probability and mathematical statistics).

_____. Populations and selection: limitations of statistics. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, London, v.156, p. 145-164,1993.

SMITH, T. M. F. To weight or not to weight, that is the question. In: VALENCIA INTERNATIONAL MEETING ON BAYESIAN STATISTICS, 3., 1987, Valencia. *Bayesian statistics 3*: proceedings...Oxford: Oxford University Press, 1988. p. 437-451.

VASCONCELLOS, M. T. L. *Metodologia do estudo nacional da despesa familiar - ENDEF*: objetivos, descrição e metodologia usadas no ENDEF. Rio de Janeiro: IBGE,1983. 134 p.

WOLTER, K. M. *Introduction to variance estimation*. New York: Springer-Verlag, 1985. 427 p.

Documentação, eventos de capacitação e disseminação de conhecimentos

Sonia Albieri

Introdução

A unidade de metodologia do Instituto Brasileiro de Geografia e Estatística - IBGE foi criada por meio da Resolução do Presidente do IBGE n. 4, de 20.05.1977. Uma descrição sobre a história da unidade pode ser encontrada na publicação *A unidade de metodologia e a evolução do uso de amostragem no IBGE* (ALBIERI, 2003), onde é feito um destaque para o fato que a sua criação e evolução ter sido fortemente ligada à utilização de amostragem em pesquisas e censos, além de outros temas tais como os procedimentos de apuração, aí incluídos os métodos de crítica e imputação e os sistemas computacionais para a execução dessas atividades. Entretanto, uma parte importante das atividades desenvolvidas pela unidade sempre esteve ligada à difusão de metodologias estatísticas e computacionais aplicadas à produção e análise de informações, tanto que, na estrutura formal da unidade vigente entre agosto de 1985 e novembro de 1986, havia uma gerência destinada exclusivamente à documentação das metodologias usadas nas pesquisas e censos, chamada Núcleo de Documentação Metodológica[1]. Além disso, por conta de sua característica específica ligada aos métodos estatísticos, a formação e aperfeiçoamento técnico dos recursos humanos do Instituto também esteve sempre explicitada nas atribuições conferidas à unidade.

A definição mais atualizada das competências da unidade de metodologia da Diretoria de Pesquisas do IBGE está na Resolução do Conselho Diretor n. 13, de 02.06.2015, que define as competências das unidades da Diretoria de Pesquisas. O Art. 12 explicita que:

> À COORDENAÇÃO DE MÉTODOS E QUALIDADE - DPE/COMEQ compete:
>
> I - Pesquisar, desenvolver, absorver, difundir, implementar e apoiar na utilização de metodologias estatísticas e computacionais aplicadas à produção e análise de informações;
>
> II - Atuar junto à DPE na articulação de seus diversos órgãos para definir normas, procedimentos operacionais, rotinas, métodos e sistemas comuns na busca da eficiência e melhoria de qualidade; e,
>
> III - Contribuir para a formação e aperfeiçoamento técnico dos recursos humanos (IBGE, 2015, p. 1).

[1] Conforme os Art. 9º e 10º da Resolução do Presidente do IBGE n. 40, de 16.08.1985, que cria o Centro de Ensino e Desenvolvimento Metodológico e dá outras providências.

Este capítulo descreve de forma resumida um conjunto de ações lideradas pela unidade de metodologia no que se refere à documentação dos métodos usados no processo de produção de informações estatísticas e aos eventos de capacitação e de disseminação de conhecimentos, tais como seminários, encontros, minicursos, participação em congressos, visitas técnicas, cursos do programa anual de treinamento do IBGE, etc.

Outros documentos com avaliações ou descrições de atividades anteriormente realizadas foram usados como referência e estão devidamente citados no decorrer do texto e nas referências bibliográficas.

Documentação

Em relação às atividades relacionadas com a documentação de métodos, processos e sistemas usados durante a produção de informações estatísticas, mais especificamente, durante a realização de pesquisas e censos, vale destacar a iniciativa de publicar o primeiro volume da Série Relatórios Metodológicos, efetivada em 1981, com a publicação de *Metodologia da pesquisa nacional por amostra de domicílios na década de 70* (1981). Esta série tem a seguinte nota de escopo: "divulga as metodologias empregadas nas diversas fases do planejamento e execução das pesquisas do IBGE" e a unidade de metodologia ainda esteve envolvida diretamente com a preparação de mais cinco volumes de divulgação, até 1989, quando a preparação da publicação passou a ser responsabilidade da unidade responsável pela própria pesquisa ou estudo, como mais uma etapa do processo de produção.

Nesses 36 anos de existência da série, foram publicados 43 volumes, descrevendo as metodologias de operações estatísticas tais como: Censo Demográfico, Censo Agropecuário, Pesquisa Mensal de Emprego - PME, Pesquisa Industrial Anual - Empresa - PIA-Empresa, Pesquisa Anual de Comércio - PAC, Pesquisa Anual de Serviços - PAS, Pesquisa Anual da Indústria da Construção - PAIC, Pesquisa Anual do Transporte Rodoviário - PATR, pesquisas agropecuárias anuais (Produção Agrícola Municipal - PAM, Produção da Pecuária Municipal - PPM, e Produção da Extração Vegetal e da Silvicultura - PEVS), Pesquisa de Orçamentos Familiares - POF, Contas Nacionais, Contas Regionais, Produto Interno Bruto - PIB, indicadores conjunturais da indústria, Pesquisa Mensal do Comércio - PMC, Pesquisa Mensal de Serviços - PMS, Índice de Preços ao Produtor - IPP, Pesquisa de Inovação - Pintec, Sistema Nacional de Índices de Preços ao Consumidor - SNIPC, Economia Informal Urbana - Ecinf, Matriz de Insumo-Produto, Projeções de População, Estimativas da População e Regionalização das Transações do Setor Público. Vale destacar que algumas dessas operações estatísticas tiveram mais de uma edição dessa publicação em função das atualizações implementadas em sua metodologia de execução.

Outro importante veículo de documentação de estudos e processos relacionados com a produção de informações é a série Textos para Discussão, da Diretoria de Pesquisas[2], que tem como característica importante o fato de ser constituída por documentos autorais, diferentemente da série anteriormente descrita, que é constituída por publicações institucionais. De janeiro de 1988 até dezembro de 1999, foram 97 números, com objetivo de divulgar estudos e outros trabalhos técnicos nas áreas social e demográfica, elaboradas no âmbito da Diretoria. A partir de 2000, após uma reformulação editorial, a série Textos para Discussão passou a ter a seguinte nota de escopo: "Divulga estudos e outros trabalhos técnicos desenvolvidos pelo IBGE ou em conjunto com outras instituições, bem como resultantes de consultorias técnicas e traduções consideradas relevantes para disseminação pelo Instituto. A série está subdividida por unidade organizacional e os textos são de responsabilidade de cada área específica". No âmbito da Diretoria de Pesquisas, de janeiro de 2000 até dezembro de 2015, foram publicados 56 números dessa nova série.

Além de ser uma publicação perfeitamente alinhada com a ideia de documentação de métodos, processos e análises referentes ao processo de produção de informações

[2] A série Textos para Discussão, da Diretoria de Pesquisas, está disponível para pesquisa no site da biblioteca do IBGE, no endereço <https://biblioteca.ibge.gov.br/>.

estatísticas, no sentido mais amplo do termo, que abrange desde a concepção, o planejamento, a execução, o processamento, a análise e a divulgação de informações, os técnicos e pesquisadores da unidade de metodologia sempre fizeram questão de produzir documentos para a série como parte de suas atividades profissionais.

A documentação é também o foco principal do sistema de metadados do IBGE, que foi iniciado na década de 1980, por iniciativa da Diretoria de Informática, quando era um sistema acessível apenas internamente no IBGE. Em 2009, o Banco de Metadados - MetaBD foi disponibilizado no portal do IBGE na Internet, contendo os metadados de referência de uma ocorrência de cada operação estatística em produção naquele ano. Atualmente, o sistema de metadados estatísticos encontra-se em fase final de um processo de reformulação e modernização, para dar conta não só dos metadados de referência, mas também dos indicadores de qualidade associados com as operações estatísticas que estão detalhados no Capítulo **Qualidade estatística** desta publicação.

Seminário IBGE

Em abril de 1997, teve início uma série semanal de seminários, intitulada Seminários LEP, sigla que vem de Laboratório de Estatísticas Públicas, uma unidade virtual, criada pela Resolução do Conselho Diretor n. 19, de 23.12.1996, vinculada à presidência do IBGE, e que atuou em curto período de tempo, até 1998. Entretanto, os Seminários LEP continuaram e continuam sendo realizados nas instalações do IBGE. Organizados pela unidade de metodologia, os seminários são abertos à participação de todos, servidores do IBGE ou convidados, não sendo necessária inscrição prévia para assistir. Os assuntos abordados são os mais variados, podendo versar sobre trabalhos finalizados ou em andamento, estudos sobre temas emergentes ou histórias consolidadas, ou mesmo curiosidades tais como alimentação saudável ou a relação entre a matemática e a música. Os palestrantes também podem ser técnicos das diversas unidades do IBGE – voluntários ou convidados – ou pesquisadores convidados de outras instituições.

Durante os anos de 2000 a 2002, a organização dos seminários passou a ser realizada pela Coordenação de Treinamento, na Presidência do IBGE, juntamente com o Departamento de Metodologia da Diretoria de Pesquisas, que continuou com as atividades de apoio na organização da infraestrutura e na divulgação. Desde 2003, a organização dos seminários voltou para o Departamento de Metodologia da Diretoria de Pesquisas, atual Coordenação de Métodos e Qualidade, ficando sob a responsabilidade da área de métodos.

Para contemplar com maior clareza sua ampla abrangência temática, e seu caráter institucional, a partir de julho de 2009, o nome foi alterado para Seminário IBGE. Atualmente, existe um sistema de reserva de Seminário IBGE disponível na página na Intranet da Diretoria de Pesquisas, que garante a disponibilidade de auditório em datas fixadas, facilitando o agendamento por parte dos interessados em apresentar um seminário. A divulgação de cada Seminário IBGE é feita por meio de publicação de mensagem na Intranet do IBGE e através de cartazes impressos, contendo título, nome do apresentador ou palestrante, instituição ou unidade do IBGE do palestrante, e um resumo do assunto a ser tratado. Os cartazes impressos são afixados nos quadros de aviso físicos dos diversos prédios do IBGE, no Rio de Janeiro.

Vale dizer que, em outubro de 2012, foi publicado o n. 42 da série Textos para Discussão da Diretoria de Pesquisas, intitulado *Seminários IBGE: 15 anos disseminando conhecimento* (ALBIERI, 2012). Ao todo, de abril de 1997 a junho de 2017, foram realizados 475 seminários dessa série.

Seminário de Metodologia do IBGE - SMI

Em 2012, o IBGE promoveu o seu primeiro Seminário de Metodologia do IBGE - SMI, com o objetivo de propiciar espaço e oportunidade para discussão e reflexão sobre os avanços, desafios e perspectivas da metodologia relacionada à produção de informações.

O SMI reúne pesquisadores e técnicos do IBGE, de organizações públicas e privadas, representantes dos institutos nacionais de estatística de países da região, representantes de entidades nacionais, estaduais e municipais de estatística e de geociências e membros da comunidade acadêmica nacional e internacional interessados no tema.

O evento tem periodicidade anual, sendo que foram realizadas cinco edições, de 2012 a 2016 e a edição de 2017 já está com seu planejamento iniciado. O período de realização e o tema principal de cada edição são apresentados a seguir, mas uma descrição das características e das atividades desenvolvidas em cada uma das edições até SMI 2016[3] são detalhadas por Albieri e outros (2017).

- 1º Seminário de Metodologia do IBGE - SMI2012 e 11ª Reunião IASI sobre Estatística Pública, realizado de 5 a 9 de novembro de 2012, com o tema: Preservação, disseminação e confidencialidade de dados.
- 2º Seminário de Metodologia do IBGE - SMI2013, realizado de 25 a 29 de novembro de 2013, com o tema: Agregando valor aos dados: combinação, modelagem e análise de informações e estatísticas públicas.
- 3º Seminário de Metodologia do IBGE - SMI2014, realizado de 5 a 7 de novembro de 2014, com o tema: Desafios e oportunidades para a obtenção de dados.
- 4º Seminário de Metodologia do IBGE - SMI2015, realizado de 1 a 4 de dezembro de 2015, com o tema: Modernização e qualidade na produção de informações.
- 5º Seminário de Metodologia do IBGE - SMI2016, realizado de 7 a 10 de novembro de 2016, com o tema: Integração de dados estatísticos e geoespaciais e visualização de informações.
- 6º Seminário de Metodologia do IBGE - SMI2017, programado para 7 a 10 de novembro de 2017, com o tema: Censos.

Organização de eventos científicos

No âmbito de suas atividades de pesquisa de novos métodos e sistemas e com o objetivo de promover intercâmbio entre pesquisadores internos e externos ao IBGE e técnicos de instituições similares, além de estimular as participações individuais em congressos internacionais, a unidade de metodologia tem sido responsável ou vem participando da organização de eventos científicos nacionais e internacionais no Brasil, que estão descritos a seguir.

- Workshop sobre Métodos de Crítica e Imputação de Dados – promovido pela Diretoria de Pesquisas do IBGE em conjunto com a Escola Nacional de Ciências Estatísticas - ENCE do IBGE. Foi realizado nas dependências da ENCE, no Rio de Janeiro, de 12 a 16 de fevereiro de 1990. Além de conferências e apresentações orais, foram realizados quatro minicursos relacionados com o tema. Contou com a participação de convidados internacionais, como conferencistas e instrutores, e a participação de diversos especialistas no tema provenientes de institutos pesquisa ou de institutos oficiais de estatística de países tais como Suécia, Espanha, Estados Unidos, Canadá e Uruguai.
- Seminário Internacional sobre Metodologias para Pesquisas Domiciliares por Amostragem e 3ª Reunião sobre Estatística Pública do Instituto Interamericano de Estatística (Inter-American Statistical Institute - IASI) – promovido pelo IBGE em conjunto com o IASI, foi realizado nas dependências do Centro de Documentação e Disseminação de Informações do IBGE, de 25 a 27 de junho de 2001.
- 9º Seminário do IASI de Estatísticas Aplicadas: Estatística na Educação e Educação na Estatística, promovido pelo IBGE em conjunto com o IASI. Foi realizado nas dependências do Instituto de Matemática Pura e Aplicada - IMPA, no Rio de Janeiro, de 7 a 10 de julho de 2003.

[3] Mais informações sobre o evento estão disponíveis na Internet, no endereço <http://eventos.ibge.gov.br/smi2017>.

• Seminário Internacional sobre Crítica e Imputação – promovido pela Diretoria de Pesquisas em conjunto com a ENCE, foi realizado nas dependências do Centro de Documentação e Disseminação de Informações, no Rio de Janeiro, de 28 de novembro a 02 de dezembro de 2005.

• Seminário sobre Métodos Estatísticos para a Produção de Pesquisas: pensar o futuro após 30 anos da unidade de metodologia no IBGE - MEP 30 – foi promovido pela Diretoria de Pesquisas, nas dependências do Centro de Documentação e Disseminação de Informações, no Rio de Janeiro, nos dias 22 e 23 de maio de 2007.

• 1ª Escola de Amostragem e Metodologia de Pesquisa - Esamp foi promovida pelo IBGE, em conjunto com Fundação Joaquim Nabuco - Fundaj, realizada nas dependências do Centro de Documentação e Disseminação de Informações, no Rio de Janeiro, de 21 a 23 de novembro de 2007.

A 1ª Esamp teve como principal objetivo oferecer, pela primeira vez no Brasil, uma oportunidade para congregar estatísticos, pesquisadores e profissionais de pesquisa social das universidades e de diversos órgãos produtores de informação, como o IBGE e a Fundaj, visando discutir suas experiências à luz dos mais recentes desenvolvimentos metodológicos em planejamento amostral e análise de dados de levantamentos amostrais. O sucesso do evento motivou a realização de outras edições da Esamp, apresentadas a seguir, sempre com a participação da unidade de metodologia do IBGE e da ENCE na organização dos eventos.

• 2º Esamp e 1º International Workshop on Surveys for Policy Evaluation – foram promovidos pela Universidade Federal do Rio Grande do Norte - UFRN, em conjunto com o Sebrae-RN, realizada em Natal, na sede do Sebrae-RN, de 3 a 6 de novembro de 2009.

• 3º Esamp e 2º International Workshop on Surveys for Policy Evaluation – foram promovidos pela Universidade Federal de Juiz de Fora - UFJF, no Instituto de Ciências Exatas da UFJF, de 22 a 25 de novembro de 2011.

• 4º Esamp e 3º International Workshop on Surveys for Policy Evaluation – foi promovida pela Universidade de Brasília - UnB, realizada nas dependências da UnB em Brasília, de 5 a 8 de novembro de 2014.

• 5º Esamp e 4º Workshop Internacional sobre Pesquisas para Avaliação de Políticas Públicas – sob a responsabilidade da Universidade Federal do Mato Grosso - UFMT e da Associação Brasileira de Estatística - ABE, realizada no espaço Cenarium Rural, em Cuiabá, Mato Grosso, de 17 a 20 de outubro de 2017.

Encontros da Coordenação de Métodos e Qualidade

O primeiro Encontro da unidade de metodologia surgiu em 2002, quando o então Departamento de Metodologia, da Diretoria de Pesquisas, recebeu novos servidores do concurso público realizado em 2001 e os gerentes juntamente sentiram a necessidade de dar conhecimento de seus projetos e atividades aos novos servidores, bem como conhecer as características técnicas e a experiência acadêmica ou profissional desses novos servidores. As competências da unidade definidas na resolução de sua criação foram determinantes na definição do formato da iniciativa, a de um seminário técnico, com um ou mais apresentadores por assunto.

Os Encontros têm por objetivo permitir o intercâmbio de conhecimentos, a divulgação das atividades em desenvolvimento na Unidade, a integração entre os técnicos das gerências que compõem a Unidade, a abertura de canal de comunicação entre os servidores entre si e entre os servidores e as chefias, além de permitir que um servidor se apresente para participar do desenvolvimento de alguma atividade específica e apresente suas queixas e sugestões de qualquer natureza relacionadas com as atividades, a gestão dessas atividades e o ambiente de trabalho onde as atividades são desenvolvidas.

Uma característica importante desse evento é a sua realização em local diferente do ambiente de trabalho diário, para permitir o completo afastamento das questões do dia a dia e a concentração total no evento. Além disso, os intervalos para café e os almoços são planejados para serem aproveitados em conjunto pela equipe, para estimular o convívio social, possibilitando a humanização dos assuntos tratados nas conversas informais que fujam dos assuntos puramente técnicos.

O projeto foi vencedor do 1º Concurso Práticas Inovadoras em Gestão do IBGE, na área de Gestão e Desenvolvimento de Pessoas. A inovação do projeto dentro do IBGE se deu pela implementação, no ambiente de trabalho, de um tipo de atividade bastante conhecido e realizado nos meios acadêmicos, qual seja a realização de seminários ou congressos temáticos para intercâmbio de conhecimentos, além de promover a socialização entre membros de equipes diferentes, que não costumam se relacionar no dia a dia do trabalho, por não compartilharem do mesmo projeto.

Os Encontros não possuem periodicidade definida. Foram realizadas 7 edições do evento, em 2002, 2003, 2006, 2010, 2011, 2014 e 2016, com duração variando de 1 a 3 dias.

Além desses encontros internos da unidade de metodologia, nos dias 2 e 3 de junho de 2015, foi realizado o 1º Encontro ENCE-Coordenação de Métodos e Qualidade, como um fórum de discussão e intercâmbio de conhecimentos, envolvendo as equipes de professores da ENCE e servidores da Coordenação.

Um dos objetivos deste encontro foi aproximar a área de pesquisas da ENCE com a área de estudos da Coordenação, que tem entre suas competências a absorção de conhecimento e de metodologias para o aprimoramento do processo de produção de informações estatísticas. Especificamente, busca-se identificar temas de interesse dessas duas áreas e avaliar a possibilidade de realizar trabalhos em conjunto ou de forma integrada.

Um Encontro envolvendo não só a ENCE e a Coordenação de Métodos e Qualidade, mas também a Coordenação de Trabalho e Rendimento, também da Diretoria de Pesquisas, foi realizado nos dias 22 e 23 de agosto de 2017, ampliando assim o escopo de projetos a serem estudados pelas áreas.

Ciclo de Palestras sobre Métodos de Pesquisa IBGE

Com o objetivo de inserir os novos servidores da Diretoria de Pesquisas, que iniciaram suas atividades em 2016, no conjunto de atividades centrais da Diretoria, qual seja, a produção de informações, a Coordenação de Métodos e Qualidade, organizou um ciclo de palestras sobre as diversas etapas de um processo de produção de pesquisas estatísticas[4]. As apresentações foram feitas no período de 31 de outubro de 2016 a 20 de março de 2017, nas tardes de segunda-feira, no auditório do IBGE, e foram transmitidas pela TV IBGE e por *Web Cast*, atingindo assim os servidores do IBGE das diversas unidades, inclusive as sedes das Unidades Estaduais e das Agências do IBGE espalhadas pelo país. Os seminários foram abertos e diversos servidores com mais tempo no IBGE e mesmo servidores de outras diretorias participaram eventualmente de algum seminário. O material usado durante as apresentações e os vídeos das palestras foram disponibilizados na Intranet da Diretoria de Pesquisas, e pretende-se que seja utilizado para a organização de um curso no âmbito do Programa Anual de Capacitação do IBGE, no formato presencial e on-line.

Capacitação

Entre as competências atribuídas à Coordenação de Métodos e Qualidade está a de "contribuir para a formação e aperfeiçoamento técnico dos recursos humanos"

[4] O ciclo de palestras teve como base a disciplina de Metodologia de Pesquisa ministrado pela pesquisadora da Coordenação Maria Luiza Zacharias, no programa de pós-graduação da ENCE.

(IBGE, 2015, p. 1). Assim, além das atividades de documentação e disseminação de conhecimentos através da preparação de artigos e participação em eventos científicos, destaca-se a atuação dos técnicos da unidade de metodologia como:

- Professores colaboradores dos programas de graduação e de pós-graduação da ENCE;
- Instrutores dos cursos de treinamento do programa de treinamento do IBGE, conduzido também pela ENCE;
- Instrutores do Curso de Desenvolvimento de Habilidades em Pesquisa - CDHP[5], um dos principais cursos de formação de técnicos do IBGE, que contou com a participação de diversos técnicos da unidade de metodologia em sua implementação, e também na execução das diversas edições do curso desde 1997.

Congressos e visitas técnicas

Por conta de sua característica específica ligada aos métodos estatísticos voltados para a produção de informações, é fundamental que os técnicos e pesquisadores lotados na unidade de metodologia estejam atualizados nos diversos campos de atuação da área. Assim, a participação em congressos, nacionais e internacionais, e as visitas técnicas aos institutos de pesquisa, sejam órgãos produtores oficiais de estatística ou instituições de pesquisa acadêmica, é atividade não só desejada como necessária.

E assim vem sendo desde sua criação. É impossível listar todos os eventos dessa natureza, mas alguns merecem destaque. No campo internacional, tem havido participação no Congresso Mundial de Estatística (World Statistics Congress), promovido pelo Instituto Internacional de Estatística (International Statistical Institute - ISI), desde 1853, e na Conferência Européia sobre Qualidade em Estatísticas Oficiais, que vem sendo realizada desde o início dos anos 2000. No campo nacional, é importante destacar o Simpósio Nacional de Probabilidade e Estatística - Sinape, realizado desde 1974, e, mais recentemente, a Esamp, ambos promovidos pela ABE.

Esses eventos congregam uma grande quantidade de participantes, que incluem estatísticos acadêmicos, de órgãos do governo, da iniciativa privada e de diversos institutos de pesquisas.

Congressos e seminários dessa natureza são importantíssimos para estabelecer contatos com profissionais, inclusive de outros países, dedicados a problemas de mesma natureza, o que facilita em muito os contatos e intercâmbios futuros. Além disso, ressalte-se o investimento que a Instituição faz em seus melhores e promissores quadros técnicos na área de metodologia estatística.

As visitas técnicas a outros institutos produtores de informações oficiais, para conhecimento e debate sobre métodos e processos de produção, ou mesmo de sistemas específicos, e as visitas de consultores estrangeiros especialistas em temas de interesse voltados para a produção e análise de informações oficiais também tem sido de grande valor para o aperfeiçoamento dos técnicos e a melhoria da qualidade do processo de produção de informações estatísticas. Também nesse caso é impossível listar os eventos dessa natureza ocorridos nesses 40 anos.

Conclusões

Como pode ser visto, a documentação dos métodos usados no processo de produção de informações estatísticas teve início no começo da década de 1980, como parte importante das atividades da unidade de metodologia criada em 1977, que cumpriu seu papel de conscientizar as diversas unidades produtoras da Diretoria

[5] Mais informações sobre o CDHP podem ser encontradas na página da ENCE na Internet, no endereço <http://www.ence.ibge.gov.br/index.php/portal-cdhp/portal-cdhp-apresentacao>.

de Pesquisas do IBGE sobre a necessidade e a importância de ter todos os métodos e processos devidamente registrados, documentados e disseminados. De fato, a Coordenação de Métodos e Qualidade ainda vem exercendo esse importante papel, no desenvolvimento das atividades relacionadas com o sistema de metadados estatísticos, em fase de reformulação e modernização para a ampliação de seu escopo.

No que se refere aos eventos científicos, a realização bem sucedida de cinco edições anuais do SMI, de 2012 a 2016, permite concluir que este é um evento devidamente consolidado como uma das atividades importantes que o IBGE encontrou para discutir, com os pesquisadores internos e externos, os diversos aspectos metodológicos relacionados com a produção de informações. A quantidade e variedade de temas para serem avaliados e discutidos por todos os interessados em metodologia de produção de informações nos leva a prever vida longa para eventos dessa natureza.

E na área específica do envolvimento da unidade de métodos e qualidade com as atividades de ensino e pesquisa desenvolvidas pela ENCE, os eventos recentes, como os encontros entre a ENCE, a Coordenação de Métodos e Qualidade e a Coordenação de Trabalho e Rendimento, mostram-se promissores para a intensificação do intercâmbio, da realização de atividades em conjunto e da realização de estudos e pesquisas em temas de interesse específicos voltados para produção de informações e métodos associados.

Referências

ALBIERI, S. *A unidade de metodologia e a evolução do uso de amostragem no IBGE.* Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 42 p. (Textos para discussão, n. 12). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv2282.pdf>. Acesso em: set. 2017.

ALBIERI, S. *Seminários IBGE:* 15 anos disseminando conhecimento. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2012. 68 p. (Textos para discussão, n. 42). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv62894.pdf>. Acesso em: set. 2017.

ALBIERI, S. et al. *Atividades realizadas durante os seminários de metodologia do IBGE:* SMI2012 a SMI2016. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2017. 39 p. Disponível em: <https://eventos.ibge.gov.br/images/smi2017/Relat%C3%B3rio%20dos%20Semin%C3%A1rios%20de%20Metodologia%20do%20IBGE%202012%20a%202016.pdf>. Acesso em: out. 2017.

IBGE. *Resolução do Conselho Diretor n. 13, de 2 de junho de 2015.* Define as competências das unidades da Diretoria de Pesquisas. Rio de Janeiro, 2015.

_____. *Resolução do Conselho Diretor n. 19, de 23 de dezembro de 1996.* Estabelece o Laboratório de Estatísticas Públicas, unidade virtual dedicada à pesquisa sobre temas ligados à produção, análise e disseminação de informações e estatísticas públicas. Rio de Janeiro, 1997.

_____. Resolução n. 4, de 20 de maio de 1977. Dispõe sobre a estrutura, competência e atribuições dos órgãos de Assessoramento Superior, das Diretorias e Unidades Regionais do IBGE, e dá outras providências. *Boletim de Serviço,* Rio de Janeiro, n. 1296, p. 1, 10 jun. 1997.

_____. Resolução n. 40, de 16 de agosto de 1985. Cria o Centro de Ensino e Desenvolvimento Metodológico e dá outras providências. *Boletim de Serviço,* Rio de Janeiro, n. 1703, p. 2, 26 ago. 1985.

METODOLOGIA da pesquisa nacional por amostra de domicílios na década de 70. Rio de Janeiro: IBGE, 1981. 698 p. (Série relatórios metodológicos, v. 1). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv9765.pdf>. Acesso em: set. 2017.

O Canadian Census Edit and Imputation System - Canceis no IBGE

Ari do Nascimento Silva
Bruno Freitas Cortez

Introdução

A tarefa de análise da consistência (crítica) e imputação de dados é parte do processo de produção de resultados de censos e pesquisas no Instituto Brasileiro de Geografia e Estatística - IBGE. De forma sucinta, pode-se definir crítica como sendo a etapa que visa identificar tanto itens (variáveis) não respondidos pelo respondente como também os itens com respostas inconsistentes, segundo determinada regra. A etapa de imputação visa atribuir valores a uma ou mais variáveis para completar um registro ou corrigir inconsistências verificadas na etapa de crítica.

O objetivo deste documento é o de contextualizar o software Canceis, como e quando foi usado, sua história, sua metodologia, vantagens e desvantagens, e possíveis usos futuros. É preciso esclarecer que quando se fala em Canceis, na verdade estamos falando no binômio NIM-Canceis, sendo o New Imputation Methodology - NIM a versão inicial do Canceis.

Um pouco de história

O Canceis, um sistema de crítica e imputação de dados, ficou conhecido no IBGE durante o planejamento do Censo Demográfico 2000, através de contatos com técnicos do Statistics Canada - StatCan, notadamente Michael Bankier, considerado o seu desenvolvedor. Nessa época, estamos falando do ano de 1998, o Canceis nos foi apresentado sob a forma de um protótipo chamado NIM. Nos primeiros documentos sobre o NIM ele era um acrônimo de New Imputation Methodology, mas dadas suas características técnicas de uso de doadores para imputação, seu nome oficial foi mudado para Nearest-neighbor Imputation Methodology.

A ideia de usar o NIM era para complementar o sistema Detección e Imputación Automática de Errores para Datos Cualitativos - DIA desenvolvido pelo Instituto Nacional de Estadística - INE da Espanha e usado na crítica e imputação do Censo Demográfico 1991. O NIM atuaria na etapa de críticas entre pessoas de um mesmo domicílio, área onde o DIA era reconhecidamente limitado. Dentro deste desenho lógico, o NIM executaria a consistência estrutural

do domicílio (relações entre as pessoas), e o DIA faria todas as demais consistências. A automação desta consistência estrutural prévia era importante para agilizar o processo de produção do censo, uma vez que esta etapa era realizada de modo semimanual (as regras de crítica eram aplicadas automaticamente, mas as correções eram manuais).

O desenvolvimento do NIM, em sua versão para o *mainframe*, foi iniciada em 1992 por Bankier, e esta versão foi usada no processamento do Censo de População de 1996 do Canadá, para a crítica e imputação das variáveis demográficas de idade, sexo, estado conjugal, concubinato e relação de parentesco com a primeira pessoa. Posteriormente Bankier desenvolveu um protótipo programado em linguagem C para funcionar em ambiente de microcomputador, com uma série de modificações na metodologia e no desenho do sistema propriamente dito. Esta foi a versão do NIM que foi usada para transferência de tecnologia, com a entrega de todos os programas-fonte, o que nos possibilitou fazer as mudanças necessárias de adaptação a nossos propósitos. Não é o caso de detalhar todas as modificações implementadas nos códigos-fonte, mas em particular uma delas foi impactante porque permitiu a análise da consistência entre a existência de empregados domésticos e de parentes do empregado doméstico no domicílio. Tanto é assim que, no momento de definir qual software usar para a consistência do Censo Demográfico 2010, optamos pela versão *in-house* do NIM, e não pelo uso da versão Canceis (já disponível, mas que também não contemplava uma função de pertinência entre empregado e parente do empregado).

Em setembro de 2004, o IBGE recebeu oficialmente a versão 3.2 do Canceis, sendo que entre 2000 e 2004 tivemos em mãos várias versões do Canceis, para avaliação e testes. O software enviado pelo StatCan é bastante completo e organizado, com os programas de execução, diretórios com arquivos de exemplos e documentação correspondente, um diretório adicional para Ferramentas, e um diretório para a instalação da versão *Windows*. Esta foi a versão do Canceis usada no processamento do Censo Agropecuário 2006. Posteriormente foram recebidas versões mais recentes para processamento da Pesquisa de Orçamentos Familiares - POF, Pesquisa Nacional por Amostra de Domicílios - PNAD, e Pesquisa Mensal de Emprego - PME.

Metodologia do Canceis

Para efeitos descritivos, usaremos os registros do Censo Demográfico como exemplo, e os seus elementos de domicílio e pessoas. O Canceis trata o conjunto de registros do domicílio e as pessoas que pertencem a este domicílio como um único registro lógico.

Domicílios que satisfazem todas as regras de crítica são ditos como registros que passam nas regras, e são chamados de doadores. Em contrapartida, domicílios que falham alguma das regras de crítica são chamados de receptores. O resultado da imputação é chamado de domicílio imputado, ou domicílio híbrido, por ser uma combinação entre o doador e o receptor.

Talvez a maior diferença entre o Canceis e outros métodos baseados na teoria de Fellegi e Holt esteja na sua interpretação. Esta estabelece que, para um determinado registro com erro, se deve dar prioridade à imputação do menor número de variáveis. O Canceis, apesar de também levar em conta o critério do menor número de modificações, enfatiza primeiro a busca de um único doador, e depois usa o critério do menor número de imputações.

Objetivos do Canceis

- Preservar ao máximo as informações coletadas, de tal forma que o registro com imputação seja o mais parecido possível com o registro que falhou nas regras de crítica; ou seja, para um dado conjunto de doadores, imputar o menor número possível de variáveis.

• Utilizar um único registro doador para a imputação de dados num mesmo registro.

• Procurar com que o registro com dados imputados seja o mais parecido possível com o registro doador, em busca do objetivo de obter um registro plausível que contenha a combinação de respostas imputadas e não imputadas.

• Permitir que ações de imputações igualmente corretas, baseadas nos doadores disponíveis, tenham chances similares de seleção, para evitar um falso aumento de tamanho de grupos pouco representativos da população (exemplo: pessoas com mais de 100 anos).

As regras de crítica podem se referir a inconsistências dentro de um mesmo registro ou entre registros distintos. O primeiro caso se refere a possíveis inconsistências encontradas nas respostas de um informante, enquanto o segundo é quando dois registros distintos não contenham inconsistências internamente, mas possuam alguma inconsistência ao serem confrontados entre si.

Desta forma, caso seja utilizado o expediente de críticas entre registros há a necessidade da divisão do arquivo de dados em estratos de igual tamanho da unidade de investigação de interesse. Por exemplo, considerando um domicílio que contenha um ou mais registros de pessoas dentro dele, ao se proceder a crítica entre registros (no caso, entre pessoas), o banco de dados deve ser dividido em estratos de domicílio com mesmo número de moradores, domicílios de uma pessoa formam o estrato 1, domicílios com duas pessoas formam o estrato 2, etc.

Para cada estrato, o procedimento inicia pela identificação dos registros que passaram pelas regras de crítica, formando o conjunto de doadores, e o conjunto dos registros que falharam em alguma regra (receptores). Para cada um dos receptores, determinam-se quais dos registros que passaram pelas regras de crítica (potenciais doadores) são similares, tanto quanto possível, ao registro que falhou.

Esses registros são denominados vizinhos mais próximos (a pesquisa do vizinho mais próximo é feita no domicílio, e não na pessoa). Para isso, é feito o batimento do domicílio que falhou com cada potencial doador, e através de uma medida de distância são destacados aqueles com o maior número de variáveis categóricas possíveis coincidentes, e com a menor diferença entre as variáveis contínuas. Se existirem muitos registros nestas condições, a preferência será dada para aqueles que estejam mais próximos fisicamente (se os arquivos estiverem classificados pelo código geográfico, isto significaria, mas não necessariamente garantiria, uma proximidade geográfica).

Para cada vizinho mais próximo, são identificadas todas as possíveis ações de imputação geradas com base em todos os possíveis subconjuntos de variáveis não coincidentes, que, se imputadas, permitiriam ao domicílio passar nas regras de crítica. Duas variáveis numéricas (quantitativas) são consideradas não coincidentes dependendo de valores da distância entre elas.

Uma possível ação de imputação é considerada essencialmente nova se nenhum subconjunto de variáveis imputadas por aquela ação de imputação representar uma outra possível ação de imputação. Todas as ações de imputação que não são essencialmente novas são descartadas, pois uma ou mais variáveis estariam sendo desnecessariamente imputadas, e assim o princípio de mudar o mínimo conjunto de dados seria violado.

A distância entre o registro que falhou e cada registro que passou pelas regras de crítica é usada para identificar as ações de imputação com mínima modificação. É esperado que a distância entre uma possível ação de imputação e o registro que falhou, mais a distância entre a possível ação de imputação e o registro que passou seja igual à distância entre o registro que falhou e o registro que passou. Uma média ponderada desta distância é calculada para cada ação de imputação (D_{fpa}). As ações de imputação essencialmente novas com valores mínimos de D_{fpa} são chamadas de ações de imputação com mínima modificação (minimum change imputation actions).

$$D_{fpa} = \alpha D_{fa} + (1-\alpha) D_{ap}$$

tal que:

D_{fa} Distância entre a ação de imputação e o registro receptor (as letras *f* e *a* denotam *failed* e *action* respectivamente);

D_{ap} Distância entre a ação de imputação e o registro doador, isto é, o vizinho mais próximo usado (as letras *a* e *p* denotam *action* e *plausibility*, para plausibilidade, porque quanto menor esta distância, mais plausível será a imputação, posto que estaria mais próxima de um registro real do arquivo);

α Parâmetro de ponderação das duas distâncias, que deve estar na faixa entre (0,5 e 1]. Quanto maior for este parâmetro mais importância se dá ao critério de imputar o menor número de variáveis, isto é, o registro imputado se parecerá mais com o registro que falhou.

Uma destas ações de imputação com mínima modificação é selecionada aleatoriamente para imputar o domicílio que falhou nas regras de crítica.

Tabelas lógicas de decisão

Independente do sistema ou aplicativo utilizado na etapa de crítica, as regras de inconsistência devem ser definidas e, posteriormente, programadas de alguma forma. No caso específico do Canceis, as mesmas são dispostas no formato de Tabelas Lógicas de Decisão (Decision Logic Tables - DLTs).

Na DLT as linhas informam as proposições (equações lineares envolvendo as variáveis da pesquisa), e as colunas informam as regras de consistência. Os valores de cada célula da matriz podem ser: a) 'Y' (satisfaz); b) 'N' (não satisfaz); e c) ' ' (espaço em branco, indicando que a proposição não se aplica à regra em questão).

Para uma determinada regra, se todas as proposições são satisfeitas, então diz-se que o registro satisfaz à regra. Basta que uma proposição não seja satisfeita para que a regra não seja satisfeita.

No Canceis as regras são definidas como regras de conflito, isto é, definem as condições de erro: se uma regra é satisfeita então o registro está errado, isto é, não passa na consistência. Basta que uma regra seja satisfeita para que o registro não passe na consistência. A Tabela 1 mostra um exemplo bem simples de uma DLT, com 3 proposições nas linhas e 6 regras de consistência nas colunas. Para efeitos deste exemplo, assume-se que se uma pessoa está empregada ela deve trabalhar pelo menos algumas horas e deve receber algum salário.

Quadro 1 - Exemplo simples de DLT

	R1	R2	R3	R4	R5	R6
Está empregado	Y	Y		N		N
Horas trabalhadas > 0	N		Y		N	Y
Salário > 0		N	N	Y	Y	

Fonte: Os Autores.

A regra R1 estabelece que se a pessoa está empregada e não trabalha nenhuma hora, então há um conflito entre as duas proposições, e o registro falha. A regra R5 estabelece que se a pessoa trabalhou zero horas e recebeu um salário, também existe um conflito, e a pessoa falha na consistência.

As variáveis que aparecem nas proposições estão definidas no dicionário de dados. Estas variáveis podem ser únicas no registro (no caso das variáveis de domicílio, por exemplo), ou aparecer várias vezes (variáveis de pessoa), e neste caso diz-se que são variáveis de subnível. Para referenciar uma variável de subnível numa proposição

é necessário indexá-la com uma identificação do subnível a que se refere a variável (por exemplo, SEXO(2) significa o sexo da segunda pessoa). É possível fazer uma referência a uma subunidade (pessoa) genérica, usando os indexadores #1, #2, etc. Neste caso é necessário especificar na DLT a faixa de valores que o indexador pode receber.

Além da tabela propriamente dita, com as proposições e as regras de consistência, as DLTs têm vários parâmetros informados no cabeçalho, como seu nome, estrato, propósito, tipo, simetria e faixa de subnível. Um destes parâmetros que vale a pena mencionar é o de propósito da DLT, que determina se a DLT será usada para: a) Consistência; ou b) Seleção de Doadores. Consistência significa que as regras da tabela serão usadas para determinar se o registro passa ou falha. Seleção de Doadores significa que caso um registro não cumpra essa regra ele não será considerado como falho, mas sim como um registro atípico (outlier), e não será usado como doador. Regras de Consistência não são tão restritivas quanto as de Seleção de Doadores porque, para verificar se os registros podem ser usados como doadores, é preciso que estes registros tenham passado pelas regras de Consistência em primeiro lugar.

Canceis nas pesquisas do IBGE e no mundo

Canceis e o Censo Demográfico 2000

Na época da produção do Censo Demográfico 2000, o Canceis existia na sua versão NIM, e foi usado para a análise de consistência entre registros (relação entre as pessoas do mesmo domicílio) dos questionários da amostra. As variáveis envolvidas foram as seguintes: espécie do domicílio, sexo da pessoa, relação com o responsável pelo domicílio, relação com o responsável pela família, número da família, vive em companhia do cônjuge, natureza da última união, estado civil, faixa de idade e indicador de fecundidade.

Limitou-se a oito o número de estratos a serem trabalhados pelo NIM (desde domicílios com 1 pessoa, até domicílios com 8 pessoas). Os domicílios com mais de 8 pessoas foram tratados por um outro método, o Integrated Microcomputer Processing System - IMPS, precursor do Census and Survey Processing System - CSPro, ambos fornecidos pelo United States Census Bureau, dos Estados Unidos.

O processo consistiu na formatação do arquivo para extração das variáveis que seriam trabalhadas (somente aquelas envolvidas com a estrutura do domicílio). Em seguida, este arquivo era separado em 9 arquivos, um para cada estrato a ser tratado pelo NIM, e um único estrato com os domicílios com mais de 8 pessoas, a ser tratado pelo IMPS. Esses procedimentos foram realizados independentemente em cada um dos 67 lotes de questionários da amostra formados especificamente para essa etapa de crítica entre registros.

Deve ser destacado que para as críticas dentro do mesmo registro foi utilizado o sistema DIA, desenvolvido pelo INE da Espanha.

Canceis e o Censo Demográfico 2010

Aqui a utilização do Canceis foi bem mais extensa e interessante porque envolveu tanto o NIM como o Canceis propriamente dito. O NIM foi usado na análise da consistência entre registros do Universo (informações básicas para a totalidade da população recenseada) e da Amostra. E o Canceis foi usado na análise da consistência dentro de cada registro do Universo e da Amostra.

No caso do NIM, tal como no Censo Demográfico 2000, foram criados 9 estratos (para domicílios de 1 pessoa, 2 pessoas, etc., e 9 ou mais pessoas). O NIM cuidou dos oito primeiros estratos, e o último foi alocado ao CSPro.

No caso do Canceis não houve necessidade de criar estratos porque as consistências eram tratadas dentro de cada registro. Porém, dada a complexidade das mesmas, foram programadas várias aplicações separadas, uma para cada tema: domicílio, migração, educação, trabalho, deslocamento para trabalho e estudo, fecundidade e harmonização entre as classificações de ocupação e atividade referentes aos Censos Demográficos 2000 e 2010.

Demais pesquisas

O Canceis propriamente dito foi utilizado pela primeira vez, no IBGE, na imputação de algumas variáveis categóricas do Censo Agropecuário 2006, em oito temas distintos: dados gerais, lavoura permanente, lavoura temporária, floricultura, silvicultura, produtos da silvicultura, extração vegetal e indústria rural.

A segunda aplicação foi no módulo de Educação de Jovens e Adultos da PNAD 2007, onde os bons resultados obtidos e sua fácil implementação no processo de produção da pesquisa acarretaram na escolha da utilização deste software para a crítica e imputação da totalidade dos módulos a partir da PNAD 2008. O aplicativo continuou sendo utilizado até a última edição dessa pesquisa (PNAD 2015), como também desde a primeira edição da PNAD Contínua em 2012.

Adicionalmente, o Canceis foi utilizado pontualmente em outras pesquisas como na POF 2008-2009, no módulo referente ao questionário de características do domicílio e dos moradores, bem como para estudos metodológicos que, inclusive, não necessariamente são relacionados com o tema de crítica e imputação. Um exemplo foi o estudo sobre a avaliação do desempenho de um modelo de Censo Contínuo em estudo pelo IBGE. Neste caso o Canceis foi utilizado para achar doadores, a partir de dados da PME, que formariam uma população simulada de anos posteriores do Censo Contínuo.

O Canceis no Mundo

Além do Brasil e Canadá, país de origem do sistema, outros institutos nacionais de estatística possuem a licença para utilizá-lo. Em consulta feita ao StatCan em março de 2017, a lista de países era composta por Peru, Coréia do Sul, Rússia, Austrália, Suécia, Vietnã, Japão, Escócia, Inglaterra, Irlanda do Norte, Emirados Árabes Unidos (Emirado de Abu Dhabi).

Em adição aos institutos citados, outras agências também possuem licença para o uso do sistema. Temos como exemplo o Research Triangle Institute - RTI (sediado nos EUA, mas com sucursais em outros dez países), o Department for Communities and Local Government do Reino Unido e a Comissão Econômica das Nações Unidas para a Europa (United Nations Economic Commission for Europe - Unece).

Vantagens do Canceis

Em termos metodológicos, o Canceis é efetivamente bem mais avançado que os outros métodos citados: o Canceis, ao determinar que um registro (questionário, unidade de imputação) fere uma ou mais regras de consistência, ele busca um doador que seja um dos mais próximos (não é necessariamente o mais próximo, nem é necessariamente de um questionário anterior, já que ele procura também nos questionários doadores depois do questionário com erro, na sequência do arquivo), e que seja um dos mais próximos também na sua distância lógica, distância esta calculada baseando-se nas distâncias entre as variáveis de controle do doador e do receptor (distâncias estas que podem ser ponderadas). A escolha do doador logicamente mais próximo pode ser também parametrizada para escolher aleatoriamente um dos n possíveis doadores. Existem 13 alternativas (também parametrizadas) para as fórmulas de cálculo das distâncias, dependendo do tipo de variável (categorizada ou quantitativa ou mesmo alfanumérica) e de outras características de escolha do usuário.

O Canceis faz crítica e imputação, isto é, verifica a validade das respostas e a consistência entre elas, e se necessário, produz ações de imputação para corrigir os problemas. Para tanto, uma vantagem adicional do Canceis é que o usuário tem que definir somente as regras de invalidade e de inconsistência, não sendo necessário definir as regras de imputação, isto é, qual a maneira de corrigir o questionário, quais variáveis modificar, para quais valores, etc. O Canceis também tem um módulo de imputação determinística,

caso existam regras de imputação que precisem deste tipo de correção. E o Canceis escolhe um doador único para imputar todos os valores errôneos do questionário, evitando assim a formação de questionários-Frankstein (montados com partes de diversos questionários doadores). Esta é, do ponto de vista metodológico, a maior vantagem do Canceis.

Em resumo, o Canceis tem as seguintes vantagens metodológicas sobre outros métodos e sistemas de consistência e imputação:

- Uso de registros doadores, diminuindo consideravelmente a criação de registros não plausíveis;
- Uso de um doador único nas imputações de cada registro;
- Minimiza o número de ações de imputação para cada registro; e
- Não é necessária a definição de ações específicas de imputação, basta definir somente as regras de consistência.

Problemas e desvantagens

Embora em casos muito raros, nem sempre o método é capaz de encontrar um doador que solucione os erros existentes no registro que falhou. Esta condição é drástica porque exige um tratamento diferente (por outro aplicativo), provavelmente com outro tipo de programação, metodologia, e envolvendo mais recursos da instituição.

Além disso, existem outros casos em que um doador é encontrado, mas as distâncias calculadas são muito grandes, significando que o registro final imputado (combinação de dados entre doador e receptor) não seja plausível, poderia estar muito longe da realidade. Esta possibilidade cresce, no caso de crítica entre registros, para os estratos mais rarefeitos (domicílios com mais de 10 ou 11 pessoas, por exemplo). Uma possível solução para enfrentar esse problema seria aumentar (em número de casos) os estratos mais rarefeitos, combinando registros de todo o país, em vez de trabalhar por unidade da federação. É evidente que este tipo de solução tem outras desvantagens associadas, quais sejam a de ter que esperar que todo o país tenha sido processado na fase anterior para formar os estratos rarefeitos (problema de administração e execução do processo), e a desvantagem metodológica de usar doadores mais distantes (no sentido geográfico).

Também já foi mencionado, mas não custa repetir, que ainda que se juntem os lotes de processamento para os estratos de mais pessoas, existe um limite para este tipo de solução, e então, voltamos à necessidade de contar com um outro aplicativo para suprir esta limitação. Ou seja, é difícil pensar que o Canceis, por si só, possa resolver todos os problemas de consistência e imputação de dados.

É preciso tocar também numa questão que depende, em grande medida, do tipo de pesquisa e questionário que se vá trabalhar no Canceis: a organização dos temas de imputação e a escolha das variáveis para o cálculo das distâncias. Tomando-se como exemplo o questionário da amostra do Censo Demográfico 2000, dada sua grande quantidade de variáveis, não é razoável pensar que se possa imputar todo o questionário numa única aplicação (uma única execução do Canceis) optando-se, então, a trabalhar por temas.

Além das dificuldades devidas à quantidade e complexidade do questionário, comuns a todos os métodos usados para a avaliação de consistência, o Canceis traz uma desvantagem adicional, que é justamente devida à metodologia *per se,* resultado da separação inicial em registros doadores e receptores. Quanto maior o número de variáveis e o seu relacionamento, maior a possibilidade de encontrar inconsistências num registro, o que diminui proporcionalmente o universo de possíveis doadores, intensificando assim todos os problemas mencionados anteriormente. A solução, também óbvia, é a de separar a imputação em temas, mas isso traz como consequência a anulação de uma das vantagens principais do Canceis, que é a de usar um único doador para imputar o registro falhado: se o registro falhar em mais de um tema, ele será imputado separadamente em cada tema, provavelmente por doadores diferentes.

O futuro

O Canceis, atualmente, é o principal sistema de crítica e imputação para os levantamentos domiciliares do IBGE. Além da PNAD Contínua, que o utiliza atualmente, a expectativa é que tanto o Censo Demográfico quanto o Censo Agropecuário façam uso do mesmo em suas próximas edições.

Em paralelo estão sendo feitos estudos no sentido de aprimorar o atual uso do sistema, como o estudo de viabilidade da implementação da crítica entre registros na PNAD Contínua e a possível migração para a versão 5 do aplicativo (atualmente as pesquisas utilizam a versão 4.5). Especificamente com relação a este último tema, a migração ainda não foi recomendada uma vez que as versões mais recentes apresentam problemas de performance (tempo operacional) em grandes bases de dados, ou bases com elevado número de regras de crítica. Entretanto, a cada nova versão lançada (a atual é a 5.2) este processo será monitorado e novas avaliações serão feitas.

Referências

BANKIER, M. *Imputing numeric and qualitative census variables simultaneously*. Ottawa: Statistics Canada, 1994.

_____. *Experience with the new imputation methodology used in the 1996 canadian census with extensions for future censuses.* In: UN/ECE WORK SESSION ON STATISTICAL DATA EDITING, 1999, Rome. Working paper n. 24. Geneva: United Nations Economic Commission for Europe - Unece; New York: United Nations Statistical Commission, 1999. 9 p. Disponível em: <http://www3.istat.it/strumenti/metodi/software/MTSFload/TRATTAMENTOERRORIload/CANCEISload/24e.pdf>. Acesso em: set. 2017.

CORTEZ, B. F. Avaliação da substituição do software DIA pelo CANCEIS no processo de crítica e imputação das variáveis de fecundidade PNAD. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS 18., Águas de Lindóia, 2012. *Anais…* Aguás de Lindóia: Associação Brasileira de Estudos Populacionais - ABEP, 2012. p. 1-17. Seção temática 35. Disponível em: <http://www.abep.org.br/publicacoes/index.php/anais/article/view/2037/1995>. Acesso em: set. 2017.

CORTEZ, B. F.; FERNANDES, M. V. M. *Avaliação das mudanças operacionais necessárias para a migração dos procedimentos de crítica e imputação, com vistas à utilização do software Canceis versão 5.* Rio de Janeiro: IBGE, 2016. Não publicado.

CORTEZ, B. F.; MOREIRA, G. G.; FERNANDES, M. V. M. Descrição e avaliação do processo de imputação nos quesitos de fecundidade da PNAD 2008. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 17., Caxambu, 2010. *Anais…* Caxambu: Associação Brasileira de Estudos Populacionais - ABEP, 2010. p. 1-17. Seção temática 35. Disponível em: <http://www.abep.org.br/publicacoes/index.php/anais/article/view/2414/2368>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2000. Rio de Janeiro: IBGE, 2003. (Série relatórios metodológicos, 25). Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2000/metodologia/default.shtm>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2010. Rio de Janeiro: IBGE, 2013. 703 p. Acompanha 1 CD-ROM. (Série relatórios metodológicos, v. 41). Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/metodologia/default_metodologia.shtm>. Acesso em: set. 2017

SILVA, A. N. *Algumas considerações sobre o uso do NIM no censo demográfico 2000.* Rio de Janeiro: IBGE, 2003. 18 p. Não publicado.

_____. *Avaliação inicial do New Imputation Methodology System (NIM).* Rio de Janeiro: IBGE, 1998. 14 p. Não publicado.

_____. *Principais aspectos relacionados com o sistema CANCEIS.* Rio de Janeiro: IBGE, 2004. 13 p. Não publicado.

Amostra mestra nas pesquisas domiciliares do IBGE

Marcos Paulo Soares de Freitas

Introdução

O Instituto Brasileiro de Geografia e Estatística - IBGE, em 2002, intensificou a discussão interna sobre a realização de uma pesquisa sobre mercado de trabalho conjuntural em nível nacional, com início dos estudos de planejamento amostral da pesquisa que integraria os objetivos da Pesquisa Mensal de Emprego - PME e da Pesquisa Nacional por Amostra de Domicílios - PNAD.

Em 2004, decidiu-se ampliar essa integração, envolvendo todas as pesquisas domiciliares do Instituto, dando início às definições de um novo sistema de pesquisas domiciliares por amostragem. A integração de todas as pesquisas contempla o uso de uma infraestrutura amostral comum e aspectos relacionados a conceitos, a métodos, a procedimentos de trabalho de campo.

Para o Sistema Integrado de Pesquisas Domiciliares - SIPD foi definida a utilização de uma Amostra Mestra, que corresponde a um conjunto de unidades de área selecionadas de um cadastro único, segundo um método probabilístico de seleção, a partir da qual seja possível selecionar subamostras para atender às diversas pesquisas do sistema.

Como parte dos estudos para definição da Amostra Mestra, foi selecionada uma amostra de avaliação em 2007, baseada nos dados do Censo Demográfico 2000, da qual foram selecionadas subamostras para a Pesquisa de Orçamentos Familiares - POF 2008-2009 e para o teste da PNAD Contínua 2009. Após as avaliações realizadas com os resultados dos testes, alguns ajustes foram introduzidos nas definições da Amostra Mestra, e a amostra inicial para implantação do sistema foi selecionada em 2011, utilizando como base o Censo Demográfico 2010.

Este capítulo trata dos aspectos de amostragem considerados no desenvolvimento e definição da Amostra Mestra 2010: o plano amostral, a abrangência geográfica, a população-alvo, o tamanho e a rotação da amostra.

Plano amostral

Como parte da definição de uma Amostra Mestra, as unidades selecionadas para compor a amostra constituem as Unidades Primárias de

Amostragem - UPAs nos planejamentos amostrais das pesquisas do SIPD, que em geral são planos por conglomerados em diversos estágios. Por isso foi necessário considerar aspectos comuns desses planejamentos, como estratificação e seleção com probabilidades desiguais, para a determinação do plano amostral da Amostra Mestra.

Desta forma, a Amostra Mestra é composta por um conjunto de UPAs, que são estratificadas e selecionadas com probabilidade proporcional ao tamanho, medido pelo número de Domicílios Particulares Permanentes - DPPs, ocupados, fechados e vagos.

Abrangência geográfica e população alvo

A definição da abrangência geográfica e da população alvo da Amostra Mestra considerou as pesquisas domiciliares que integrarão o SIPD, para evitar a exclusão de parte do território ou das pessoas que compõem o âmbito de alguma destas pesquisas.

A abrangência geográfica foi definida como sendo todo o território nacional, dividido nos setores censitários da Base Operacional Geográfica[1], excluídas áreas com características especiais classificadas pelo IBGE como setores censitários de: aldeias indígenas, quartéis, bases militares, alojamentos, acampamentos, embarcações, penitenciárias, colônias penais, presídios, cadeias, asilos, orfanatos, conventos, hospitais e agrovilas de projetos de assentamentos rurais. Para a Amostra Mestra 2010, também foram excluídos os setores censitários localizados em Terras Indígenas.

E a população alvo é constituída por todas as pessoas moradoras em domicílios particulares permanentes da área de abrangência.

Cadastro de seleção

O cadastro básico tem como unidades os setores censitários, contendo informações sobre a dependência administrativa, provenientes da Base Operacional Geográfica, e características sociodemográficas, oriundas dos dados do Censo Demográfico 2010, como total de pessoas, rendimento médio, total de pessoas desocupadas entre outras.

A partir do cadastro básico foi construído o cadastro para seleção da Amostra Mestra, composto por unidades de área que podem ser diferentes das unidades básicas, ou setores censitários. A definição de unidades diferentes de setores foi necessária, pois as UPAs serão utilizadas na amostra de todas as pesquisas do SIPD, portanto é desejável que estas unidades tenham um tamanho mínimo, não muito pequeno, definido em termos de domicílios.

Como na Base Operacional Geográfica há muitos setores censitários pequenos, foi preciso realizar uma agregação de setores para composição das UPAs, de tal modo que estas possuam no mínimo 60 domicílios particulares permanentes ocupados, fechados ou vagos, número julgado suficiente para atender a demanda das pesquisas do SIPD.

Com isso, o cadastro de seleção da Amostra Mestra é composto por setores censitários ou conjuntos de setores censitários, que foram agregados com o objetivo de maximizar o número de conjuntos, juntando o mínimo possível de setores, tendo como restrições a contiguidade, o tamanho mínimo de 60 DPPs e algumas características dos setores: tipo, situação e divisão administrativa (subdistrito).

Estratificação

A estratificação das UPAs da Amostra Mestra foi definida levando-se em consideração os objetivos das diversas pesquisas que serão contempladas por esta amostra e também as questões operacionais e domínios de divulgação.

---

[1] As embaixadas, consulados e representações do Brasil no exterior são considerados território nacional, porém não são incluídos na Base Operacional Geográfica.

Estratificação por divisão administrativa

Na primeira etapa de estratificação, foram consideradas como estratos as Unidades da Federação - UFs, e outras divisões administrativas que são ou poderão ser domínios de divulgação, para garantir que sejam contempladas na amostra.

Em cada UF, os municípios, e todos as suas UPAs, foram classificados em cinco grupos:

(1) Capital;

(2) Demais municípios pertencentes à Região Metropolitana[2] - RM ou à Região Integrada de Desenvolvimento - RIDE;

(3) Municípios pertencentes a colar ou expansão metropolitana ou a outra RM;

(4) Municípios pertencentes à RIDE com sede em outra UF; e

(5) Demais municípios da UF.

Os grupos (2), (3) e (4) só existem nas UFs que possuem a divisão administrativa correspondente considerada.

Estratificação geográfica e espacial

Com a estratificação geográfica e espacial, buscou-se a garantia de um espalhamento da amostra no território, para captar características de áreas diferentes dentro dos estratos anteriores.

Nesta etapa, os municípios, e todas as suas UPAs, foram agrupados em estratos de municípios semelhantes, usando informações de meso e microrregiões e demais conhecimentos derivados de outras pesquisas da Coordenação de Geografia, da Diretoria de Geociências. E no grupo (1), as UPAs foram agrupadas diretamente, considerando as divisões internas do município, como distrito, subdistrito e bairro.

A introdução desta etapa de estratificação foi realizada para que seja possível produzir resultados para outros grupos de municípios que não os definidos como domínios de estimação no planejamento da amostra, onde os tamanhos de amostra foram controlados, desde que avaliada a precisão das estimativas ou utilizados métodos apropriados para estimação com amostras pequenas.

Estratificação por situação dos domicílios da unidade primária de amostragem

Como o IBGE divulga resultados de algumas de suas pesquisas domiciliares separados para a área urbana e para a área rural, as UPAs foram estratificadas considerando a situação dos seus domicílios, urbana e rural, para garantir a seleção de amostra em cada uma destas áreas, com controle da precisão das estimativas.

Esta estratificação foi feita dentro de cada estrato geográfico e espacial definidos na etapa anterior.

Estratificação estatística

As etapas anteriores tiveram como objetivos garantir o espalhamento da amostra no território, permitir o controle da seleção e do tamanho de amostra para possíveis divulgações de resultados. Esta última etapa, estatística, foi feita com o intuito de melhorar a precisão das estimativas obtidas com os dados das pesquisas.

A estratificação estatística considerou informações disponíveis para todas as UPAs, formando estratos homogêneos segundo estas informações: rendimento total dos domicílios e total de DPPs.

O método utilizado consiste em classificar as UPAs em grupos de tal forma que a variância do estimador do total de uma característica de interesse seja mínima, considerando o plano amostral das pesquisas domiciliares: amostragem conglomerada com seleção das UPAs com probabilidade proporcional a uma medida de tamanho (número de DPPs).

---

[2] Foram consideradas apenas as Regiões Metropolitanas que contêm o município da capital.

Para implementação do método, foi calculada uma medida de distância (dissimilaridade) para cada par de UPAs da seguinte maneira:

$$d(i,i') = N_i \cdot N_{i'} \cdot \left(\frac{Y_i}{N_i} - \frac{Y_{i'}}{N_{i'}} \right)^2$$

onde

N_i é o número de domicílios na UPA *i*;
$N_{i'}$ é o número de domicílios na UPA *i'*;
Y_i é o total da característica de interesse *Y* na UPA *i*;
$Y_{i'}$ é o total da característica de interesse *y* na UPA *i'*; e
y é o rendimento total dos domicílios.

Definiu-se em 2 ou 3 o número de estratos estatísticos formados em cada estrato de situação e como 150 UPAs o tamanho mínimo dos estratos. A implementação deste método foi feita utilizando algoritmos de otimização desenvolvidos por Montenegro e Brito (2006).

Tamanho da Amostra Mestra

Para determinação do tamanho da amostra foi preciso definir a pesquisa que demanda um tamanho de amostra maior, pois a Amostra Mestra servirá como base para todas as pesquisas do SIPD. Das pesquisas que já estavam definidas no sistema, a PNAD Contínua é a pesquisa que necessita de amostra maior, pois produz resultados trimestrais sobre mercado de trabalho e anualmente para outros temas sociodemográficos, para níveis geográficos bem desagregados.

Optou-se por calcular o tamanho de amostra como o tamanho necessário para estimar o total de pessoas desocupadas de 14 anos ou mais de idade no trimestre, que é um dos principais indicadores da PNAD Contínua, com um nível de precisão pré-determinado.

As fórmulas utilizadas são as que se seguem:

Total de interesse

$$Y = \sum_{h=1}^{L} \sum_{i=1}^{M_h} \sum_{j=1}^{N_{hi}} y_{hij} = \sum_{h=1}^{L} Y_h$$

onde

h é o índice do estrato a que pertence a UPA;
i é o índice da UPA dentro do estrato;
j é o índice do domicílio dentro da UPA;
L é o número total de estratos;
M_h é o número de UPAs na população do estrato *h*;
N_{hi} é o número de domicílios na população da UPA *i*, do estrato *h*;
Y_{hij} é o valor da variável *y* no domicílio *j* da UPA *i* do estrato *h*; e
Y_h é o total da variável *y* no estrato *h*.

E o estimador para o total, considerando o plano amostral conglomerado em dois estágios, com estratificação das UPAs, seleção das UPAs com probabilidade proporcional ao tamanho e seleção das unidades secundárias de amostragem (USAs), os domicílios, com probabilidade igual e número fixo de unidades, é dado por Kish (1965):

$$\hat{Y} = \sum_{h=1}^{L} \frac{1}{m_h} \sum_{i=1}^{m_h} \frac{\hat{Y}_{hi}}{p_{hi}}$$

e a variância deste estimador pode ser escrita como

$$V(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{m_h} \cdot \left[\sum_{i=1}^{M_h} \frac{Y_{hi}^2}{p_{hi}} - Y_h^2 + \sum_{i=1}^{M_h} \frac{N_{hi}^2}{p_{hi}} \cdot \frac{S_{yhi}^2}{n_{hi}} \cdot \frac{N_{hi} - n_{hi}}{N_{hi}} \right]$$

onde

m_h é o número de UPAs na amostra do estrato *h*;

$p_{hi} = \frac{N_{hi}}{N_h}$ é o tamanho relativo da UPA *i*, do estrato *h*, que define a probabilidade de seleção desta UPA em um sorteio com PPT com reposição;

N_h é o número de domicílios na população do estrato *h*,

$\hat{Y}_{hi} = \sum_{j=1}^{n_{hi}} w_{j|hi} \cdot y_{hij}$ é o estimador simples do total $Y_{hi} = \sum_{j=1}^{N_{hi}} y_{hij}$ da variável *y* na UPA *i* do estrato *h*;

n_{hi} é o número de domicílios na amostra da UPA *i*, do estrato *h*;

$w_{j|hi} = \frac{N_{hi}}{n_{hi}}$ é o peso do domicílio *j* dado a seleção da UPA *i* do estrato *h* ; e

$S_{yhi}^2 = \frac{1}{N_{hi} - 1} \cdot \sum_{j=1}^{N_{hi}} \left(y_{hij} - \bar{Y}_{hi} \right)^2$ é a variância da variável de interesse y na população na UPA i do estrato h .

$\bar{Y}_{hi}$ é a média da variável de interesse y na população na UPA i do estrato h .

Como a característica y , número de desocupados, foi investigada no Censo Demográfico 2010 apenas no questionário da amostra, foi necessário derivar o seguinte estimador não viciado para a variância do total, conforme Lima e Bianchini (1988):

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{m_h} \cdot \left[\sum_{i=1}^{M_h} \frac{\tilde{Y}_{hi}^2}{p_{hi}} - \tilde{Y}_h^2 + \sum_{i=1}^{M_h} \frac{N_{hi}^2}{p_{hi}} \cdot \frac{s_{yhi}^2}{n_{hi}} \cdot \frac{N_{hi} - n_{hi}}{N_{hi}} \right]$$

onde

$\tilde{Y}_{hi}^2 = N_{hi}^2 \cdot \left(\bar{y}_{hi}^2 - \frac{N_{hi} - n_{hi}^*}{N_{hi}} \cdot \frac{s_{yhi}^2}{n_{hi}^*} \right)$ é um estimador não viciado para Y_{hi}^2 ;

$\tilde{Y}_h^2 = \left(\sum_{i=1}^{M_h} N_{hi} \cdot \bar{y}_{hi} \right)^2 - \sum_{i=1}^{M_h} \left(N_{hi}^2 \cdot \frac{N_{hi} - n_{hi}^*}{N_{hi}} \cdot \frac{s_{yhi}^2}{n_{hi}^*} \right)$ é um estimador não viciado para Y_h^2;

$s_{yhi}^2 = \frac{1}{n_{hi}^* - 1} \cdot \sum_{j=1}^{n_{hi}^*} \left(y_{hij} - \bar{y}_{hi} \right)^2$; e

n^*_{hi} é o número de domicílios na amostra da UPA *i*, do estrato *h* no Censo Demográfico 2010.

E o cálculo do número de UPAs na amostra, necessário para estimar o total de interesse com a precisão pré-definida, foi feito considerando como fixo o número de domicílios a serem selecionados por UPA ($\bar{n}$), com a expressão a seguir:

$$m = \frac{N \cdot \sum_{h=1}^{H} \sum_{i=1}^{M_h} \left\{ N_{hi} \cdot \left(\bar{y}^2_{hi} - \bar{y}^2_h \right) - \left[\left(\frac{N_{hi} - n^*_{hi}}{n^*_{hi}} \right) \cdot \left(1 - \frac{N_{hi}}{N_h} \right) - \left(\frac{N_{hi} - \bar{n}}{\bar{n}} \right) \right] \cdot s^2_{yhi} \right\}}{\left(CV(\hat{Y}) \cdot \hat{Y} \right)^2}$$

onde

$CV(\hat{Y})$ é o coeficiente de variação desejado para a estimativa de interesse e

$\hat{Y}$ é a estimativa do total da característica de interesse, proveniente do Censo Demográfico 2010.

Levando em consideração os níveis de precisão para a estimativa do total de desocupados obtidos com as PNADs 2001 a 2009, definiu-se, para cada UF, o CV esperado e optou-se por selecionar 14 domicílios por UPA na PNAD Contínua.

Alocação do tamanho da amostra

A alocação do tamanho da amostra pelos estratos dentro de cada UF levou em consideração as características da amostra da PNAD Contínua e avaliações da precisão das estimativas em domínios menores que UF.

A alocação foi feita em duas etapas. Na primeira etapa, o tamanho da amostra da UF foi alocado proporcionalmente ao número de UPAs em dois grupos: um composto pelos estratos da capital e da RM, e o outro pelos demais estratos, através da expressão

$m_g = m \cdot \frac{M_g}{M}$, onde

M_g é o número de UPAs na população do grupo *g* e

M é o número de UPAs na população da UF.

Em uma segunda etapa, o tamanho da amostra do grupo dos demais estratos também foi alocado proporcionalmente ao número de UPAs nos estratos finais. E no grupo dos estratos da capital e RM, a alocação foi proporcional ao número de DPPOs. Foram utilizadas as seguintes expressões

$m_h = m_g \cdot \frac{M_h}{M_g}$ e $m_h = m_g \cdot \frac{N_h}{N_g}$, respectivamente.

Após a alocação, os tamanhos foram ajustados para que fossem no mínimo 15 UPAs nos estratos finais, devido ao esquema de rotação da amostra da PNAD Contínua, e para que fossem múltiplos de 12 UPAs nos estratos geográficos, pela distribuição de coleta nas semanas do trimestre. Por fim, pela avaliação da precisão da estimativa de pessoas desocupadas de 14 anos ou mais de idade em cada capital, em algumas delas os tamanhos tiveram que ser ajustados para que o CV esperado fosse menor que 15%.

Na Tabela 1, são apresentados os tamanhos finais de amostra, assim como os coeficientes de variação esperados para a estimativa de interesse.

Tabela 1 - Tamanho da Amostra Mestra e da PNAD Contínua necessário para estimar, trimestralmente, o total de pessoas desocupadas de 14 anos ou mais de idade e o respectivo coeficiente de variação esperado por situação do domicílio, segundo o Total Brasil, Grandes Regiões e Unidades da Federação

Nível Geográfico	Situação do domicílio								
	Total			Urbano			Rural		
	UPAs na amostra	Domicílios na amostra	CV (%)	UPAs na amostra	Domicílios na amostra	CV (%)	UPAs na amostra	Domicílios na amostra	CV (%)
Brasil	**15 096**	**211 344**	**1,3**	**11 187**	**156 618**	**1,3**	**3 909**	**54 726**	**3,8**
Norte	**1 896**	**26 544**	**3,4**	**1 319**	**18 466**	**3,4**	**577**	**8 078**	**13,8**
11 – RO	264	3 696	8,0	177	2 478	8,3	87	1 218	26,2
12 – AC	276	3 864	9,4	191	2 674	10,0	85	1 190	25,5
13 – AM	360	5 040	6,4	260	3 640	6,5	100	1 400	29,8
14 – RR	156	2 184	9,9	116	1 624	9,8	40	560	42,4
15 – PA	504	7 056	5,9	321	4 494	5,8	183	2 562	18,3
16 – AP	108	1 512	8,8	93	1 302	8,9	15	210	53,5
17 - TO	228	3 192	**7,9**	161	2 254	8,2	67	938	27,4
Nordeste	**4 908**	**68 712**	**1,9**	**3 137**	**43 918**	**2,0**	**1 771**	**24 794**	**5,0**
21 – MA	900	12 600	4,5	463	6 482	5,0	437	6 118	10,0
22 – PI	324	4 536	7,1	191	2 674	7,6	133	1 862	17,5
23 – CE	780	10 920	4,4	549	7 686	4,7	231	3 234	12,2
24 – RN	300	4 200	5,8	224	3 136	6,3	76	1 064	15,0
25 – PB	384	5 376	5,9	262	3 668	6,4	122	1 708	14,0
26 – PE	600	8 400	4,4	421	5 894	4,7	179	2 506	13,1
27 - AL	564	7 896	4,5	374	5 236	4,9	190	2 660	12,5
28 – SE	288	4 032	6,1	189	2 646	6,6	99	1 386	16,6
29 – BA	768	10 752	4,2	464	6 496	4,5	304	4 256	11,5
Sudeste	**4 092**	**57 288**	**2,3**	**3 389**	**47 446**	2,4	**703**	**9 842**	**7,7**
31 – MG	1 104	15 456	3,8	798	11 172	3,9	306	4 284	14,1
32 – ES	600	8 400	4,4	493	6 902	4,5	107	1 498	18,0
33 – RJ	1 164	16 296	3,7	1 088	15 232	3,8	76	1 064	15,7
35 – SP	1 224	17 136	3,8	1 010	14 140	3,9	214	2 996	11,2
Sul	**2 664**	**37 296**	**2,6**	**2 078**	**29 092**	2,7	**586**	**8 204**	**9,6**
41 – PR	828	11 592	4,3	641	8 974	4,5	187	2 618	14,9
42 – SC	948	13 272	4,5	757	10 598	4,7	191	2 674	16,0
43 – RS	888	12 432	4,4	680	9 520	4,5	208	2 912	17,5
Centro-Oeste	**1 536**	**21 504**	**3,1**	**1 264**	**17 696**	3,1	**272**	**3 808**	**13,7**
50 – MS	336	4 704	6,1	270	3 780	6,3	66	924	25,1
51 – MT	396	5 544	6,1	311	4 354	6,3	85	1 190	24,6
52 – GO	528	7 392	5,4	426	5 964	5,5	102	1 428	23,1
53 – DF	276	3 864	6,2	257	3 598	6,3	19	266	28,1

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Métodos e Qualidade.

Seleção da Amostra Mestra

Para a seleção da amostra optou-se por Amostragem de Pareto PPT (COSTA, 2007), que combina a técnica de números aleatórios permanentes com o tamanho relativo da UPA no grupo de rotação (número de domicílios da UPA dividido pelo número de domicílios no grupo de rotação).

Como a PNAD Contínua possui um esquema de rotação da amostra de domicílios, foi necessário dividir a população e, consequentemente, a amostra, em grupos de rotação, para facilitar a operacionalização desse esquema. Assim, a amostra foi selecionada dentro de cada um desses grupos, seguindo os seguintes passos:

1 – Adicionar a informação do tamanho de amostra por estrato final ao cadastro de UPAs;
2 – Contar o número de UPAs por grupo;
3 – Dividir o tamanho da amostra de cada estrato final pelo número de grupos daquele estrato, ou seja, calcular o tamanho da amostra em cada grupo. A princípio a amostra de todos os grupos é igual, mas no caso da divisão não ser exata, os grupos com mais setores serão também os com maior tamanho de amostra;
4 – Sortear um número aleatório, denominado ALEAT, entre 0 e 1 para cada UPA. Este número será permanentemente atrelado à UPA;
5 – Definir a variável de tamanho, número de Domicílios Particulares Permanentes Ocupados - DPPO, de cada UPA, limitando-a entre 50 e 500 para evitar probabilidades muito baixas ou muito altas de seleção;
6 – Calcular a proporção, p_i, de DPPO da UPA i dentro do grupo;
7 – Calcular o valor de Q_i de cada UPA de acordo com a fórmula:

$$Q_i = \frac{ALEAT - ALEAT \times p_i}{p_i - ALEAT \times p_i}$$

8 – Ordenar as UPAs dentro de cada grupo de maneira crescente pelo Q_i e selecionar para a Amostra Mestra as primeiras de cada grupo, de acordo com os tamanhos previamente calculados;
9 – Substituir uma UPA selecionada se esta tiver uma quantidade de DPPO menor que 40, pela UPA seguinte na ordem de seleção.

Rotação da Amostra Mestra

O método escolhido de seleção da amostra permite a incorporação de atualizações no cadastro de seleção, acompanhando a evolução do crescimento das UPAs e mudanças na Base Operacional Geográfica, além de permitir a renovação controlada da amostra.

A cada trimestre serão trocadas no máximo 2,5% das UPAs, o que resultará em uma substituição quase completa da amostra em 10 anos, quando é prevista uma avaliação mais detalhada do plano da Amosta Mestra.

Desta forma, a Amostra Mestra para um determinado ano é composta pelas UPAs selecionadas para o primeiro trimestre mais as UPAs selecionadas para entrarem nos 3 trimestres seguintes (3*2,5% = 7,5%). As UPAs de um trimestre compõem a amostra da PNAD Contínua daquele trimestre, e a subamostra de uma outra pesquisa do SIPD será selecionada da amostra do trimestre mais próximo do prazo final de definição do planejamento da pesquisa. Todas as UPAs da amostra das pesquisas do sistema já terão sido alvo de investigação na PNAD Contínua.

Optou-se por levantar as atualizações continuamente, seja de quantitativo de domicílios, seja da Base Operacional Geográfica, consolidá-las uma vez por ano, incorporando-as ao cadastro, e selecionar em janeiro a Amostra Mestra correspondente aos 2 últimos trimestres do ano e os 2 primeiros trimestres do ano seguinte.

Em decorrência da incorporação das alterações da Base Operacional Geográfica, a rotação da amostra de UPAs por trimestre passou a ser maior que 2,5%, pois em grupos onde as UPAs permaneceriam na amostra apenas trocando domicílios (17,5%), foi preciso trocar também algumas UPAs, devido a mudanças na formação dos setores que as compunham.

Conclusões

O desenho amostral da Amostra Mestra visa atender várias pesquisas, buscando satisfazer os seus diversos objetivos, e com isso é de se esperar que o desenho amostral não seja o melhor para cada pesquisa isoladamente, mas seja bom para todas.

O SIPD começou a ser implantado com a realização da PNAD Contínua a partir do quarto trimestre de 2011 e outras duas pesquisas já tiveram sua amostra selecionada: Pesquisa Nacional de Saúde - PNS 2013 e POF prevista para ir a campo em 2017.

Apesar da implementação do sistema já estar em andamento, alguns aspectos ainda precisam ser definidos ou consolidados: avaliação da incorporação das atualizações da Base Operacional Geográfica, consolidação da criação do código único para os domicílios, integração dos programas de seleção, de rotação, de substituição da amostra, consolidação dos procedimentos do cálculo dos fatores de expansão, entre outros.

Referências

ANTONACI, G.; SILVA, D. B. N. Emparelhamento de domicílios e pessoas na pesquisa mensal de emprego e cálculo da autocorrelação da característica desocupação. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 17., 2006, Caxambu. *Anais...* São Paulo: Associação Brasileira de Estatística - ABE, 2006.

ASSUNÇÃO, R. M. *Análise de conglomerados espaciais para uso em amostragem estratificada com probabilidade de seleção proporcional ao tamanho*: relatório final BRA/97/013. Brasília, DF: Instituto de Pesquisa Econômica Aplicada - IPEA, 2000. Projeto Rede de Pesquisas e Desenvolvimento de Políticas Econômicas, Sistema Rede IPEA. Projeto 6: Desenvolvimento e absorção de novas tecnologias de produção de informações.

BIANCHINI, Z. M.; ALBIERI, S. E. *Principais aspectos da amostragem da pesquisas domiciliares do IBGE*: revisão 2002. Rio de Janeiro: IBGE, 2003. 27 p. (Textos para discussão. Diretoria de Pesquisas, n. 8). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv66373.pdf>. Acesso em: set. 2017.

BUSSAB, W. O.; DINI, N. P. Pesquisa de emprego e desemprego SEADE/DIEESE: regiões homogêneas da Grande São Paulo. *Revista da Fundação SEADE*, São Paulo: Fundação Sistema Estadual de Análise de Dados - Seade, v. 1, n.3, p. 5-11, set./dez., 1985.

CARRILHO, A.; NTHABISENG, M. *Sampling, weighting and standard error estimation methodology for the labour force survey conducted in september 2001*. Trabalho apresentado no Workshop on Survey Sample Designs, realizado em Windhoek, Namíbia, 2002.

COCHRAN, W. G. *Sampling techniques*. 3rd ed. New York: Wiley, c1977. 428 p. (Wiley series in probability and mathematical statistics). Disponível em: <https://archive.org/details/Cochran1977SamplingTechniques_201703>. Acesso em: set. 2017.

COSTA, G.T. L. da. *Coordenação de amostras PPT em pesquisas repetidas, utilizando o método de amostragem de Pareto*. 2007. 94 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro. 2007.

FREITAS, M. P. S.; LILA, M. F. *Uma proposta de dimensionamento de amostra para a pesquisa domiciliar contínua*: versão preliminar. Rio de Janeiro: IBGE, 2004. Não publicado.

FREITAS, M. P. S. *Estratificação para a amostra de uma pesquisa domiciliar sobre mercado de trabalho*. 2002. 98 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro. 2002.

KISH, L. *Survey sampling*. New York: Wiley, [1965]. 643 p.

LILA, M. F.; FREITAS, M. P. S. *Estimação de intervalos de confiança para estimadores de diferenças temporais na pesquisa mensal de emprego*. Rio de Janeiro: IBGE, 2007. 101 p. (Textos para discussão. Diretoria de Pesquisas, n. 22). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv35915.pdf>. Acesso em: set. 2017.

FREITAS, M. P. S. de et al. *Amostra mestra para o sistema integrado de pesquisas domiciliares*. Rio de Janeiro: IBGE, 2007. 67 p. (Textos para discussão. Diretoria de Pesquisas, n. 23). Disponível em: <http://www.ibge.gov.br/home/estatistica/indicadores/sipd/texto_discussao_23.pdf>. Acesso em: set. 2017.

FREITAS, M. P. S. de; ANTONACI, G .A. *Sistema integrado de pesquisas domiciliares*: amostra mestra 2010 e amostra da PNAD contínua. Rio de Janeiro: IBGE, 2014. 36 p. (Textos para discussão. Diretoria de Pesquisas, n. 50). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv86747.pdf>. Acesso em: set. 2017.

LIMA, M. I. F.; BIANCHINI, Z. M. Estudos para o dimensionamento da amostra da pesquisa de orçamentos familiares. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 7., 1986, Campinas. *Resumos...* São Paulo: Associação Brasileira de Estatística - ABE, 1988.

MONTENEGRO, F. M. T.; BRITO, J. A. de M. Um algoritmo genético para o problema de agrupamento. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 38., 2006, Goiânia. *Anais…* Rio de Janeiro: Sociedade Brasileira de Pesquisa Operacional - Sobrapo, 2006. p. 1471-1480. Disponível em: <http://din.uem.br/sbpo/sbpo2006/pdf/arq0196.pdf>. Acesso em: set. 2017.

PETTERSSON, H. Designg of master sampling frames and master samples for household surveys. In: HOUSEHOLD sample surveys in developing and transition countries . New York: United Nations, Department of Economic and Social Affairs, 2005. p. 71-94. (Studies in methods. Series F, n. 96).Disponível em: <https://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf>. Acesso em: set. 2017.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. H. *Model assisted survey sampling*. New York: Spring-Verlag, 1992. 694 p. (Springer series in survey statistics).

SILVA, P. L. N. *Algumas idéias para a revisão das pesquisas domiciliares por amostragem do IBGE*. Rio de Janeiro: IBGE, 2001. Não publicado.

SILVA, P. L. N. et al. *Aspectos sobre a estrutura longitudinal no contexto da pesquisa sobre mercado de trabalho*. Rio de Janeiro: IBGE, 1998. Não publicado.

TURNER, A. G. *Sampling frames and master sample*. New York: United Nations, Statistical Division, 2003. 26 p. Disponível em: <https://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_3.pdf>. Acesso em: set. 2017.

YANSANEH, I. S.; FULLER, W. A. Optimal recursive estimation for repeated surveys. *Survey Methodology*, Ottawa: Statistics Canada - StatCan, v. 24, n. 1, p. 31-40, June 1998. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/1998001/article/3907-eng.pdf>. Acesso em: set. 2017.

Evolução dos aspectos metodológicos na investigação por amostragem dos Censos Demográficos

Sonia Albieri

Introdução

A discussão sobre o uso de amostragem na coleta de informações em Censos Demográficos passa sempre por duas grandes definições: o plano amostral e a fração (ou frações) de amostragem. A pesquisa por amostragem probabilística na coleta dos Censos Demográficos no Brasil teve início em 1960 e foi realizada, desde então, em todos os Censos decenais. Apenas as duas operações de Contagem da População 1996 e 2007 não fizeram uso de amostragem na coleta das informações, devido ao número restrito de variáveis que investigaram, em função do próprio objetivo desse tipo de operação ser apenas a contagem da população.

É esse procedimento que vem permitindo, quando se realiza o Censo, a ampliação e o aprofundamento dos temas abordados para obtenção de informações mais detalhadas sobre as condições de vida da população nos municípios e localidades, sem aumentar os custos de forma insustentável ou assoberbar o trabalho de coleta de dados a ponto de colocar em risco a qualidade das informações.

Desta forma, durante a coleta do Censo Demográfico 2010, tal como nos Censos anteriores, foram usados dois modelos de questionário, sendo, em cada domicílio, aplicado somente um dos modelos. Um deles, denominado Questionário Básico, é simplificado; o outro, bem mais extenso e detalhado, o Questionário da Amostra, foi aplicado em domicílios selecionados através de amostragem probabilística. Este último contém todas as perguntas do Questionário Básico, e mais um conjunto de quesitos sobre temas como educação, religião, deficiência, migração, fecundidade, trabalho e rendimento, entre outros.

Assim, todas as perguntas do Questionário Básico também estão contidas no Questionário da Amostra, de forma que essas perguntas comuns foram aplicadas a todos os domicílios e pessoas moradoras (residentes) no País. Já os quesitos específicos do Questionário da Amostra foram aplicados apenas nos domicílios selecionados para amostra e seus respectivos moradores.

O conjunto de informações comuns aos dois questionários, o Básico e o da Amostra, constitui as informações básicas censitárias obtidas para 100% da população, o que se convencionou chamar de Conjunto Universo.

Este capítulo descreve, de forma resumida, os aspectos relacionados com o uso da tecnologia de amostragem para essa investigação. Os detalhes metodológicos em geral e específicos de amostragem das operações censitárias podem ser encontrados nas publicações relacionadas na lista de **Referências** ao final do documento.

Coleta de dados

Para a realização dos Censos, o território nacional é dividido em partições geográficas denominadas setores censitários, de tal forma que seus limites respeitam as divisões internas dos municípios em zonas urbanas e rurais, e em distritos e subdistritos, caso existam. O setor censitário foi planejado de forma a que um entrevistador consiga realizar a operação de coleta no período de realização do Censo. Em 1991, foram definidos 163 266 setores censitários. Em 2000, eram 215 811 setores e em 2010 foram definidos 310 120 setores com população.

Até o Censo Demográfico 2000, os domicílios particulares foram selecionados por amostragem sistemática em cada setor censitário. As famílias ou pessoas sós moradoras em domicílios coletivos (alojamentos estudantis, quartéis, prisões, hospitais, orfanatos, conventos, etc.) foram selecionadas, também de forma sistemática, independentemente da seleção de domicílios particulares, usando a mesma fração amostral definida para o setor a que pertence cada domicílio coletivo.

No Censo Demográfico 2010, a coleta de dados foi realizada usando computadores de mão, o que facilitou a implementação de novo esquema de seleção de domicílios para a amostra, por meio de um programa de seleção aleatória por grupo de domicílios, de acordo com a fração amostral definida para o setor.

Assim, por exemplo, em um setor com fração amostral de 20%, em vez de selecionar 1 em cada 5 domicílios, o algoritmo prevê a seleção de 2 em cada 10, aleatoriamente no grupo formado. Dessa forma, não existe um salto constante entre os números de ordem de unidades selecionadas em sequência, dificultando a introdução de vícios operacionais por parte dos entrevistadores.

Plano amostral

O plano amostral adotado em cada Censo foi o mesmo definido em 1960, a saber: seleção de domicílios particulares e de unidades domiciliares em domicílios coletivos, independentemente em cada setor censitário, o que equivale a um plano de amostragem estratificada, onde os estratos são os setores, e a seleção de domicílios é feita com equiprobabilidade em cada estrato, sendo que a investigação é feita para todas as pessoas moradoras nos domicílios selecionados para a amostra.

Os estudos realizados sobre o plano amostral utilizado nos Censos Demográficos estão indicados nas **Referências** ao final deste capítulo, com destaque para o trabalho de Albieri e Bianchini (2015) e as publicações do IBGE *Metodologia do censo demográfico 2000* (2003) e *Metodologia do censo demográfico 2010* (2016), de onde foram extraídas muitas das informações para elaboração deste texto.

Uma vantagem desse plano amostral, motivo pelo qual não foi alterado ao longo do tempo, é que permite atender à demanda por informações em níveis geográficos formado por agregados de poucos setores (estratos) e à necessidade de informações da amostra em todos os setores, para o planejamento de amostras das pesquisas domiciliares no período intercensitário.

Tamanho da amostra

O tamanho da amostra decorre da fração amostral definida para a coleta de dados do Questionário da Amostra. Nos Censos Demográficos 1960, 1970 e 1980, foi utilizada uma única fração amostral de 25% dos domicílios. Durante o planejamento do Censo Demográfico 1991, o assunto foi tratado no âmbito de uma comissão interna, criada especialmente para esse fim, que encomendou estudos específicos visando avaliar o efeito da redução da fração amostral que vinha sendo adotada nos Censos anteriores (25%). A conclusão dos estudos apontou alternativas que levaram à definição das frações amostrais usadas em 1991, com o emprego de duas frações amostrais diferentes de acordo com o tamanho do município, medido em função da projeção de população para a data de referência do Censo: 20% para os municípios com até 15.000 habitantes e 10% para os demais municípios, acarretando uma fração geral de cerca de 11% dos domicílios e das também pessoas. No Censo Demográfico 2000, foram usadas as mesmas frações amostrais que em 1991.

Para o Censo Demográfico 2010, os estudos tiveram início em 2005, no âmbito do projeto denominado Estudos de Modalidades Alternativas de Censos Demográficos - EMACD, pelo grupo de trabalho criado para estudar os aspectos de Amostragem, Estimação e Acumulação de Informações.

Um ponto que foi muito discutido durante as etapas de planejamento dos Censos Demográficos 1991 e 2000, que influenciou fortemente a definição das frações amostrais usadas, refere-se às questões operacionais relacionadas com a utilização de mais de uma fração amostral, dependendo do tamanho do município. Havia uma preocupação muito grande com os controles para garantir a aplicação correta da fração definida para cada município. Isso determinou que, naqueles Censos, o número de frações diferentes ficasse reduzido a dois, mesmo havendo evidências das vantagens de se ter mais uma ou mais duas frações distintas. Nesse sentido, o uso de computadores de mão representou um facilitador, uma vez que os controles puderam ser feitos de forma automática e centralizada. E a adoção de mais de duas frações amostrais passou a ser uma alternativa de fato.

Assim, no Censo Demográfico 2010, para a parte do levantamento pesquisada por amostragem, foram aplicadas cinco frações de amostragem, considerando os tamanhos dos municípios em termos da população estimada em 1º de julho de 2009, tal como apresentada a seguir: 50% para os municípios com até 2 500 habitantes; 33% para os municípios com mais de 2 500 e até 8 000 habitantes; 20%, para os municípios com mais de 8 000 e até 20 000 habitantes; 10% para os municípios com mais de 20 000 e até 500 000 habitantes; e 5% para os municípios com população superior a 500 000 habitantes.

Para os 40 municípios com mais de 500 000 habitantes, foi avaliada a possibilidade de aplicação de frações amostrais diferentes em cada uma de suas divisões administrativas intramunicipais (distritos e subdistritos), de forma a permitir a divulgação de estimativas e de microdados nesses níveis geográficos. Em 18 desses municípios, houve a necessidade de aumento da fração amostral, definida dentre as especificadas, em pelo menos uma subdivisão. Nos demais municípios dessa classe (22 municípios), a fração amostral de 5% foi mantida, pois, para sete deles, não há subdivisão administrativa na base territorial 2010 e, para os 15 restantes, o tamanho esperado da amostra resultante em cada subdivisão já contempla o tamanho mínimo estabelecido para a divulgação das estimativas para todas as subdivisões existentes.

A aplicação dessas frações de amostragem fez com que a dimensão da amostra do Censo resultasse robusta o suficiente, propiciando medidas de precisão adequadas para níveis geográficos variados, como forma de atender às demandas por informações municipais e, dependendo da dimensão populacional do município, até mesmo para áreas menores.

Apenas como curiosidade, os tamanhos de amostra nos últimos três Censos Demográficos foram:

Censo 1991 – 4.024.553 domicílios e 17.045.712 pessoas;

Censo 2000 – 5.304.711 domicílios e 20.274.412 pessoas;

Censo 2010 – 6.192.332 domicílios e 20.635.472 pessoas.

Expansão da amostra

Para a expansão da amostra é necessário definir basicamente três elementos: o método de cálculo dos pesos a serem associados a cada unidade da amostra, as áreas de ponderação onde serão calculados os pesos, e o cálculo das medidas de precisão associadas às estimativas obtidas com os pesos calculados. Os itens a seguir tecem considerações sobre esses elementos para os Censos Demográficos brasileiros.

Método de cálculos dos pesos

A etapa de expansão da amostra consiste em aplicar um método de cálculo de pesos, ou fatores de expansão, que leve em consideração o plano amostral e a existência de informações sobre algumas características conhecidas da população em nível censitário para servirem como variáveis auxiliares no processo.

O método de expansão da amostra adotado nos Censos Demográficos 1960 e 1970 foi denominado Grupos de Controle, cujo estimador corresponde ao estimador simples da pós-estratificação, onde os estratos ou grupos de controle são definidos por meio do cruzamento de características para as quais se tem conhecimento dos dados de universo. Nesse processo, em cada grupo de controle, um peso ou fator de expansão é calculado pela razão entre o número de pessoas no universo e na amostra, encontrados no grupo, respeitados os critérios de tamanho mínimo do número de pessoas no grupo e tamanho máximo do peso obtido.

Em 1980, o método de expansão utilizado foi denominado Processo Iterativo de Estimação por Totais Marginais - Pietom, caso particular do *Raking Ratio Estimation Procedure* - RREP (METODOLOGIA..., 1983). Esse processo faz uso iterativo de estimadores de razão em duas dimensões, com o objetivo de usar uma maior quantidade de pós-estratos, ou seja, uma matriz de estratos com um grande número de células, e diminuir os problemas que surgem na pós-estratificação simples com a definição de estratos pequenos.

Em 1991, para a determinação dos pesos foi usado o método denominado no IBGE por Mínimos Quadrados Generalizados em duas etapas - MQG2, uma adaptação do *Generalized Least Squares Estimation Procedure* - GLSEP, de Bankier, Rathwell e Majkowski (1992), e as variáveis auxiliares utilizadas foram definidas dentre aquelas investigadas para 100% da população, no próprio Censo Demográfico. Esse procedimento de estimação de regressão atribui um único peso fracionário a cada domicílio e a cada um de seus moradores, sendo importante destacar essas duas situações novas em relação aos Censos anteriores: o peso fracionário e único para domicílios, famílias e pessoas.

Em 2000 e em 2010, o procedimento usado para a determinação dos pesos para a expansão da amostra foi o mesmo do de 1991, porém simplificado, passando o ajuste do modelo de regressão a ser realizado em apenas uma etapa em cada área de ponderação.

Áreas de ponderação

O procedimento de estimação de totais foi aplicado em cada área de ponderação separadamente. Uma área de ponderação é um conjunto de setores censitários que corresponde à menor área geográfica onde é aplicado o método de cálculo de pesos (procedimento de calibração dos pesos) e para a qual são calculadas e divulgadas as estimativas provenientes da amostra.

A definição das áreas de ponderação também sofreu alterações a cada Censo. Em 1960 e 1970, as áreas de ponderação eram os municípios. Em 1980, para um conjunto de municípios grandes, o distrito, o subdistrito, a região administrativa, a zona administrativa, ou agregados destes, foram definidos como áreas de ponderação, desde que tivessem população igual ou superior a 5.000 pessoas. Essas informações estão detalhadas na publicação *Metodologia do censo demográfico de 1980* (1983).

Em 1991, além de cada município, a princípio, poder ser uma área de ponderação, foi definido um procedimento que considerava cada divisão administrativa interna dos municípios como uma área de ponderação, obedecido o critério de tamanho mínimo definido de 4.000 pessoas, seguindo a experiência do Canadá. Posteriormente, nesses

municípios, as áreas resultantes foram divididas em agregados de setores de tamanhos aproximadamente iguais em número de setores, considerando também o tamanho mínimo definido. Nesse processo de construção dos agregados de setores, não foram feitas restrições de contiguidade dos setores nem de homogeneidade em relação a alguma característica, o que acarretou a formação de áreas não necessariamente contínuas e sem significado geográfico ou administrativo.

Para os Censos Demográficos 2000 e 2010, foram usados métodos e sistemas automáticos de formação de áreas de ponderação que conjugam os seguintes critérios: tamanho, que foi definido em 400 domicílios particulares ocupados na amostra, para permitir estimativas com qualidade estatística em áreas pequenas; contiguidade, no sentido de serem constituídas por conjuntos de setores limítrofes com sentido geográfico; e homogeneidade em relação a um conjunto de características populacionais e de infraestrutura conhecidas. Além disso, alguns municípios grandes definiram as áreas de ponderação em função de suas áreas de planejamento municipal.

Precisão das estimativas

Os erros amostrais podem ser avaliados através das estimativas dos coeficientes de variação ou dos erros-padrão calculados a partir das estimativas das variâncias. É possível estimar os erros amostrais de acordo com a metodologia usada na obtenção dos pesos, ou seja, considerando o estimador definido para o cálculo das estimativas.

No caso do Censo Demográfico 1980, os cálculos dos coeficientes de variação foram feitos para um conjunto de estimativas seguindo a fórmula própria de variância associada ao estimador decorrente do processo de expansão usado, o Pietom. Para a divulgação da precisão das estimativas, foi feito um modelo de regressão para o ajuste do coeficiente de variação estimado em função do tamanho da estimativa, considerando o conjunto de variáveis categóricas para as quais foram calculados os erros amostrais usando a fórmula própria. E foram divulgadas as estimativas dos coeficientes de regressão, de forma a permitir aos usuários a obtenção de um valor aproximado para o coeficiente de variação de uma estimativa de interesse.

No caso dos Censos Demográficos 1991, 2000 e 2010, o método direto é bastante complexo (SÄRNDAL; SWENSSON; WRETMAN, 1992) e pode ser implementado usando, por exemplo, o pacote *survey* do programa estatístico R. Mas também foi usado apenas para avaliação interna.

Como a amostra usada no Censo Demográfico é bastante grande e os domicílios se distribuem de forma aleatória dentro de cada setor censitário, pode-se aproximar o cálculo do erro padrão, segundo Cochran (1977), usando as fórmulas da amostragem aleatória simples sem reposição.

Considerando isso e a complexidade das fórmulas associadas ao estimador de regressão utilizado nos últimos três Censos Demográficos, o IBGE vem adotando um método simples e rápido para obtenção de uma aproximação do erro padrão da estimativa, que pode ser usado para a construção de intervalos aproximados com níveis de confiança fixados.

Assim, nas divulgações de resultados provenientes da investigação por amostragem, são divulgadas tabelas onde são apresentados valores de erros-padrão calculados para alguns valores de estimativas de características de pessoas e de domicílios, considerando a aproximação para amostragem aleatória simples sem reposição nos diversos níveis geográficos de divulgação dos resultados do Censo: Brasil, Grandes Regiões e Unidades da Federação.

Conclusões

Com a descrição sucinta apresentada, pode-se perceber claramente a evolução na utilização de amostragem probabilística na coleta de dados dos Censos Demográficos brasileiros, sendo que foram dados destaques para: o esquema de seleção da amostra,

as frações amostrais definidas, em função do tamanho do município, e o método de cálculo de pesos para a expansão da amostra e obtenção das estimativas. No que se refere à avaliação da precisão das estimativas, pode-se dizer que já é possível estimar medidas de precisão, por meio do erro padrão associado a cada estimativa, considerando o plano amostral e o método de cálculo de pesos adotados. Porém, ainda não está totalmente solucionada a questão da divulgação dessas medidas de precisão, ou mesmo a divulgação de informações que permitam ao usuário dos microdados da amostra do Censo Demográfico calcularem adequadamente, e não por aproximação, os erros amostrais associados às estimativas de interesse.

Referências

ALBIERI, S.; BIANCHINI, Z. M. *Principais aspectos de amostragem das pesquisas domiciliares do IBGE:* revisão 2015. Rio de Janeiro: IBGE, 2015. [54] p. (Texto para discussão. Diretoria de Pesquisas, n. 55). Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=294403>. Acesso em: set. 2017.

BANKIER, M. D.; RATHWELL, S.; MAJKOWSKI, M. *Two step generalized least squares estimation in the 1991 canadian census.* Ottawa: Statistics Canada, c1992. 26 p. (Working paper. Methodology branch, v. 92). Disponível em: <http://publications.gc.ca/collections/collection_2017/statcan/11-613/CS11-613-92-7-eng.pdf>. Acesso em: set. 2017.

COCHRAN, W. G. *Sampling techniques.* 3rd ed. New York: Wiley, c1977. 428 p. (Wiley series in probability and mathematical statistics). Disponível em: <https://archive.org/details/Cochran1977SamplingTechniques_201703>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico de 1980. Rio de Janeiro: IBGE, 1983. 478 p. (Série relatórios metodológicos, v. 4). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv13083.pdf>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2000. Rio de Janeiro: IBGE, 2003. 574 p. (Série relatórios metodológicos, v. 25). Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2000/metodologia/default.shtm >. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2010. 2. ed. Rio de Janeiro: IBGE, 2016. 720 p. Acompanha 1 CD-ROM. (Série relatórios metodológicos, v. 41). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv95987.pdf>. Acesso em: set. 2017.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. H. *Model assisted survey sampling.* New York: Springer-Verlag, 1992. 694 p. (Springer series in statistics).

Estimação de moradores em domicílios fechados no Censo Demográfico 2010

Antonio José Ribeiro Dias
Alexandre dos Reis Santos

Motivação

As unidades domiciliares pesquisadas nos Censos Demográficos e em Contagens da População são classificadas em categorias de acordo com a situação de seus moradores na data de referência da coleta, a saber: domicílios ocupados (particulares ou coletivos); domicílios fechados; domicílios vagos; e domicílios de uso ocasional. A operação censitária visa obter informações das pessoas moradoras nos domicílios classificados nas duas primeiras categorias (domicílios ocupados e domicílios fechados).

Os domicílios classificados como fechados são aqueles que sabidamente possuíam moradores na data de referência, mas que não tiveram entrevista realizada para o preenchimento das informações do questionário, independentemente do motivo da não realização da entrevista, que pode ser tanto por recusa dos moradores em prestar informações como alguma dificuldade do recenseador em estabelecer contato com os informantes (ou seja, a ausência de pessoas no domicílio nos momentos das visitas do recenseador).

Antes de 2007, nas divulgações de resultados de Censos Demográficos, os totais da população para cada um dos municípios brasileiros foram sempre divulgados considerando apenas os domicílios ocupados (particulares e coletivos) na data de referência da operação censitária, sendo que os moradores nos domicílios fechados eram simplesmente ignorados, levando a uma subestimação do total populacional.

As informações sobre o número de domicílios fechados, vagos e de uso ocasional, que também são divulgadas, são usadas, em conjunto com outras informações disponíveis, para a avaliação da qualidade da cobertura das operações censitárias e, neste sentido, elas contribuem indiretamente para os procedimentos de avaliação das estimativas municipais de população.

Pela primeira vez em 2007, o IBGE decidiu adotar o procedimento de estimar a parcela da população moradora nos domicílios fechados em cada um dos municípios abrangidos pela operação da Contagem da População 2007.

Foi realizada uma revisão bibliográfica para identificar se e como outros países que realizam censos de população tratavam essa questão[1]; e

[1] Foram consultados documentos descrevendo os procedimentos adotados, por exemplo, no México e na Austrália.

a definição do procedimento a ser utilizado, naquela ocasião, foi feita após a avaliação da disponibilidade de informações que pudessem auxiliar no processo de estimação.

A primeira hipótese considerada referia-se ao padrão de tamanho dos domicílios, em número de moradores, daqueles classificados como fechados. Teriam eles a mesma distribuição de tamanho do que aqueles que foram entrevistados sem dificuldades? A hipótese de que essas distribuições são distintas é bastante razoável e o que se seguiu foram estudos para tentar determinar quais características seriam as definidoras da mudança de padrão e qual seria o padrão da distribuição do número de moradores nos domicílios classificados como fechados. Assim, admitiu-se que os domicílios fechados possuem alguma característica em sua composição, principalmente no número de moradores, que implicou na dificuldade do recenseador para realizar a entrevista ou na recusa do informante, e na sua classificação como fechado após o término do período de coleta.

Após adotar pela primeira vez o procedimento de estimação do número de moradores em domicílios fechados em 2007, decidiu-se por fazê-lo, também, no Censo Demográfico 2010. Isto porque considerou-se que essa estimação representa um aperfeiçoamento metodológico na produção de estimativas populacionais provenientes de Censos e Contagens da População. Assim, no Censo Demográfico 2010, com o mesmo objetivo de diminuir a diferença entre o quantitativo populacional recenseado e o efetivo, o IBGE novamente adotou um procedimento de estimação de moradores em domicílios fechados. Para servir como variável auxiliar no processo de imputação usado durante o tratamento dos domicílios fechados, foram registradas todas as alterações da espécie[2] dos domicílios durante o período da coleta; por espécie entende-se: unidade não residencial, domicílio ocupado, domicílio fechado, domicílio vago, domicílio de uso ocasional.

Com isso, esperava-se poder discriminar as características dos domicílios fechados, em relação às características das demais espécies, principalmente a de ocupados com entrevistas realizadas, de forma a obter variáveis auxiliares que de fato melhorassem a qualidade do processo de estimação desejado.

Vale destacar que essa prática é adotada internacionalmente, dentre outros, por países como México, Canadá e Austrália.

Metodologia

Existem diversas possibilidades para a imputação de moradores nos domicílios fechados, onde não se conseguiu realizar entrevista. No censo mexicano de 2010, foi verificada a média de moradores daqueles domicílios onde foram necessárias três ou mais visitas para a realização da entrevista e, a partir desse estudo, decidiu-se por imputar três moradores em cada um dos domicílios fechados. Em seguida foi imputado o sexo de acordo com a distribuição observada na população entrevistada em cada uma das entidades federativas. As demais variáveis não foram imputadas sendo designadas como "não especificado" (SÍNTESIS..., 2011). No censo de 2011 da Austrália foi adotado um procedimento hotdecking em duas etapas. Na primeira etapa foi imputado o número de pessoas por sexo, elegendo um domicílio doador entre os domicílios com o mesmo tipo de estrutura e geograficamente próximos ao domicílio a ser imputado. Na segunda etapa foram imputadas as principais variáveis demográficas a partir da escolha de um doador entre os domicílios entrevistados onde todos os moradores responderam às características demográficas; as estruturas dos domicílios deviam ser semelhantes; deviam ter exatamente o mesmo número de moradores homens e mulheres do domicílio a ser imputado; deviam estar localizados geograficamente o mais próximo possível do domicílio a ser imputado (DERIVATIONS..., 2011; 2016).

No caso da estimação do número de moradores para os domicílios fechados do Censo Demográfico 2010, admitiu-se que o padrão dos domicílios fechados é diferente do padrão

[2] Inicialmente chegou a ser definido que seriam gravadas data e hora de todas as tentativas de entrevistas frustradas e da bem-sucedida que corresponde à entrevista realizada. Entretanto, não foi possível dispor do registro da data e hora das tentativas frustradas, mas ficou o registro da alteração de espécie do domicílio.

dos domicílios ocupados, que foram efetivamente investigados, no que se refere ao número de moradores do domicílio. Ou seja, admitiu-se que os domicílios fechados possuíam características em sua composição, principalmente em relação ao número de moradores, que implicou a dificuldade do recenseador para realizar a entrevista e, consequentemente, na sua classificação como fechado após o término do período de coleta.

Para avaliar essa hipótese, os domicílios particulares ocupados foram estratificados segundo a sua localização. Os estratos foram definidos, para cada Unidade da Federação, considerando a situação do domicílio, urbana ou rural, e a classe de tamanho populacional do município ao qual pertence.

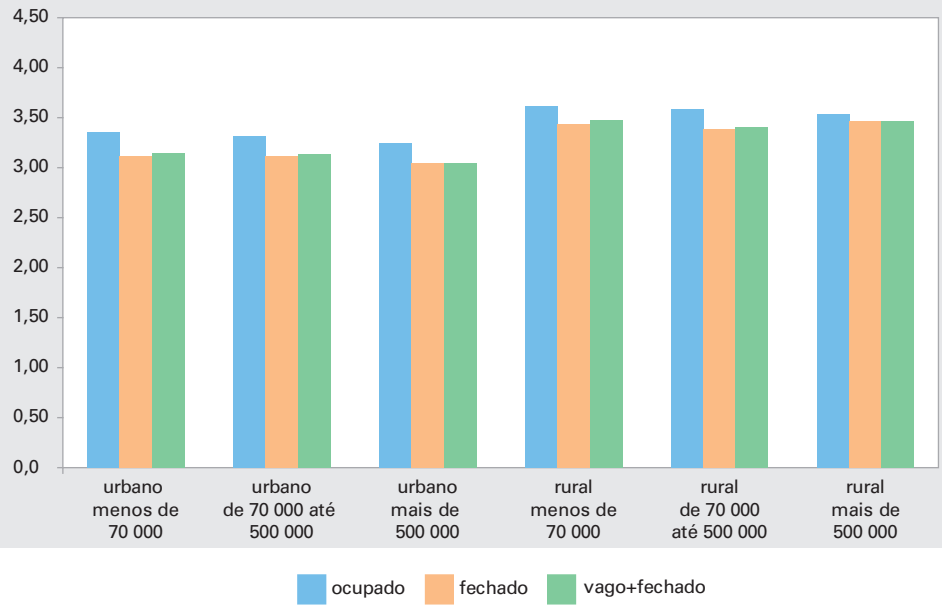
Foram criadas três classes de tamanho, a saber: municípios com menos de 70.000 habitantes; municípios de 70.000 a menos de 500.000 habitantes e municípios com 500.000 habitantes ou mais. Os municípios com 500.000 habitantes ou mais, foram tratados individualmente, enquanto os demais foram considerados em seus respectivos estratos de tamanho. Não fizeram parte da análise os domicílios de Setores Censitários localizados em Terras Indígenas, que foram objeto de um tratamento à parte, além dos domicílios com mais de 10 moradores, para garantir a robustez do método. Foram obtidas as distribuições do número de moradores em domicílios particulares ocupados por estrato em dois conjuntos de entrevistas realizadas, a saber:

1. Domicílios particulares ocupados que tiveram entrevista realizada logo na primeira tentativa;
2. Domicílios particulares ocupados inicialmente classificados como fechados ou vagos (estes últimos também considerados por terem sido erroneamente classificados como vagos), mas que posteriormente tiveram entrevista realizada.

Em cada estrato, a análise das duas distribuições do número de moradores confirmou a hipótese, verificando-se, de modo geral, um menor número médio de moradores nos domicílios apontados em (2) do que em (1).

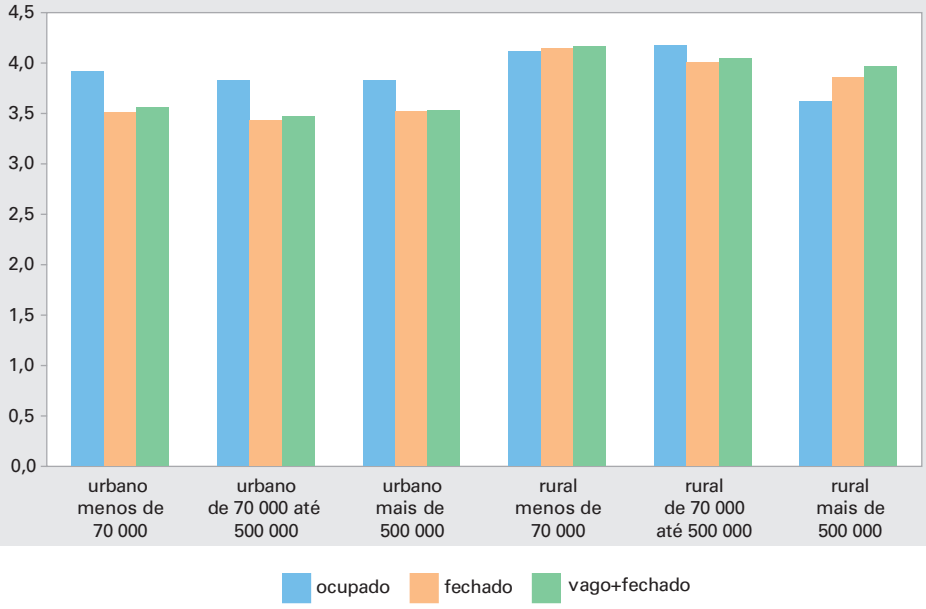
Os Gráficos 1 até 6, a seguir, que apresentam os gráficos para o Brasil e cada uma das cinco Grandes Regiões, evidenciam a conclusão acima.

Gráfico 1 - Brasil - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



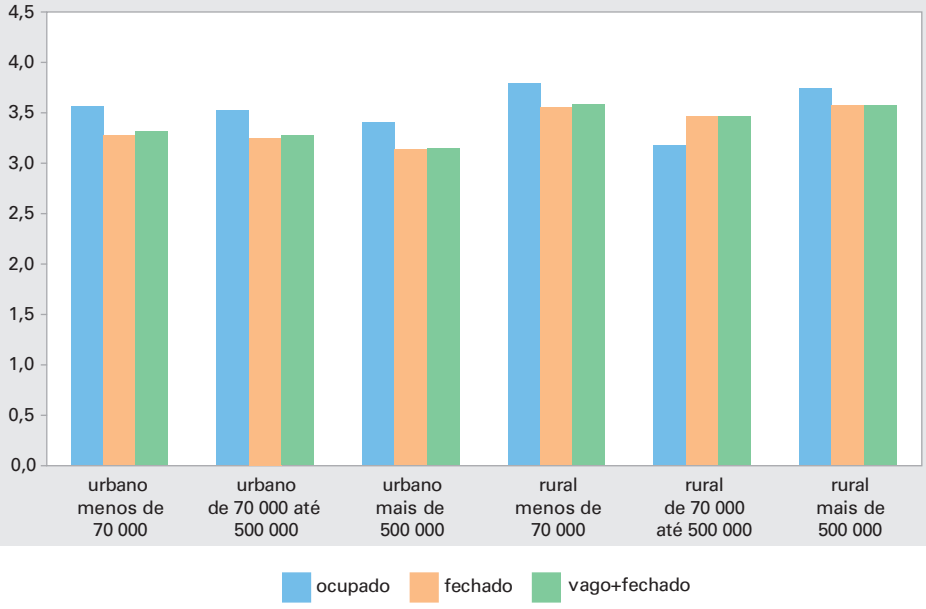
Fonte: IBGE, Censo Demográfico 2010.

Gráfico 2 - Região Norte - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



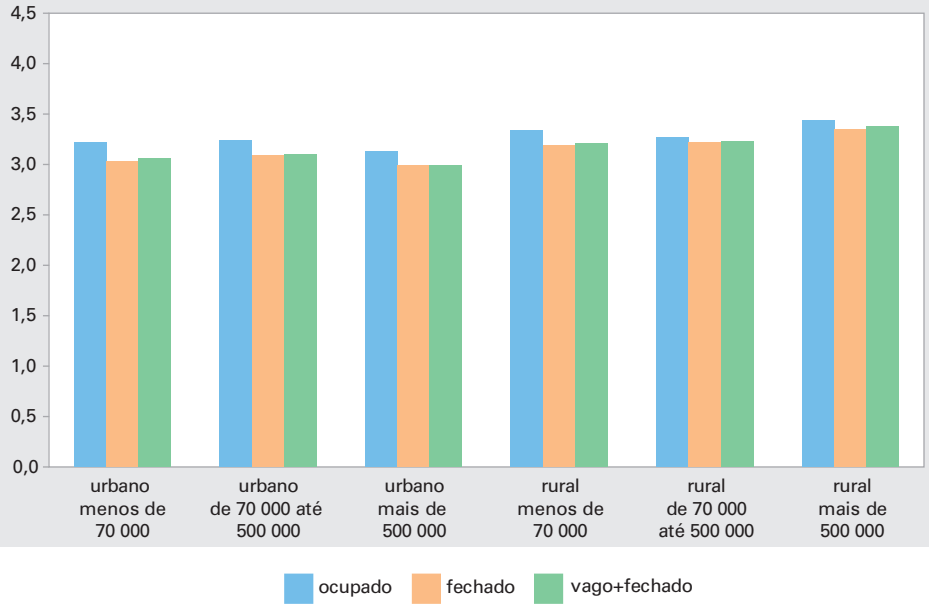
Fonte: IBGE, Censo Demográfico 2010.

Gráfico 3 - Região Nordeste - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



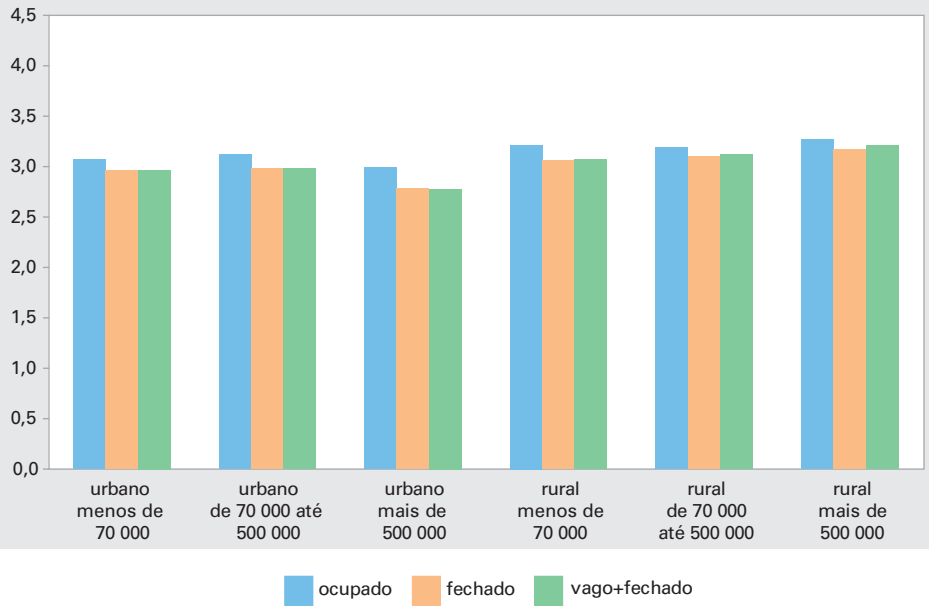
Fonte: IBGE, Censo Demográfico 2010.

Gráfico 4 - Região Sudeste - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



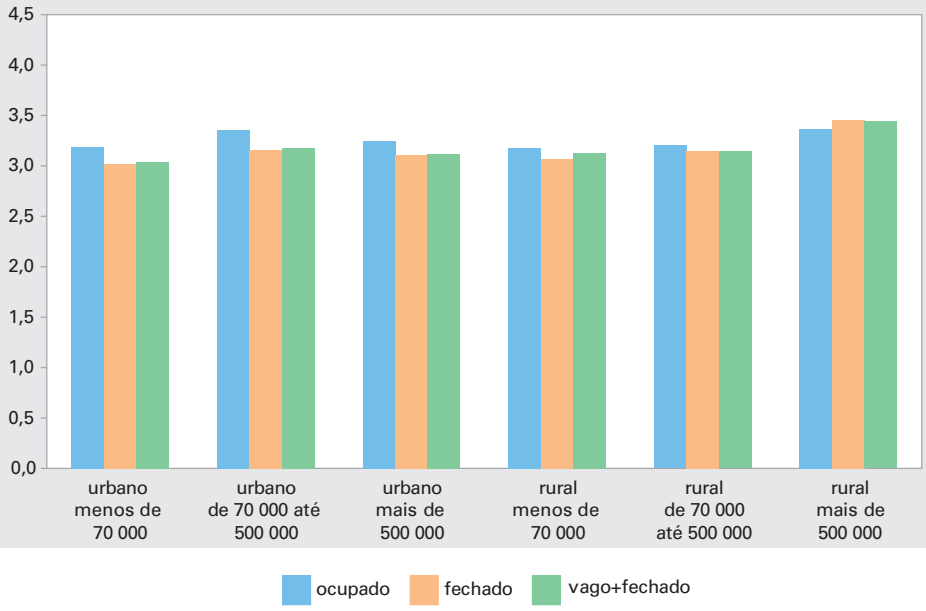
Fonte: IBGE, Censo Demográfico 2010.

Gráfico 5 - Região Sul - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



Fonte: IBGE, Censo Demográfico 2010.

Gráfico 6 - Região Centro-Oeste - Média de moradores nos domicílios, por estrato, segundo a espécie do domicílio, situação e população do município



Fonte: IBGE, Censo Demográfico 2010.

A classificação de um domicílio na categoria de fechado é equivalente a considerá-lo como uma não resposta, que é um dos erros não amostrais mais comuns na realização de uma pesquisa, seja ela censitária ou por amostragem, que pode ocorrer por variados motivos como a recusa por parte do informante e a impossibilidade de localização do informante durante o período de coleta. Há muitas formas diferentes de se lidar com a não resposta.

Uma delas é a que utiliza procedimentos de imputação, onde se atribui informações individuais às unidades sem informação. O pressuposto básico do procedimento de imputação é que a perda de dados foi aleatória e, se não for, que o padrão de não resposta seja conhecido ou, pelo menos, estimado, para ser considerado durante o tratamento da não resposta por imputação.

Para estimar o número de moradores em domicílios fechados para cada município abrangido pelo Censo Demográfico 2010, definiu-se cada domicílio fechado como uma não resposta cujo atributo necessário é o número de moradores. O tratamento adotado para essa não resposta foi um procedimento de imputação por meio de seleção aleatória de um domicílio doador entre um conjunto de possíveis doadores, formado pelos domicílios particulares ocupados classificados como em (2), tendo sido adotada, ainda, a estratificação de domicílios acima descrita. A exceção a essa regra foram os domicílios rurais de municípios com mais de 500.000 habitantes, onde, por uma questão de robustez do método, foram considerados como possíveis doadores todos os domicílios particulares ocupados, devido à pouca ocorrência de doadores rurais.

Em termos operacionais, o procedimento consistiu em imputar, para cada município, tantos domicílios quantos os classificados como fechados, com número de moradores de acordo com a distribuição obtida pelo conjunto de domicílios do estrato correspondente.

Assim, o total de moradores estimados no conjunto de domicílios fechados de cada município foi obtido pela soma dos moradores nos domicílios imputados.

O procedimento de estimação foi aplicado aos domicílios efetivamente fechados, após todas as tentativas de obtenção da entrevista, que correspondem, para todo o território nacional, a 1,55% do total de domicílios ocupados pesquisados pelo Censo Demográfico 2010.

Foram imputadas, a partir das informações do domicílio doador, todas as variáveis do questionário básico, ou seja, as variáveis investigadas para todas as pessoas da população, exceto as variáveis relativas à população indígena, como língua indígena e etnia.

Nas Tabelas 1 e 2, a seguir, são apresentados os números finais, absolutos e relativos, de domicílios fechados e moradores imputados nos mesmos, por sexo e Unidade da Federação.

Tabela 1 - Censo Demográfico de 2010 - Domicílios fechados e pessoas imputadas nos domicílios fechados, por Unidade da Federação

Unidades da Federação	Domicílios		Pessoas		Domicílios Fechados (%)	Pessoas Imputadas (%)
	Total	Fechados	Total	Imputadas		
Brasil	**58 051 449**	**899 152**	**190 755 799**	**2 795 533**	**1,55**	**1,47**
Rondônia	468 316	6 292	1 562 409	19 584	1,34	1,25
Acre	193 692	5 469	733 559	18 116	2,82	2,47
Amazonas	806 974	22 985	3 483 985	83 877	2,85	2,41
Roraima	117 965	1 124	450 479	4 159	0,95	0,92
Pará	1 877 876	35 912	7 581 051	126 574	1,91	1,67
Amapá	158 453	6 466	669 526	21 692	4,08	3,24
Tocantins	402 257	2 024	1 383 445	6 726	0,50	0,49
Maranão	1 661 659	35 142	6 574 789	124 231	2,11	1,89
Piauí	852 506	4 288	3 118 360	14 742	0,50	0,47
Ceará	2 380 173	24 504	8 452 381	84 512	1,03	1,00
Rio Grande do Norte	906 488	13 853	3 168 027	43 282	1,53	1,37
Paraíba	1 090 463	3 496	3 766 528	11 465	0,32	0,30
Pernambuco	2 574 137	45 949	8 796 448	144 321	1,79	1,64
Alagoas	851 101	6 126	3 120 494	19 780	0,72	0,63
Sergipe	595 769	9 393	2 068 017	28 999	1,58	1,40
Bahia	4 126 224	60 412	14 016 906	186 488	1,46	1,33
Minas Gerais	6 111 179	75 194	19 597 330	229 685	1,23	1,17
Espírito Santo	1 113 408	21 630	3 514 952	67 590	1,94	1,92
Rio de Janeiro	5 299 014	126 392	15 989 929	361 465	2,39	2,26
São Paulo	13 053 253	230 099	41 262 199	698 135	1,76	1,69
Paraná	3 340 516	32 737	10 444 526	102 243	0,98	0,98
Santa Catarina	2 015 139	17 230	6 248 436	51 545	0,86	0,82
Rio Grande do Sul	3 653 000	19 541	10 693 929	55 090	0,53	0,52
Mato Grosso do Sul	775 003	11 711	2 449 024	35 844	1,51	1,46
Mato Grosso	932 110	22 806	3 035 122	69 740	2,45	2,30
Goiás	1 909 041	35 632	6 003 788	107 454	1,87	1,79
Distrito Federal	785 733	22 745	2 570 160	78 194	2,89	3,04

Fonte: IBGE, Censo Demográfico 2010.

Nota: No total estão incluídos os domicílios particulares permanentes ocupados com e sem entrevistas realizadas (fechados), domicílios particulares improvisados ocupados e domicílios coletivos com morador.

Tabela 2 - Censo Demográfico de 2010 – Pessoas imputadas nos domicílios fechados, por sexo e Unidade da Federação

Unidades da Federação	Homens		Mulheres		Imputados (%)	
	Total	Imputados	Total	Imputadas	Homens	Mulheres
Brasil	**93 406 990**	**1 354 472**	**97 348 809**	**1 441 061**	**1,45**	**1,48**
Rondônia	795 157	10 025	767 252	9 559	1,26	1,25
Acre	368 324	9 139	365 235	8 977	2,48	2,46
Amazonas	1 753 179	41 947	1 730 806	41 930	2,39	2,42
Roraima	228 859	2 102	221 620	2 057	0,92	0,93
Pará	3 821 837	63 159	3 759 214	63 415	1,65	1,69
Amapá	335 135	10 972	334 391	10 720	3,27	3,21
Tocantins	702 424	3 380	681 021	3 346	0,48	0,49
Maranão	3 261 515	61 137	3 313 274	63 094	1,87	1,90
Piauí	1 528 422	7 169	1 589 938	7 573	0,47	0,48
Ceará	4 120 088	40 131	4 332 293	44 381	0,97	1,02
Rio Grande do Norte	1 548 887	21 190	1 619 140	22 092	1,37	1,36
Paraíba	1 824 379	5 459	1 942 149	6 006	0,30	0,31
Pernambuco	4 230 681	69 099	4 565 767	75 222	1,63	1,65
Alagoas	1 511 767	9 451	1 608 727	10 329	0,63	0,64
Sergipe	1 005 041	13 876	1 062 976	15 123	1,38	1,42
Bahia	6 878 266	89 994	7 138 640	96 494	1,31	1,35
Minas Gerais	9 641 877	110 886	9 955 453	118 799	1,15	1,19
Espírito Santo	1 731 218	33 184	1 783 734	34 406	1,92	1,93
Rio de Janeiro	7 625 679	170 974	8 364 250	190 491	2,24	2,28
São Paulo	20 077 873	335 811	21 184 326	362 324	1,67	1,71
Paraná	5 130 994	49 857	5 313 532	52 386	0,97	0,99
Santa Catarina	3 100 360	25 242	3 148 076	26 303	0,81	0,84
Rio Grande do Sul	5 205 057	26 240	5 488 872	28 850	0,50	0,53
Mato Grosso do Sul	1 219 928	17 613	1 229 096	18 231	1,44	1,48
Mato Grosso	1 549 536	35 773	1 485 586	33 967	2,31	2,29
Goiás	2 981 627	53 205	3 022 161	54 249	1,78	1,80
Distrito Federal	1 228 880	37 457	1 341 280	40 737	3,05	3,04

Fonte: IBGE, Censo Demográfico 2010.

Na base de dados do Censo Demográfico 2010 é possível identificar os domicílios imputados através da variável "espécie de unidade visitada", onde os domicílios efetivamente entrevistados aparecem como "domicílio particular permanente ocupado" enquanto que os imputados são classificados como "domicílio particular permanente ocupado sem entrevista realizada". Dessa maneira é possível aos usuários dos dados do Censo Demográfico 2010 optarem por algum tratamento alternativo, a sua escolha, em relação aos moradores dos domicílios onde não foi possível realizar a entrevista.

Considerações finais

A partir da Contagem da População 2007 o IBGE passou a imputar moradores aos domicílios fechados. Esse procedimento melhora as estimativas das variáveis referentes ao total populacional, já que antes havia, de fato, uma subestimação, pois só eram considerados os domicílios ocupados e com entrevistas realizadas durante a coleta dos dados na operação censitária. Tal procedimento é, também, adotado por outros países, como o México, a Austrália, o Canadá, entre outros.

A imputação consiste em "completar" os bancos de dados preenchendo as ausências de informação (*missing data*) com estimativas, geralmente baseadas nas informações das unidades para as quais não há falta de dados. Isso facilita os procedimentos de análise a serem utilizados, pois passa-se a ter um banco completo. É fundamental, porém, que os métodos utilizados na imputação sejam bem documentados e que os valores imputados possam ser claramente identificados, permitindo que os usuários possam, inclusive, fazer as suas próprias imputações, caso desejem, ou trabalhar apenas com os registros completos, desprezando os que sofreram alguma imputação.

Referências

2011 CENSUS item edit and imputation process. London: Office for National Statisitcs - ONS, 2012. (2011 census: methods and quality report). 23 p. Disponível em: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/index.html>. Acesso em: set. 2017.

ALBIERI, S. *A ausência de resposta em pesquisas:* uma aplicação de métodos de imputação. Rio de Janeiro: Instituto de Matemática Pura e Aplicada - IMPA, 1992. 138 p. (Informes de matemática. Série D, 48). Originalmente apresentada como dissertação de Mestrado no Instituto, Rio de Janeiro, 1989. Disponível em: <ftp://ftp.dpe.ibge.gov.br/Dissertacao-tese/Dissertacao_Sonia_Albieri.pdf>. Acesso em: set. 2017.

DERIVATIONS and imputations. In: CENSUS dictionary, 2006 (reissue). Canberra: Australian Bureau of Statistics - ABS, 2011. Glossary. Disponível em: <http://www.abs.gov.au/AUSSTATS/abs@.nsf/bb8db737e2af84b8ca2571780015701e/973aeb439071a9e5ca25720a000dcfe6!OpenDocument>. Acesso em: set. 2017.

_____. In: CENSUS dictionary, 2011. Canberra: Australian Bureau of Statistics - ABS, 2016. Glossary. Disponível em: <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter2910201>. Acesso em: set. 2017.

DIAS, A. J. R.; ALBIERI, S. Uso de imputação em pesquisas domiciliares. In: ENCONTRO NACIONAL DE ESTUDOS E POPULAÇÕES, 8., 1992, São Paulo. *Anais...* Brasília, DF: Associação Brasileira de Estudos Populacionais - ABEP, 1992. p. 11-26. Disponível em: <http://www.abep.org.br/publicacoes/index.php/anais/article/view/562/542>. Acesso em: set. 2017.

DICK, P. The census of Canada: the dwelling classification study. In: JOINT STATISTICAL MEETINGS, 3., 2002, New York. *Proceedings...* Alexandria [Estados Unidos]: American Statistical Association - ASA, 2002. p. 782-787.

GARCÍA RUBIO, E.; VILLÁN CRIADO, I. Dia: descripción del sistema. In:_____. *Sistema DIA:* sistema de detección e imputación automática de errores para datos cualitativos. Madrid: Instituto Nacional de Estadística - INE, 1988. v. 1.

PESSOA, D. G. C.; MOREIRA, G. G.; SANTOS, A. dos R. *Imputação de rendimentos no questionário da amostra do censo demográfico 2000.* Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2003. 19 p.

PESSOA, D. G. C.; SANTOS, A. dos R. *Imputação da variável de rendimento dos responsáveis por domicílios:* conjunto universo do censo demográfico 2000. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2004. 15 p.

RAHMAN, N.; GOLDRING, S. Modelling census household non-response. In: ISI WORLD STATISTICS CONGRESS, 56., 2007, Lisboa; SATELLITE MEETING ON INNOVATIVE METHODOLOGIES FOR CENSUSES IN THE NEW MILLENNIUM, 2007, Lisboa. *Proceedings...* The Hague [Holanda]: International Statistical Institute - ISI, 2007. Disponível em: <https://unstats.un.org/unsd/censuskb20/Attachments/2007GBR_ISI-GUIDbaf5209280564a0391f2edbaedd0daeb.pdf >. Acesso em: set. 2017.

SÍNTESIS metodológica **y** conceptual del censo de población **y** vivienda 2010. Aguascalientes [México]: Instituto Nacional de Estadística **y** Geografía - Inegi, 2011. 81 p. Disponível em: <http://siga.conagua.gob.mx/Censo_2010/Doc/sm_cpv2010.pdf>. Acesso em: set. 2017.

Avaliação empírica do estimador de variância do método do conglomerado primário para estimadores de totais na Pesquisa Nacional por Amostra de Domicílios - PNAD Contínua

Pedro Luis do Nascimento Silva
Sâmela Batista Arantes
Roberta Carneiro de Souza

Introdução

Uma parte importante de qualquer análise de dados amostrais é obter as estimativas de variância, isto é, conhecer a precisão das estimativas pontuais. Atualmente, uma abundância de microdados de pesquisas amostrais tem sido disponibilizada pelas agências oficiais de estatística, particularmente no caso das pesquisas domiciliares. Este é também o caso e a prática do Instituto Brasileiro de Geografia e Estatística - IBGE, que tem disponibilizado os microdados de suas pesquisas amostrais domiciliares desde os anos 90.

A Pesquisa Nacional por Amostra de Domicílios - PNAD Contínua, iniciada em 2012, adota um plano amostral complexo, estratificado e conglomerado em dois estágios, onde em cada estrato as Unidades Primárias de Amostragem - UPAs são selecionadas pelo método de Probabilidade Proporcional ao Tamanho - PPT de Pareto sem reposição. Este método é diferente do método usado anteriormente pelo IBGE nas pesquisas amostrais domiciliares para a seleção de UPAs, que era Amostragem PPT Sistemática (FREITAS et al., 2007).

Desde o início da implementação da pesquisa, a variância dos estimadores de totais da PNAD Contínua tem sido estimada usando o método do Conglomerado Primário - CP (HANSEN; HURWITZ; MADOW, 1953), que faz essa estimação tratando a seleção de UPAs no primeiro estágio como se tivesse sido feita usando um método de amostragem PPT com reposição (PESSOA; SILVA, 1998). Este método tem sido bastante usado na prática em muitas outras pesquisas domiciliares similares, mas pouco se sabe a respeito do desempenho empírico deste método de estimação sob o desenho amostral usado na PNAD Contínua.

Considerando que a PNAD Contínua é uma pesquisa muito importante para o sistema estatístico brasileiro, é importante assegurar que suas estimativas de variância tenham boa acurácia, e espera-se também que sejam aproximadamente não viciadas. Isto é relevante para estabelecer se a estimação de variância hoje praticada no IBGE, a qual pode ser facilmente implementada usando software de análise de pesquisas amostrais, poderia ser recomendada para uso geral na análise dos microdados da PNAD Contínua que são disponibilizados gratuitamente na web para análises secundárias.

Este capítulo, portanto, investiga o desempenho do estimador de variância do método CP para estimar totais sob o desenho amostral utilizado pela PNAD Contínua.

No tópico **Plano amostral da PNAD Contínua** descrevemos os detalhes do desenho amostral usado na PNAD Contínua. No tópico seguinte, **Estimação de totais sob amostragem estratificada e conglomerada em dois estágios** mostramos as expressões do estimador de total e de sua variância sob o método do Conglomerado Primário. No tópico **Simulação e resultados** descrevemos um estudo de simulação realizado para avaliar o desempenho do estimador de variância do método CP e apresentamos seus resultados. Nas **Considerações finais** concluímos este capítulo com um resumo das conclusões e algumas indicações de trabalhos futuros.

A PNAD Contínua

A PNAD Contínua é uma pesquisa nacional trimestral realizada pelo IBGE para obter informações sobre a força de trabalho e indicadores socioeconômicos usados para monitorar o desenvolvimento do país.

Plano amostral da PNAD Contínua

As Unidades Primárias de Amostragem - UPAs são os setores censitários ou grupos de setores censitários. Da mesma forma que as estratégias amostrais previamente usadas pelo IBGE em pesquisas domiciliares, a PNAD Contínua usa estratificação geográfica e estatística das UPAs para espalhamento da amostra, buscando assim aumentar a precisão e também assegurar que há amostra suficiente nos domínios de interesse. Dentro de cada estrato, as UPAs são selecionadas pelo método PPT de Pareto (ROSÉN, 1997), onde a medida de tamanho é o número de domicílios segundo o último Censo Demográfico disponível. Dentro de cada UPA, 14 domicílios (as Unidades Secundárias de Amostragem - USAs) são selecionadas usando Amostragem Aleatória Simples - AAS sem reposição. Os detalhes da estratificação e alocação da amostra estão disponíveis na publicação *Pesquisa nacional por amostra de domicílios contínua: notas metodológicas*. Em geral, o tamanho total da amostra da PNAD Contínua em cada trimestre é de aproximadamente 15 100 UPAs e 211 000 domicílios (PESQUISA..., 2014).

A Amostragem de Pareto é um dos métodos de amostragem por ordenação, disponíveis para seleção com probabilidade proporcional ao tamanho sem reposição (ROSÉN, 1997). Este método pode ser usado para coordenar[1] amostras no tempo ou entre pesquisas diferentes. O tamanho de amostra pode ser prefixado e variâncias podem ser estimadas com boas propriedades (COSTA, 2007). Métodos de amostragem por ordenação consistem em associar um número aleatório a cada unidade populacional, calcular um número aleatório modificado como função da probabilidade de inclusão, ordenar as unidades por esse número aleatório modificado, e então selecionar as unidades com os menores números aleatórios modificados. A amostragem PPT de Pareto para seleção de UPAs numa população qualquer é descrita por Arantes e Silva (2013).

Estimação de totais sob amostragem estratificada e conglomerada em dois estágios

Considere a estimação do total populacional de uma variável y e também da variância desse estimador sob amostragem estratificada e conglomerada em dois

[1] A coordenação de amostras consiste no controle da sobreposição amostral de unidades em períodos consecutivos. Em geral é fixado um percentual de unidades que se pretende sobrepor, como por exemplo, na PNAD Contínua, onde há sobreposição amostral de 80% dos domicílios em trimestres consecutivos.

estágios supondo AAS no segundo estágio. Quando a estratificação ocorre com H estratos, o estimador de Horvitz-Thompson para o total é dado por:

$$\hat{Y}=\sum_{h=1}^{H}\sum_{i\in s_h}\frac{\hat{Y}_{hi}}{\pi_{hi}} \tag{1}$$

onde $\hat{Y}_{hi}=\frac{N_{hi}}{n_{hi}}\sum_{j\in a_{hi}} y_{hij}$ estima o total $Y_{hi}=\sum_{j\in U_{hi}} y_{hij}$ da variável y na UPA i do estrato h, e $\pi_{hi}=m_h\frac{N_{hi}}{N_h}$ é a probabilidade de inclusão na amostra da UPA i do estrato h, m_h é o tamanho da amostra de UPAs no estrato h, N_{hi} é o tamanho da UPA i do estrato h, N_h é a soma dos tamanhos das UPAs no estrato h, s_h é o grupo de UPAs selecionadas na amostra do estrato h e a_{hi} é o grupo de USAs, unidades amostrais secundárias, selecionadas na amostra da UPA i do estrato h.

O estimador de variância atualmente usado na PNAD Contínua é baseado no Método do Conglomerado Primário, e é dado por:

$$\hat{V}_{CP}(\hat{Y})=\sum_{h=1}^{H}\frac{1}{m_h(m_h-1)}\sum_{i\in s_h}\left(\frac{\hat{Y}_{hi}}{p_{hi}}-\hat{Y}_h\right)^2 \tag{2}$$

onde $p_{hi}=\frac{N_{hi}}{N_h}$ e $\hat{Y}_h=\frac{1}{m_h}\sum_{i\in s_h}\frac{\hat{Y}_{hi}}{p_{hi}}$.

Simulação e resultados

Um estudo de simulação foi desenvolvido para avaliar o desempenho do estimador de variância do método CP definido em (2). Esta simulação foi realizada com os dados do Censo Demográfico 2010. Isto nos permitiu obter várias amostras do banco de dados do censo, calcular as estimativas de cada amostra e comparar as estimativas com os valores populacionais.

Por questões de custo computacional, o estudo foi limitado a uma única Unidade da Federação, Minas Gerais. Minas Gerais possui 31.168 UPAs, das quais 1.104 UPAs foram incluídas na amostra da PNAD Contínua, e 404 municípios com, pelo menos, uma UPA na amostra. Em Minas Gerais foram usados 49 estratos finais, considerando todas as variáveis de estratificação e o grupo de rotação das UPAs.

Os motivos da escolha de Minas Gerais para a simulação foram o tamanho do estado, que contém cerca de 850 municípios e aproximadamente 20 milhões de habitantes, e também sua diversidade socioeconômica. Por esses motivos, essa Unidade da Federação é frequentemente utilizada para estudos comparativos de desenhos amostrais e estimadores.

O conjunto de UPAs foi construído e estratificado exatamente como foi feito para seleção da amostra da PNAD Contínua. Em seguida foram selecionadas 500 amostras desse conjunto de forma similar ao que ocorre no desenho amostral da PNAD Contínua, ou seja, as amostras simuladas possuem o mesmo tamanho e mesma alocação do desenho amostral da PNAD Contínua. Para cada uma das 500 amostras de UPAs, foram selecionas as USAs em cada UPA da mesma forma que são selecionadas na PNAD Contínua. Essas 500 réplicas da amostra da PNAD Contínua foram usadas para estimar totais e respectivas variâncias de algumas variáveis disponíveis no Censo Demográfico e definidas na Tabela 1, considerando o método descrito. Todas as variáveis foram observadas em Domicílios Particulares Permanentes - DPPs.

Tabela 1 - Lista das variáveis consideradas no estudo de simulação

Variável	Descrição
V1	Renda Domiciliar total mensal dos moradores com 10 anos ou mais
V2	Total de moradores analfabetos com 10 anos ou mais
V3	Indicador de que o domicílio possui eletricidade
V4	Indicador de que o domicílio tem dois ou mais banheiros
V5	Indicador de que o chefe do domicílio é mulher
V6	Número de homens no domicílio
V7	Número de mulheres no domicílio
V8	Número de moradores de 14 anos ou mais no domicílio

Fonte: IBGE, Censo Demográfico 2010.

A Tabela 2 contém os Coeficientes de Variação - CVs para estimar totais obtidos das 500 réplicas da amostra da PNAD Contínua para a Unidade da Federação de Minas Gerais. A coluna 'CV da simulação' contém os valores dos coeficientes de variação do estimador de total das variáveis estimados através da simulação, considerando a variabilidade das estimativas de total ao longo das 500 réplicas.

A coluna 'CV médio Conglomerado Primário' contém as médias das 500 estimativas de CVs obtidas nas réplicas, quando para cada uma das réplicas a variância do estimador de total foi estimada usando (2) e o total foi estimado usando (1).

Tabela 2 - Coeficientes de Variação para estimadores dos totais das variáveis selecionadas

Variável	CV da simulação (%)	CV médio Conglomerado Primário (%)	Diferença (%) Conglomerado Primário – Simulação
V1	2,84	2,82	(-) 0,65
V2	2,91	2,92	0,48
V3	0,11	0,11	2,71
V4	2,54	2,61	2,77
V5	1,58	1,56	(-) 1,33
V6	0,68	0,69	2,54
V7	0,61	0,63	4,15
V8	0,46	0,48	4,74
Média			1,93

Fonte: Simulações dos Autores com base no Censo Demográfico 2010.

Observa-se que as médias das estimativas de CV obtidas com o método do Conglomerado Primário apresentam pequenas diferenças em comparação com os coeficientes de variação estimados por simulação para o estimador de total. Em termos relativos, a maior diferença não chegou a 5% do valor do CV da simulação. Para seis das variáveis consideradas, a pequena diferença foi no sentido de sobrestimar o CV do estimador do total, o que seria esperado em função do método do Conglomerado Primário ignorar que a seleção das UPAs foi feita sem reposição.

Uma das preocupações a respeito da qualidade dos estimadores de variância é a consistência. Isso levou a uma investigação mais minuciosa desse aspecto. Para isso selecionamos o maior estrato de Minas Gerais com 2.072 UPAs e investigamos o comportamento do estimador de variância para as variáveis V1 (renda) e V2 (analfabetismo) à medida que aumentamos o tamanho da amostra de UPAs nesse estrato. Foram realizadas 100 réplicas da amostra com vários tamanhos para avaliar o comportamento dos estimadores dos coeficientes de variação. Os resultados estão nas Tabelas 3 e 4.

Tabela 3 - Tamanhos de amostra de UPAs e Cvs do total da renda domiciliar para 100 réplicas

Maior estrato em MG da PNADC: Mh = 2.072 UPAs, mh = 59				
Tamanho da amostra de UPAs	Fração amostral (%)	CV da simulação (%)	CV médio Conglomerado Primário (%)	Diferença (%) Conglomerado Primário – Simulação
10	0,5	23,76	16,42	(-) 30,89
20	1,0	17,53	13,04	(-) 25,61
40	1,9	13,28	10,07	(-) 24,17
100	4,8	9,70	7,94	(-) 18,14
200	9,7	5,27	5,45	3,42
400	19,3	4,01	4,05	1,00
800	38,6	3,06	3,04	(-) 0,65
1000	48,3	2,45	2,70	10,20

Fonte: Simulações dos Autores com base no Censo Demográfico 2010.

Para estimação do total da variável Renda, nota-se que o viés relativo da estimação do CV é negativo e relevante para amostras de tamanhos até 100. Para os tamanhos de amostra de UPAs de 200 a 800, o viés relativo da estimação do CV é positivo mas pequeno, e diminui, chegando perto de zero com amostras de 400 ou 800 UPAs. Entretanto, para amostras de 1.000 UPAs, o viés relativo voltou a crescer, chegando a 10,2%. O método pareceu não funcionar muito bem quando os tamanhos da amostra de UPAs no estrato foram menores que 200 para a variável total da renda domiciliar.

Tabela 4 - Tamanhos de amostra de UPAs e CVs do total de analfabetos para 100 réplicas

Maior estrato em MG da PNADC: Mh = 2.072 UPAs, mh = 59				
Tamanho da amostra de UPAs	Fração amostral (%)	CV da simulação (%)	CV médio Conglomerado Primário (%)	Diferença (%) Conglomerado Primário – Simulação
10	0,5	32,02	31,14	(-) 2,75
20	1,0	22,76	22,60	(-) 0,70
40	1,9	16,92	16,71	(-) 1,24
100	4,8	11,64	10,28	(-) 11,68
200	9,7	7,36	7,40	0,54
400	19,3	5,11	5,19	1,57
800	38,6	3,33	3,64	9,31
1000	48,3	2,98	3,30	10,74

Fonte: Simulações dos Autores com base no Censo Demográfico 2010.

Para estimação do total da variável Total de analfabetos, nota-se que o viés relativo da estimação do CV é negativo e pequeno para amostras de tamanhos até 40. Para amostras de 100 UPAs, o viés relativo da estimação do CV foi mais intenso, embora ainda negativo (-11,68%). Para amostras de 200 ou 400 UPAs, o viés relativo da estimação do CV ficou perto de zero. Para amostras de 800 ou 1.000 UPAs, o viés relativo voltou a crescer, chegando perto de 10% nos dois casos.

Estes resultados sugerem que a estimação de variância pelo método CP não é consistente, pois para os maiores tamanhos de amostra o viés torna-se positivo e mais importante que para tamanhos de amostra menores. Como o tamanho de amostra a partir do qual o viés se tornou mais importante variou de uma variável para a outra, não há como sugerir tamanhos de amostra para os quais o viés fique desprezível para qualquer variável.

Na amostra da PNAD Contínua em Minas Gerais, o número de UPAs selecionadas em cada estrato varia muito (Tabela 5). A maior parte dos estratos possuem tamanho de amostra igual a 16 UPAs (24 estratos).

Tabela 5 - Tamanho de Amostra de UPAs nos estratos da PNAD Contínua em Minas Gerais

Número de UPAs na amostra	4	15	16	17	19	20	21	22	24	26	28	29	35	37	40	41	42	45	46	59
Número de estratos	1	2	24	1	1	2	2	2	1	2	1	1	1	1	1	1	1	1	2	1

Fonte: Simulações dos Autores com base no Censo Demográfico 2010.

Considerações finais

No estudo consideramos o desenho amostral da PNAD Contínua com os tamanhos de amostras que ocorrem na prática, número de estratos finais, tamanhos da amostra de UPAs nos estratos e tamanhos de amostras dentro das UPAs. Dessa forma, as evidências encontradas são válidas para os tamanhos de amostras praticados atualmente na PNAD Contínua.

No conjunto da amostra para Minas Gerais, o método do Conglomerado Primário apresentou boa aproximação para estimar variâncias (e coeficientes de variação) dos estimadores de total sob o método de sorteio da PNAD Contínua, para as oito variáveis analisadas. Já as análises feitas para investigar a consistência do estimador de variância do método CP sugerem que pode haver viés importante para amostras muito pequenas ou muito grandes, dependendo da variável.

Estudos adicionais devem ser realizados para avaliar o desempenho do estimador de variância do método CP para outros parâmetros e variáveis de interesse, tais como razões, por exemplo.

Referências

ARANTES, S. B.; SILVA, P. L. N. Planejamento de amostras domiciliares no Brasil explorando a malha setorial do Censo Demográfico 2010. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 74, n. 239, p. 101-130, jul./dez. 2013. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

CENSO demográfico 2010. Rio de Janeiro: IBGE, [2017]. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm>. Acesso em: set. 2017.

COSTA, G.T. L. da. *Coordenação de amostras PPT em pesquisas repetidas, utilizando o método de amostragem de Pareto*. 2007. 94 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro. 2007.

FREITAS, M. P. S. de et al. *Amostra mestra para o sistema integrado de pesquisas domiciliares*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2007. 67 p. (Textos para discussão, n. 23). Disponível em: <http://www.ibge.gov.br/home/estatistica/indicadores/sipd/texto_discussao_23.pdf>. Acesso em: set. 2017.

FREITAS, M. P. S. de; ANTONACI, G. de A. *Sistema integrado de pesquisas domiciliares*: amostra mestra 2010 e a amostra da PNAD contínua. 36 p. (Texto para discussão. Diretoria de Pesquisas, n. 50). Rio de Janeiro: IBGE, 2014. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv86747.pdf>. Acesso em: set. 2017.

HANSEN, M. H.; HURWITZ, W. N.; MADOW, W. G. *Sample survey methods and theory*. New York: Wiley, 1953. 2 v. (Wiley publications in statistics).

LILA, M. F. *Estimação de variâncias em pesquisas amostrais domiciliares*. 2004. 121 f. Dissertação (Mestrado)-Programa de Mestrado em Estudos populacionais e Pesquisas Sociais, Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2004.

PESQUISA nacional por amostra de domicílios contínua: notas metodológicas. v. 1. Rio de Janeiro: IBGE, 2014. 47 p. Disponível em: <ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf>. Acesso em: set. 2017.

PESSOA, D. G. C.; SILVA, P. L. do N. Análise de dados amostrais complexos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 13., 1998, Caxambu. *Anais...* São Paulo: Associação Brasileira de Estatística - ABE, 1998. 170 p. Disponível em: <http://www.ie.ufrj.br/download/livro.pdf>. Acesso em: set. 2017.

ROSÉN, B. A user's guide to Pareto πps *sampling*, *R&D Report*: research, methods, development, Stockholm: Statistiska centralbyrån, n. 6, Dec. 2000. 10 p. Disponível em: <http://www.scb.se/contentassets/14f5e346f4814dd0acd52d10b23286c6/rnd-report-2000-06-green.pdf>. Acesso em: set. 2017.

_____. On sampling with probability proportional to size, *Journal of Statistical Planning and Inference*, Amsterdam: Elsevier, v. 62, n. 2, p. 159-191, 1997.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. H. *Model assisted survey sampling*. New York: Springer-Verlag, 1992. 694 p. (Springer series in statistics).

Método de otimização aplicado à estratificação de unidades primárias de amostragem

José André de Moura Brito

Introdução

Este capítulo apresenta um novo método de otimização que foi desenvolvido para a resolução de um problema de amostragem probabilística de grande relevância no âmbito do IBGE, qual seja, a estratificação de unidades primárias de amostragem, doravante denotadas por UPAs.

Este problema está associado à fase de planejamento de pesquisas domiciliares por amostragem, para integração de algumas pesquisas do IBGE. Mais especificamente, considera-se o uso de um mesmo cadastro de seleção e de uma amostra em comum, denominada amostra mestra. Esta amostra corresponde a um conjunto de unidades de área selecionadas de um cadastro, segundo um método probabilístico de seleção, a partir do qual seja possível selecionar subamostras para atender diversas pesquisas. No que concerne à amostra mestra, as UPAs correspondem a agregados de setores censitários que, de acordo com Freitas e outros (2007), são definidos considerando critérios específicos.

Em uma fase posterior, são definidos os estratos estatísticos, sendo cada estrato formado por um número mínimo de UPAs. Ainda neste sentido, a definição da alocação das UPAs, aos seus respectivos estratos, tem por base a avaliação de uma expressão de variância. Destarte, quanto menor o valor dessa expressão, mais homogêneos são considerados os estratos.

Por sua vez, a definição dos estratos está associada à resolução de um problema clássico de otimização de alta complexidade computacional, denominado problema de agrupamento com soma mínima ou, de acordo com Hansen e Jaumard (1997), clique de soma mínima.

Esta associação entre esses dois problemas restringe a aplicação de métodos de enumeração exaustiva ou implícita para a resolução do problema de estratificação de UPAs. Isto também sinaliza que não existe nenhum algoritmo que produza o ótimo global (estratos com a menor variância possível) em tempo computacional factível, mesmo considerando todo progresso quanto à aplicação de métodos do tipo Branch e Bound, Plano de Cortes e Programação Dinâmica (WOLSEY, 1998), para resolver diversos problemas de otimização combinatória.

Considerando a necessidade de produzir soluções de boa qualidade para este problema, em tempo computacional factível, apresentamos, neste trabalho, um algoritmo que foi desenvolvido com base no estudo de um método de otimização denominado Greedy Randomized Adaptive Search Procedure - GRASP (RESENDE; RIBEIRO, 2014).

Além desta **Introdução**, nos tópicos seguintes são apresentados, respectivamente, os conceitos básicos de análise de agrupamentos, uma descrição detalhada do problema de estratificação de UPAs, o método de otimização que foi considerado para resolução do problema, bem como o algoritmo proposto a partir do estudo desse método. Concluindo o trabalho, no último tópico são apresentadas algumas considerações derivadas da análise de um conjunto de experimentos computacionais realizados com dados associados à base operacional geográfica do Censo Demográfico 2010.

Análise de agrupamentos

De acordo com Hair e outros (2009) e Kaufman e Rousseeuw (1990), a análise de agrupamentos é uma técnica de análise multivariada que combina um conjunto de métodos que são utilizados, com o objetivo, de agrupar os n objetos de uma base dados em k grupos. Ainda segundo Hair e outros (2009, p. 430):

> A análise de agrupamentos classifica objetos de modo que cada objeto é semelhante aos outros no agrupamento com base em um conjunto de características escolhidas. Os agrupamentos resultantes de objetos devem exibir elevada homogeneidade interna (dentro dos agrupamentos) e elevada heterogeneidade externa (entre agrupamentos).

Atualmente, a análise de agrupamentos é utilizada nas mais variadas áreas do saber (HAIR et al., 2009) como, por exemplo: Estatística, Inteligência Artificial, Computação, Economia, Marketing, Sociologia, Medicina etc.

Basicamente, os métodos de análise de agrupamento são concebidos com vistas à resolução do problema de agrupamento clássico: Dado um conjunto X formado por n objetos, $X = \{x_1, x_2,...,x_n\}$ com s atributos (variáveis), tal que $x_i=(x_{i1},x_{i2},...,x_{is})$, e sendo o número de grupos igual a um inteiro k, deve-se construir k grupos $G_1, G_2,..., G_k$, tal que sejam observadas as três restrições abaixo:

$$G_1 \cup G_2 \cup ... \cup G_k = X \qquad (1)$$

$$G_i \cap G_j = \varnothing \quad \text{i, j} = 1,\ldots,\text{k} \ , \text{i} < \text{j} \qquad (2)$$

$$|G_i| \geq 1 \qquad \text{i} = 1,\ldots,\text{k} \qquad (3)$$

A restrição (1) indica que a união dos grupos corresponde ao conjunto X, a restrição (2) indica que um objeto pertence a exatamente um grupo e a restrição (3) garante que cada grupo tem, pelo menos, um objeto.

A construção solução de um agrupamento (partição), ou seja, a alocação dos objetos aos grupos, depende, basicamente, da medida de distância (associada às s variáveis) utilizada e da função objetivo $f(.)$ considerada para a avaliação da homogeneidade dos grupos, caracterizando, consequentemente, como agrupamento ótimo, aquele associado à solução que tem o menor valor de função objetivo.

A obtenção do agrupamento ótimo associado aos grupos mais homogêneos, segundo a função objetivo definida a priori, é uma tarefa computacionalmente difícil. A explicação para isso é que o número soluções possíveis para este problema é impactado, diretamente, pelo número de objetos (n) da base de dados associada à aplicação. Ou seja, o número de soluções possíveis para o problema de agrupamento definido a partir das restrições (1), (2) e (3) pode ser calculado a partir do número de *Stirling* de segundo tipo, conforme Johnson e Wichern (2002). Por exemplo, se n=30 e k=2, o número de soluções a serem consideradas é de 536.870.911. Mantendo o mesmo número de grupos, e apenas dobrando o número de objetos, há mais de meio quatrilhão de soluções possíveis. Esses valores crescem exponencialmente à medida que n aumenta.

Dessa forma, a aplicação de um método enumeração exaustiva, que avalie todas soluções deste problema, ou seja, todas as possíveis alocações dos objetos aos grupos, e determine a solução ótima, não é algo factível computacionalmente. Ainda em relação a esta questão, em alguns problemas de agrupamento, como os descritos nos trabalhos de Vinod (1969), Rao (1971), Hansen e Jaumard (1997), Cruz (2010), Brito, Semaan e Brito (2015), pode-se utilizar formulações de programação matemática. A aplicação dessas formulações é viável, apenas, para problemas de pequeno porte (n<100), tendo em vista que o número de variáveis e/ou restrições destas formulações cresce substancialmente à medida que o número de objetos aumenta. Por sua vez, isto impacta diretamente no tempo de processamento consumido para resolver a formulação e produzir o ótimo global, mediante a utilização de algum *solver* de otimização.

Levando em conta a complexidade desses problemas, busca-se, através de algoritmos heurísticos, a produção de soluções viáveis que sejam de boa qualidade, em relação à coesão dos grupos, e que demandem baixo tempo computacional, quando comparado ao tempo demandado por uma formulação de programação matemática ou método de enumeração.

Considerando a associação entre as inúmeras aplicações desta técnica e todos os avanços computacionais, nas últimas décadas, tem sido proposta uma ampla gama de novos algoritmos heurísticos, como, por exemplo, nos trabalhos de Brito, Semaan e Brito (2015), Brusco e Steinley (2007), Carrizosa, Mladenovic e Todosijevic (2013), Cruz (2010), Friedman e Meulman (2004), Naldi (2011), Santia, Aloise e Blanchardc (2016) e Semaan (2013).

Problema de estratificação de unidades primárias de amostragem

O IBGE trabalha com pesquisas que têm um desenho amostral que incorpora, em pelo menos um de seus estágios, as estratificações geográfica e estatística (COCHRAN, 1977). A determinação dos estratos estatísticos, por conseguinte, tem uma ligação direta com problemas de agrupamento de difícil solução. Esta dificuldade é decorrente de uma substancial quantidade de dados que devem ser agrupados e da utilização de restrições particulares para cada tipo de estratificação estatística, conforme comentado em Brito e Montenegro (2010).

Um bom exemplo da utilização estratificação estatística diz respeito ao sistema de pesquisas domiciliares por amostragem para integração de algumas pesquisas do IBGE, considerando o uso de um mesmo cadastro de seleção e de uma amostra em comum, denominada amostra mestra. Segundo Freitas e outros (2007), esta amostra corresponde a um conjunto de unidades de área selecionadas de um cadastro, segundo um método probabilístico de seleção, a partir do qual seja possível selecionar subamostras para atender às diversas pesquisas. Estas subamostras podem ser selecionadas de forma independente ou com certo controle para que tenham ou não algumas unidades coincidentes.

A população alvo da amostra mestra inclui toda a população a ser investigada em todas as pesquisas, sendo constituída pelos moradores residentes em todos os domicílios na área que constitui a abrangência geográfica definida a seguir. A abrangência geográfica da amostra mestra considera o âmbito das diversas pesquisas que farão uso dessa amostra comum. Assim sendo, não se pode deixar de incluir qualquer parte do território que seja contemplado por algumas das pesquisas.

A abrangência geográfica da amostra mestra é constituída pelos setores censitários da Base Operacional Geográfica de todo o território nacional.

Um importante aspecto para seleção desta amostra diz respeito à definição de suas unidades primárias de amostragem (UPAs). Mais especificamente, em pesquisas domiciliares, estas unidades são definidas, em geral, por unidades de área com um determinado tamanho mínimo populacional calculado em termos de domicílios ou pessoas. As UPAs podem ser definidas como sendo as unidades básicas do cadastro mestre, ou podem corresponder a agregações contíguas destas como, por exemplo, uma divisão administrativa.

Como na Base Operacional Geográfica de 2010 haviam muitos setores censitários pequenos, foi realizada uma agregação de setores censitários para a composição das UPAs, de tal modo que estas possuíssem um número de domicílios suficientes para atender a demanda das pesquisas a serem integradas, reunidas no chamado Sistema Integrado de Pesquisas Domiciliares - SIPD, do IBGE.

De acordo com Freitas e outros (2007) e Brito, Montenegro e Freitas (2011), no que diz respeito à amostra mestra, cada uma das UPAs é composta por setores censitários contíguos e cuja soma dos seus Domicílios Particulares Permanentes - DPPs é maior ou igual 60.

Definidas as UPAs, deve ser efetuada uma estratificação estatística, o que implica agregar as UPAs em estratos que sejam homogêneos. Esta homogeneidade é avaliada a partir do cálculo de uma expressão de variância que incorpora a renda total dos domicílios e o total de domicílios particulares permanentes (DPPs), associados aos setores censitários de cada uma das UPAs.

Adicionalmente, cada estrato deve ter um número mínimo de 150 UPAs e a alocação das UPAs aos estratos é realizada de forma a minimizar a variância do estimador de total da renda domiciliar, considerando o plano amostral normalmente utilizado nas pesquisas domiciliares, qual seja, a amostragem conglomerada (COCHRAN, 1977) com seleção de UPAs com probabilidade proporcional ao número de DPPs.

Posto isso, busca-se minimizar a seguinte expressão de variância dentro de cada um dos estratos E_h ($h=1,...,L$):

$$V_h = \sum_{\forall i,j \in E_h} d_{ij} = N_i N_j \left(\frac{Y_i}{N_i} - \frac{Y_j}{N_j} \right)^2, \quad h=1,2,\ldots,L \qquad (4)$$

sendo N_i = número de DPPs na i-ésima UPA, N_j = número de DPPs na j-ésima UPA, Y_i = renda total domiciliar na i-ésima UPA e Y_j = renda total domiciliar na j-ésima UPA.

A menos da restrição de capacidade, associada ao número mínimo de UPAs por estrato, o problema de estratificação definido acima corresponde a um conhecido problema da literatura, qual seja, o problema de agrupamento com soma mínima de distâncias. Este problema é abordado, por exemplo, nos trabalhos de Brito e Brito (2008), Hansen e Jaumard (1997) e Nascimento, Toledo e Carvalho (2010).

Uma primeira alternativa à resolução deste problema consiste na adaptação e aplicação da formulação de programação inteira mista, apresentada nos trabalhos de Brito e Brito (2008) e Nascimento, Toledo e Carvalho (2010).

$$\text{Minimizar} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} x_{ij} \qquad (5)$$

$$\text{Sujeito a} \sum_{i=i}^{L} y_i^h = 1, \quad i=1,\ldots,n \qquad (6)$$

$$\sum_{i=1}^{n} y_i^h \geq 150, \quad h=1,...,L \qquad (7)$$

$$y_i^h + y_j^h - 1 \leq x_{ij}, h=1,\ldots,L, i=1,\ldots,n-1, j=i+1,\ldots,n \qquad (8)$$

$$x_{ij} \geq 0, i=1,\ldots,n-1, j=i+1,\ldots,n \qquad (9)$$

$$y_i^h \in \{0,1\}, i=1,\ldots,n, h=1,\ldots,L \qquad (10)$$

A função minimizada em (5) corresponde à função objetivo definida na equação (4). A variável binária y_i^h assume valor 1 quando a i-ésima UPA é alocada ao h-ésimo estrato

E_h ($h=1,...,L$) e zero caso contrário. A restrição (6) garante que cada UPA i ($i=1,...,n$) deve ser alocada a exatamente um dos estratos E_h (h=1,...,L), a restrição (7) certifica que cada estrato Eh tem, pelo menos, 150 UPAs. As restrições (8) e (9) garantem que x_{ij} assume valor um se, simultaneamente, y_i^h e y_j^h assumem valor 1, ou seja, se as UPAs i e j estão alocadas a um mesmo estrato, o que implica contabilizar a distância d_{ij} entre das UPAs no mesmo estrato.

De acordo com o número de UPAs (n), esta formulação pode ter um substancial número de variáveis e restrições, fato que restringe a sua aplicação a problemas de pequeno porte, ou seja, bases de dados com poucas UPAs (da ordem centenas). Todavia, mesmo considerando este caso, e disponibilizando um tempo de processamento razoável (da ordem de horas, dias ou até meses) à aplicação da formulação, o máximo que se obtém para este tipo de problema são soluções correspondentes a ótimos locais.

Por sua vez, em muitos casos, estes ótimos locais não são de boa qualidade, no que diz respeito à homogeneidade dos estratos ou, de forma mais geral, à homogeneidade dos grupos.

Esta observação é ratificada nos trabalhos de Brito e Brito (2008), Brito, Montenegro e Freitas (2011), e Nascimento, Toledo e Carvalho (2010). Nestes três trabalhos são propostos algoritmos heurísticos para o problema de agrupamento com soma mínima de distâncias. Em particular, o trabalho de Brito, Montenegro e Freitas (2011) traz a proposta de dois algoritmos heurísticos que foram utilizados, à época, para efetuar a estratificação das UPAS da amostra mestra, sejam eles: um algoritmo baseado na metaheurística ILS (GLOVER; KOCHENBERGER, 2003) e outro no algoritmo de otimização microcanônica (MONTENEGRO; TORREÃO; MACULAN, 2003), doravante denotado algoritmo OM.

Método de otimização

Segundo Resende e Ribeiro (2014), o GRASP consiste em um método iterativo que poder ser aplicado para resolver diversos problemas de otimização e que combina dois procedimentos, a saber: construção e busca local.

Em linhas gerais, durante m iterações desse método, aplica-se o procedimento de construção, seguido do procedimento de busca local. A finalidade do procedimento de construção é produzir uma solução (ótimo local) so que será melhorada, mediante a aplicação do procedimento de busca local, que tende a produzir uma solução s* de qualidade superior à solução s_o no que diz respeito ao valor da função objetivo (f(.)). A melhor solução obtida após as m iterações do GRASP será a solução do problema de otimização em questão. A Figura 1 ilustra a aplicação do GRASP, considerando um problema de minimização.

Figura 1 – Algoritmo GRASP

```
f_melhor ← +∞    (m = número de iterações)
PARA k ← 1 ATÉ m FAÇA
    s_o ← construção()
    s* ← busca_local(s_o)
    SE (f(s*) < f_melhor) ENTÃO
        s_melhor ← s*    (solução do problema)
        f_melhor ← f(s*) (valor da função objetivo)
    FIM-SE
FIM-PARA
RETORNE s_melhor, f_melhor
```

Elaboração do Autor com base no algoritmo GRASP

A construção de cada solução s_o é baseada uma lista de candidatos (LC) formada por todos os elementos que, se incorporados em s_o, produzem uma solução viável. Definida a LC, deve-se avaliar todos os seus elementos através de uma função g(.), que permite determinar o custo de se adicionar um novo elemento $i \in LC$ na solução s_o. Ainda neste sentido, de forma a dar maior variabilidade e qualidade às soluções que serão produzidas na fase de construção, deve-se definir uma lista de candidatos restrita (LCR), formada pelos melhores elementos avaliados na LC, através de uma função g(.).

Para a construção da LCR, em cada iteração da fase de construção, denotam-se, respectivamente, por g_{min} e g_{max} os menores e maiores acréscimos produzidos pela inserção de um elemento $i \in LC$ em s_o segundo a avaliação de g(.). A partir da aplicação dessa função, e da utilização dos valores g_{min} e g_{max}, define-se:

$$LCR = \left\{ i \in LC \mid g(i) \leq g_{min} + \alpha \left(g_{max} - g_{min} \right),\ \alpha \quad [0,1] \right\} \qquad (11)$$

O parâmetro α determina quais elementos da LC serão adicionados à LCR.

Algoritmo proposto

O presente tópico traz a proposta do algoritmo GRASP que foi implementado para resolução do problema de estratificação das UPAS. É feita uma descrição da forma de representação de uma solução viável para o problema, do cálculo da função objetivo e dos procedimentos de construção e de busca local que foram adotados.

Representação das soluções e função objetivo

Um primeiro passo à implementação do algoritmo heurístico diz respeito à representação das soluções viáveis para o problema de otimização em questão. Levando em conta que o problema de estratificação das UPAs corresponde a um problema de agrupamento, cada solução viável para este problema foi representada por um vetor s com n posições correspondentes ao número de UPAS e cada posição desse vetor contém um valor entre 1 e L, que indica a qual estrato a i-ésima UPA está inicialmente alocada.

No que diz respeito à função objetivo, de forma a otimizar o seu cálculo, constrói-se, a priori, uma matriz D_{nxn} que contém, em cada uma de suas entradas d_{ij}, o valor correspondente à distância entre duas UPAs *i* e *j* (vide equação 4), supondo que estas UPAs estejam alocadas a um mesmo estrato.

Considerando a matriz D e uma solução viável denotada por s, avalia-se quais são as UPAs que estão em cada um dos L estratos E_1, E_2,...,E_L e calcula-se a soma das distâncias d_{ij} entre todas as UPAs, tomadas duas a duas, dentro de cada um dos estratos. Considerando um pequeno exemplo hipotético com L=2 e n=7, temos que s={1,2,1,1,2,1,2} corresponde a uma solução para o problema. O estrato 1 (E_1) é formado pelas UPAs 1,3,4,6 e o estrato 2 (E_2) é formado pelas UPAs 2, 5 e 7. Assim sendo, os valores da função objetivo nos estratos 1 e 2 seriam, respectivamente, dados pelas seguintes somas: $d_{13}+d_{14}+d_{16}+d_{34}+d_{36}+d_{46}$ e $d_{25}+d_{27}+d_{57}$.

Procedimento de construção

De forma a produzir um vetor solução s, no primeiro passo do procedimento de construção são selecionados L.k (k inteiro positivo entre 2 e 3) valores entre 1 e n, sendo esses valores correspondentes às UPAs que, inicialmente, serão alocadas aos L estratos E_1, E_2,...,E_h,...,E_L, de forma que $|E_h|=k$ (k=1,...,L).

As demais r=(n-L.k) UPAs são alocadas aos estratos de forma iterativa, utilizando a lista de candidatos (LC) e lista de candidatos restrita (LCR). Particularmente, define-se como U o conjunto constituído pelas r UPAs que ainda não foram alocadas, tal que: $U=\{u_1,u_2,...,u_r\}$ (U corresponde à lista de candidatos - LC). Para cada uma das UPAS $u_i \in U$, calcula-se, o valor da função $g(u_i)$, sendo $g(u_i)=mínimo_{1,....,L}(du_iE_h)$ (distância (variância) entre a UPA u_i e as UPAs alocadas, até o presente momento, a cada um dos

estratos E_h (h=1,...,L). Ou seja, g(u_i) corresponde ao menor incremento que pode ser obtido em um dos estratos, mediante a inclusão da UPA u_i.

Em um passo posterior, considerando cada um dos elementos de U e o cálculo de g(u_i), define-se um vetor g=(gu_1,gu_2,...,gu_i,...,gu_r) e um vetor a=(a_1,a_2,...,a_i,...,a_r). O vetor a está associado ao vetor g e contém, em cada uma de suas posições, um valor entre 1 e L correspondente ao estrato mais próximo E_h que UPA ui pode ser alocada. Em seguida, calcula-se o máximo e mínimo deste vetor (g_{max} e g_{min}) e define-se a lista de candidatos restrita (LCR) que corresponde a um subconjunto de U, denotado por U', tal que:

$$U' = \left\{ u_i \in U \mid g(u_i) \leq g_{min} + \alpha\left(g_{max} - g_{mim}\right) \right\} \qquad (12)$$

U' contém um subconjunto de UPAs que correspondem às melhores UPAS, isto, àquelas que produzem o menor acréscimo à variância do estimador de total da renda domiciliar definido na equação (4).

Finalmente, seleciona-se aleatoriamente de U', um elemento u_i e aloca-se ao estrato E_h (de acordo com o valor de a_i) que tem o número de UPAs inferior a 150 (critério de capacidade). Feito isso, atualiza-se o conjunto U (exclui-se de U a UPA u_i) e repete-se o cálculo de g(u_i) considerando as (r-1) UPAs de U. Este procedimento é aplicado r vezes até todas as UPAs tenham sido alocadas aos estratos. Cabe observar que, a partir do momento que todos os estratos têm, pelo menos, 150 UPAs, a alocação considera, apenas, o critério de homogeneidade (redução da variância).

Uma vez aplicado o procedimento de construção, produz-se um vetor solução s_o que contém a informação da alocação das n UPAs aos L estratos. Sobre este vetor são aplicados dois procedimentos de busca local denotados, respectivamente, por busca_local1 (BL1) e busca_local2 (BL2).

Procedimentos de busca local

O procedimento BL1 consiste, basicamente, em realocar cada uma das UPAs ao seu estrato mais próximo. Ou seja, partindo da solução s_o produzida a partir da aplicação do procedimento de construção, toma-se cada uma das n UPAs (u_1,u_2,...,u_i,...,u_n) e avalia-se a sua distância a cada um dos L estratos, mediante o seguinte cálculo: d_{ui}=mínimo$_{h=1,...,L}$($d_{ui}E_h$) (i=1,...,n).

Caso a distância d_{ui} seja inferior à distância da UPA u_i ao estrato E_h ao qual esta UPA está atualmente alocada (caracterizando redução na variância), troca-se a UPA de estrato e atualiza-se o vetor s_o, produzindo um vetor solução denotado por s'. Essa alocação é feita levando em conta a restrição do número mínimo de UPAs por estrato.

O procedimento BL2 trabalha de forma similar ao procedimento BL1, efetuando a troca de UPAs entre os estratos de forma a reduzir o valor da variância. Em linhas gerais, tomando a solução s' produzida a partir da aplicação do procedimento BL1, são selecionados, aleatoriamente, dois estratos E_q e E_t (q,t$\in$ {1,2,...,L}) e, seleciona-se de cada um desses estratos, uma amostra aleatória (subconjunto de UPAs) de tamanho n' (n'$\in${2,3,...,25}). Em seguida, as UPAs associadas à amostra de E_q são trocadas com as UPAs associada à amostra de E_t e vice-versa, definindo um novo vetor s'', ou seja, implicando nova configuração para os estratos. Finalmente, calcula-se o valor da função objetivo associada ao vetor s''. Caso haja redução no valor de f(s'') em relação ao valor de f(s'), atualiza-se s' com s'' e atualiza-se o valor da função objetivo. Este procedimento de seleção de estratos e seleção de amostra de UPAs para troca é repetido por p vezes, sendo p=2.000.

Experimentos computacionais

Com o objetivo de avaliar o algoritmo GRASP, foram realizados experimentos computacionais considerando a aplicação desse algoritmo e do algoritmo OM[1] proposto

[1] Foi escolhido o algoritmo OM em detrimento ao algoritmo ILS, tendo em vista que, segundo Brito, Montenegro e Freitas (2011), o algoritmo OM produziu soluções de melhor qualidade quando comparado ao ILS.

por Brito, Montenegro e Freitas (2011), além da formulação apresentada na seção que traz a descrição do problema de estratificação de UPAs.

A formulação foi implementada utilizando o *software* LINGO (versão 14.0) e os algoritmos GRASP[2] e OM foram implementados em linguagem R. No que diz respeito ao algoritmo GRASP, os parâmetros foram os seguintes: m=50 e α=0.1. Todos os experimentos foram efetuados em um computador com 16GB de memória RAM e dotado de seis processadores AMD FX-6300 de 3.5 GHz e sistema operacional Windows 7 (64 bits).

De forma a avaliar a performance do algoritmo GRASP frente ao algoritmo OM, no que diz respeito à qualidade das soluções produzidas, foi realizado um conjunto de experimentos computacionais com 50 problemas teste. Esses problemas estão associados a 50 arquivos contendo, cada um, uma lista de UPAs com seus respectivos números de domicílios e rendas totais.

Estes arquivos correspondem a um subconjunto de todos os arquivos da amostra mestra, observando que o menor arquivo processado continha 302 UPAs e o maior 1.857 UPAs. Além disso, o número de estratos construídos variou entre 2 e 5 e número mínimo de UPAs por estrato foi de 150.

Em relação à formulação, mesmo disponibilizando um tempo de processamento de oito horas, os resultados produzidos quanto ao valor da função objetivo foram de qualidade bem inferior aos resultados produzidos pelos algoritmos GRASP e OM.

No que diz respeito aos dois algoritmos, foi observada uma pequena vantagem do algoritmo GRASP em relação ao algoritmo OM. Mais especificamente, o algoritmo GRASP produziu soluções com valor de função objetivo (variância) menor ou igual ao valor da função objetivo observado no algoritmo OM, implicando estratos mais homogêneos. Em relação à eficiência, os tempos de processamento demandados pelo algoritmo GRASP para resolver os 50 problemas teste variaram entre 2 e 39 minutos, ficando próximos dos tempos demandados pelo algoritmo OM.

Neste trabalho foi apresentado um método de otimização para resolver um problema de alta relevância no âmbito do IBGE, qual seja, o problema de estratificação de unidades primárias de amostragem. Conforme comentado, este problema está associado a um problema de otimização de alta complexidade computacional. Por conta disso, foi desenvolvido um algoritmo heurístico baseado em método de otimização denominado GRASP. No que diz respeito aos experimentos computacionais reportados nesta seção, este algoritmo produziu soluções de boa qualidade às expensas de um tempo computacional aceitável.

Os resultados satisfatórios apresentados para instâncias de porte variado (número de UPAS entre centenas e milhares) indicam que o algoritmo apresentado neste trabalho pode ser uma boa alternativa à resolução do problema de estratificação de UPAs.

Referências

BRITO, J. A. de M.; BRITO, L. R. Algoritmos VNS e genéticos aplicados ao problema de agrupamento com soma mínima de distâncias. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 40., 2008, João Pessoa, *Anais...* Rio de Janeiro: Sociedade Brasileira de Pesquisa Operacional - Sobrapo, 2008. p. 1150-1161. Disponível em: <http://www.din.uem.br/sbpo/sbpo2008/pdf/arq0167.pdf>. Acesso em: set. 2017.

BRITO, J. A. de M.; MONTENEGRO, F. M. T. Um algoritmo VNS aplicado ao problema de definição de áreas de ponderação. In: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 13., 2010, Rio de Janeiro. *Anais...*, Rio de Janeiro: Marinha do Brasil, 2010. Disponível em: <https://www.marinha.mil.br/spolm/sites/www.marinha.mil.br.spolm/files/74163.pdf>. Acesso em: set. 2017.

[2] Os procedimentos de construção e busca local deste algoritmo foram paralelizados utilizando a pacote snowfall do R.

BRITO, J. A. de M.; MONTENEGRO, F. M. T; FREITAS, M. P. S. *Algoritmos de otimização aplicados à estratificação da amostra mestra*. Trabalho apresentado na III Escola de Amostragem e Metodologia de Pesquisa - ESAMP, realizado em Juiz de Fora, 2011.

BRITO, J. A. de M.; SEMAAN, G. da S.; BRITO, L. R. Resolução do problema dos k-medoids via algoritmo genético de chaves aleatórias viciadas. *Revista Pesquisa Naval*, Brasília: Secretaria de Ciência, Tecnologia e Inovação da Marinha, v. 27, p. 126-142, 2015. Disponível em: <https://www.marinha.mil.br/dgdntm/sites/www.marinha.mil.br.dgdntm/files/flipping_book/rpn27/index.html#overlay-context=informativo>. Acesso em: set. 2017.

BRUSCO, M.J.; STEINLEY, D. A comparison of heuristic procedure for minimum within-cluster sums of squares partitioning. *Psychometrika*, New York: Springer-Verlag, v. 72, n. 4, p. 583–600, 2007.

CARRIZOSA, E.; MLADENOVIC, N.; TODOSIJEVIC, R. Variable neighborhood search for minimum sum-of-squares clustering on networks. *European Journal of Operational Research*, New York: Elsevier, v. 230, n. 2, p. 356-363, 2013.

COCHRAN, W. G. *Sampling techniques*. 3rd ed. New York: Wiley, c1977. 428 p. (Wiley series in probability and mathematical statistics). Disponível em: <https://archive.org/details/Cochran1977SamplingTechniques_201703>. Acesso em: set. 2017.

CRUZ, M. D. O *Problema de clusterização automática*. 2010. 120 p. Tese (Doutorado em Engenharia de Sistemas e Computação) - Coordenação dos Programas de Pós-Graduação de Engenharia - Coppe, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010. Disponível em: <http://objdig.ufrj.br/60/teses/coppe_d/MarceloDibCruz.pdf>. Acesso em: set. 2017.

FREITAS, M. P. S. de et. al. *Amostra mestra para o sistema integrado de pesquisas domiciliares*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2007. 67 p. (Textos para Discussão, n. 23).

FRIEDMAN, J. H.; MEULMAN J. J. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society. Series B Methodological*, London, v. 66, n. 4, p. 815-849, 2004. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2004.02059.x/full>. Acesso em: set. 2017.

GLOVER, F. E.; KOCHENBERGER, G. A. (Ed.). *Handbook of metaheuristic*. Norwell: Kluwer Academic Publishers, 2003. 557 p.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data*: an introduction to cluster analysis. New York: Wiley, 1990.

HAIR, J. F. et al. Análise multivariada de dados. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

HANSEN, P.; JAUMARD, B. Cluster analysis and mathematical programming. *Mathematical Programming*, New York: Springer-Verlag, v. 79, n. 1-3, p. 191-215, 1997.

JOHNSON, A. R.; WICHERN, D. W. *Applied multivariate statistical analysis*. 5. ed. Englewood: Prentice Hall, 2002. 767 p.

MONTENEGRO, F.; TORREÃO, J. R. A.; MACULAN, N. Microcanonical optimization algorithm for the Euclidean Steiner problem in R^n with application to phylogenetic inference, *Physical Review E*, New York: American Physical Society, v. 68, n. 5, nov. 2003.

NALDI, M. C. Técnicas de combinação para agrupamento centralizado e distribuído de dados. 2011. 245 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional)-Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo - USP, São Carlos, 2011. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-16032011-113154/pt-br.php>. Acesso em: set. 2017.

NASCIMENTO, M. C. V.; TOLEDO, F. M. B.; CARVALHO, A. C. P. L. F. de. Investigation of a new GRASP-based clustering algorithm applied to biological data. *Computers & Operations Research*, New York: Pergamon Press, v. 37, n. 8, p. 1381-1388, 2010.

RAO, M. R. Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, American Statistical Association, Alexandria [Estados Unidos], v. 66, n. 335, p. 622-626, Sep. 1971.

RESENDE, M. G. C.; RIBEIRO C. C. GRASP: greedy randomized adaptive search procedures. In: BURKE, E.; KENDALL, G. (Ed.). *Search methodologies*. 2nd. Boston: Springer, 2014. p. 287-310.

SANTI, E.; ALOISE, D.; BLANCHARD, S. J. A model for clustering data from heterogeneous dissimilarities. *European Journal of Operational Research*, New York: Elsevier, v. 253, n. 3, p. 659-672, 2016.

SEMAAN, G. S. *Algoritmos para o problema de agrupamento automático*. 2013. 184 p. Tese (Doutorado)-Departamento de Ciência da Computação, Universidade Federal Fluminense - UFF, Niterói, 2013.

VINOD, H. D. Integer programming and theory of grouping. *Journal of the American Statistical Association*, American Statistical Association, Alexandria [Estados Unidos], v. 64, n. 326, p. 506-517, 1969.

WOLSEY, L. A. *Integer programming*. New York: Wiley, 1998. 288 p.

Indicadores de pobreza nos municípios de Minas Gerais: comparação de métodos de estimação em pequenas áreas

Nícia Custódio Hansen Brendolin
Debora Ferreira de Souza
Viviane Cirillo Carvalho Quintaes
Djalma Galvão Carneiro Pessoa
Solange Correa Onel

Introdução

Institutos de estatísticas oficiais cada vez mais se deparam com grande demanda por informações detalhadas e precisas, por outro lado sofrem com a constante restrição financeira na produção das pesquisas amostrais. Neste sentido, pesquisadores de vários órgãos produtores de estatísticas oficiais, atualmente, estudam metodologias de estimação em pequenos domínios a fim de fornecer estimativas para áreas geográficas ou domínios menores com precisão controlada, sem aumento de custos.

Para estimar pobreza em pequenas áreas, são utilizadas metodologias que combinam as informações coletadas em pesquisas domiciliares amostrais de múltiplos propósitos com a ampla cobertura geográfica dos Censos Demográficos.

Um método de estimação em pequenas áreas amplamente utilizado por institutos de estatística é o desenvolvido em Fay e Herriot (1979). Tal metodologia é baseada em modelo linear em nível de área com efeitos aleatórios. Já, o *Mapa de pobreza e desigualdade: municípios brasileiros 2003* publicado em 2008 pelo Instituto Brasileiro de Geografia e Estatística - IBGE (MAPA..., 2008) adotou o método, em nível de unidade, proposto por Elbers, Lanjouw e Lanjouw (2003) e teve por finalidade fornecer uma descrição detalhada da distribuição espacial da pobreza no país.

Recentemente, Molina e Rao (2010) desenvolveram uma metodologia, também em nível de unidade, na qual utilizam o máximo possível da informação amostral para estimação dos indicadores de pobreza nas pequenas áreas. Segundo os autores, esta proposta resulta em menores erros quadráticos médios quando comparados aos obtidos a partir do método apresentado por Elbers, Lanjouw e Lanjouw (2003).

O objetivo principal deste trabalho é comparar os três métodos de estimação em pequenas áreas citados anteriormente. Para o estudo, foram extraídas subamostras de uma população a partir do desenho amostral utilizado na *Pesquisa de orçamentos familiares 2008-2009: despesas, rendimentos e condições de vida* (2010) e implementaram-se as três metodologias nestas subamostras. A amostra do Censo Demográfico 2010 do Estado de Minas Gerais, excluindo-se os domicílios com rendimento total igual a zero, foi

considerada a população. Ao final, compararam-se as estimativas da incidência de pobreza obtidas pelas metodologias a partir das subamostras com aquelas calculadas na população em termos de erro quadrático médio relativo, vício relativo e ordenação de áreas. Esta foi feita utilizando-se o coeficiente de correlação de postos de Spearman. Os resultados foram avaliados tanto para municípios quanto para microrregiões. Mais detalhes sobre o estudo podem ser encontrados no trabalho de Souza e outros (2014). Tais resultados podem fornecer subsídios para escolher uma metodologia adequada para estimação de indicadores necessários para a produção de mapa de pobreza no país.

Este capítulo está organizado em 5 tópicos. O tópico **Indicadores FGT e seus estimadores diretos** apresenta os indicadores usuais de pobreza e seus estimadores diretos. O tópico **Métodos para estimação de indicadores de pobreza em pequenas áreas** está subdividido em três partes. Na primeira, é apresentado o modelo descrito em Fay e Herriot (1979), na segunda parte, definem-se as especificações do método desenvolvido por Elbers, Lanjouw e Lanjouw (2003) e na terceira parte, descreve-se a metodologia proposta por Molina e Rao (2010). O tópico **Aplicação** detalha a forma de obtenção das subamostras usadas no experimento, os resultados obtidos e as possíveis comparações entre os métodos. Por fim, no tópico **Conclusões**, são apresentadas as considerações finais.

Indicadores FGT e seus estimadores diretos

Considere uma população finita de tamanho N particionada em D áreas de tamanhos. N_1,... N_D. Suponha que E_{dj} seja uma medida de bem-estar adequada do indivíduo j na pequena área d, tal como renda ou despesa, e seja z uma dada linha de pobreza; ou seja, o limiar para E_{dj} abaixo do qual uma pessoa é considerada pobre. A família de indicadores de pobreza, chamada FGT, introduzida por Foster, Greer e Thorbecke (1984), para cada pequena área d, é definida como:

$$FGT_{cd}=\frac{1}{N_d}\sum_{j=1}^{N_d} FGT_{cdj}=\frac{1}{N_d}\sum_{j=1}^{N_d}\left(\frac{z-E_{dj}}{z}\right)^c I\left(E_{dj}<z\right),\ c=0,1,2,\ d=1,\ldots,D$$

em que $I\left(E_{dj}<z\right)=1$ se $E_{dj}<z$ (pessoa em situação de pobreza) e $I\left(E_{dj}<z\right)=0$ se $E_{dj}\geqslant z$ (pessoa fora de situação de pobreza). Para *c=0* tem-se a proporção de indivíduos pobres na área *d*, também chamada de incidência de pobreza. A medida FGT quando *c=1* é chamada de hiato de pobreza, o qual é a proporção média em que a medida de bem-estar deve ser aumentada de forma a acabar com a pobreza. Quando a *c=2* medida é chamada de severidade da pobreza. Este indicador eleva ao quadrado os hiatos de pobreza enfatizando a pobreza extrema.

Os estimadores diretos dos indicadores de pobreza FGT são dados por:

$$fgt_{cd}^{w}=\frac{1}{\hat{N}_d}\sum_{j\in s_d} fgt_{cdj}^{w}=\frac{1}{\hat{N}_d}\sum_{j\in s_d} w_{dj}\left(\frac{z-E_{dj}}{z}\right)^c I\left(E_{dj}<z\right),\ c=0,1,2,\ d=1,\ldots,D. \quad (1)$$

em que W_{dj} é o peso amostral (inverso da probabilidade de inclusão) do indivíduo *j* da área *d* e $\hat{N}_d=\Sigma_{j\in s_d} w_{dj}$ é o estimador direto do tamanho populacional N_d da área *d*.

Tamanhos amostrais limitados em algumas áreas impedem o uso de estimadores como (1). Para obtenção de estimadores confiáveis para estas áreas torna-se necessária a utilização de técnicas de estimação em pequenas áreas (RAO, 2003). Tais técnicas

melhoram os procedimentos de estimação através de modelos que estabelecem alguma relação entre as áreas, baseados em informações auxiliares (provenientes do Censo Demográfico e/ou registros administrativos) relacionadas às variáveis de bem-estar de interesse. Estes modelos fornecem estimadores "indiretos" que fazem uso de dados de outras áreas correlatas, os quais podem reduzir drasticamente os erros de estimação.

Métodos para estimação de indicadores de pobreza em pequenas áreas

Método Fay e Herriot

O método proposto por Fay e Herriot (1979) consiste em um modelo linear com efeito aleatório de área o qual é útil nos casos em que as variáveis auxiliares estão disponíveis no nível de área ou quando não é possível ligar as informações das unidades amostrais com os microdados do Censo e com os registros administrativos.

Suponha que $\theta_d = g(\overline{Y}_d)$ seja uma função da medida $\overline{Y}_d$ para a qual se deseja obter uma estimativa e que esteja relacionada às variáveis auxiliares de área $\boldsymbol{x}_d = (x_{1d}, \ldots, x_{pd})'$ através do seguinte modelo linear:

$$\theta_d = \boldsymbol{x}_d'\boldsymbol{\beta} + v_d, \; d = 1, \ldots, D, \tag{2}$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é o vetor dos coeficientes da regressão. Além disso, os v_d*s* são efeitos aleatórios de área independentes e identicamente distribuídos com $E(v_d) = 0$ e $Var(v_d) = \sigma_v^2$. Frequentemente, assume-se a normalidade destes efeitos aleatórios.

Suponha que das *D* áreas existentes na população apenas *D** são selecionadas. Neste caso, assume-se o modelo da forma (2) para a população. Além disso, assume-se que estejam disponíveis os estimadores diretos $\hat{\overline{Y}}_d$ para as pequenas áreas, supondo a seguinte relação

$$\hat{\theta}_d = g(\hat{\overline{Y}}_d) = \boldsymbol{x}_d'\boldsymbol{\beta} + v_d + e_d, \; d = 1, \ldots, D^*,$$

na qual os erros amostrais e_d são independentes com $E(e_d | \theta_d) = 0$ e $Var(e_d | \theta_d) = \sigma_{e,d}^2$ conhecida.

Este modelo é estimado utilizando-se as informações das áreas amostradas e, para cada domínio, é obtido o seguinte estimador composto:

$$\theta_d^{EBLUP} = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d)\tilde{\theta}_d, \; d = 1, \ldots, D^*,$$

em que $\hat{\theta}_d$ é o estimador direto, $\tilde{\theta}_d = \boldsymbol{x}_d'\hat{\boldsymbol{\beta}}$ é o estimador sintético e $\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_{e,d}^2 + \hat{\sigma}_v^2}$, com $\hat{\sigma}_v^2$ e $\hat{\sigma}_{e,d}^2$ estimadores de σ_v^2 e $\sigma_{e,d}^2$, respectivamente. Nas áreas sem amostra é utilizado apenas o estimador sintético.

O estimador para σ_v^2 pode ser obtido através de:

$$E\left[\sum_d (\hat{\theta}_d - \boldsymbol{x}_d'\tilde{\boldsymbol{\beta}})^2 / (\sigma_{e,d}^2 + \sigma_v^2)\right] = E\left[h(\sigma_v^2)\right] = D^* - p,$$

em que $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\sigma_v^2)}$. O estimador $\hat{\sigma}_v^2$ é obtido resolvendo-se a equação $h(\sigma_v^2) = D^* - p$ de modo iterativo e supondo que $\hat{\sigma}_v^2 = 0$ quando não existe solução positiva.

Método Elbers, Lanjouw e Lanjouw

O método proposto por Elbers, Lanjouw e Lanjouw (2003) consiste em estimar, na mesma amplitude territorial do Censo Demográfico, uma informação que resuma o bem-estar social obtida a partir de uma pesquisa por amostragem. Dessa forma, é possível utilizar a ampla base geográfica do Censo para produzir estimativas dos principais índices de pobreza, com suas precisões, para níveis geográficos menores.

Define-se Y_{dj} como sendo função de uma variável de bem-estar do domicílio j pertencente à área d. O modelo pode ser escrito da seguinte forma:

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj},\ u_d \sim N(0, \sigma_u^2) \ \text{ e } \ e_{dj} \sim N(0, \sigma_{e,dj}^2) \qquad (3)$$

em que x_{dj} é um vetor contendo os valores observados das variáveis comuns no Censo e na pesquisa amostral, para o domicílio j da área d, j=1,..., n_d e d=1,..., D*. Note que o componente de erro é decomposto em duas partes: (i) uma associada ao domicílio e; (ii) outra associada ao nível geográfico (área) onde se localiza o domicílio (por exemplo, município, distrito, etc.), com u_d e e_{dj} independentes.

A estimativa da variância de u_d, $\hat{\sigma}_u^2$, é obtida regredindo-se os resíduos do modelo (3) nas variáveis indicadoras dos níveis de locação e então os parâmetros do modelo original são estimados através de mínimos quadrados generalizados.

Admitindo-se que os erros do nível domiciliar e_{dj} sejam heterocedásticos, Elbers, Lanjouw e Lanjouw (2003) sugerem estimar a regressão logística:

$$\ln\left(\frac{e_{dj}^2}{A - e_{dj}^2}\right) = \mathbf{z}'_{dj}\boldsymbol{\alpha} + r_{dj}, \qquad (4)$$

e estimam a variância no nível de domicílio segundo a fórmula:

$$\hat{\sigma}_{e,dj}^2 \approx \left[\frac{AB}{1+B}\right] + \frac{1}{2}\,var(r)\left[\frac{AB(1-B)}{(1+B)^3}\right],$$

onde $A = 1{,}05\,max\left\{e_{dj}^2\right\}$, $B = \exp\left\{\mathbf{z}'_{dj}\hat{\boldsymbol{\alpha}}\right\}$, *var (r)* é o erro quadrático dos resíduos da regressão logística estimada e $\mathbf{z}_{dj}$ é um vetor de variáveis explicativas.

Elbers, Lanjouw e Lanjouw (2003) derivaram uma estimativa da variância do efeito de área u_d:

$$var(\hat{\sigma}_u^2) \approx \sum_d 2\left\{a_d^2\left[(\hat{\sigma}_u^2)^2 + (\hat{\tau}_d^2)^2 + 2\hat{\sigma}_u^2\hat{\tau}_d^2\right] + b_d^2\frac{(\hat{\tau}_d^2)^2}{n_d - 1}\right\}$$

em que $\hat{\tau}_d^2 = \frac{1}{n_d(n_d-1)}\sum_j (e_{dj} - e_{d.})^2$, $e_{d.} = 1/n_d \sum_j e_{dj}$, $a_d = w_d / \sum_j w_j(1-w_j)$,

$b_d = w_d(1-w_d)/\sum_j w_j(1-w_j)$, $\hat{\sigma}_u^2 = \sum_d a_d \eta_{d.}^2 - \sum_d b_d \hat{\tau}_d^2$, $\eta_{d.}^2 = u_d + e_{d.}$, $w_d = \sum_j w_{dj}/n_d$,

n_d é o total de pessoas associado à área d e w_j é o peso do domicílio j.

Uma série de simulações é conduzida de tal forma que, para cada simulação l, l=1,... L, um conjunto de parâmetros é gerado a partir de distribuições derivadas das estimativas encontradas anteriormente. Os coeficientes $\tilde{\boldsymbol{\beta}}^{(l)}$ são obtidos da distribuição

normal multivariada cujos parâmetros são os coeficientes e matriz de covariâncias estimados pelo modelo (3). De forma análoga, os coeficientes $\tilde{\boldsymbol{\alpha}}^{(l)}$ são gerados a partir do modelo (4). Adicionalmente, $\tilde{\sigma}_u^{2(l)}$, um valor simulado para a variância do componente de efeito de área é gerado a partir de uma distribuição gama com média $\hat{\sigma}_u^2$ e variância $var(\hat{\sigma}_u^2)$. Combinando os coeficientes $\tilde{\boldsymbol{\alpha}}^{(l)}$ com os dados do Censo, obtém-se um valor para a variância do componente de erro do domicílio. Para cada domicílio do Censo, são extraídos $\tilde{u}_d^{(l)}$ e $\tilde{e}_{dj}^{(l)}$ de suas distribuições correspondentes e então a variável dependente é gerada através da equação $\tilde{y}_{dj}^{(l)} = \boldsymbol{x}_{dj}^{'censo}\tilde{\boldsymbol{\beta}}^{(l)} + \tilde{u}_d^{(l)} + \tilde{e}_{dj}^{(l)}$, $j=1,\ldots, N_d$ e $d=1,\ldots,D$. A partir da variável de bem-estar simulada, calculam-se os indicadores de pobreza L vezes. As estimativas pontuais dessas medidas são dadas pelas médias de todas as L iterações. Este procedimento permite que sejam estimadas as precisões das estimativas.

Método Molina e Rao

O método descrito por Molina e Rao (2010) combina os dados de domicílio da pesquisa aos dados do Censo e, através de modelos hierárquicos de dois níveis com efeitos aleatórios no intercepto, prevê a variável de bem-estar.

Seja $\boldsymbol{y}_d = \left(Y_{d1},\ldots,Y_{dN_d}\right)' = \left(\boldsymbol{y}_{ds}', \boldsymbol{y}_{dr}'\right)'$ o vetor cujos componentes são funções da variável de bem-estar do domínio d com N_d unidades da população, onde $\boldsymbol{y}_{ds}$ é um subvetor de elementos da pesquisa e $\boldsymbol{y}_{dr}$ é um subvetor de elementos não existentes na amostra. O modelo pode ser escrito da seguinte forma:

$$Y_{dj} = \boldsymbol{x}_{dj}'\boldsymbol{\beta} + u_d + e_{dj},\ u_d \overset{iid}{\sim} N(0,\sigma_u^2),\ e_{dj} \overset{iid}{\sim} N(0,\sigma_e^2), \tag{5}$$

em que o efeito de área u_d e os erros e_{dj} são independentes para $j=1,\ldots, N_d$ e $d=1,\ldots, D$. Os vetores $\boldsymbol{y}_d$, $d=1,\ldots$, D, sob o modelo (5) são independentes com $\boldsymbol{y}_d \sim N\left(\boldsymbol{\mu}_d, V_d\right)$, no qual $\boldsymbol{\mu}_d = X_d\boldsymbol{\beta}$ e $V_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2 I_{N_d}$.

A distribuição dos elementos fora da amostra, $\boldsymbol{y}_{dr}$, dados os elementos pertencentes à amostra, $\boldsymbol{y}_{ds}$, é dada por $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds} \sim N\left(\boldsymbol{\mu}_{dr|s}, V_{dr|s}\right)$ em que o vetor de médias condicionais e a matriz de covariâncias são dados por:

$$\boldsymbol{\mu}_{dr|s} = X_{dr}\boldsymbol{\beta} + \sigma_u^2 \mathbf{1}_{N_d - n_d}\mathbf{1}_{n_d}' V_{ds}^{-1}\left(\boldsymbol{y}_s - X_s\boldsymbol{\beta}\right)$$
$$V_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}' + \sigma_e^2 I_{N_d-n_d},$$

com $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d)$. Observe que a matriz $V_{dr|s}$ corresponde à matriz de covariâncias de um vetor $\boldsymbol{y}_{dr}$ gerado pelo modelo

$$\boldsymbol{y}_{dr} = \boldsymbol{\mu}_{dr|s} + \upsilon_d \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{dr}, \tag{6}$$

com efeitos aleatórios υ_d e $\boldsymbol{\varepsilon}_{dr}$ independentes satisfazendo $\upsilon_d \sim N\left(0, \sigma_u^2(1-\gamma_d)\right)$ e $\boldsymbol{\varepsilon}_{dr} \sim N\left(\mathbf{0}_{N_d-n_d}, \sigma_e^2 I_{N_d-n_d}\right)$.

Utilizam-se as expressões anteriores para gerar variáveis normais univariadas, $\upsilon_d \sim N\left(0, \sigma_u^2(1-\gamma_d)\right)$ e $\varepsilon_{dj} \sim N\left(0,\sigma_e^2\right)$, independentemente, para $j \in r_d$ e então obter as respostas Y_{dj} usando a equação (6) através do valor conhecido $\boldsymbol{\mu}_{dr|s}$. Na prática, todos os parâmetros desconhecidos do modelo β, σ_u^2 e σ_e^2 são substituídos por estimadores adequados, assim, as variáveis Y_{dj} são geradas a partir das correspondentes distribuições normais estimadas de $\hat{\upsilon}_d$ e $\hat{e}_{dj}$ e as medidas de pobreza correspondentes à área d, $d=1,\ldots, D$, são calculadas L vezes através da seguinte expressão:

$$fgt_{cd} = \frac{1}{N_d}\left[\sum_{j \in s_d} fgt^{w}_{cdj} + \sum_{j \in r_d} fgt_{cdj}\right].$$

Os indicadores de pobreza são estimados utilizando aproximação de Monte Carlo (ROBERT; CASELLA, 2004), os L vetores gerados $\boldsymbol{y}_r^{(l)}$, $l = 1, \ldots, L$, e os dados da amostra $\boldsymbol{y}_s$.

No caso em que não é possível associar as unidades do Censo às da pesquisa e nem todas as áreas de interesse estão na pesquisa, os autores sugerem uma adaptação que consiste em calcular a média de Y_{dj} e a variância do erro aleatório da seguinte forma: $\hat{\mu}_{dj} = \boldsymbol{x}_{dj}^{'censo}\hat{\boldsymbol{\beta}} + \hat{u}_d$ e $\hat{\sigma}_v^2 = \hat{\sigma}_u^2(1 - \hat{\gamma}_d)$, para as áreas pertencentes à amostra; $\hat{\mu}_{dj} = \boldsymbol{x}_{dj}^{'censo}\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}_v^2 = \hat{\sigma}_u^2$, para as áreas fora da amostra.

Aplicação

Seleção das subamostras no Censo Demográfico

Para avaliação das metodologias apresentadas no tópico anterior, foram selecionadas 400 subamostras de uma população, seguindo o desenho amostral da Pesquisa de Orçamentos Familiares - POF que foi implementado seguindo o esquema de estratificação descrito na publicação *Pesquisa de orçamentos familiares 2008-2009: despesas, rendimentos e condições de vida*, do IBGE. O objetivo foi comparar, em termos de vício, Erro Quadrático Médio - EQM relativos e ordenação de áreas, as estimativas dos indicadores incidência e hiato de pobreza obtidas pelas três metodologias com os respectivos valores diretamente estimados a partir dos dados populacionais.

Da amostra do Censo Demográfico 2010 do Estado de Minas Gerais de domicílios particulares permanentes dos setores dos tipos normal e aglomerado subnormal, excluíram-se os domicílios localizados em terras indígenas, totalizando 763.505 domicílios. Destes, excluíram-se os domicílios com rendimento total igual a zero, cerca de 3,11%, devido ao fato dos processos de modelagem adotados pelos métodos não possuírem tratamento específico para concentração de zeros na variável resposta. Ao final, restaram 739.762 domicílios que constituíram a população do estudo. A escolha de Minas Gerais ocorreu em virtude deste ser um estado que se assemelha e reflete as condições socioeconômicas do país.

A área de interesse neste estudo é o município por ser uma unidade administrativa importante para implementação de políticas públicas. Vale ressaltar que, de acordo com o plano amostral adotado, não se tem garantia de que os 853 municípios de Minas Gerais estejam contemplados nas 400 subamostras de domicílios. Apenas 14 municípios tinham domicílios selecionados em todas as subamostras e 96 apareciam, pelo menos, em 301 subamostras.

A ausência de áreas de interesse na amostra bem como o pequeno número de unidades amostradas em áreas selecionadas podem prejudicar o desempenho dos métodos, fazendo com que estes forneçam estimativas viciadas e/ou com baixa precisão. A agregação de áreas pode minimizar esse problema devido ao aumento do tamanho de amostra nos domínios. A fim de verificar esta hipótese, as comparações entre os métodos também foram realizadas para microrregiões as quais são grupos de municípios. Das 66 microrregiões de Minas Gerais, 40 possuíam observações em todas as subamostras. As demais foram encontradas em mais de 155 subamostras.

Resultados

Nesta parte, apresentam-se os resultados da aplicação dos métodos FH (FAY; HERRIOT, 1979), ELL (ELBERS; LANJOUW; LANJOUW, 2003) e MR (MOLINA; RAO, 2010), para municípios e microrregiões, nas 400 subamostras da população. No primeiro método, objetivou-se estimar a proporção de pessoas pobres e o hiato

de pobreza em cada domínio, portanto utilizaram-se como variáveis respostas dos modelos as estimativas diretas destas medidas obtidas através dos dados das subamostras. A variável de bem-estar considerada, nos métodos ELL e MR, foi a renda domiciliar *per capita*, pois esta estava disponível nos dados da população e os indicadores derivados puderam ser comparados com aqueles simulados pelos métodos. Para cada subamostra, um modelo de regressão específico foi selecionado para os métodos ELL e MR. Para o método FH, foi escolhido um modelo diferente para cada medida de pobreza e tipo de área. As variáveis utilizadas nos modelos representam características domiciliares, demográficas e educacionais, incluindo informações agregadas no nível de setor censitário.

Para cada subamostra e método (ELL e MR), foram geradas 1000 replicações da renda domiciliar *per capita* para todos os domicílios da população. A linha de pobreza utilizada foi de R$ 255,00, metade do salário mínimo vigente em 2010. Na aplicação do método MR, utilizou-se a adaptação sugerida pelos autores e descrita anteriormente para o caso de ausência de áreas de interesse na amostra. Os resultados dos métodos FH, ELL e MR foram obtidos utilizando rotinas desenvolvidas em SAS 9.2, pacote PovMap 1.2 e código desenvolvido em R 2.14.1, respectivamente.

A seguir são apresentadas as comparações dos resultados obtidos pelos três métodos com os valores da população em termos de erro quadrático médio relativo, vício relativo e ordenação de áreas.

Erro quadrático médio e vício relativos

Para cada método e área foram calculadas as seguintes medidas:

(a) média das estimativas de pobreza: $fgt_{cd}^{m} = (1/400)\sum_{i=1}^{400} fgt_{icd}^{m}$;

(b) vício relativo: $VR_{cd}^{m} = (1/400)\sum_{i=1}^{400} \left(fgt_{icd}^{m} - FGT_{cd}^{pop} \right) / FGT_{cd}^{pop}$;

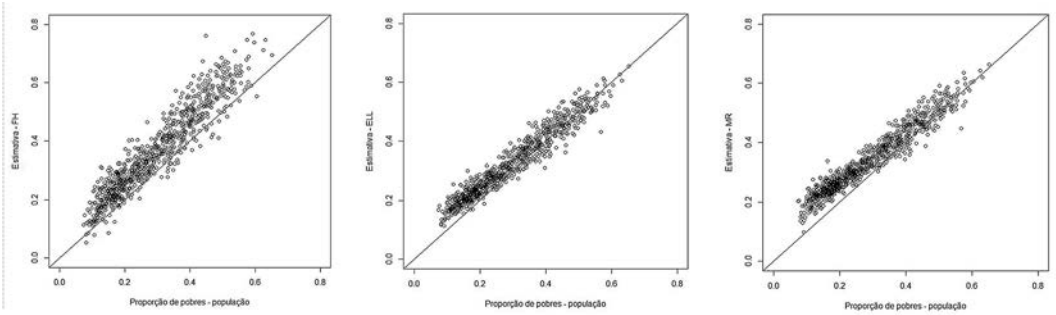
(c) erro quadrático médio relativo: $EQMR_{cd}^{m} = (1/400)\sum_{i=1}^{400} \left[\left(fgt_{icd}^{m} - FGT_{cd}^{pop} \right) / FGT_{cd}^{pop} \right]^2$.

A quantidade fgt_{icd}^{m} é a estimativa do indicador *c* obtida pelo método *m* na área *d* para a subamostra *i* e FGT_{cd}^{pop} é o indicador *c* obtido diretamente com base na população para a área *d*.

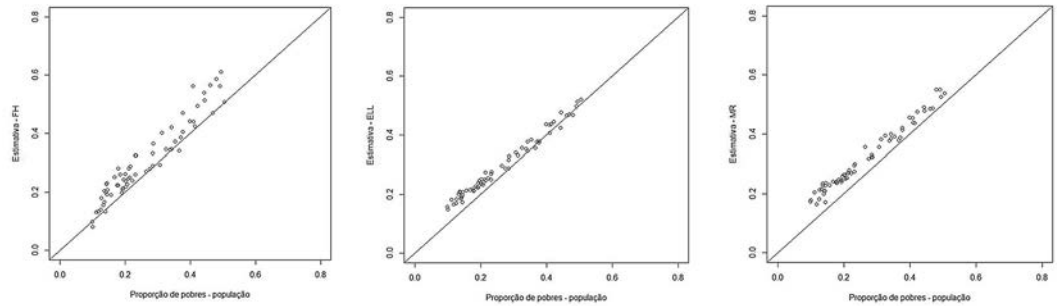
A Figura 1 compara as estimativas obtidas pelos métodos de estimação com aquelas calculadas diretamente na população para municípios e microrregiões. Pode-se observar que as estimativas dadas pelos métodos ELL e MR tendem a ficar mais próximas dos indicadores populacionais à medida que seus valores crescem. As estimativas fornecidas pelos métodos FH e MR foram, em geral, maiores que aquelas dadas pelo método ELL.

Como pode ser visto na Tabela 1, considerando as estimativas municipais, os métodos FH, ELL e MR apresentaram vícios relativos médios estimados de 28,11%, 17,42% e 27,49%, respectivamente, e EQMs relativos médios estimados de 21,13%, 7,71% e 15,40%, respectivamente, mostrando, portanto, que os métodos testados apresentam desempenhos diferentes com relação a vício e EQM relativos, mas todos superestimaram os valores da população. Os métodos FH e MR mostraram-se um pouco piores com relação a vício relativo, quando comparados ao método ELL. Em relação ao erro médio quadrático relativo, o método FH apresentou valores bem mais altos do que os demais métodos. No caso das microrregiões, houve redução nos valores dos EQMs e vícios relativos. O método FH apresentou, em média, os valores mais altos de EQMs relativos e o método MR apresentou o vício relativo médio mais alto.

Figura 1 - Estimativas de proporções de pobres obtidas pelos métodos FH, ELL e MR versus indicadores populacionais para os municípios e microrregiões de Minas Gerais



(a) Municípios



(b) Microrregiões

Fonte: Comparações dos Autores com base nos dados do Censo Demográfico 2010.

Tabela 1 - Estatísticas descritivas dos EQMs e vícios relativos das proporções de pobres obtidos pelos métodos para os municípios e microrregiões de Minas Gerais

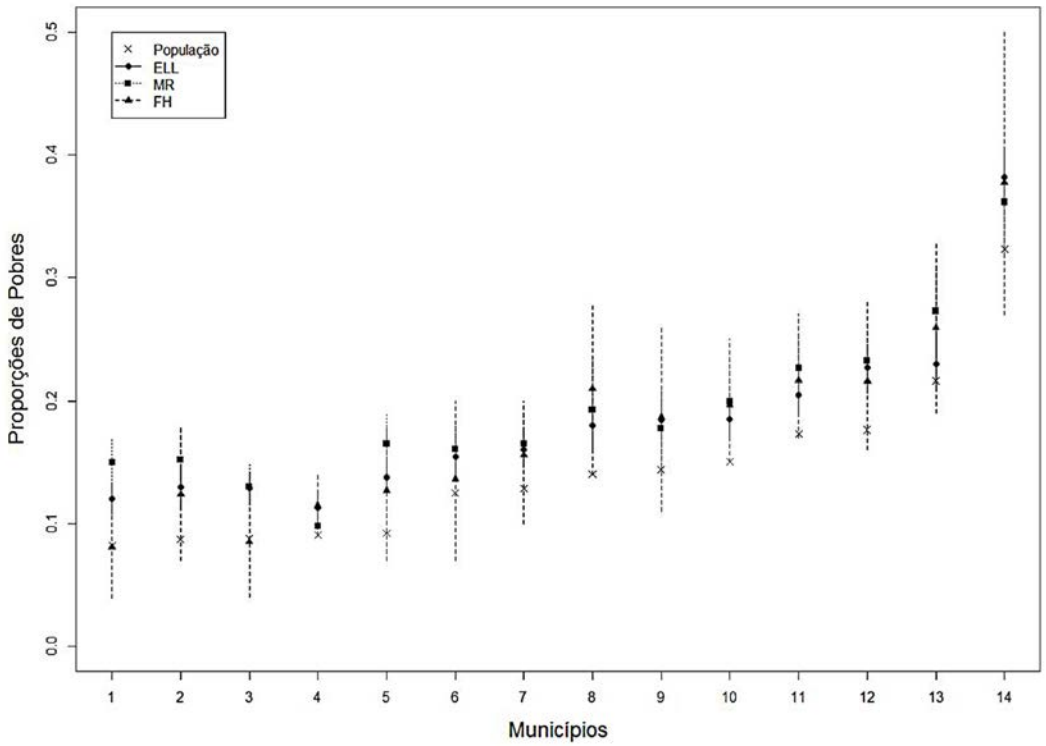
Municípios										
Resumo	Estimativa (em %)				EQM Relativo (em %)			Vício Relativo (em %)		
	População	FH	ELL	MR	FH	ELL	MR	FH	ELL	MR
Mínimo	7,32	5,33	11,29	9,81	1,20	0,04	0,05	(-) 43,61	(-) 24,02	(-) 21,34
Mediana	26,99	33,20	29,94	32,18	12,12	1,85	4,27	24,71	12,50	19,79
Média	29,07	35,97	32,16	34,30	21,13	7,71	15,40	28,11	17,42	27,49
Máximo	65,25	76,79	65,45	66,36	326,40	186,70	320,30	145,20	134,10	177,00
Microrregiões										
Resumo	Estimativa (em %)				EQM Relativo (em %)			Vício Relativo (em %)		
	População	FH	ELL	MR	FH	ELL	MR	FH	ELL	MR
Mínimo	9,93	8,02	14,91	16,42	2,18	0,04	0,20	(-) 20,22	(-) 4,34	3,21
Mediana	21,98	27,57	25,35	27,24	9,40	2,33	5,71	16,81	14,95	22,93
Média	26,35	30,77	29,56	31,84	14,29	6,38	12,61	18,99	18,45	28,25
Máximo	50,54	61,12	51,99	54,92	59,08	40,16	71,74	59,27	63,04	83,61

Fonte: Os Autores com base nos dados do Censo Demográfico 2010.

As Figuras 2 e 3 a seguir contêm as proporções de pobres populacionais e as estimadas pelos métodos com seus respectivos intervalos de confiança de 95% para os municípios que apareceram nas 400 subamostras e para as microrregiões de Minas Gerais, respectivamente. Os extremos dos intervalos de confiança de cada município e microrregião foram determinados a partir dos quantis 0,025 e 0,975 das distribuições das estimativas encontradas pelos métodos nas subamostras. O comportamento

observado na Figura 1 pôde ser visto mesmo nos municípios que apareceram em todas as subamostras. Na maioria dos municípios, os intervalos dos métodos ELL e MR não contêm os indicadores populacionais. O método FH apresentou estimativas muito diferentes por subamostra para estes municípios e, consequentemente, intervalos de confiança muito amplos que abrangeram os valores populacionais.

Figura 2 - Proporções de pobres populacionais e estimadas pelos métodos com seus respectivos intervalos de confiança de 95% para os municípios de Minas Gerais que apareceram nas 400 subamostras



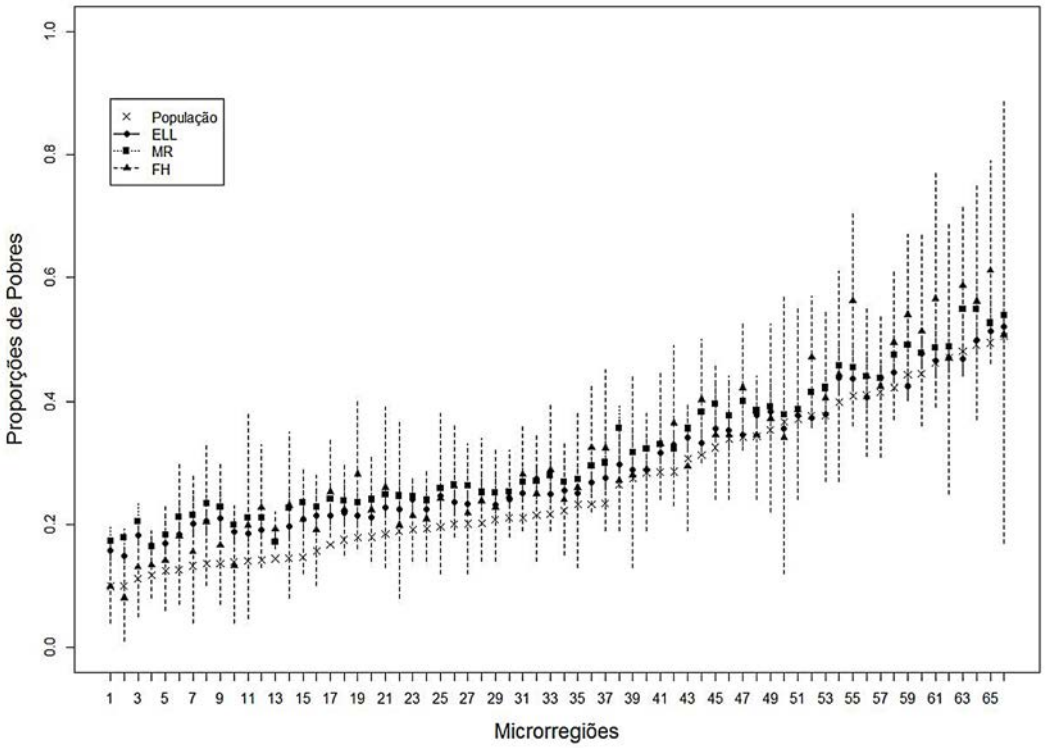
Fonte: Elaboração dos Autores com base nos dados do Censo Demográfico 2010.

A estimativa obtida pelo método MR para o município de Belo Horizonte, representado na 4ª posição da Figura 2, encontra-se mais próxima à proporção populacional em relação às estimativas obtidas pelos métodos FH e ELL. De maneira geral, os intervalos de confiança dos métodos ELL, MR e FH continham, respectivamente 30,7%, 21,2% e 89,3% dos valores populacionais dos municípios.

Considerando a Figura 3, os indicadores populacionais das microrregiões estão contidos em 97,0%, 22,7% e 13,6% dos intervalos de confiança dos métodos FH, ELL e MR, respectivamente. Em algumas microrregiões, nota-se, também, maior proximidade entre as estimativas dos métodos, como, por exemplo, na 13ª microrregião do gráfico a qual inclui o município de Belo Horizonte. As microrregiões que menos apareceram nas subamostras, 270 e 155 vezes, correspondem às posições 50 e 66, respectivamente.

Resultados análogos foram encontrados para o indicador hiato de pobreza.

Figura 3 - Proporções de pobres populacionais e estimadas pelos métodos com seus respectivos intervalos de confiança de 95% para as microrregiões de Minas Gerais



Fonte: Elaboração dos Autores com base nos dados do Censo Demográfico 2010.

Ordenação de áreas

Nesta parte, comparam-se as ordenações dos municípios e microrregiões a partir das proporções de pobres estimadas por cada um dos três métodos com as ordenações populacionais por meio do coeficiente de correlação de Spearman. Mais detalhes sobre o assunto são apresentados por Lehmann e D'Abrera (1976).

Denote por $(R_1,S_1),\ldots,(R_D,S_D)$ os postos das medidas de pobreza estimadas das áreas. O coeficiente de correlação de postos de Spearman é definido por:

$$r=\frac{\sum_{d=1}^{D}(R_d-\bar{R})(S_d-\bar{S})}{\sqrt{\sum_{d=1}^{D}(R_d-\bar{R})^2(S_d-\bar{S})^2}},$$

em que $\bar{R}=\sum_{d=1}^{D}R_d/D$ e $\bar{S}=\sum_{d=1}^{D}S_d/D$. Este coeficiente pode ser usado como uma medida de concordância de duas ordenações de objetos segundo dois critérios diferentes. Observe que quando as duas ordenações coincidem exatamente r=1 e que quando as ordenações são exatamente contrárias r=-1.

Para cada subamostra, foi medida a concordância entre a ordenação dos municípios ou microrregiões obtida com base nas estimativas FH, ELL e MR com a da população por meio do coeficiente de correlação de Postos de Spearman. Sejam, r_i(FH, POP), r_i(ELL, POP) e r_i(MR, POP), respectivamente, os coeficientes de correlação de postos entre as estimativas da proporção de pobres estimadas pelos métodos FH, ELL e MR e os valores populacionais correspondentes para cada subamostra i, i=1,... ,400. Analogamente, define-se $r_i(ELL, MR)$, $r_i(FH,ELL)$ e $r_i(FH, MR)$ como a correlação de postos entre as estimativas dadas pelos métodos. A Tabela 2 apresenta resumos estatísticos das 400 replicações das correlações de postos para municípios e microrregiões.

Tabela 2 - Distribuição da correlação de postos de Spearman por municípios e microrregiões

Áreas	Correlação	Min.	Média	Max.	Correlação	Min.	Média	Max.
Municípios	$r_i(FH,POP)$	0,77	0,86	0,92	$r_i(ELL,MR)$	0,97	0,99	1,00
	$r_i(ELL,POP)$	0,95	0,96	0,97	$r_i(FH,ELL)$	0,78	0,86	0,92
	$r_i(MR,POP)$	0,95	0,96	0,97	$r_i(FH,MR)$	0,78	0,86	0,92
Microrregiões	$r_i(FH,POP)$	0,73	0,85	0,95	$r_i(ELL,MR)$	0,92	0,97	0,99
	$r_i(ELL,POP)$	0,98	0,99	0,99	$r_i(FH,ELL)$	0,72	0,85	0,94
	$r_i(MR,POP)$	0,94	0,97	0,99	$r_i(FH,MR)$	0,74	0,86	0,95

Fonte: Os Autores com base nos dados do Censo Demográfico 2010.

É possível ver que há bastante concordância entre as ordenações definidas a partir dos valores estimados para os municípios de Minas Gerais pelos métodos ELL e MR e dos valores populacionais, sendo quase todos os valores das correlações de postos maiores que 0,95. O mesmo comportamento não é observado em relação ao método FH. Em geral, há maior concordância entre a ordenação dos municípios na população e a obtida a partir do método ELL. Considerando as microrregiões, há uma diferença maior entre as correlações de postos obtidas para o método ELL em relação às obtidas considerando o método MR.

Com relação à ordenação de municípios e microrregiões, os valores dos coeficientes de correlação de postos mostraram que, no caso de municípios, há maior concordância entre os métodos ELL e MR do que entre cada um dos três métodos e a população. Considerando as microrregiões, há maior concordância entre o método ELL e a população. Os mesmos resultados foram obtidos para as estimativas de hiato de pobreza.

Conclusões

A principal proposta deste capítulo foi comparar abordagens para estimação em pequenas áreas disponíveis na literatura atual. Para tanto, foram selecionadas, utilizando o desenho amostral da Pesquisa de Orçamentos Familiares - POF, 400 subamostras de uma população baseada nos dados do Estado de Minas Gerais do Censo Demográfico 2010. As três metodologias foram aplicadas aos dados das subamostras e as estimativas das medidas de pobreza obtidas pelos métodos foram comparadas com os valores calculados diretamente na população em termos de vício e EQM relativos e ordenação de áreas.

Cada método possui sua dificuldade de aplicação a dados reais. No método FH é preciso dispor de variáveis explicativas advindas de Censos Demográficos e registros administrativos. Na prática, a existência dessas variáveis em níveis menores é restrita. Outra desvantagem do método é a avaliação dos erros das estimativas ser possível apenas para as áreas contidas na amostra. Quanto aos métodos ELL e MR, são necessárias regressoras que existam na pesquisa amostral e no Censo Demográfico no nível domiciliar. Na realidade, a tarefa de reunir tais variáveis não é simples, uma vez que pode não ser possível a compatibilização de conceitos de variáveis importantes para medir o bem-estar do domicílio. Além disso, muitas vezes são afetadas por não-resposta, inviabilizando seu uso nos modelos. O método ELL calcula as variâncias das estimativas diretamente a partir das simulações, pelo método de Monte Carlo. No entanto, esta forma de obtenção dos erros das estimativas é questionada por Molina e Rao (2010). Em relação aos erros do método MR, os autores propõem um método *bootstrap* paramétrico que demanda grande tempo computacional, mas suposto eficaz na redução do erro quadrático médio. Este estudo não avaliou o desempenho dos métodos em relação aos erros quadráticos médios das estimativas propostos por seus respectivos autores.

O presente estudo indicou que os valores de vício e EQM relativos foram maiores para os municípios que apresentaram valores mais baixos de incidência de pobreza. Os métodos de estimação em pequenas áreas superestimaram os valores de proporção de pobres resultantes da população. Observou-se que os EQMs e os vícios relativos médios desses indicadores foram maiores para o método FH. No caso das microrregiões, o método MR foi o que apresentou vício relativo médio maior e o método FH apresentou o maior EQM relativo médio. Como esperado, os valores de EQMs e vícios relativos foram menores para as microrregiões que os valores obtidos para os municípios.

Os métodos produziram estimativas mais próximas entre si nas microrregiões. Na maior parte das áreas, os intervalos de confiança de 95% dos métodos ELL e MR não contiveram os indicadores populacionais. O método FH forneceu intervalos de confiança com grandes amplitudes que contiveram os valores populacionais na maioria das áreas.

As comparações realizadas em termos de ordenação de municípios, utilizando as estimativas de proporção de pobres, mostraram que as diferenças encontradas no desempenho dos métodos ELL e MR foram menores que as diferenças encontradas nas comparações de vício e EQM relativos. O método que apresentou ordenações de municípios mais diferentes da ordenação da população foi o FH, com correlação média de 0,86. No caso das microrregiões, notou-se uma maior diferença entre os métodos de estimação em pequenas áreas. O método ELL apresentou uma concordância maior com os valores populacionais.

Os métodos estudados foram capazes de identificar os municípios mais pobres já que apresentaram ordenações concordantes com a população, por outro lado superestimaram os indicadores de pobreza. Em termos de combate à pobreza se faz necessário estimar tais medidas com precisão para melhor identificar a parcela de domicílios mais pobres.

Referências

CENSO demográfico 2010. Rio de Janeiro: IBGE, [2017]. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/default.shtm>. Acesso em: set. 2017.

ELBERS, C.; LANJOUW, J. O.; LANJOUW, P. Micro–level estimation of poverty and inequality. *Econometrica*, New York: The Econometric Society, v. 71, n. 1, p. 355–364, Jan. 2003.

FAY, R. E.; HERRIOT, R. A. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, Alexandria [Estados Unidos]: American Satatistical Association – ASA, v. 74, n. 366a, p. 269–277, 1979. Disponível em: <http://amstat.tandfonline.com/toc/uasa20/74/366a?nav=tocList>. Acesso em: set. 2017.

FOSTER, J.; GREER, J.; THORBECKE, E. A class of decomposable poverty measures. *Econometrica*, New York: The Econometric Society, v. 52, n. 3, p. 761-766, May 1984.

MAPA de pobreza e desigualdade: municípios brasileiros 2003. IBGE, 2008. 1 DVD. Disponível em: <https://biblioteca.ibge.gov.br/biblioteca-catalogo.html?view=detalhes&id=241385>. Acesso em: set. 2017.

PESQUISA de orçamentos familiares 2008-2009: despesas, rendimentos e condições de vida. Rio de Janeiro: IBGE, 2010. 215 p. Acompanha 1 CD-ROM. Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv45130.pdf>. Acesso em: set. 2017.

LEHMANN, E. L.; D'ABRERA, H. J. M. *Nonparametrics*: statistical methods based on ranks. San Francisco: Holden-Day; New York; London: McGrae-Hill, [1976]. 457 p.

MOLINA, I.; RAO, J. N. K. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, Ottawa: Statistical Society of Canada, v. 38, n. 3, p. 369-385, Sep.

2010. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/cjs.v38:3/issuetoc>. Acesso em: set. 2017.

RAO, J. N. K. *Small area estimation*. Hoboken: Wiley, c2003. 313 p. (Wiley series in survey methodology).

ROBERT, C. P.; CASELLA, G. *Monte Carlo statistical methods*. 2nd. ed. New York: Springer, c2004. 645 p. (Springer texts in statistics).

SOUZA, D. F. et al. *Indicadores de pobreza nos municípios de Minas Gerais*: comparação de métodos de estimação em pequenas áreas. Rio de Janeiro: IBGE, 2014. 53p. (Texto para discussão. Diretoria de Pesquisas, n. 49). Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=286731>. Acesso em: set. 2017.

Alguns aspectos de amostragem da Pesquisa Nacional de Saúde do Escolar - PeNSE

Antonio José Ribeiro Dias
André Wallace Nery da Costa

Introdução

A Pesquisa Nacional de Saúde do Escolar - PeNSE, desde a sua primeira realização em 2009, vem sendo executada em convênio entre o Instituto Brasileiro de Geografia e Estatística - IBGE e o Ministério da Saúde, contando, também, com a colaboração do Ministério da Educação através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP que disponibilizou as informações necessárias para a construção dos cadastros de escolas para a seleção das amostras. Até o presente momento foram realizadas três edições da pesquisa, em 2009, 2012 e 2015, com algumas alterações entre elas. Neste texto serão focadas, principalmente, as mudanças metodológicas no que se refere à questão da amostragem.

Na fase de transição entre a infância e a vida adulta, ocorre uma série de alterações biológicas, cognitivas, sociais e comportamentais que terão impacto importante na vida futura dessas pessoas. Uma forte exposição a diversos fatores de risco comportamentais, como a iniciação sexual, uso do tabaco, do álcool e mesmo de drogas ilícitas, alimentação inadequada, sedentarismo pode ocorrer nessa época. Muitas doenças que lideram as causas de morte na idade adulta, não só no Brasil, como o diabetes ou as doenças cardiovasculares têm sua origem associada a esses fatores. Em relação à iniciação sexual, o Fundo das Nações Unidas para a Infância (United Nations Children's Fund - UNICEF) recomenda investir no estudo sobre a saúde sexual e reprodutiva no início da adolescência, principalmente visando a orientação e o cuidado no sentido de evitar a gravidez precoce e o aparecimento das doenças sexualmente transmissíveis - DST.

É, também, importante investigar as causas externas (lesões, acidentes, violências etc.) que são as principais causas de mortes e de sequelas ou incapacidades que venham afetar a vida de adolescentes e jovens. Estas causas de morte refletem a exposição a situações de risco vividas pelos adolescentes e que podem ser prevenidas, em grande parte, por mudanças no ambiente social e no comportamento desta parcela da população.

A PeNSE é a primeira iniciativa em nosso país de investigar diretamente junto aos adolescentes os fatores de risco e proteção, visando dar importante

subsídio aos gestores e responsáveis pela criação e condução de políticas públicas para proteção e monitoramento da saúde desses adolescentes. Além dos gestores públicos, os dados coletados pela pesquisa estão disponíveis para qualquer usuário interessado no assunto, que pode obtê-los gratuitamente no portal do IBGE na Internet.

Uma importante inovação introduzida desde a primeira PeNSE foi quanto à forma de coleta dos dados. Tradicionalmente, as pesquisas realizadas pelo IBGE são do tipo entrevista onde um pesquisador da instituição faz as perguntas ao informante e preenche um questionário, seja ele em papel ou, como ocorre atualmente, num equipamento portátil de entrada de dados ou Dispositivo Móvel de Coleta - DMC, como um *tablet* ou um *smartphone*. Nas PeNSE o método utilizado para a coleta dos dados é o da auto-entrevista, onde cada um dos estudantes selecionados para a amostra da pesquisa recebe um DMC e responde ao questionário sem a participação direta de um entrevistador. Um servidor do IBGE apenas seleciona as turmas cujos alunos são convidados a responderem a pesquisa, distribuem os dispositivos de coleta e, eventualmente, auxiliam os respondentes com esclarecimentos de dúvidas a respeito do questionário.

População de estudo

Em concordância com o tópico anterior, a população a ser amostrada, sujeita aos fatores de risco em questão, por recomendação da Organização Mundial da Saúde - OMS (World Health Organization - WHO) é aquela entre 13 e 15 anos de idade. Como uma aproximação dessa população, foi decidido investigar uma amostra de estudantes que estivessem cursando o 9º ano (ou a antiga 8ª série) do Ensino Fundamental, que concentra quase que a totalidade dos meninos e meninas da faixa etária de interesse, além de propiciar facilidades operacionais para atingir os adolescentes a serem pesquisados, por estarem agrupados em turmas dentro das escolas, que são facilmente identificáveis. Além disso o INEP realiza anualmente o Censo Escolar, permitindo a obtenção de um bom cadastro para a seleção da amostra.

Existem algumas escolas com poucos alunos cursando o 9º ano do Ensino Fundamental. Nesse caso o investimento de recursos para pesquisar uma dessas escolas não compensaria em termos da quantidade de informação obtida, o que levou à decisão de eliminar do cadastro as escolas com menos de 15 alunos matriculados no 9º ano do Ensino Fundamental. Os alunos do turno noturno costumam ser mais velhos, geralmente não pertencendo à faixa etária de interesse da pesquisa e, portanto, as turmas desse turno foram retiradas do cadastro de seleção.

Com essas considerações, a população de estudo da pesquisa, ou população amostrada, foi formada pelos estudantes cursando o 9º ano do Ensino Fundamental no ano de referência da pesquisa, em turmas dos turnos diurnos nas escolas públicas ou privadas com 15 ou mais alunos matriculados nessa etapa de ensino, segundo o cadastro de seleção das escolas, construído a partir dos dados disponíveis do Censo Escolar mais recente realizado pelo INEP.

O plano amostral básico

O plano amostral utilizado para a seleção dos estudantes participantes da pesquisa, nas três rodadas já realizadas, foi, basicamente, um plano de amostragem estratificada de conglomerados, em vários estágios de seleção, onde foram selecionadas escolas, nessas escolas foram selecionadas turmas de 9º ano, e foram convidados, para participar da pesquisa, todos os alunos das turmas selecionadas. A princípio poderia ser selecionada apenas uma amostra de alunos em cada turma, mas para simplificar o gerenciamento da coleta de dados optou-se pela não introdução de mais um estágio de

seleção, já que poderia ser complicado explicar aos alunos porque alguns responderiam ao questionário e outros não.

Em cada uma das realizações da pesquisa foram introduzidas alterações no plano amostral, principalmente quanto à estratificação e ao aumento da abrangência da pesquisa em relação aos domínios pesquisados, que serão comentadas nos tópicos seguintes.

Aspectos de amostragem da PeNSE 2009

Como o cadastro mais recente e disponível de todos os escolares matriculados no 9º ano do Ensino Fundamental no conjunto das escolas brasileiras era referente ao Censo Escolar 2007, o que impossibilitava a seleção direta de uma amostra de escolares, o que se propôs para a pesquisa em questão foi um plano com amostragem de conglomerados em dois estágios, onde no primeiro estágio foi selecionada uma amostra de escolas e, no segundo estágio, foram selecionadas turmas do 9º ano do Ensino Fundamental das escolas selecionadas. A amostra de escolares foi formada, portanto, por todos os escolares das turmas selecionadas na amostra de escolas. As escolas foram estratificadas, inicialmente, levando-se em conta sua localização geográfica e dependência administrativa. No caso, a estratificação geográfica correspondeu às capitais das Unidades da Federação e o Distrito Federal, totalizando 27 estratos. Em cada um destes, as escolas com classes do 9º ano do Ensino Fundamental foram agrupadas em escolas privadas ou públicas (federais, estaduais ou municipais).

O tamanho da amostra foi calculado para fornecer estimativas de proporções (ou prevalências) de algumas características de interesse, em cada um dos estratos geográficos, com um erro máximo de 0,03 em valor absoluto com nível de confiança de 95%. Para garantir que isto ocorra, foi dimensionada a amostra, considerando que a prevalência (proporção) é da ordem de 0,5 (ou 50%), pois para proporções desse valor, a variância dos estimadores amostrais é máxima. Os agrupamentos formados pelo cruzamento dos estratos geográficos com a dependência administrativa (pública ou privada) e tamanho, medido pelo número de turmas de 9º ano, das escolas foram utilizados apenas para alocação da amostra, de maneira a garantir a presença de escolas públicas e privadas na amostra, de forma proporcional a sua existência no cadastro de seleção. Em cada estrato geográfico, a amostra foi obtida em dois estágios, como descrito anteriormente. A seleção das escolas, foi pelo método de seleção com probabilidades proporcionais ao tamanho, utilizando-se como medida de tamanho o número total de turmas do 9º ano do Ensino Fundamental de cada escola, conforme o cadastro de seleção.

Cada uma das escolas selecionadas nesse primeiro estágio foi visitada para construção de uma lista atualizada de turmas do 9º ano do Ensino Fundamental existentes em 2009. Após a obtenção dessas listas, foram selecionadas, por Amostragem Aleatória Simples - AAS, as turmas do 9º ano do Ensino Fundamental a serem efetivamente pesquisadas em cada uma das escolas selecionadas no primeiro estágio. Nas escolas com até duas turmas de 9º ano foi selecionada uma turma para a amostra, enquanto que nas escolas com três ou mais turmas de 9º ano foram selecionadas duas turmas.

É sabido que o emprego de planos amostrais conglomerados geralmente resulta em redução de custos para amostras de igual tamanho total em comparação com a AAS, por concentrar a amostra nos conglomerados selecionados, reduzindo seu espalhamento geográfico. Por outro lado, o impacto sobre a precisão costuma ser negativo, no sentido de que amostras conglomeradas de igual tamanho que uma AAS leva a estimadores com maior variância. Uma medida usual do impacto do emprego de amostragem conglomerada (ou de amostragem usando planos complexos, em geral) é o chamado Efeito do Plano Amostral - EPA, definido como a razão entre a variância do estimador sob o plano conglomerado dividido pela variância do estimador sob AAS de igual tamanho. Portanto, para estimar uma proporção da ordem de 50%, com uma dada margem de erro, k, e nível de confiança

de 95%, pode-se estimar um tamanho (no caso em número de escolares) para um plano amostral conglomerado em dois estágios e seleção com probabilidades proporcionais a uma medida de tamanho pela fórmula:

$$n_{PPT} = epa \frac{\frac{1{,}96}{4k^2}}{1 + \frac{1}{N}\frac{1{,}96}{4k^2}},$$

onde n_{ppt} é o tamanho da amostra de escolares num determinado estrato geográfico, N é o total de escolares nesse mesmo estrato geográfico e *epa* é uma estimativa do efeito de conglomeração ou Efeito do Plano Amostral - EPA, pelo fato de ser utilizada uma amostra de conglomerados e não uma Amostra Aleatória Simples - AAS de escolares. Em outras palavras, como, geralmente, a amostragem de conglomerados é menos precisa, é necessário uma amostra maior para se obter uma precisão equivalente à AAS.

As estimativas para o efeito de conglomeração para proporções utilizadas em cada estrato geográfico foram obtidas a partir das variáveis socioeconômicas da pesquisa do Sistema de Avaliação da Educação Básica - SAEB 2003. Os valores para diversas variáveis, em cada estrato geográfico, foram analisados e decidiu-se utilizar o valor equivalente ao terceiro quartil da distribuição dos EPA, assegurando que aproximadamente 75% das variáveis teriam estimativas de precisão igual ou maior que a especificada para o dimensionamento da amostra.

Para determinar o número de turmas a ser selecionado em cada estrato, o tamanho da amostra no segundo estágio, divide-se n_{ppt} pelo número médio de escolares por turma do 9º ano do Ensino Fundamental, conforme obtido do cadastro de seleção da amostra de escolas. O número de escolas do primeiro estágio é obtido dividindo-se o número de turmas da amostra pelo número médio de turmas das escolas do cadastro em cada estrato.

Como se vê, o dimensionamento da amostra é feito do final para o início, calculando-se o tamanho da amostra em termos das unidades elementares, neste caso o número de alunos, depois o número de unidades do segundo estágio, turmas, finalmente obtendo-se o número de escolas a serem selecionadas. É bom ressaltar que o número de escolas selecionadas é realmente o calculado no dimensionamento, porém o número de unidades do segundo estágio, turmas, e o tamanho final em termos de unidades elementares, escolares, são conhecidos apenas ao final da seleção, já que o número de turmas vai depender do tamanho das escolas selecionadas e o número de escolares dependerá do tamanho das turmas selecionadas, sendo os números obtidos no dimensionamento apenas valores esperados.

Como na PeNSE é desejado estimar com precisão equivalente para cada um dos estratos geográficos, o dimensionamento foi feito de forma independente para cada um desses estratos. Os tamanhos de amostra calculados para cada estrato geográfico foram distribuídos proporcionalmente pelos estratos de alocação, que levaram em conta a dependência administrativa e o tamanho das escolas.

A Tabela 1 mostra os valores dos tamanhos de amostra calculados como mostrado anteriormente e os valores efetivamente obtidos pela coleta nas escolas.

O cadastro utilizado para a seleção da amostra de escolas foi baseado nas escolas listadas no Censo Escolar 2007 e a coleta realizada de março a junho de 2009. Os números de alunos matriculados e daqueles que frequentam regularmente as aulas foram fornecidos pelas escolas no momento da pesquisa.

No total dos 27 estratos geográficos, 501 escolares presentes no dia da realização da pesquisa em suas respectivas turmas se recusaram a responder ao questionário. Além disso, alguns respondentes deixaram de informar a variável sexo.

Tabela 1 - Tamanhos de amostra calculados e coletados, por estágio de seleção, segundo as capitais e Distrito Federal - 2009

Capitais e Distrito Federal	Calculados			Coletados				
						Escolares		
	Escolas (1)	Turmas (2)	Escolares matriculados (2)	Escolas	Turmas	Matriculados (3)	Frequentes (3)	Presentes no dia da pesquisa (3)
Total	**1 507**	**2 270**	**72 596**	**1 453**	**2 175**	**72 782**	**68 735**	**63 411**
Porto Velho	56	86	2 363	53	76	2 361	2 123	2 120
Rio Branco	46	72	2 495	43	67	2 256	2 256	1 967
Manaus	42	70	2 391	38	59	2 323	2 039	1 977
Boa Vista	49	82	2 340	48	80	2 316	2 029	1 997
Belém	46	72	2 440	46	73	2 678	2 626	2 189
Macapá	58	88	2 988	56	84	2 813	2 805	2 498
Palmas	41	54	2 000	41	56	1 843	1 627	1 615
São Luis	66	92	3 058	65	79	3 092	2 711	2 670
Teresina	60	82	2 604	59	78	2 582	2 276	2 276
Fortaleza	52	76	2 575	46	72	2 588	2 584	2 340
Natal	62	90	2 853	61	86	2 980	2 933	2 597
João Pessoa	70	98	3 061	67	87	2 768	2 356	2 321
Recife	56	92	3 336	52	79	3 044	3 044	2 561
Maceió	46	70	2 334	40	59	2 304	1 985	1 948
Aracaju	64	88	2 777	62	82	2 787	2 614	2 344
Salvador	48	78	2 466	45	78	2 567	2 216	2 198
Belo Horizonte	68	114	3 467	65	111	3 433	3 115	3 105
Vitória	64	92	2 922	61	83	2 519	2 489	2 259
Rio de Janeiro	56	84	2 866	56	91	3 397	3 257	2 984
São Paulo	52	88	2 835	52	92	3 178	3 086	2 680
Curitiba	44	80	2 578	44	80	2 684	2 679	2 397
Florianópolis	67	88	2 519	66	88	2 553	2 544	2 225
Porto Alegre	52	74	2 001	52	67	1 934	1 873	1 727
Campo Grande	60	78	2 180	58	76	2 420	2 151	2 141
Cuiabá	50	72	2 029	50	76	2 285	2 243	2 014
Goiânia	76	114	3 856	73	112	3 727	3 727	3 291
Distrito Federal	56	96	3 262	54	94	3 350	3 347	2 970

Fontes: 1. IBGE, Pesquisa Nacional de Saúde do Escolar 2009. 2. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, Censo Escolar 2007.

(1) Tamanhos de amostra calculados. (2) Valores esperados. (3) Informações obtidas nas escolas pesquisadas.

Para determinação dos pesos amostrais divulgados e utilizados na produção das estimativas, decidiu-se utilizar apenas as informações dos estudantes das turmas selecionadas de 9º ano do Ensino Fundamental para a pesquisa que concordaram em responder ao questionário e informaram seu sexo. Os pesos amostrais foram calculados de maneira a representar os alunos matriculados no 9º ano do Ensino Fundamental que frequentam regularmente as aulas.

Como no dia da pesquisa alguns escolares não compareceram às aulas, deixando de responder ao questionário, e alguns dos respondentes podem não ter informado o sexo ou idade, tais correções têm de ser feitas. Dessa maneira o peso do aluno *i*, da turma *j* da escola *k*, para um determinado estrato geográfico (capital ou Distrito Federal), foi calculado como:

$$w_{ijk} = w_k \frac{T_k F_{jk} R_{jk}}{t_k R_{jk} S_{jk}} = w_k \frac{T_k F_{jk}}{t_k S_{jk}}, \forall\ i = 1,2,\ldots,S_{jk},$$

onde:

w_k é o peso da escola k, dado pelo inverso da sua probabilidade de seleção;

T_k é o número total de turmas do 9º ano do Ensino Fundamental na escola k;

t_k é o número de turmas do 9º ano do Ensino Fundamental na escola k selecionadas para a amostra;

F_{jk} é o número total de escolares na turma j da escola k, que frequentam regularmente as aulas;

R_{jk} é o número total de escolares na turma j da escola k, que frequentam regularmente as aulas e responderam à pesquisa;

S_{jk} é o número total de escolares na turma j da escola k, que frequentam regularmente as aulas, responderam a pesquisa e informaram o sexo.

Ressalte-se que para a PeNSE 2009 pode-se obter estimativas para os 27 estratos geográficos definidos e agregações dos mesmos, porém não se pode estimar parâmetos para o conjunto do país.

Aspectos de amostragem da PeNSE 2012

A amostra da PeNSE 2012 foi dimensionada para estimar parâmetros populacionais, com a estimativa da respectiva precisão, em diversos domínios geográficos: cada uma das 26 capitais e o Distrito Federal, cada uma das 5 grandes regiões geográficas (Norte, Nordeste, Sudeste, Sul e Centro-Oeste), além do país como um todo. Também é possível obter estimativas para agregações desses domínios, desde que façam sentido.

Os municípios brasileiros foram, então, estratificados em 32 estratos geográficos da seguinte forma:

- Cada um dos 26 municípios das capitais e o Distrito Federal foi definido como um estrato geográfico;
- Os demais municípios foram agrupados em sua respectiva Grande Região geográfica formando 5 estratos.

Os procedimentos de seleção da amostra para os estratos das capitais e Distrito Federal foram os mesmos utilizados para a PeNSE 2009.

Para os estratos formados pelos municípios que não são capitais, agrupados por grandes regiões geográficas, foi introduzido mais um estágio de seleção, já que a seleção de escolas diretamente em cada estrato poderia levar a um grande espalhamento da amostra, o que poderia dificultar e onerar bastante as atividades da coleta dos dados.

A seleção de municípios como um estágio intermediário apresentava o problema da grande heterogeneidade de tamanho dos municípios, medido pelo número de turmas de 9º ano do Ensino Fundamental no Censo Escolar 2010, além de, também, espalhar a amostra. A solução adotada foi agrupar os municípios seguindo critérios de homogeneidade e vizinhança, obtendo grupos de 300 a 600 turmas aproximadamente e, em seguida, selecionar uma amostra desses grupos de municípios em cada grande região geográfica. Nos agrupamentos de municípios selecionados foram selecionadas as escolas para a amostra da PeNSE 2012.

O procedimento adotado ajudou a reduzir o custo total da coleta, diminuiu o tempo e simplificou o controle dos trabalhos de campo, já que as escolas selecionadas pertenciam a municípios vizinhos.

Para os estratos das capitais e Distrito Federal foram selecionadas diretamente amostras de escolas e em cada escola, após atualização do cadastro das turmas de 9º ano do Ensino Fundamental, foi selecionada uma amostra de turmas, onde todos seus alunos presentes no dia da pesquisa foram convidados a participar da pesquisa.

Assim, para as capitais e Distrito Federal as unidades primárias de amostragem - UPAs foram as escolas, enquanto que para os demais estratos as UPAs foram os grupos de municípios selecionados.

O cadastro de seleção da amostra foi construído a partir das informações das escolas listadas no Censo Escolar 2010 e, como na PeNSE anterior, só foram consideradas as turmas dos turnos diurnos das escolas com pelo menos 15 estudantes matriculados. O dimensionamento das amostras em cada estrato foi equivalente ao utilizado na PeNSE 2009.

A Tabela 2 mostra os resultados do dimensionamento das amostras e os números finais efetivamente obtidos após a coleta, por estrato geográfico.

Tabela 2 - Tamanhos de amostra calculados e coletados, por estágio de seleção, segundo os estratos geográficos - 2012

Estrato geográfico	Calculados			Coletados					
						Escolares			
	Escolas (1)	Turmas (2)	Escolares matriculados (2)	Escolas	Turmas	Matriculados (3)	Frequentes (3)	Presentes no dia da pesquisa (3)	Respondentes
Total	**3 004**	**4 288**	**131 741**	**2 842**	**4 091**	**134 310**	**132 123**	**110 873**	**109 104**
Não Capitais									
Norte	292	389	11 272	270	345	11 089	10 914	9 303	8 802
Nordeste	308	385	11 150	279	352	11 376	11 183	9 493	9 460
Sudeste	257	371	11 867	247	362	12 604	12 527	10 145	9 945
Sul	237	348	9 384	235	351	10 171	10 167	8 943	8 731
Centro-Oeste	357	540	15 094	342	462	14 638	14 026	11 149	11 021
Capitais									
Porto Velho	53	78	2 404	49	76	2 549	2 531	2 072	2 002
Rio Branco	41	60	2 575	38	60	2 085	2 072	1 737	1 735
Manaus	44	67	2 376	41	64	2 518	2 517	2 043	2 019
Boa Vista	48	82	2 432	47	86	2 450	2 450	2 042	2 027
Belém	47	68	2 440	44	68	2 509	2 339	1 978	1 974
Macapá	58	92	3 016	56	90	3 006	2 959	2 441	2 437
Palmas	45	64	2 165	42	61	2 140	2 087	1 789	1 778
São Luis	65	89	3 081	62	86	3 127	3 094	2 676	2 675
Teresina	63	83	2 624	55	80	2 740	2 718	2 482	2 377
Fortaleza	54	78	2 564	52	79	2 728	2 656	2 270	2 266
Natal	65	89	2 836	63	90	2 943	2 943	2 385	2 384
João Pessoa	83	102	3 038	75	97	3 222	3 221	2 612	2 610
Recife	71	97	3 289	70	99	3 718	3 705	3 010	3 006
Maceió	48	68	2 304	40	60	2 239	2 143	1 825	1 819
Aracaju	66	90	2 780	63	92	3 196	3 167	2 648	2 640
Salvador	47	75	2 470	46	77	2 589	2 454	2 070	2 064
Belo Horizonte	68	109	3 457	68	102	3 284	3 270	2 759	2 754
Vitória	69	97	2 808	64	88	2 526	2 501	2 161	2 140
Rio de Janeiro	58	85	2 869	54	84	2 900	2 851	2 424	2 413
São Paulo	50	86	2 832	50	88	3 070	3 007	2 464	2 408
Curitiba	46	79	2 581	43	76	2 477	2 463	2 174	2 153
Florianópolis	68	89	2 662	67	105	3 011	2 992	2 771	2 539
Porto Alegre	54	69	1 987	52	64	1 752	1 742	1 456	1 455
Campo Grande	56	74	2 194	55	76	2 355	2 215	1 987	1 953
Cuiabá	47	70	2 082	43	65	2 136	2 123	1 545	1 539
Goiânia	82	121	3 841	77	112	3 792	3 727	3 055	3 044
Distrito Federal	57	94	3 267	53	94	3 370	3 359	2 964	2 934

Fontes: 1. IBGE, Pesquisa Nacional de Saúde do Escolar 2012. 2. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, Censo Escolar 2012.

(1) Tamanhos de amostra calculados. (2) Valores esperados. (3) Informações obtidas nas escolas pesquisadas.

Do total de escolas selecionadas, 162 não foram pesquisadas por diversos motivos, como a inexistência de turmas do 9º ano, menos de 15 alunos matriculados no 9º ano, escolas desativadas, recusa por parte da direção da escola ou impedimento de acesso à escola. O fato do cadastro de seleção ter uma defasagem em relação ao período de coleta pode ser entendido como a causa de algumas dessas perdas. Tais perdas estão dentro dos padrões de pesquisas semelhantes realizadas tanto pelo IBGE como por outras instituições de estatística.

Como mostra a Tabela 2, segundo as informações das escolas pesquisadas, de um total de 134 310 alunos matriculados nas turmas selecionadas, 132 123 frequentavam habitualmente as aulas e, destes, 110 873 estiveram presentes no dia da pesquisa. Dos estudantes presentes, 1 769 não foram considerados no cálculo dos pesos amostrais, por se recusarem a participar da pesquisa ou não terem informado as variáveis sexo e idade, correspondendo a aproximadamente 1,6% dos presentes. As ausências e recusas fizeram com que a amostra de alunos que efetivamente responderam ao questionário fosse cerca de 83% dos alunos matriculados no 9º ano do Ensino Médio e que frequentavam habitualmente as aulas nas turmas selecionadas.

Os pesos amostrais foram determinados levando em conta as perdas relatadas e as correções para representarem os alunos matriculados no 9º ano que frequentavam habitualmente as aulas. Assim o cálculo dos pesos amostrais, para os alunos de um determinado estrato, foram feitos de acordo com as fórmulas seguintes:

$$w_{ijkl} = w_i\, w_{ij}\, w_{ijk}, l = 1,2,\ldots, P_{ijk},$$

$w_i = 1,$ para capitais e Distrito Federal

$w_i = \dfrac{T}{2\,T_i},$ representa o peso da UPA *i* para demais estratos geográficos

$$w_{ij} = \frac{T_i}{n_i\, T_{ij}}$$

$$w_{ijk} = \frac{T_{ij}}{n_{ij}} \frac{F_{ijk}}{P_{ijk}}$$

onde:

w_{ij} é o peso amostral da escola *j*, da capital ou UPA *i*;

w_{ijk} é o peso amostral da turma *k*, da escola *j*, da capital ou UPA *i*;

T é o número total de turmas de 9º ano do estrato;

T_i é o número total de turmas de 9º ano da capital ou UPA *i*;

T_{ij} é o número total de turmas de 9º ano da escola *j*, da capital ou UPA *i*;

n_i é o tamanho da amostra de escolas da capital ou UPA *i*;

n_{ij} é o tamanho da amostra de turmas da escola *j*, da capital ou UPA *i*;

P_{ijk} é o número de alunos respondentes da turma *k*, escola *j*, capital ou UPA *i*;

F_{ijk} é o número de alunos frequentes da turma *k*, escola *j*, capital ou UPA *i*;

Para os estratos das capitais e Distrito Federal, as unidades primárias de amostragem são as escolas, enquanto que para os demais estratos as unidades primárias de amostragem são os agrupamentos de municípios selecionados.

Foram utilizadas os dados sobre totais de turmas e alunos matriculados nas escolas atualizados pelo Censo Escolar 2012, realizado pelo INEP, para ajustar as probabilidades de seleção das escolas.

Aspectos de amostragem da PeNSE 2015

A principal inovação da PeNSE 2015 em relação às anteriores foi a investigação de duas amostras independentes. Uma, referenciada como Amostra 1, foi equivalente às amostras das pesquisas anteriores visando manter a série e a comparabilidade entre as pesquisas dos escolares do 9º ano do Ensino Fundamental, enquanto que a outra, Amostra 2, foi uma amostra mais abrangente em termos de idade dos estudantes pesquisados, pois procurou abranger os adolescentes entre 13 e 17 anos de idade, investigando alunos que frequentavam do 6º ao 9º ano do Ensino Fundamental e da 1ª a 3ª séries do Ensino Médio.

Em termos de cobertura geográfica, a Amostra 1 foi planejada para estimar parâmetros populacionais para as capitais e o Distrito Federal, Unidades da Federação, grandes regiões geográficas e Brasil, além de agregações de interesse desses domínios. A Amostra 2 foi definida de modo a fornecer estimativas apenas para as grandes regiões geográficas e Brasil.

A Amostra 1, como nas PeNSE anteriores, seguiu as recomendações da Organização Mundial da Saúde - OMS (World Health Organization - WHO) quanto à utilização de uma amostra de estudantes de 9º ano do Ensino Fundamental que, no Brasil, concentra mais de 80% dos adolescentes na faixa etária dos 13 aos 15 anos, e estão expostos aos fatores de risco de interesse da pesquisa. Por outro lado, essa etapa de ensino não se mostra adequada para investigar a faixa etária mais ampliada de 13 a 17 anos, pois a presença de alunos de 16 e 17 anos no 9º ano do Ensino Fundamental é bastante residual. Foi então necessário ampliar a pesquisa em relação às etapas de ensino a serem amostradas pela Amostra 2. Analisando a informações do Censo Escolar realizado pelo INEP no ano de 2013, verificou-se que cerca de 95% dos estudantes na faixa etária de interesse estavam concentrados entre o 6º e o 9º anos do Ensino Fundamental e entre a 1ª e 3ª séries do Ensino Médio, das escolas públicas e privadas em todo o território brasileiro. Os cadastros utilizados para seleção de ambas as amostras pesquisadas foram formados pelas escolas de ensino regular, listadas pelo Censo Escolar 2013, que possuíam turmas referentes às etapas de ensino de interesse para cada uma das duas amostras. No caso da Amostra 2, foram também incluídas as escolas com turmas de interesse do turno noturno, já que, neste caso, a presença de alunos na faixa etária estudada é significativa.

No Quadro 1 são apresentadas as principais características das duas amostras utilizadas na PeNSE 2015.

A Amostra 1, como se vê no Quadro 1, foi dimensionada para estimar parâmetros (proporções ou prevalências) populacionais para as 26 capitais e o Distrito Federal, as 26 Unidades da Federação, e, consequentemente, as agregações de interesse desses domínios como, por exemplo, as 5 grandes regiões geográficas e a totalidade do País. Dessa forma, foram criados 53 estratos geográficos onde foram dimensionadas e selecionadas as amostras para a pesquisa:

- cada uma das capitais e o Distrito Federal foram definidos como um estrato geográfico;
- os demais municípios foram agrupados em 26 estratos geográficos, representando cada uma das Unidades da Federação, excluídas as respectivas capitais.

Quadro 1 - Comparação entre as principais características das duas amostras da PeNSE 2015

Principais características	Amostras	
	1	2
População-alvo	Escolares frequentando o 9º ano (antiga 8ª série) do ensino fundamental	Escolares na faixa etária de 13 a 17 anos
População pesquisada	Escolares frequentando o 9º ano (antiga 8ª série) do ensino fundamental (turnos matutino e vespertino) de escolas públicas e privadas com pelo menos 15 escolares matriculados na série escolhida	Escolares frequentando do 6º ao 9º ano do ensino fundamental (antigas 5ª a 8ª séries) e da 1ª a 3ª série do ensino médio (turnos matutino, vespertino e noturno), de escolas públicas e privadas com pelo menos 15 escolares matriculados no conjunto de séries escolhidas
Unidade informante	Todos os escolares das turmas selecionadas para a amostra e o responsável pela escola	Todos os escolares das turmas selecionadas para a amostra e o responsável pela escola
Unidade de análise	Todos os escolares das turmas selecionadas para a amostra que concordaram em participar da pesquisa e informaram sexo e idade	Todos os escolares de 13 a 17 anos das turmas selecionadas para a amostra que concordaram em participar da pesquisa e informaram sexo e idade
Abrangência geográfica	Brasil, Grandes Regiões, Unidades da Federação e Municípios das Capitais	Brasil e Grandes Regiões
Comparação recomendada	Municípios das Capitais e Distrito Federal para as edições de 2009 e 2012 da pesquisa, e somente Grandes Regiões e Brasil, para a edição de 2012	Indicadores por grupos de idade (13 a 17 anos), conforme a Global School-based Student Health Survey - GSHS
Total planejado de municípios na amostra	675	179
Total planejado de escolas na amostra	3 160	380
Total de turmas selecionadas na amostra	4 159	653
Questionários coletados	102 301	16 608
Questionários válidos (1)	102 072	16 556

Fonte: IBGE, Pesquisa Nacional de Saúde do Escolar 2015.

(1) Foi considerado válido o questionário do estudante que concordou em participar da pesquisa e informou seu sexo e idade.

Em cada um dos 53 estratos geográficos foi dimensionada e selecionada uma amostra de escolas de um cadastro formado pelas escolas com turmas do 9º ano do Ensino Fundamental, segundo as informações fornecidas pelo Censo Escolar 2013. Foram excluídas do cadastro as escolas com menos de 15 alunos de 9º ano do ensino fundamental, bem como as turmas do turno noturno. Após a seleção as escolas foram visitadas e foi feita a atualização das informações das suas turmas de 9º ano, para a seleção de uma amostra de turmas e, como nas pesquisas anteriores, todos os alunos das turmas selecionadas foram convidados a responder ao questionário. A seleção das turmas foi feita por amostragem aleatória simples entre as turmas de cada escola da amostra.

Nos estratos geográficos foram criados os estratos de alocação, considerando a dependência administrativa (pública ou privada) das escolas e sua classe de tamanho em termos de turmas de 9º ano, considerando as escolas com até duas turmas e as com três ou mais turmas. Assim cada estrato geográfico teria quatro estratos de alocação, porém em alguns dos 53 estratos geográficos não existiam escolas em todos os estratos de alocação, o que levou à definição de 207 destes estratos.

Para os estratos geográficos formados pelas capitais e Distrito Federal, a seleção das escolas foi direta em cada estrato de alocação, com probabilidades proporcionais ao número de turmas de 9º ano do Ensino Fundamental informado no Censo Escolar.

Para os demais estratos, com o intuito de controlar o espalhamento geográfico das escolas selecionadas, as Agências do IBGE, que se constituem de grupos de municípios vizinhos e são unidades utilizadas no gerenciamento das atividades de campo das pesquisas, foram utilizadas como uma das fases de seleção. Em cada Unidade da Federação foram selecionadas 20% das Agências, sendo no mínimo duas por Unidade da Federação quando possível, já que em algumas Unidades menos populosas só existe uma Agência. A seleção foi feita com probabilidades proporcionais ao número de turmas de 9º ano da Agência. Após a seleção foi necessário fazer algumas agregações de Agências, pois em alguns estratos de alocação das mesmas não existiam, ou existiam poucas escolas, notadamente nos estratos de escolas privadas. Após essas agregações, a seleção das escolas foi feita em cada estrato de alocação definido, da mesma forma que nos estratos das capitais e Distrito Federal.

A determinação dos tamanhos das amostras para os estratos geográficos foi feita de maneira similar às pesquisas anteriores, e alocadas aos estratos de alocação de maneira proporcional.

A Tabela 3 mostra os valores dos tamanhos das amostras calculados e os valores efetivamente coletados para cada estrato geográfico da PeNSE 2015, Amostra 1.

Vê-se que houve uma perda de aproximadamente 3,8% das escolas, que não foram coletadas por diversos motivos: não possuíam turmas do 9º ano no momento da coleta, estavam desativadas, recusaram participar da pesquisa, impossibilidade de acesso, entre outros. O número de turmas coletadas sofreu um decréscimo de 5,9% em relação ao número esperado. Em relação aos estudantes, era esperado que o número de alunos que costumavam frequentar as aulas fosse menor que o número de matriculados, e que o número de respondentes da pesquisa fosse menor que o de frequentes (14,9%).

Para a determinação dos pesos amostrais, como nas pesquisas anteriores, foram considerados os questionários dos estudantes que concordaram em participar da pesquisa e informaram as variáveis sexo e idade. Os 229 casos de questionários não válidos foram mantidos na base de dados, para efeito de documentação, com peso amostral zero. A seleção das escolas foi feita baseada num cadastro formado a partir das informações do Censo Escolar 2013. Ao final da coleta dos dados, em 2015, já estavam disponíveis os primeiros resultados do Censo Escolar 2015, que serviram para a atualização do cadastro da pesquisa.

Tabela 3 - Tamanho das amostras de escolas, turmas e alunos matriculados no 9º ano do ensino fundamental, segundo as Unidades da Federação e respectivos estratos geográficos - Amostra 1 da PeNSE 2015

(continua)

Unidades da Federação	Estratos Geográficos	Calculados		
		Escolas	Turmas	Alunos matriculados
Brasil		**3 160**	**4 418**	**128 027**
Rondônia	Porto Velho	46	73	2 188
	Interior	66	92	2 249
Acre	Rio Branco	42	71	2 242
	Interior	64	81	2 077
Amazonas	Manaus	37	59	2 139
	Interior	58	73	1 968
Roraima	Boa Vista	47	85	2 224
	Interior	62	77	1 659
Pará	Belém	46	65	2 262
	Interior	57	78	2 385
Amapá	Macapá	56	90	2 710
	Interior	57	80	2 364
Tocantins	Palmas	43	60	1 961
	Interior	70	81	2 165
Maranhão	São Luís	56	79	2 660
	Interior	84	101	2 821
Piauí	Teresina	58	76	2 317
	Interior	85	95	2 318
Ceará	Fortaleza	50	72	2 127
	Interior	72	86	2 303
Rio Grande do Norte	Natal	59	82	2 425
	Interior	82	97	2 685
Paraíba	João Pessoa	68	87	2 757
	Interior	83	107	2 964
Pernambuco	Recife	60	87	2 864
	Interior	69	93	2 981
Alagoas	Maceió	42	60	2 127
	Interior	53	67	2 122
Sergipe	Aracaju	58	81	2 478
	Interior	80	94	2 595
Bahia	Salvador	44	73	2 148
	Interior	61	84	2 369
Minas Gerais	Belo Horizonte	64	101	2 872
	Interior	79	112	3 180
Espírito Santo	Vitória	64	89	2 355
	Interior	78	111	2 978
Rio de Janeiro	Rio de Janeiro	54	78	2 427
	Interior	64	89	2 568
São Paulo	São Paulo	44	76	2 434
	Interior	48	80	2 504
Paraná	Curitiba	43	74	2 162
	Interior	55	86	2 475
Santa Catarina	Florianópolis	52	83	2 433
	Interior	74	107	2 744
Rio Grande do Sul	Porto Alegre	53	68	1 727
	Interior	64	77	1 876
Mato Grosso	Cuiabá	45	63	1 881
	Interior	59	75	1 959
Mato Grosso do Sul	Campo Grande	41	65	1 898
	Interior	57	82	2 144
Goiás	Goiânia	71	104	3 399
	Interior	86	124	3 454
Distrito Federal	Brasília	50	88	2 903

Tabela 3 - Tamanho das amostras de escolas, turmas e alunos matriculados no 9º ano do ensino fundamental, segundo as Unidades da Federação e respectivos estratos geográficos - Amostra 1 da PeNSE 2015

(conclusão)

Unidades da Federação	Estratos Geográficos	Coletados				
		Escolas	Turmas	Alunos		
				Matriculados	Frequentes	Respondentes
Brasil		**3 040**	**4 159**	**124 227**	**120 122**	**102 301**
Rondônia	Porto Velho	46	72	2 119	2 026	1 732
	Interior	62	73	1 812	1 747	1 506
Acre	Rio Branco	41	65	2 039	2 003	1 723
	Interior	60	73	1 825	1 820	1 526
Amazonas	Manaus	35	57	2 106	1 992	1 645
	Interior	57	73	2 013	1 970	1 716
Roraima	Boa Vista	44	78	2 239	2 182	1 838
	Interior	60	73	1 615	1 491	1 280
Pará	Belém	46	64	2 312	2 189	1 864
	Interior	54	75	2 363	2 320	1 972
Amapá	Macapá	51	85	2 590	2 551	2 153
	Interior	53	76	2 187	2 116	1 867
Tocantins	Palmas	40	52	1 766	1 697	1 478
	Interior	65	75	2 060	2 008	1 677
Maranhão	São Luís	54	71	2 259	2 220	1 920
	Interior	80	95	2 514	2 490	2 115
Piauí	Teresina	58	74	2 320	2 319	2 019
	Interior	84	91	2 221	2 169	1 890
Ceará	Fortaleza	46	68	2 020	1 965	1 644
	Interior	69	84	2 362	2 317	2 018
Rio Grande do Norte	Natal	59	74	2 334	2 333	1 948
	Interior	80	98	2 998	2 973	2 413
Paraíba	João Pessoa	68	89	3 036	2 854	2 423
	Interior	79	102	3 087	2 841	2 381
Pernambuco	Recife	56	81	2 720	2 642	2 168
	Interior	67	96	3 059	2 992	2 391
Alagoas	Maceió	41	60	2 063	2 005	1 694
	Interior	49	59	2 079	2 041	1 694
Sergipe	Aracaju	54	74	2 305	2 204	1 893
	Interior	79	87	2 432	2 389	2 097
Bahia	Salvador	44	69	2 187	2 096	1 750
	Interior	58	78	2 364	2 311	1 968
Minas Gerais	Belo Horizonte	62	101	3 151	3 087	2 692
	Interior	76	107	3 094	3 023	2 575
Espírito Santo	Vitória	62	83	2 273	2 251	1 962
	Interior	76	108	3 037	3 024	2 632
Rio de Janeiro	Rio de Janeiro	51	79	2 503	2 483	2 123
	Interior	62	90	2 753	2 640	2 134
São Paulo	São Paulo	43	74	2 356	2 175	1 899
	Interior	48	75	2 311	2 145	1 803
Paraná	Curitiba	42	68	2 092	2 030	1 771
	Interior	53	83	2 449	2 390	2 099
Santa Catarina	Florianópolis	51	70	1 947	1 874	1 650
	Interior	73	90	2 262	2 195	1 965
Rio Grande do Sul	Porto Alegre	46	52	1 316	1 200	1 015
	Interior	62	68	1 671	1 595	1 373
Mato Grosso	Cuiabá	44	60	1 886	1 799	1 536
	Interior	56	70	1 994	1 829	1 543
Mato Grosso do Sul	Campo Grande	39	65	2 003	1 898	1 605
	Interior	56	80	2 124	1 971	1 632
Goiás	Goiânia	68	97	3 295	3 140	2 634
	Interior	83	114	3 423	3 237	2 731
Distrito Federal	Brasília	48	84	2 881	2 863	2 524

Fontes: 1. IBGE, Pesquisa Nacional de Saúde do Escolar 2015. 2. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, Censo Escolar 2013.

A partir dos dados do Censo Escolar 2015 foi construído um cadastro atualizado utilizando os mesmos filtros que serviram para construção do cadastro de seleção da amostra de escolas. Nesse processo não foram encontradas 50 escolas pertencentes à amostra selecionada, cujos dados foram coletados e correspondiam a 1 300 questionários. Algumas dessas escolas constavam das bases do INEP como paralisadas, extintas ou com informações que as excluiriam do cadastro da Amostra 1, porém, no campo, foram localizadas e pesquisadas. Decidiu-se por incluir tais escolas no novo cadastro com as informações originais do cadastro de seleção. Com o cadastro atualizado foi possível calcular os pesos amostrais das escolas incorporando o tratamento de não resposta pela ausência de informações para as 120 escolas selecionadas que não foram pesquisadas. Foi apurado o número efetivo de escolas pesquisadas por estrato de expansão, definidos da mesma forma que os estratos de alocação. Em quatro estratos de expansão não houve questionário preenchido. As escolas desses estratos foram agrupadas no cadastro e na amostra em estratos que levaram em conta apenas o estrato geográfico e a dependência administrativa.

Para os estratos geográficos formados pelos municípios fora das capitais, os pesos foram calculados de forma aproximada supondo a seleção de escolas diretamente dos estratos geográficos. Assim as escolas foram consideradas como unidade primária de seleção em todos os casos. Para a correção dos pesos devido à perda de unidades amostrais, foram considerados os tamanhos de amostras efetivamente pesquisadas nos estratos.

O peso do aluno *k* da turma *j* da escola *i* de um determinado estrato de expansão, foi calculado como:

$$w_{ijk} = w_i \, w_{ij} \frac{M^*_{ij}}{v_{ij}}$$

onde:

w_i é o peso da escola *i* do estrato, dado por:

$$w_i = \frac{T}{nT_i}$$

T é o número de total de turmas de 9º ano cadastradas no estrato;
n é o tamanho da amostra de escolas no estrato; e
T_i é o número de turmas de 9º ano cadastradas na escola *i* do estrato;
w_{ij} é o peso da turma *j* da escola *i* do estrato, dado por:

$$w_{ij} = \frac{T^*_i}{t_i}$$

T^*_i é o número de turmas de 9º ano da escola *i* do estrato, atualizado pela pesquisa;
t_i é o número de turmas de 9º ano selecionadas na escola *i* do estrato;
M^*_{ij} é o número de alunos matriculados no 9º na turma *j* da escola *i* do estrato, atualizado pela pesquisa; e
v_{ij} é o número de questionários válidos da turma *j* da escola *i* do estrato.

Para que as estimativas de totais de alunos matriculados no 9º ano do Ensino Fundamental refletissem os totais apurados pelo Censo Escolar 2015, decidiu-se pela calibração dos pesos amostrais.

Os pesos calibrados, para um determinado estrato geográfico, foram obtidos pela fórmula:

$$wc_{ijk} = w_{ijk} \frac{M}{\hat{M}}$$

onde:

w_{cijk} é o peso calibrado para o aluno *k* da turma *j* da escola *i* da amostra do estrato;

w_{ijk} é o peso sem calibração para o aluno *k* da turma *j* da escola *i* da amostra do estrato;

M é o número de matriculados no 9º ano no estrato geográfico, dado pelo Censo Escolar 2015; e

$\hat{M}$ é o número de matriculados no 9º ano no estrato geográfico, estimado pela pesquisa utilizando os pesos sem calibração.

Os pesos calibrados sofreram um último ajuste para que estimem o número de alunos matriculados no 9º ano do Ensino Fundamental que habitualmente frequentavam as aulas, como nas PeNSE 2009 e 2012. Esses foram os pesos finais atribuídos aos questionários considerados válidos de acordo com os critérios já citados anteriormente. Foram calculados utilizando a fórmula:

$$w^{*}_{ijk} = wc_{ijk} \frac{F_{ij}}{M^{*}_{ij}}$$

onde:

w^{*}_{ijk} é o peso ajustado para o aluno *k* da turma *j* da escola *i* do estrato;

wc_{ijk} é o peso calibrado para o aluno *k* da turma *j* da escola *i* da do estrato;

F_{ij} é o número de alunos frequentes da turma *j* da escola *i* do estrato;

M^{*}_{ij} é o número de alunos matriculados, atualizado pela pesquisa, da turma *j* da escola *i* do estrato.

Na PeNSE 2015 foi selecionada uma segunda amostra, referenciada como Amostra 2, com o objetivo de estimar porcentagens ou prevalências para as cinco grandes regiões geográficas e Brasil, para estudantes na faixa etária de 13 a 17 anos. Portanto as Regiões Norte, Nordeste, Sudeste, Sul e Centro-Oeste foram definidas como estratos geográficos. O planejamento amostral da Amostra 2 foi semelhante ao da Amostra 1, inclusive no que diz respeito à utilização das Agências do IBGE no controle do espalhamento da amostra. As escolas selecionadas para esta amostra foram visitadas e foi feita a atualização cadastral relativa às turmas de alunos do 6º ano da Ensino Fundamental até a 3ª série de Ensino Médio. Para cada escola da amostra foi selecionada uma amostra de turmas e todos os seus estudantes foram convidados a responder ao questionário da pesquisa, idêntico ao da Amostra 1.

O Quadro 2 relaciona todas as etapas de ensino consideradas na seleção da Amostra 2.

Quadro 2 - Etapas de ensino consideradas para seleção das escolas da Amostra 2 da PeNSE 2015

Etapas de ensino	
Código	Descrição
8	Ensino fundamental de 8 anos 5ª série
9	Ensino fundamental de 8 anos 6ª série
10	Ensino fundamental de 8 anos 7ª série
11	Ensino fundamental de 8 anos 8ª série
19	Ensino fundamental de 9 anos 6º ano
20	Ensino fundamental de 9 anos 7º ano
21	Ensino fundamental de 9 anos 8º ano
41	Ensino fundamental de 9 anos 9º ano
25	Ensino médio 1ª série
26	Ensino médio 2ª série
27	Ensino médio 3ª série
29	Ensino médio não seriado
30	Curso técnico integrado (Ensino médio integrado) 1ª série
31	Curso técnico integrado (Ensino médio integrado) 2ª série
32	Curso técnico integrado (Ensino médio integrado) 3ª série
35	Ensino médio normal/magistério 1ª série
36	Ensino médio normal/magistério 2ª série
37	Ensino médio normal/magistério 3ª série

Fonte: IBGE, Pesquisa Nacional de Saúde do Escolar 2015.

Os tamanhos de amostra foram calculados para cada estrato geográfico de maneira análoga à Amostra 1, também considerando um erro amostral de 3%, em valor absoluto, para estimar proporções da ordem de 50%, com nível de confiança de 95%, e um efeito de conglomeração médio aproximado de 3. As amostras foram alocadas nos estratos de alocação que levaram em conta o estrato geográfico, a dependência administrativa (pública ou privada) e o tamanho da escola medido pelo número de turmas das etapas de ensino de interesse (menos de 8 turmas ou 8 turmas ou mais).

Os tamanhos calculados e efetivos das amostras por região geográfica são mostrados na Tabela 4.

Tabela 4 - Tamanho das amostras de escolas, turmas e alunos matriculados, do 6º ano do ensino fundamental à 3ª série do ensino médio, segundo as grandes regiões geográficas - Amostra 2

Grandes Regiões	Calculado			Coletado				
	Escolas	Turmas	Alunos matriculados	Escolas	Turmas	Alunos		
						Matriculados	Frequentes	Respondentes
Brasil	**380**	**652**	**19 558**	**371**	**653**	**20 516**	**19 402**	**16 608**
Norte	76	131	4 150	74	127	4 127	3 777	3 195
Nordeste	77	127	3 991	75	131	4 305	4 161	3 478
Sudeste	70	125	3 832	68	122	4 052	3 849	3 292
Sul	82	143	3 803	81	141	3 954	3 704	3 219
Centro-Oeste	75	126	3 782	73	132	4 078	3 911	3 424

Fontes: 1. IBGE, Pesquisa Nacional de Saúde do Escolar 2015. 2. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, Censo Escolar 2013.

Não foram coletados os dados referentes a nove escolas, aproximadamente 2,4% das escolas selecionadas. Já o número de turmas selecionadas superou o previsto em duas das 5 regiões geográficas. O número de matrículas informado pelas escolas também superou em 4,9% o esperado. Segundo as escolas da amostra visitadas pela pesquisa, o número de alunos que frequentam habitualmente as aulas era 5,4% menor que os matriculados. O número de respondentes da pesquisa foi 14,4% menor que o de alunos frequentes. A seleção das escolas para as duas amostras da PeNSE 2015 foi feita de maneira independente, sendo que 31 escolas e oito turmas foram selecionadas para ambas amostras. Essa coincidência foi irrelevante para os 194 alunos envolvidos, pois cada um respondeu apenas um questionário que na apuração foi incluído nas duas bases de dados.

Para o cálculo dos pesos da Amostra 2, foram desprezados 52 questionários de alunos que se recusaram a participar, não informaram idade ou sexo. Na base de dados os registros correspondentes foram incluídos com peso zero.

Os pesos amostrais foram obtidos de forma semelhante aos da Amostra 1 e calibrados por idade dentro de cada estrato geográfico, de acordo com as informações do Censo Escolar 2015. Assim, os pesos calibrados foram determinados por:

$$wc_{ijkl} = w_{ijkl} \frac{M_l}{\hat{M}_l}$$

onde:

w_{cijkl} é o peso calibrado para o aluno k da idade l da turma j da escola i do estrato geográfico;

w_{ijkl} é o peso sem calibração para o aluno k da idade l da turma j da escola i do estrato geográfico;

M_l é o número de alunos matriculados da idade *l* do estrato geográfico, dado pelo Censo Escolar 2015; e

$\hat{M}_l$ é o número de alunos matriculados da idade *l* do estrato geográfico, estimado pela pesquisa utilizando os pesos não calibrados.

Os pesos calibrados foram, em seguida, ajustados para fornecer estimativas para os alunos matriculados e que compareciam habitualmente às aulas. Para isso os pesos calibrados dos alunos foram multiplicados pelo quociente entre o número de alunos frequentes de sua turma e o número de alunos matriculados na turma, atualizados pela pesquisa. A fórmula aplicada, em cada estrato geográfico, para gerar os pesos definitivos para cada questionário válido da pesquisa foi $w^*_{ijkl} = wc_{ijkl} F_{ij} / M^*_{ij}$.

Considerações finais

A PeNSE é uma pesquisa que inovou com a introdução de um novo método de coleta de informações onde o aluno selecionado para responder ao questionário da pesquisa o faz sem a necessidade da intermediação de um entrevistador, o que pode auxiliar na questão do sigilo e eliminar algum possível constrangimento no caso de questões que possam ser consideradas mais delicadas por parte do respondente.

A implantação da pesquisa de forma paulatina em relação à abrangência geográfica da mesma, permitiu ajustes no plano amostral e nas questões operacionais da pesquisa.

Para as futuras edições, nos parece fundamental que se discuta qual deve ser a periodicidade para a aplicação da pesquisa que seja adequada para a captação das mudanças nos hábitos e fatores de risco de interesse a que estão sujeitos os adolescentes, alvo do estudo.

Referências

COCHRAN, W. G. *Sampling techniques*. 3rd ed. New York: Wiley, c1977. 428 p. (Wiley series in probability and mathematical statistics). Disponível em: <https://archive.org/details/Cochran1977SamplingTechniques_201703>. Acesso em: set. 2017.

PESQUISA nacional de saúde do escolar 2009. Rio de Janeiro: IBGE, 2009. 144 p. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/pense/default.shtm>. Acesso em: set. 2017.

PESQUISA nacional de saúde do escolar 2012. Rio de Janeiro: IBGE, 2013. 254 p. Acompanha 1 CD-ROM. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/pense/2012/default.shtm>. Acesso em: set. 2017.

PESQUISA nacional de saúde do escolar 2015. Rio de Janeiro: IBGE, 2016. 126 p. Disponível em: <https://www.ibge.gov.br/home/estatistica/populacao/pense/2015/default.shtm>. Acesso em set. 2017.

PESSOA, D. G. C.; SILVA, P. L. do N. Análise de dados amostrais complexos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 13., 1998, Caxambu. *Anais...* São Paulo: Associação Brasileira de Estatística - ABE, 1998. 170 p. Disponível em: <http://www.ie.ufrj.br/download/livro.pdf>. Acesso em: set. 2017.

SILVA, P. L. do N.; MOURA, F. A. da S. E*feito de conglomeração da malha setorial do censo demográfico de 1980*. Rio de Janeiro: IBGE, 1986. 115 p. (Textos para discussão, n. 32). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv66426.pdf>. Acesso em: set. 2017.

SITUAÇÃO mundial da infância 2011: relatório. Brasília, DF: Fundo das Nações Unidas para a Infância - Unicef, 2011. 137 p. Disponível em: <http://www.unicef.org/brazil/pt/br_sowcr11web.pdf>. Acesso em: set. 2017.

Qualidade estatística no IBGE

Maria Luiza Barcellos Zacharias

Introdução

A qualidade das informações e estatísticas produzidas sempre foi um compromisso e uma preocupação constante no IBGE. Como principal produtor de estatísticas nacionais, o instituto provê, para diversos segmentos da sociedade civil e para órgãos governamentais de todas as esferas, informações que são balizadoras e, muitas vezes, determinantes na tomada de decisões por parte de seus usuários. A informação é o principal produto do IBGE e, portanto, sua qualidade deve ser assegurada e acrescida de atributos para garantir sua confiabilidade.

Uma abordagem sistemática e mais estruturada voltada para o controle e a gestão da qualidade nos institutos de estatística é relativamente recente. Tomando por base os Princípios Fundamentais das Estatísticas Oficiais (Fundamental Principles of National Official Statistics), adotados na 28ª sessão da Comissão de Estatística das Nações Unidas, realizada em 1994 e endossados pela Assembleia Geral em 2014 (UNITED NATIONS, 2014) vários arcabouços foram desenvolvidos por organizações internacionais produtoras de estatísticas, com o objetivo de estruturar modelos de garantia de qualidade. Um passo importante neste sentido foi a criação de um grupo formado por especialistas dos principais institutos europeus, Leadership Group on Quality, que de 1999 a 2001 discutiu e propôs um conjunto de recomendações e propostas visando ao desenvolvimento de um framework consistente para lidar com aspectos de qualidade. Um dos principais produtos resultantes deste trabalho foi o Código de Conduta das Estatísticas Europeias (European Statistics Code of Practice), adotado pelo Sistema Estatístico Europeu (European Statistical System - ESS), em 2005[1], que embasou, posteriormente, a elaboração de instrumento semelhante voltado para a região da América Latina e Caribe.

Acompanhando esse movimento internacional, o IBGE tem dado alguns passos importantes com interesse específico em garantir a qualidade de seus produtos e processos, em suas diversas dimensões, tais como relevância, precisão, pontualidade, comparabilidade, coerência, acessibilidade e a transparência. Um desses passos foi a criação, em 2007, de uma unidade

[1] Esse documento foi revisado em 2011, com uma versão em português (EUROSTAT, 2011).

organizacional dedicada especificamente para lidar com aspectos relacionados à qualidade no Instituto, inserida na atual unidade responsável pelo estudo, aplicação e difusão de metodologias estatísticas, a Coordenação de Métodos e Qualidade. Outro marco importante para fomentar a contínua aplicação dos melhores métodos e práticas na produção estatística foi a edição do Código de Boas Práticas das Estatísticas do IBGE.

Este artigo tem por objetivo apresentar algumas relevantes iniciativas que foram adotadas e outras que estão em andamento para manter elevada a qualidade da informação produzida pelo IBGE, em especial ao longo dos últimos 10 anos, desde a implantação da unidade responsável pela gestão da qualidade estatística no IBGE.

Criação da Gerência de Qualidade Estatística

De modo a instituir formalmente a gestão da qualidade estatística, a partir de setembro de 2007, o IBGE passou a contar, em sua estrutura, com uma unidade responsável pelas atividades relacionadas à qualidade dos processos e produtos estatísticos, subordinada à Coordenação de Métodos e Qualidade, novo nome do Departamento de Metodologia. O Conselho Diretor do IBGE, por meio da Resolução RCD n. 13, de 02.06.2015, instituiu formalmente a Gerência de Qualidade Estatística com as atribuições de "pesquisar, avaliar, adaptar, propor e difundir métodos e técnicas relacionados com a qualidade dos processos de produção, análise e disseminação de informações estatísticas; e coordenar as atividades de implementação desses métodos e técnicas".

O desenvolvimento das atividades da recém-criada unidade foi impulsionado pela participação intensa nas discussões do Grupo de Trabalho 8 – Qualidade Total das Estatísticas, do Projeto de Cooperação CE-Mercosul em Matéria Estatística II, entre a União Europeia e países do Mercosul. Este grupo de trabalho foi criado a partir de recomendações de um estudo metodológico sobre sistema de indicadores de qualidade para os institutos de estatística, desenvolvido entre 2002 e 2003, pelos países envolvidos no âmbito do primeiro projeto de cooperação União Europeia-Mercosul. O objetivo principal desse projeto era contribuir para o processo de integração do Mercosul e ao fortalecimento de suas instituições nacionais e regionais, mediante o desenvolvimento, aperfeiçoamento, harmonização e integração da produção e divulgação de estatísticas. Dentre as atividades do Grupo de Trabalho 8 estava o desenvolvimento, sob o ponto de vista técnico e metodológico, de ações voltadas para a adoção de práticas enquadradas no conceito de gestão de qualidade por parte dos Institutos Nacionais de Estatística. O principal produto resultante desse trabalho conjunto foi a definição de um conjunto reduzido de indicadores, a ser utilizado para medir a qualidade dos dados produzidos e realizar seu acompanhamento ao longo do tempo, com base nas características particulares de cada Instituto Nacional de Estatística do Mercosul e na factibilidade de aplicação dos mesmos. Este produto foi formalizado por meio da publicação *Conjunto mínimo de indicadores padrão de qualidade a ser aplicado no Mercosul*, editado pela Diretoria de Pesquisas do IBGE (CONJUNTO..., 2014).

Além das atividades descritas acima, e logo após sua instalação em 2007, a Gerência de Qualidade Estatística também investiu fortemente na melhoria do sistema de metadados e na padronização da documentação das operações estatísticas visando sua futura disseminação na Internet.

A documentação é uma atividade amplamente reconhecida como um aspecto relacionado à qualidade. É uma ferramenta para aprimorar a qualidade dos processos estatísticos bem como para ajudar os usuários a recuperar as informações de interesse. Por isso, os Institutos Nacionais de Estatística têm investido cada vez mais no desenvolvimento de sistemas de informação para gestão dos metadados de suas operações estatísticas.

Metadados podem ser entendidos como "dados que descrevem os dados", ou seja, são informações e detalhamentos necessários para a efetiva compreensão dos dados. O registro dessas informações permite que usuários avaliem a forma como o dado foi obtido e sua qualidade, o que amplia seu potencial de uso e preserva a memória e o conhecimento sobre o fazer da pesquisa.

Assim, buscando estar aderente às dimensões de qualidade estatística de Acessibilidade e Transparência, em 2009, o sistema de informática foi concluído permitindo disponibilizar, no portal do IBGE, os metadados das operações estatísticas mais recentes em um relatório padrão de documentação. Até então, os metadados estavam disponíveis apenas para uso interno, na Intranet da Instituição. O sistema de metadados viria a ser reformulado a partir de 2014, ampliando seu escopo e funcionalidades, e contribuindo para modificar, de forma definitiva, a cultura interna sobre a importância dos metadados na produção estatística.

O Código de Boas Práticas das Estatísticas do IBGE

Dando prosseguimento aos trabalhos iniciados em 2007 pela Gerência de Qualidade Estatística, em 2013, o IBGE publicou o Código de Boas Práticas das Estatísticas do IBGE (IBGE, 2013), divulgando um conjunto de diretrizes, princípios e indicadores sobre as melhores práticas adotadas pelo Instituto na produção de estatísticas oficiais, tomando como referência o *Código regional de buenas prácticas em estadística para América Latina y el Caribe*, da Comissão Econômica para a América Latina e o Caribe (Comisión Económica para América Latina y el Caribe - Cepal). O Código estabelece 17 princípios e 80 indicadores, em três seções: ambiente institucional e coordenação, processos estatísticos e produtos estatísticos, que podem ser utilizados para monitorar a sua implementação, bem como para acompanhar e avaliar a qualidade da produção estatística do IBGE pela sociedade.

Os princípios do Código de Boas Práticas das Estatísticas do IBGE, que refletem os Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas, são agrupados nas três seções, como mostra o Quadro 1:

Quadro 1 - Seções e princípios do Código de Boas Práticas das Estatísticas do IBGE

Seções	Princípios
1 - Ambiente institucional e coordenação	1 - Independência institucional 2 - Coordenação do Sistema Estatístico Nacional 3 - Mandato estatístico de coleta de dados 4 - Confidencialidade estatística 5 - Uso eficiente dos recursos 6 - Compromisso com a qualidade 7 - Imparcialidade e objetividade 8 - Cooperação e participação internacional
2 - Processos Estatísticos	9 - Metodologia sólida 10 - Processos estatísticos adequados 11 - Solicitação de informação não excessiva 12 - Relação entre custo e eficácia
3 - Produtos estatísticos	13 - Relevância 14 - Precisão e acurácia 15 - Oportunidade e pontualidade 16 - Coerência e comparabilidade 17 - Acessibilidade e transparência

Fonte: IBGE. Código de boas práticas das estatísticas do IBGE. Rio de Janeiro, 2013. 48 p. Disponível em: <ftp://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>. Acesso em: set. 2017.

Para cada princípio é estabelecido um conjunto de indicadores de boas práticas. O princípio 6, que trata do Compromisso com a Qualidade, estabelece: "O IBGE deve trabalhar, coordenar e cooperar com as demais entidades produtoras de estatísticas no âmbito do Sistema Estatístico Nacional, de acordo com normas, princípios e padrões internacionais" (IBGE, 2013, p. 23).

Os seguintes indicadores de boas práticas foram definidos para esse princípio:

> 6.1 O IBGE deve ter uma política e um modelo de qualidade claramente definidos e documentados.
>
> 6.2 A qualidade do produto estatístico deve ser avaliada periodicamente, de acordo com orientações internas e normas internacionais.
>
> 6.3 Deve-se promover e incentivar, sistematicamente, uma cultura de melhoria contínua da produção estatística do Instituto.
>
> 6.4 Deve haver processos documentados para avaliar e controlar a qualidade em cada uma das etapas do processo estatístico do Instituto (IBGE, 2013, p. 23).

O desenvolvimento do Código de Boas Práticas das Estatísticas do IBGE como elemento central de sua política de qualidade, por si só, demonstra a preocupação do IBGE com a qualidade de seus produtos. Contudo, para buscar a efetividade das práticas apresentadas, é importante o desdobramento do Código em um plano de trabalho que objetive o monitoramento da qualidade e o aprimoramento contínuo do desempenho do Instituto nos aspectos avaliados.

Nesse sentido, alguns projetos e atividades começaram a ser desenvolvidos em 2016, visando garantir a aderência do Instituto ao seu Código, bem como para institucionalizar uma cultura de monitoramento da qualidade de cada operação estatística em seus diferentes estágios de produção. Um desses projetos, já em andamento, consiste na elaboração de textos explicativos sobre o contexto e o objetivo de cada indicador, com a definição de critérios mensuráveis – qualitativos e quantitativos – para seu acompanhamento e avaliação. Com esses critérios definidos, torna-se possível estabelecer a rotina de monitoramento e de avaliação periódica da aderência ao Código, buscando-se garantir, assim, a qualidade dos produtos e processos estatísticos de forma sistemática.

O Código de Boas Práticas das Estatísticas do IBGE também pode ser considerado como um importante passo no sentido de melhorar a governança e a consistência das estatísticas oficiais no Brasil, pois além de fomentar a discussão sobre a cultura de qualidade na produção estatística dos diversos órgãos do Sistema Estatístico Nacional, também subsidia a criação de uma versão mais completa do Código, abrangendo a produção estatística de todo o sistema. Especificamente, o objetivo é introduzir um entendimento comum da qualidade entre todos os produtores de estatísticas no Brasil e também alinhar as práticas nacionais às normas internacionais. Para a realização desta futura edição do Código de Boas Práticas do Sistema Estatístico Nacional, o IBGE tem promovido fóruns de discussão com outras instituições responsáveis pela produção de estatísticas oficiais, inicialmente no nível do governo federal.

Dando seguimento à divulgação do Código de Boas Práticas das Estatísticas do IBGE, um conjunto de guias, contendo diretrizes e protocolos, foram elaborados e divulgados para fornecer orientação e apoio relacionados a aspectos específicos do processo de produção e disseminação de informações, de acordo com os princípios estabelecidos pelo Código.

Guias e protocolos

Um dos principais objetivos do IBGE é garantir que as informações produzidas sejam disponibilizadas de forma imparcial para honrar o direito dos cidadãos à informação pública, de acordo com o primeiro dos Princípios Fundamentais das Estatísticas Oficiais - Relevância, imparcialidade e igualdade de acesso. Para divulgar amplamente este compromisso entre seus técnicos e cidadãos brasileiros, em outubro de 2014, o IBGE disponibilizou, no seu Portal da Internet, o documento intitulado *Princípios Fundamentais das Estatísticas Oficiais: orientações para divulgações de resultados pelo IBGE*, contendo algumas

recomendações básicas que se relacionam com a apresentação e divulgação de estatísticas oficiais, tanto na forma escrita como falada, quanto ao aspecto da imparcialidade, visando à implementação de critérios que garantam sua observância (IBGE, 2014). Essas orientações vão ao encontro do estabelecido no Código de Boas Práticas das Estatísticas do IBGE do IBGE, especificamente no que diz respeito a dois princípios, a saber: Imparcialidade e Objetividade (Princípio 7) e Acessibilidade e Transparência (Princípio 17).

Em abril de 2015, o IBGE publicou o guia *Procedimentos para lidar com erros de divulgação de dados e informações estatísticas*, a fim de garantir um gerenciamento uniforme de eventuais ocorrências de erros (IBGE, 2015). É importante definir e divulgar, para a sociedade em geral, procedimentos padronizados para lidar com erros de divulgação de informações, erros esses que ocorrem mesmo tendo sido consideradas todas as medidas para garantir a qualidade da produção das informações. De acordo com o Código de Boas Práticas das Estatísticas do IBGE, no seu princípio 15 – Oportunidade e Pontualidade, o indicador 15.4 estabelece que: "os erros descobertos nas estatísticas publicadas devem ser corrigidos o mais cedo possível e divulgados" (IBGE, 2013, p. 29). Para manter a confiança nas estatísticas oficiais em caso de erros, é importante que o IBGE reaja de forma compatível com a natureza e gravidade do erro, de maneira uniforme e transparente. Desta forma, o guia descreve os procedimentos a serem adotados, compreendendo uma tipologia de erros e a reação do Instituto a cada tipo de erro detectado, considerando as diferentes formas de disseminação de resultados estatísticos. Este protocolo foi fortemente inspirado por um guia semelhante publicado pelo Federal Statistical Office of Germany (2013), o instituto de estatística da Alemanha.

Além desses, foram aprovados em 2015 dois novos protocolos, reforçando o compromisso de estabelecer procedimentos padronizados, um voltado para a revisão de dados estatísticos e outro, para lidar com o mau uso dos dados estatísticos.

A *Política de revisão de dados divulgados das operações estatísticas do IBGE* descreve os procedimentos em relação ao planejamento, tempo de implementação e finalidade das mudanças nos dados estatísticos. Segundo o documento:

> Por revisão de dados entende-se toda e qualquer revisão programada de dados numéricos, em que são disponibilizadas novas informações, que não estavam acessíveis quando da primeira divulgação, tais como: um dado tardio que substitui uma não resposta; um dado corrigido pelo próprio informante; ou um conjunto de dados que foi submetido a processo de crítica e imputação (IBGE, 2015, p. 5).

E também, a inclusão de uma fonte de dados adicional ou a transição para um novo período-base, por exemplo.

No IBGE, cada operação estatística adota sua própria política de revisão de dados de acordo com o tipo de informação que produz. Assim, por exemplo, o Índice Nacional de Preços ao Consumidor - INPC, usado como índice oficial de inflação, não está sujeito a revisões, por razões legais. Por outro lado, algumas pesquisas de empresas revisam suas estimativas iniciais porque são incluídos novos dados – mais completos ou adicionais - que não estavam disponíveis quando da primeira divulgação das estatísticas. Assim, a política de revisão para os produtos sujeitos a revisões programadas e está divulgada no portal do IBGE na Internet, em conformidade com o que dispõe o Código de Boas Práticas das Estatísticas do IBGE em seu Princípio 17 – Acessibilidade e transparência, indicador 17.1: "deve-se garantir a todos os usuários o acesso livre e igualitário às estatísticas oficiais por meio de procedimentos claramente estabelecidos e conhecidos" (IBGE, 2013, p. 31).

O guia *Procedimentos para lidar com o mau uso dos dados e informações estatísticas e geoespaciais do IBGE*, define orientações relacionadas ao Princípio Fundamental das Estatísticas Oficiais 4, que estabelece que os órgãos estatísticos têm o direito de comentar sobre interpretação errônea e uso indevido de estatísticas, o que também é tratado no indicador 1.7 do Princípio 1 – Independência Institucional do Código de Boas Práticas das Estatísticas do IBGE: "As autoridades superiores da produção de estatística do IBGE, quando for o caso, devem realizar e orientar comentários públicos sobre questões estatísticas, incluindo críticas e usos inadequados das estatísticas oficiais" (IBGE, 2013, p. 19). Tipos de mau uso de dados e informações estatísticas incluem: descartar dados desfavoráveis, adulterar sentido ou interpretação de quesitos, desconsiderar ou interpretar de forma equivocada os conceitos usados para as variáveis envolvidas,

fazer generalização excessiva na análise de dados, fazer interpretação equivocada de relatórios de erros estimados, fazer uso de falsa causalidade, etc. Assim, este protocolo define os procedimentos a serem adotados pelo IBGE nos casos em que for detectado um mau uso, uma interpretação equivocada ou mesmo uma reação exagerada, por parte da mídia, da sociedade ou do governo, em relação aos dados e informações estatísticas que produz e divulga.

Novo Banco de Metadados - MetaBD do IBGE

De acordo com os princípios que regem as atividades estatísticas internacionais, faz parte das boas práticas, em matéria de acessibilidade e transparência, documentar os conceitos, definições e classificações, bem como os procedimentos utilizados ao longo das operações estatísticas, tornando essas informações acessíveis ao público. O Princípio 17 do Código de Boas Práticas das Estatísticas do IBGE, que trata de Acessibilidade e Transparência, em seu indicador 17.2, estabelece que "as estatísticas oficiais e seus respectivos metadados devem ser divulgados de forma clara e precisa aos usuários, a fim de facilitar sua correta interpretação e comparações significativas" (IBGE, 2013, p. 31).

Para tanto, as boas práticas recomendam o uso de normas e padrões internacionais, a fim de promover a comparabilidade dos dados produzidos por diferentes agências, como preconizado pelo Princípio Fundamental das Estatísticas Oficiais 9 – Uso de Padrões Internacionais: "A utilização de conceitos, classificações e métodos internacionais pelos órgãos de estatística de cada país promove a coerência e a eficiência dos sistemas de estatística em todos os níveis oficiais" (NAÇÕES UNIDAS, 1994, p. 3).

Sem padrões comuns, a comparabilidade dos dados produzidos por diferentes institutos, dentro e fora do país, seria impossível. A comparabilidade é uma dimensão importante da qualidade, pois dados não comparáveis, perdem utilidade, e o Instituto que os produz perde relevância.

Consistente com essas boas práticas, em 2014, o IBGE iniciou uma ampla reformulação de seu sistema de metadados estatísticos para um sistema mais abrangente, introduzindo uma série de melhorias. A principal delas se refere à atualização do sistema informático para uma arquitetura tecnológica mais moderna, reunindo, em uma única plataforma, diversos módulos e aplicações relacionados à consulta e atualização dos metadados, que até 2014 funcionavam de forma independente e com baixo nível de integração. Esta nova arquitetura do sistema, que unifica várias aplicações em um portal acessado via Intranet e Internet, permite a seus usuários o gerenciamento, atualização e consulta de todas as classes de metadados existentes – variáveis, agregados, classificações, dicionários e microdados -, na medida em que integra as bases de dados disponíveis no acervo digital institucional com seus respectivos metadados e dicionários de dados. Além disso, permite plena integração do banco de metadados com os dados agregados presentes no Sistema IBGE de Recuperação Automática - Sidra, plataforma que disponibiliza, no portal do IBGE na Internet, gratuitamente tabelas contendo resultados agregados de estudos e pesquisas realizados pelo IBGE. A reformulação do sistema buscou, também, atingir aos seguintes objetivos:

- Melhoria da navegabilidade do sistema de atualização e consulta de metadados;
- Incorporação de funcionalidades para desenvolvimento de um sistema para gestão da qualidade;
- Aderência a padrões internacionais (Data Documentation Initiative - DDI, Statistical Data and Metadata Exchange - SDMX e Generic Statistical Information Model - GSIM)
- Padronização de nomes e conceitos de variáveis;
- Maior integração entre as áreas produtoras e entre dados de pesquisas.

Inicialmente, o conteúdo do sistema de metadados era direcionado para atender à sua finalidade primordial, qual seja, fornecer aos usuários as informações necessárias para interpretar os dados estatísticos divulgados pelo IBGE. Assim, até 2015, apenas os metadados de referência compunham o sistema, abrangendo informações, principalmente, sobre os métodos utilizados nas operações estatísticas, descrevendo seus aspectos chave, como metodologia, periodicidade, períodos de referência, tipo de operação estatística, unidades estatísticas, cobertura geográfica, público-alvo, etc. da pesquisa. A partir de 2016, o novo Banco de Metadados - MetaBD do IBGE teve seu escopo ampliado, passando a incorporar três novos módulos para abrigar outras categorias de metadados, além dos metadados de referência já existentes: os metadados estruturais, para descrição de informações, conceitos e classificações associados a variáveis primárias e agregadas; os metadados de processos, relacionados à descrição detalhada das etapas e subetapas desenvolvidas ao longo do processo de produção de informações; e os metadados de qualidade, descrevendo informações para avaliar a qualidade dos produtos estatísticos, com relação a diversos aspectos de qualidade, como descrito no próximo tópico, **Indicadores de qualidade**. Até o momento, esses módulos estão disponíveis apenas internamente, com previsão de estarem disponíveis para o público em geral a partir de 2018.

Com a reformulação e ampliação do sistema que abriga os metadados estatísticos, espera-se melhorar a qualidade da documentação estatística fornecida aos usuários dos dados, por meio de um arcabouço estruturado para organizar e divulgar informações sobre o conteúdo e a estrutura da informação estatística produzida pelo IBGE, elaborado com base em padrões e modelos internacionais e espelhado nas melhores práticas de modernização da produção estatística.

Indicadores de qualidade

A avaliação da qualidade dos dados é um dos aspectos centrais do trabalho de um instituto de estatística, destacado pelo Código de Boas Práticas das Estatísticas do IBGE. A Seção 3 do Código estabelece que as estatísticas oficiais devem satisfazer às necessidades dos usuários e cumprir as normas de qualidades definidas para os produtos estatísticos. Os princípios 13 a 17 do Código estabelecem a necessidade de se proceder a uma avaliação permanente dos vários componentes da qualidade do produto, como a relevância, precisão e acurácia, oportunidade e pontualidade, coerência e comparabilidade, e acessibilidade e transparência. Recomenda-se que tal avaliação seja realizada com base em indicadores de qualidade, que são elementos específicos e mensuráveis voltados para mensurar e avaliar a qualidade dos produtos ou processos estatísticos, a partir de vários aspectos.

Uma primeira iniciativa voltada para avaliação da qualidade dos produtos do IBGE foi a elaboração do *Conjunto mínimo de indicadores padrão de qualidade a ser aplicado no Mercosul*, mencionado anteriormente, a qual não foi implementada de forma efetiva. Como já mencionado, um dos objetivos da reformulação do sistema de metadados era a integração de um sistema para a gestão da qualidade, para controle e monitoramento da qualidade dos processos e dos produtos estatísticos.

Assim, em 2016, o Banco de Metadados - MetaBD do IBGE sofreu uma adequação para incorporar os metadados associados às operações estatísticas, o que gerou uma oportunidade para validar e atualizar o conjunto de indicadores de qualidade a serem adotados, buscando alinhamento aos padrões internacionalmente recomendados. Um pouco antes, em 2014, a Eurostat, autoridade estatística da União Europeia, publicou um manual contendo um conjunto padronizado de metadados de qualidade, contemplando dimensões e aspectos relevantes para avaliar a qualidade dos produtos estatísticos, chamado Single Integrated Metadata Structure - SIMS (TECHNICAL..., 2014). Uma das vantagens desta estrutura de metadados é o fato de estar alinhada ao padrão do Statistical Data and Metadata eXchange - SDMX, que estabelece mecanismos padronizados para o compartilhamento de dados e metadados, e ao qual o IBGE, no âmbito dos sistemas informáticos de metadados, já vem trabalhando para estar plenamente aderente a esse padrão, atendendo às recomendações internacionais.

Assim, em 2017, definiu-se uma proposta de atributos de metadados e indicadores de qualidade, apresentada no Quadro 2, tomando por base os manuais *Conjunto mínimo de indicadores padrão de qualidade a ser aplicado no* MERCOSUL, do MERCOSUL, e *Technical manual of the single integrated metadata structure (SIMS)*, da EUROSTAT (CONJUNTO..., 2014; TECHNICAL..., 2014), que deverão ser preenchidos pelos responsáveis das operações estatísticas visando posterior disponibilização de relatórios de qualidade padronizados na Internet.

Esta iniciativa se configura, portanto, em um passo importante para a efetiva implementação de um conjunto de indicadores que serão usados para continuamente monitorar e avaliar os diversos componentes da qualidade dos produtos estatísticos, como estabelecido pelo Código de Boas Práticas das Estatísticas do IBGE.

Quadro 2 - Estrutura atual dos metadados de qualidade do produto

Dimensão	Indicador/Atributo de qualidade
Relevância	Necessidade dos usuários
	R1-Índice de satisfação do usuário
Acurácia, precisão e exatidão	Erro amostral
	A1-Coeficiente de variação
	Erro não amostral
	A2a-Taxa de resposta de unidade
	A3a-Taxa de resposta por pergunta
	A4a-Taxa de imputação das principais variáveis
	Erro de cobertura
	A5-Taxa de excesso de cobertura
	Erro de processamento
	Revisão de dados
	A6. Revisão de dados – Tamanho médio
Oportunidade e pontualidade	Oportunidade
	OP2-Intervalo de tempo entre o final do período de referência e a data da publicação dos resultados
	Pontualidade
	OP1-Pontualidade no calendário de uma publicação
Acessibilidade e clareza	Publicações
	Tabelas
	Microdados
	Documentação metodológica
	Documentação sobre qualidade
	AT1b-Acessos via Web
	AT2-Taxa de completude da informação de metadados para as estatísticas
Comparabilidade	Comparabilidade ao longo do tempo
	C1-Longitude das séries temporais comparáveis
Coerência	Coerência entre estatísticas com periodicidades diferentes
Confidencialidade	Confidencialidade – tratamento dos dados
Gestão da qualidade	Avaliação da qualidade

Fonte: IBGE. Banco de metadados - Metabd. Rio de Janeiro, [2017]. Disponível em: <https://metadados.ibge.gov.br/consulta/default.aspx>. Acesso em: set. 2017.

Projeto de modernização, aplicações do GSBPM e estações de qualidade

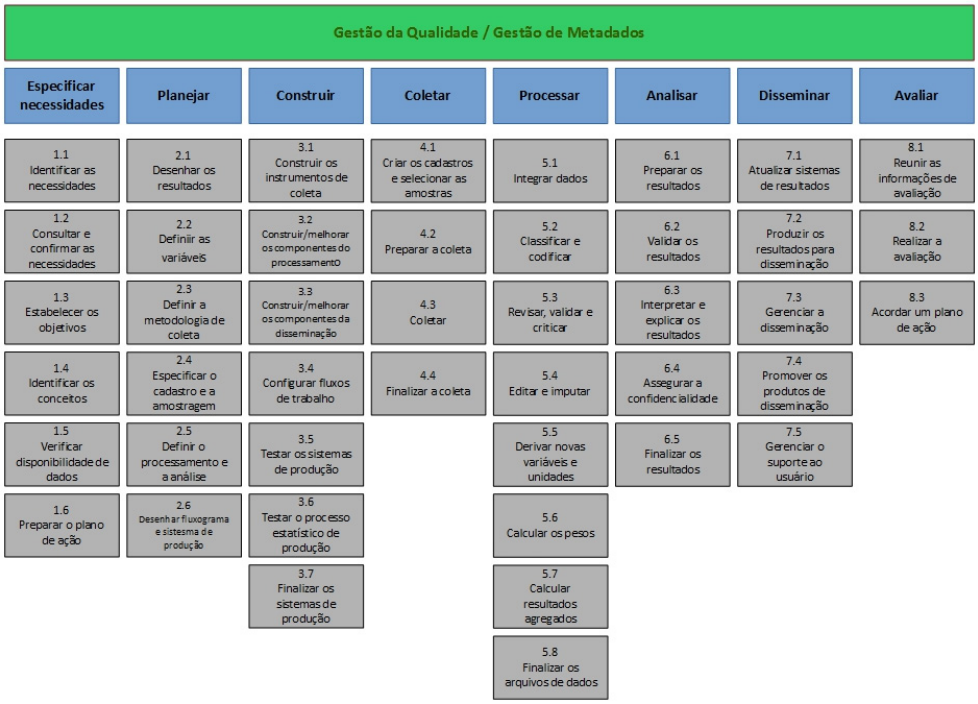
Nos últimos anos, o IBGE vem se dedicando ao estudo de iniciativas e práticas internacionais voltadas para a modernização da produção estatística, visando ao alcance de eficiência, transparência e à melhoria da qualidade de seus produtos e serviços, em face dos grandes desafios que o cenário atual apresenta.

Neste sentido, a Diretoria de Pesquisas, por meio de sua Gerência de Qualidade Estatística, com apoio da Gerência de Desenvolvimento Organizacional, da Diretoria Executiva, vem conduzindo o projeto de Modernização da Produção Estatística, sob o amparo de recomendações e de arcabouços internacionais, baseados, fundamentalmente, no uso de padrões e modelos genéricos.

O Modelo Genérico do Processo de Produção Estatística - (Generic Statistical Business Process Model – GSBPM, em tradução livre), tem sido utilizado como quadro de referência para nortear o projeto de modernização da produção estatística. O GSBPM descreve e define o conjunto de processos de ações necessárias para produzir estatísticas oficiais, por meio de uma estrutura padrão e terminologia harmonizada para ajudar as organizações de estatística a modernizar seus processos de produção estatística, bem como para compartilhar métodos e componentes. O GSBPM, apresentado na Figura 1, é composto por três níveis:

- Nível 0: o processo de produção estatística (uma pesquisa, por exemplo)
- Nível 1: os processos, que identificam as grandes fases (etapas)
- Nível 2: os subprocessos, que são um conjunto de tarefas.

Figura 1. Modelo Genérico do Processo de Produção Estatística – GSBPM (Generic Statistical Business Process Model)



Adaptado de Generic statistical business process model: GSBPM: version 5.0 (2013, p. 10).

O GSBPM tem sido usado nos institutos de estatística para a integração de padrões de dados e metadados, como um modelo para a documentação de processos, para a harmonização de métodos estatísticos e ferramentas de informática, e para fornecer um quadro de referência para a avaliação e melhoria da qualidade dos processos, dentre outras aplicações.

O IBGE decidiu adotar o GSBPM 5.0 (GENERIC..., 2013) como a pedra fundamental para a modernização do Instituto, alinhando-se ao trabalho internacional coordenado pela Comissão Econômica das Nações Unidas para a Europa (United Nations Economic Commission for Europe - UNECE), para promover modernização baseada em padrões, no âmbito do Projeto Modernstats[2]. Três iniciativas diferentes estão sendo desenvolvidas para melhorar a qualidade dos dados dos produtos do IBGE e a eficiência de seus programas estatísticos, envolvendo o uso do GSBPM.

A primeira aplicação do GSBPM é voltada diretamente para a melhoria da qualidade dos dados, atuando como base para a documentação e na revisão das práticas adotadas nas diversas operações estatísticas, visando identificar quais processos têm maior risco de ocorrência de erros. No IBGE, todas as pesquisas têm seus próprios controles de qualidade de dados e práticas de garantia de qualidade. No entanto, tais ações não são padronizadas nem documentadas e organizadas de forma estruturada.

Como acontece em qualquer organização grande e complexa, podem ocorrer problemas durante o processo de produção, resultando na ocorrência de erros. A maioria dos erros é detectada internamente, antes da publicação das estatísticas. No entanto, recentemente, houve alguns episódios de erros detectados apenas após sua divulgação, com diferentes graus de impacto no domínio público, resultando em necessidade de correções e suscitando questionamentos relativos à credibilidade do órgão. Para reduzir a chance de ocorrência de situações semelhantes no futuro, o IBGE decidiu implementar práticas de gestão de melhor qualidade através do desenvolvimento e uso da estratégia de mitigação de riscos, chamada de "Estações de Qualidade" (tradução livre do termo em inglês *Quality Gates*), inspirada no trabalho desenvolvido pelo Australian Bureau of Statistics - ABS. "Estações de Qualidade" consistem em um conjunto de critérios de aceitação da qualidade, colocados em pontos estratégicos no processo de produção, para facilitar a detecção, discussão e solução de problemas, com a finalidade de melhorar a qualidade final da produção estatística. Para detecção de erros no processo de produção estatística, o GSBPM pode ser usado como um guia para mapear as atividades dos processos ao longo da produção estatística, a fim de identificar onde pontos de verificação devem ser colocados ao longo de um processo para identificar erros ou problemas antecipadamente (PINK, 2010).

A segunda aplicação da GSBPM visa à eficiência dos processos. O IBGE enfrenta o desafio de manter a qualidade de seus produtos de informação ao mesmo tempo em que utiliza cada vez menos recursos. Buscando otimizar e padronizar os processos envolvidos na produção estatística, recentemente foi realizado um estudo piloto para integrar os processos de trabalho das pesquisas anuais por empresas, usando o GSBPM como referência.

Assim, no 2º semestre de 2016, o estudo piloto, aplicado à realidade das pesquisas anuais por empresas (Pesquisa Industrial Anual - PIA, Pesquisa Anual do Comércio - PAC, Pesquisa Anual da Indústria da Construção - PAIC, Pesquisa Anual de Serviços - PAS), teve por objetivo estudar, adequar e mapear os processos de produção estatística utilizando o GSBPM como modelo de referência para o mapeamento, com o desenvolvimento das seguintes ações:

- Mapeamento e descrição detalhada de processos das pesquisas e operações estatísticas, usando o GSBPM como modelo de referência;
- Elaboração dos desenhos de processos utilizando o software Bizagi;
- Avaliação da situação atual visando ao aprimoramento para esses processos, com identificação de mecanismos de melhoria contínua;
- Levantamento e avaliação de riscos existentes nos processos;
- Definição de indicadores para medir o desempenho dos processos.

Como resultado desse estudo piloto, foi elaborada uma proposição de metodologia de controle de processos e recomendações para implementação de uma estratégia de qualidade para a produção estatística das pesquisas em questão.

---

[2] Para mais detalhes, consultar: <http://www1.unece.org/stat/platform/display/hlgbas/HLG-MOS+Presentations>.

Finalmente, a terceira aplicação do GSBPM é na documentação das operações estatísticas, gerando, em consequência, os metadados de processos. Como já mencionado, a reformulação do atual sistema de metadados objetiva ampliar seu escopo, incorporando informações que abrangem todo o processo estatístico, o que requer o mapeamento e a descrição de todos os processos e subprocessos. Pretende-se, com isto, transformar o sistema de metadados em um abrangente repositório de documentação, não apenas para registrar os metadados de referência, estruturais e de qualidade, mas também para armazenar informações, documentos e materiais produzidos ao longo dos processos de produção de todas as operações estatísticas, fundamentais para controlar e garantir sua qualidade.

Conclusão

O produto do IBGE é a informação. A confiança na qualidade da informação produzida é chave fundamental para sua sobrevivência. Se houver dúvida sobre as informações fornecidas, a credibilidade do Instituto pode ser questionada e sua reputação como órgão independente e fonte de informações confiáveis pode ficar abalada. Neste sentido, uma gestão da qualidade sistemática e efetiva é essencial para assegurar a qualidade dos processos e produtos da organização, razão pela qual o órgão tem investido continuamente no desenvolvimento de ações concretas a fim de melhorar os aspectos relacionados à relevância, precisão, acurácia, oportunidade, pontualidade, comparabilidade, coerência, acessibilidade e transparência de suas estatísticas. Nos últimos dez anos, especialmente, com a criação de uma unidade especialmente voltada para a gestão da qualidade estatística, o IBGE redobrou seus esforços, voltando sua atenção para diferentes aspectos da qualidade da produção estatística, levando em consideração as melhores práticas e os princípios das estatísticas oficiais. Além disto, o Instituto tem se amparado fortemente na adoção de padrões e recomendações internacionais, buscando se manter alinhado ao estado da arte em termos de gestão da qualidade estatística. Embora muitas ações já tenham sido implantadas ao longo dos últimos anos, esforços significativos ainda estão em desenvolvimento visando modernizar os processos de produção estatística, para que o IBGE possa dar um salto de qualidade, colocando-o em um nível de excelência, lado a lado com os institutos de estatística de primeiro mundo.

Referências

COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE. Código regional de buenas prácticas en estadísticas para América Latina y el Caribe. Santiago de Chile: CEPAL, 2011. Aprovado na Sexta Reunión de la Conferencia Estadística de las Américas - CEA-CEPAL, realizada em Bávaro, República Dominicana, em novembro de 2011. 21 p. Disponível em: <http://repositorio.cepal.org/bitstream/handle/11362/16422/FILE_148023_es.pdf?sequence=1&isAllowed=y>. Acesso em: set. 2017.

CONJUNTO mínimo de indicadores padrão de qualidade a ser aplicado no Mercosul. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2014. 19 p. (Textos para discussão, n. 52). Disponível em: < https://biblioteca.ibge.gov.br/visualizacao/livros/liv87506.pdf >. Acesso em: set. 2017.

ESS handbook for quality reports. Luxembourg: Statistical Office of the European Union - Eurostat, 2014. 162 p. (Eurostat manuals and guidelines). Disponível em: <http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf>. Acesso em: set. 2017.

EUROSTAT. *Código de conduta das estatísticas europeias*: pelos institutos de estatística nacionais e comunitários. Luxembourg, 2011. Adotado pelo Sistema Estatístico Europeu em 28 de setembro de 2011. 8 p. Disponível em: < http://ec.europa.eu/eurostat/documents/3859598/5922361/10425-PT-PT.PDF>. Acesso em: set. 2017.

FEDERAL STATISTICAL OFFICE OF GERMANY. *How to deal with publication errors*: guidelines. Wiesbaden: Statistisches Bundesamt, 2013. 8 p. Disponível em: <https://www.destatis.de/EN/Methods/Quality/Publication_Errors.pdf;jsessionid=6612A0AE2C8DD572860839B535357E86.cae2?__blob=publicationFile>. Acesso em: set. 2017.

GENERIC statistical business process model: GSBPM: version 5.0. Paris: United Nations Economic Commission for Europe - Unece, 2013. 30 p. Disponível em: <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0?preview=/97356247/97519760/GSBPM%205_0.docx>. Acesso em: set. 2017.

IBGE. *Código de boas práticas das estatísticas do IBGE*. Rio de Janeiro, 2013. 48 p. Disponível em: <ftp://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>. Acesso em: set. 2017.

_____. *Política de revisão de dados divulgados das operações estatísticas do IBGE*. Rio de Janeiro, 2015. 8 p. Disponível em <http://www.ibge.gov.br/home/disseminacao/eventos/missao/politica_revisao_dados.pdf>. Acesso em: set. 2017.

_____. *Princípios fundamentais das estatísticas oficiais*: orientações para divulgações de resultados pelo IBGE. Rio de Janeiro, [2014]. 5 p. Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/principios_fundamentais_orientacoes_divulgacoes.shtm>. Acesso em: set. 2017.

_____. *Procedimentos para lidar com erros de divulgação de dados e informações estatísticas do IBGE*. Rio de Janeiro, 2015. 22 p. Disponível em <http://www.ibge.gov.br/home/disseminacao/eventos/missao/Procedimentos_ebook.pdf>. Acesso em: set. 2017.

_____. *Procedimentos para lidar com o mau uso dos dados e informações estatísticas e geoespaciais do IBGE*. Rio de Janeiro, 2016. 16 p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv98006.pdf>. Acesso em: set. 2017.

_____. *Resolução do Conselho Diretor n. 13, de 2 de junho de 2015*. Define as competências das unidades da Diretoria de Pesquisas. Rio de Janeiro, 2015.

NAÇÕES UNIDAS. Comissão de Estatística. *Princípios fundamentais das estatísticas oficiais*. Rio de Janeiro: IBGE, [2017]. 3 p. Adotados na sessão espacial da Comissão de Estatística das Nações Unidas, Nova Iorque, em 11-15 de abril de 1994. Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/principios_fundamentais_estatisticas.shtm> . Acesso em: set. 2017.

PINK, B. *Quality management of statistical processes using quality gates*. Canberra: Australian Bureau of Statistics - ABS, 2010. 24 p. (Information paper). Disponível em: <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1540.0Dec%202010?OpenDocument>. Acesso em: set. 2017.

TECHNICAL manual of the single integrated metadata structure (SIMS). Luxembourg: Statistical Office of European Union - Eurostat, 2014. 49 p. Disponível em: <http://ec.europa.eu/eurostat/documents/64157/4373903/03-Single-Integrated-Metadata-Structure-and-its-Technical-Manual.pdf/6013a162-e8e2-4a8a-8219-83e3318cbb39>. Acesso em: set. 2017.

UNITED NATIONS. Economic and Social Council. Fundamental principles of official statistics. New York, 2014. 2 p. Adotado pela Resolução A/RES/68/261 da Assembleia Geral das Nações Unidas, Nova Iorque, em 24 de janeiro de 2014. Disponível em: <http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf >. Acesso em: set. 2017.

UNITED NATIONS. *Statistical Division. Implementation of the fundamental principles of official statistics*. New York, 2013. 37 p. (Background document). Relatório apresentado na 44th session, agenda item 3(b), 26 Fev. a 1 March 2013. Disponível em: <https://unstats.un.org/unsd/statcom/doc13/BG-FP.pdf>. Acesso em: set. 2017.

Modelos de séries temporais para pesquisas amostrais repetidas

(edição fac-similar)

Eduardo Santiago Rosseti
Denise Britz do Nascimento Silva

Introdução

Várias das séries temporais utilizadas por diversos setores da sociedade, seja como fonte de informação para o conhecimento da realidade, seja como base de dados para pesquisa científica, são provenientes de pesquisas repetidas no tempo. Tais pesquisas, que em geral investigam as mesmas unidades em diversas ocasiões, são desenvolvidas visando à produção de informações que permitam o estudo da evolução de uma população ou de um fenômeno ao longo do tempo.

Na análise de séries temporais, é de extremo interesse a decomposição das séries observadas em componentes de tendência e sazonalidade. No caso de pesquisas amostrais repetidas, deve-se considerar o efeito que o desenho amostral da pesquisa exerce sobre a série observada, influenciando diretamente a autocorrelação da série e, consequentemente, o modelo de decomposição a ser utilizado.

Duncan e Kalton (1987), bem como Kalton e Citro (1993), introduziram uma tipologia para as pesquisas repetidas segundo a forma de inclusão de unidades na amostra. As pesquisas são classificadas de acordo com o nível de sobreposição de suas amostras ao longo do tempo. No caso de pesquisas com alguma sobreposição destacam-se: as pesquisas de painel fixo, nas quais as mesmas unidades amostrais são investigadas nas diversas ocasiões (amostras totalmente coincidentes ao longo do tempo), e as pesquisas de painéis rotativos, nas quais parte da amostra é mantida fixa entre duas ou mais ocasiões.

Em pesquisas de painéis rotativos, uma parte da amostra é mantida durante um período determinado, enquanto o restante é substituído por um novo conjunto de unidades amostrais. Com isso, existe uma sobreposição ou coincidência entre as unidades amostras em ocasiões sucessivas da pesquisa (sobreposição parcial das amostras ao longo do tempo). Os conjuntos mutuamente exclusivos de unidades amostrais que entram e saem da amostra de forma coordenada, de acordo com um esquema de rotação, são denominados painéis rotativos. Os diferentes esquemas de

rotação induzem complexas estruturas de correlação nas séries temporais de pesquisas repetidas. Adicionalmente, erros amostrais estão associados aos resultados "cuja estrutura de correlação e variabilidade precisam ser incorporadas na modelagem" (SILVA; CRUZ, 2002, p. 6).

Se o objetivo dos pesquisadores ou usuários das estatísticas públicas for obter informações sobre a evolução do verdadeiro valor populacional, é necessário decompor a série observada em dois processos distintos. O primeiro que representa o valor populacional (ou sinal) e o segundo associado ao erro amostral imposto pelo mecanismo gerador dos resultados. Smith (1999) ressalta que o analista deve ter cuidado uma vez que, de acordo com o padrão de sobreposição e rotação das amostras entre as diversas ocasiões da pesquisa, a estrutura de correlação dos erros amostrais se confunde com a estrutura de correlação da série temporal de observações, o que pode gerar distorção nas análises.

Sendo assim, neste capítulo são apresentados modelos para séries temporais de pesquisas amostrais repetidas e sua aplicação na análise da Pesquisa Mensal de Emprego - PME do IBGE que, durante vários anos, foi a principal fonte de informação sobre a dinâmica temporal do mercado de trabalho do país. Os modelos propostos, definidos de acordo com a representação de Espaço de Estados (HARVEY, 1989), permitem incorporar as características do desenho da pesquisa na análise da série histórica.

Modelos de espaço de estados

Os modelos de Espaço de Estados têm suas raízes na Engenharia e na Física nas quais um sistema dinâmico pode ser representado por um modelo matemático que descreve um processo real, caracterizado por valores de entrada, valores de saída e uma função de transferência.

Uma forma de descrever um sistema dinâmico é utilizando a representação em espaço de estados. A forma geral do Modelo de Espaço de Estados (MEE) contempla séries temporais multivariadas $\{\boldsymbol{y}_t\}$ com $t = 1, \dots, T$ e M elementos (ou séries).

Considere uma série temporal multivariada $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T$. Seja o vetor de observações $\boldsymbol{y}_t = (y_{1t}, y_{2t}, \dots, y_{Mt})'$ a cada instante $t = 1, \dots, T$ e denote por $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_T$ a sequência de vetores dos componentes não-observáveis de $\boldsymbol{y}_t$, denominados vetores de estados, tais que $\boldsymbol{\alpha}_t = (\alpha_{1t}, \alpha_{2t}, \dots, \alpha_{Mt})'$. O foco da modelagem é a realização de inferências sobre o vetor de estados $\boldsymbol{\alpha}_t$ com base na informação contida nos valores observados $\boldsymbol{y}_t$.

Um MEE é definido por duas equações. A primeira equação, denominada equação de observação representa a relação entre as observações e os estados atuais das componentes não-observáveis. A segunda, dita equação de transição (ou do sistema), descreve a forma como as componentes não-observáveis evoluem estocasticamente ao longo do tempo.

A equação de observação é dada por:

$$\boldsymbol{y}_t = \boldsymbol{H}_t \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \tag{1}$$

onde $\boldsymbol{y}_t$ é um vetor com dimensão $(M x\, 1)$, $\boldsymbol{H}_t$ é uma matriz $(M x\, n)$, $\boldsymbol{\alpha}_t$ é um vetor $(n\, x\, 1)$ e $\boldsymbol{\varepsilon}_t$ é um vetor $(M\, x\, 1)$ cujos elementos são não correlacionados no tempo, com média zero e matriz de covariância $\boldsymbol{U}_t$, tal que $E(\boldsymbol{\varepsilon}_t) = 0$, $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' = \boldsymbol{U}_t)$ e $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_r') = 0$, $\forall t \neq r$.

A equação de transição expressa a relação entre os vetores de estado $\boldsymbol{\alpha}_t$ e $\boldsymbol{\alpha}_{t-1}$, não-observáveis, por um processo markoviano de primeira ordem:

$$\boldsymbol{\alpha}_t = \boldsymbol{T}_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{G}_t \boldsymbol{\eta}_t \tag{2}$$

onde $\boldsymbol{T}_t$ e $\boldsymbol{G}_t$ são matrizes com dimensões (n x n) e (n x g), respectivamente, $\boldsymbol{\eta}_t$ é um vetor (M x 1) cujos elementos são não correlacionados no tempo, com média zero e matriz de covariância $\boldsymbol{Q}_t$, tal que: $E(\boldsymbol{\eta}_t) = 0$, $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = \boldsymbol{Q}_t$ *e* $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_r') = 0$, $\forall t \neq r$.

Para finalizar a especificação de um MEE, são necessárias duas hipóteses adicionais:

- o vetor inicial de estados $\boldsymbol{\alpha}_0$ é normalmente distribuído com média e matriz de covariância dadas por: $E(\boldsymbol{\alpha}_0) = \hat{\boldsymbol{\alpha}}_{0|0}$ e $COV(\boldsymbol{\alpha}_0) = \boldsymbol{P}_{0|0}$; e
- as perturbações aleatórias $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são não correlacionadas no tempo e também não correlacionadas com o vetor inicial de estados $\boldsymbol{\alpha}_0$, ou seja, $E(\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_t') = 0$, $E(\boldsymbol{\eta}_t\boldsymbol{\alpha}_0') = 0$ e $E(\boldsymbol{\varepsilon}_t\boldsymbol{\alpha}_0') = 0$, $\forall t = 1,2,\dots,T$.

As matrizes que determinam as equações do sistema $\boldsymbol{H}_t$, $\boldsymbol{U}_t$, $\boldsymbol{T}_t$, $\boldsymbol{G}_t$ e $\boldsymbol{Q}_t$, geralmente dependem de um conjunto de parâmetros desconhecidos que necessitam ser estimados. As etapas do processo de previsão de $\boldsymbol{y}_t$, e de estimação das componentes não-observáveis $\boldsymbol{\alpha}_t$, são realizadas de forma recursiva com base no Filtro de Kalman (HARVEY, 1989).

Seja $\boldsymbol{D}_t$ um vetor aleatório que contém toda a informação disponível até o tempo t, tal que $\boldsymbol{D}_t = (\boldsymbol{y}_1', \boldsymbol{y}_2', \dots, \boldsymbol{y}_t')$. A distribuição condicional $P(\boldsymbol{\alpha}_t|\boldsymbol{D}_{t-1})$ contém toda a informação que $\boldsymbol{D}_{t-1} = (\boldsymbol{y}_1', \dots, \boldsymbol{y}_{t-1}')$ fornece sobre $\boldsymbol{\alpha}_t$. Nas etapas de previsão e extração de sinal é preciso obter, respectivamente, $E(\boldsymbol{\alpha}_t|\boldsymbol{D}_{t-1})$ e $E(\boldsymbol{\alpha}_t|\boldsymbol{D}_t)$ para $\forall t$. Define-se $\hat{\boldsymbol{\alpha}}_{t|t-1} = (\boldsymbol{\alpha}_t|\boldsymbol{D}_{t-1})$ e, analogamente, $\hat{\boldsymbol{\alpha}}_{t|t} = E(\boldsymbol{\alpha}_t|\boldsymbol{D}_{t-1}, \boldsymbol{y}_t) = E(\boldsymbol{\alpha}_t|\boldsymbol{D}_t)$, com matrizes de covariância $\boldsymbol{P}_{t|t-1} = V(\boldsymbol{\alpha}_t|\boldsymbol{D}_{t-1})$ e $\mathbf{P}_{t|t} = \mathrm{V}(\boldsymbol{\alpha}_t|\mathbf{D}_t)$. As etapas do processo de estimação pelo Filtro de Kalman são definidas a seguir.

- **Previsão**: corresponde a antever o comportamento futuro de $\boldsymbol{\alpha}_t$ e $\boldsymbol{y}_t$ para algum momento $t^* > t$. Considere, no tempo $t-1$, que $\boldsymbol{D}_{t-1}$ é observado mas os estados $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_{t-1}$ são desconhecidos. Supondo que o estimador $\hat{\boldsymbol{\alpha}}_{t-1|t-1}$ de $\boldsymbol{\alpha}_{t-1}$ está disponível, as equações de previsão do filtro de Kalman fornecem um estimador ótimo $\hat{\boldsymbol{\alpha}}_{t|t-1} = E(\boldsymbol{\alpha}_{t-1}|\boldsymbol{D}_{t-1})$ para $\boldsymbol{\alpha}_t$, e uma previsão para $\boldsymbol{y}_t$ dada por $\hat{\boldsymbol{y}}_{t|t-1} = E(\boldsymbol{y}_t|\boldsymbol{D}_{t-1})$, antes de observar $\boldsymbol{y}_t$, $\forall t$.
- **Extração de Sinal***:* corresponde a recuperar, no tempo t, informação sobre quantidades não-observáveis do sistema ($\boldsymbol{\alpha}_t$) utilizando apenas as informações disponíveis até o tempo atual $\boldsymbol{D}_t = (\boldsymbol{y}_1', \boldsymbol{y}_2', \dots, \boldsymbol{y}_t')$**.** Dado $\boldsymbol{y}_t$, o estimador $\hat{\boldsymbol{\alpha}}_{t|t}$ de $\boldsymbol{\alpha}_t$ é $E(\boldsymbol{\alpha}_t|\boldsymbol{y}_t)$, da distribuição condicional obtida de resultados conhecidos da distribuição normal multivariada (RAO, 1973) aplicada em $P(\boldsymbol{\alpha}_t, \boldsymbol{y}_t|\boldsymbol{D}_{t-1})$.
- **Suavização:** corresponde a recuperar uma informação sobre quantidades não-observáveis ($\boldsymbol{\alpha}_t$) utilizando toda informação disponível anterior e posterior ao tempo t. Sendo assim, a recuperação ocorre após o tempo t com base em $\boldsymbol{D}_T = (\boldsymbol{y}_1', \boldsymbol{y}_2', \dots, \boldsymbol{y}_T')$. Este procedimento inicia no tempo T, calculando-se $\hat{\boldsymbol{\alpha}}_{T|T}$ para então produzir as estimativas referentes aos períodos $T, T-1, \dots, 1$.

Segundo Durbin e Koopman (2001), a principal vantagem de um MEE é sua abordagem geral que permite representar uma grande quantidade de modelos, inclusive os modelos ARIMA (BOX; JENKINS; REINSEL, 1994), modelos estruturais com nível, tendência e sazonalidade, coeficientes de regressão e de intervenção. Outro destaque é sua flexibilidade para incorporar mudanças na estrutura da série ao longo do tempo. Mais detalhes sobre a representação de uma série temporal em MEE podem ser encontrados em Harvey (1989), Durbin e Koopman (2001), Commandeur e Koopman (2007) e Wei (1990).

O filtro de Kalman é uma ferramenta recursiva capaz de fornecer informações sobre o vetor de estados $\boldsymbol{\alpha}_t$ dado $\boldsymbol{D}_t = (\boldsymbol{y}_1', \boldsymbol{y}_2', \dots, \boldsymbol{y}_t')$ pela função de probabilidade $P(\boldsymbol{\alpha}_t|\boldsymbol{D}_t)$. O melhor estimador para $\boldsymbol{\alpha}_t$, que minimiza o Erro Quadrático Médio

(EQM), é obtido por $E(\boldsymbol{\alpha}_t|\boldsymbol{D}_t)$. As equações do filtro de Kalman fornecem, portanto, o estimador linear de mínimo erro quadrático para $\boldsymbol{\alpha}_t$ dado $\boldsymbol{D}_t$. Caso as pertubações aleatórias, e o vetor de estado inicial $\boldsymbol{\alpha}_0$, sejam normalmente distribuídos, as equações do filtro de Kalman fornecem o estimador de mínimo erro quadrático médio para $\boldsymbol{\alpha}_t$ dado $\boldsymbol{D}_t$, como apresentado em Harvey(1989) e Silva e Cruz (2002).

Neste capítulo, considera-se que as perturbações aleatórias são normalmente distribuídas, tal que o MEE tem a seguinte formulação:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \quad , \quad \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{U}_t), \\ \boldsymbol{\alpha}_t &= \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{G}_t\boldsymbol{\eta}_t \quad , \quad \boldsymbol{\eta}_t \sim N(0, \boldsymbol{Q}_t). \end{aligned} \tag{3}$$

Para inicialização do filtro assume-se que $\boldsymbol{\alpha}_0 \sim N(\boldsymbol{\alpha}_0, \boldsymbol{P}_0)$, com $\boldsymbol{\alpha}_0 = 0$ e $\boldsymbol{P}_0 = 10^5\boldsymbol{I}$.

Ao assumir a normalidade das perturbações aleatórias, a função de verossimilhança do modelo é obtida através da decomposição de erros de previsão (HARVEY, 1989). As matrizes do sistema de equações do modelo em (3) dependem de parâmetros, denotados por $\boldsymbol{\Omega}$, que precisam ser estimados. No caso de análise de séries temporais, as observações $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T$ não são independentemente distribuídas e deve-se utilizar a função de probabilidade condicional $P(\boldsymbol{y}_t|\boldsymbol{D}_{t-1})$, como definido em Harvey (1989), tal que:

$$P(\boldsymbol{D}_t|\boldsymbol{\Omega}) = P(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T|\boldsymbol{\Omega}) = \prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{D}_{t-1}) \tag{4}$$

Considerando a suposição de normalidade do modelo, a função de *log*-verossimilhança é dada por:

$$\ell(\boldsymbol{D}_t, \boldsymbol{\Omega}) = -\frac{MT}{2} ln\, 2\pi - \frac{1}{2}\sum_{t=1}^{T}\left\{ln\left|\boldsymbol{F}_{t|t-1}\right| + \left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}\right)' \boldsymbol{F}_{t|t-1}^{-1}\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}\right)\right\}$$

$$\text{com} \qquad \hat{\boldsymbol{y}}_{t|t-1} = E(\boldsymbol{y}_t|\boldsymbol{D}_{t-1}) \qquad \text{e} \qquad \boldsymbol{F}_{t|t-1} = V(\boldsymbol{y}_t|\boldsymbol{D}_{t-1}) \quad . \tag{5}$$

Modelos para o processo $\{\boldsymbol{y}_t\}$

Denote por $\boldsymbol{\theta}_t$ o parâmetro de interesse num dado tempo t e considere uma pesquisa amostral realizada em intervalos equidistantes $t = 1, \dots, T$. Seja $\boldsymbol{e}_t$ o erro amostral da estimativa $\boldsymbol{y}_t$ obtida para $\boldsymbol{\theta}_t$ a partir da pesquisa. Uma pesquisa repetida produz uma série temporal de estimativas $\{\boldsymbol{y}_t\}$ da série de parâmetros $\{\boldsymbol{\theta}_t\}$. Nesse caso, é necessário considerar a combinação de dois modelos para decompor a série $\{\boldsymbol{y}_t\}$ em processos distintos: um para descrever a evolução das quantidades não-observáveis do sinal $\{\boldsymbol{\theta}_t\}$ e outro para representar a autocorrelação dos erros amostrais $\{\boldsymbol{e}_t\}$ das estimativas. Com isso, a série de observações pode ser decomposta em sinal e ruído tal que:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \tag{6}$$

Os trabalhos de Blight e Scott (1973) e Scott e Smith (1974) foram pioneiros na utilização da abordagem de séries para estimação do valor desconhecido de $\boldsymbol{\theta}_t$ em pesquisas repetidas empregando a teoria de extração de sinal na presença de ruído. Os modelos de espaços de estados com componentes de tendência e sazonalidade para o sinal $\boldsymbol{\theta}_t$ (HARVEY, 1989) foram introduzidos por Binder e Hidiroglou (1988), Binder e Dick (1989), Pfeffermann (1991) e Tiller (1992) na estimação e análise de pesquisas repetidas, dando origem a uma série de novos trabalhos (PFEFFERMANN; BELL; SIGNORELLI, 1996; SILVA, 1996; PFEFFERMANN; FEDER; SIGNORELLI, 1998 e SILVA; SMITH, 2001). A ampla utilização dessa metodologia em diversos

países, bem como novos desenvolvimentos, podem ser encontrados, por exemplo, em: Silva e Cruz (2002), Pfeffermann e Tiller (2006), Brakel e Krieg (2009), Krieg e Brakel (2012), Hernandez (2012), Rosseti (2013) e Bollineni-Balabay, Brakel e Palm (2016, 2017).

Silva e Cruz (2002) destacam que, sob a ótica da estimação em pesquisas repetidas, o MEE permite aprimorar a estimativa corrente da pesquisa com base em toda informação passada disponível e, sob a ótica da análise de séries temporais, representa a evolução do sinal $\{\boldsymbol{\theta}_t\}$ considerando a autocorrelação dos erros amostrais induzida pelas características do desenho da pesquisa referentes à rotação e sobreposição da amostra. Por isso, a definição de um modelo para o processo $\{\boldsymbol{y}_t\}$, e sua formulação em espaço de estados, deve incorporar elementos que representem os processos $\{\boldsymbol{\theta}_t\}$ e $\{\boldsymbol{e}_t\}$. O uso da abordagem de espaço de estados fornece flexibilidade na especificação dos modelos de séries temporais para ambos os processos. No caso do sinal, a escolha do modelo baseia-se na análise do fenômeno de interesse e no objetivo da modelagem. Por outro lado, para o processo dos erros amostrais, estimativas da função de autocorrelação são utilizadas na identificação de modelos VARMA (REINSEL, 1993; SILVA, 1996; SILVA; SMITH, 2001).

Modelo para o sinal $\{\boldsymbol{\theta}_t\}$

No caso de séries temporais provenientes de pesquisas amostrais, o alvo de inferência é o processo $\{\boldsymbol{\theta}_t\}$ mas observa-se apenas o processo $\{\boldsymbol{y}_t\}$. Pode-se supor que o comportamento da série publicada é dominado pelo sinal, já que as estimativas são obtidas por amostras probabilísticas desenhadas de modo a manter sob controle o erro amostral. Sendo assim, a definição do modelo para o sinal baseia-se na experiência de analisar a série de estimativas e cabe ao analista decidir entre a aplicação de modelos VARMA ou estruturais. Em geral, opta-se por modelos estruturais que permitem a decomposição da série em componentes de interesse, tais como a tendência e a sazonalidade. O modelo usualmente proposto para o sinal $\{\boldsymbol{\theta}_t\}$ é um modelo estrutural básico multivariado definido por:

$$\begin{aligned}
\boldsymbol{\theta}_t &= \boldsymbol{L}_t + \boldsymbol{S}_t + \boldsymbol{\varepsilon}_t \quad , \quad \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Sigma}_\varepsilon) \\
\boldsymbol{L}_t &= \boldsymbol{L}_{t-1} + \boldsymbol{R}_{t-1} + \boldsymbol{\eta}_L \quad , \quad \boldsymbol{\eta}_L \sim N(0, \boldsymbol{\Sigma}_L) \\
\boldsymbol{R}_t &= \boldsymbol{R}_{t-1} + \boldsymbol{\eta}_R \quad , \quad \boldsymbol{\eta}_R \sim N(0, \boldsymbol{\Sigma}_R) \\
\boldsymbol{S}_t &= -\sum_{j=1}^{s-1} \boldsymbol{S}_{t-j} + \boldsymbol{\eta}_S \quad , \quad \boldsymbol{\eta}_S \sim N(0, \boldsymbol{\Sigma}_S)\,.
\end{aligned} \tag{7}$$

Sua representação em modelo espaço de estados é dada por:

$$\begin{aligned}
\boldsymbol{\theta}_t &= \boldsymbol{H}^{(\boldsymbol{\theta})}\, \boldsymbol{\alpha}_t^{(\boldsymbol{\theta})} + \boldsymbol{\varepsilon}_t \quad , \quad \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Sigma}_\varepsilon), \\
\boldsymbol{\alpha}_t^{(\boldsymbol{\theta})} &= \boldsymbol{T}^{(\boldsymbol{\theta})}\, \boldsymbol{\alpha}_{t-1}^{(\boldsymbol{\theta})} + \boldsymbol{G}^{(\boldsymbol{\theta})}\, \boldsymbol{\eta}_t^{(\boldsymbol{\theta})} \quad , \quad \boldsymbol{\eta}_t^{(\boldsymbol{\theta})} \sim N(0, \boldsymbol{Q}_t).
\end{aligned} \tag{8}$$

$$\boldsymbol{\alpha}_t^{(\boldsymbol{\theta})} = [L_{1t} \quad \ldots \quad L_{Mt} \quad R_{1t} \quad \ldots \quad R_{Mt} \quad S_{1t} \quad \ldots \quad S_{Mt} \quad \ldots \quad S_{M,t-s-2}]'$$

$$\boldsymbol{H}^{(\boldsymbol{\theta})} = [1 \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0] \otimes \boldsymbol{I}_{(M)}$$

$$\boldsymbol{T}^{(\boldsymbol{\theta})} = \begin{bmatrix} 1 & 1 & & \\ 0 & 1 & & \boldsymbol{0}_{(2\times s-1)} \\ & & -1 & -1 \quad \ldots \quad -1 \quad -1 \\ & & 1 & 0 \quad \ldots \quad 0 \quad 0 \\ \boldsymbol{0}_{(2\times s-1)} & & 0 & 1 \quad \ldots \quad 0 \quad 0 \\ & & \vdots & \vdots \quad \ddots \quad \vdots \quad \vdots \\ & & 0 & 0 \quad \ldots \quad 1 \quad 0 \end{bmatrix} \otimes \boldsymbol{I}_{(M)}$$

$$\boldsymbol{G}^{(\boldsymbol{\theta})} = \begin{bmatrix} \boldsymbol{I}_{(3)} \\ \boldsymbol{0}_{(s-2\times 3)} \end{bmatrix} \otimes \boldsymbol{I}_{(M)}$$

$$\boldsymbol{\eta}_t^{(\boldsymbol{\theta})} = [\boldsymbol{\eta}_L \quad \boldsymbol{\eta}_R \quad \boldsymbol{\eta}_S]' \text{ com } \boldsymbol{Q}_t = \begin{bmatrix} \boldsymbol{\Sigma}_L & & \boldsymbol{0} \\ & \boldsymbol{\Sigma}_R & \\ \boldsymbol{0} & & \boldsymbol{\Sigma}_S \end{bmatrix}$$

Modelo para o erro amostral $\{\boldsymbol{e_t}\}$

A especificação do modelo de séries temporais para os erros amostrais é realizada de acordo com o desenho amostral da pesquisa, com o esquema de rotação da amostra e com o nível de sobreposição das amostras ao longo do tempo. Empregam-se, em geral, modelos VARMA (ou ARMA no caso de séries univariadas) para descrever o processo do erro amostral, sendo necessário então estimar a função de autocorrelação dos erros amostrais para identificação do modelo adequado. Destaca-se que o processo dos erros amostrais $\{\boldsymbol{e}_t\}$ é não-observável. Silva e Cruz (2002) apresentam uma revisão das diferentes abordagens disponíveis para estimação da função de autocorrelação dos erros amostrais e a utilizada neste trabalho está descrita no tópico Modelo para o sinal $\{\boldsymbol{\theta}_t\}$. Para completar a formulação do modelo (7), assume-se, por exemplo, um modelo VAR(p) para o processo multivariado dos erros amostrais:

$$\boldsymbol{e}_t = \boldsymbol{\Phi}_1\boldsymbol{e}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{e}_{t-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{e}_{t-p} + \boldsymbol{\eta}_e \quad , \quad \boldsymbol{\eta}_e \sim N(0, \boldsymbol{\Sigma}_e)\,, \tag{9}$$

cuja representação em modelo espaço de estados é dada por:

$$\begin{aligned} \boldsymbol{e}_t &= \boldsymbol{H}^{(e)}\,\boldsymbol{\alpha}_t^{(e)} \quad , \\ \boldsymbol{\alpha}_t^{(e)} &= \boldsymbol{T}^{(e)}\,\boldsymbol{\alpha}_{t-1}^{(e)} + \boldsymbol{G}^{(e)}\,\boldsymbol{\eta}_e \quad , \quad \boldsymbol{\eta}_e \sim N(0, \boldsymbol{\Sigma}_e). \end{aligned} \tag{10}$$

$$\boldsymbol{T}^{(e)} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \\ \vdots & \boldsymbol{I}_{(p-1)\times(p-1)} \\ \vdots & \\ \boldsymbol{\Phi}_p & \boldsymbol{0}_{1\times(p-1)} \end{bmatrix} \quad , \quad \boldsymbol{H}^{(e)} = [1,0,\ldots,0]\,, \; \boldsymbol{G}^{(e)} = 1 \quad \text{e}$$

$$\boldsymbol{\alpha}_t^{(e)} = (\boldsymbol{\alpha}_{1t}, \ldots, \boldsymbol{\alpha}_{pt})' \text{ com } \boldsymbol{\alpha}_{1t} = \boldsymbol{e}_t\,.$$

Modelo de extração de sinal para $\{\boldsymbol{y_t}\}$

Definidos os modelos para o sinal $\{\boldsymbol{\theta}_t\}$ e para o erro amostral $\{\boldsymbol{e}_t\}$, o modelo para a série de observações $\{\boldsymbol{y}_t\}$ é dado por:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\theta}_t + \boldsymbol{e}_t \\ \boldsymbol{\theta}_t &= \boldsymbol{L}_t + \boldsymbol{S}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{L}_t &= \boldsymbol{L}_{t-1} + \boldsymbol{R}_{t-1} + \boldsymbol{\eta}_L \\ \boldsymbol{R}_t &= \boldsymbol{R}_{t-1} + \boldsymbol{\eta}_R \\ \boldsymbol{S}_t &= -\sum_{j=1}^{s} \boldsymbol{S}_{t-j} + \boldsymbol{\eta}_S \\ \boldsymbol{e}_t &= \boldsymbol{\Phi}_1\boldsymbol{e}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{e}_{t-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{e}_{t-p} + \boldsymbol{\eta}_e \end{aligned} \tag{11}$$

com a seguinte representação em modelo espaço de estados:

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \boldsymbol{\alpha}_t^{(\theta)} \\ \dots \\ \boldsymbol{\alpha}_t^{(e)} \end{bmatrix}, \qquad \boldsymbol{H}_t = \begin{bmatrix} \boldsymbol{H}_t^{(\theta)} & \vdots & \boldsymbol{H}_t^{(e)} \end{bmatrix}, \qquad \boldsymbol{T}_t = \begin{bmatrix} \boldsymbol{T}_t^{(\theta)} & \vdots & \mathbf{0} \\ \dots & \dots & \dots \\ \mathbf{0} & \vdots & \boldsymbol{T}_t^{(e)} \end{bmatrix},$$

$$\boldsymbol{G}_t = \begin{bmatrix} \boldsymbol{G}_t^{(\theta)} \\ \dots \\ \boldsymbol{G}_t^{(e)} \end{bmatrix}, \qquad \boldsymbol{\eta}_t = [\boldsymbol{\eta}_L \quad \boldsymbol{\eta}_R \quad \boldsymbol{\eta}_S \quad \vdots \quad \boldsymbol{\eta}_e]', \qquad \boldsymbol{V}(\boldsymbol{\eta}_t) = \begin{bmatrix} \boldsymbol{\Sigma}_L & & & \mathbf{0} \\ & \boldsymbol{\Sigma}_R & & \\ & & \boldsymbol{\Sigma}_S & \\ \mathbf{0} & & & \boldsymbol{\Sigma}_e \end{bmatrix}.$$

É importante observar que o vetor de estados $\boldsymbol{\alpha}_t$ inclui componentes de dois processos: o que representa a evolução do sinal $\{\boldsymbol{\theta}_t\}$ e o correspondente ao erro amostral $\{\boldsymbol{e}_t\}$. Permite assim a obtenção de estimativas das componentes não-observáveis do sinal $\boldsymbol{\alpha}_t$ livres das flutuações induzidas pelo desenho amostral da pesquisa repetida. Silva e Cruz (2002) descrevem quatro passos para a estimação do modelo para a série de observações $\{\boldsymbol{y}_t\}$:

- estimação dos parâmetros e da ordem do modelo VAR(p) para o processo do erro amostral realizada externamente ao filtro de Kalman (detalhes no tópico Modelo para o sinal $\{\boldsymbol{\theta}_t\}$);
- estimação dos hiperparâmetros $(\boldsymbol{\Sigma}_L, \boldsymbol{\Sigma}_R, \boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_e)$ pela maximização da verossimilhança apresentada na equação 5;
- extração das componentes não observáveis do sinal e suas respectivas estimativas suavizadas; e
- avaliação da qualidade do ajuste do modelo.

Cabe destacar que essa metodologia pode ser aplicada ao tratamento de diversas questões relacionadas à estimação em pesquisas repetidas, à análise de séries multivariadas e séries de composição, e à estimação em pequenas áreas ou domínios.

Estimação da função de autocorrelação dos erros amostrais

Por simplicidade, a partir desta seção considera-se o caso de uma série temporal univariada $\{y_t\}$ obtida de uma pesquisa amostral com painéis rotativos. Seja y_t uma estimativa não viciada do parâmetro de interesse θ_t tal que:

$$y_t = \theta_t + e_t \quad \text{com } E(e_t) = 0 \ \forall t \quad . \tag{12}$$

Ressalta-se que y_t denota tanto um estimador para θ_t, quanto sua estimativa após a efetiva realização da pesquisa. Considere também a situação usual em pesquisas repetidas com K painéis rotativos na qual cada painel $(k = 1, \dots, K)$ fornece uma estimativa $y_t^{(k)}$ não viciada do parâmetro θ_t, e as estimativas estão sujeitas a erros amostrais $e_t^{(k)}$ que são correlacionados segundo uma estrutura de correlação imposta pelo mecanismo de seleção/rotação da amostra, tal que:

$$y_t^{(k)} = \theta_t + e_t^{(k)} \quad \text{com } E\left(e_t^{(k)}\right) = 0 \ \forall t, k \quad . \tag{13}$$

A cada ocasião da pesquisa, a estimativa final y_t do parâmetro θ_t é uma combinação das estimativas $y_t^{(k)}$ (denominadas estimativas elementares).

O método de obtenção da função de autocorrelação depende da disponibilidade de informações e de microdados da pesquisa repetida. Se o analista tem acesso apenas às estimativas publicadas y_t, a única maneira de especificar modelos para os erros amostrais é elaborando hipóteses sobre os efeitos do desenho amostral na função de autocorrelação. Por outro lado, se os microdados da pesquisa, ou as estimativas elementares $y_t^{(k)}$, estão disponíveis é possível calcular estimativas

diretas da função de autocorrelação dos erros amostrais. Silva (1996) apresenta uma discussão sobre os diferentes métodos de estimação da função de correlação dos erros amostrais em pesquisas repetidas. No caso da série temporal univariada em (12) e (13), o modelo proposto para o processo do erro amostral é um modelo ARMA(p,q). Apresenta-se, a seguir, a metodologia desenvolvida por Pfeffermann, Bell e Signorelli (1996) e Pfeffermann, Feder e Signorelli (1998) para estimar as Funções de Autocorrelação -FAC e Funções de Autocorrelação Parcial - FACP do processo $\{e_t\}$ com base nos pseudo-erros (*pseudo panel errors*) definidos pelos autores como:

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t \ \text{ com } \ y_t = \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} \tag{14}$$

Assim, o pseudo-erro associado ao k-ésimo painel pode ser interpretado como o desvio entre a estimativa elementar e o valor observado y_t da série. Considerando $y_t^{(k)}$ um estimador não viciado do verdadeiro valor populacional θ_t como em (13), tem-se que:

$$\begin{aligned} \tilde{e}_t^{(k)} &= y_t^{(k)} - y_t = y_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} \\ &= \left(y_t^{(k)} - \theta_t - \left(\frac{1}{K}\sum_{k=1}^{K} y_t^{(k)} - \theta_t \right) \right) \\ &= e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)} = e_t^{(k)} - e_t \end{aligned} \tag{15}$$

O método proposto para a estimação da função de autocorrelação dos erros amostrais baseia-se em duas hipóteses:

$$\begin{aligned} &(a)\ COV\left(e_{t-h}^{(j)}, e_t^{(k)}\right) = 0 \ \text{ se } \ j \neq k\,, \ \ \forall t, h\ ; \\ &(b)\ COV\left(e_{t-h}^{(k)}, e_t^{(k)}\right) = \gamma_h^{(k)}\,, \ \forall t, h \text{ e } k = 1,2,\dots,K. \end{aligned} \tag{16}$$

A suposição (a) implica que os erros amostrais associados às estimativas elementares de painéis não coincidentes (quando não há sobreposição de amostras) são não correlacionados. A suposição (b) indica estacionaridade para o processo $\left\{e_t^{(k)}\right\}$, com $k = 1,2,\dots,K$, pois assume-se que a autocovariância dos erros amostrais de estimativas elementares não varia ao longo do tempo, depende apenas do painel e da defasagem no tempo (lag). A partir das equações (15) e (16), tem-se:

$$\begin{aligned} \gamma_h &= COV(e_{t-h}, e_t) = COV\left(\frac{1}{K}\sum_{j=1}^{K} e_{t-h}^{(j)} \ , \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)} \right) \\ &= \frac{1}{K^2}\, COV\left(\sum_{k=1}^{K} e_{t-h}^{(k)}, e_t^{(k)} \right) = \frac{1}{K^2}\gamma_h^{(k)} \end{aligned} \tag{17}$$

sendo $\gamma_h^{(k)}$ a autocovariância de lag h do processo do erro amostral $e_t^{(k)}$ referente ao k-ésimo painel e γ_h a autocovariância de lag h do processo do erro amostral e_t. Além disso, é possível obter a autocovariância $C_h^{(k)}$ dos pseudo-erros $\tilde{e}_t^{(k)}$ associados ao k-ésimo painel com base nas equações (14) e (15):

$$
\begin{aligned}
C_h^{(k)} &= COV\left(\tilde{e}_{t-h}^{(k)}\ ,\ \tilde{e}_t^{(k)}\right) = COV\left(e_{t-h}^{(k)} - e_{t-h}\ ,\ e_t^{(k)} - e_t\right) \\
&= COV\left(e_{t-h}^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_{t-h}^{(k)}\ ,\ e_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} e_t^{(k)}\right) \\
&= COV\left(e_{t-h}^{(k)}, e_t^{(k)}\right) - \frac{1}{K}\sum_{j=1}^{K} COV\left(e_{t-h}^{(k)}, e_t^{(j)}\right) + \\
&- \frac{1}{K}\sum_{j=1}^{K} COV\left(e_{t-h}^{(j)}\ ,\ e_t^{(k)}\right) + \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} COV\left(e_{t-h}^{(i)}, e_t^{(j)}\right)
\end{aligned}
\tag{18}
$$

Aplicando-se as hipóteses (a) e (b) na equação (18) obtém-se:

$$
\begin{aligned}
C_h^{(k)} &= \gamma_h^{(k)} - \frac{1}{K}\gamma_h^{(k)} - \frac{1}{K}\gamma_h^{(k)} + \frac{1}{K^2}\sum_{j=1}^{K}\gamma_h^{(j)} \\
&= \left[1 - \frac{2}{K}\right]\gamma_h^{(k)} + \frac{1}{K^2}\sum_{j=1}^{K}\gamma_h^{(j)} \\
&= \left[1 - \frac{1}{K}\right]^2\gamma_h^{(k)} + \frac{1}{K^2}\sum_{j\neq k}^{K}\gamma_h^{(j)}
\end{aligned}
\tag{19}
$$

Utilizando as equações (17) e (19), pode-se demonstrar que:

$$
\begin{aligned}
\sum_{k=1}^{K} C_h^{(k)} &= \sum_{k=1}^{K}\left[1 - \frac{1}{K}\right]^2\gamma_h^{(k)} + \sum_{k=1}^{K}\frac{1}{K^2}\sum_{j\neq 1}^{K}\gamma_h^{(j)} \\
&= \left[1 - \frac{1}{K}\right]^2\sum_{k=1}^{K}\gamma_h^{(k)} + \frac{K-1}{K^2}\sum_{k=1}^{K}\gamma_h^{(k)} \\
&= \frac{K^2 - K}{K^2}\sum_{k=1}^{K}\gamma_h^{(k)} = (K^2 - K)\gamma_h \\
&= (K^2 - K)\ COV(e_{t-h}, e_t)
\end{aligned}
\tag{20}
$$

Finalmente, a estimativa da função de autocorrelação dos erros amostrais ρ_h é dada por:

$$
\begin{aligned}
\rho_h &= \frac{\sum_{k=1}^{K} C_h^{(k)}}{\sum_{k=1}^{K} C_0^{(k)}} = \frac{(K^2 - K)COV(e_{t-h}, e_t)}{\sqrt{(K^2 - K)COV(e_t, e_t)(K^2 - K)COV(e_{t-h}, e_{t-h})}} \\
&= \frac{COV(e_{t-h}, e_t)}{\sqrt{COV(e_t, e_t)COV(e_{t-h}, e_{t-h})}} = \frac{\gamma_h}{\gamma_0}
\end{aligned}
\tag{21}
$$

A função de autocorrelação dos erros amostrais é então estimada calculando-se a função de autocovariância amostral das séries de pseudo-erros e combinando-as de acordo com (21). Estimativas da função de autocorrelação parcial dos erros amostrais são obtidas, a partir das estimativas de autocorrelação da série, empregando-se as equações de Yule-Walker (WEI, 1990) que fornecem também estimativas dos parâmetros do modelo para o processo de erro amostral, conforme apresentado por Silva e Cruz (2002).

A Pesquisa Mensal de Emprego - PME

Esta seção contém um breve resumo sobre a Pesquisa Mensal de Emprego - PME do IBGE cuja análise da série temporal está apresentada a seguir. No período de 1980 até fevereiro de 2016, a pesquisa foi a principal fonte de informação para produção de indicadores mensais sobre a força de trabalho que permitiram acompanhar a evolução do mercado de trabalho em diferentes regiões metropolitanas do país (PESQUISA..., 2007).

A amostra da PME era, desde 2002, composta por oito painéis rotativos e possuía esquema de rotação do tipo 4-8-4, conforme especificado no Quadro 1. Os domicílios de um dado painel permaneciam na amostra por quatro meses consecutivos (para a primeira, segunda, terceira e quarta visitas), eram retirados da amostra nos oito meses subsequentes, e retornavam à amostra para a quinta, sexta, sétima e oitava visitas, sendo então retirados definitivamente da pesquisa.

Quadro 1 – Exemplo de painéis, grupos de rotação e números de visitas na amostra da PME

Mês	Painéis A								Painéis B							
	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	B4	B5	B6	B7	B8
Jan					4	3	2	1								
Fev						4	3	2	1							
Mar							4	3	2	1						
Abr								4	3	2	1					
Mai									4	3	2	1				
Jun	5									4	3	2	1			
Jul	6	5									4	3	2	1		
Ago	7	6	5									4	3	2	1	
Set	8	7	6	5									4	3	2	1
Out		8	7	6	5									4	3	2
Nov			8	7	6	5									4	3
Dez				8	7	6	5									4
Jan					8	7	6	5								
Fev						8	7	6	5							
Mar							8	7	6	5						
Abr								8	7	6	5					
Mai									8	7	6	5				
Jun										8	7	6	5			
Jul											8	7	6	5		
Ago												8	7	6	5	
Set													8	7	6	5
Out														8	7	6
Nov															8	7
Dez																8

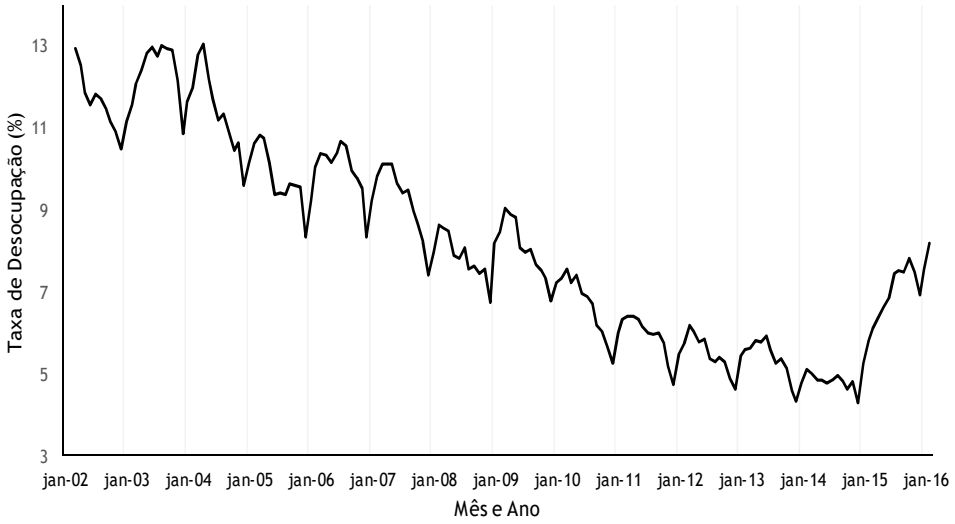
Fonte: Elaborado pelos Autores com base na metodologia da Pesquisa Mensal de Emprego.

As unidades primárias de amostragem da pesquisa (os setores censitários) foram alocadas em grupos mutuamente exclusivos, denominados grupos de rotação. O número de grupos de rotação deve ser sempre igual ao número total de ocasiões que um domicílio permanece na pesquisa. No caso da PME, os setores censitários selecionados para a amostra foram divididos em oito grupos de rotação. A cada mês, dois painéis eram retirados da amostra (os painéis que haviam completado a quarta e a oitava visitas) e substituídos por painéis de domicílios dos mesmos grupos de

rotação aos quais pertenciam (domicílios dos mesmos setores censitários). Os respectivos substitutos eram: um novo painel (para substituir os domicílios já visitados em oito ocasiões da pesquisa) e um painel composto de domicílios que retornavam à amostra depois de oito meses de ausência (em troca do painel cujos domicílios já haviam sido visitados em quatro ocasiões da pesquisa).

Neste esquema de rotação, havia uma sobreposição de 75% das unidades domiciliares na amostra em meses consecutivos e a sobreposição da amostra era de 50% em pares de anos consecutivos (PESQUISA..., 2007). O desenho amostral da PME atribuía, então, uma complexa estrutura de correlação aos erros amostrais, que deve ser incorporada na modelagem da série pois, caso contrário, se confunde com a evolução das componentes do sinal (SILVA; CRUZ, 2002). A Figura 1 apresenta a série de estimativas da taxa de desocupação obtida na PME no âmbito deste estudo, para o total das regiões metropolitanas: Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Porto Alegre.

Figura 1 - Taxa de Desocupação do Total das Regiões Metropolitanas



IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Mensal do Emprego 2002- 2016. Dados trabalhados pelos Autores.

Modelo de espaço de estados para a série da taxa de desocupação da PME

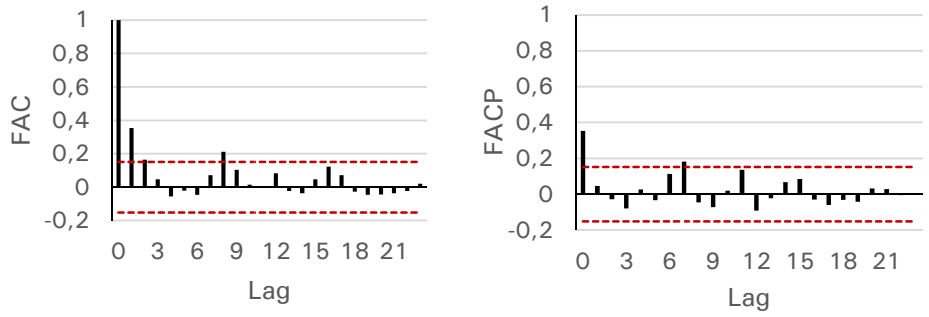
Sob a abordagem de extração de sinal na presença de ruído, a série observada da taxa de desocupação $\{y_t\}$ é representada pela soma de duas componentes: $y_t = \theta_t + e_t$, sendo θ_t o sinal, valor populacional da taxa de desocupação, e e_t o erro amostral. O processo do sinal $\{\theta_t\}$ será representado por um modelo estrutural básico com componentes de tendência e sazonalidade, e o processo do erro amostral $\{e_t\}$ por um modelo AR(p), numa versão univariada do modelo definido em (11).

$$\begin{aligned}
\theta_t &= L_t + S_t + \varepsilon_t && , \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
L_t &= L_{t-1} + R_{t-1} + \eta_L && , \quad \eta_L \sim N(0, \sigma_L^2) \\
R_t &= R_{t-1} + \eta_R && , \quad \eta_R \sim N(0, \sigma_R^2) \\
S_t &= -\sum_{j=1}^{11} S_{t-j} + \eta_S && , \quad \eta_S \sim N(0, \sigma_S^2) \\
e_t &= \phi_1 e_{t-1} + \cdots + \phi_p e_{t-p} + \eta_e && , \quad \eta_e \sim N(0, \sigma_e^2)
\end{aligned} \tag{22}$$

A estimação da ordem e dos coeficientes do modelo AR(p) do erro amostral $\{e_t\}$ é feita com base na análise dos pseudo-erros univariados como descrito na seção Estimação da Função de autocorrelação dos erros amostrais. As estimativas das taxas de desocupação para cada um dos oito grupos rotacionais da PME, estimativas elementares $y_t^{(k)}$, foram obtidas com microdados da PME disponíveis na página do IBGE. Foi necessário construir as oito séries mensais de estimativas elementares e calcular as autocovariâncias $C_h^{(k)}$ dos pseudo-erros $\tilde{e}_t^{(k)}$ associados aos painéis de cada um dos grupos rotacionais. A função de autocorrelação dos erros amostrais foi então estimada combinando-se as funções de autocovariância amostral das séries de pseudo-erros de acordo com (21).

A análise gráfica da FAC e da FACP, apresentadas nas Figuras 2 e 3, sugere a escolha de um modelo AR(1) para representar a evolução estocástica dos erros amostrais $\{e_t\}$.

Figura 2 – Estimativas da FAC e FACP dos erros amostrais $\{e_t\}$ e limites dos intervalos de confiança



Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

O modelo AR(1) selecionado para representar a evolução estocástica dos erros amostrais é dado por $e_t = \phi_t e_{t-1} + \eta_t^e$ com $\eta_t^e \sim N(0, \sigma_e^2)$. Os valores dos coeficientes estimados utilizando-se as equações de Yule-Walker (MORETTIN; TOLOI, 2006) são:

$$e_t = 0{,}3531 e_{t-1} + \eta_t^e \ , \text{ com } \hat{\sigma}_e^2 = 0{,}04245 \qquad (23)$$

A estimação do conjunto de hiperparâmetros $\boldsymbol{\Omega} = (\sigma_L^2, \sigma_R^2, \sigma_S^2, \sigma_\varepsilon^2)$ do modelo estrutural do sinal $\{\theta_t\}$ foi realizada por maximização da função de *log*-verossimilhança como definida em (5). A Tabela 1 apresenta as estimativas dos hiperparâmetros e as estimativas dos erros padrões aparecem entre parênteses.

Tabela 1 - Estimativas dos hiperparâmetros para os modelos ajustados à série da taxa de desocupação da PME

Modelo	$\hat{\sigma}_L^2$	$\hat{\sigma}_R^2$	$\hat{\sigma}_S^2$	$\hat{\sigma}_\varepsilon^2$
MEB	0,073092 (7,6 e-05)	0,000138 (0)	0,000205 (0)	0 (0)
MEB Reestimado	0,079419 (8,1 e-05)	-	-	-
MEB + AR(1)	0,014163 (6,5 e-05)	0,002095 (1 e-06)	2.9 e-05 (0)	0 (0)
MEB + AR(1) Reestimado	-	0,003736 (2 e-06)	-	-

Fonte: Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal do Emprego 2002-2016.

Os resultados indicam que apenas a componente de tendência L_t do Modelo Estrutural Básico - MEB, e o termo associado ao incremento da tendência R_t do modelo MEB + AR(1), podem ser considerados estocásticos (com variância significativamente diferente de zero). Dessa forma, os modelos foram reestimados excluindo-se os demais hiperparâmetros.

A Tabela 2 contém as estatísticas usuais para diagnóstico dos modelos com base nas inovações padronizadas $v_t = F_{t|t-1}^{-1}\left(y_t - \hat{y}_{t|t-1}\right)$ e nos seguintes testes: Jarque-Bera (HARVEY, 1989), para a hipótese de normalidade, teste H para homoscedasticidade (COMMANDEUR; KOOPMAN, 2007), e o teste de *Ljung-Box* para a hipótese de independência (DURBIN; KOOPMAN, 2001). Os resultados indicam que não há evidências para rejeitar as suposições de normalidade, homoscedasticidade e independência das inovações obtidas dos modelos MEB e MEB+AR(1).

Tabela 2 - Estatísticas e p-valores associados aos testes de diagnósticos dos modelos

Hipótese	Teste	Modelo MEB	Modelo MEB+AR(1)
		Estatística (P-valor)	Estatística (P-valor)
Independência	Ljung-Box	18,84 (0,76)	26,92 (0,31)
Homoscedasticidade	H(39)	0,81 (0,77)	0,65 (0,94)
Normalidade	Jarque-Bera	1,99 (0,37)	1,04 (0,59)

Fonte: Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal do Emprego 2002-2016.

Estimativas das componentes estruturais do sinal são obtidas como combinações lineares dos elementos que compõem o vetor de estados $\boldsymbol{\alpha}_t$. Definem-se vetores $\boldsymbol{W}$, tal que $w_t = \boldsymbol{W}\boldsymbol{\alpha}_t$, cujo estimador é $\hat{w}_t = \boldsymbol{W}\,\hat{\boldsymbol{\alpha}}_{t|T}$, com variância $V(\hat{w}_t) = \boldsymbol{W}\,V(\hat{\boldsymbol{\alpha}}_{t|T})\boldsymbol{W}'$. O Quadro 2 apresenta as componentes de interesse e os respectivos vetores $\boldsymbol{W}$.

Quadro 2 – Componentes estruturais e vetores W

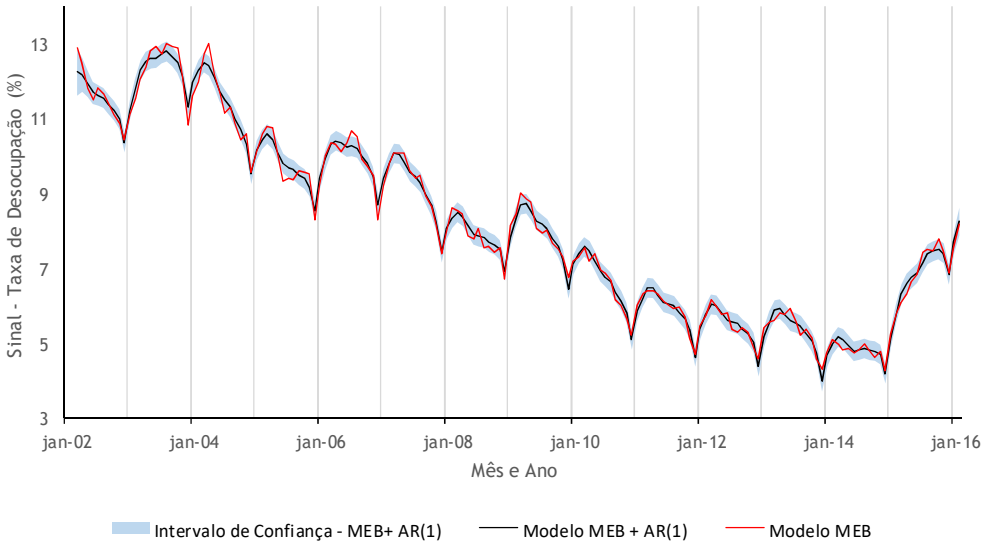
Componente Estrutural	Componente do Vetor de Estados	**W**
Sinal	$\theta_t = L_t + S_t$	[1 0 1 0 0 0 0 0 0 0 0 0 0 0]
Tendência	L_t	[1 0 0 0 0 0 0 0 0 0 0 0 0 0]
Incremento	R_t	[0 1 0 0 0 0 0 0 0 0 0 0 0 0]
Sazonalidade	S_t	[0 0 1 0 0 0 0 0 0 0 0 0 0 0]
Série Dessazonalizada	$L_t + e_t$	[1 0 0 0 0 0 0 0 0 0 0 0 0 1]

Fonte: Os Autores.

O principal objetivo desta abordagem é aprimorar a estimação das componentes estruturais não observáveis do sinal $\{\theta_t\}$ por uso das estimativas suavizadas $\hat{\boldsymbol{\alpha}}_{t|T}$ ou filtradas $\hat{\boldsymbol{\alpha}}_{t|t}$. Isto ocorre por dois motivos principais: a especificação de um modelo para o processo do erro amostral e utilização de toda informação disponível $(t = 1, \ldots, T)$ no caso das estimativas suavizadas. Como pode ser observado na Figura 3, as estimativas suavizadas para o sinal $\{\theta_t\}$ no modelo MEB+AR(1) apresentam menos irregularidades do que as estimativas suavizadas obtidas pela direta aplicação de um MEB à serie de estimativas amostrais $\{y_t\}$. Há então evidências de que as irregularidades estão associadas ao processo do erro amostral que causa ruído na estimativa do sinal caso não seja devidamente tratado. Destaca-se que o modelo MEB + AR(1) considera a estrutura de correlação do erro

amostral $\{e_t\}$ em sua especificação devido à inclusão do modelo autoregressivo AR(1) cujos coeficientes foram estimados com base nos pseudo-erros.

Figura 3 - Estimativas suavizadas do sinal da taxa de desocupação do Total das Regiões Metropolitanas

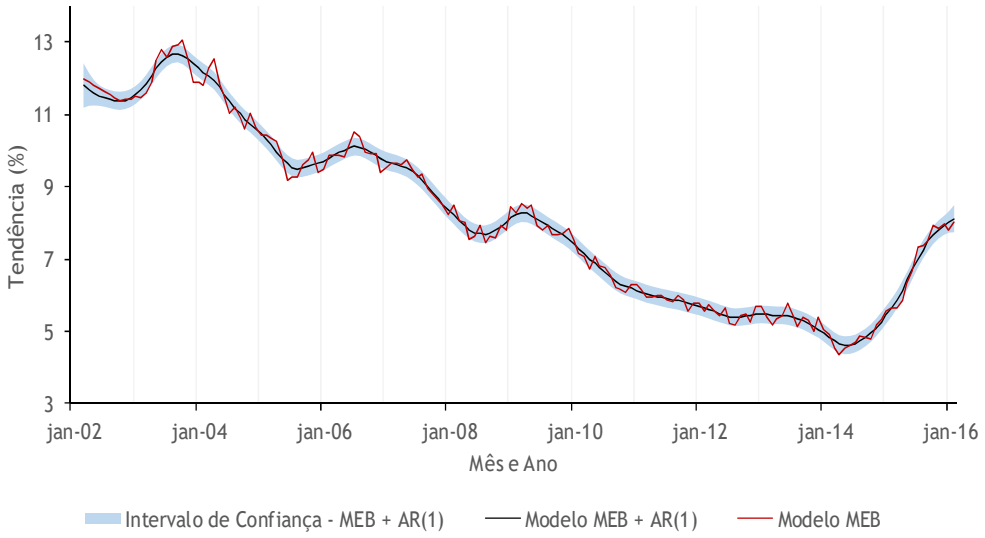


Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

É conhecido na literatura que a estrutura de autocorrelação dos erros amostrais e_t pode ser absorvida pelas componentes estruturais do sinal θ_t, principalmente pela tendência L_t, no caso de modelos estruturais que não levam em conta que a série temporal é proveniente de uma pesquisa amostral. A Figura 4 apresenta as estimativas da tendência suavizada obtidas dos modelos MEB e MEB + AR(1).

As estimativas de tendência obtidas pelo modelo MEB + AR(1) também são menos irregulares, indicando que as flutuações oriundas do desenho amostral da pesquisa são removidas ou têm menos influência nas estimativas calculadas. Observa-se que as estimativas da tendência do modelo MEB estão contidas no intervalo de confiança das estimativas do modelo BSM + AR(1). Isto pode indicar que, apesar de adicionar ruído, a estrutura de correlação dos erros amostrais tem efeito limitado no componente da tendência da taxa de desocupação. Silva e Cruz (2002) e Rosseti (2013) chegaram a conclusão semelhante ao aplicar a mesma metodologia para taxa de desocupação de PME. Cabe ressaltar que existem diversos trabalhos que apontam a existência de movimentos espúrios na tendência ao ignorar a estrutura de correlação do erro amostral e_t como, por exemplo, Tiller (1992) e Silva e Smith (2001).

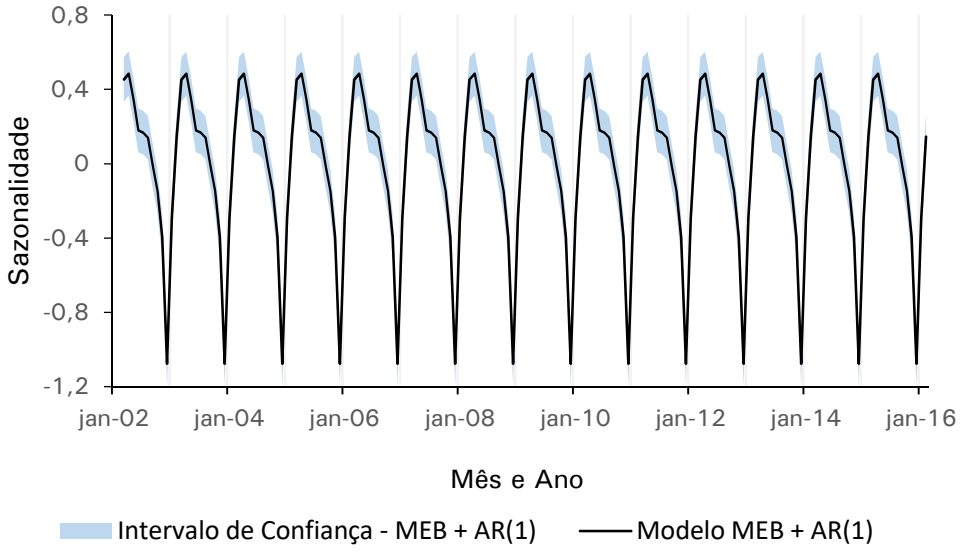
Figura 4 - Estimativas suavizadas de tendência da taxa de desocupação do Total das Regiões Metropolitanas



Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

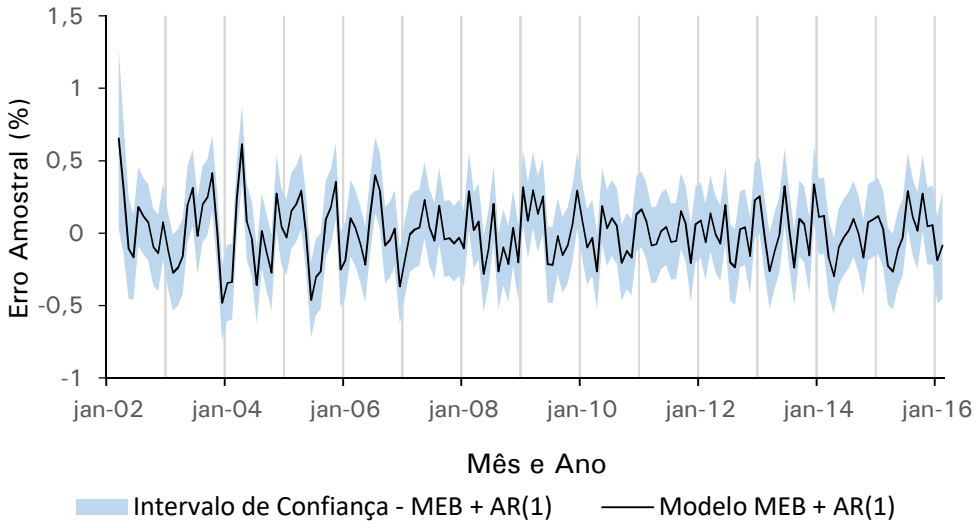
O uso de modelos estruturais possibilita o cálculo de estimativas para a série dessazonalizada, dada por $L_t + e_t$. As Figuras 5, 6 e 7 apresentam, respectivamente, as estimativas da componente sazonal S_t, do erro amostral e_t e da série dessazonalizada do modelo BSM + AR(1). É importante comentar que a estimativa da série dessazonalizada incorpora ruído (o erro amostral). Por isso, institutos nacionais de estatísticas de vários países usam a tendência como o mais importante indicador para a análise de séries econômicas.

Figura 5 - Estimativas suavizadas da componente sazonal da Taxa de Desocupação do Total das Regiões Metropolitanas



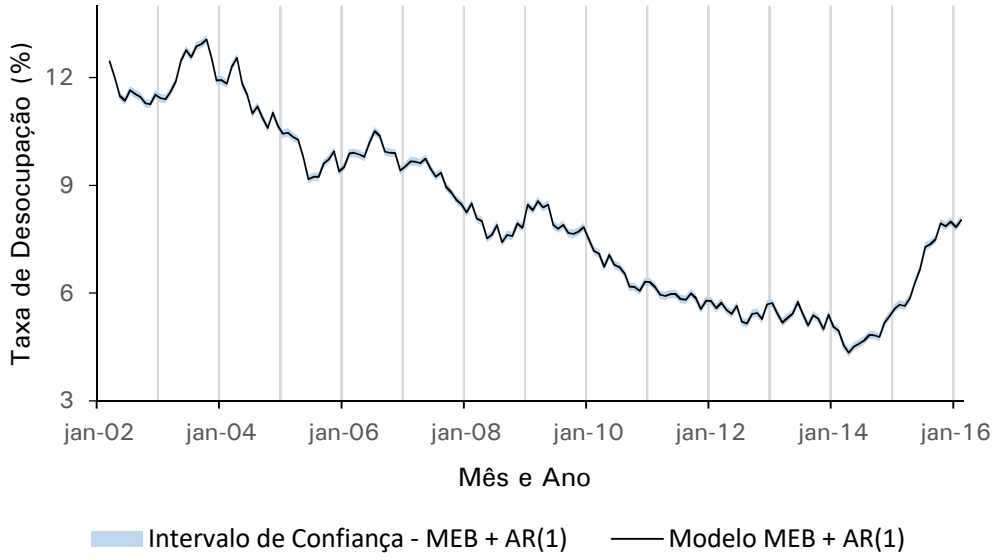
Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

Figura 6 - Estimativas suavizadas do erro amostral da Taxa de Desocupação do Total das Regiões Metropolitanas



Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

Figura 7 - Estimativas suavizadas da série dessazonalizada da Taxa de Desocupação do Total das Regiões Metropolitanas



Estimativas elaboradas pelos Autores com base nos dados da Pesquisa Mensal de Emprego 2002-2016.

Considerações finais

A metodologia apresentada neste capítulo já é atualmente utilizada em diversos países, principalmente por institutos nacionais de estatística responsáveis pela produção e análise de séries temporais. Tal procedimento, porém, não é automático. Para seu uso adequado, demanda o conhecimento da teoria estatística de análise de séries temporais, de modelos de espaço de estados, bem como da teoria de amostragem probabilística.

Por outro lado, é reconhecida a popularidade de métodos automáticos de ajustamento sazonal, como o X-11 e variantes (entre eles o X-13-ARIMA). Cabe destacar, entretanto, que estes métodos não permitem a decomposição da série observada entre sinal e ruído, fazendo que o analista possa confundir flutuações devidas ao do plano amostral da pesquisa com a verdadeira evolução do fenômeno de interesse. Os resultados aqui obtidos demonstram a importância de incorporar o efeito da correlação do erro amostral nos modelos de séries temporais de pesquisas amostrais. Isto deve ser feito pois existem evidências de que a estrutura de correlação imposta aos erros amostrais pelo desenho amostral com painéis rotativos da pesquisa tem influência na estimação das componentes estruturais associadas ao sinal da série.

Finalmente, destaca-se a flexibilidade da modelagem proposta e indica-se como trabalho futuro a elaboração de modelos de séries temporais para a Pesquisa Nacional por Amostra de Domicílios - PNAD Contínua do IBGE.

Referências

BINDER, D. A.; DICK, J. P. Modelling and estimation for repeated surveys. *Survey Methodology*, v. 15, n. 1, p. 29-45, June 1989. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/1989001/article/14579-eng.pdf>. Acesso em: set. 2017.

BINDER, D. A.; HIDIROGLOU, M. A. *Sampling in time*. In: KRISHNAIAH, P. R.; RAO, C. R. (Ed.). Sampling. New York: Elsevier Science, 1988. p. 187-211. (Handbook of statistics, v. 6).

BLIGHT, B. J. N.; SCOTT, A. J. A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*. Series B, Methodological. London, v. 35, n. 1, p. 61-66, 1973.

BOLLINENI-BALABAY, O.; BRAKEL, J. van den; PALM, F. Multivariate state space approach to variance reduction in series with level and variance breaks due to survey redesigns. *Journal of the Royal Statistical Society*. Series A, Statistics in Society. London, v. 179, n. 2, p. 377-402, Feb. 2016.

_____. State space time series modelling of the dutch labour force survey: model selection and mean squared errors estimation. *Survey Methodology*, Ottawa, v. 43, n. 1, p. 41-67, June 2017. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-eng.pdf>. Acesso em: set. 2017.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time series analysis: forecasting and control*. 3rd ed. Englewood Cliffs: Prentice-Hall, 1994. 598 p.

BRAKEL, J. van den; KRIEG, S. Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, v. 35, n. 2, p. 177-190, Dec. 2009. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11040-eng.pdf>. Acesso em: set. 2017.

COMMANDEUR, J. J. F.; KOOPMAN, S. J. *An introduction to state space time series analysis*. Oxford: Oxford University Press, 2007. 174 p.

DUNCAN, G. J.; KALTON, G. Issues of design and analysis of surveys across time. *International Statistical Review*, Voorburg, v. 55, n. 1, p. 97-117, Apr. 1987.

DURBIN, J.; KOOPMAN, S. J. *Time series analysis by state space methods*. Oxford: Oxford University Press, 2001. 253 p.

HARVEY, A. C. *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press, 1989. 554 p.

HERNANDEZ, A. L. *Estimação em pesquisas repetidas empregando o filtro GLS*. 2012. 147 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática, Estatística, e Computação, Universidade Estadual de Campinas - Unicamp, Campinas, 2012. Disponível em: <http://repositorio.unicamp.br/bitstream/REPOSIP/305866/1/LunaHernandez_Angela_M.pdf>. Acesso em: set. 2017.

KALTON, G.; CITRO, C. F. Panel surveys: adding the fourth dimension. *Survey Methodology*, Ottawa, v. 19, n. 2, p. 205-215, Dec. 1993. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/1993002/article/14452-eng.pdf>. Acesso em: set. 2017.

KRIEG, S.; BRAKEL, J. van den. Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis*, London, v. 56, n. 10, p. 2918-2933, Oct. 2012.

MORETTIN, P.; TOLOI, C. M. de C. *Análise de séries temporais*. 2.ed. São Paulo: Blucher, 2006. 564 p.

PESQUISA mensal de emprego. 2. ed. Rio de Janeiro: IBGE, 2007. 89 p. (Série relatórios metodológicos, v. 23). Disponível em: <ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Mensal_de_Emprego/Metodologia_da_Pesquisa/srmpme_2ed.pdf. Acesso em: set. 2017.

PFEFFERMANN, D. Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, Alexandria [Estados Unidos]: American Statistical Association - ASA, v. 9, n. 2, p. 163-175, 1991.

PFEFFERMANN, D.; BELL, P.; SIGNORELLI, D. Labour force trend estimation in small areas. In: ANNUAL RESEARCH CONFERENCE AND TECHNOLOGY INTERCHANGE, 1996, Virginia. *Proceedings*… Washington, DC: Bureau of the Census, 1996. p. 407-431. Disponível em: <https://babel.hathitrust.org/cgi/pt?id=umn.31951d01238781t;view=1up;seq=1>. Acesso em: set. 2017.

PFEFFERMANN, D.; FEDER, M.; SIGNORELLI, D. Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, Virginia, v. 16, n. 3, p. 339-348, July 1998.

PFEFFERMANN, D.; TILLER, R. Small-area estimation with state: space models subject to benchmark constraints. *Journal of the American Statistical Association*, Alexandria [Estados Unidos], v. 101, n. 476, p. 1387-1397, Dec. 2006.

REINSEL, G. C. *Elements of multivariate time series analysis*. New York: Springer-Verlag,1993. 358 p

ROSSETI, E. S. *Análise de séries temporais da pesquisa mensal de emprego com incorporação dos efeitos do plano amostral*. 2013. 92 p. Dissertação (Mestrado em Estudos Populacionais e Pesquisas Sociais) - Programa de Pós-Graduação em Estudos Populacionais e Pesquisas Sociais, Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2013. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2013/Resumo_Dissertacao_2013_Eduardo_Santiago_Rosseti.pdf>. Acesso em: set. 2017.

SCOTT, A. J.; SMITH, T. M. F. Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, Virginia, v. 69, n. 347, p. 674-678, Sep. 1974.

SILVA, D. B. do N. *Modelling compositional time series from repeated surveys*. 1996. 237 p. Tese (Doutorado em Estatística)-Faculty of Mathematical Studies, University of Southampton, Southampton, 1996.

SILVA, D. B. do N.; CRUZ, M. M. *Séries temporais de pesquisas amostrais periódicas. São Paulo:* Associação Brasileira de Estatística - ABE, 2002. 141 p.

SILVA, D. B. do N.; SMITH, T. M. F. Modelling compositional time series from repeated surveys. *Survey Methodology*, Ottawa, v. 27, n. 2, p. 205-215, Dec. 2001. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/2001002/article/6097-eng.pdf>. Acesso em: set. 2017.

SMITH, T. M. F. Defining parameters of interest in longitudinal studies and some implications for design. In: INTERNATIONAL STATISTICAL INSTITUTE SESSION, 52., 1999, Helsinki. *Bulletin of the International Statistical Institute*. The Hague [Holanda]: International Statistical Institute, 1999. v. 58, n. 2, p. 7-10. Disponível em: <https://www.stat.fi/isi99/proceedings/arkisto/varasto/smit0949.pdf>. Acesso em: set. 2017.

TILLER, R. B. Time series modelling of sample data from the U.S. current population survey. *Journal of Official Statistics*, Stockholm: Statistics Sweden, v.8, n. 2, p. 149-166, 1992.

WEI, W. W. S. *Time series analysis: univariate and multivariate methods*. Redwood City, Calif.: Addison-Wesley, 1990. 478 p.

Os métodos estatísticos no IBGE: trajetória desde 1997 e perspectivas

Pedro Luis do Nascimento Silva

Introdução

Vivemos num mundo em que a produção e disponibilidade de dados e informações alcançou níveis sem precedentes. Estes dados são demandados e utilizados para conhecer o mundo em que vivemos, para informar a tomada de decisões e para avaliar os impactos dessas decisões. Apesar disso, ainda há lacunas substanciais de informações e também questões de qualidade referentes a grande parte das informações disponíveis. Por estes motivos, nunca esteve em tanta evidência e demanda o conhecimento e a necessidade de desenvolvimento da Ciência Estatística.

Nesse contexto mais geral, a produção de informações e estatísticas públicas relevantes, atuais, confiáveis e de qualidade, de forma ágil e eficiente, requer métodos e tecnologias sofisticados, que demandam constante atualização, seja pela incorporação ou adaptação de conhecimentos disponíveis, seja pela pesquisa na fronteira de conhecimento para desenvolver novos métodos e tecnologias.

No Instituto Brasileiro de Geografia e Estatística - IBGE, os últimos vinte anos foram um período de muitas transformações no processo de produção e disseminação de informações e estatísticas públicas. Por um lado, houve grande aumento de produtividade sustentado, em grande parte, por maior automação dos processos, mas, ao mesmo tempo, por uma mudança qualitativa do perfil dos servidores e funcionários.

Migramos de um ambiente em que se usava predominantemente questionários em papel com captura das informações via digitação ou escaneamento ótico, para um novo ambiente em que a captura das informações já é feita predominantemente em forma digital, usando Dispositivos Móveis de Coleta - DMCs e/ou a Internet. O contingente de funcionários de nível médio envolvidos em tarefas repetitivas (tais como digitação e crítica de informações coletadas) se reduziu substancialmente, sendo parcialmente substituído por funcionários de nível superior mais qualificados, envolvidos principalmente em atividades de planejamento das pesquisas, e de análise e interpretação dos dados e informações coletados.

É nesse cenário que o IBGE de hoje se posiciona como uma moderna agência de pesquisa e produção de informações e estatísticas públicas (também oficiais), voltada para satisfazer as crescentes demandas do Estado e da Sociedade, tanto em nível nacional, como na esfera internacional.

As transformações que o IBGE vivenciou foram, em grande medida, amparadas por desenvolvimentos na área dos métodos estatísticos e sua aplicação aos processos de produção, análise e disseminação de dados e informações estatísticas.

Na sequência, esse texto apresenta a trajetória do emprego de métodos estatísticos na produção e análise das estatísticas oficiais do IBGE desde 1997, em tópicos focando três aspectos principais: **Planejamento amostral; Estimação** e **Modelagem e análise**.

Planejamento amostral

Pesquisas econômicas

O planejamento amostral de pesquisas do IBGE teve alguns avanços expressivos desde 1997. Na área das estatísticas econômicas, o destaque principal ficou por conta da implementação das pesquisas estruturais (anuais) nos segmentos de Indústria, Comércio e Serviços, que substituíram os Censos Econômicos quinquenais a partir de 1996. O planejamento amostral elaborado para a Pesquisa Industrial Anual - Empresa, PIA-Empresa, Pesquisa Industrial Anual - Produto, PIA-Produto e Pesquisa Anual de Comércio - PAC, implementado para as pesquisas com ano de referência 1996, levadas a campo a partir de 1997, foi descrito por Silva e outros (1998).

Estas pesquisas formaram o pilar da reestruturação das estatísticas econômicas dos segmentos de Indústria, Construção Civil, Comércio e Serviços a partir de 1996. Desde então, o IBGE orientou toda a produção de suas estatísticas econômicas com base em pesquisas amostrais, seja de caráter estrutural através das pesquisas anuais, seja de caráter conjuntural, nas pesquisas de periodicidade mais frequente, tipicamente mensais. Além das pesquisas econômicas estruturais e conjunturais, o IBGE passou também a elaborar anualmente estatísticas derivadas dos dados do seu Cadastro Central de Empresas - Cempre, com a primeira edição desta série referente ao ano 2000 publicada em 2002, e incluindo algumas tabelas selecionadas referentes aos anos de 1996-1999 (ESTATÍSTICAS..., 2002).

Complementando o sistema, a Pesquisa Anual de Serviços - PAS foi implantada com ano de referência 1998, e coleta em 1999 (PESQUISA..., 2001). A Pesquisa Mensal do Comércio - PAC foi iniciada em 1995, mas só tornada de abrangência nacional a partir de 2000 (PESQUISA..., 2015a). A Pesquisa Mensal de Serviços - PMS foi iniciada em janeiro de 2011, passando a divulgar indicadores a partir de 2012 (PESQUISA..., 2015b).

Entre as inovações metodológicas importantes referentes a planejamento amostral que foram sendo aplicadas em todas ou quase todas as pesquisas da área econômica, destacam-se:

a) Investimento na consolidação de bases cadastrais de abrangência e cobertura nacionais, estáveis, e com mecanismos de atualização frequente e bem amparados por dados de registros administrativos e das próprias pesquisas amostrais;
b) Adoção da Classificação Nacional de Atividades Econômicas - CNAE que foi transformada no padrão nacional, usada amplamente não só pelo IBGE mas pelos demais agentes públicos que compilam registros administrativos e fiscais; o emprego desta classificação favoreceu o aproveitamento dos registros administrativos como fonte de dados para atualização cadastral;
c) Implementação do mecanismo de atribuição de números aleatórios permanentes às empresas, como ferramenta para facilitar a coordenação e rotação de amostras ao longo do tempo (OHLSSON, 1995); e
d) Coleta de dados via questionários eletrônicos autopreenchidos pelos respondentes, inicialmente em disquetes, e posteriormente via Internet.

Vários estudos foram feitos visando ao aprimoramento da estratificação para as amostras das pesquisas econômicas estruturais. Todavia, os resultados destes estudos não foram ainda apropriados nessas pesquisas. Vale mencionar, entre outros, os seguintes trabalhos de investigação comparativa de métodos de estratificação:

- Azevedo (2004) comparou o desempenho de dois métodos alternativos para estratificação ótima em populações assimétricas, tomando como exemplo uma população de fazendas produtoras de café; entretanto os métodos considerados seriam igualmente aplicáveis a pesquisas de empresas ou estabelecimentos como as que o IBGE faz na indústria, comércio e serviços;
- Almeida (2007) ampliou a comparação de métodos para estratificação ótima em populações assimétricas considerando novos métodos, e aplicações a dados de empresas industriais;
- Brito e outros (2010) propuseram uma implementação de estratificação ótima sob alocação proporcional, e ilustraram o método com dados de algumas populações assimétricas de pesquisas do IBGE;
- Veiga (2015) revisitou o tema da estratificação ótima, expandindo a coleção de métodos comparados, e também o conjunto de populações consideradas na comparação; e
- Brito e outros (2015) apresentaram uma formulação de programação inteira para resolver o problema da alocação ótima de amostras estratificadas, para respostas multivariadas.

Pesquisas domiciliares

Na área das pesquisas amostrais domiciliares, há três focos de inovação metodológica que merecem atenção:

a) Aprofundamento do uso da amostragem nos Censos Demográficos decenais;

b) Reformulação da Pesquisa Mensal de Emprego - PME, nova metodologia, a partir de 2002; e

c) Implantação da Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua a partir de 2012.

Os Censos Demográficos brasileiros utilizaram amostragem para coleta de informações socioeconômicas mais detalhadas desde a edição de 1960. Mas começando em 1991, com a introdução de duas frações amostrais distintas dependendo do tamanho do município, houve uso mais intensivo de amostragem na coleta do questionário mais detalhado, reduzindo em consequência a parcela da população pesquisada com esse questionário. Nos Censos Demográficos 2000, e posteriormente 2010, o uso de mais frações amostrais permitiu reduzir ainda mais a parcela da população pesquisada com o questionário longo, ao mesmo tempo em que favoreceu a obtenção de resultados de melhor qualidade para os pequenos municípios, onde as frações amostrais foram maiores que em edições anteriores do Censo Demográfico (METODOLOGIA..., 2003, 2016).

O IBGE considerou também alternativas para a forma de condução do Censo Demográfico decenal, tendo investido na apreciação de um projeto para a utilização do modelo de censo contínuo. Embora tenha optado por não implementar mudanças dessa natureza para a rodada de 2020, os estudos realizados foram inovadores e permitiram conhecer com detalhes as várias alternativas disponíveis e que vem sendo usadas por outros países[1].

[1] Mais detalhes sobre os estudos e resultados do projeto de pesquisa realizado pelo IBGE sobre o tema podem ser encontrados no endereço <http://www.ibge.gov.br/home/estatistica/populacao/censo_continuo/default.shtm>, no portal do IBGE, na Internet.

Quanto à PME, nova metodologia, implementada a partir de março de 2002, a principal novidade foi a correção do mecanismo de rotação da amostra de domicílios. Essa correção compreendeu o emprego de oito grupos de rotação, como necessário para implementar de forma adequada o mecanismo 4-8-4, evitando assim a introdução de desequilíbrios na amostra que poderiam resultar em viés de grupos de rotação, conforme descrito em Pfeffermann, Silva e Freitas (2000).

Mas a grande novidade no âmbito das pesquisas amostrais domiciliares foi a substituição da Pesquisa Nacional por Amostra de Domicílios - PNAD, cuja série foi encerrada com a edição de 2015 da pesquisa, e da PME, cuja série foi encerrada em fevereiro de 2016, pela PNAD Contínua, cuja série foi iniciada em 2012 (PESQUISA..., 2014).

Quanto ao planejamento amostral, a PNAD Contínua inovou em relação à PNAD em vários aspectos importantes:

a) A pesquisa empregou amostragem conglomerada em apenas dois estágios de seleção, conseguindo com isso um espalhamento muito maior da amostra no território nacional;
b) O sorteio das Unidades Primárias de Amostragem - UPAs – (setores censitários ou agregados destes formados para ter um tamanho mínimo igual a 60 domicílios) foi feito com amostragem com Probabilidades Proporcionais ao Tamanho - PPT segundo método de Pareto (ROSÉN, 1997, 2000); esse método de sorteio das UPAs favoreceu também a introdução de mecanismo para controle da rotação das UPAs ao longo do tempo, o que vai permitir a renovação parcial da amostra de UPAs a cada nova rodada da pesquisa; o estudo precursor de Costa (2007) pavimentou o caminho para a adoção desta abordagem na seleção e manutenção da amostra mestra do Sistema Integrado de Pesquisas Domiciliares do IBGE;
c) Foi empregada uma estratificação mais detalhada das unidades primárias de amostragem, permitindo algum ganho de eficiência; um estudo precursor da estratificação para uma pesquisa como essa foi feito por Freitas (2002);
d) No segundo estágio, o sorteio dos domicílios particulares permanentes ocupados dentro de cada UPA da amostra passou a ser de tamanho fixo (14 domicílios por UPA), e a ser efetuado por amostragem aleatória simples, considerando informações atualizadas do Cadastro Nacional de Endereços para Fins Estatísticos - Cnefe; o uso do número fixo de domicílios corrigiu um velho problema da PNAD e PME, que era o crescimento descontrolado das amostras nos setores, causador de cargas de trabalho desiguais para os entrevistadores;
e) Ao usar um tamanho um pouco menor de amostra por UPA (14 em vez de 16 como nas pesquisas substituídas), junto com a eliminação do estágio de sorteio de municípios, a PNAD Contínua aumentou o espalhamento da amostra e reduziu o efeito de conglomeração, aumentando a precisão das estimativas amostrais; e
f) A utilização do esquema de rotação 1-2(5) – um mês na pesquisa, dois fora, repetido cinco vezes – favoreceu o acompanhamento dos domicílios e moradores em intervalos regulares, por cinco trimestres; este arranjo é inédito em pesquisas domiciliares de abrangência nacional no Brasil, e favorece a investigação de temas que precisem do acompanhamento longitudinal das unidades pesquisadas (domicílios e/ou seus moradores); favorece também a obtenção de resultados mais frequentes para as variáveis contidas no questionário básico da pesquisa, e a introdução de pesquisas suplementares que se aproveitem das múltiplas ocasiões em que cada domicílio é pesquisado durante sua permanência na amostra; o aproveitamento pleno dos resultados desta inovação ainda é uma promessa, pois o IBGE ainda não divulgou junto dos microdados públicos uma chave de identificação de registros de pessoas que permita fazer o acompanhamento longitudinal ao longo das várias rodadas da pesquisa, mas há promessa de que tal recurso venha a ser disponibilizado nos próximos meses.

Pesquisas agropecuárias

Uma área na qual houve pouco avanço na utilização de métodos mais modernos de amostragem foi a das estatísticas agropecuárias. O IBGE vem se preparando para uma reformulação de seu sistema de estatísticas agropecuárias com as discussões do projeto de Reformulação das Pesquisas Agropecuárias, disponível no endereço <http://www.ibge.gov.br/home/estatistica/indicadores/prpa/default.shtm>, na Internet. Todavia, a implantação do Sistema Nacional de Pesquisas por Amostragem de Estabelecimentos Agropecuários - SNPA é ainda uma promessa, dependente da realização do Censo Agropecuário 2017, para viabilizar sua concretização (PROPOSTA..., 2011).

Infraestrutura para pesquisas amostrais

Uma das áreas em que o aprimoramento da infraestrutura de pesquisa foi notável foi a digitalização da cartografia e bases operacionais disponíveis para apoiar os censos de população e agropecuário, bem como todas as pesquisas amostrais domiciliares. A digitalização da cartografia chegou tarde ao IBGE, em comparação com outros institutos nacionais de estatística, mas sua chegada viabilizou avanços importantes no planejamento, suporte e controle da execução da coleta, processamento e disseminação dos dados de muitas pesquisas, mas notadamente nos censos demográficos.

A disseminação de resultados dos censos passou a tornar disponíveis produtos que se mostraram valiosos do ponto de vista da comunidade de usuários. Entre os produtos que se beneficiaram de maneira importante da cartografia digital se encontram:

a) Arquivos agregados por setores censitários, agora acompanhados de arquivos digitais que permitem mapear por setores censitários inúmeras variáveis cujos resultados são tabulados nesse nível, disponíveis no endereço <ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/>;
b) Arquivos de microdados da amostra, com a identificação de áreas de ponderação, que permitem tabular todas as variáveis do questionário da amostra por áreas de ponderação e posteriormente mapear tais resultados nesse nível de agregação, disponíveis no endereço <http://www.ibge.gov.br/home/estatistica/populacao/censo2010/resultados_gerais_amostra/resultados_gerais_amostra_tab_uf_microdados.shtm>; e
c) Resultados divulgados através da Grade Estatística 2010, elaborada para o Censo Demográfico 2010, onde pela primeira vez os usuários terão a liberdade de elaborar estatísticas sobre total de população e de domicílios para áreas livremente definidas a partir dos recortes mínimos definidos para a grade, disponíveis no endereço <http://mapasinterativos.ibge.gov.br/grade/default.html>.

Estimação

No campo dos métodos ligados à estimação com base em pesquisas amostrais houve muitos avanços. Tanto a PNAD quanto a PME passaram a utilizar etapas de tratamento de não resposta por reponderação simples em nível de setor censitário. No caso da PNAD, isto se deu a partir das PNADs da década de 2000. No caso da PME, a partir da reformulação para introdução da Nova Metodologia, a partir de 2002. Em rodadas anteriores destas pesquisas não havia tal etapa de ponderação, e a única correção dos pesos amostrais básicos era feita com base no estimador tipo razão para fazer com que os pesos totalizassem as populações totais projetadas por unidade da federação ou pós-estratos tais como as regiões metropolitanas, nos estados onde estas são consideradas no planejamento amostral. Uma consequência da introdução de reponderação por setor censitário é a redução de viés devido à não resposta diferencial entre setores de diferentes tipos ou características.

Apesar deste avanço no caso das duas pesquisas, o problema da não resposta diferencial seguiu sem um tratamento mais completo até a extinção das duas pesquisas. No caso da PNAD, Miguel Ruiz (2015) mostrou que havia indícios de viés causado pela não resposta diferencial nas PNADs de 2005 a 2013. Mostrou também que tal viés poderia ser parcialmente compensado com a utilização de estimadores de calibração que considerassem as distribuições por sexo e idade da população, além das projeções dos totais populacionais por grandes estratos geográficos, únicos dados auxiliares que eram regularmente considerados na ponderação das duas pesquisas. Todavia, a série da PNAD foi encerrada sem que seu processo de ponderação fosse modificado para incorporar o tratamento proposto por Miguel Ruiz (2015).

No caso da PME, indícios de não resposta diferencial foram identificados em diversas ocasiões. Uma dificuldade central foi relatada por Pfeffermann, Silva e Freitas (2000), que observaram efeitos importantes de viés de grupos de rotação nas estimativas das taxas de desocupação da pesquisa na década de 1990. Tais efeitos eram, segundo os autores, consequência da implementação equivocada do esquema de rotação da pesquisa, que usava apenas quatro grupos de rotação, em vez dos oito necessários para uma distribuição equilibrada da amostra de cada mês conforme o tempo de permanência dos domicílios pesquisados na amostra. A implementação da PME Nova Metodologia a partir de 2002 corrigiu este defeito da pesquisa, mas nenhuma medida foi tomada para alertar os usuários ou aplicar correções nas séries divulgadas para as décadas anteriores.

Figueiredo (2003) estudou em detalhes os padrões de observação e não resposta ao longo das oito ocasiões que um domicílio poderia participar da amostra da PME, tendo fornecido evidências da presença de viés de grupos de rotação e iniciado estudo de métodos que permitiriam mitigar esse viés.

Lapa (2015) apontou indícios de não resposta diferencial em relação à situação ocupacional, ao estimar os fluxos brutos desta variável entre meses adjacentes usando dados da PME para a Região Metropolitana de Salvador no período janeiro de 2008 a julho de 2013. Em seu trabalho, Lapa (2015) utilizou uma abordagem proposta por Gutiérrez (2014), também descrita em Gutiérrez, Trujillo e Silva (2014), para estimar os fluxos brutos de uma variável categórica com dados de uma pesquisa amostral complexa, sujeita a não resposta diferencial. Esse trabalho corroborou as evidências iniciais disponíveis em Figueiredo (2003) e em Gutiérrez (2014) de que as probabilidades de resposta em meses adjacentes na PME dependem da situação ocupacional dos indivíduos. Não houve, entretanto, apropriação pela PME dos métodos e ferramentas desenvolvidos para mitigar este problema até o encerramento da pesquisa em fevereiro de 2016.

Tanto a PNAD como a PME incorporaram, ao longo da sua última década, métodos de imputação para tratamento de não resposta de itens. No caso da PME Nova Metodologia (PESQUISA..., 2007), o sistema Detección e Imputación Automática de Errores para Datos Cualitativos - DIA) (GARCÍA RUBIO; VILLÁN CRIADO, 1988) passou a ser a principal ferramenta usada para crítica e imputação de dados, lidando tanto com a não resposta de itens, mas também com as combinações de valores inaceitáveis. Já a imputação dos quesitos de renda passou a ser feita usando Árvores de Regressão (BREIMAN et al., 1993) para formar classes de imputação, combinada com a seleção probabilística de registros doadores em cada classe construída através da árvore (PESQUISA..., 2007; PESSOA; SILVA; SANTOS, 2000).

Por se tratar de uma pesquisa repetida de painéis rotativos, a PME permitia a observação repetida de uma amostra de domicílios e de seus moradores. Embora não tenha sido desenhada para ser uma pesquisa longitudinal, e por isto não fazer o esforço de seguir os indivíduos selecionados ao longo do tempo, esta possibilidade de observação repetida dos mesmos indivíduos criou potencial para várias análises de natureza longitudinal. Entre as ferramentas desenvolvidas para explorar este potencial, vale citar as estratégias de pareamento estatístico de indivíduos feitas com dados da PME, entre as quais merecem destaque Lopes (2002), Figueiredo (2003), Ribas e Soares (2009) e Gutiérrez e Ortiz (2012).

A estrutura de repetição da amostra foi aproveitada por Teixeira Júnior (2015) para a produção de pesos longitudinais para estimação e análise de dados da PME. Segundo Teixeira Júnior (2015, f. viii):

> Pesos longitudinais foram produzidos para três diferentes escopos de análise: evolução mensal, anual e entre a primeira e última visita ao informante (painel completo). Os resultados [...] demonstram a importância da produção e utilização de pesos longitudinais, indicando que o uso dos pesos publicados pela pesquisa não corrigem adequadamente o viés de não resposta para o caso de estudos com foco em análises longitudinais.

Um estudo feito por Santos, Brito e Silva (2004) analisou a possibilidade de utilização de métodos de calibração composta (FULLER; RAO, 2001; GAMBINO; KENNEDY; SINGH, 2001) para a PME. A calibração composta é uma técnica que permite incorporar informações auxiliares via estimadores tipo regressão em pesquisas repetidas com sobreposição amostral. Para sua implementação, é necessário fazer o pareamento de observações em ocasiões adjacentes da pesquisa. Os resultados obtidos nos estudos iniciais não foram promissores, e a ideia não foi adiante em termos de sua aplicação na pesquisa.

Na área das pesquisas econômicas estruturais, foi incorporada a aplicação de métodos de tratamento de não resposta e ponderação das amostras com base em estimadores de calibração do tipo regressão, que aproveitam informação populacional auxiliares disponíveis no cadastro para combater viés e reduzir variância (SILVA et al., 1999). O emprego destes tipos de estimadores tem permitido obter estimativas de melhor precisão que as que seriam derivadas de estimadores que não tirassem proveito das ricas informações auxiliares disponíveis nos cadastros usados para as pesquisas de empresas.

Nascimento (2002) investigou o emprego de estimadores do tipo de calibração composta para a Pesquisa Industrial Mensal de Dados Gerais - PIM-DG. O estudo concluiu que tais estimadores tinham viés menor que os estimadores em uso naquela pesquisa, na ocasião. Uma vantagem adicional é que tais estimadores permitem a combinação de dados das pesquisas conjunturais, cujas amostras são tipicamente menores, com os dados de pesquisas estruturais (anuais), cujas amostras são geralmente bem maiores. A combinação de dados é feita de tal forma que são aproveitadas as informações das pesquisas anuais na ponderação das pesquisas conjunturais (mensais), levando à melhoria tanto das estimativas de variações de curto como de médio/longo prazo a partir das pesquisas conjunturais. Apesar das evidências favoráveis apontadas no estudo, esta abordagem de estimação não foi incorporada à pesquisa.

Um dos desafios que vêm sendo enfrentados pelo IBGE, e também pelas demais agências produtoras de informações estatísticas, é um aumento sem precedentes da demanda por informações. Holt (2007) destaca a situação como sendo o "desafio Olímpico das estatísticas oficiais: mais amplas, mais detalhadas; mais frequentes e atuais; melhores e mais baratas". Nesse contexto, uma das direções de expansão da demanda é a da produção de informações para domínios ou áreas menores, para as quais as amostras disponíveis são pequenas demais para fornecer estimativas diretas com a precisão desejada.

Rao (2003) apresenta as abordagens disponíveis para estimação em pequenas áreas, que contém as ferramentas disponíveis para ajudar a tentar satisfazer a demanda acima descrita. Em *Mapa de pobreza e desigualdade: municípios brasileiros* (2008) há uma primeira aplicação em larga escala de métodos de estimação para pequenas áreas, ao elaborar o chamado 'mapa de pobreza' com estimativas para indicadores selecionados de pobreza em nível de municípios. Este projeto requereu a combinação de dados da Pesquisa de Orçamentos Familiares - POF 2002-2003 com dados da amostra do Censo Demográfico 2000.

Os métodos empregados pelo IBGE no projeto do mapa de pobreza foram propostos por Elbers, Lanjouw e Lanjouw (2002). Molina e Rao (2010) mostraram que é possível obter estimadores mais eficientes para o mesmo tipo de problema. Antonaci (2012) e Antonaci, Silva e Moura (2013) compararam os dois métodos citados no contexto da elaboração de mapas de pobreza com os mesmos dados usados pelo IBGE. Os resultados mostraram a superioridade do método denominado *Hierarchical Bayes* - HB

proposto por Molina e Rao (2010) para esta aplicação. Até o momento, o IBGE não voltou a produzir novas estimativas de mapas de pobreza, apesar de terem ficado disponíveis os dados da POF 2008-2009.

Na área das pesquisas econômicas, Neves (2012) estudou métodos de estimação em pequenos domínios aplicados à PAS do IBGE. Os pequenos domínios considerados foram determinadas atividades de serviços localizadas nas Regiões Norte, Nordeste e Centro-Oeste do país. Moura, Neves e Silva (2017) aprimoraram os métodos de Neves (2012), e ilustraram como métodos de estimação baseados em modelos para dados assimétricos ou com caudas pesadas podem ser úteis quando os dados têm distribuições que não são bem aproximadas pela distribuição normal. Apesar dos progressos na apropriação dos modernos métodos de estimação para pequenas áreas, a abordagem considerada não foi ainda aplicada para a produção de estimativas que o IBGE planeje publicar como parte dos resultados das pesquisas econômicas estruturais.

Modelagem e análise

Uma questão onde houve grande progresso nos últimos 20 anos foi a de modelagem e análise dos dados das pesquisas do IBGE. A publicação do artigo de Pessoa, Silva e Duarte (1997), do livro de Pessoa e Silva (1998) e a realização do minicurso sobre Análise de dados amostrais complexos no 13º Simpósio Nacional de Probabilidade e Estatística - Sinape, em 1998, foram marcos importantes para a disseminação e adoção de práticas mais adequadas de análise de microdados de pesquisas amostrais no IBGE e no Brasil.

A ampliação da prática de disseminação de microdados anonimizados de pesquisas domiciliares tais como a PNAD, PME, POF e da amostra do Censo Demográfico usando a Internet, e a disponibilidade de software especializado que permite a fácil aplicação de modelos e métodos que incorporam os efeitos da amostragem complexa nas análises propiciaram grande aumento da utilização dos microdados pela comunidade de usuários das pesquisas do IBGE. Esse aumento do uso gerou um aumento do valor dos dados, mediante sua exploração e aproveitamento em estudos analíticos, em contraste com os usos mais descritivos das décadas anteriores.

Entre os muitos usos inovadores de métodos capazes de incorporar os efeitos do plano amostral complexo vale citar, por seu pioneirismo, os seguintes:

- Correa (2001) empregou modelos lineares hierárquicos para relacionar o índice de massa corporal com variáveis preditoras usando dados da Pesquisa Sobre Padrões de Vida - PPV do IBGE;
- Cruz (2001) estudou a estimação de variâncias para séries temporais dessazonalizadas pelo método X-12 ARIMA, considerando o desenho amostral das pesquisas geradoras das séries; Cruz e Silva (2002) avançaram no tema oferecendo um primeiro curso sobre a análise de séries temporais de pesquisas amostrais durante o 15º Simpósio Nacional de Probabilidade e Estatística - Sinape, em 2002;
- Leite (2001) aplicou modelos logísticos para analisar a situação ocupacional de crianças e adolescentes nas Regiões Sudeste e Nordeste do Brasil, utilizando informações da PNAD 1999, incorporando aspectos da amostragem complexa nas análises; e
- Rodrigues (2003) analisou a estrutura salarial revelada pela PPV incorporando pesos e plano amostral.

Os trabalhos acima foram os primeiros de uma longa lista de análises de dados de pesquisas amostrais do IBGE levadas a cabo por alunos do mestrado da Escola Nacional de Ciências Estatísticas - ENCE. A criação deste mestrado, e a introdução de uma disciplina de Análise de dados amostrais, em sua grade curricular, serviram de plataforma para ampliar a disseminação do conhecimento sobre as ferramentas, métodos e modelos disponíveis.

Conclusões e desafios

Este artigo descreveu em parte a trajetória do emprego de métodos estatísticos na produção e análise das estatísticas oficiais do IBGE desde 1997, com foco nas questões de planejamento amostral, estimação e análise de dados amostrais. Tais áreas foram, desde sua criação, pontos fortes da atuação da área de métodos do IBGE.

Muitos avanços da pesquisa nessas áreas continuam a ser feitos, e muitas questões permanecem como oportunidades para aprimoramento do emprego da amostragem nas pesquisas do IBGE.

Entre as novidades e desafios mais recentes se coloca a necessidade de buscar a combinação de fontes de dados, sejam elas convencionais (pesquisas amostrais, censos e registros administrativos, etc.) ou novidades (*Big Data* ou dados orgânicos diversos). Tais estratégias de combinação de informações serão essenciais no futuro para permitir o atendimento das crescentes demandas sobre o sistema de estatísticas oficiais.

Como revela a revisão aqui apresentada, a combinação de pesquisa dirigida às necessidades dos projetos e atividades do IBGE, combinada com o ensino e a disseminação dos melhores métodos e práticas formam pilares importantes do trabalho da área de métodos e qualidade do IBGE.

É nesse contexto rico e desafiador que a área de métodos do IBGE terá que continuar evoluindo, de modo que possa continuar sua trajetória de sucesso e bons serviços prestados ao IBGE.

Referências

ALBIERI, S.; BIANCHINI, Z. M.; CARDOSO, R. L. *Pesquisa domiciliar sobre padrões de vida*: planejamento da amostra. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1995. 25 p.

ALMEIDA, A. G. R. d'. *Métodos para delimitação de estratos em populações assimétricas*: uma comparação. 2007. 145 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2007. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2007/Resumo_Dissertacao_2007__Adriane_Gonzalez.pdf>. Acesso em: set. 2017.

ANTONACI, G. de A. *Comparação de métodos para estimação de índices de pobreza em pequenas áreas*. 2012. 100 p. Dissertação (Mestrado)- Instituto de Matemática, Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro, 2012. Disponível em: <http://www.pg.im.ufrj.br/teses/Estatistica/Mestrado/138.pdf>. Acesso em: set. 2017.

ANTONACI, G.; SILVA, P. L. do N.; MOURA, F. A. da S. A comparison of methods to estimate poverty indexes in small samples. In: ISI WORLD STATISTICS CONGRESS, 59., 2013, Hong Kong. *Proceedings*… The Hague [Holanda]: International Statistical Institute - ISI, 2013. p. 3737-3742. Disponível em: <https://www.isi-web.org/index.php>. Acesso em: set. 2017.

AZEVEDO, R. V. de. *Estudo comparativo de métodos de estratificação ótima de populações assimétricas*. 2004. 121 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2004. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2004/rosemary_vallejo_de_azevedo_TC.pdf>. Acesso em: set. 2017.

BREIMAN, L. et al. *Classification and regression trees*. Boca Raton: Chapman & Hall, 1993. 358 p.

BRITO, J. A. de M. et al. Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, Ottawa: Statistics Canada, v. 41, n. 2, p. 427-442, Dec. 2015. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/index-eng.htm>. Acesso em: set. 2017.

BRITO, J. et al. An exact algorithm for the stratification problem with proportional allocation. *Optimization Letters*, New York: Springer, v. 4, n. 2, p. 185-195, May 2010. Disponível em: <https://link.springer.com/journal/volumesAndIssues/11590>. Acesso em: set. 2017.

CORREA, S. T. *Modelos lineares hieráquicos em pesquisas por amostragem*: relacionando o índice de massa corporal com as variáveis da pesquisa sobre padrões de vida. 2001. 110 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2001. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2001/solange_correa_TC.pdf>. Acesso em: set. 2017.

COSTA, G. T. L. da. *Coordenação de amostras PPT em pesquisas repetidas, utilizando o método de amostragem de Pareto*. 2007. 94 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2007.

CRUZ, M. M. *Estimação de variâncias para séries dessazonalizadas pelo método X-12 ARIMA, considerando o desenho amostral*. 2001. 207 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2001. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2001/marcelo_martins_cruz_TC.pdf>. Acesso em: set. 2017.

CRUZ, M. M.; SILVA, D. B. do N. *Séries temporais de pesquisas amostrais periódicas*. São Paulo: Associação Brasileira de Estatística - ABE, 2002. 141 p. Minicurso apresentado no 15º Simpósio Nacional de Probabilidade e Estatística - Sinape, realizado em Águas de Lindóia, 2002.

ELBERS, C.; LANJOUW, J. O.; LANJOUW, P. Micro-level estimation of welfare. *Policy Research Working Paper*, Washington, DC: World Bank, n. 2911, 2002.

ESTATÍSTICAS do cadastro central de empresas 2000. Rio de Janeiro: IBGE, 2002. 223 p. Acompanha 1 CD-ROM. Disponível: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv1039.pdf>. Acesso em: set. 2017.

FIGUEIREDO, J. da S. *Não-resposta diferencial e tendenciosidade de rotação na Pesquisa Mensal de Emprego do IBGE*. 2003. 150 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2003.

FREITAS, M. P. S. de. *Estratificação para a amostra de uma pesquisa domiciliar sobre mercado de trabalho*. 2002. 114 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2002. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2002/Resumo_Dissertacao_2002_Marcos_Paulo_Soares_de_Freitas.pdf>. Acesso em: set. 2017.

FULLER, W. A.; RAO, J. N. K. A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, Ottawa: Statistics Canada, v. 27, n. 1, p. 45-51, June 2001. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/index-eng.htm>. Acesso em: set. 2017.

GAMBINO, J.; KENNEDY, B.; SINGH, M. P. Regression composite estimation for the Canadian Labour Force Survey: evaluation and implementation. *Survey Methodology*, Ottawa: Statistics Canada, v. 27, n. 1, p. 65-74, June 2001. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/index-eng.htm>. Acesso em: set. 2017.

GARCÍA RUBIO, E.; VILLÁN CRIADO, I. *Sistema DIA*: sistema de detección e imputación automática de errores para datos cualitativos. Madrid: Instituto Nacional de Estadística - INE, 1988.

GUTIÉRREZ ROJAS, H. A. *Modelos para estimar cambios brutos en encuestas rotativas com ausencia de respuesta em diseños de muestreo complejos*. 2014. 165 p. Tese (Doutorado)- Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, 2014. Disponível em: <http://www.bdigital.unal.edu.co/39564/1/hugoandresgutierrezrojas.pdf>. Acesso em: set. 2017.

GUTIÉRREZ, A.; ORTIZ, J. Emparejamiento de paneles y clasificación de la ausencia de respuesta en la Pesquisa Mensal de Emprego usando funciones en R. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 73, n. 237, p. 75-118, jul./dez. 2012. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

GUTIÉRREZ, A.; TRUJILLO, L.; SILVA, P. L. do N. The estimation of gross flows in complex surveys with random nonresponse. *Survey Methodology*, Ottawa: Statistics Canada, v. 40, n. 2, p. 285-321, Dec. 2014. Disponível em: <http://www.statcan.gc.ca/pub/12-001-x/index-eng.htm>. Acesso em: set. 2017.

HOLT, D. T. The official statistics olympic challenge: wider, deeper, quicker, better, cheaper. *The American Statistician*, Alexandria [Estados Unidos]: American Statistical Association - ASA, v. 61, n. 1, 2007.

LAPA, P. P. de A. *Estimando fluxos das situações ocupacionais com pesquisas amostrais repetidas*. 2015. 157 p. Dissertação (Mestrado) - Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2015. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2015/Dissertacao_priscilapagung.pdf.>. Acesso em: set. 2017.

LEITE, P. G. P. G. *Análise da situação ocupacional de crianças e adolescentes nas Regiões Sudeste e Nordeste do Brasil*: utilizando informações da PNAD 1999. 2001. 159 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2001. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2001/phillippe_george_pereira_guimaraes_leite_TC.pdf>. Acesso em: Set. 2017.

LOPES, M. D. *Avaliação de desgaste de painéis em estudos longitudinais*: uma aplicação na Pesquisa Mensal de Emprego (PME/IBGE). 2002. 106 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2002. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2002/marcio_duarte_lopes_TC.pdf>. Acesso em: set. 2017.

MAPA de pobreza e desigualdade: municípios brasileiros 2003. Rio de Janeiro: IBGE, 2008. 1 DVD.

METODOLOGIA do censo demográfico 2000. Rio de Janeiro: IBGE, 2003. (Série relatórios metodológicos, 25). Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2000/metodologia/default.shtm>. Acesso em: set. 2017.

METODOLOGIA do censo demográfico 2010. 2. ed. 720 p. Rio de Janeiro: IBGE, 2016. (Série relatórios metodológicos, v. 41). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv95987.pdf>. Acesso em: set. 2017.

MIGUEL RUIZ, C. M. *Explorando alternativas para a calibração dos pesos amostrais da Pesquisa Nacional por Amostra de Domicílios*. 2015. 111p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2015. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2015/disertacao_final_Charles_Ruiz.pdf.>. Acesso em: set. 2017.

MOLINA, I.; RAO, J. N. K. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, Ottawa: Statistical Society of Canada, v. 38, n. 3, p. 369-385, Sep. 2010. Disponível em: <http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1708-945X>. Acesso em: set. 2017.

MOURA, F. A. S.; NEVES, A. F.; SILVA, D. B. do N. Small area models for skewed Brazilian business survey data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, London: Royal Statistical Society - RSS, June 2017. Disponível em: <http://rss.onlinelibrary.wiley.com/hub/journal/10.1111/(ISSN)1467-985X/>. Acesso em: set. 2017.

NASCIMENTO, V. L. da S. A. do. *Estimação em pesquisas repetidas no tempo*: uma aplicação à Pesquisa Industrial Mensal de Dados Gerais do IBGE. 2002. 122 p. Dissertação (Mestrado)-Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2002. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2002/vera_lucia_da_silva_augusto_do_nascimento_TC.pdf>. Acesso em: set. 2017.

NEVES, A. F. A. Estimação em pequenos domínios aplicada à pesquisa anual de serviços 2008. 195 p. Dissertação (Mestrado)-Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2012. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2012/Dissertacao_2012_Andre_Felipe_Azevedo_Neves.pdf>. Acesso em: set. 2017.

OHLSSON, E. Coordination of samples using permanent random numbers. In: COX, B. et al. (Ed.). Business survey methods. New York: Wiley, c1995. p. 153-169.

PESQUISA ANUAL DE SERVIÇOS 1998-1999. Rio de Janeiro: IBGE, 2001. v.1. Disponível em: <https://ww2.ibge.gov.br/home/estatistica/indicadores/servicos/pms/default.shtm>. Acesso em: set. 2017.

PESQUISA mensal de comércio. 4. ed. Rio de janeiro: IBGE, 2015a. 64 p. (Série relatórios metodológicos, v. 15). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv93106.pdf>. Acesso em: set. 2017.

PESQUISA mensal de emprego. 2. ed. Rio de Janeiro: IBGE, 2007. 89 p. (Série relatórios metodológicos, v. 23). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv37313.pdf>. Acesso em: set. 2017.

PESQUISA mensal de serviços. Rio de Janeiro: IBGE, 2015b. 50 p. (Série relatórios metodológicos, v. 42). Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv93108.pdf>. Acesso em: set. 2017.

PESQUISA nacional por amostra de domicílios contínua: notas metodológicas. v. 1. Rio de Janeiro: IBGE, 2014. 47 p. Disponível em: <ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf>. Acesso em: ago. 2016.

PESSOA, D. G. C.; SILVA, P. L. do N. *Análise de dados amostrais complexos*. Rio de Janeiro, 1998. 170 p. Disponível em: <http://www.ie.ufrj.br/download/livro.pdf>. Acesso em: set. 2017.

PESSOA, D. G. C.; SILVA, P. L. do N.; DUARTE, R. P. N. Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 58, n. 210, p. 53-75, jul./dez. 1997. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

PESSOA, D. G. C.; SILVA, P. L. do N.; SANTOS, A. R. Imputação para não-resposta parcial de renda na pesquisa mensal de emprego. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2000. Relatório técnico.

PFEFFERMANN, D.; SILVA, P. L. do N.; FREITAS, M. P. S. de. *Implicações do esquema de rotação da pesquisa mensal de emprego do IBGE na qualidade das estimativas publicadas*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 2000. Não publicado.

PROPOSTA de Sistema Nacional de Pesquisas por Amostragem de Estabelecimentos Agropecuários - SNPA: concepção geral e conteúdo temático: 2a versão. Rio de Janeiro: IBGE, 2011. 47 p. (Texto para discussão). Disponível em: <https://ww2.ibge.gov.br/home/estatistica/indicadores/prpa/SNPA_concepcao_e_conteudo2av.pdf>. Acesso em: set. 2017.

RAO, J. N. K. *Small area estimation*. Hoboken, N. J.: Wiley, 2003. 313 p. (Wiley series in survey methodology).

RIBAS, R. P.; SOARES, S. S. D. Sobre o painel da Pesquisa Mensal de Emprego - PME do IBGE: problemas e soluções para o emparelhamento usando microdados. *Revista Brasileira de Estatística*, Rio de Janeiro: IBGE, v. 70, n. 233, p. 75-108, jul./dez. 2009. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7111>. Acesso em: set. 2017.

RODRIGUES, S. de C. *Análise da estrutura salarial revelada pela PPV incorporando peso e plano amostral*. 2003. 122 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2003. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2003/simone_de_castro_rodrigues.pdf>. Acesso em: set. 2017.

ROSÉN, B. A user's guide to Pareto πps sampling, R&D Report: research, methods, development, Stockholm: Statistiska centralbyrån, n. 6, Dec. 2000. 10 p. Disponível em: <http://www.scb.se/contentassets/14f5e346f4814dd0acd52d10b23286c6/rnd-report-2000-06-green.pdf>. Acesso em: set. 2017.

ROSÉN, B. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, Amsterdam: Elsevier, v. 62, n. 2, p. 159-191, Aug. 1997. Disponível em: <https://www.journals.elsevier.com/journal-of-statistical-planning-and-inference>. Acesso em: set. 2017.

SANTOS, D. P. dos; BRITO, J. A. de M.; SILVA, P. L. do N. *Estimação da taxa de desemprego aberto, com dados da Pesquisa Mensal de Emprego, no período 1998-2000, utilizando calibração composta*. Rio de Janeiro: IBGE, 2004.

SILVA, P. L. do et al. *Procedimentos de estimação utilizados na Pesquisa Industrial Anual e na Pesquisa Anual do Comércio*. Rio de Janeiro: IBGE, 1999. 14 p. Não publicado.

SILVA, P. L. do N. et al. *Planejamento amostral para as pesquisas anuais da indústria e do comércio*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1998. 40 p. (Textos para discussão, n. 92). Disponível em: <ftp://ftp.dpe.ibge.gov.br/gdi/texdisc/texdisc92-98.pdf>. Acesso em: set. 2017.

TEIXEIRA JÚNIOR, A. E. *Produção de pesos longitudinais para estimação e análise de dados da Pesquisa Mensal de Emprego do IBGE*. 2015. 170 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2015. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2015/Dissertacao_AntonioEtevaldo.pdf>. Acesso em: set. 2017.

VEIGA, T. M. da. *Um estudo comparativo de métodos para delimitação de estratos em populações assimétricas*. 2015. 97 p. Dissertação (Mestrado)- Escola Nacional de Ciências Estatísticas - ENCE, Rio de Janeiro, 2015. Disponível em: <http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2015/Dissertacao_Tomas.pdf>. Acesso em: set. 2017.

Sobre os autores

Alexandre dos Reis Santos

Graduado em Estatística pela Escola Nacional de Ciências Estatísticas - ENCE e Mestre em Estatística pelo Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ. Servidor do IBGE desde fevereiro de 2002, tendo atuado até abril de 2017 na unidade de metodologia da Diretoria de Pesquisas. Atualmente é lotado da Supervisão de Pesquisas Sociais da Unidade Estadual do IBGE no Espírito Santo.

Antonio José Ribeiro Dias

Possui graduação em Estatística pela Universidade Estadual de Campinas - Unicamp, aperfeiçoamento em Estatística pela Associação Instituto Nacional de Matemática Pura e Aplicada - IMPA e Mestrado em Engenharia de Sistemas pelo Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia - Coppe, da UFRJ. Servidor do IBGE desde janeiro de 1981, tendo atuado a maior parte do tempo na unidade de metodologia da Diretoria de Pesquisas. Atuou, também, durante mais de 20 anos como professor colaborador da Graduação em Estatística da ENCE, na área de amostragem. Atualmente é Gerente de Metodologia Estatística da Coordenação de Métodos e Qualidade.

Ari do Nascimento Silva

Graduado em engenharia pela UFRJ, e Mestre em Ciência da Computação pela Universidade da Califórnia em Los Angeles (University of California - UCLA). Tem como principais áreas de interesse Bancos de Dados Estatísticos, Desenvolvimento de Software, Gerência de Grandes Projetos em Computação, Informação sobre Dados de População e Sistema de Informações Geográficas - SIG (Geographic Information System - GIS).

Sua experiência no Brasil começou em 1970 na Petrobras, em desenvolvimento de software, e depois, em 1972, no IBGE, onde trabalhou, entre outros, como analista de sistemas, chefe de desenvolvimento para processamento de censos, e encerrou ali a carreira como consultor na Comeq. Trabalhou por 12 anos na Organização das Nações Unidas - ONU, no Centro Latinoamericano

de Demografia (Centro Latinoamericano y Caribeño de Demografía - Celade), em Santiago do Chile, como chefe da divisão de processamento de dados e assessor em processamento de dados demográficos para todos os países da região. Participou desde seu início, e ainda participa ativamente no projeto de desenvolvimento da Recuperação de Dados para Áreas Pequenas por Microcomputador (Retrieval of Data for Small Areas by Microcomputer - Redatam). Atualmente trabalha como consultor privado em processamento de censos e pesquisas.

Bruno Freitas Cortez

Possui graduação em Estatística pela ENCE, MBA em Finanças e Mercado de Capitais pela Universidade Veiga de Almeida - UVA e Mestrado em Estudos Populacionais e Pesquisas Sociais pela ENCE. Servidor do IBGE desde novembro de 1999, contratado temporariamente como Analista Censitário e a partir de fevereiro de 2002 servidor efetivo concursado, tendo atuado todo este tempo na unidade de metodologia da Diretoria de Pesquisas. Atua também como professor colaborador nos programas de treinamento do IBGE nos cursos de estatística básica e crítica e imputação de dados com o sistema Canadian Census Edit and Imputation System - Canceis. Atualmente trabalha na Gerência de Metodologia Estatística da Coordenação de Métodos e Qualidade.

Débora Ferreira De Souza

Possui graduação, mestrado e doutorado em Estatística pelo Instituto de Matemática da UFRJ. Servidora do IBGE desde janeiro de 2002, atualmente lotada na Coordenação de Métodos e Qualidade, também atuou na Gerência Técnica do Censo Demográfico, ambas unidades da Diretoria de Pesquisas.

Denise Britz do Nascimento Silva

Possui graduação em Estatística pela ENCE, Mestrado em Estatística pela UFRJ e Doutorado em Estatística pela University of Southampton. Foi Coordenadora Geral da ENCE de setembro de 2011 a agosto de 2014. Atualmente é Coordenadora de Graduação e pesquisadora da ENCE com trabalhos nos seguintes temas: análise de dados amostrais, estimação em pequenas áreas/domínios, modelagem estatística, análise de séries temporais de pesquisas amostrais e métodos para pesquisas quantitativas. No âmbito internacional, a autora é membro do Instituto Internacional de Estatística (International Statistical Institute - ISI), trabalhou na área de metodologia estatística do Office for National Statistics do Reino Unido e foi professora da University of Southampton. Foi eleita para a Presidência da International Association of Survey Statisticians - IASS, com mandato de quatro anos, sendo dois anos como Presidente eleita (2017 - 2019) e dois anos como Presidente (2019 - 2021).

Djalma Galvão Carneiro Pessoa

Possui graduação em Engenharia Civil pela Universidade Federal de Pernambuco - UFPE, Mestrado em Matemática pelo Departamento de Matemática do Instituto Tecnológico de Aeronáutica - ITA, Ph.D em Estatística pela University of California, Berkeley. Foi professor do Departamento de Matemática do ITA, Pesquisador Associado do IMPA, Diretor da ENCE e Diretor Geral do IBGE. Atuou durante vários anos como consultor da Coordenação de Métodos e Qualidade.

Eduardo Santiago Rosseti

Possui graduação em Matemática pela Universidade Federal de Juiz de Fora - UFJF, Mestrado em Estudos Populacionais e Pesquisas Sociais pela ENCE. Trabalhou no IBGE no período entre 2009 e 2011 e prestou consultoria para órgãos internacionais como o Fundo de População das Nações Unidas (United Nations Population Fund - UNFPA)

e a Organização Pan-Americana de Saúde - OPAS (Pan American Health Organization - PAHO). Atualmente é Técnico de Projetos Pleno na Fundação Getúlio Vargas - FGV com trabalhos nas seguintes áreas: análise de dados amostrais, modelagem estatística, projeções de população, métodos demográficos, análise de séries temporais e métodos para pesquisas quantitativas.

José André de Moura Brito

Tem bacharelado em Matemática (1997) pela UFRJ, Mestrado (1999) e Doutorado (2004), ambos em Engenharia de Sistemas e Computação (Otimização) pela UFRJ, e Pós-Doutorado em Otimização (2008) pela Universidade Federal Fluminense - UFF. Atualmente é professor da ENCE, onde leciona disciplinas na graduação e no mestrado. Também é editor associado e editor da área de Estatísticas Oficiais da *Revista Brasileira de Estatística* e membro do grupo de pesquisas Gestão da Informação através da Inteligência Computacional. Tem experiência nas áreas de Otimização, Estatística e Computação, atuando principalmente nos seguintes temas: Metaheurísticas, Programação Inteira, Otimização Combinatória, Programação Não Diferenciável, Análise de Algoritmos, Análise de Agrupamentos e Amostragem.

Marcos Paulo Soares de Freitas

Possui graduação em Estatística e Mestrado em Estudos Populacionais e Pesquisas Sociais, Área de Concentração em Pesquisas Sociais e Amostragem pela ENCE. Tecnologista em Informações Geográficas e Estatísticas da Coordenação de Métodos e Qualidade, da Diretoria de Pesquisas do IBGE, tendo atuado no planejamento amostral de diversas pesquisas e, atualmente, participa de grupos de trabalho relacionados com a avaliação da qualidade da produção estatística.

Maria Luíza Barcellos Zacharias

Bacharel em Estatística pela ENCE e Doutora em Administração pelo Instituto de Pesquisa e Pós-graduação em Administração de Empresas - Coppead da UFRJ. Servidora do IBGE desde junho de 1980, foi Gerente do Cadastro Central de Empresas - Cempre por 18 anos, tendo se dedicado a estudos sobre métodos de pesquisas e levantamentos a partir de 2009, quando passou a integrar a equipe da Coordenação de Métodos e Qualidade da Diretoria de Pesquisas. Desde 2013, atua como professora colaboradora no Mestrado e Doutorado do Programa de Pós-Graduação em População, Território e Estatísticas Públicas da ENCE, lecionando as disciplinas "Qualidade de dados em pesquisas" e "Métodos para Pesquisas e Levantamentos". Desde 2014 é Gerente de Qualidade Estatística da referida Coordenação, responsável pelo planejamento e implementação de ações voltadas para a gestão da qualidade da produção estatística do IBGE.

Maurício Teixeira Leite de Vasconcellos

Estatístico graduado pela ENCE e D.Sc. em Saúde pública pela Escola Nacional de Saúde Pública - ENSP da Fundação Oswaldo Cruz - Fiocruz. Começou sua carreira no IBGE, em 1971, trabalhando na equipe de coordenação do Estudo Nacional da Despesa Familiar - Endef e, em 1981, foi para o Departamento de Coordenação de Métodos. Foi superintendente da Superintendência de Estatísticas Industriais, Comerciais e dos Serviços e chefe do Departamento de Censo Demográfico. Participou da equipe responsável pela estruturação do Centro de Documentação e Disseminação de Informações, sendo transferido para o antigo Departamento de Contas Nacionais. Em 1993, retornou para a área de métodos da Diretoria de Pesquisas do IBGE, onde trabalhou até 2003, quando foi transferido para a ENCE. Em 2009, aposentou-se, mas continua a trabalhar no IBGE como colaborador voluntário no programa de pós-graduação da

ENCE. Durante este período, atuou, também, como consultor da Organização das Nações Unidas para a Alimentação e a Agricultura (Food and Agriculture Organization of the United Nations - FAO), em diferentes países e projetos, da Organização Mundial da Saúde - OMS (World Health Organization - WHO) e do Grupo Banco Mundial (World Bank Group). Atua na interseção entre as áreas de saúde pública e de estatística, em particular com métodos de avaliação nutricional, desenho de amostras probabilísticas, métodos de pesquisa e estratégias para avaliação de projetos sociais.

Nícia Custódio Hansen Brendolin

Possui graduação e mestrado em Estatística pela UFRJ. Servidora do IBGE desde março de 2009, tendo atuado a maior parte do tempo na Coordenação de Métodos e Qualidade, da Diretoria de Pesquisas. Atuou, também, como professora colaboradora da Graduação em Estatística da ENCE na área de modelos lineares generalizados.

Pedro Luis do Nascimento Silva

Bacharel em Estatística pela ENCE (1980), Mestre em Matemática Aplicada - Estatística pelo IMPA (1988) e Doutor em Estatística pela University of Southampton (1996). É Pesquisador Titular da ENCE. Presidiu o ISI (2015-2017). Tem larga experiência nas áreas de amostragem e métodos de pesquisa, análise de dados amostrais complexos, pesquisas amostrais domiciliares, estimação de variâncias, estimadores de calibração, crítica e imputação de dados, estimação para pequenos domínios, pesquisas amostrais na avaliação de políticas públicas, estatísticas oficiais.

Roberta Carneiro de Souza

Possui graduação em Estatística pela UFRJ e mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio. Doutoranda em Engenharia Civil pela Coppe, da UFRJ. Atuou por 13 anos como estatística no mercado segurador, sendo os últimos anos como gerente técnica e atuarial. Em 2013 lecionou estatística na UFF como professora substituta da graduação. Servidora do IBGE desde junho de 2014, na unidade de metodologia da Diretoria de Pesquisas. Atua, também, como professora-adjunta da graduação em administração da Escola Superior Nacional de Seguros - ESNS na área de estatística desde 2008.

Sâmela Batista Arantes

Possui graduação em Estatística pela Universidade Federal de Minas Gerais - UFMG e mestrado em Estudos Populacionais e Pesquisas Sociais pela ENCE. Servidora do IBGE desde março de 2012, tendo atuado na unidade de metodologia da Diretoria de Pesquisas. Também tem atuado como monitora na pós-graduação da ENCE desde 2012 na área de Estatística e Amostragem.

Solange Correa Onel

Graduada em Estatística pela UFRJ (1996), Mestrado em Estudos Populacionais e Pesquisas Sociais (sub-área Amostragem) pela ENCE (2001) e Doutorado em Estatística, University of Southampton, Reino Unido (2008). Servidora do IBGE desde fevereiro de 1997, exercendo cargos de Tecnologista de Informações Geográficas e Estatísticas e Gerente de Pesquisa e Desenvolvimento da Coordenação de Métodos e Qualidade da Diretoria de Pesquisas. Atuou, também, como Coordenadora Geral da ENCE do IBGE. Licenciada do IBGE desde setembro de 2012, atualmente exerce o cargo de gerente acadêmica do Data Science Campus, Office for National Statistics, Reino Unido. Áreas de interesse: análise de dados amostrais complexos, planejamento amostral, modelos de estimação em pequenas áreas e métodos de reamostragem aplicados a modelos hierárquicos.

Sonia Albieri

Possui graduação em Estatística pela Unicamp e Mestrado em Estatística pelo IMPA. Servidora do IBGE desde setembro de 1981, tendo atuado a maior parte do tempo na unidade de metodologia da Diretoria de Pesquisas. Desde março de 2000, atua como Coordenadora de Métodos e Qualidade.

Viviane Cirillo Carvalho Quintaes

Possui graduação em Estatística (1999) pela ENCE e Mestrado em Engenharia de Produção (2003) pela UFRJ. Atualmente é Tecnologista em informação geográfica e estatística na Coordenação de Métodos e Qualidade. Tem experiência em probabilidade e estatística, com foco em modelagem e em métodos de estimação em pequenas áreas.

Zélia Magalhães Bianchini

É graduada em Matemática pela Universidade Estadual Paulista "Júlio de Mesquita Filho" - Unesp, em São José do Rio Preto em São Paulo, e Mestre em Estatística pelo IMPA no Rio de Janeiro. Servidora do IBGE de agosto de 1978 a agosto de 2016, tendo atuado grande parte do tempo na unidade de metodologia da Diretoria de Pesquisas, que chefiou de 1992 a janeiro de 1999. Atuou como orientadora técnica na área de metodologia, especialmente, em amostragem, processo de produção de informações, confidencialidade, princípios fundamentais das estatísticas oficiais e qualidade em estatísticas oficiais. Atuou, também, como professora colaboradora da Graduação em Estatística da ENCE do IBGE, na área de amostragem. Foi Assessora e Diretora Adjunta da Diretoria de Pesquisas de 1999 até sua aposentadoria em 2016, acumulando vasta experiência como representante do IBGE em diversos fóruns nacionais e internacionais.

Equipe técnica

Diretoria de Pesquisas

Coordenação de Métodos e Qualidade

Sonia Albieri

Planejamento e organização da publicação

Sonia Albieri
Antonio José Ribeiro Dias

Elaboração de textos

Alexandre dos Reis Santos
Antonio José Ribeiro Dias
Ari do Nascimento Silva
Bruno Freitas Cortez
Débora Ferreira de Souza
Denise Britz do Nascimento Silva
Djalma Galvão Carneiro Pessoa
Eduardo Santiago Rossetti
José André de Moura Brito
Marcos Paulo Soares de Freitas
Maria Luíza Barcellos Zacharias
Mauricio Teixeira Leite de Vasconcellos
Nícia Custódio Hansen Brendolin
Pedro Luis do Nascimento Silva
Roberta Carneiro de Souza
Sâmela Batista Arantes
Solange Correa Onel
Sonia Albieri
Viviane Cirillo Carvalho Quintaes
Zélia Magalhães Bianchini

Projeto Editorial

Centro de Documentação e Disseminação de Informações

Coordenação de Produção

Marise Maria Ferreira

Gerência de Editoração

Estruturação textual
Katia Vaz Cavalcanti
Marisa Sigolo

Diagramação tabular e de gráficos
Leonardo Ferreira Martins
Solange Maria Melo de Oliveira

Diagramação textual
Maria da Graça Fernandes de Lima

Programação visual da publicação
Luiz Carlos Chagas Teixeira

Produção do *e-book*
Roberto Cavararo

Gerência de Documentação

Pesquisa e normalização documental
Ana Raquel Gomes da Silva
Juliana Chagas Moreira
Juliana da Silva Gomes
Kleiton Moura Silva (Estagiário)
Lioara Mandoju
Nadia Bernuci dos Santos
Solange de Oliveira Santos
Vera Lúcia Punzi Barcelos Capone

Elaboração de quartas capas
Ana Raquel Gomes da Silva
Juliana da Silva Gomes

Gerência de Gráfica
Ednalva Maia do Monte

Impressão e acabamento
Newton Malta de Souza Marques
Helvio Rodrigues Soares Filho

Série Documentos para Disseminação
ISSN 0103-6335

1- O IBGE e o atendimento à sociedade: (prefácio ao projeto técnico CDDI), de Nelson de Castro Senra e Lídia Vales de Souza.
ISBN 85-240-0329-4. 1990. 43 p.

2 – Projetos de disseminação: contribuição ao estabelecimento de uma metodologia, de Cláudio Alex Fagundes da Silva.
ISBN 85-240-0355-3. 1990. 29 p.

3 – Pensando a disseminação de informações: (o caso do IBGE), de Nelson de Castro Senra.
ISBN 85-240-0459-2. 1993. 39 p.

4 – Memória institucional do IBGE: em busca de um referencial teórico, de Icléia Thiesen Magalhães Costa.
ISBN 85-240-0446-0. 1992. 40 p.

Subsérie Memória Institucional

ISSN 0103-6459

1 – Teixeira de Freitas: pensamento e ação, de Mario Augusto Teixeira de Freitas.
Organizado pelo Setor de Memória Institucional.
ISBN 85-240-0351-0. 1990. 140 p.

3 – Pró-censo: algumas notas sobre os recursos para o processamento de dados nos recenseamentos do Brasil, de Francisco Romero Feitosa Freire.
ISBN 85-240-0460-6. 1993. 53 p.

4 – A criação do IBGE no contexto da centralização política do Estado Novo, de Eli Alves Penha.
ISBN 85-240-0463-0. 1993. 123 p.

5 – IBGE: um retrato histórico, de Jayci de Mattos Madeira Gonçalves.
ISBN 85-240-0542-4. 1995. 61 p.

6 – Síntese histórica da formação dos Estados, Distrito Federal e Território da República Federativa dos Estados Unidos do Brasil e divisas inter-estaduais, de Ildefonso Escobar.
ISBN 85-240-0545-9. 1995. 144 p.

7 – O pensamento de Fábio de Macedo Soares Guimarães: uma seleçãode textos.
Organizado por Nelson de Castro Senra.
ISBN 85-240-3868-3. 2006. 282 p.

8 – Isaac Kerstenetzky: legado e perfil.
Organizado por Nelson de Castro Senra.
ISBN 85-240-3900-0. 2006. 213 p.

9 – Giorgio Mortara: ampliando os horizontes da demografia brasileira.
Organizado por Nelson de Castro Senra.
ISBN 85-240-3937-9. 2007. 105 p.

10 – A estatística brasileira e o Esperanto: uma história centenária: 1907-2007.
Organizado por Nelson de Castro Senra.
ISBN 85-240-3944-7. 2007. 161 p.

11 – Bulhões Carvalho, um médico cuidando da estatística brasileira.
Organizado por Nelson de Castro Senra.
ISBN 978-85-240-3982-9. 2007. 433 p.

12 – Embaixador Macedo Soares: um príncipe da conciliação: recordando o primeiro presidente do IBGE.
Organizado por Nelson de Castro Senra.
ISBN 978-85-240-4008-5. 2008. 331 p.

13 - O IBGE na história do municipalismo e sua atuação nos municípios: o pensamento de Teixeira de Freitas e de Rafael Xavier.
Organizado por Nelson de Castro Senra.
ISBN 978-85-240-4017-7. 2008. 432 p.

14 - Lyra Madeira, um mestre da demografia brasileira.
Organizado por Nelson de Castro Senra.
ISBN 978-85-240-4032-0. 2008. 134 p.

15 - Teixeira de Freitas, Um Cardeal da Educação Brasileira: sua atualidade intelectual.
Organizado por Nelson de Castro Senra.
ISBN 978-85-240-4052-8. 2008. 266 p.

16 - Geografia e Geopolítica: a contribuição de Delgado de Carvalho e Therezinha de Castro.
Organizado por Marco Aurelio Martins Santos
ISBN 978-85-240-4084-9. 2009. 432 p.

17 - Evolução da divisão territorial do Brasil 1872-2010.
ISBN 978-85-240-4208-9. 2011. 264 p.

18 - Christovam Leite de Castro e a Geografia no Brasil.
Organizado por Leandro Malavota.
ISBN 978-85-240-4274-4. 2013. 340 p.

19 - ENDEF
Organizado por Leandro Malavota.
ISBN 978-85-240-4330-7. 2014. 340 p.

20 - PNAD
Organizado por Leandro Malavota, Luigi Bonafé e Vera Abrantes.
ISBN 978-85-240-4364-2. 2015. 202 p.

21 - Indicadores Sociais: passado, presente e futuro
Organizado por André Simões e
Antônio Carlos Alkmim.
ISBN 978-85-240-4424-3. 2017. 174p.

Subsérie Fontes de Documentação
ISSN 0103-6459

1 – A indexação do banco de metadados do IBGE, de Philippe Jean Damian, Marília de Almeida March e Vera Lucia Cortes Abrantes.
ISBN 85-240-0475-4. 1993. 25 p.

Se o assunto é **Brasil**,
procure o **IBGE**.

 /ibgecomunica  /ibgeoficial  /ibgeoficial  /ibgeoficial

www.ibge.gov.br 0800-721-8181

40 Anos da Unidade de Métodos Estatísticos do IBGE

Alguns Passos

Com o lançamento do presente volume da série Memória Institucional, o IBGE comemora os 40 anos de existência de sua unidade de metodologia estatística, cuja atuação, ao longo desse período, trouxe significativos aprimoramentos e avanços, tanto no que se refere aos métodos estatísticos e sistemas computacionais utilizados pelo Instituto para a produção, análise e disseminação, como no que diz respeito à qualidade estatística dos dados.

Para a empreitada, foram convidados diversos pesquisadores da área, que, na condição de servidores atuais ou pregressos da Coordenação de Métodos e Qualidade, e das unidades que a antecederam com outras denominações, produziram, especialmente para este livro comemorativo, uma coletânea de artigos que reportam algumas das múltiplas atividades que vêm sendo realizadas pela unidade, desde sua criação, em maio de 1977, quando integrou-se à estrutura organizacional.

Dada a diversidade de temas metodológicos afetos à produção de informações estatísticas em um instituto de pesquisas do porte do IBGE, por certo seria ambicioso tentar abarcá-los em um único volume. Este livro, portanto, não exaure todas as áreas de estudo em que esses dedicados pesquisadores estão ou estiveram envolvidos ao longo de tal jornada. Ao promover a divulgação de parte dos trabalhos realizados e dos significativos avanços ocorridos no IBGE no que se refere a métodos e qualidade estatísticos, este livro rende, antes de tudo, uma singela homenagem a todos aqueles – técnicos e pesquisadores – que colaboraram de alguma forma, nesses 40 anos, para o avanço metodológico na Instituição.

Esta publicação também está disponível no portal do IBGE.

ISBN 978-85-240-4430-4
9 788524 044304