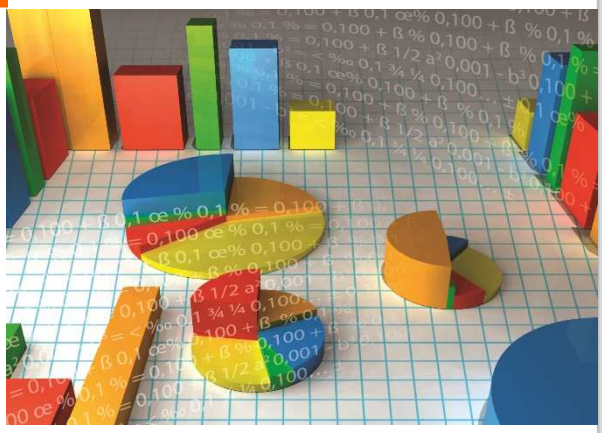
EDIÇÃO ESPECIAL – Celebrando Djalma Pessoa

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 2675-3243



Volume 79

Número 246

Janeiro/Junho 2021

IBGE
Instituto Brasileiro de Geografia e Estatística

Instituto Brasileiro de Geografia e Estatística – IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

Volume 79 - Número 246 – Janeiro/Junho 2021

ISSN 2675-3243

R. Bras. Estat., Rio de Janeiro, v.79, n.246, p. 1-162, Jan/Jun 2021

Instituto Brasileiro de Geografia e Estatística – IBGE

@IBGE. 2021

Revista Brasileira de Estatística, ISSN 2675-3243

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE
Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site https://rbes.ibge.gov.br/.

Os artigos submetidos à RBE não devem ter sido publicados em outros periódicos.

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

Informações Adicionais

Revista Brasileira de Estatística/IBGE - v.1, n.1 (jan/mar. 1940) – Rio de Janeiro: IBGE, 1940.v. trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística. Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989). Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN publicação online 2675-3243, a partir de 2019.
I. Estatística – Periódicos. II. IBGE. III. Associação Brasileira de Estatística.

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

Sumário

Nota dos Editores do Número Especial

A publicação deste Número Especial da Revista Brasileira de Estatística (RBEs) faz parte das homenagens prestadas ao Prof. Djalma Galvão Carneiro Pessoa, por suas inúmeras contribuições à Estatística no Brasil. Entre muitas outras atividades em que exerceu papel de destaque, o Prof. Djalma foi editor responsável pela RBEs num período de renovação da revista, tendo dado novos rumos e revitalizado a sua publicação.

O primeiro artigo, de Kévin Rodrigues, traz uma breve resenha sobre "A vida acadêmica e a influência do estatístico Djalma Galvão Carneiro Pessoa", destacando a trajetória percorrida e as principais contribuições do Prof. Djalma. Este é seguido por um artigo de Zélia Bianchini, onde a autora aborda as boas práticas relacionadas com a credibilidade das estatísticas oficiais, apresenta o histórico dos Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas, descreve aspectos da sua implementação, interpretação e de sua relação com os códigos de boas práticas.

Na sequência, Tereza Barbosa e Davi Alves buscam apresentar caminhos para formar estatísticos conscientes da realidade brasileira, através do uso de bases de dados oficiais e de bibliotecas do sistema R. Esse artigo tem grande afinidade com as muitas contribuições do Prof. Djalma no campo do ensino e da promoção do uso do R e da análise correta dos dados do IBGE.

O quarto artigo, de Guilherme Jacob e Anthony Damico, sobre "O futuro do pacote *convey*", revela os planos dos coautores do Prof. Djalma no desenvolvimento dessa importante ferramenta para análises estatísticas visando à estimação de indicadores de desigualdade e pobreza a partir de dados de amostras complexas. Na sequência, Leiliane Oliveira e Pedro Silva apresentam uma "Comparação de métodos de amostragem aplicáveis a pesquisas para índices de preços ao consumidor", trazendo à luz resultados inovadores sobre Amostragem Bidimensional, planejamento amostral pouco conhecido, mas com potencial para aplicação nessa área de produção das estatísticas oficiais.

Seguindo essa linha de análises comparativas, João Gabriel Malaguti, Samuel Cândido e Marcel Vieira apresentam uma "Comparação de planos amostrais probabilísticos alternativos com um plano amostral não probabilístico", num contexto em que têm crescido as demandas por aproveitar melhor dados de fontes alternativas e, dentre estas, inclusive dados de amostras não probabilísticas.

Andrea Diniz, José André Brito e Flavio Bolliger assinam junto com o Prof. Djalma o próximo artigo. Os autores insistiram nessa homenagem, reconhecendo a contribuição relevante do Prof. Djalma para o desenvolvimento da tese de doutorado da primeira autora. O artigo "Comparando o desempenho de métodos de pareamento para integrar dados sobre a agropecuária" é mais um exemplo de estudo comparativo de métodos, uma das áreas em que o Prof. Djalma fez inúmeras contribuições, particularmente no seu período como consultor do IBGE.

Encerrando o número, Kelly Gonçalves, Fernando Moura e Helio Migon apresentam “Aplicações de estimadores lineares de Bayes para previsão em população finita”, conectando essa importante área de aplicações de amostragem e produção das estatísticas oficiais com desenvolvimentos recentes da inferência Bayesiana, ampliando assim o campo de aplicações desta.

Os artigos que compõem este número buscam dar um panorama da variedade de interesses e colaborações do Prof. Djalma para o desenvolvimento da Estatística e suas aplicações. No conjunto, estes artigos fazem justa memória às contribuições do Prof. Djalma à Estatística brasileira e, em particular, ao IBGE, instituição à qual dedicou seus últimos anos de atividades profissionais.

Boa leitura.

Antonio José Ribeiro Dias
Pedro Luis do Nascimento Silva
Sonia Albieri
Zélia Magalhães Bianchini

Editores do Número Especial

Revista Brasileira de Estatística, V.79, N. 246, p.7-14, janeiro-junho 2021

A VIDA ACADÊMICA E A INFLUÊNCIA DO ESTATÍSTICO DJALMA GALVÃO CARNEIRO PESSOA

Kévin Allan Sales Rodrigues

kevin@usp.br

Universidade de São Paulo

Resumo: Neste artigo homenageamos o egrégio e insigne estatístico brasileiro, Prof. Dr. Djalma Galvão Carneiro Pessoa. No texto abordamos, a trajetória de Djalma por meio de uma breve resenha, além dos seus principais feitos à estatística brasileira. Como objetivo secundário, desejamos contribuir, humildemente, para a recuperação de uma parte da memória da sociedade estatística nacional, isto é, a História da Estatística no Brasil.
Palavras-chave: Djalma Galvão Carneiro Pessoa; Estatística no Brasil; Amostragem; IMPA; ENCE

Abstract: In this article we pay tribute to the egregious and distinguished Brazilian statistician, Prof. Dr. Djalma Galvão Carneiro Pessoa. In the text, we approach Djalma's trajectory through a brief review, in addition to his main achievements in Brazilian statistics. As a secondary objective, we wish to humbly contribute to the recovery of part of the memory of the national statistical society, that is, the History of Statistics in Brazil.
Keywords: Djalma Galvão Carneiro Pessoa; Statistics in Brazil; Sampling; IMPA; ENCE

1. INTRODUÇÃO

O objetivo deste artigo é homenagear um dos mais importantes estatísticos brasileiros, o Prof. Dr. Djalma Galvão Carneiro Pessoa que é uma referência para a comunidade universitária brasileira e, em particular, para a comunidade estatística brasileira. Assim contribuiremos para a recuperação da memória do desenvolvimento da estatística no Brasil, pois a história do Prof. Djalma é entrelaçada à história da estatística no Brasil. O estatístico que homenageamos contribuiu muito com suas orientações, publicações, aulas, palestras, livros, amizade, humildade e sabedoria à ENCE, aos seus alunos e à sociedade estatística brasileira. A ENCE — Escola Nacional de Ciências Estatísticas tem dado forte contribuição para a formação da comunidade estatística brasileira e serve de referência para os demais cursos de estatística do Brasil.

Esse artigo foi profundamente inspirado nas palavras da nota de falecimento escrita pelo Prof. Dr. Mauricio Teixeira Leite de Vasconcellos (Vasconcellos, 2020), que é presidente da SCIENCE — Sociedade para o Desenvolvimento da Pesquisa Científica, na nota de falecimento da SCIENCE; assim alguns trechos deste artigo foram retirados da nota de falecimento supracitada.

2. FORMAÇÃO ACADÊMICA

A formação acadêmica de Djalma foi bem fluida e multidisciplinar, ele iniciou o curso de engenharia civil na UFPE em 1959 e o concluiu em 1963; em 1966 concluiu o mestrado em matemática aplicada no ITA — Instituto Tecnológico de Aeronáutica com dissertação intitulada “Sistemas de Funções Independentes em Probabilidades”; posteriormente, iniciou o doutorado em estatística na University of California – Berkeley tendo o concluído em 1970 com tese intitulada “Asymptotically Minimax Fixed Length Confidence Intervals”. Aproximadamente uma década após a obtenção do título de doutor, Djalma executou seu estágio pós-doutoral na Stanford University de 1979 a 1980. Além disso, o Prof. Djalma foi um dos primeiros doutores em estatística formados no exterior. Note que durante sua formação acadêmica, incluindo o estágio pós-doutoral, Djalma passou por 4 instituições distintas; veremos a seguir que esta notável capacidade de adaptação de Djalma a diferentes instituições em diferentes localidades geográficas se estende inclusive a sua atuação profissional.

3. CONTRIBUIÇÕES ACADÊMICAS, CIENTÍFICAS E ADMINISTRATIVAS

Djalma Galvão Carneiro Pessoa participou do período de desenvolvimento e efervescência criativa de formação da comunidade estatística brasileira e, de formação do ambiente universitário nacional, por meio dos seus trabalhos acadêmicos e sua incansável dedicação à estatística nacional.

Djalma também teve atuação importante no IBGE e no IMPA e influência em outras instituições, como o ITA. No ITA, Djalma foi "auxiliar de ensino", que era o primeiro degrau da carreira docente do ITA. Na ocasião Djalma foi vinculado ao Departamento de Matemática do ITA, mas manteve esta posição por pouco tempo, pois logo em seguida foi cursar seu doutorado nos Estados Unidos e quando retornou com o título de doutor passou pouco tempo no ITA e logo seguiu sua carreira no IMPA.

Acerca da trajetória do Prof. Djalma, Vasconcellos (2020) afirmou:

> "No IBGE, (o Prof. Djalma) dirigiu a ENCE, como disse sua atual Diretora, em um dos períodos mais difíceis de sua história, soube tirá-la da crise que a ameaçava de extinção e pavimentou o caminho da renovação, a modernização da graduação em Estatística, a consolidação do papel da ENCE na capacitação de servidores do IBGE e, sobretudo, o resgate da prática da pesquisa da Escola, que propiciou a criação de cursos de pós-graduação *lato sensu* e posteriormente, de mestrado e doutorado. Dirigiu a ENCE de 28 de outubro de 1986 até a sua aposentadoria em 1995, com um breve período de afastamento, enquanto esteve a frente da então Diretoria de Planejamento e Coordenação do IBGE, entre 21 de junho de 1992 e 5 outubro de 1993. A partir de 1997, quando sua falta ficou evidente, passou a ser consultor do IBGE em atividades ligadas ao desenvolvimento e à aplicação de métodos estatísticos, ao ensino e à pesquisa. Nessa função de consultor, colaborou decisivamente na realização do Censo Demográfico 2000, tendo orientado diversos projetos relacionados com aplicação de métodos estatísticos inovadores no censo. Colaborou decisivamente na orientação profissional de diversos estatísticos jovens que ingressaram no IBGE nesse período, dando a estes a necessária segurança para

> inovar nas aplicações e avanços propostos, sem perder de vista o rigor e cuidado necessários no caso de aplicações do porte das requeridas em um censo populacional e pesquisas nacionais, nunca se afastando da ENCE e colaborando com sua pós-graduação."

Além dos importantes cargos de gestão e administração no âmbito do IBGE ocupados, em virtude do seu incontestável mérito, por Djalma, ele também teve sua parcela de influência na criação do curso de doutorado dentro do programa de pós-graduação em População, Território e Estatísticas Públicas (Vasconcellos, 2020).

Conforme Vasconcellos (2020) destaca,

> "Ele (Djalma) foi o líder do grupo que fundou a Associação Brasileira de Estatística — ABE, sendo o seu sócio número 1. Também presidiu a diretoria responsável pela instalação da ABE, entre 1984 e 1986. Todavia, muito antes disso, esteve diretamente envolvido com a criação e implantação do Simpósio Nacional de Probabilidade e Estatística — SINAPE, que até hoje é o principal simpósio da área de Probabilidade e Estatística no Brasil, tendo sido membro da comissão organizadora do 2º SINAPE."

Atualmente o SINAPE é organizado pela ABE bienalmente.

O SINAPE é um importante fórum para a difusão dos avanços da Estatística nas diversas áreas do conhecimento, além de ser um espaço de debates de políticas públicas para a Ciência e Tecnologia. Note que a semente plantada por Djalma dá frutos até hoje, pois o SINAPE já é um evento realizado há mais de 40 anos. A cada edição o SINAPE tem sido realizado em um Estado brasileiro: 20º SINAPE (em 2012) foi realizado em João Pessoa/PB, o 21º SINAPE (em 2014) em Natal/RN e o 22º SINAPE (em 2016), em Porto Alegre/RS. Além do SINAPE, a ABE organiza outros eventos de tradição e relevância nacional ligados à estatística, como a ESTE — Escola de Séries Temporais e a EMR — Escola de Modelos de Regressão. Tanto a EMR quanto a ESTE gozam do mesmo prestígio que o SINAPE em termos de tradição e relevância nacional para os pesquisadores de modelos de regressão e séries temporais, respectivamente. Essas e outras informações acerca dos eventos organizados ou apoiados pela ABE podem ser vistas diretamente na página da ABE (2021).

Em 2010, o Prof. Djalma recebeu a mais alta honraria concedida pela ABE, o Prêmio ABE. Criado em 2002, o "Prêmio ABE" homenageia durante cada edição do SINAPE, um(a) renomado(a) estatístico(a) brasileiro(a) que tenha contribuído de forma significativa para o desenvolvimento do ensino e da pesquisa em Estatística no Brasil. O(A) premiado(a) deve ter se integrado de forma ímpar à formação de capital humano especializado na área, através de orientação de dissertações e trabalhos científicos, ou que tenha sido membro atuante de iniciativas de grande impacto para os estatísticos do país, podendo ter se distinguido também pela presença constante e efetiva durante a existência da Associação Brasileira de Estatística (ABE). O "Prêmio ABE" objetiva estimular as atividades de reflexão e pesquisa em Estatística no Brasil e visa a fazer parte da história da ABE.

Vasconcellos (2020) lembra ainda que:

> "… ele (Djalma) contribuiu de forma decisiva para a implantação de vários programas de pós-graduação em Estatística no Brasil. No IMPA, foi professor do Mestrado em Estatística ali existente da segunda metade dos anos 1970 até 1986, tendo contribuído com a formação e orientação de muitos dos que depois se tornaram líderes da área de Estatística no Brasil. Ele sempre foi um entusiasta pelo saber, e desenvolveu grande interesse pelo uso de sistemas de computação estatística de código aberto. Após se aposentar, aprendeu o *software* R e passou a disseminar essa ferramenta através de cursos e do desenvolvimento de bibliotecas de funções especializadas para analisar dados de pesquisas amostrais complexas, além de colaborar e compartilhar seus conhecimentos com diversos autores de bibliotecas do R. De fato, ele foi um dos precursores do uso do R no Brasil."

A disseminação do R nos bacharelados em estatística do Brasil foi muito benéfica, tanto para o ensino e a prática da estatística em si quanto para fins teóricos, em estudos de simulação, por exemplo. Além disso, vale a pena destacar que o R é um *software* livre gratuito, isto é, todo o R e seus pacotes estão escritos em código aberto e qualquer um pode alterar estes códigos criando seus próprios pacotes e até mesmo suas próprias versões do R; ao contrário de outros *softwares* estatísticos que exigem pagamento de licenças e

não permitem que o usuário veja o código utilizado nas diversas funções estatísticas oferecidas pelo *software*.

É importante destacar o pacote *convey: Income Concentration Analysis with Complex Survey Samples* do R, que tem o Prof. Djalma como principal autor. O pacote *convey* está disponível gratuitamente em https://CRAN.R-project.org/package=convey e é mantido pelo CRAN (*The Comprehensive R Archive Network*), que é um repositório de pacotes do R mantidos pela equipe responsável pelo desenvolvimento e manutenção do R. De forma geral, este pacote estima medidas de pobreza, desigualdade e bem-estar no contexto de amostragens complexas. Amostragens complexas são amostragens que envolvem a identificação e coleta de dados de uma amostra de unidades populacionais por meio de múltiplos estágios ou fases de identificação e seleção. Essa área de amostragem complexa é uma das especialidades do Prof. Djalma.

Segundo o currículo Lattes do Prof. Djalma (Pessoa, 2017), ele concluiu a orientação de 9 mestrandos, 3 do programa de pós-graduação em Estudos Populacionais e Pesquisas Sociais e 6 do Mestrado em Matemática Aplicada do IMPA. Apesar dos cargos administrativos que ocupou, da responsabilidade de orientar mestrandos e lecionar, Djalma foi autor de cinco importantes artigos científicos: Pessoa (1973), James et al. (1992), Pessoa et al. (1997), Silva et al. (2002) e Silva et al. (2015). Além disso, foi autor de um importante livro de estatística da década de 70 intitulado “Estatística Não-Paramétrica” (Pessoa,1977). Djalma foi um pioneiro na escrita de livros de estatística em português. Este livro foi muito importante devido ao fato de livros com temáticas específicas de estatística em português serem praticamente inexistentes na década de 70, isto é, a literatura era restrita a livros introdutórios de estatística em português e a livros mais específicos em inglês. Dessa forma ele contribuiu e muito, tanto para o ensino e a pesquisa da estatística no Brasil, porque naquela época os computadores não eram financeiramente acessíveis aos estudantes de graduação como são hoje e a internet ainda estava em desenvolvimento, ou seja, não existia como existe nos moldes contemporâneos. Outros dois livros foram publicados pelo Prof. Djalma, “Análise de Dados Estruturados” (Beltrão & Pessoa,1988) e “Análise de Dados Amostrais Complexos” (Pessoa & Silva,1998); esses dois livros estão mais ligados a prática estatística desenvolvida por Djalma no IBGE e a sua principal linha de pesquisa. Em particular, o livro “Análise de Dados Amostrais Complexos” continua sendo uma importante referência até hoje em dia a quem deseja trabalhar com inferência em amostragens que envolvem múltiplos estágios de identificação e seleção das observações que comporão a

amostra. Sempre preocupado em contribuir com a literatura estatística em português, Djalma foi um dos dois tradutores do livro “Encontros com o acaso: um primeiro curso de análise de dados e inferência” (Wild & Seber, 2004). Outra contribuição importante de Djalma foi sua atuação como Editor-Responsável da Revista Brasileira de Estatística durante o período de 1987 a 1992, e posteriormente, contribuiu como Editor de Estatísticas Oficiais da revista de 1996 a 2002.

4. COMENTÁRIOS FINAIS

Este texto poderia conter mais detalhes sobre a passagem de Djalma pelo IMPA ou sobre os documentos elaborados por Djalma enquanto o mesmo atuava na função de consultor do IBGE, mas o propósito deste texto não é apresentar todos os mínimos detalhes da trajetória de Djalma e nem todas as incontáveis contribuições de Djalma à sociedade estatística brasileira (precisaríamos do espaço correspondente a uma enciclopédia inteira se este fosse o objetivo do presente texto). O propósito deste texto é relembrar uma pequena amostra aleatória de quem foi Djalma e assim permitir que os estatísticos das futuras gerações possam inferir a importância que Djalma tem e teve na estatística brasileira.

Pelas suas inúmeras contribuições em diversos aspectos à comunidade estatística brasileira, nunca esqueceremos o que o Prof. Djalma fez à ciência estatística. Esperamos que nas próximas gerações de estatísticos surjam mais “Djalmas” para que a estatística continue prosperando no Brasil da mesma forma que prosperou no passado.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ABE (2021). **ABE**: Associação Brasileira de Estatística, c2021. Página inicial.

Disponível em: <http://www.redeabe.org.br/site/>. Acesso em: 09 de jan. de 2021.

Beltrão, K.I. & Pessoa, D.G.C. (1988). *Análise de Dados Estruturados.* Rio de Janeiro: Instituto de Matemática Pura e Aplicada (IMPA).

James, B.R., James, K.L. & Pessoa, D.G.C. (1992). An Optimal C(α) Test of the Average Precipitation in Randomized Cloud-Seeding Experiments. Brazilian Journal of Probability and Statistics, 6(2), 139-151.

Pessoa, D.G.C. (1973). A Property of the Chow-Robins Procedure For Fixed Length Confidence Intervals. Boletim da Sociedade Brasileira de Matemática (Cessou em 2001. Cont. ISSN 1678-7544 Bulletin Brazilian Mathematical Society (Impresso)), 4(1), 27-30.

Pessoa, D.G.C. (1977). *Estatística Não-Paramétrica*. Rio de Janeiro: Instituto de Matemática Pura e Aplicada (IMPA).

Pessoa, D.G.C. (2017). **Currículo do sistema currículo Lattes**. [Brasília], 11 de abr. de 2017. Disponível em: <http://lattes.cnpq.br/7156389974781144>. Acesso em: 09 de jan. de 2021.

Pessoa, D.G.C. & Silva, P.L.N. (1998). *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística (ABE).

Pessoa, D.G.C., Silva, P.L.N., & Duarte, R.P.N. (1997). Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. Revista Brasileira de Estatística, 59(210), 53-76.

Silva, A.D., Freitas, M.P.S., & Pessoa, D.G.C. (2015). Assessing Coverage of the 2010 Brazilian Census. Statistical Journal of the IAOS, 31(2), 215-225.

Silva, P.L.N., Pessoa, D.G.C. & Lila, M.F. (2002). Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência & Saúde Coletiva, 7(4), 659-670.

Vasconcellos, M.T.L. (2020). **SCIENCE**: Sociedade para o Desenvolvimento da Pesquisa Científica, c2020. Nota de falecimento sobre Djalma. Disponível em: <https://science.org.br/science/noticia_djalma.php>. Acesso em: 09 de jan. de 2021.

Wild, C. J. & Seber, G. A. F. (2004). *Encontros com o acaso: um primeiro curso de análise de dados e inferência* (tradução de Cristiana Filizola Carneiro Pessoa & Djalma Galvão Carneiro Pessoa). Rio de Janeiro, Livros Técnicos e Científicos.

AGRADECIMENTOS e COLABORAÇÕES

O autor agradece ao CNPq pelo apoio financeiro. O autor também agradece pela oportunidade oferecida pela Revista Brasileira de Estatística de celebrar a memória do Prof. Dr. Djalma Galvão Carneiro Pessoa. Além disso, os comentários fornecidos pelo parecerista foram inestimáveis à versão final deste artigo.

Revista Brasileira de Estatística, V.79, N. 246, p.15-39, janeiro-junho 2021

A IMPLEMENTAÇÃO E PRÁTICAS DOS PRINCÍPIOS FUNDAMENTAIS DAS ESTATÍSTICAS OFICIAIS

Zélia Magalhães Bianchini

zelia.bianchini@gmail.com

Resumo: As estatísticas oficiais são essenciais para os sistemas de informação das sociedades democráticas, servindo aos governos e ao público dados que podem ajudar a compreender e tomar decisões: sobre a economia, a população, a sociedade e o ambiente. Um dos principais desafios para os produtores de estatísticas oficiais é comunicar com precisão as informações que produzem, otimizando o foco, sem, no entanto, violar o princípio da imparcialidade. O valor e o uso prático das estatísticas oficiais dependem da confiança dos usuários de que não haja qualquer interferência política ou interesses privilegiados. A credibilidade é baseada na reputação e confiança nos dados, sustentada por todos os produtos da instituição produtora. Portanto, as boas práticas levam à construção da imagem da instituição e à conquista da credibilidade. O objetivo deste artigo é abordar as boas práticas relacionadas com a credibilidade das estatísticas oficiais, com base na sua principal referência - os Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas. O artigo contém o histórico dos Princípios Fundamentais, aspectos da implementação, a relação com os códigos de boas práticas, a descrição, motivação, interpretação e as boas práticas de cada princípio, considerações e ações recomendadas.
Palavras-chave: estatísticas oficiais; credibilidade; imparcialidade; integridade; qualidade; confidencialidade

Abstract: Official statistics are essential to the information systems of democratic societies, serving governments and the public with data that can help understand and make decisions: about the economy, the population, the society and the environment. One of the main challenges for the producers of official statistics is to accurately communicate the information they produce, optimizing the focus, without, however, violating the principle of impartiality. The value and practical use of official statistics depend on users' confidence that there is no political interference or privileged interests. Credibility is based on the reputation and trust in data, supported by all the products of the producing institution. Therefore, good practices lead to building the institution's image and achieving credibility. The purpose of this article is to address good practices related to the credibility of official statistics, based on its main reference - the Fundamental Principles of Official Statistics of the United Nations. The article contains the history of the Fundamental Principles, aspects of implementation, the relationship with the codes of good practice, the description, motivation, interpretation and good practices of each principle, considerations and recommended actions.
Keywords: official statistics; credibility; impartiality; integrity; quality; confidentiality

1. INTRODUÇÃO

Em primeiro lugar, cabe a menção sobre o significado de “estatísticas oficiais”. Na literatura estatística, tal expressão é utilizada sem estabelecer uma definição básica. Recentemente, um conceito de estatísticas oficiais foi apresentado na versão provisória da 4ª edição do Manual de Organização Estatística, disponível para comentários em United Nations (2021). Tal conceito é baseado nos Princípios Fundamentais das Estatísticas Oficiais da Organização das Nações Unidas - ONU, os quais buscam assegurar que as estatísticas oficiais sigam critérios profissionais e científicos com padrões bem definidos.

O IBGE, no Código de Boas Práticas das Estatísticas, divulgado em 2013, adota a seguinte definição: “As Estatísticas Oficiais são informações produzidas e disseminadas por agências governamentais, em bases regulares, regidas pela legislação em matéria estatística e/ou regulamentos administrativos, sujeitas ao cumprimento de um sistema padronizado de conceitos, definições, unidades estatísticas, classificações, nomenclaturas e códigos, visando: retratar as condições econômicas, sociais e ambientais; fornecer subsídios para o planejamento, execução e acompanhamento de políticas públicas; propiciar suporte técnico para as tomadas de decisões; e consolidar o exercício da cidadania.”

A produção de estatísticas oficiais é uma cadeia complexa de operações. Começa com investigações sobre as necessidades de informação de usuários, de tal forma que possa gerar resultados que cumpram as demandas de muitos usuários e que não sejam direcionadas exclusivamente a um grupo de usuários, envolvendo a definição das prioridades e o equilíbrio dos recursos disponíveis.

Para uma visão de conjunto veja o Modelo Genérico do Processo de Produção Estatística (*Generic Statistical Business Process Model – GSBPM*), definido pela *United Nations Economic Commission for Europe – UNECE* (GSBPM, 2019). Tal modelo descreve de forma abrangente as atividades do processo de produção estatística. As etapas envolvem a especificação das necessidades, planejamento, elaboração dos instrumentos de coleta e sistemas, coleta dos dados, processamento (aí incluídos: integração/organização dos dados, classificação e codificação, crítica e tratamentos dos dados, derivação de variáveis, cálculo dos pesos ou fatores de expansão da amostra (no caso de pesquisa por amostragem), agregação dos resultados e preparação dos arquivos de dados), análise dos dados, disseminação e avaliação dos resultados. Conforme indicado neste modelo, o processo de produção engloba atividades de supervisão e controle em todas as etapas e a gestão da qualidade é transversal perpassando todas as etapas definidas no modelo.

Como os usuários não podem replicar facilmente essa complexa cadeia de operações, é essencial a confiança pública de que os resultados publicados sejam confiáveis e imparciais. Para isso, é preciso ter confiança de que os produtores atuam profissionalmente em todas as situações, mesmo quando os resultados das estatísticas oficiais se constituem em más notícias para atores da cena política. Esta atuação é caracterizada pela independência profissional, imparcialidade, transparência nos métodos adotados, decisões sobre o tempo e as formas de disseminação livres de qualquer interferência que possa enviesar, distorcer ou ocultar os resultados de alguma maneira.

A independência é crucial para a qualidade, a credibilidade e a integridade das estatísticas oficiais. As estatísticas devem ser protegidas de pressão política, servir a interesses múltiplos e ser regidas por abordagens que garantam neutralidade, objetividade, igualdade de acesso, alta qualidade e confidencialidade. Se for esse o caso, informantes, usuários, mídia, políticos e cidadãos podem confiar nas estatísticas fornecidas.

A credibilidade é baseada na reputação e confiança nos dados, sustentada por todos os produtos da instituição produtora. Portanto, as boas práticas levam à construção da imagem da instituição e à conquista da credibilidade.

As estatísticas oficiais estão operando em um ambiente competitivo e desafiador, que mudou significativamente nos últimos 20 anos e se acelerou muito nos últimos anos. Para os usuários tradicionais, o valor e a importância das estatísticas oficiais são incontestáveis. Porém, as revoluções das mídias digitais e sociais propiciam que mais pessoas tenham acesso instantâneo a várias fontes de dados para além das estatísticas oficiais. Além disso, o fenômeno das “notícias falsas” está crescendo cada vez mais, fatos que vêm exigir mais esforços e transparência na produção das estatísticas oficiais.

O objetivo deste artigo é abordar as boas práticas relacionadas com a credibilidade das estatísticas oficiais, com base na sua principal referência - os Principios Fundamentais das Estatísticas Oficiais das Nações Unidas (UNSD, 1994). O artigo está estruturado com o histórico dos Princípios Fundamentais, aspectos da implementação, a relação com os códigos de boas práticas, a descrição, motivação, interpretação e as boas práticas de cada princípio, considerações e ações recomendadas.

A justificativa para a inclusão deste artigo nesta Edição Especial intitulada “Celebrando Djalma Pessoa” está baseada no seu exemplo de profissionalismo, competência e rigor no trato da estatística e das questões metodológicas. Suas contribuições nos serviços prestados ao IBGE foram de

destacada relevância, tanto no papel de educador, compartilhando conhecimento e formando quadros técnicos, como na capacidade de colocar em prática avanços metodológicos inéditos.

A propósito, cabe mencionar a publicação nos Relatórios Técnicos da Escola Nacional de Ciências Estatísticas - ENCE - de uma tradução do Prof. Djalma Pessoa do artigo de Lars Thygesen intitulado "Comercializando estatísticas oficiais sem vender a alma" (IBGE, 1993). "A alma neste caso significa tudo que as estatísticas oficiais representam: Integridade, Imparcialidade, Credibilidade, e o direito de todo cidadão de livre acesso a uma base comum de conhecimento."

2. HISTÓRICO DOS PRINCÍPIOS FUNDAMENTAIS

A necessidade de um conjunto de princípios para a governança das estatísticas oficiais tornou-se evidente no final da década de 1980, quando, especialmente na Europa central e ex-União Soviética os sistemas políticos e econômicos tinham sido transformados e haviam surgido várias nações novas. Era essencial garantir que os sistemas estatísticos nacionais nesses países seriam capazes de produzir dados adequados e confiáveis, de acordo com certos padrões profissionais e científicos. Com esta finalidade, a Conferência dos Estatísticos Europeus desenvolveu e adotou os Princípios Fundamentais das Estatísticas Oficiais, em 1992.

Estatísticos em outras partes do mundo logo perceberam que os princípios tinham importância global muito mais ampla. Após um processo de consulta internacional, um marco na história da comunidade estatística internacional foi atingido quando a Comissão de Estatística das Nações Unidas, na sua sessão extraordinária de abril de 1994, adotou o mesmo conjunto de princípios - como os **Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas** (UNSD, 1994).

Na sua 42ª sessão, em 2011, a Comissão de Estatística discutiu os Princípios Fundamentais das Estatísticas Oficiais e entendeu que os princípios ainda eram tão adequados quanto no passado e que não havia necessidade de revisá-los. A Comissão recomendou, no entanto, que um grupo de Amigos do Presidente (*Friends of the Chair*) revisasse e atualizasse o preâmbulo dos Princípios Fundamentais, a fim de considerar os desenvolvimentos posteriores ao momento em que os princípios foram formulados, bem como desenvolvesse um guia prático para a implementação dos Princípios Fundamentais, incluindo recomendações e procedimentos para garantir independência dos sistemas estatísticos nacionais.

Em 2013, a Comissão de Estatística adotou o preâmbulo revisado e os Princípios Fundamentais foram endossados pelo Conselho Econômico e Social das Nações Unidas - ECOSOC: Resolução 2013/21 (julho/2013).

Outro marco na história para a comunidade estatística internacional foi alcançado, no mais alto nível político, quando, em 29 de janeiro de 2014, a Assembleia Geral das Nações Unidas adotou uma resolução sobre os Princípios Fundamentais (Resolução 68/261). Ver United Nations (2014).

O 20º aniversário da adoção dos Princípios Fundamentais, comemorado no início de 2014, foi usado como uma ocasião para promover ainda mais sua implementação mundial. A Comissão acolheu o primeiro esboço das diretrizes de implementação e convidou os países a enriquecerem essas diretrizes com mais comentários e apresentações de boas práticas experimentadas por eles. No processo de consulta que se seguiu, quase 40 países contribuíram e as diretrizes de implementação finais foram apresentadas à Comissão de Estatística na sua 46ª sessão (UNSD, 2015), em março de 2015.

A parte I das diretrizes de implementação compreende para cada princípio: descrição e objetivo, escopo da aplicação e os riscos no caso de ocorrer violação do princípio. A parte II trata de como garantir independência ao Sistema Estatístico Nacional.

Na reunião da Comissão de Estatística de 2020, foi aprovada a inclusão de dois capítulos nas diretrizes de implementação dos princípios (UNSD, 2020). O primeiro capítulo foi incluído para avaliar os níveis de cumprimento dos princípios, identificar áreas para desenvolvimento, fornecer caminhos para a melhoria contínua e orientar para padrões aos quais os Institutos Nacionais de Estatística - INEs podem aspirar. O segundo capítulo suplementar foi incluído para tratar da aplicação dos princípios quando do uso de fontes alternativas de dados: não oficiais e não tradicionais.

A próxima avaliação dos Princípios Fundamentais está prevista para 2024, por ocasião do 30º aniversário da sua adoção.

3. OS PRINCÍPIOS FUNDAMENTAIS E OS CÓDIGOS DE BOAS PRÁTICAS

Os Princípios Fundamentais foram estruturados como um instrumento universal para todas as culturas, sistemas políticos e épocas e constituem um pilar do Sistema Estatístico Global. Eles foram formulados de forma flexível, de tal forma que os países decidem o caminho mais conveniente de implementação para cada um deles. O caminho escolhido deve cobrir todo o sistema de estatísticas oficiais e não só áreas selecionadas das atividades do Instituto

Nacional de Estatística. A maioria dos princípios é direcionada para o arcabouço institucional e a integridade dos produtos e seu *staff*.

A adoção e o cumprimento dos Princípios Fundamentais trazem benefícios significativos para produzir estatísticas oficiais de alta qualidade, confiáveis e imparciais.

A Divisão de Estatística das Nações Unidas (*United Nations Statistics Division – UNSD*) disponibiliza vários materiais no site http://unstats.un.org/unsd/methods/statorg/default.htm, que abrangem práticas das estatísticas oficiais. Parte deste site é dedicada ao texto dos Princípios Fundamentas das Estatísticas Oficiais e aos elementos definidores e essenciais por trás de cada princípio.

O referido site da UNSD inclui o "Banco de Dados de Boas Práticas", criado em 2000, que fornece material de referência sobre os Princípios Fundamentais. Contém, em particular, exemplos de políticas e práticas em vários países para a implementação dos vários elementos dos princípios, incluindo links relevantes para sites de agências estatísticas.

Outra referência importante, no site da UNSD, é o "Manual de Organização Estatística", que trata dos fundamentos dos sistemas nacionais de estatísticas oficiais: princípios gerais, coleta de dados e políticas dos informantes, princípios de organização e gestão e diretrizes de disseminação. Uma versão provisória da 4ª edição, revisada a partir da edição do Manual de 2003, encontra-se disponível para comentários em United Nations (2021).

Com base nos Princípios Fundamentais foram desenvolvidos códigos de boas práticas, que surgem como instrumentos para colocar em prática os Princípios Fundamentais para o desenvolvimento da atividade estatística.

Em 2001, o Serviço de Estatística da União Europeia (*Statistical Office of the European Communities – EUROSTAT*) emitiu a Declaração de Qualidade do Sistema Estatístico Europeu que foi a base para o desenvolvimento do Código de Boas Práticas das Estatísticas Europeias, adotado em 2005 e que teve a última versão revisada em 2017 (EUROSTAT, 2018).

O Código Regional de Boas Práticas das Estatísticas para a América Latina e o Caribe foi elaborado no contexto do Memorando de Entendimento entre a União Europeia e a Comissão Econômica para a América Latina e o Caribe – CEPAL. O Código foi aprovado na Conferência de Estatística das Américas - CEA-CEPAL, em 2011 (CEPAL, 2011).

O Código de Boas Práticas das Estatísticas do IBGE tomou por base o Código Regional de Boas Práticas das Estatísticas para a América Latina e o

Caribe e foi divulgado em 2013. Trata-se de um instrumento orientador e regulador, estruturado por princípios e indicadores de boas práticas. Os indicadores devem ser entendidos como orientações e diretrizes a serem observadas e seguidas e não como instrumento de mensuração que quantifica um resultado. Tem como finalidade nortear o aperfeiçoamento contínuo das atividades de produção das estatísticas, assegurando o fortalecimento institucional. Trata-se de um marco na formalização do compromisso com a qualidade e a transparência da informação estatística no âmbito da Instituição. Para mais informações, ver IBGE (2013).

Os dez Princípios Fundamentais recomendados pelas Nações Unidas estão contemplados no Código de Boas Práticas das Estatísticas do IBGE.

4. MOTIVAÇÃO, INTERPRETAÇÃO E PRÁTICAS DOS PRINCÍPIOS FUNDAMENTAIS

Esta seção está estruturada com subseções para apresentação da descrição, motivação, interpretação e práticas de cada um dos dez Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas. No Apêndice 1 são apresentados o preâmbulo, revisado e adotado em 2013, e a lista com a descrição de cada um dos Princípios Fundamentais.

4.1 PRINCÍPIO 1 - RELEVÂNCIA, IMPARCIALIDADE E IGUALDADE DE ACESSO

“Princípio 1 - As estatísticas oficiais constituem um elemento indispensável no sistema de informação de uma sociedade democrática, oferecendo ao governo, à economia e ao público dados sobre a situação econômica, demográfica, social e ambiental. Com esta finalidade, os órgãos oficiais de estatística devem produzir e divulgar, de forma imparcial, estatísticas de utilidade prática comprovada, para honrar o direito do cidadão à informação pública.”

A produção de estatísticas oficiais é relevante e legítima se o grau em que atendem às necessidades atuais e potenciais de vários grupos de usuários alcançar os seguintes objetivos:

- Contribuir para informar o público sobre o desempenho dos governos quanto às responsabilidades assumidas (informação básica é direito do cidadão);

- Fornecer informações aos governos e outros órgãos estatais (Banco Central, Congresso etc.) para a tomada de decisão e a orientação de suas ações no campo das políticas públicas (econômicas, sociais, territoriais e ambientais);

- Prover as informações para o acompanhamento de compromissos internacionais assumidos pelo País (por exemplo: educação, saúde, Objetivos de Desenvolvimento Sustentável - ODS);

- Fornecer informações para a comunidade acadêmica realizar estudos e propiciar avanços científicos, tecnológicos e culturais.

Portanto, as informações são relevantes se atendem às necessidades dos usuários. Nesse sentido, há necessidade de transformar as demandas dos usuários em conceitos mensuráveis para coleta e disseminação de dados. Isso requer a decisão sobre o que produzir, qual a periodicidade e qual a desagregação das informações.

Nas diretrizes de implementação dos Princípios Fundamentais destaca-se o papel da consulta aos usuários, para promover a confiança e maximizar o valor público das informações, através do contato, o envolvimento de forma eficaz e regular (por exemplo: conselhos consultivos, comitês interinstitucionais, setoriais, fóruns, grupos de trabalho e reuniões) e a promoção do intercâmbio com as universidades. A importância do *feedback* e massa crítica de analistas experientes também reforçam o papel da relevância.

Para que as estatísticas oficiais sejam benéficas e relevantes para a sociedade, para o debate sobre políticas e a tomada de decisões, elas devem ser conhecidas, compreendidas, comunicadas e bem utilizadas. Nesse sentido, destaca-se a necessidade de facilitar a interpretação correta e o uso adequado dos dados.

A comunicação vem crescendo em importância e há uma percepção do valor de uma comunicação eficaz para manter e aumentar a relevância das estatísticas oficiais na sociedade de hoje. As agências de estatística devem se empenhar em comunicar de maneira cada vez mais eficaz para ajudar os formuladores de políticas, a mídia e o público em geral a compreender e fazer pleno uso das informações em todos os níveis de tomada de decisão. A estratégia de comunicação deve estar alinhada com os valores e princípios que sustentam as organizações estatísticas e que são definidos pelos Princípios Fundamentais (UNECE, 2019; UNSD, 2018).

Manter a relevância das informações requer revisão da produção e a realização periódica de encontros com produtores e usuários.

No que diz respeito às condições para a confiança nas estatísticas oficiais, destacam-se a independência profissional, a competência científica do pessoal e o reconhecimento de que as estatísticas são produzidas e disseminadas com imparcialidade.

A imparcialidade é um direito de todos (sociedade e governo) à informação pública de qualidade e de utilidade, com garantia de igualdade de acesso e sem nenhuma interferência no retrato produzido ou na interpretação, que possa influenciar na isenção com que o dado deve ser apresentado.

A igualdade de acesso às estatísticas oficiais é um importante pilar para imparcialidade e independência, reconhecida como boa prática para aderir aos Princípios Fundamentais.

A prática de precedência de estatísticas oficiais, para o governo e/ou para a imprensa (embargo), varia entre e dentro dos países. Há um grande debate sobre essa questão com argumentos a favor e contra a precedência. Ver, por exemplo, os artigos de Georgiou (2020); Alleva & Gandolfo (2020); Bruun & Mikkelä (2020); e Trewin (2020) publicados em um número especial sobre o tema pelo *Statistical Journal of the International Association of Official Statistics – IAOS,* em 2020.

A precedência das informações para o governo é um legado de tempos quando as estatísticas oficiais eram as estatísticas "do governo“, que não é mais o caso. As estatísticas oficiais são um bem público e pertencem a todos os usuários. Para que a integridade das estatísticas oficiais seja mantida, qualquer arranjo de precedência para o governo ou embargo com a imprensa deve ser limitado, bem justificado, controlado, publicado e administrado com muito cuidado, com as responsabilidades claras e perda de privilégio e ações se houver violações.

A questão está em assegurar, efetivamente, que o órgão produtor de estatística seja imparcial. Ainda que a apreciação dos usuários com relação aos resultados e suas opiniões sobre as possíveis implicações sejam diferentes entre si, todos devem reconhecer os resultados como retrato da realidade. Assim sendo, o papel da análise de dados na divulgação das estatísticas oficiais do país é viabilizar a melhor compreensão possível das informações, sem implicações valorativas. A credibilidade e o entendimento dos dados que estão sendo apresentados são os aspectos prioritários da divulgação.

Como destacado em IBGE (2014), a análise de dados é o processo pelo qual se dá ordem, estrutura, interpretação e significado aos dados. A análise permite diferentes abordagens e requer extremo cuidado na interpretação, de forma que os comentários analíticos devem mostrar a significância, importância e relevância das informações, de uma maneira fácil de entender, concisa, simples e interessante, observando as seguintes características:

- Objetividade - evitar incluir na análise elementos subjetivos, sujeitos a visões políticas ou teóricas distintas. Qualquer que seja o resultado, ele deve ser apresentado de forma a retratar fielmente o levantamento do dado, independentemente das preferências individuais ou impactos para alguns usuários (ser "boa" ou "má" notícia);

- Neutralidade – na interpretação dos dados, as declarações devem ser feitas sem refletir conflitos de interesse, evitando que possam ser interpretadas como comentário político;

- Não deduzir determinados resultados sem comprovação direta dos dados - limitar a análise aos dados levantados e ao período de referência das informações que estão sendo divulgadas e aos valores da série histórica. Não especular sobre resultados futuros dos dados que estão sendo divulgados ou de fatores que irão influenciá-los.

Cabe registrar as publicações da UNECE que fornecem guias sobre estratégias para apresentar, passar a mensagem e comunicar as estatísticas de forma eficaz (UNECE, 2009a, b; UNECE, 2011; UNECE, 2014).

A imparcialidade na produção e disseminação de estatísticas oficiais pode ser reconhecida através de algumas práticas descritas nas seguintes ações:

- Dar acesso à mesma informação e ao mesmo tempo para todos os usuários (igualdade de acesso e simultaneidade na disseminação).
- Garantir o uso público das informações produzidas, independentemente de qual tenha sido a fonte de financiamento.
- Implantar e executar uma política de disseminação sistemática. Todos os resultados liberados têm de ser acessíveis ao público. Não há resultados caracterizados como oficiais que sejam para uso exclusivo do governo ou de qualquer outro agente.
- Estabelecer calendários de divulgação que indicam antecipadamente a divulgação das diversas estatísticas.
- Explicitar as razões para qualquer alteração na data de lançamento anunciada nos calendários de divulgação, que só deve ocorrer a partir de critérios técnicos.
- Ser absolutamente livre de interferências externas, sobre o quê e como divulgar.
- Divulgar os resultados considerando os critérios de natureza estritamente estatística, ou seja, acompanhados de documentação sobre as práticas estatísticas, tais como: fontes de informação, conceitos, classificações,

métodos e processos adotados para permitir a correta interpretação dos dados pelos usuários.

- Comunicar amplamente aos usuários, antecipadamente, sobre mudanças metodológicas ou de conteúdo ou de formato das divulgações.

4.2 PRINCÍPIO 2 - PADRÕES PROFISSIONAIS E ÉTICA

"Princípio 2 - Para manter a confiança nas estatísticas oficiais, os órgãos de estatística devem tomar decisões de acordo com considerações estritamente profissionais, aí incluídos os princípios científicos e a ética profissional, para a escolha dos métodos e procedimentos de coleta, processamento, armazenamento e divulgação dos dados estatísticos."

O Princípio 2 incorpora os conceitos de independência técnica e independência profissional e amplia ainda mais o elemento de imparcialidade do Princípio 1.

Trata-se de "como" as atividades de estatística devem centrar-se exclusivamente na obtenção de uma representação mais confiável do fenômeno do mundo real com os recursos disponíveis. Os métodos estatísticos, padrões internacionais e boas práticas são os melhores guias para a independência técnica.

A independência técnica precisa ser traduzida em salvaguardas institucionais. Essas salvaguardas têm de proteger a Instituição, a direção e os técnicos de pressões políticas e outras relativas a decisões sobre o "como", especialmente durante o processo de disseminação das informações.

Os usuários devem ser consultados, mas as decisões devem ser tomadas pelo órgão produtor de estatística.

Qualquer falha em manter os padrões profissionais e a ética pode comprometer a confiança do público e a capacidade das estatísticas oficiais de fornecer informações confiáveis. Ameaças aos padrões profissionais, aos princípios científicos e à ética apresentam um sério desafio à confiança com a qual os produtos estatísticos são mantidos. Mesmo pequenas ofensas ou acusações sem fundamento podem levar a danos à imagem e à reputação da instituição a longo prazo, com sérias implicações e difícil reversão. As agências estatísticas são desafiadas a salvaguardar sua independência e objetividade, sendo transparentes em sua metodologia e se comunicando abertamente com o público.

O uso da rede internacional de sociedades nacionais de estatística e do *International Statistical Institute - ISI* pode ser útil para a organização de eventos para apoiar os Institutos Nacionais de Estatística quando eles têm que enfrentar

uma situação difícil. Laços fortes com os usuários e as mídias geralmente são bons aliados e a existência de código de ética profissional dos servidores também são elementos importantes para garantir a independência técnica e a integridade das estatísticas oficiais.

Entidades de várias esferas (conselhos ou comitês internacionais, regionais ou nacionais, conselhos estatísticos ou associações profissionais, grupos consultivos, especialistas e consultores) fornecem apoio, assessoria e revisão especializada; compartilham avanços recentes em práticas científicas, metodologia e conceitos; monitoram as práticas e conduta ética profissional e relatam violações dos princípios éticos.

4.3 PRINCÍPIO 3 - RESPONSABILIDADE E TRANSPARÊNCIA

"Princípio 3 - Para facilitar uma interpretação correta dos dados, os órgãos de estatística devem apresentar informações de acordo com normas científicas sobre fontes, métodos e procedimentos estatísticos."

O Princípio 3 trata da responsabilidade e transparência e é considerado como corolário da independência técnica. Neste sentido, os produtores de estatísticas têm que ser: totalmente transparentes, para todos os usuários, sobre os métodos adotados (metadados e metodologias publicamente acessíveis); plenamente responsáveis perante o público pelas decisões tomadas; totalmente responsáveis pelos resultados divulgados; e totalmente transparentes sobre a qualidade dos resultados que publica.

Para ser capaz de avaliar a qualidade e os limites das interpretações, os produtores têm que estar engajados regularmente em atividades de análise e ter o *know-how* necessário sobre técnicas analíticas.

A implementação institucional do Princípio 3 envolve a transparência: nos padrões e métodos utilizados; nos conceitos, definições e classificações utilizados; em quaisquer mudanças importantes na metodologia; e nas regras e diretrizes sobre o acesso às informações.

Os Princípios 2 e 3 chamam a atenção para os fundamentos das estatísticas oficiais em métodos e procedimentos de coleta e apuração de dados, bem como a importância de fornecer informações aos usuários com práticas de documentação metodológica facilmente acessível, tais como: metadados, manuais sobre conceitos e definições, coleta, processamento, análise e apresentação de dados; e regras e diretrizes sobre revisões e política de erros.

Em questões de confiança nas estatísticas, os Princípios 2 e 3 ressaltam a natureza crítica dos esforços da Instituição para garantir a integridade e demonstrar a qualidade das estatísticas.

Cabe registrar o manual das Nações Unidas, intitulado *United Nations National Quality Assurance Frameworks Manual for Official Statistics – UN NQAF Manual* (United Nations, 2019), como uma importante referência internacional para orientar os países na implementação de uma estrutura nacional de garantia de qualidade, inclusive para novas fontes de dados, novos provedores de dados e dados e estatísticas sobre os indicadores ODS. O referido manual foi aprovado pela Comissão de Estatística em 2019, substituindo o modelo genérico e as diretrizes de 2012.

4.4 PRINCÍPIO 4 - PREVENÇÃO DO MAU USO DOS DADOS

"Princípio 4 - Os órgãos de estatística têm direito de comentar interpretações errôneas e utilização indevida das estatísticas."

O risco de que as estatísticas sejam mal utilizadas pode resultar em decisões equivocadas e causar grandes danos à credibilidade da Instituição, impactando negativamente na confiança dos usuários. O mau uso pode ser o resultado de um planejamento, implementação e comunicação inadequados ou uma tentativa deliberada de enganar.

As fontes de uso indevido são as seguintes: dado errado, impreciso ou incompleto; domínio dos novos canais de comunicação, com opiniões formadas sobre crenças em vez de fatos precisos; manipulação / falsificação; e a falta de literacia estatística e erros de interpretação.

A prevenção do mau uso dos dados deve levar em conta as seguintes práticas: certificar sobre a qualidade dos resultados antes do lançamento; documentar com clareza para permitir a correta interpretação dos dados pelos usuários e menção à adoção das recomendações internacionais; anunciar alterações na metodologia com antecedência da divulgação através de mensagens na Internet; e capacitar os usuários e a mídia quando há nova metodologia ou mudança metodológica.

Outra prática recomendada é definir procedimentos nos casos em que seja detectado um mau uso, uma interpretação equivocada ou mesmo uma reação exagerada, por parte da mídia, da sociedade ou do governo, em relação às estatísticas oficiais divulgadas. Ver, por exemplo, em IBGE (2016), o guia com os procedimentos para lidar com o mau uso dos dados e informações estatísticas e geoespaciais do IBGE. Há um outro guia elaborado (IBGE, 2015), com o objetivo de definir procedimentos de forma a assegurar que todas as ocorrências de erros de divulgação de informações estatísticas detectadas sejam tratadas uniformemente e de forma transparente.

Além disso, a desinformação e uso indevido de estatísticas podem ser evitados diante das seguintes ações: valorizar e melhorar a disseminação das estatísticas existentes; oferecer novos produtos direcionados para diferentes usuários; expandir ainda mais a cultura de mídia social, assegurando a presença onde as discussões acontecem e as opiniões são formadas; comunicar se as estatísticas oficiais estão sendo mal utilizadas e corrigir fatos imprecisos; trabalhar com verificadores de fatos; construir parcerias e colaborações; contribuir para a literacia estatística; e monitorar a confiança e responder às lacunas de credibilidade identificadas.

O órgão de estatística deve ter o direito de comentar sobre o uso incorreto ou interpretação incorreta das estatísticas oficiais em público. É um direito reagir e não uma obrigação. Comentar de forma eficaz quando, por exemplo: uma mídia importante, em seus comentários, não observa limites explícitos de interpretações; a tomada de decisão é baseada em estatísticas inapropriadas ou duvidosas ou quando os limites de uso anunciados são desconsiderados; e em casos de abuso.

É importante reconhecer que a Instituição pode correr riscos ao entrar em um debate sobre o uso indevido de suas estatísticas, então as respostas precisam ser cuidadosamente medidas. Dependendo da resposta, pode ser considerado defensivo ou não imparcial, por isso a hesitação em se envolver no debate. Muitos casos são comentados por outros, antes mesmo que a Instituição tome a iniciativa.

4.5 PRINCÍPIO 5 - EFICIÊNCIA

“Princípio 5 - Os dados utilizados para fins estatísticos podem ser obtidos a partir de diversos tipos de fontes, sejam pesquisas estatísticas ou registros administrativos. Os órgãos de estatística devem escolher as fontes levando em consideração a qualidade, oportunidade, custos e ônus para os informantes.”

Os tipos de levantamentos estatísticos, sejam censitários, por amostragem ou sondagem têm a finalidade exclusiva de produção de estatísticas, enquanto nos registros administrativos, coletados principalmente para fins administrativos, as informações são obtidas de instituições públicas com finalidade de gestão, controle e geração de base de dados.

A produtividade na geração de estatísticas depende de fatores como modelos de organização e gestão, pessoal suficiente e qualificado, tecnologia de acordo com os padrões de produção definidos e os orçamentos adequados para as metas de produção.

A carga de resposta para o informante tem de ser avaliada previamente e levar em consideração que: todas as pesquisas têm de ser testadas (provas-pilotos); os informantes devem ser informados sobre o fundamento legal, finalidade da pesquisa e sobre as medidas de proteção da confidencialidade; o questionário deve ser adequado com perguntas compreensíveis e não-intrusivas; e as taxas de resposta devem ser cuidadosamente monitoradas.

Deve-se fazer um esforço contínuo para utilizar ou desenvolver técnicas que reduzam o volume de informações solicitadas aos informantes e otimizar os processos. Neste sentido, são necessários esforços para melhorar o potencial estatístico dos registros administrativos e reduzir os custos com pesquisas. O uso direto dos registros administrativos pode não ser possível, se o conteúdo não for consistente com os requisitos estatísticos. O ideal é desenvolver os registros administrativos, desde o início, com base em requisitos estatísticos.

O processo de gestão eficiente dentro ou entre instituições requer mecanismos de comunicação; monitoramento e controle; e procedimentos de articulação.

Um outro aspecto a ser mencionado é a correspondência ou ligação de dados como uma maneira eficiente de usar as estruturas existentes para novas estatísticas.

As novas fontes de dados, chamadas de *Big Data*, abrem novas oportunidades para estatísticas oficiais. Hackl (2016) oferece uma breve visão geral de projetos e pilotos de *Big Data* conduzidos em estatísticas oficiais em nível nacional e internacional. As experiências destes projetos indicam que não existe uma abordagem metodológica uniforme para a utilização dos novos dados nos vários domínios estatísticos.

A utilização de *Big Data* nas estatísticas oficiais envolve vários desafios, sendo necessárias soluções para questões metodológicas, tecnológicas e padrões. Não se pode esperar que *Big Data* substitua, a curto prazo, as fontes de dados das pesquisas, mas deve ser avaliada sua utilidade na produção de certas estatísticas oficiais.

4.6 PRINCÍPIO 6 – CONFIDENCIALIDADE

“Princípio 6 - Os dados individuais coletados pelos órgãos de estatística para elaboração de estatísticas, sejam referentes a pessoas físicas ou jurídicas, devem ser estritamente confidenciais e utilizados exclusivamente para fins estatísticos.”

O sigilo estatístico visa proteger a privacidade das unidades individuais cujas informações são coletadas e processadas. São abordados dois enfoques:

o primeiro, sobre o uso de dados protegidos de unidades individuais, por produtores de estatísticas oficiais, apenas para fins estatísticos; e o segundo, a não revelação pelos produtores de estatísticas oficiais, direta ou indiretamente, das características sobre as unidades protegidas a terceiros, de tal forma que permita derivar informações adicionais sobre uma unidade protegida.

Proteger a privacidade, impedindo a revelação da identidade, é um direito dos cidadãos e entidades pesquisadas.

Garantir a confiança e cooperação dos informantes para obter as informações é crucial para a produção de estatísticas oficiais, que depende da relação de troca entre a obrigatoriedade de informar *versus* a garantia do sigilo das informações individualizadas. A agência de estatística deve ser percebida como responsável por manter em sigilo todos os registros individuais e essa garantia deve ser dada para que os informantes sintam confiança e tenham atitude cooperativa, fornecendo a matéria prima básica para a produção de estatísticas.

As agências de estatística são desafiadas a investir em mecanismos de proteção da confidencialidade em cada etapa do processo estatístico de produção: durante a coleta e o processamento de dados; para publicação de dados agregados, através do uso de técnicas para limitação da revelação da identidade; ao liberar microdados para uso público ou para fins de pesquisa. Em IBGE (2018) encontram-se documentados os procedimentos adotados para garantir o cumprimento do princípio da confidencialidade, bem como da legislação nacional sobre o assunto.

Outro aspecto diz respeito à proteção dos dados em termos de segurança física das instalações, controle de acesso, defesa contra hackers, transmissão dos dados com criptografia, remoção de atributos de identificação e proteção contra adulteração dos arquivos de dados.

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 de informações implica o risco de revelação de dados identificáveis e confidenciais. Neste contexto, existe um dilema sobre até onde vão os benefícios e riscos de disponibilizar informações e o desafio de gerenciar o conflito entre o direito à privacidade e o direito de acesso à informação.

Cabe registrar também a atenção a ser dada ao compromisso de proteção do sigilo das informações individualizadas *versus* abordagens de

acesso e uso de registros administrativos, ligação de bases de dados e o georreferenciamento das informações.

4.7 PRINCÍPIO 7 – LEGISLAÇÃO

"Princípio 7 - As leis, regulamentos e medidas que regem a operação dos sistemas estatísticos devem ser tornadas de conhecimento público."

Os produtores de estatísticas oficiais devem ser regidos por uma legislação que especifique o seu objetivo e que possivelmente incentive a qualidade da produção de estatística. É essencial que leis e regulamentos cubram todas as atividades e responsabilidades dos envolvidos no Sistema Estatístico Nacional.

Não só a lei estatística em si, mas também o programa estatístico e todo o conjunto de normas (portarias, regimentos etc.), relatórios anuais, relatórios de avaliação e auditorias de atividades estatísticas devem ser públicos.

Alinhando-se com os Princípios Fundamentais, foi desenvolvido por uma força tarefa da UNECE um guia que orienta os países no processo de rever e revisar a legislação estatística necessária para apoiar a modernização dos sistemas estatísticos e para a valorização das estatísticas oficiais (UNECE, 2018).

4.8 PRINCÍPIO 8 - COORDENAÇÃO NACIONAL

"Princípio 8 - A coordenação entre os órgãos de estatística de um país é indispensável, para que se obtenha coerência e eficiência no sistema estatístico."

A tarefa de coordenação, atribuída ao Instituto Nacional de Estatística de cada país, deve envolver todos os produtores de estatísticas oficiais e incluir: a definição de normas, como por exemplo as classificações; a geração de cadastros de base estatística para seleção de amostras; apoio e aconselhamento em metodologia e boas práticas; a garantia de que as exigências das estatísticas oficiais, especialmente as boas práticas relacionadas com os Princípios Fundamentais, sejam implementadas; a criação de um programa estatístico abrangente e livre de duplicação; a decisão sobre o que é o resultado oficial, se houver resultados divergentes compilados a partir de fontes diferentes; e a organização de reuniões periódicas com os outros produtores.

A padronização das definições, conceitos e classificações é uma ferramenta importante de coordenação estatística para permitir a articulação, comparações no tempo e a ligação de dados de fontes diferentes.

4.9 PRINCÍPIO 9 - USO DE PADRÕES INTERNACIONAIS

"Princípio 9 - 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."

A colaboração entre diferentes agências é facilitada pelo uso de padrões, que fornecem uma estrutura comum de entendimento e um compartilhamento de vocabulário, que formam a base da comunicação, principalmente entre diferentes línguas.

A participação ativa na criação e revisão das normas, o *feedback* sobre as necessidades de informações nacionais e questões de implementação são fundamentais para manter a relevância dos padrões internacionais.

A utilização de padrões internacionais em nível nacional é fundamental para melhorar a comparabilidade internacional, fomentar a coerência e eficiência dos sistemas estatísticos, aumentar a imparcialidade das decisões sobre o "como", especialmente quando há controvérsia.

As recomendações internacionais fornecem orientações e assistência aos países e não são impostas. Cada país implementa de acordo com as suas condições e fontes de dados existentes. Seguem alguns exemplos de disponibilidade de recomendações internacionais temáticas: Sistema de Contas Nacionais (Nações Unidas); Sistema de Contas Econômicas Ambientais (Nações Unidas); Classificações (Nações Unidas); Censos de População e Habitação (Nações Unidas); Estatísticas do Trabalho (Organização Internacional do Trabalho - OIT); Estatísticas de Educação (Organização das Nações Unidas para a Educação, a Ciência e a Cultura - UNESCO).

Cabe destacar os padrões internacionais, definidos pela UNECE, para apoiar uma maior colaboração entre agências estatísticas: *Generic Statistical Business Process Model - GSBPM* – modelo para a documentação de processos, para a harmonização de métodos e ferramentas de informática e para avaliação e melhoria da qualidade; e o *Generic Statistical Information Model - GSIM* - modelo que fornece uma linguagem comum para descrever definições, atributos e relacionamentos internacionalmente aceitos, que são usadas na produção de estatísticas oficiais.

4.10 PRINCÍPIO 10 - COOPERAÇÃO INTERNACIONAL

"Princípio 10 - A cooperação bilateral e multilateral na esfera da estatística contribui para melhorar as estatísticas oficiais em todos os países."

Dada a velocidade das mudanças na sociedade, a evolução precisa ser intensificada e um grande facilitador é o intercâmbio de experiências entre institutos de estatística e agências internacionais. A experiência internacional dos muitos avanços proporciona uma grande oportunidade para acompanhamento e implementação das melhores práticas.

Diante de problemas com questão da aplicação dos Princípios Fundamentais, tais como interferência, uma ajuda da comunidade internacional das estatísticas oficiais deve ser considerada.

A troca contínua de conhecimentos é pré-requisito essencial para garantir a eficácia em matéria de Cooperação Estatística, para ambos os lados. Assim, são recomendadas visitas oficiais de gestores de instituições estatísticas, reuniões para troca de conhecimento, acolhimento de estagiários de outros países e participação em programas de treinamento internacional para intercâmbio de melhores práticas.

5. CONSIDERAÇÕES E AÇÕES RECOMENDADAS

Os benefícios da aplicação dos Princípios Fundamentais podem ser resumidos em: cumprimento adequado do papel de informar a sociedade acerca dos fatos sociais, econômicos e ambientais relevantes; reconhecimento e fortalecimento da credibilidade das estatísticas oficiais; proteção para as estatísticas oficiais e os profissionais que as produzem; conscientização sobre a importância da missão institucional associada com a produção de estatísticas oficiais; garantia de qualidade e controle de qualidade das estatísticas oficiais.

Para serem eficazes, os Princípios Fundamentais precisam ser respeitados por todas as partes interessadas e em todos níveis políticos, sendo, para isso, muito útil o desenvolvimento de códigos para colocar em prática os Princípios Fundamentais.

A seguir, são recomendadas algumas ações que promovem os Princípios Fundamentais.

- Conscientizar as equipes envolvidas com a produção, análise e disseminação de informações oficiais, sobre a finalidade dos Princípios Fundamentais no reconhecimento e fortalecimento da credibilidade das estatísticas oficiais.
- Elaborar manual para a implementação de cada princípio incluindo: o que todos têm que observar em seu trabalho; o que tem de ser recusado; o que pode ser feito se determinados critérios não forem cumpridos; como dar transparência para decidir casos limites e desenvolver a jurisprudência.

• Buscar o aperfeiçoamento e modernização das práticas e da estratégia dos métodos de disseminação e comunicação, com o alinhamento do foco e reafirmando a aderência aos Princípios Fundamentais.

• Ampliar a transparência e o controle externo sobre produtos e processos.

• Disseminar os Princípios Fundamentais para os produtores e usuários de informações oficiais.

• Discutir com os demais produtores de estatísticas oficiais sobre as deficiências do sistema estatístico, explorar as boas práticas e a promoção da integridade e qualidade das estatísticas oficiais, visando o fortalecimento e a eficácia do Sistema Estatístico Nacional.

• Divulgar (periodicamente) relatórios do cumprimento das boas práticas, aderentes aos Princípios Fundamentais, com o objetivo de informar e dar transparência sobre as ações de melhoria.

Diante da ampliação de outros tipos de produtos não reconhecidos como estatísticas oficiais, as práticas a serem aplicadas a esses outros produtos devem ser as mesmas adotadas para as estatísticas oficiais, no que couber, a fim de não prejudicar a reputação geral e a credibilidade da Instituição.

As conexões entre as estatísticas oficiais e a sociedade do futuro certamente vão requerer esforços para enfrentar novos desafios diante: das inovações que acontecem em termos de contextos, questões, fontes de dados, produtores, meios e usuários; do aumento de elementos fora do controle; da necessidade de novos métodos, aprimoramento da comunicação e o foco na preservação e fortalecimento da credibilidade.

A digitalização e a globalização estão criando novas oportunidades e riscos, num contexto em que as informações têm um papel importante se produzidas com qualidade que as torne confiáveis e tenham um propósito específico. As estatísticas oficiais devem buscar ativamente a modernização de seus métodos na produção e comunicação, centrados na qualidade e credibilidade, visando atender às expectativas dos usuários.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Alleva, G. & Gandolfo, M. (2020). Pre-release access to official statistics: The deep difference between granting privileged pre-release access to the government and to the media. Statistical Journal of the IAOS, 36(2), 335–339. Disponível em: <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200626>. Acesso em: janeiro 2021.

Bruun, M. & Mikkelä, H. (2020). Pre-release access to official statistics: The case of Finland. Statistical Journal of the IAOS, 36(2), 327–330. Disponível em:

<https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200621>. Acesso em: janeiro 2021.

CEPAL. (2011). *Código regional de buenas prácticas em estadísticas para América Latina y el Caribe*. Santiago de Chile: CEPAL. 21p. Disponível em: <https://repositorio.cepal.org/bitstream/handle/11362/16422/FILE_148023_es.pdf?sequence=1&isAllowed=y>. Acesso em: abril 2021.

EUROSTAT. (2018). *European Statistics Code of Practice – Revised edition 2017*. Luxembourg. 19p. Disponível em: <https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-18-142>. Acesso em: abril 2021.

Georgiou, A. V. (2020) Prerelease access to official statistics is not consistent with professional ethics. Statistical Journal of the IAOS, 36(2), 313–325. Disponível em: <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200620>. Acesso em: janeiro 2021.

GSBPM. (2019). *Generic Statistical Business Process Model: GSBPM: version 5.1*. United Nations Economic Commission for Europe - UNECE. 32p. Disponível em: <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1>. Acesso em: janeiro 2021.

Hackl, P. (2016). Big Data: What can official statistics expect? Statistical Journal of the IAOS, 32(1), 43-52. Disponível em: <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji965>. Acesso em: janeiro 2021.

IBGE. (1993). *Comercializando estatísticas oficiais sem vender a alma.* Relatórios Técnicos Nº 08/93, Escola Nacional de Ciências Estatísticas – ENCE. Por Lars Thygesen, em *Bulletin of the International Statistical Institute*, 49th Session, Book 2, p 193-203. Invited paper meeting; 17.1. Tradução Djalma G. C. Pessoa. Disponível em: <https://biblioteca.ibge.gov.br/biblioteca-catalogo?id=223999&view=detalhes>. Acesso em: janeiro 2021.

IBGE. (2013). *Código de boas práticas das estatísticas do IBGE.* Rio de Janeiro. 43p. Disponível em: <https://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>. Acesso em: janeiro 2021.

IBGE. (2014). Princípios fundamentais das estatísticas oficiais: orientações para divulgações dos resultados pelo IBGE. Rio de Janeiro. 5p. Disponível em: <https://www.ibge.gov.br/institucional/codigos-e-principios.html?option=com_content&view=article&id=16150>. Acesso em: janeiro 2021.

IBGE. (2015). *Procedimentos para lidar com erros de divulgação de dados e informações estatísticas do IBGE.* Rio de Janeiro. 22p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv94492.pdf>. Acesso em: janeiro 2021.

IBGE. (2016). *Procedimentos para lidar com o mau uso dos dados e informações estatísticas e geoespaciais do IBGE.* Rio de Janeiro. 16p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv98006.pdf>. Acesso em: janeiro 2021.

IBGE. (2018). *Confidencialidade no IBGE: procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas.* Rio de Janeiro. 86p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101636.pdf>. Acesso em: janeiro 2021.

Trewin, D. (2020). Discussant comments on pre-release access to official statistics. Statistical Journal of the IAOS, 36(2), 331–334. Disponível em: <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200628>. Acesso em: janeiro 2021.

United Nations. (2014). Fundamental Principles of Official Statistics, United Nations. Resolution adopted by the General Assembly on 29 January 2014. Disponível em: <https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf>. Acesso em: janeiro 2021

United Nations. (2019). *United Nations National Quality Assurance Frameworks Manual for Official Statistics.* New York. 123p. Disponível em: <https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual>. Acesso em: janeiro 2021.

United Nations. (2021). *The Handbook on Management and Organization of Statistical Systems*. Initial draft version, 4th edition. New York. 576p. Disponível em: <https://unstats.un.org/capacity-development/handbook/index.cshtml>. Acesso em: janeiro 2021.

UNECE. (2008). How Should a Modern National System of Official Statistics Look? Geneva. (Basic text written by Heinrich Bruengger, Director of the Statistical Division of UNECE). 24p. Disponível em: <http://www.unece.org/fileadmin/DAM/stats/documents/applyprinciples.e.pdf>. Acesso em: janeiro 2021.

UNECE. (2009a). *Making Data Meaningful Part 1: A guide to writing stories about numbers.* Geneva. 21p. Disponível em: <https://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part1_English.pdf>. Acesso em: janeiro 2021.

UNECE. (2009b). *Making Data Meaningful Part 2: A guide to presenting statistics.* Geneva. 52p. Disponível em: <http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf>. Acesso em: janeiro 2021.

UNECE. (2011). *Making Data Meaningful Part 3: A guide to communicating with the media.* Geneva. 55p. Disponível em: <http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part3_English_Print.pdf>. Acesso em: janeiro 2021.

UNECE. (2014). *Making Data Meaningful Part 4: A guide to statistical literacy.* Geneva. 65p. Disponível em: <http://www.unece.org/fileadmin/DAM/stats/publications/2013/Making_Data_Meaningful_4.pdf>. Acesso em: janeiro 2021.

UNECE. (2018). *Guidance on Modernizing Statistical Legislation.* 132p. Disponível em: <https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20183.pdf>. Acesso em: fevereiro 2021.

UNECE. (2019). *Strategic Communications Framework for Statistical Institutions.* 67th Plenary session Conference of European Statisticians. Item 4 (e) of the provisional agenda. Paris. 56p. Disponível em: <www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2019/7_Strategic_commmunication_framework_for_consultation.pdf>. Acesso em: janeiro 2021.

UNSD. (1994). Fundamental Principles of Official Statistics. New York. Disponível em: <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>. Acesso em: janeiro 2021.

UNSD. (2015). *Fundamental Principles of Official Statistics: Implementation guidelines.* New York. 115p. Disponível em: <http://unstats.un.org/unsd/dnss/gp/impguide.aspx>. Acesso em: janeiro 2021.

UNSD. (2018). Communicating data and statistics: Bringing trusted and actionable data to the public, the media and policy-makers. High Level Forum on Official Statistics. Side events. Forty-ninth session Statistical Commission. New York. Disponível em: <https://unstats.un.org/unsd/statcom/49th-session/side-events/20180305-3A-high-level-forum-on-official-statistics>. Acesso em: janeiro 2021.

UNSD. (2020). Supplementary chapter to the Fundamental Principles of Official Statistics Implementation Guidelines. Fifty-first session Statistical Commission. Item 3(q) of the provisional agenda. New York. 16p. Disponível em: <http://unstats.un.org/unsd/statcom/51st-session/documents/BG-Item3q-Supplementary_chapter-E.pdf>. Acesso em: janeiro 2021.

APÊNDICE 1

PRINCÍPIOS FUNDAMENTAIS DAS ESTATÍSTICAS OFICIAIS

A seguir, são apresentados o preâmbulo, revisado adotado em 2013, e a descrição de cada um dos dez Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas, de acordo com a tradução do IBGE, disponível em https://www.ibge.gov.br/institucional/codigos-e-principios.html.

Preâmbulo

A Comissão de Estatística

Considerando recentes resoluções[1] da Assembleia Geral e Conselho Econômico e Social, as quais destacam a importância fundamental das estatísticas oficiais para a agenda de desenvolvimento global,

[1] Inclui a resolução nº 64/267 da Assembleia Geral, sobre o Dia Mundial da Estatística, e as resoluções nº 2006/6 e nº 2005/13, do Conselho Econômico e Social, respectivamente em reforço da capacidade estatística e sobre o Programa Mundial de Censo de População e Habitação.

Levando em conta o papel fundamental da informação estatística oficial de alta qualidade nas análises e tomadas de decisão política em prol do desenvolvimento sustentável, da paz e da segurança, bem como para o conhecimento recíproco e comércio entre Estados e povos de um mundo cada vez mais conectado, que exige abertura e transparência,

Levando em conta também que a confiança essencial do público na integridade dos sistemas estatísticos oficiais e nas estatísticas depende, em grande medida, do respeito pelos valores e princípios fundamentais, que são a base de qualquer sociedade que busca compreender a si mesma e respeitar os direitos dos seus membros, e neste contexto, que a independência profissional e prestação de contas de agências estatísticas são cruciais,

Salientando que, para serem eficazes, os valores e princípios fundamentais que regem o trabalho estatístico devem ser garantidos por arcabouço legal e institucional e respeitados em todos os níveis políticos e por todos os intervenientes dos sistemas estatísticos nacionais.

Reafirma os Princípios Fundamentais das Estatísticas Oficiais definidos a seguir.

Princípio 1 - Relevância, imparcialidade e igualdade de acesso

As estatísticas oficiais constituem um elemento indispensável no sistema de informação de uma sociedade democrática, oferecendo ao governo, à economia e ao público dados sobre a situação econômica, demográfica, social e ambiental. Com esta finalidade, os órgãos oficiais de estatística devem produzir e divulgar, de forma imparcial, estatísticas de utilidade prática comprovada, para honrar o direito do cidadão à informação pública.

Princípio 2 - Padrões profissionais e ética

Para manter a confiança nas estatísticas oficiais, os órgãos de estatística devem tomar decisões de acordo com considerações estritamente profissionais, aí incluídos os princípios científicos e a ética profissional, para a escolha dos métodos e procedimentos de coleta, processamento, armazenamento e divulgação dos dados estatísticos.

Princípio 3 - Responsabilidade e transparência

Para facilitar uma interpretação correta dos dados, os órgãos de estatística devem apresentar informações de acordo com normas científicas sobre fontes, métodos e procedimentos estatísticos.

Princípio 4 - Prevenção do mau uso dos dados

Os órgãos de estatística têm direito de comentar interpretações errôneas e utilização indevida das estatísticas.

Princípio 5 - Eficiência

Os dados utilizados para fins estatísticos podem ser obtidos a partir de diversos tipos de fontes, sejam pesquisas estatísticas ou registros administrativos. Os órgãos de estatística devem escolher as fontes levando em consideração a qualidade, oportunidade, custos e ônus para os informantes.

Princípio 6 - Confidencialidade

Os dados individuais coletados pelos órgãos de estatística para elaboração de estatísticas, sejam referentes a pessoas físicas ou jurídicas, devem ser estritamente confidenciais e utilizados exclusivamente para fins estatísticos.

Princípio 7 - Legislação

As leis, regulamentos e medidas que regem a operação dos sistemas estatísticos devem ser tornadas de conhecimento público.

Princípio 8 - Coordenação nacional

A coordenação entre os órgãos de estatística de um país é indispensável, para que se obtenha coerência e eficiência no sistema estatístico.

Princípio 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.

Princípio 10 - Cooperação internacional

A cooperação bilateral e multilateral na esfera da estatística contribui para melhorar as estatísticas oficiais em todos os países.

AGRADECIMENTOS

A autora agradece ao IBGE, que proporcionou a oportunidade de aprendizagem sobre o tema ao longo de sua carreira, à Sonia Albieri e aos revisores pelos comentários e contribuições para uma exposição mais compreensível do artigo. As falhas que porventura persistem são por conta da autora.

Revista Brasileira de Estatística, V.79, N. 246, p.40-63, janeiro-junho 2021

AS BASES DE DADOS OFICIAIS E AS BIBLIOTECAS DO R: INSTRUMENTOS PARA O ENSINO DE UMA ESTATÍSTICA CÍVICA

Maria Tereza Serrano Barbosa

tserranobarbosa@gmail.com

Universidade Federal do Estado do Rio de Janeiro

Davi da Silveira Barroso Alves

davi.alves@uniritotec.br

Universidade Federal do Estado do Rio de Janeiro

Resumo: Este artigo tem o objetivo de demonstrar caminhos para formar estatísticos conscientes da realidade brasileira. Para isso, ele se alinha por um lado aos que consideram importante a ampliação do uso das bases de dados oficiais e as bibliotecas do R nas disciplinas e por outro, aos que se preocupam com o desconhecimento dos usuários a respeito de conceitos importantes da Estatística que levam à falta de replicabilidade das pesquisas ou interpretações incorretas dos seus resultados. A atividade proposta utiliza a Base de dados da PNAD COVID e as bibliotecas do R *survey* e *convey* para estimar e discutir como o cálculo de indicadores incorporando o plano amostral pode contribuir com uma formação tecnicamente sólida e socialmente referenciada.
Palavras-chave: Estatísticas oficiais; Educação Estatística; desigualdade; estatística cívica

Abstract: This article aims to demonstrate ways to form statisticians aware of the Brazilian reality. For this, it aligns on the one hand with those who consider important the expansion of the use of official databases and R libraries in disciplines and on the other, to those who are concerned about the lack of knowledge of users about important concepts of statistics that lead to the lack of replicability of research or incorrect interpretations of their results. The proposed activity uses the PNAD COVID Database and the R survey and convey libraries in order to estimate and discuss how the calculation of indicators incorporating the sampling plan can contribute to a technically sound and socially referenced education.
Keywords: Official statistics; Statistical Education; inequality; civic statistics

1. INTRODUÇÃO

Este artigo destaca a importância de incluir, entre os objetivos de aprendizagem das disciplinas de Estatística, a contribuição com a formação de cidadãos conscientes da realidade brasileira. Para isso, apresenta o exemplo de uma atividade pedagógica que permite a compreensão dos conceitos mais simples ou mais complexos de forma contextualizada. No seu planejamento utilizaram-se materiais de minicursos, artigos, o livro e bibliotecas do R produzidos pelo Professor Djalma Pessoa.

O legado do Prof. Djalma para a Estatística brasileira vai muito além dos trabalhos desenvolvidos sozinho, com orientandos ou colegas. Nos cinquenta anos após o seu doutorado, não lhe faltaram entusiasmo nem dedicação para aprender, ensinar, mostrar a importância social da Estatística, o potencial de suas aplicações e ainda implementar rotinas computacionais no R. Além de ter contribuído direta ou indiretamente para a formação de várias gerações de estatísticos brasileiros, também foi responsável, através da sua atuação, por fortalecer e divulgar uma formação ética e com compromisso social. Sempre preocupado com a qualidade no ensino, a motivação dos alunos e a aplicação social da Estatística, o Professor Djalma foi um Educador estatístico muito antes do surgimento desta área de pesquisa, foco principal das reflexões aqui realizadas.

A área da Educação Estatística surgiu há mais de trinta anos e vem produzindo estudos e discussões que indicam a necessidade de mudanças nas disciplinas introdutórias de Estatística nos cursos superiores em geral, mas também nas graduações em Estatística. Dentre as mais importantes reflexões sobre este tema estão as realizadas por Moore (1998) e que passaram a ser uma referência obrigatória para os principais autores da área da Educação Estatística. Apesar de Moore ter reforçado a ideia de um primeiro curso introdutório mais atraente, com pouca formalidade e estimulante intelectualmente, sua principal e original tese é de que a Estatística é uma das artes liberais. Este termo define uma metodologia de ensino na Idade Média que buscava uma formação multidisciplinar que adequasse o intelecto à realidade do mundo. Assim, segundo Moore (1998), o pensamento estatístico na sociedade moderna deveria ter o mesmo valor para a formação da cidadania que o treinamento de falar em público tinha na Roma de Cícero. Para ele, trabalhar com dados é tanto arte quanto ciência e que nós aprendemos não apenas dominando métodos formais, mas seguindo exemplos de professores atuais e mestres do passado.

Partindo destas importantes reflexões, aqui serão destacadas duas outras justificativas recentes que advogam as mudanças nas disciplinas iniciais de Estatística. As trazidas por Steel et al. (2019) que partem da discussão sobre a crise atual atravessada pela ciência devido à incapacidade de replicação de resultados científicos. Estes autores defendem uma formação com foco no pensamento estatístico e na interação entre estudantes com diferentes tipos de conhecimentos, argumentando que os conhecimentos processuais dos estudantes não garantem que consigam pensar e compreender o mundo a partir do que os resultados dos seus estudos informam. Um segundo campo de reflexões trazidas por outro conjunto de pesquisadores (Engel, 2014, 2017; Gal & Ograjenšek, 2017; Nicholson et al., 2018) defende a necessidade de que o ensino da Estatística contribua com a formação de cidadãos com mais engajamento e participação em sociedades democráticas. Algumas destas reflexões discutem o conceito de *Estatísticas cívicas* que deram origem inclusive a um projeto Europeu denominado *Pro Civic Statistics* ("ISLP — ProCivicStat", 2018) com o objetivo de promover o engajamento cívico a partir da utilização de bases de dados reais para investigar a realidade social.

No Brasil, a utilização de dados reais no desenvolvimento de projetos em sala de aula tem sido bastante difundida como uma alternativa pedagógica, pela área da Educação Estatística, tanto no Ensino Básico (Barberino & Magalhães, 2016) quanto no Ensino Superior (Barbosa et al., 2016; Campos & Coutinho, 2019). Estas abordagens têm favorecido o letramento estatístico que, segundo definição de Gal (2002), pode ser resumido como: *i)* a habilidade de interpretar e avaliar criticamente uma informação estatística; e *ii)* a habilidade de discutir e comunicar estas informações, dando sua opinião a respeito delas.

Durante a pandemia da COVID-19, a necessidade do letramento estatístico ficou evidente quando a mídia, além de apresentar gráficos diários, entrevistava especialistas sobre resultados de modelos estatísticos, das pesquisas por amostra para estimar a prevalência de infecções ou dos estudos clínicos para medir a eficácia das vacinas. Neste momento em que o Brasil já tem mais de 400 mil mortos e reflexos econômicos e sociais em uma magnitude que atinge principalmente a população mais vulnerável (Nassif et al., 2020), esta necessidade fica ainda mais clara. O impacto na sociedade desta grave crise sanitária e econômica pode servir como uma oportunidade para ampliar a compreensão a respeito dos elementos essenciais para uma formação estatística cívica.

Para isso acontecer, os cursos introdutórios nas graduações e pós-graduações em Estatística poderiam priorizar o exercício da construção do

pensamento estatístico em uma análise de dados reais que afetem a sociedade, em detrimento à priorização de uma formação técnica abstrata.

Nesta direção, Weiland (2017) propôs uma extensão da definição de letramento estatístico, incluindo a habilidade de ler e escrever tanto a palavra quanto o mundo, como cidadãos críticos. Ele defendeu que os estudantes devem ter oportunidades de abordar questões sociopolíticas complexas em conjunto com a aprendizagem de poderosos conceitos e práticas estatísticas.

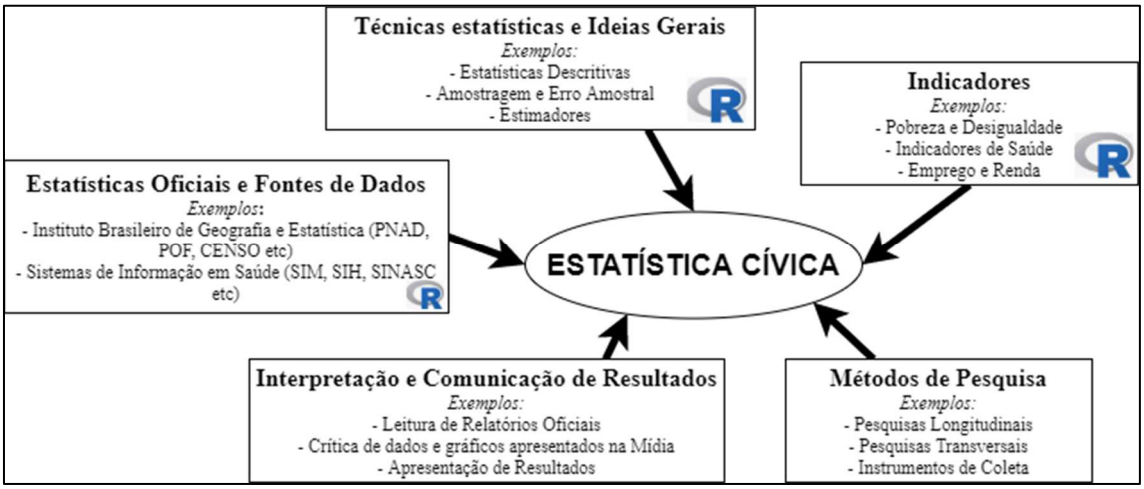
Os caminhos didáticos para atingir este tipo de letramento não são únicos e nem são fáceis, mas devem ser experimentados com vistas a uma formação mais condizente com as necessidades da ciência e da sociedade.

Uma metodologia com vistas a uma formação crítica cidadã dos estudantes foi apontada por Barbosa et al. (2019) que mostraram como as etapas do ciclo investigativo de uma pesquisa poderiam estimular o debate de temas da realidade social dos estudantes e com isso tornar a disciplina mais interessante. A questão que se coloca aqui é: a realidade social pode ser uma motivação para a aprendizagem de Estatística, ou é a Estatística que pode ser utilizada para motivar os estudantes a conhecerem a realidade da sociedade em que vivem?

Uma resposta para esta questão pode ter sido apontada por Gal & Ograjenšek (2017), que ao identificarem a ausência de uma formação em estatísticas oficiais, tanto nos públicos mais amplos, quanto nos professores de disciplinas introdutórias nos cursos de graduação, propõem um modelo com seis elementos para formação no que eles denominam de letramento em estatísticas oficiais.

Com base neste modelo e nas reflexões realizadas por Weiland (2017) e Nicholson et al. (2018), aqui foi feita uma proposta adaptada à realidade das estatísticas oficiais brasileiras para o desenvolvimento de uma Estatística cívica que considere questões sociopolíticas, conceitos estatísticos e procedimentos práticos de aquisição e análise de dados (**Figura 1**). Para os procedimentos práticos priorizou-se a utilização de bibliotecas para a aquisição de dados, incorporação do plano amostral e cálculo de indicadores desenvolvidas para o programa R, gratuito e de acesso aberto.

Figura 1 – Elementos fundamentais para uma formação estatística cívica adaptada à realidade brasileira



O objetivo deste artigo é o de demonstrar a possibilidade de uma formação Estatística cívica através da proposição de uma atividade baseada em um exemplo deste modelo apresentado. As contribuições do Prof. Djalma Pessoa, tanto na descrição de porque é necessária a incorporação do plano amostral nas estimativas, como também no desenvolvimento de uma biblioteca do R que permite o cálculo de indicadores de pobreza e desigualdade incorporando o plano amostral, foram as principais fontes inspiradoras.

O artigo foi dividido em seis seções. Esta primeira faz uma apresentação geral do tema com vistas a justificar o modelo proposto e as demais seções contextualizam e materializam alguns dos seus elementos. Na seção 2 é abordada a importância das estatísticas oficiais para a sociedade. A seção 3 tem o objetivo de contextualizar a agenda 2030, descrever seu primeiro objetivo e discutir a respeito de decisões técnicas que podem interferir nos valores dos indicadores a serem monitorados. A seção 4 trata da importância da incorporação do plano amostral na análise dos dados. Na seção 5 são descritas a base de dados e as bibliotecas do R utilizadas para incorporar o plano amostral e calcular indicadores de pobreza e desigualdade. E a seção 6 descreve uma atividade com suas várias etapas.

2. A IMPORTÂNCIA DAS ESTATÍSTICAS OFICIAIS PARA A SOCIEDADE

Um maior conhecimento a respeito da realidade brasileira através de indicadores demográficos, de saúde, etnia, filiação religiosa, pobreza entre outros tem sido possível a partir dos avanços metodológicos e tecnológicos que permitiram não só uma maior produção, como o acesso público aos microdados

advindos das mais diversas fontes e pesquisas. Isto reforça as reflexões realizadas por Gal & Ograjenšek (2017) a respeito da necessidade da formação em estatísticas oficiais.

A importância das estatísticas oficiais foi destacada por Feijó & Valente (2006) ao considerarem que as mudanças tecnológicas levaram à possibilidade da produção de informações mais rápidas para as sociedades modernas e fizeram com que o conhecimento passasse a ter um papel crucial como base do desenvolvimento e da hegemonia mundial. Neste contexto, relembraram os princípios estabelecidos pela Organização das Nações Unidas (ONU) para orientar o funcionamento de um Sistema Nacional de Estatística e o papel do Instituto Brasileiro de Geografia e Estatística (IBGE), como órgão produtor e coordenador da produção estatística nacional.

Entre os princípios estabelecidos pela ONU em 1994 e ratificados em 2014, destacam-se aqui os três primeiros: o que estabelece as estatísticas oficiais como um elemento indispensável para o sistema de informação de uma sociedade democrática, provendo o Governo, a economia e o público com dados sociais, econômicos e ambientais; o que estabelece a necessidade das agências estatísticas agirem de forma ética e científica nas definições de métodos para coleta, processamento, armazenamento e apresentação de dados; e finalmente, o terceiro princípio que trata da necessidade da transparência na apresentação das estatísticas, apresentando-as de acordo com as normas científicas sobre as fontes, métodos e procedimentos (United Nations, 2013).

Estes princípios, assim como os outros, têm incentivado o desenvolvimento e aperfeiçoamento constante das metodologias de produção das estatísticas oficiais e a ampla disponibilização de microdados. A importância deles foi reforçada quando os 193 estados membros da Organização das Nações Unidas estabeleceram os Objetivos do Desenvolvimento Sustentável (ODS), em uma agenda para 2030, a partir do monitoramento de indicadores ambientais, sociais e econômicos. No Brasil, a tarefa de produzir estes indicadores de forma periódica, confiável e com qualidade e disponibilização de dados desagregados é coordenada pelo IBGE, que precisa fazer amplo trabalho de articulação interinstitucional (Kronemberger, 2019).

3. INDICADORES DE POBREZA E DESIGUALDADE NA AGENDA 2030

A agenda 2030 para o Desenvolvimento Sustentável foi adotada por 193 Estados membros da Organização das Nações Unidas (ONU). Desta agenda constam os Objetivos do Desenvolvimento Sustentável (ODS) que orientaram as

metas e indicadores para serem monitorados nas dimensões econômica, social e ambiental. Esta agenda é constituída por 17 Objetivos de Desenvolvimento Sustentável (ODS), 169 metas e 232 indicadores. A criação, compatibilização com a realidade brasileira, acompanhamento e avaliação destes indicadores, em nível regional e nacional constitui, portanto, um imenso desafio com necessidade de interlocução entre um grande número de instituições. Uma descrição desta complexidade e desafios pode ser encontrada em Kronemberger (2019).

O primeiro objetivo (ODS1) é o da "erradicação da pobreza em todas as suas formas e em todos os lugares". Neste objetivo existem duas metas, mas aqui será abordada apenas a primeira: "Até 2030, erradicar a pobreza extrema para todas as pessoas em todos os lugares, atualmente medida como pessoas vivendo com menos de US$ 1,25 por dia".

Para avaliar o cumprimento desta meta, os países precisam monitorar o percentual da sua população abaixo da linha internacional de pobreza extrema, por sexo, idade, status de ocupação e localização geográfica (urbano/rural). Para isso, é necessário que estimem regularmente os rendimentos domiciliares *per capita* auferidos por sua população, definam o valor de corte na moeda local e deflacionem este valor durante o acompanhamento. Entre os elementos essenciais propostos para uma formação cívica, segundo Gal & Ograjenšek (2017), estaria a compreensão a respeito das diversas fontes de variabilidade destas estimativas. Importante distinguir quando a origem desta variabilidade é de uma opção conceitual ou metodológica, ou de outros fatores, mais ou menos controláveis. Os efeitos das diferentes opções no cálculo desse indicador podem ser abordados em sala de aula a partir de estudos que avaliam o cumprimento desta meta e chegar a diferentes valores, ou a partir de exercícios práticos com as bases de dados oficiais.

O Instituto de Pesquisa Econômica Aplicada (IPEA), ao apresentar as adequações necessárias para o acompanhamento das metas globais da Agenda 2030 à realidade brasileira propõe utilizar como referência o dólar PPC (paridade do poder de compra) que elimina as diferenças de custo de vida entre os países. Assim, a meta 1.1 para o Brasil, segundo IPEA (2018), seria: "Até 2030, erradicar a pobreza extrema para todas as pessoas em todos os lugares, medida como pessoas vivendo com menos de PPC$ 3,20 *per capita* por dia". Segundo este mesmo documento, este valor, em 2016, equivaleria a R$ 226,14 por mês, com um percentual de 12,55% da população abaixo desta linha.

A atividade proposta será baseada neste primeiro objetivo (ODS 1) e o cálculo dos seus indicadores serão realizados a partir da biblioteca *convey* (Pessoa et al., 2020).

4. IMPORTÂNCIA DO PLANO AMOSTRAL NAS ANÁLISES DOS DADOS

As noções básicas de amostragem fazem parte de qualquer disciplina de Estatística e constam dos livros textos utilizados nas graduações de áreas não pertencentes às ciências exatas, tais como as das ciências sociais ou ciências da saúde. Apenas com os conceitos básicos é possível apresentar de forma intuitiva como fazer a expansão de uma amostra aleatória simples, onde cada unidade tem a mesma probabilidade de seleção, a partir da multiplicação do valor de uma variável pelo peso amostral. No entanto, a necessidade de se recorrer a procedimentos de estratificação das unidades, a vários estágios de seleção, pode ser apresentada informalmente, mas gera uma variedade de subscritos nas fórmulas dos estimadores e de suas variâncias necessárias para as estimativas nas amostras mais complexas. Esta complexidade também faz com que se perca a intuição a respeito do significado do peso amostral e da necessidade do peso amostral ser incorporado às estimativas de parâmetros populacionais e das variâncias amostrais destes estimadores. A apresentação das fórmulas desses diversos estimadores e de suas variâncias só é realizada em cursos mais formais de amostragem nas graduações de Estatística ou nas pós-graduações de áreas correlatas e, na maioria das vezes, sem fazer uma contextualização que incentive uma discussão a respeito da realidade descrita pelas estimativas obtidas.

Em contraposição a isso, percebe-se que os principais usuários das estatísticas oficiais produzidas nas mais diversas áreas não foram preparados para identificarem as limitações e alcances de um processo de estimação por amostra. Talvez por isso, durante muito tempo, as pesquisas que utilizavam dados oriundos de amostras complexas, disponibilizados nos microdados, não consideravam que tanto as estimativas quanto as medidas de precisão seriam influenciadas conjuntamente pela estratificação, conglomeração e pesos.

No Brasil, a ausência da possibilidade de incorporar o desenho amostral nas estimativas realizadas pela maioria dos pacotes estatísticos foi identificada por Pessoa et al. (1997) que, além de avaliarem o efeito desta simplificação nas estimativas dos parâmetros de um modelo, apresentaram os fundamentos teóricos para a incorporação do desenho amostral da PNAD 90. Com maior detalhamento técnico a respeito deste tema, Silva et al. (2002) partem do Plano amostral na PNAD 98, utilizam seus dados e analisam o efeito nas estimativas de variabilidade ao incorporar o Plano amostral. Por ter conseguido apresentar o problema e a solução, de forma didática e relativamente simples aos usuários de dados da PNAD, esta referência já foi citada segundo o google acadêmico

por 193 trabalhos científicos. A importância deste tema para os usuários das mais diversas áreas faz com que a inclusão desta referência em qualquer disciplina de amostragem ou mesmo de modelagem ajudem os estudantes a terem uma compreensão menos abstrata do problema. Além disso, o recente livro de Pessoa & Silva (2018a) também conjuga a teoria e prática necessárias a uma abordagem didática bastante eficiente.

5. BASE DE DADOS DA PNAD COVID19 E BIBLIOTECAS DO R

As atividades propostas neste artigo utilizarão a Base de dados da PNAD COVID19 e as bibliotecas do R *survey* (Lumley, 2020) e *convey* (Pessoa et al., 2020), que permitem obter estimativas de indicadores oriundos de amostras complexas, incorporando as informações do desenho amostral.

As descrições estão feitas de uma forma que os *scripts* possam ser replicados ou adaptados em salas de aula por todos aqueles que possuírem algum conhecimento do R. Estes exemplos práticos favorecem a aprendizagem com o foco em alguns dos seis elementos apontados por Gal & Ograjenšek (2017) para o letramento dos cidadãos. Mais especificamente, destaca-se a importância do conhecimento das estatísticas oficiais, suas bases de dados, processo de pesquisa e importância do plano amostral assim como a possibilidade de comparar e interpretar os indicadores de pobreza.

5.1 DESCRIÇÃO DA BASE PNAD COVID

A PNAD COVID é uma versão da Pesquisa Nacional por Amostra de Domicílios Contínua realizada pelo IBGE durante a pandemia do novo coronavírus. O levantamento de dados foi realizado durante os meses de maio a novembro de 2020 através de entrevistas telefônicas atingindo quase 50 mil domicílios por semana em todo o território nacional (IBGE, 2020a) através de uma amostra fixa, de maneira que todos os domicílios entrevistados no mês inicial permaneceram na amostra nos meses posteriores (IBGE, 2020b).

A pesquisa utilizou a base da amostra de domicílios da PNAD Contínua e teve, no final do processo, 193.662 domicílios selecionados para sua amostra, representando cerca de 92% da amostra base. Segundo o IBGE (2020c), “A amostra original da PNAD Contínua foi obtida por um plano amostral conglomerado em dois estágios com estratificação das unidades primárias de amostragem (UPAs). No primeiro estágio foram selecionadas UPAs com probabilidade proporcional ao número de domicílios dentro de cada estrato definido. No segundo estágio foram selecionados 14 domicílios particulares permanentes (que podem estar ou não ocupados) dentro de cada UPA da amostra do primeiro estágio. O sorteio dos domicílios em cada UPA foi feito por

amostragem aleatória simples, considerando os endereços listados no Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) atualizado para cada UPA."

Para a coleta de dados, foi utilizado um instrumento dividido em duas partes, uma relacionada à saúde de todos os moradores dos domicílios da amostra e outra relacionada a informações sobre ocupação, trabalho, busca por trabalho e rendimento, abordando ainda os rendimentos não oriundos do trabalho, como aluguéis, aposentadoria e auxílios (IBGE, 2020a).

Seus dados podem ser obtidos utilizando a biblioteca R denominada *COVIDIBGE* (Assunção & Hidalgo, 2020). Esta biblioteca facilita o usuário pois permite que, ao importar a base de dados requerida, as recodificações das variáveis sejam realizadas conforme o dicionário de dados[1] a partir do argumento *labels = TRUE* e que o plano amostral seja incorporado com uso do argumento *design=TRUE*, criando o objeto necessário para as análises na biblioteca *survey*, que será apresentada na próxima seção.

A seguir, o *script* R para criação de um arquivo de dados covidbr06 da base de dados da PNAD COVID referente ao mês de junho de 2020:

Instalando e carregando o pacote e fazendo o download do arquivo de dados referente ao mês de junho de 2020 da PNAD COVID:

```
install.packages("COVIDIBGE")

install.packages("survey")

library(COVIDIBGE)

library(survey)

covidbr06 <- get_covid(year=2020,month=6, labels = TRUE, deflator =
TRUE, design = TRUE, savedir = tempdir())
```

Outras bases de dados do IBGE precisarão de download e importação direta, de recodificação conforme dicionário da pesquisa e criação do objeto da biblioteca *survey*. O roteiro disponível em http:// davibalves.github.io/roteiro_artigo_rbe_celebrando_Djalma_Pessoa apresenta o *script* de um exemplo de download, importação da base de dados da PNAD COVID preparação do objeto *survey design*.

[1]https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_PNAD_COVID19/Microdados/Documentacao/Dicionario_PNAD_COVID_062020.xls

5.2. BIBLIOTECA *SURVEY* DO R PARA INCORPORAR O PLANO AMOSTRAL

A biblioteca *survey* desenvolvida para o software estatístico R por Thomas Lumley permite a incorporação das informações do plano amostral às análises de dados de amostras complexas. Segundo Lumley (2004), a biblioteca tem como propósitos principais: I) a possibilidade de aliar as informações dos metadados do desenho amostral à base de dados em um objeto que integra os ajustes necessários para a correta análise de dados, permitindo a extração de subconjuntos do banco de dados preservando os metadados através de *subsets* do objeto criado, garantindo estimativas corretas para as subpopulações; e II) fornecer estimativas válidas de variância para as estatísticas calculadas utilizando este objeto.

O pacote fornece ferramentas para a implementação de análises amplamente realizadas pelo R, como a obtenção de tabelas de frequência, medidas resumo e ajuste de modelos realizando as estimativas corretas a partir do plano amostral (Lumley, 2004). No roteiro disponível em http:// davibalves.github.io/roteiro_artigo_rbe_celebrando_Djalma_Pessoa se encontram exemplos de *scripts* para estimar médias e proporções utilizando a biblioteca.

5.3. BIBLIOTECA CONVEY DO R PARA CALCULAR OS INDICADORES DE POBREZA

A biblioteca *convey* do R, desenvolvida por Djalma Pessoa, Anthony Damico e Guilherme Jacob, permite estimar indicadores de pobreza, desigualdade e bem-estar, considerando as informações do desenho amostral de amostras complexas, de maneira integrada ao pacote *survey*. Mais de vinte indicadores podem ser obtidos com o uso do pacote, desde indicadores de pobreza absoluta monetária, como a proporção da população abaixo da linha da pobreza, indicadores de desigualdade de renda como a disparidade salarial de gênero e o índice de Gini, além de indicadores multidimensionais (Jacob et al., 2020).

Para estimar os indicadores utilizando a biblioteca *convey* é necessário preparar um objeto da biblioteca *survey* com o desenho amostral (*survey design*) utilizando a função *convey_prep*. Nos *scripts* da atividade apresentados a seguir são mostrados os procedimentos para o preparo do objeto para uso das funções da biblioteca *convey* e a estimação de indicadores de pobreza e extrema pobreza em subgrupos específicos da população e o roteiro de análise disponível em

http://davibalves.github.io/roteiro_artigo_rbe_celebrando_Djalma_Pessoa/ mostra exemplos de estimação do Índice de Gini.

6. PROPOSTA DE ATIVIDADE PARA PROMOÇÃO DA ESTATÍSTICA CÍVICA EM SALA DE AULA

A depender da disciplina ou do grau do curso, a atividade aqui descrita pode ser realizada na sua totalidade, incluindo todos os conceitos teóricos envolvidos em uma estimação a partir de uma amostragem complexa, ou apenas parcialmente com ênfase nas interpretações das estimativas com vistas a demonstrar efeitos práticos ao não se levar em conta o desenho amostral, não se definir corretamente o indicador ou não se levar em consideração a deflação de valores monetários. Ela deve ser vista apenas como um exemplo que pode ser ampliado, reduzido ou modificado para outros temas, banco de dados ou conceitos estatísticos. O importante é que a sua realização promova discussões tanto a respeito da realidade social brasileira, quanto a respeito das limitações e alcances advindos de uma única análise estatística.

A atividade propõe a comparação de indicadores de pobreza e extrema pobreza com e sem o auxílio emergencial instituído a partir da epidemia da COVID-19. Os objetivos, etapas e pré-requisitos necessários estão descritos no **Quadro 1**. Ela foi pensada para provocar o debate sobre a situação de pobreza no Brasil antes e após a pandemia e possibilitar que os estudantes façam uma análise com dados reais e depois apresentem seus resultados.

Quadro 1 - Descrição resumida do Plano de aula para a atividade

Atividade	Objetivos	Etapas de realização	Conceitos
1. Comparação de indicadores de pobreza e extrema pobreza com e sem o auxílio emergencial obtidos a partir da PNAD COVID	i) Incentivar o debate sobre a realidade brasileira; ii) Demonstrar a importância das estatísticas oficiais; iii) Exercitar as habilidades de produzir, interpretar e comunicar resultados estatísticos; iv) Avaliar o efeito nas estimativas da incorporação do Plano amostral; v) Contribuir com a formação estatística cívica.	i) Discussão em grupo; ii) Definição dos objetivos da análise; iii) Criação da base de dados e verificação das variáveis; iv) Cálculo dos indicadores e efeito do plano amostral; v) Elaboração de um relatório com as conclusões; vi) Seminário de apresentação.	i) Amostragem e erros amostrais; ii) Estatística descritiva.

Durante sua execução os estudantes seriam incentivados a avaliar a situação do país em relação ao cumprimento da meta 1 da agenda 2030, expandir as análises por gênero, raça, localização geográfica e região, e discutir o impacto do auxílio emergencial na proporção de pobreza em cada grupo específico.

Por ser destinada a motivar os mais diversos perfis de professores e alunos, as etapas relacionadas ao cálculo do indicador de pobreza com e sem o auxílio emergencial na população como um todo e em grupos específicos da população serão descritas de forma bastante detalhada para permitir a ampliação dos usuários destas bases oficiais.

6.1. DESENVOLVIMENTO DA ATIVIDADE

6.1.1. ETAPA 1: LEITURA E DISCUSSÃO DE TEXTOS

Para a realização da atividade os estudantes devem ser inicialmente divididos em grupos que receberão dois textos temáticos com questões motivadoras como as apresentadas no **Quadro 2**, desenvolvidas com base nos artigo de Santos (2012), Monte (2020) e Haeberlin & Silva (2019). Enquanto o primeiro artigo permitirá uma discussão mais ampla e profunda a respeito da objetividade e subjetividade por trás das mensurações da pobreza, abordando os diferentes limiares para defini-las, os outros dois utilizam o limiar proposto pelo Banco Mundial para analisar o impacto de políticas de transferência de renda na diminuição da pobreza no Brasil. Conforme o curso e disciplina, os textos podem ser um artigo científico, um relatório, uma notícia na imprensa ou apenas um trecho extraído de um destes materiais.

Quadro 2 – Exemplos de possíveis questões Motivadoras

1. *A quem foi destinado o auxílio emergencial e qual foi o motivo da sua criação?*
2. *Qual a importância das estatísticas produzidas pelo IBGE para a sociedade brasileira?*
3. *Quais as diferenças conceituais existentes entre os indicadores de pobreza utilizados no Brasil?*
4. *Qual foi o efeito do auxílio emergencial na renda domiciliar per capita?*
5. *O Brasil ainda pode cumprir a meta de erradicação da pobreza?*
6. *Quais grupos da população foram mais beneficiados pelo auxílio?*
7. *Qual o impacto da redução dos valores do auxílio emergencial na proporção de pessoas em situação de pobreza?*

Desta maneira, os alunos poderão imergir na temática da situação da pobreza no Brasil e suas especificidades regionais, vislumbrando a possibilidade de conhecer melhor a realidade na qual estão inseridos, a partir de dados públicos de acesso aberto, e apoderar-se do processo de pesquisa, refletindo

sobre quais perguntas podem ser formuladas e quais as considerações necessárias para o desenvolvimento das análises para respondê-las de uma perspectiva teórica.

6.1.2. ETAPA 2: DEFININDO OS OBJETIVOS E METODOLOGIA DA ANÁLISE

Após a leitura e discussão do (s) texto (s), os estudantes devem receber o objetivo da atividade: "*comparar os indicadores de pobreza com e sem o auxílio emergencial obtidos a partir da PNAD COVID*" e cada grupo deve estabelecer os objetivos das análises que irão realizar, organizando e planejando os procedimentos necessários para sua execução. A formulação clara dos objetivos de um estudo, bem como a descrição da metodologia a ser utilizada, vêm sendo apontadas pela área da Educação Estatística e reforçada por Steel et al (2019) como fundamentais no processo científico.

Em seguida, para subsidiar a definição metodológica, devem ser apresentados aos textos que abordem a importância da incorporação do plano amostral na estimação de indicadores, tais como o prefácio do livro "Análise de Dados Amostrais Complexos" (Pessoa & Silva, 2018b), que introduz de maneira mais ampla estes conceitos e o artigo "Análise de dados da PNAD: incorporando a estrutura do plano amostral" (Silva et al., 2002), que permite tanto formalizar o processo de incorporação do plano amostral no cálculo das estimativas e medidas de variabilidade quanto abordar o efeito do plano amostral nestas estimativas de maneira prática.

De acordo com a profundidade com que o tema e as técnicas de análise serão explorados poderá ser solicitado que especifiquem quais variáveis serão analisadas, como os indicadores de renda e desigualdade serão calculados, quais as considerações teóricas para suas escolhas e quais os procedimentos necessários para obtê-los.

Nesta etapa podem ser formalizados os conceitos relacionados à estimação utilizando dados de amostra complexa e as bibliotecas necessárias para a obtenção de indicadores incorporando o plano amostral, fornecendo materiais de referência que apresentem estes conceitos, bem como os procedimentos de análise e os *scripts* de programação em R para desenvolvê-las, como a apresentação do minicurso ministrado por Djalma Pessoa e André Costa no Seminário de Metodologia do IBGE (Pessoa & Costa, 2013) e o laboratório de R do livro "Análise de Dados Amostrais Complexos" (Pessoa & Silva, 2018c).

A seguir serão apresentados, a título de exemplo, os procedimentos e *scripts* R para o desenvolvimento de análises que contemplem os seguintes

objetivos: I) "*Analisar o impacto do Auxílio Emergencial na proporção de pobres no Brasil em junho de 2020 na população em geral"*; e II) "*Avaliar o impacto do Auxílio Emergencial na proporção de pobres no estado do Rio de Janeiro em junho de 2020 segundo localização geográfica (urbana/rural)"*. Também serão apresentados os resultados de uma análise com o objetivo de "*Comparar o impacto do Auxílio Emergencial na proporção de pobres e extremamente pobres no Rio de Janeiro em junho e novembro de 2020*".

6.1.3. ETAPA 3: CRIAR A BASE DE DADOS E DEFINIR AS VARIÁVEIS PARA OS INDICADORES

Nesta etapa os alunos deverão adquirir e preparar a base de dados para a implementação das análises necessárias para contemplar os objetivos propostos na etapa anterior. Para tanto, deverão buscar informações referentes à coleta de dados da pesquisa, buscando compreender como as variáveis relacionadas ao rendimento foram obtidas e quais os procedimentos para obtenção da informação necessária para as estimativas dos indicadores.

Neste exemplo, os indicadores de pobreza e extrema pobreza a serem calculados referem-se à porcentagem de pessoas cuja renda domiciliar *per capita* mensal está abaixo respectivamente de US$ 5,50 e US$ 1,90 por dia, segundo definido pelo Banco mundial. O cálculo será realizado considerando e desconsiderando o rendimento referente ao Auxílio Emergencial. A base de dados utilizada será a "covidbr06" apresentada na seção 5.1, referente aos dados da PNAD COVID do Brasil no mês de junho de 2020.

Para isso, é necessário conhecer, a partir do questionário da pesquisa, a identificação de todas as variáveis que contêm as informações dos diversos tipos de rendimentos e saber também que:

- A base de dados da PNAD COVID apresenta as informações de rendimento referente ao trabalho e demais fontes para cada pessoa do domicílio;
- O rendimento do trabalho foi coletado para todos os indivíduos com 14 ou mais anos de idade que exerciam atividade remunerada no mês de referência;
- Os rendimentos das demais fontes (pensão, bolsas, auxílios, aluguéis, entre outros) comuns a todos os moradores do domicílio, constam de maneira repetida em cada linha referente aos residentes daquele domicílio;

Assim, neste exemplo, a variável utilizada para obter a proporção de indivíduos abaixo da linha de pobreza e extrema pobreza será o rendimento

domiciliar *per capita* calculado a partir da soma das rendas oriundas do trabalho e a média dos valores de demais rendimentos de acordo com o número de moradores no domicílio.

Para o cálculo da renda domiciliar *per capita* será utilizada a biblioteca *dplyr* para agregar as rendas segundo domicílio, identificados pela Unidade Primária de Amostragem (variável "UPA") e pela ordem do domicílio sorteado na respectiva UPA (variável "V1008" que pode assumir valores de 1 a 14). Para esta atividade serão criadas duas rendas domiciliares *per capita*: i) "renda_dom_per_com", que inclui o Auxílio Emergencial; e ii) "renda_dom_per_sem", desconsiderando esse auxílio. Estas duas variáveis de renda serão adicionadas ao objeto do desenho amostral "covidbr06" e logo em seguida será realizado o preparo deste objeto para o uso da biblioteca *convey*.

Obtendo a renda domiciliar per capita com e sem o auxílio emergencial

```
install.pakages("dplyr")

library("dplyr")

covidbr06$variables %>%
  group_by(UPA,V1008,D0013,D0023,D0033,D0043,D0053,D0063,D0073) %>%
  summarise(n_moradores=n(),renda_trab=sum(C011A12,na.rm=T)) %>%
  mutate(
  renda_trab_per=renda_trab/n_moradores,D0013_per=D0013/n_moradores,
  D0023_per=D0023/n_moradores,D0033_per=D0033/n_moradores,
  D0043_per=D0043/n_moradores,D0053_per=D0053/n_moradores,
  D0063_per=D0063/n_moradores,D0073_per=D0073/n_moradores)%>%
  rowwise() %>%
  mutate(
  renda_fam_per_com=sum(c_across(renda_trab_per:D0073_per),na.rm=T),
  renda_fam_per_sem=sum(c_across(renda_trab_per:D0073_per)[-6],
  na.rm=T)) %>%
  ungroup() %>% select(UPA,V1008,n_moradores,renda_fam_per_com,
  renda_fam_per_sem)-> renda_domiciliar
```

Adicionando as rendas ao objeto do desenho amostral

```
covidbr06$variables<-left_join(covidbr06$variables,renda_domiciliar)
```

Preparando o objeto para o uso da biblioteca *convey*

```
covidbr06_prep <- convey_prep(covidbr06)
```

Para esta etapa, conforme o curso e conhecimento do R, o professor poderia, com vistas a tornar mais ágil a atividade, ter criado objetos *.RData* de cada Unidade da Federação já com as variáveis de renda domiciliar calculadas e preparadas para utilizar a biblioteca *convey*, isto porque para o uso da mesma

é necessário que o objeto *svydesign* seja preparado pela função *convey_prep* antes de que qualquer subconjunto seja definido.

Contudo, ainda que seja fornecida a base de estados ou regiões específicas, é importante que os alunos tenham a oportunidade de discutir e compreender o cálculo das variáveis de rendimento que serão utilizadas, de maneira que possam interpretar o cálculo dos indicadores de pobreza e discutirem outros caminhos possíveis para obtê-los.

Neste exemplo, será descrita a criação de um subconjunto com os dados do Rio de Janeiro a partir do objeto "covidbr06" utilizando a biblioteca *COVIDIBGE*:

```
Selecionar o subconjunto do objeto do plano amostral "survey design"
utilizando o Rio de Janeiro como exemplo:
covidrj06_prep <- subset(covidbr06_prep,UF=="Rio de Janeiro")
```

6.1.4. ETAPA 4: CALCULANDO OS INDICADORES A PARTIR DA BIBLIOTECA *CONVEY*

Para o cálculo da porcentagem de indivíduos abaixo da linha de pobreza e extrema pobreza, com e sem o auxílio emergencial, é necessário incorporar o desenho amostral, o que pode ser realizado utilizando a função *svyfgt* da biblioteca *convey* que estima o indicador de pobreza de Foster-Greer-Thorbecke (Foster et al., 1984). Este indicador foi construído com a possibilidade de ponderar a distância entre a renda dos indivíduos que vivem na pobreza e o limiar que define a pobreza. No entanto, ao utilizar a função *svyfgt* com argumento *g=0* estima-se a proporção da população abaixo da linha de pobreza sem considerar esta distância. O argumento *type_tresh=*"*abs*" indica que o limite para a situação de pobreza será um valor absoluto definido no argumento *abs_tresh* (Pessoa et al., 2020).

Nesta atividade os valores limites da renda para definir a proporção de pobreza e extrema pobreza serão de R$ 446 e R$ 154 conforme Monte (2020), que se baseia no valor proposto pelo Banco Mundial, realizando a conversão da moeda de acordo com a paridade do poder de compra inflacionado para o IPCA para o mês de junho de 2020. Outros limiares para definir a população abaixo da linha de pobreza podem ser discutidos, apresentando-se outras referências, como os das metas dos objetivos de desenvolvimento sustentável apresentados anteriormente ou a metodologia oficial no Brasil que, conforme Santos (2012) utiliza como referência o Salário Mínimo.

Os comandos a seguir implementam o cálculo da proporção de indivíduos abaixo da linha de pobreza em junho com e sem o Auxílio Emergencial no Brasil. Para o Rio de Janeiro este cálculo foi realizado segundo a localização geográfica. Para estimar a extrema pobreza deve-se apenas substituir o valor de corte na função *svyfg*t.

1 - Estimando a proporção de pessoas abaixo da linha de pobreza incorporando o desenho amostral com a biblioteca *convey* no Brasil para a população em geral em junho de 2020

`1.1` – Desconsiderando o auxílio emergencial

```
svyfgt(~renda_fam_per_sem,g=0,type_tresh="abs",abs_thresh=446,covidbr06_prep)
```

```
##                              fgt0       SE
## renda_fam_per_sem 0.35791 0.0022
```

1.2 – Considerando o auxílio emergencial

```
svyfgt(~renda_fam_per_com,g=0,type_tresh="abs",abs_thresh=446,covidbr6_prep)
```

```
##                              fgt0     SE
## renda_fam_per_com 0.21881 0.002
```

2 – Estimando a proporção de pessoas abaixo da linha de pobreza incorporando o desenho e peso amostral com a biblioteca *convey* no Rio de Janeiro segundo localização geográfica da moradia (Urbana/Rural) em junho de 2020

2.1 – Desconsiderando o auxílio emergencial

```
svyfgt(~renda_fam_per_sem,g=0,type_tresh="abs",abs_thresh=446,subset(covidrj6_prep,V1022=="Urbana"))
```

```
##                              fgt0     SE
## renda_fam_per_sem 0.29283 0.007
```

```
svyfgt(~renda_fam_per_sem,g=0,type_tresh="abs",abs_thresh=446,subset(covidrj6_prep,V1022=="Rural"))
```

```
##                              fgt0     SE
## renda_fam_per_sem 0.49091 0.034
```

2.2 – Considerando o auxílio emergencial

```
svyfgt(~renda_fam_per_com,g=0,type_tresh="abs",abs_thresh=446,subset(covidrj6_prep,V1022=="Urbana"))
```

```
##                              fgt0     SE
## renda_fam_per_com 0.18435 0.0057
```

```
svyfgt(~renda_fam_per_com,g=0,type_tresh="abs",abs_thresh=446,subset(covidrj6_prep,V1
022=="Rural"))
##                                fgt0      SE
## renda_fam_per_com 0.29906 0.0342
```

6.1.5. ETAPA 5: INTERPRETAÇÃO E COMUNICAÇÃO DOS RESULTADOS

Após as análises, o que se espera é que os relatórios dos diversos grupos e os seminários de apresentação demonstrem o nível de conhecimento obtido.

Em um primeiro nível da avaliação, deve-se verificar se os relatórios conseguiram apresentar, interpretar e fazer uma crítica dos resultados obtidos, como por exemplo, os apresentados na Tabela 1. Além da construção da tabela com os resultados principais, ao saberem descrever as porcentagens de pobreza e extrema pobreza com e sem o auxílio emergencial e fazerem comparações entre o Brasil e o Rio de Janeiro, discutindo o impacto do auxílio em populações com características distintas, os estudantes demonstrariam ter alcançado as habilidades necessárias ao letramento estatístico de acordo com o proposto por Gal (2002).

Tabela 1 - Proporção da população abaixo da linha da pobreza e extrema pobreza no Brasil e no estado do Rio de Janeiro em junho de 2020

	Indicador	Com o Auxílio Emergencial (%) (ep)	Sem o Auxílio Emergencial (%) (ep)
Brasil	Pobreza	21,88 (0,002)	35,79 (0,002)
	Extrema Pobreza	3,33 (0,000)	15,13 (0,001)
Rio de Janeiro	Pobreza	18,79 (0,006)	29,90 (0,007)
	Extrema Pobreza	3,60 (0,002)	12,77 (0,004)

Em um outro nível de avaliação, pode-se verificar o quanto as questões motivadoras no início da atividade e a leitura dos artigos e textos disponibilizados contribuíram para que o letramento estatístico alcançado pudesse ser considerado como o proposto por Weiland (2017), que o definiu como a habilidade de ler e escrever tanto a palavra quanto o mundo, como cidadãos críticos. Neste nível de letramento, outros fatores determinantes para a desigualdade no Brasil, tais como cor ou raça, gênero e região, que podem ser abordados adaptando-se os *scripts* apresentados neste artigo e no roteiro

http://davibalves.github.io/roteiro_artigo_rbe_celebrando_Djalma_Pessoa, poderiam ser explorados para ampliar o potencial da atividade, tanto nos aspectos técnicos (formulação da pergunta, planejamento das etapas da análise e sua implementação), quanto na formação cívica, ao explorarem e discutirem a pobreza por diferentes perspectivas. Estas reflexões junto com a familiaridade com os conceitos estatísticos, com a utilização de base de dados e bibliotecas do R, complementariam a formação estatística cívica.

Em um outro nível de profundidade, espera-se que os alunos sejam capazes de discutir os resultados relacionando-os com a questão de pesquisa e a opção metodológica utilizada, e que consigam avaliar a importância da incorporação do plano amostral nas estimativas e suas variabilidades. Espera-se ainda que percebam as diferentes possibilidades de cálculos da renda domiciliar e diferentes pontos de corte para definir pobreza e extrema pobreza, e o potencial de se avaliar outros indicadores para conhecer as desigualdades brasileiras. Ao alcançarem estas habilidades devem estar aptos a realizar outros estudos, com esta ou com outras bases de dados oficiais, e também mais próximos de uma formação estatística que leve em consideração uma cadeia de conhecimentos mais eficazes, conforme preocupações abordadas por Steel et al. (2019).

7. CONCLUSÃO

Considerando a importância da Educação Estatística para a ciência e para uma sociedade democrática, as reflexões aqui realizadas reforçam também as argumentações de Horton & Hardin (2015) a respeito da importância de uma nova geração de estatísticos, mas vão além disso e se aliam, por um lado aos que estão preocupados com os atuais tempos desafiadores e enfatizam que a responsabilidade social pode e deve fazer parte das preocupações no ensino-aprendizagem, buscando promover o engajamento cívico (Engel, 2014, 2017; Nicholson et al., 2018), e por outro, aos que consideram que os conceitos e práticas estatísticas podem ser abordados junto de questões sociais mais complexas (Weiland, 2017).

Aqui foi proposta uma adaptação ao modelo de Gal & Ograjenšek (2017) com vistas a uma formação estatística cívica. A atividade descrita pode ser aplicada durante uma disciplina em um curso de graduação ou de pós-graduação de Estatística a partir da inclusão de conceitos mais complexos, tanto em relação aos estimadores amostrais, quanto, por exemplo aos estimadores de parâmetros em um processo de modelagem. No entanto, pode ser adaptada para disciplinas de Introdução em cursos das mais variadas áreas, tais como as que utilizam o

desenvolvimento de projetos a partir de análise de dados com o programa R (Barbosa et al., 2019).

Seu planejamento foi baseado nas reflexões de Gal & Ograjenšek (2017) sobre a importância de formar pessoas que conheçam bem as estatísticas oficiais e considerou também as preocupações de Steel et al. (2019) que advogam a necessidade de apresentar aos estudantes como o pensamento estatístico pode ser desafiante, que erros comuns podem ser evitados e qual a responsabilidade dos cientistas, incluindo os estatísticos na comunicação correta dos resultados e incertezas das pesquisas.

Utilizou também como referência e inspiração as primeiras e mais importantes produções a respeito da necessidade de incorporação do Plano amostral nas estimativas (Pessoa et al., 1997; Silva et al., 2002) que levaram ao desenvolvimento das bibliotecas do R que permitem estimar corretamente a partir de considerar o plano amostral. Entre elas, a biblioteca *convey* (Pessoa et al., 2020) também mereceu destaque por permitir estimar variados indicadores de pobreza e desigualdade social para o Brasil e para vários estratos sociais.

Também foi relevante o uso de microdados de uma pesquisa recente e bem documentada, que abordou o grave momento sanitário, social e econômico que enfrentamos no país e permite um maior engajamento dos estudantes em conhecer sua própria realidade. Neste sentido, usando a mesma base de dados, é possível explorar outros aspectos relevantes como o desemprego, o desenvolvimento de sintomas da COVID e o acesso e uso de serviços de saúde, que são de interesse das mais diversas áreas.

Hoje existem inúmeras bibliotecas desenvolvidas para o programa R destinadas a facilitar o acesso e utilização dos microdados oriundos de diversas pesquisas do IBGE, como a PNADcIBGE (Braga et al., 2020) que permite a importação de dados da PNAD Contínua, PNSIBGE para dados da Pesquisa Nacional de Saúde (Assunção et al., 2020a) e a POFIBGE para dados da Pesquisa de Orçamentos Familiares (Assunção et al., 2020b) que podem e devem ser exploradas em atividades como a aqui proposta. Elas permitem a ampliação do número de usuários destas bases e um constante aprimoramento metodológico que, além de fortalecerem o Sistema de estatísticas oficiais, podem contribuir com o engajamento cívico decorrente de uma maior conscientização da realidade brasileira a partir dos seus indicadores.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Assunção, G., Hidalgo, L. (2020). COVIDIBGE: Downloading, Reading and Analysing PNAD COVID19 Microdata. Disponível em: <https://CRAN.R-project.org/package=COVIDIBGE>.

Assunção, G., Hidalgo, L., Braga, D. (2020a). PNSIBGE. Disponível em: <https://CRAN.R-project.org/package=PNSIBGE>

Assunção, G., Hidalgo, L., Braga, D. (2020b) PNSIBGE. POFIBGE. Disponível em: <https://CRAN.R-project.org/package=POFIBGE>

Barberino, M. R. B., Magalhães, M. N. (2016). Aprendizagem de Estatística por meio de projetos no Ensino Médio da escola pública. Educação Matemática Pesquisa, 18(3), 1223-1243.

Barbosa, M.T.S., Velasque, L.S., Silva, A. S. (2016) O letramento estatístico na formação dos professores; um tutorial metodológico. VIDYA, 36(2), 397-408.

Barbosa, M. T. S., Ross, S. D., Simões, B. F., Velasque, L. S., Silva, A. S., Assunção, M. B. M. & Tuttman, M. (2019). Educação estatística com formação cidadã em uma pesquisa quanti-ação no ensino superior. Revista de Ensino de Ciências e Matemática (RENCIMA), 10, 114-125.

Braga, D.; Assunção, G.; Hidalgo, L. (2020) PNADcIBGE. Disponível em: <https://CRAN.R-project.org/package=PNADcIBGE>

Campos, C. R, Coutinho, C. Q. S. (2019). Mathematical modelling and statistical literacy in teaching of graphs. Revista Eletrônica de Educação Matemática (REVEMAT), v.14, p.1-20. http:// doi.org/105007/1981-1322.

Engel, J. (2014). Open data, civil society and monitoring progress: challenges for statistics education. In ICOTS9. Disponível em: https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_4F4_ENGEL.pdf

Engel, J. (2017). Statistical literacy for active citizenship: a call for data science education. Statistics Education Research Journal, 16(1), 44-49.

Feijó, C., Valente, E. (2006). As estatísticas oficiais e o interesse público. In II Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais. Rio de Janeiro, IBGE.

Foster, J., Greer, J. & Thorbecke, E. (1984) A Class of Decomposable Poverty Measures. Econometrica, 52(3), 761–766.

Feijó, C. Valente, C (2005) As estatísticas oficiais e o interesse público. Bahia Análise & Dados, Salvador, 15(1), 43-54, jun. 2005. Disponível em: http://www.icad.puc-rio.br/cfeijo/pdf/artigofeijo_e_valente.pdf.

Gal, Iddo. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. International Statistical Review, 70,1-25.

Gal, I. & Ograjenšek, I. (2017). Official statistics and statistics education: Bridging the gap. Journal of Official Statistics, 33(1), 79-100.

Haeberlin, M.P., Silva, R.S. (2019). Erradicação da pobreza: contribuições do programa de transferência de renda bolsa família para o cumprimento do ODS1 (Objetivo de Desenvolvimento Sustentável 1) da agenda 2030 da ONU. Revista de Direitos Sociais, Seguridade e Previdência Social 5(2), 45 - 60.

Horton, N. J., Hardin, J. S. (2015). Teaching the Next Generation of Statistics Students to "Think With Data": Special Issue on Statistics and the Undergraduate Curriculum. The American Statistician, 69(4), 259–265.

IBGE. (2020a) Pesquisa Nacional por Amostra de Domicílios - PNAD COVID19. Disponível em: <https://www.ibge.gov.br/estatisticas/investigacoes-experimentais/estatisticas-experimentais/27946-divulgacao-semanal-pnadcovid1?t=o-que-e&utm_source=covid19&utm_medium=hotsite&utm_campaign=covid_19>

IBGE. (2020b) PNAD COVID19 : plano amostral e ponderação. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101726>

IBGE. (2020c) Relatório IBGE : pareamento de dados PNAD COVID19. Coleção Ibgeana. Rio de Janeiro: IBGE, Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101725>

IPEA (2018). ODS - Metas Nacionais dos Objetivos de Desenvolvimento Sustentável. Disponível em: <https://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/180801_ods_metas_nac_dos_obj_de_desenv_susten_propos_de_adequa.pdf>

ISLP – ProCivicStat (2018). Disponível em: <https://iase-web.org/islp/pcs/>

Jacob, G., Damico, A., Pessoa, D. (2020). Poverty and Inequality with Complex Survey Data. Disponível em: <https://guilhermejacob.github.io/context/index.html#introduction>

Kronemberger, D.M.P. (2019). Os desafios da construção dos indicadores ODS globais. Indicadores de Sustentabilidade, 40-45.

Lumley, T. (2004). Analysis of Complex Survey Samples. Journal of Statistical Software, 9(1), 1–19.

Lumley, T. (2020). Survey: analysis of complex survey samples.

Monte, P.A. (2020). Auxílio emergencial e seu impacto na redução da desigualdade e pobreza. In XXV Encontro Regional de Economia, Área 1 – Economia Regional, código JEL R11, H5332.

Moore, D.S. (1998). Statistics Among the Liberal Arts. Journal of the American Statistical Association, 93,1253–1259.

Nassif L.P., Carvalho, L., Rawet, L. (2020). Multi-dimensional inequality and Covid-19 in Brazil. Investigacion Economica, 80(315), 33-58.

Nicholson, J., Gal, I., & Ridgway, J. (2018). Understanding Civic Statistics: A Conceptual Framework and its Educational Applications. A product of the ProCivicStat Project. Retrieved (Date) from: http://IASE-web.org/ISLP/PCS.

Pessoa, D., Costa, A. (2013). Introdução à análise de dados amostrais complexos. In Seminário de Metodologia do IBGE - SMI 2013. Disponível em: <https://eventos.ibge.gov.br/images/smi2013/downloads/MC2/CursoMC2.pdf>.

Pessoa, D., Damico, A., Jacob, G. (2020). Convey: Estimation of indicators on social exclusion and poverty and its linearization, variance estimation. Disponível em: <https://github.com/DjalmaPessoa/convey/>.

Pessoa, D. G. C., Silva, P. N., Duarte, R. P. N. (1997). Análise Estatística de Dados de Pesquisas por Amostragem: Problemas no Uso de Pacotes-Padrão. Revista Brasileira de Estatística, 58(210), 53-76.

Pessoa, D., Silva, P. N. (2018a). *Análise de Dados Amostrais Complexos*. Disponível em: <https://djalmapessoa.github.io/adac/>

Pessoa, D., Silva, P. N. (2018b). Prefácio. In Análise de Dados Amostrais Complexos. Disponível em: <https://djalmapessoa.github.io/adac/>

Pessoa, D., Silva, P. N. (2018c). Laboratório de R do Capítulo 1. In Análise de Dados Amostrais Complexos. Disponível em: https://djalmapessoa.github.io/adac/

Santos, H.D. (2012). Da objetividade à objetivação: Conceitos, categorias e significados. Estatística e Sociedade, 2, 97-111.

Silva, P.L.N., Pessoa, D.G.N., Lila, M.F., (2002). Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência & Saúde Coletiva. 7(4), 659-670.

Steel, E.A., Liermann, P., Guttorp, P. (2019) Beyond Calculations: A Course in Statistical Thinking. The American Statistician, 73:sup1, 392-401, DOI: 10.1080/00031305.2018.1505657

United Nations. (2013). Fundamental Principles of Official Statistics. Disponível em: <https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf>

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. Educational Studies in Mathematics, 96(1), 33-47. https://doi.org/10.1007/s10649-017-9764-5

Revista Brasileira de Estatística, V.79, N. 246, p.64-74, janeiro-junho 2021

O FUTURO DO PACOTE *CONVEY*

Guilherme A. Pinheiro Jacob

guilhermejacob91@gmail.com

Escola Nacional de Ciências Estatísticas (ENCE)

Anthony J. Damico

ajdamico@gmail.com

Kaiser Family Foundation

Resumo: As contribuições do Professor Djalma Pessoa não se limitaram a desenvolvimentos teóricos. Uma de suas últimas contribuições foi o desenvolvimento do pacote *convey*, voltado para a estimação de medidas de desigualdade e pobreza com amostras complexas. Para homenagear seu trabalho, os autores deste artigo pretendem continuar melhorando o pacote. Este artigo descreve três possíveis adições ao pacote: (1) produção de estimativas do Efeito do Plano Amostral para estimadores de medidas de pobreza e desigualdade; (2) uso de Simulações de Monte Carlo e inclusão de seus resultados na documentação; e (3) melhora do manuseio de funções de influência, melhorando sua integração com o pacote *survey*.
Palavras-chave: desigualdade; pobreza, amostragem complexa; estatística computacional.

Abstract: Professor Djalma Pessoa's contributions were not limited to theoretical developments. One his last contributions was the development of the convey package, aimed at estimation of inequality and poverty measures with complex sample surveys. To honour his work, the authors will keep improving the package. This article describes three possible additions to the package: (1) the production of DEFF estimates for inequality and poverty measures estimators; (2) the use of Monte Carlo simulations and the inclusion of these results in the package documentation; and (3) to improve the handling of influence functions, for better integration with the survey package.
Keywords: inequality, poverty, complex sample surveys, computational statistics.

1. O ESTADO ATUAL DO PACOTE *CONVEY*

As contribuições do professor Djalma Pessoa para a estatística não se limitaram aos desenvolvimentos teóricos. Uma de suas últimas contribuições foi o desenvolvimento do pacote *convey* (Pessoa et al., 2020), um conjunto de algoritmos em linguagem R com a finalidade de estimar medidas de desigualdade e pobreza com amostras complexas. Os autores deste artigo são profundamente gratos por ter contribuído nesta tarefa, uma oportunidade que, sob as cuidadosas diretrizes do professor Djalma, nos permitiu aprimorar nosso entendimento sobre amostragem e estatística oficial.

Por volta de junho de 2020, os autores do pacote voltaram a discutir sobre possíveis melhorias. O presente artigo apresenta nossas discussões sobre o futuro do pacote *convey*. A Seção 2 descreve o estado atual do pacote, sua integração com o pacote *survey* (Lumley, 2004) e o paradigma estatístico adotado para estimação de variância em planos amostrais complexos. A Seção 3 discute o Efeito do Plano Amostral (EPA) e como ele pode ser útil tanto para o planejamento de pesquisas amostrais, quanto para a análise de dados amostrais. A Seção 4 explora o manuseio de funções de influência no pacote *convey* e discute como melhorar sua integração com o pacote *survey*.

O pacote *convey* é um *wrapper* para o pacote *survey* (Lumley, 2004, 2010), focado na estimação de medidas de desigualdade e pobreza, como o índice de Gini, os índices de Entropia Generalizada (Cowell, 1980; Cowell & Kuga, 1981a; b), a classe de índices de Foster et al. (1984), entre outros. O ponto central da inferência baseada no plano amostral é que as informações sobre o plano amostral devem ser consideradas em todas as análises. Por exemplo, a exclusão de observações de um conjunto de dados não seria particularmente problemática em uma amostra aleatória simples (AAS), mas pode mudar a contagem de unidades primárias de amostragem (UPA) em planos amostrais complexos.

O pacote *survey* foi desenvolvido para este tipo de análise utilizando R (R CORE TEAM, 2020), uma linguagem de estatística computacional baseada em programação orientada a objetos. O paradigma de inferências baseadas no plano amostral é implementado usando *objetos de plano amostral*, objetos que contém o conjunto de dados dos respondentes e as informações sobre o plano amostral, essenciais para os procedimentos de estimação. Procedimentos especiais para estimação de médias e totais, ajuste de modelos de regressão, estimação em domínios etc. são realizados a partir destes objetos, reduzindo a possibilidade de uso incorreto por parte do usuário. Embora haja certo custo ao

aprender suas funcionalidades e sintaxe, o pacote *survey* é uma importante ferramenta para o analista de pesquisas amostrais.

Os principais métodos para estimação de variâncias de funções não-lineares com planos amostrais complexos (Tillé, 2020; Wolter, 2007) podem ser divididos em dois grupos: analíticos e por reamostragem. Métodos analíticos se baseiam em fórmulas que aproximam a variância (assintótica) de um estimador não-linear. Por outro lado, métodos de reamostragem se baseiam em pseudo-replicações do plano amostral, tomando as variâncias das estimativas destas réplicas para aproximar a variância da estimativa principal.

De um ponto de vista operacional, métodos analíticos requerem a solução de derivadas possivelmente não-triviais, enquanto métodos baseados em reamostragem podem ser computacionalmente custosos. Além disso, as propriedades dos métodos de reamostragem variam: por exemplo, o método *jackknife* pode gerar estimativas não-consistentes para variâncias de estimadores de quantis, enquanto o método de *Balanced Repeated Replications* gera estimativas consistentes para tais estimadores sob condições razoáveis (Shao, 1996, p. 214). Esta limitação decorre do fato de que o quantil é uma função não suave; como nota Rao (1997, p. 151), este problema se estende para funções baseadas em quantis, como a medida de polarização de Foster-Wolfson (Foster & Wolfson, 2009; Wolfson, 1994) e medidas de pobreza que utilizam linhas de pobreza baseadas na mediana.

Muitos destes procedimentos já estão disponíveis no pacote *survey*, mas alguns métodos específicos não estão. Algumas destas ferramentas que não estavam disponíveis envolvem estimadores não-lineares para medidas de desigualdade e pobreza. O pacote *convey* é uma tentativa de preencher esta lacuna, adicionando funções para estimar uma gama de medidas de desigualdade, pobreza e polarização.

Existem diversas publicações envolvendo a tarefa de estimar medidas de desigualdade sob planos amostrais complexos. Kovačević & Binder (1997) desenvolveram métodos de estimação para seis medidas de desigualdade e polarização sob a abordagem das equações estimadoras (Binder, 1991; Binder & Patak, 1994; Godambe & Thompson, 1986). Deville (1999) desenvolveu uma abordagem analítica baseada em funções de influência para estimadores não-lineares em planos amostrais complexos, usando o índice de Gini como um exemplo. Osier (2009) aplicou esta abordagem para estimação de diversas medidas de desigualdade e pobreza. Adicionalmente, Verma & Betti (2011) desenvolveram fórmulas de linearização para várias medidas de desigualdade

e pobreza. Como explicou Deville (1999), em geral, as fórmulas de linearização podem ser entendidas como funções de influência.

Uma característica importante do pacote *convey* é sua capacidade de lidar com linhas de pobreza relativas. Quando as medidas de pobreza se baseiam em linhas de pobreza relativas, estas linhas de pobreza também são estimativas. Considerando a classe de medidas de Foster et al. (1984) com linhas relativas, Preston (1995) mostrou que a variabilidade amostral da linha de pobreza deve ser considerada na variabilidade da própria medida de pobreza. Ignorar este fato geralmente resulta em estimativas demasiado conservadoras da variância. À primeira vista, esta conclusão pode parecer contraintuitiva, mas uma vez que a variável linearizada é descrita como a soma do componente da medida de pobreza considerando fixa a linha de pobreza e de outro componente decorrente da linha de pobreza relativa, é evidente que a variância desta soma é reduzida pela covariância (positiva) entre os componentes.

Enquanto Preston (1995) considerou este problema sob AAS, este método de amostragem raramente é utilizado na prática. Grandes pesquisas amostrais são realizadas utilizando algum tipo de plano amostral complexo, envolvendo probabilidades desiguais de seleção, ajustes de não-resposta, estratificação e/ou conglomeração (Cochran, 1977; Deaton, 2019; Kish, 1965). Berger & Skinner (2003) e Osier (2009) estenderam a abordagem de Preston (1995) para planos amostrais complexos.

No entanto, este método adiciona uma complicação operacional para análises de domínios: enquanto as linhas de pobreza costumam ser estimadas considerando toda a população, os domínios de análise são um subconjunto da população. Ou seja: mesmo que o analista tente estimar a medida de pobreza de uma cidade específica, a linha de pobreza relativa depende das rendas da população inteira, incluindo outras cidades; isto deve ser levado em consideração para a estimação da variância. O pacote *convey* resolve este problema mantendo separadamente as informações necessárias para estimação de medidas e de linhas de pobreza, permitindo maior flexibilidade na estimação para domínios.

2. EFEITO DO PLANO AMOSTRAL

O Efeito do Plano amostral (EPA) é um conceito central em amostragem, sendo uma ferramenta importante para o planejamento e análise de pesquisas amostrais, indicando o fator pelo qual um plano amostral

complexo aumenta ou diminui a variabilidade amostral em comparação com uma AAS.

Em suma, O EPA é a razão entre a variância de um estimador considerando o plano amostral efetivamente utilizado e a variância equivalente sob AAS (Kish, 1965, p. 162). Ou seja: o EPA é uma medida resumo do efeito do plano amostral sobre a precisão de uma estimativa. Em comparação com AAS, geralmente temos que: amostras estratificadas possuem menor variabilidade, com EPA < 1; por outro lado, amostras por conglomerado geralmente resultam em EPA > 1, em decorrência da correlação intraclasse. Para um plano amostral que envolve estratificação e conglomeração, o efeito da conglomeração costuma superar o da estratificação, resultando em EPA > 1. Esta medida é particularmente importante em pesquisas que investigam uma vasta gama de temas, já que o EPA pode variar de estimativa para estimativa; i.e., na mesma pesquisa, o EPA da altura média em metros dos respondentes é geralmente diferente do EPA do peso médio em quilogramas dos mesmos respondentes, por causa das interações entre o plano amostral e estas variáveis.

O EPA pode ser interpretado de duas maneiras. Para o planejamento de pesquisas amostrais, ele demonstra o impacto de um plano amostral sobre a eficiência; i.e., EPA = 2 implica que uma AAS com metade do tamanho de determinado plano amostral seria tão eficiente para determinada estimativa. Por causa de sua relação com a correlação intraclasse[1] (Kish, 1965, p. 163–164), o EPA pode também ser utilizado no planejamento de pesquisas futuras.

Do ponto de vista de análise de dados amostrais, o EPA[2] expressa o impacto de ignorar o plano amostral sobre a avaliação da precisão de uma estimativa; i.e., EPA = 2 implica que a estimativa "ingênua" de variância sob AAS subestima a variância "efetiva" da estimativa por um fator de 2. Esta é precisamente a diferença entre o método de Preston (1995) e os métodos de Deville (1999) e Berger & Skinner (2003).

O pacote *survey* já produz estimativas do EPA para vários estimadores, enquanto o pacote *convey* ainda não possui esta funcionalidade. Por causa de

1 Ou, como explicou Kish (1965), a taxa de homogeneidade "roh".

2 Como explicou Mukhopadhyay (2010, p. 30–37), o EPA do "planejamento" e da "análise" não são tecnicamente idênticos. O EPA do "planejamento" considera as variâncias propriamente ditas sob o plano amostral e sob AAS, enquanto o EPA da "análise" considera as estimativas das variâncias sob o plano amostral e AAS. Por esta razão, o EPA "analítico" também é chamado de Efeito da Má Especificação (do plano amostral). No entanto, esta distinção não é importante na prática, já que uma é estimada pela outra com viés aproximadamente nulo.

sua centralidade em pesquisas amostrais, seria útil adicionar esta funcionalidade para medidas de desigualdade e pobreza.

3. SIMULAÇÃO DE MONTE CARLO

A maioria dos métodos de estimação para estimadores não-lineares e não-regulares se baseiam em hipóteses de normalidade e aproximações assintóticas. É importante verificar o funcionamento destes métodos em diferentes populações e planos amostrais. Também seria importante entender como a não-normalidade da distribuição de renda afeta viés e variância dos estimadores.

Para avaliar o Erro Quadrático Médio (EQM) dos estimadores, pode-se recorrer aos experimentos de Monte Carlo. Em suma, isso significa simular uma população finita, dela extrair diversas amostras e, finalmente, avaliar o viés e variância da distribuição de estimadores. Estas simulações também permitem avaliar a probabilidade de cobertura dos intervalos de confiança; i.e., avaliar se a cobertura do intervalo de confiança é suficientemente próxima da taxa de cobertura nominal. Este método é bastante flexível: por exemplo, o pesquisador pode variar o tamanho da população, modelo da distribuição de renda, plano amostral, tamanho de amostras etc.

Isto permite que o analista tenha uma ideia melhor sobre a confiabilidade dos procedimentos de estimação adotados no pacote *convey*. Consequentemente, seria possível decidir se funções atualmente classificadas como experimentais deveriam ser mantidas ou não. No entanto, este procedimento é computacionalmente pesado, devendo ser realizado apenas quando o código de uma função é substancialmente modificado. Além disso, os resultados deste procedimento devem ser descritos de maneira simples, de modo que o usuário médio possa entendê-los e tomar uma decisão informada acerca da viabilidade do método.

Chang et al. (1992) explicaram como simulações de Monte Carlo podem ser um recurso interessante para compreender planos amostrais. Mais recentemente, Goga & Ruiz-Gazen (2010) descreveram como implementar este método para amostras complexas nos ambientes R e SAS. Morris et al. (2019) forneceram diretrizes sobre descrição de procedimentos de simulação, gerenciamento de recursos computacionais e comunicação eficaz dos resultados da simulação.

Diante disso, pretende-se adicionar os resultados de simulações de Monte Carlo dos estimadores implementados à documentação do pacote *convey*, de forma que o usuário possa avaliar qual estimador utilizar.

4. FUNÇÕES DE INFLUÊNCIA

Como explicou Tillé (2020 Cap. 15), o termo “linearização” reúne uma variedade de métodos, incluindo equações estimadoras (Binder, 1991), funções de influência (Deville, 1999), e outras variantes (Binder, 2004; Demnati & Rao, 2004; Graf, 2011).

Recentemente, a nova versão do pacote *survey* adiciona várias melhorias, incluindo melhorias no manuseio de funções de influência. A consequência mais importante desta adição é que ela permite estimar matrizes de variância-covariância para estimativas em diversos domínios usando a função *svyby*. Isso implica, por exemplo, que o usuário pode considerar as covariâncias em pesquisas de painéis rotativos, produzindo estimativas de variações com maior precisão. Embora esta integração não esteja disponível no momento de publicação da versão mais recente, o pacote *convey* já havia começado a carregar as funções de influência nos objetos de estimativa. Portanto, em tese, esta integração seria mais fácil de ser realizada.

Além da estimação de variâncias, funções de influência têm outros usos analíticos. Cowell & Victoria-Feser (1996) usaram funções de influência para entender a robustez de medidas de desigualdade a outliers, demonstrando que estas medidas podem ser muito sensíveis à contaminação de dados nas caudas das distribuições de renda. Cowell & Victoria-Feser (2003) também as usaram para derivar variâncias assintóticas com dados censurados ou truncados.

Outra possibilidade é indicada pela regressão de funções de influência recentralizadas, discutidas por Firpo et al. (2009) e Essama-Nssah & Lambert (2011). Esta técnica permite entender a correlação entre uma medida de desigualdade ou pobreza e covariáveis usando regressão. A tarefa de modelar uma função de influência costuma ser realizada através da modelagem indireta da variável subjacente sobre a qual a medida de desigualdade foi calculada; i.e., a modelagem da função de influência do índice de Gini da desigualdade de salários é feita modelando a distribuição de salários. Isto abre uma possibilidade interessante para abordagens *model-assisted*, considerando sua robustez à má especificação do modelo (Pessoa & Silva, 2018; Pfeffermann, 1993, 2011). Firpo et al. (2018) também propuseram uma extensão da decomposição Oaxaca-Blinder para estatísticas não-lineares usando funções de influência recentralizadas.

Assim, as funções implementadas no pacote terão o argumento *influence*, indicando a opção de salvar as estimativas das funções influência no objeto de estimação *cvystat* resultante. Isto permitiria pronta integração com o

pacote *survey*, além de permitir que o usuário acesse estas estimativas facilmente, caso tenha interesse em usá-las em alguma análise específica. A opção também tornaria mais eficiente o consumo de memória, já que os usuários que não estão interessados nestas estimativas não guardariam uma informação desnecessária para a análise pretendida na memória do computador.

5. CONCLUSÃO

Desde seu lançamento, o pacote *convey* passou a ser amplamente utilizado por pesquisadores de diversas áreas, ensejando interações proveitosas entre amostristas, economistas e econometristas aplicados. Neste artigo, apresentamos um plano de possíveis desenvolvimentos.

A versão atual aponta para três possibilidades principais: (1) estimação de EPAs; (2) simulações de Monte Carlo; e (3) manuseio de funções de influência.

Estimativas do EPA podem ser úteis no planejamento e na análise de pesquisas amostrais. No entanto, o pacote *convey* não produz estimativas do EPA para estimadores de medidas de desigualdade e pobreza. Diante da importância desta informação, esta funcionalidade seria uma importante adição ao pacote.

Medidas de pobreza e desigualdade são geralmente funções não-lineares das observações na população, resultando em estimadores que são também funções não-lineares das observações na amostra. Simulações de Monte Carlo são um método útil para avaliar o EQM, viés e variância de tais estimadores sob uma variedade de planos amostrais. Esta ferramenta poderá ajudar os usuários a escolher qual plano amostral ou estimador utilizar em suas aplicações. Este método também é recurso interessante para aprender sobre amostragem, produzindo diversos exemplos de conceitos como viés, consistência e variância amostral.

Funções de influência já fazem parte do pacote *convey*. No entanto, elas não estão completamente integradas ao pacote *survey*. Melhorias no manuseio de funções de influência são um desenvolvimento importante, já que elas apresentam possibilidades analíticas tanto para a estimação de variâncias, quanto para analisar robustez/sensibilidade de medidas de desigualdade e pobreza.

Para homenagear o trabalho do professor Djalma Pessoa, pretendemos continuar melhorando o pacote *convey* com a adição das funcionalidades aqui descritas.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Berger, Y. G.; Skinner, C. J. (2003). Variance estimation for a low income proportion. Journal of the Royal Statistical Society: Series C (Applied Statistics), 52(4), 457–468.

Binder, D. A. (1991). Use of Estimating Functions for Interval Estimation from Complex Surveys. Proceedings of the Survey Research Methods Section. Anais…Chicago: American Statistical Association.

___ (2004). Linearization methods for single phase and two-phase samples: A cookbook approach. Survey Methodology, 22(1), 17–22.

Binder, D. A.; Patak, Z. (1994). Use of Estimating Functions for Estimation from Complex Surveys. Journal of the American Statistical Association, 89(427), 1035–1043.

Chang, T. C.; Lohr, S. L.; Mclaren, C. G. (1992). Teaching Survey Sampling Using Simulation. The American Statistician, 46(3), 232–237.

Cochran, W. G.(1977). *Sampling Techniques*. New York: John Wiley & Sons.

Cowell, F. A. (1980). Generalized entropy and the measurement of distributional change. European Economic Review, 13(1), 147–159.

Cowell, F. A.; Kuga, K. (1981a). Inequality measurement: An axiomatic approach. European Economic Review, 15(3), 287–305.

Cowell, F. A.; Kuga, K. (1981b). Additivity and the entropy concept. Journal of Economic Theory, 25(1), 131–143.

Cowell, F. A.; Victoria-Feser, M.-P. (1996). Robustness Properties of Inequality Measures. Econometrica, 64(1), 77–101.

Cowell, F. A.; Victoria-Feser, M.-P. (2003). Distribution-Free Inference for Welfare Indices under Complete and Incomplete Information. The Journal of Economic Inequality, 1(3), 191–219.

Deaton, A. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Washington, D.C.: World Bank, 2019.

Demnati, A.; Rao, J. N. K.(2004). Linearization variance estimators for survey data. Survey Methodology, 30(1), 17–26.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology, 25(2), 193–203.

Essama-Nssah, B.; Lambert, P. J. (2011). Influence functions for distributional statistics. [s.l.] ECINEQ, Society for the Study of Economic Inequality. Disponível em: <https://ideas.repec.org/p/inq/inqwps/ecineq2011-236.html>.X

Firpo, S. P.; Fortin, N. M.; Lemieux, T. (2009). Unconditional Quantile Regressions. Econometrica, 77(3), 953–973.

Firpo, S. P.; Fortin, N. M.; Lemieux, T. (2018). Decomposing Wage Distributions Using Recentered Influence Function Regressions. Econometrics, 6(2).

Foster, J. E.; Wolfson, M. C. (2009). Polarization and the decline of the middle class: Canada and the U.S. The Journal of Economic Inequality, 8(2), 247–273.

Foster, J.; Greer, J.; Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. Econometrica, 52(3), 761–766.

Godambe, V. P.; Thompson, M. E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review / Revue Internationale de Statistique, 54(2), 127–138.

Goga, C.; Ruiz-Gazen, A. (2010). The use of Monte Carlo simulations in teaching survey sampling (C. Reading, Ed.)Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8). Anais…Voorburg, The Netherlands: International Association for Statistical Education, International Statistical Institute.

Graf, M. (2011). Use of Survey Weights for the Analysis of Compositional Data. *In*: Compositional Data Analysis. [s.l.] John Wiley & Sons, Ltd, 114–127.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kovačević, M.; Binder, D. A. (1997). Variance Estimation for Measures of Income Inequality and Polarization - the Estimating Equations Approach. Journal of Official Statistics, 13(1), 41–58.

Lumley, T. (2004). Analysis of Complex Survey Samples. Journal of Statistical Software, 9(1), 1–19.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Hoboken, New Jersey: John Wiley & Sons.

Morris, T. P.; White, I. R.; Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074–2102.

Mukhopadhyay, P. (2010). *Complex Surveys: Analysis of Categorical Data*. Singapore: Springer Singapore.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. Survey Research Methods, 3(3), 167–195.

Pessoa, D.; Damico, A.; Jacob, G. (2020). convey: Estimation of indicators on social exclusion and poverty and its linearization, variance estimationComprehensive R Archive Network - CRAN.

Pessoa, D.; Silva, P. L. N. (2018). *Análise de Dados Amostrais Complexos*.

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. International Statistical Review / Revue Internationale de Statistique, 61(2), 317–337.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? Survey Methodology, 37(2), 115–136.

Preston, I. (1995). Sampling Distributions of Relative Poverty Statistics. Journal of the Royal Statistical Society. Series C (Applied Statistics), 44(1), 91–99.

Rao, J. N. K. (1997). Resampling Methods for Complex Survey Data. (C. Malaguerra, S. Morgenthaler, & E. Ronchetti, Eds.)Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth. Anais…Basel: Birkhäuser Basel, 1997.

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Shao, J. (1996). Resampling Methods in Sample Surveys. Statistics, 27(3:4), 203–237.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations*. Hoboken, NJ: John Wiley & Sons.

Verma, V.; Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. Journal of Applied Statistics, 38(8), 1549–1576.

Wolfson, M. C. (1994). When Inequalities Diverge. The American Economic Review, 84(2), 353–358.

Wolter, K. *Introduction to Variance Estimation*. 2. ed. New York: Springer, 2007.

Revista Brasileira de Estatística, V.79, N. 246, p.75-101, janeiro-junho 2021

COMPARAÇÃO DE MÉTODOS DE AMOSTRAGEM APLICÁVEIS A PESQUISAS PARA ÍNDICES DE PREÇOS AO CONSUMIDOR

Leiliane da Silva Oliveira

leilianeoliveira.est@gmail.com

Pedro Luis do Nascimento Silva

pedronsilva@gmail.com

Escola Nacional de Ciências Estatísticas

Resumo: A estimação (cálculo) de Índices de Preços ao Consumidor (IPCs) é geralmente realizada a partir de amostras de locais de compra e de produtos, podendo tais amostras serem consideradas como provenientes de uma população bidimensional. A Amostragem de Conglomerados (AC) é comumente usada nas pesquisas conduzidas para estimação de IPCs. A Amostragem Bidimensional (AB) pode ser uma alternativa conveniente em algumas situações de pesquisa. No entanto, essa metodologia (AB) é por vezes confundida com a Amostragem Conglomerada, é pouco difundida, tem literatura escassa e poucos estudos empíricos sobre a variância de seus estimadores. Esse trabalho tem como objetivo comparar a Amostragem Bidimensional e a Amostragem Conglomerada no contexto da seleção de amostras para estimação de índices de preços ao consumidor. A fonte de dados deste trabalho é proveniente do projeto *World Food Programme* e utilizou-se o *software* R. A comparação entre os planos amostrais foi realizada calculando a variância do estimador do IPC a cada mês. Para isso, foram obtidas expressões aproximadas da variância do estimador do Índices de Preços de Laspeyres para cada tipo de amostragem.
Palavras-chave: Amostragem Bidimensional; Amostragem Conglomerada; Índices de Preços ao Consumidor.

Abstract: The estimation (calculation) of Consumer Price Indices (CPIs) is usually performed from outlet and product samples, which can be considered samples from two-dimensional populations. Cluster Sampling (CS) is commonly used in surveys conducted for the estimation of CPIs. Two-Dimensional Sampling - or Cross-Classified Sampling - (CCS) may be convenient in some survey situations. However, this methodology (CCS) is sometimes confused with Cluster Sampling, it is not widespread, has scarce literature and few empirical studies on the variance of its estimators. This work aims to compare Two-Dimensional Sampling and Cluster Sampling in the context of sample selection for estimating consumer price indices. The data source used for this work comes from the World Food Program project and we use the R software. The comparison between the sampling plans is performed calculating the variance of the CPI estimator for each month. For this, approximate expressions for the variance of the Laspeyres Price Index estimator were obtained for each type of sampling.
Keywords: Cluster Sampling; Cross-Classified Sampling; Consumer Price Indices.

1. INTRODUÇÃO

Índices de Preços ao Consumidor (IPCs) são medidas econômicas que representam as variações médias dos preços de uma determinada cesta de produtos. No Brasil, o Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA), são insumos importantes para as negociações de reajuste salarial e ajustes contratuais (Fava, 2007).

Para Bethlehem (2009), população bidimensional é aquela em que um elemento pode ser identificado por uma sequência de dois rótulos identificadores, um para cada dimensão. No caso das pesquisas para produção de IPCs, as dimensões em questão são os locais de compra e os produtos.

No Brasil, a pesquisa que levanta os preços para cálculo do INPC e IPCA emprega amostras de locais de compra e de produtos (IBGE, 2013), podendo tais amostras serem consideradas como extraídas por Amostragem Conglomerada.

Na Suécia se utiliza Amostragem Bidimensional nas pesquisas para elaboração dos IPCs. No entanto, esse tipo de amostragem é por vezes confundido com a Amostragem Conglomerada (Juillard, 2016) no momento da estimação de variâncias. O motivo principal da confusão é que na Amostragem Bidimensional, a seleção de unidades numa população bidimensional é conduzida de forma independente ao longo das linhas e colunas. Na Amostragem Conglomerada (em 2 estágios), considerando a seleção de linhas como primeiro estágio, a seleção de colunas é feita condicionalmente dentro das linhas selecionadas no primeiro estágio. Esta diferença de planejamento tem implicações na variância dos estimadores, e a confusão feita na hora da estimação de variâncias é motivada pela facilidade em contar com ferramentas prontas em vários pacotes estatísticos (R, SAS, STATA etc.) para estimação no caso da Amostragem Conglomerada. Facilidades similares para lidar com a Amostragem Bidimensional não estão disponíveis.

No Brasil nunca foi feita análise da possibilidade de empregar Amostragem Bidimensional para a pesquisa que levanta os preços para obtenção do INPC e IPCA. Assim, o objetivo deste trabalho é comparar a Amostragem Bidimensional e a Amostragem Conglomerada no contexto da seleção de amostras para estimação desses IPCs.

Para comparar os dois planos amostrais no contexto do IPC brasileiro foi necessário desenvolver expressões para a variância do estimador do índice sob os dois planos. Este desenvolvimento foi feito usando linearização de Taylor, e

inspirado na descrição desse método contida em Pessoa e Silva (1998). Djalma Pessoa foi grande incentivador do emprego dos melhores métodos estatísticos no IBGE, e este é um dos exemplos de pesquisa inspirado por suas ideias e contribuições.

Este artigo está dividido em mais cinco seções além desta primeira. Na seção 2, apresentamos o contexto de pesquisas para estimação de Índices de Preços ao Consumidor (IPCs). Na seção 3 tratamos das ideias de Amostragem relevantes para esse contexto, e introduzimos a Amostragem Bidimensional. Na seção 4 apresentamos a metodologia e dados que usaremos para comparar o desempenho dos métodos de amostragem considerados no artigo. Na seção 5 apresentamos os resultados da comparação realizada, incluindo análises de eficiência relativa e partições da variância total. A seção 6 destaca os principais achados e sinaliza algumas possibilidades de trabalhos futuros.

2. ÍNDICES DE PREÇOS AO CONSUMIDOR

A variação dos IPCs é acompanhada por diversos segmentos da sociedade (governo, empresas, consumidores). Um IPC representa a variação média dos preços dos produtos de uma cesta de produtos e serviços especificada.

Os IPCs são estimados com base em dados amostrais provenientes de uma Amostra Conglomerada. No Brasil, foi aplicada a Pesquisa de Locais de Compra (PLC) aos domicílios selecionadas na Pesquisa de Orçamentos Familiares (POF). A partir dessa pesquisa, foi criado o cadastro de locais de compra a serem posteriormente selecionados para a amostra de locais onde é feita a coleta dos preços. Nos locais selecionados, realiza-se o cadastramento de produtos comercializados no local que compõem a cesta padrão de consumo das famílias. Estes produtos então passam a compor o cadastro de produtos, de onde são selecionadas as amostras de produtos que têm seus preços coletados em cada local (IBGE, 2013). Esse modo de obtenção da amostra se assemelha à Amostragem Conglomerada em 2 estágios (AC2).

A coleta dos preços é realizada mensalmente, de forma contínua ao longo do mês. De forma geral, os índices de preços são uma média ponderada dos relativos dos preços dos produtos, com pesos dados em função da quantidade consumida num certo período de referência. O IBGE utiliza a metodologia de Laspeyres, onde a ponderação é realizada num período base, ou seja, usa pesos que refletem as quantidades consumidas no período base (IBGE, 2013).

A construção de cadastros de locais de compra e de produtos ser realizada através de pesquisas com domicílios selecionados na POF implica num processo de amostragem de locais (e depois, de produtos) que não corresponde exatamente ao método de Amostragem Conglomerada em 2 estágios. O principal motivo para isso é que o cadastro de locais não é exaustivo. Isto torna difícil a estimação dos erros amostrais das estimativas dos IPCs (Fava, 2007). Apesar disso, neste trabalho, tomamos a AC2 como o método 'padrão' contra o qual será comparada a alternativa da Amostragem Bidimensional.

É neste ponto relacionado à amostragem de locais e de produtos para a elaboração do índice de preços que esse trabalho se propõe a contribuir. Fazemos uma comparação de um processo de Amostragem Conglomerada com a Amostragem Bidimensional para a seleção de locais e produtos a terem os preços coletados.

2.1 ÍNDICE DE PREÇOS DE LASPEYRES CONSIDERANDO UM CADASTRO DE LOCAIS DE COMPRA E UM CADASTRO DE PRODUTOS

O índice de preços pode ser calculado simplesmente como uma média aritmética dos relativos de preços de produtos (Fisher,1922). Essa seção tem como objetivo alcançar a expressão de cálculo do índice de preços considerando os preços em nível de produto×local, ou seja, uma expressão que considere a variação de preços de produtos em cada local de compra.

Para medir variação de poder de compra, é recomendável atribuir pesos aos produtos de forma que reflitam sua importância nos gastos das famílias. Assim, o índice de preços pode ser escrito como uma média aritmética ponderada dos relativos de preços, onde a ponderação é dada de forma proporcional aos gastos com cada produto no período base.

A média aritmética ponderada é dada pela expressão (1) e os pesos pela expressão (2):

$$IPL_{t/0} = 100 \sum_{j=1}^{M} W_{0j} R_{tj} = 100 \sum_{j=1}^{M} W_{0j} \frac{P_{tj}}{P_{oj}} \quad (1)$$

$$W_{oj} = \frac{Q_{oj} P_{oj}}{\sum_{k=1}^{M} Q_{ok} P_{0k}} \quad (2)$$

onde P_{0j} é o preço do produto j no mês de referência (mês 0), P_{tj} é o preço do produto j no mês t, com $t = 0,1,2,\cdots$; Mé o número de produtos considerados e Q_{oj} é a quantidade consumida do produto j no mês de referência.

O índice de preços de Laspeyres corresponde ao índice de preço que reflete os gastos no período base, conforme a expressão (1) (Fisher, 1922). Após manipulações algébricas, nota-se que o Índice de Preços de Laspeyres é a razão

entre os gastos totais da família nos dois períodos, fixadas as quantidades consumidas no período base.

As expressões (1) e (2) consideram apenas os produtos. Neste trabalho será considerado o nível de agregação por locais de compra. Assim após algumas manipulações algébricas, descritas em Oliveira (2018), pode-se escrever o índice de preços de Laspeyres considerando um nível de agregação, os locais de compra:

$$IPL_{t/0} = 100\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{oij}P_{0ij}\left(\frac{P_{tij}}{P_{0ij}}\right)}{\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{oij}P_{0ij}} = 100\sum_{i=1}^{N}\sum_{j=1}^{M} W_{0ij}^{*}\left(\frac{P_{tij}}{P_{0ij}}\right) \quad (3)$$
$$= 100\sum_{i=1}^{N}\sum_{j=1}^{M} W_{oij}^{*}R_{tij}$$

onde M é o número de produtos considerados, N é o número de locais de compra, P_{0ij} é o preço do produto j no local de compra i no mês de referência (mês 0), P_{tij} é o preço do produto j no local de compra i no mês t, com $t = 0,1,2,\cdots$; Q_{oij} é a quantidade consumida do produto j no local de compra i no mês de referência e

$$W_{0ij}^{*} = \frac{Q_{0ij}P_{0ij}}{\sum_{h=1}^{N}\sum_{k=1}^{M} Q_{0hk}P_{0hk}} \quad (4)$$

é a estrutura de ponderação considerando as quantidades consumidas de cada produto em cada local de compra i e $R_{tij} = \frac{P_{tij}}{P_{0ij}}$ é o relativo dos preços, ou seja, a razão entre os preços do mês t e do mês de referência para cada produto j em cada local.

Portanto, considerando os relativos de preços em nível de produto×local, os pesos do índice de Laspeyres também dependem do local. E a expressão (3) considera a diferença de preços e de consumo de produtos em cada local de compra. O índice de preços pode ser visto como um total, ou seja:

$$IPL_{t|0} = 100\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{tij} \quad (5)$$

onde, $Y_{tij} = W_{0ij}^{*}R_{tij}$

As expressões (1), (3) e (5) estabelecem parâmetros populacionais, ou seja, expressam o índice de preços de uma população como um total populacional que poderia, em tese, ser calculado se fosse possível conduzir um censo da população para observar todos os valores Y_{tij}. No entanto, extrair informações populacionais pode ser uma tarefa demorada, cara ou até mesmo inviável. Então, pode-se buscar estimativas dos índices de preços ao consumidor através de amostras adequadas das populações de produto×local. A seção seguinte aborda métodos de amostragem que podem ser aplicados para a extração de amostras que produzam boas estimativas dos IPCs.

3. AMOSTRAGEM

O procedimento de amostragem tem como objetivo "obter informações sobre o todo, baseando-se no resultado de uma amostra" (Bolfarine & Bussab, 2005, p. 1). Essa amostra deve ser capaz de permitir a generalização de estimativas para toda a população, com o menor erro amostral, considerando o custo, tempo e restrições operacionais. Além disso, deve permitir a mensuração de precisão das estimativas.

Existem diversos planos amostrais, dentre eles os probabilísticos (processo de seleção aleatorizado) e os não-probabilísticos. Dentre os planos amostrais probabilísticos, uma série de características, quando agrupadas, formam diferentes planos. Essas características dizem respeito à probabilidade de seleção da unidade amostral (igual ou distinta), tipo de unidade amostral (elementar ou conjunto de unidades elementares), divisão da população em estratos, número de estágios de seleção, e forma de seleção das unidades amostrais (aleatória ou sistemática) (Bolfarine & Bussab, 2005).

Dentre as diversas técnicas de amostragem existentes, uma delas, a Amostragem Conglomerada (AC), é comumente usada em pesquisas para estimação de IPCs. A Amostragem Bidimensional (AB) pode ser uma alternativa conveniente à AC em algumas situações de pesquisa.

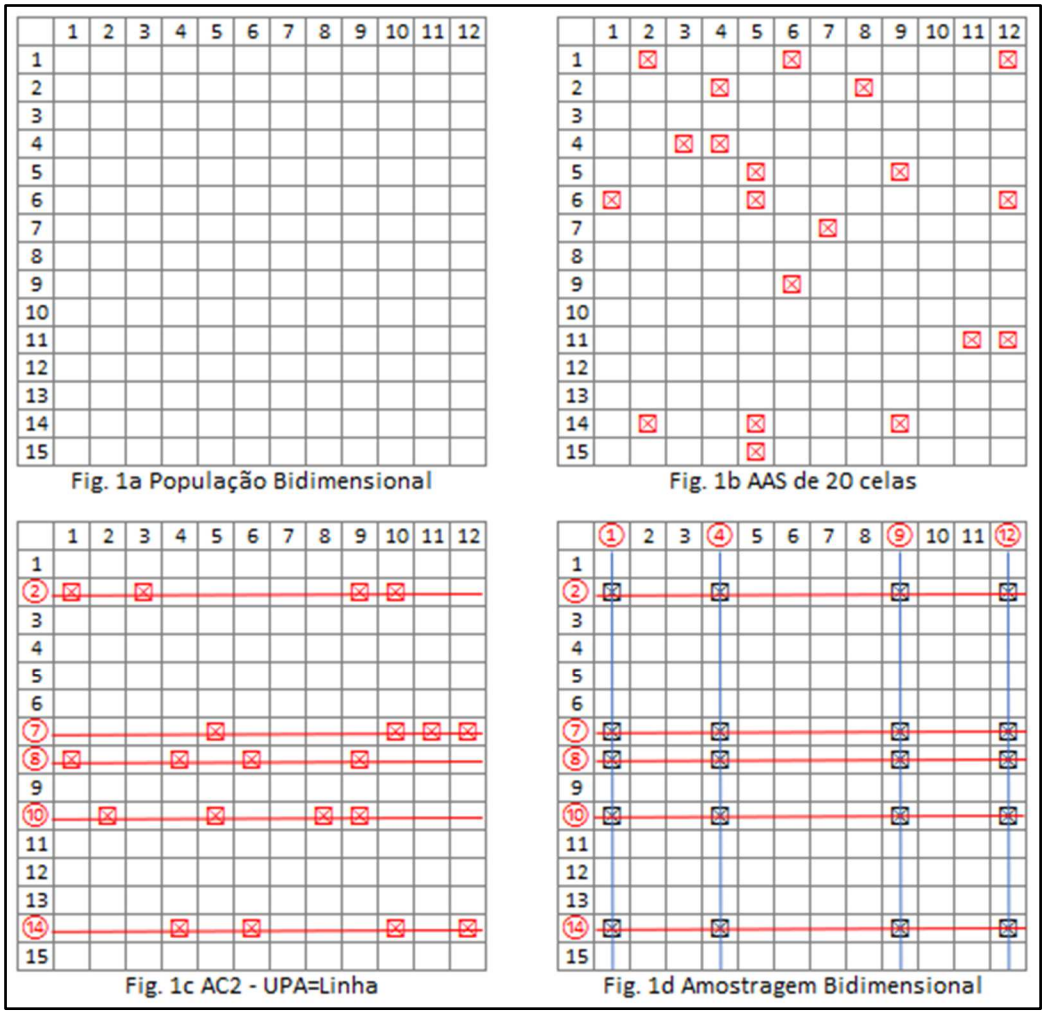
Juillard (2016) apresentou, de forma didática, três opções para amostragem em populações bidimensionais[1] (a população objetivo da pesquisa é classificada em duas dimensões, representadas por linhas e colunas). A primeira opção de amostragem é a seleção direta das observações nas celas, como se fosse selecionado um par de coordenadas que identificam a observação nas duas dimensões, como podemos observar pela Figura 1b.

A segunda opção é a seleção em dois estágios (Figura 1c), onde são selecionados conjuntos de observações (conglomerados, que no exemplo são as linhas) de uma dimensão e posteriormente, para esses conjuntos selecionados, selecionam-se as observações ao longo da segunda dimensão. A terceira opção é a Amostragem Bidimensional, onde a seleção em cada uma das duas dimensões é realizada de forma independente (Figura 1d). Posteriormente, as amostras das duas dimensões são cruzadas, gerando as celas que pertencem à amostra bidimensional[2].

[1] Bethlehem (2009), no seu capítulo 5, também apresenta as formas de amostragem em populações bidimensionais de forma bem simples.
[2] Neste trabalho, optamos por chamar a amostragem cruzada (CCS) por amostragem bidimensional.

Ohlsson (1996) definiu a Amostragem Bidimensional como o cruzamento das amostras independentes de linhas e de colunas, também chamada de Amostragem Cruzada, ou em inglês '*Cross-Classified Sampling*' (CCS).

Figura 1 - Ilustração de uma população bidimensional e das suas opções de amostragem



Fonte - Adaptado de Juillard (2016)

O procedimento de seleção de conglomerados pode ser realizado de acordo com um método de seleção qualquer. Após a seleção dos conglomerados, todas as unidades pertencentes ao mesmo são pesquisadas (Amostragem Conglomerada em 1 estágio), ou pode-se selecionar elementos dos conglomerados já pertencentes à amostra por algum método de amostragem (Amostragem Conglomerada em 2 ou mais estágios[3]).

No contexto das pesquisas para estimação dos IPCs do IBGE, os locais de compra podem ser considerados como conglomerados os quais são selecionados por algum método de amostragem. Num segundo estágio, em cada um dos locais de compra da amostra do primeiro estágio, os produtos são listados e depois selecionados por um método especificado. Isso configura uma

[3] Indicada quando as unidades dentro de cada conglomerado são parecidas entre si.

espécie de plano amostral similar à amostragem de conglomerados em dois estágios, com estratificação por área de pesquisa. Nessa abordagem, a seleção dos produtos a pesquisar é feita dentro (condicionalmente) de cada local de compra selecionado no primeiro estágio. Juillard (2016) ressaltou que é comum os usuários confundirem a Amostragem Bidimensional com a AC2.

Para a ilustração desse erro que provoca vieses na estimação de variância, Juillard (2016), utilizando dados da pesquisa francesa sobre infância, fez uma decomposição da variância, analogamente à decomposição pela ANOVA, para ambos os planos amostrais. Concluiu que a diferença está na quantidade de termos na decomposição da variância, o que é dado pelo efeito da independência entre as amostras de cada dimensão na Amostragem Bidimensional.

Utilizando os mesmos dados franceses, Juillard et al. (2017) apresentaram propostas de estimadores de variância sob o plano amostral bidimensional para estimadores de total do tipo Horvitz-Thompson.

No contexto de IPCs, a Amostragem Bidimensional seria o cruzamento de amostras de locais de compra com amostras de produtos. Dalén & Ohlsson (1995) apresentam expressões para estimadores de totais e correspondentes variâncias para Amostragem Bidimensional no contexto do IPC da Suécia. Norberg (2004) apresentou, de forma detalhada, como era feito o cálculo dos índices de preços desse país, e propôs um número adicional de estimadores para a variância.

Uma vantagem de AB é que não há a necessidade de um cadastro de celas (produtos×locais), apenas um cadastro de locais de compra e um outro, independente, de produtos. Além disso, conforme Bethlehem (2009,p. 125) é possível ter o controle sobre as amostras das duas dimensões. No caso do IPC, é possível ter controle sobre a amostra de locais e sobre a amostra de produtos, diferente da AC2, na qual o controle é exercido apenas sobre a amostra de locais de compra. Na AB o tamanho da amostra de produtos (que será o mesmo para cada local) pode ser planejado e previamente conhecido, permitindo assim produzir estatísticas a respeito dos produtos selecionados com uma precisão especificada.

Vale ressaltar que a grade de dados (cruzamento de linhas e colunas) não precisa ser completa, no entanto como as análises, neste trabalho, foram realizadas considerando conglomerados do mesmo tamanho, a grade é completa. O Quadro 1 apresenta essa comparação entre AC2 e AB de forma resumida.

Quadro 1 - Comparação entre as características de AC2 e AB

AC2	AB
Cadastro de locais de compra Cadastro de celas (produtos × locais) elaborado por amostragem, nos locais de compra selecionados	Cadastro de locais de compra independente do cadastro de produtos
Controle da amostra de locais de compra	Controle da amostra de locais de compra
	Controle da amostra de produtos
	Permite estatísticas de produtos específicos com precisão especificada

4. DADOS E CRITÉRIOS USADOS NA COMPARAÇÃO

A base de dados necessária para a realização deste trabalho deve ser composta pelo cruzamento de informações de preços por produtos e por locais de compra, onde o número de locais de compra e de produtos seja grande o suficiente para retirarmos amostras. Deve conter os preços de produtos em todos os locais e em todos os períodos de coleta, de forma que em cada local tenha informações de todos os produtos. A fonte de dados deste trabalho é proveniente do projeto *World Food Programme* (WFP), que disponibiliza seus dados na plataforma aberta de compartilhamento de dados *Humanitarian Data Exchange* (HDX).

É uma base com preços de 321 produtos e 1449 mercados (locais de compra). No entanto, nem todos os produtos existem em todos os mercados, e apesar dos dados terem uma periodicidade mensal existem meses em que os preços não foram coletados para determinados produtos, sendo necessário um tratamento de dados faltantes.

Para gerar a base de dados necessária para o desenvolvimento deste trabalho, o tratamento desses dados foi realizado no *software R Core Team* (2016). Envolveu etapas de recortes de elegibilidade, conversão dos preços para dólar, padronização das unidades de medida, análise de *outliers*, imputação e ponderação (de forma a gerar uma base completa que chamaremos de população)[4].

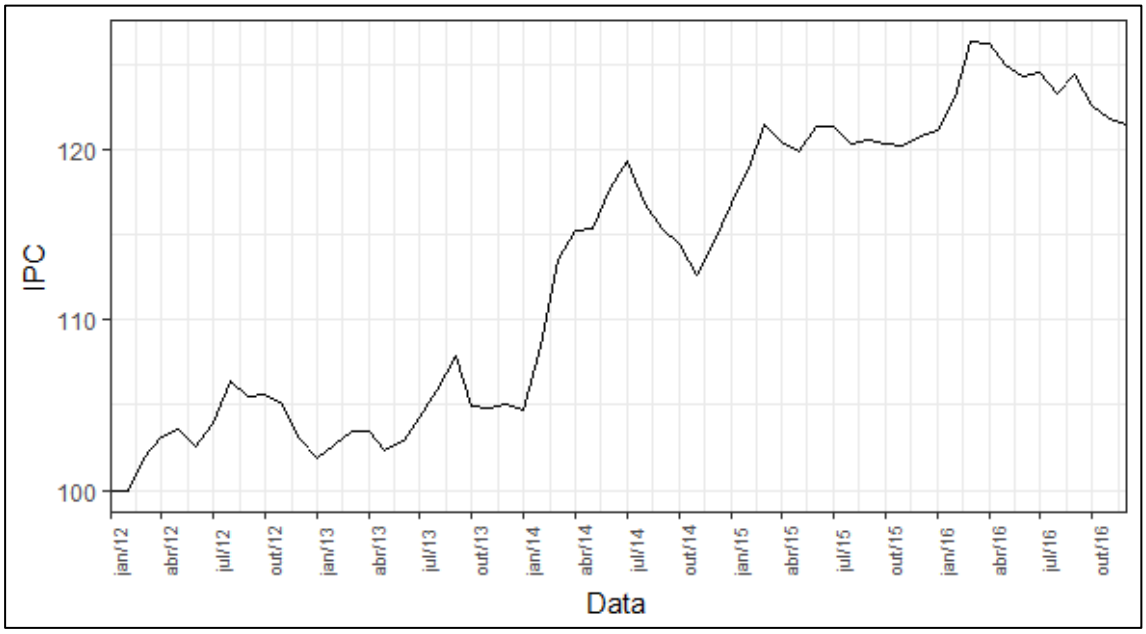
A base de preços final após os tratamentos descritos possui 15.606.420 registros que representam as 60 observações (as observações são mensais do

[4] Para mais detalhes sobre o tratamento da base de dados, consulte Oliveira (2018).

período de janeiro de 2012 a dezembro de 2016) de preços de cada um dos 263 produtos selecionados em cada um dos 989 locais de compra elegíveis.

Usando a expressão (5) foi calculado o índice de preços com os dados populacionais para todos os meses da base de dados, ou seja, para o período de janeiro de 2012 a dezembro de 2016. O Gráfico 1 apresenta o índice de preços com a variação de preços com base na população de preços elaborada para o trabalho.

Gráfico 1 - Índice de Preços ao Consumidor da População de Pesquisa para o período de 01/2012 a 12/2016



Fontes: WFP. Microdados do World Food Programme 2012-2016.

A comparação entre os planos amostrais AC2 e AB foi realizada calculando a variância do estimador do Índices de Preços ao Consumidor a cada mês. Para isso, foram obtidas expressões aproximadas da variância do estimador do Índices de Preços de Laspeyres para cada tipo de amostragem, que podem ser consultados no APÊNDICE A.

Com as expressões das variâncias aproximadas dos estimadores do IPC, calculamos os seus valores para os diferentes cenários de amostragem obtidos variando o método de amostragem (AC2 ou AB), o tamanho da amostra total, da amostra de produtos e da amostra de locais. Através do Coeficiente de Variação (CV), disponível no APÊNDICE B, verificamos em quais situações um plano amostral é mais indicado que o outro, visando estabelecer recomendações para o planejamento de amostras visando à estimação de IPCs.

5.COMPARAÇÃO EMPÍRICA DOS MÉTODOS

As variâncias do estimador de IPC para a amostragem conglomerada e para a amostragem bidimensional foram calculadas para alguns cenários de tamanhos amostrais. Considerando o tamanho amostral final de unidades elementares (m_0), foram analisadas as variâncias para as amostras com os seguintes tamanhos: 13.150, 26.300, 39.450, 52.600, 65.750 e 78.900. Em relação ao tamanho da amostra de produtos, foram criados cenários que variaram pela seleção de 100% dos produtos a 10 % dos produtos, ou seja, os tamanhos de m iguais a: 263, 237, 210, 184, 158, 132, 105, 79, 53, 26.

Para cada um dos valores de m_0, foram geradas combinações de tamanhos de amostra de locais de compra (n) e de produtos (m), onde $n \times m \simeq m_0$, de forma a permitir a comparação dos planos amostrais segundo m_0. Essa configuração permitiu avaliar o desempenho dos planos amostrais em relação ao tamanho final da amostra bem como avaliar o desempenho de acordo com a combinação de n e m para cada m_0.

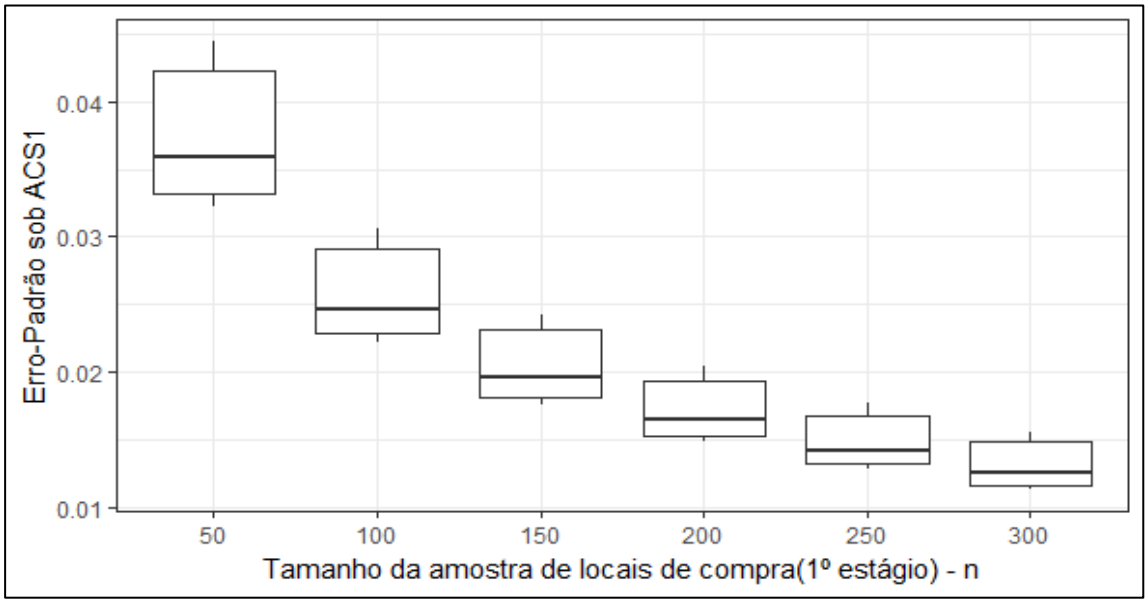
5.1 ANÁLISE DAS VARIÂNCIAS SOB AMOSTRAGEM CONGLOMERADA EM 1 ESTÁGIO

Começando as análises pela amostragem conglomerada simples em um estágio (ACS1), o Gráfico 2 apresenta a distribuição do erro-padrão do estimador de Horvitz-Thompson do índice de preços ao consumidor sob ACS1 para diversos tamanhos de amostras de conglomerados. Cada Boxplot deste gráfico representa a distribuição do erro-padrão do estimador para os 59 meses estudados (de 02/2012 a 12/2016)[5]. Observa-se que, à medida que o tamanho das amostras de locais de compra (n) aumenta, o erro-padrão do estimador diminui.

Além disso, pode-se observar que a distribuição dos erros-padrão para cada tamanho de amostra de conglomerados é mais homogênea para n maiores. Esse resultado está de acordo com a teoria, uma vez que se espera reduzir a variabilidade do estimador à medida que aumenta o tamanho amostral.

[5] O mês de janeiro de 2012 não entra na análise por ser o mês de referência do índice de preços ao consumidor.

Gráfico 2 - Distribuição, ao longo dos meses, do Erro-Padrão do Estimador de Horvitz-Thompson do IPC sob ACS1



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

5.2 COMPARAÇÃO DOS MÉTODOS DE AMOSTRAGEM CONGLOMERADA E BIDIMENSIONAL (POR MÊS)

Iniciando as análises de comparação dos planos amostrais por conglomerado e bidimensional, foram primeiramente analisados os resultados para um mês específico (02/2012), de forma a entender o comportamento de cada plano para cada cenário. A Tabela 1 apresenta os erros-padrões (EP_ACS2 e EP_ABS) e Coeficientes de Variação (%) (CV_ACS2 e CV_ABS) do estimador do índice de preços para o primeiro mês subsequente ao mês de referência do índice, que é o mês janeiro de 2012, para os planos amostrais ACS2 e ABS.

Tabela 1 - Comparação dos planos amostrais para o estimador do índice de preços de fevereiro de 2012 por tamanho amostral final (m_0)

(Continua)

m_0	n (1)	m	EP_ACS2	EP_AB	CV_ACS2 (%)	CV_AB (%)
13.150	50	263(2)	0,0323	0,0323	3,23	3,23
13.035	55	237	0,0373	0,1401	3,73	14,02
13.230	63	210	0,0414	0,2093	4,14	20,93
13.064	71	184	0,0455	0,2716	4,55	27,17
13.114	83	158	0,0490	0,3371	4,90	33,71
13.200	100	132	0,0521	0,4113	5,21	41,14
13.125	125	105	0,0555	0,5059	5,55	50,60

Tabela 1 - Comparação dos planos amostrais para o estimador do índice de preços de fevereiro de 2012 por tamanho amostral final (m_0)

(Continua)

m_0	n (1)	m	EP_ACS2	EP_AB	CV_ACS2 (%)	CV_AB (%)
13.114	166	79	0,0585	0,6290	5,85	62,91
13.144	248	53	0,0612	0,8200	6,12	82,00
13.156	506	26	0,0640	1,2431	6,40	124,33
26.300	100	263(2)	0,0222	0,0222	2,22	2,22
26.307	111	237	0,0257	0,1381	2,57	13,81
26.250	125	210	0,0289	0,2080	2,89	20,80
26.312	143	184	0,0316	0,2706	3,16	27,06
26.228	166	158	0,0342	0,3363	3,42	33,63
26.268	199	132	0,0366	0,4106	3,66	41,07
26.250	250	105	0,0389	0,5054	3,89	50,55
26.307	333	79	0,0410	0,6286	4,10	62,86
26.288	496	53	0,0430	0,8196	4,30	81,97
39.450	150	263(2)	0,0176	0,0176	1,76	1,76
39.342	166	237	0,0206	0,1375	2,06	13,75
39.480	188	210	0,0232	0,2075	2,32	20,76
39.376	214	184	0,0255	0,2703	2,55	27,03
39.500	250	158	0,0276	0,3360	2,76	33,60
39.468	299	132	0,0295	0,4104	2,95	41,05
39.480	376	105	0,0314	0,5052	3,14	50,53
39.421	499	79	0,0332	0,6284	3,32	62,85
39.432	744	53	0,0348	0,8195	3,48	81,96
52.600	200	263(2)	0,0148	0,0148	1,48	1,48
52.614	222	237	0,0174	0,1371	1,74	13,71
52.500	250	210	0,0198	0,2073	1,98	20,73
52.624	286	184	0,0217	0,2701	2,17	27,01
52.614	333	158	0,0236	0,3359	2,36	33,59
52.536	398	132	0,0253	0,4103	2,53	41,04
52.605	501	105	0,0270	0,5051	2,70	50,52
52.614	666	79	0,0285	0,6283	2,85	62,84
65.750	250	263(2)	0,0128	0,0128	1,28	1,28
65.649	277	237	0,0152	0,1369	1,52	13,69
65.730	313	210	0,0173	0,2072	1,73	20,72

Tabela 1 - Comparação dos planos amostrais para o estimador do índice de preços de fevereiro de 2012 por tamanho amostral final (m_0)

(Conclusão)

m_0	n (1)	m	EP_ACS2	EP_AB	CV_ACS2 (%)	CV_AB (%)
65.688	357	184	0,0192	0,2700	1,92	27,00
65.728	416	158	0,0208	0,3358	2,08	33,58
65.736	498	132	0,0224	0,4102	2,24	41,03
65.730	626	105	0,0239	0,5051	2,39	50,51
65.728	832	79	0,0253	0,6283	2,53	62,84
78.900	300	263(2)	0,0113	0,0113	1,13	1,13
78.921	333	237	0,0135	0,1368	1,35	13,68
78.960	376	210	0,0155	0,2071	1,55	20,71
78.936	429	184	0,0172	0,2699	1,72	27,00
78.842	499	158	0,0188	0,3357	1,88	33,58
78.936	598	132	0,0202	0,4102	2,02	41,02
78.855	751	105	0,0216	0,5050	2,16	50,51

Fontes: WFP. Microdados do *World Food Programme* 2012-2016.
Nota: (1) As linhas onde a combinação de n e m resultava em n>N>989 foram suprimidas.
(2) Quando m=263, o plano é na verdade ACS1 com sorteio apenas de locais.

Cada bloco da Tabela 1 mostra os resultados dos cenários para cada valor do tamanho total da amostra (m_0). Observa-se que o tamanho final da amostra sofre pequenas variações devido à combinação de valores inteiros de n e m, mas os tamanhos são próximos ao m_0 inicialmente estabelecido. Comparando o CV para m=263, que corresponde à ACS1, com os CVs dos demais tamanhos de amostra de produtos, uma comparação entre os planos ACS1 e ACS2 está sendo feita. Pode-se então, observar que em todos os cenários ACS2 é menos eficiente que ACS1. Isso acontece pois em ACS1 não há o componente de variância proveniente dos produtos como em ACS2, e consequentemente a variância do estimador em ACS1 é sempre menor do que a variância do estimador em ACS2.

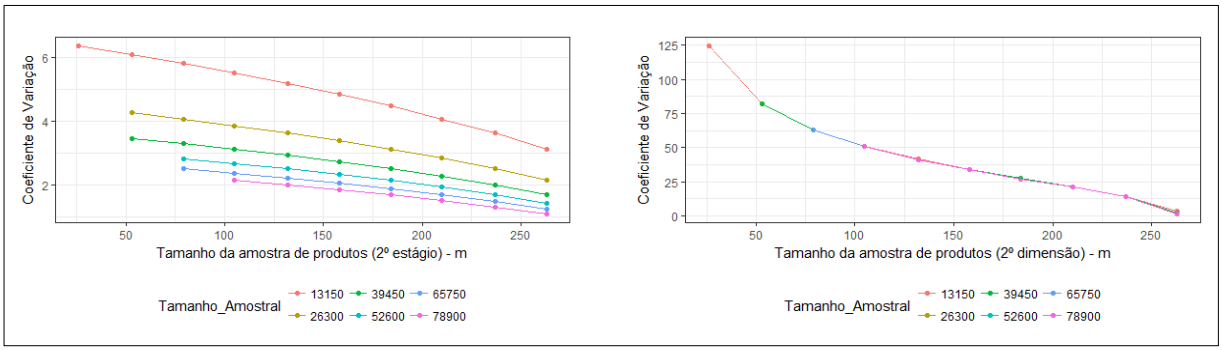
Observando os resultados para o cenário de m_0=13.150 (as 10 primeiras linhas), pode-se notar que a Amostragem Conglomerada em dois estágios (ACS2) é mais eficiente do que a Amostragem Bidimensional (AB), pois seus coeficientes de variação são sempre bem menores que os da AB. A mesma conclusão é obtida para os demais valores de m_0.

Fazendo uma análise comparativa entre os resultados de cada cenário de m_0, pode-se observar que, para ACS2, em todas as combinações de n e m, os estimadores são mais precisos à medida que m_0 aumenta. Por exemplo, observando os cenários onde m=132, que corresponde a aproximadamente 50% dos produtos, o erro-padrão de ACS2 diminui à medida em que mais locais de compra são selecionados (que é o mesmo que aumentar o tamanho total m_0, pois se mantém a relação de $m_0 = n \times m$).

Já para AB não há mudanças na precisão do estimador à medida que m_0 aumenta. Novamente, observando o cenário onde m=132, à medida em que mais locais de compra são selecionados, o valor do erro-padrão se mantém praticamente o mesmo.

O Gráfico 3 mostra o coeficiente de variação do estimador para o mês de junho de 2014. Basicamente traz a mesma informação da Tabela 1, porém de forma visual e do estimador do índice de alguns meses à frente do mês de referência. Neste gráfico, pode-se observar, novamente, que o estimador é mais preciso à medida que m_0 aumenta, para o caso de ACS2, independente da combinação de n e m. Bem como, é mais preciso à medida em que mais produtos são selecionados dentro de cada cenário de m_0. Para ABS há mudanças na precisão do estimador apenas quando mais produtos são selecionados, independentemente do número de locais de compra selecionados.

Gráfico 3 - Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), segundo o tamanho amostral final (m_0), junho de 2014



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

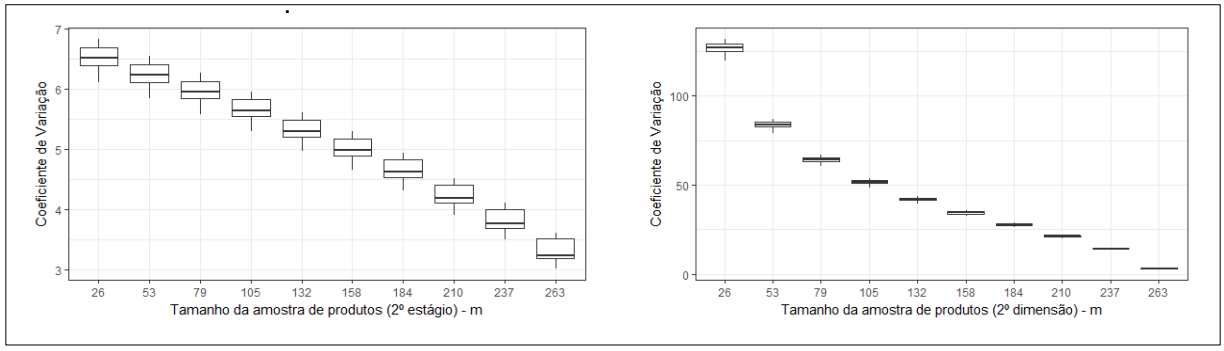
Tomando como referência padrões de qualidade sugeridos pelo IBGE (Albieri, 2006) para estimativas produzidas em pesquisas amostrais, CVs maiores que 15% levam a classificar estimativas como imprecisas. Nesse contexto, a Amostragem Bidimensional só seria capaz de produzir estimativas de qualidade aceitável quando o número de produtos amostrados (n) for muito grande, próximo do número total de produtos. Na aplicação considerada, o plano Amostral Conglomerado em 2 estágios é amplamente superior para a estimação do IPC.

5.3 COMPARAÇÃO DOS MÉTODOS DE AMOSTRAGEM CONGLOMERADA E BIDIMENSIONAL (AO LONGO DOS MESES)

A base de dados é composta por 60 meses e, assim, retirando o mês de referência (01/2012), foram calculados os índices de preços para todos os meses como mostrado no Gráfico 1. Para cada um desses meses, foi calculada a variância do estimador do IPC, assim pode-se verificar o comportamento dos planos amostrais no decorrer dos meses. Os gráficos a seguir mostram a distribuição dos coeficientes de variação do estimador sob cada plano amostral no decorrer dos meses, segundo o tamanho amostral final (m_0) e as combinações dos tamanhos de amostra de locais de compra (n) e de amostra de produtos (m).

O Gráfico 4 mostra que, para o tamanho amostral final (m_0) de 13.150 combinações de produto×local, a distribuição do coeficiente de variação do estimador de IPC sob ACS2 é, em média, bem menor do que o coeficiente de variação do mesmo estimador sob ABS. O que mais chama atenção é a grande diferença escalar dos coeficientes de variação de cada plano amostral. Enquanto sob ACS2 o CV não ultrapassa 7% para qualquer tamanho amostral de produtos, o CV sob ABS é sempre maior que 10%, chegando a ultrapassar 100% de variação para o tamanho de amostra de produtos de m=26 unidades.

Gráfico 4 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 13.150 combinações de produto × local

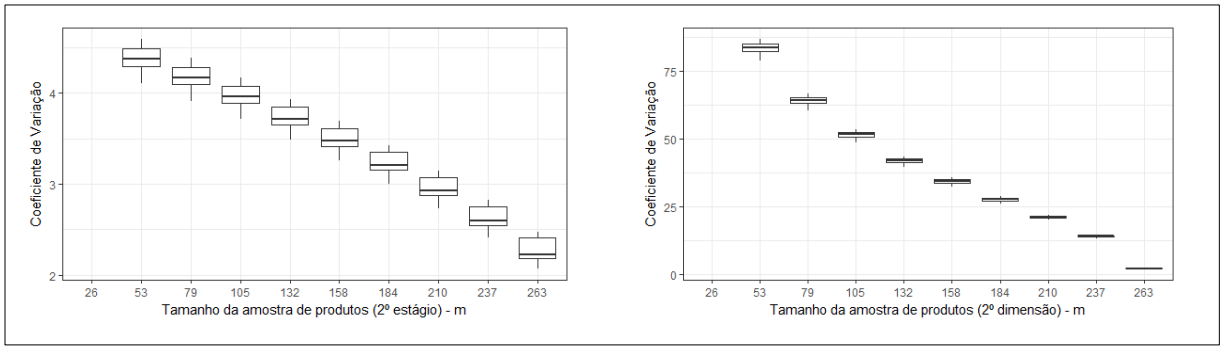


Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

O Gráfico 5 mostra que, para o tamanho amostral final (m_0) de 26.300 produtos, assim como para m_0=13.150, a distribuição do coeficiente de variação do estimador de IPC sob ACS2 é, em média, bem menor do que o coeficiente de variação do mesmo estimador sob ABS. Para este tamanho de m_0, sob ACS2 a média do CV não ultrapassa 5% para qualquer tamanho amostral de produtos, exceto para m=26 que não foi considerado por ser um cenário onde número de locais de compras selecionados era maior do que o número na população. Já a média do CV sob ABS é sempre maior que 10%.

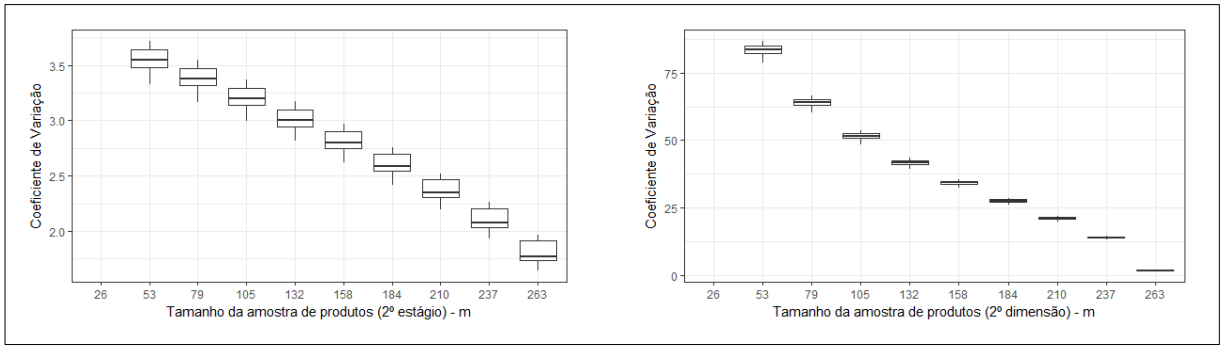
O Gráfico 6 mostra que, para o tamanho amostral final (m_0) de 39.450 produtos, o comportamento do coeficiente de variação ao longo dos meses é o mesmo que nos tamanhos amostrais m_0 já analisados até aqui. Para este tamanho de m_0, sob ACS2 a média do CV não ultrapassa 4% para qualquer tamanho amostral de produtos, além de ser possível observar uma redução da média quando m_0 aumenta. Já a média do CV, sob ABS, é sempre maior que 10% e com valores mais elevados que os de ACS2, apesar da redução com o aumento de m_0.

Gráfico 5 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 26.300 combinações de produto×local



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

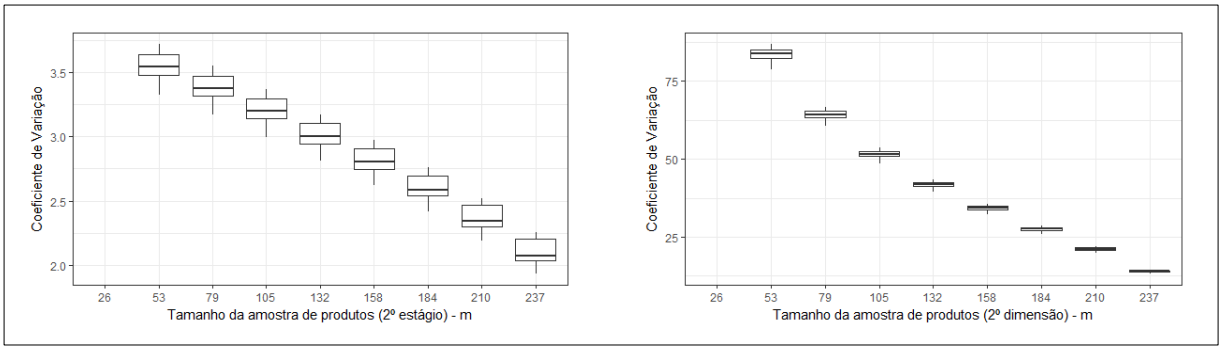
Gráfico 6 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 39.450 combinações de produto×local



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

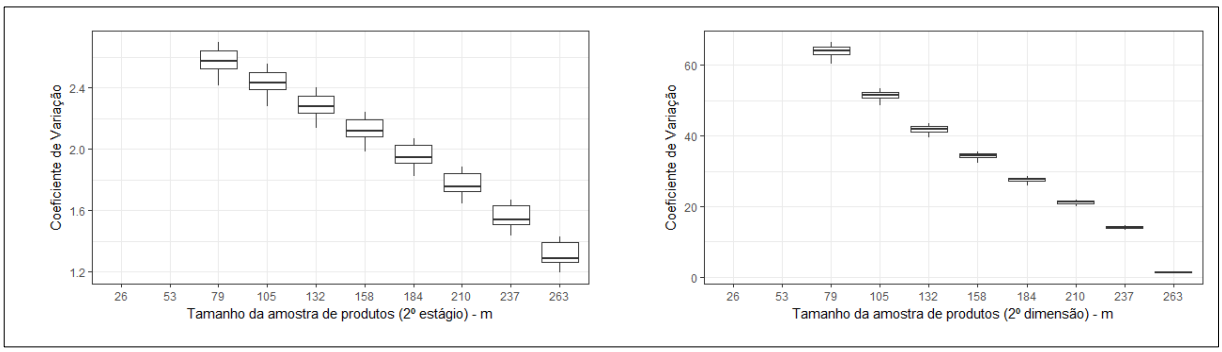
Os Gráficos (7), (8) e (9) mostram a distribuição dos CVs ao longo dos meses para os tamanhos amostrais finais de 52.600, 65.750 e 78.900, respectivamente, confirmando o mesmo comportamento dos métodos amostrais verificado até aqui.

Gráfico 7 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 52.600 combinações de produto×local



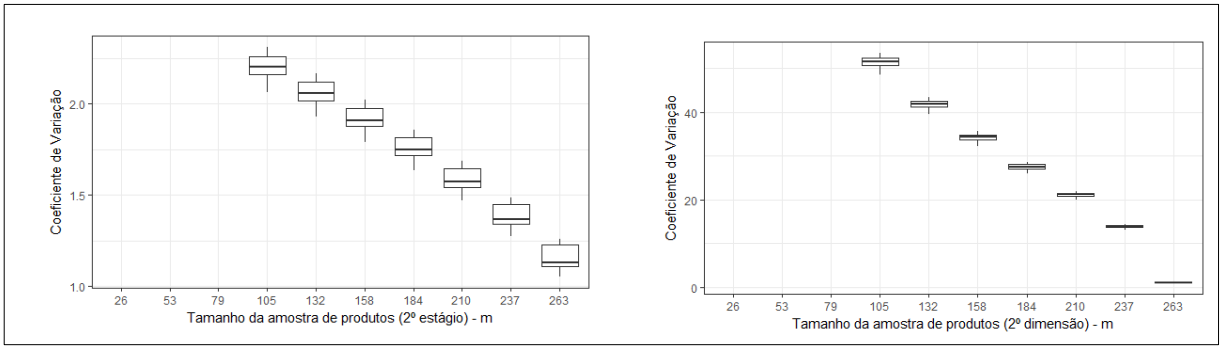
Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

Gráfico 8 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 65.750 combinações de produto×local



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

Gráfico 9 - Distribuição, ao longo dos meses, do Coeficiente de Variação (%) do estimador do IPC sob ACS2(esquerda) e ABS (direita), para o tamanho amostral final (m_0) de 78.900 combinações de produto×local



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

Em relação à variabilidade dos CVs, pode-se observar, em todos os gráficos desta subseção que, sob ACS2, a variação dos CVs ao longo dos meses é levemente maior à medida em que o número de produtos selecionados aumenta. Vale lembrar que, aumentando o número de produtos selecionados, o número de locais de compra selecionados diminui, uma vez que o tamanho da amostra final é mantido constante. Assim, pode-se pensar que existe uma participação importante da variância do primeiro estágio na variância total do estimador sob ACS2. O mesmo não é possível verificar no gráfico de ABS devido à grande diferença de escala dos valores dado o tamanho de amostra de produtos. Assim, uma análise da participação da variância de cada estágio de seleção (e dimensão) na variância total do estimador se torna necessária.

5.4 PARTICIPAÇÃO DA VARIÂNCIA DOS PRODUTOS NA VARIÂNCIA TOTAL DO ESTIMADOR DE IPC

O comportamento distinto entre os CVs de ACS2 e ABS conduz a questionar qual dos dois estágios (ou dimensões) tem maior influência sobre a variância do estimador do IPC.

A partir das expressões das variâncias dos estimadores para cada plano amostral, pode-se definir as participações da variância dos locais de compra e dos produtos na variância total do estimador. Para ACS2, a expressão (6) pode ser dividida em duas parcelas:

$$V_{ACS2}(\hat{L}_{t|0}) = V_{ACS1}(\hat{L}_{t|0}) + \frac{N}{n}\sum_{i\in L}\frac{M_i^2}{m_i}(1-f_2)\,S_i^2 \qquad (6)$$
$$= V_{UPA} + V_{USA}$$

onde V_{UPA} é a variância do primeiro estágio, ou seja, é a variância proveniente do sorteio dos locais de compra e V_{USA} é a variância do segundo estágio, ou seja, é a variância proveniente do sorteio dos produtos dentro dos locais de compra selecionados. Assim, podemos definir a participação da variância do segundo estágio na variância total do estimador como:

$$PV_{USA} = \frac{V_{USA}}{V_{ACS2}(\hat{L}_{t|0})} \qquad (7)$$

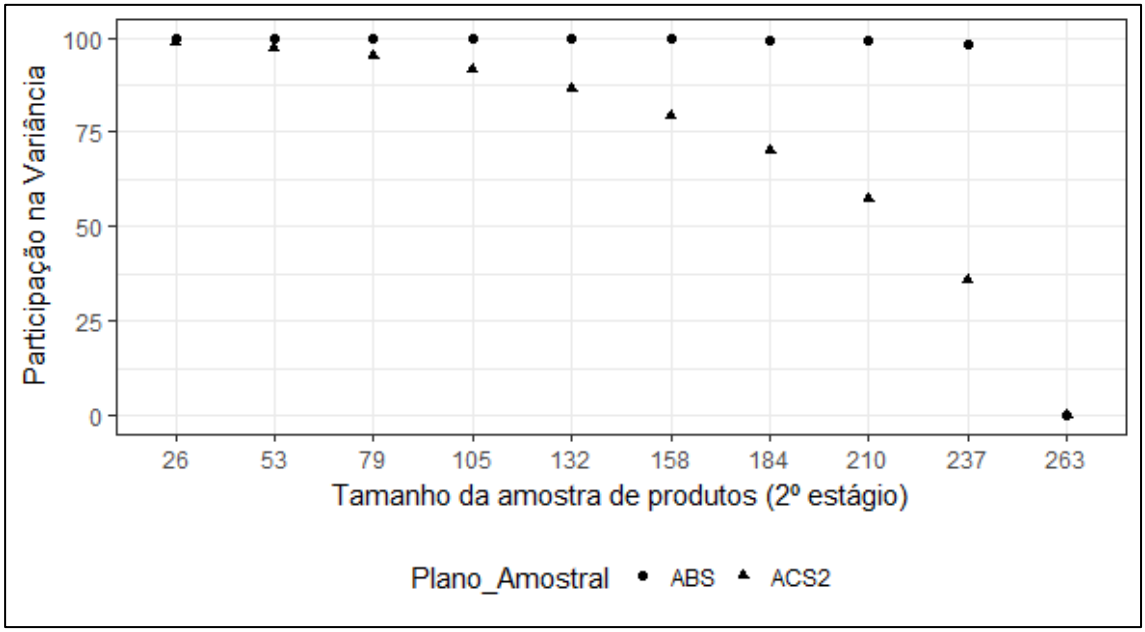
Para ABS, a expressão da variância do estimador já está subdivida em três parcelas, conforme a expressão (16) do Apêndice A. Assim, podemos definir a participação da variância da segunda dimensão (dos produtos) na variância total como:

$$PV = \frac{V^L}{V_{ABS}(\hat{L}_{t|0})_{2ºdim}} \qquad (8)$$

Pode-se observar pelo Gráfico 10, que apresenta a média da participação da variância dos produtos na variância total do estimador para os 59 meses que, para ACS2, a participação da variância do segundo estágio reduz-se à medida em que mais produtos são selecionados. O que ocorre neste caso é que selecionando mais produtos e mantendo o tamanho amostral final (m_0) constante, o número de locais de compra selecionados reduz e com isso a variância proveniente do primeiro estágio aumenta. Já para ABS a participação da variância dos produtos na variância total é sempre constante e próxima dos 100%. Isso nos leva a concluir que a variação dos produtos tem maior impacto na variância do estimador do que a variância dos locais de compra, para o caso da ABS.

O resultado para ABS surpreendeu, pois era esperado um comportamento semelhante ao método de amostragem conglomerada, onde o local de compra também tem uma participação elevada na variância do estimador. Levando em consideração o contexto dos nossos dados, essa conclusão faz sentido. Por uma percepção do dia-a-dia, dado um produto específico, o seu preço varia de local de compra para local de compra. No entanto, dado um local de compra específico, a variabilidade de preços de produtos dentro deste local de compra é maior, uma vez que no local de compra há diversos produtos com preços bem distintos entre si. Assim, a variabilidade dos produtos ter uma forte participação na variação total do estimador é plausível sob ABS.

Gráfico 10 - Média, ao longo dos meses, da Participação (%) da variância do segundo estágio (2ª dimensão) na variância total do estimador do IPC



Fontes: WFP. Microdados do *World Food Programme* 2012-2016.

6. CONCLUSÃO E CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem para o cálculo da variância do estimador de Índices de Preços ao Consumidor sob planos amostrais aplicáveis a populações bidimensionais. A abordagem baseou-se na comparação de eficiência, do ponto de vista de precisão de estimadores, dos planos amostrais conglomerado e bidimensional. Concluiu-se, por essa perspectiva de eficiência, que o plano amostral bidimensional é bem menos eficiente que o plano amostral conglomerado em dois estágios que usa os locais de compra como UPAs, no contexto da estimação de IPCs definidos via índices de Laspeyres.

Uma conclusão, um tanto surpreendente, foi a de que a variância dos preços entre os locais de compra tem pouca importância na variância total do estimador, sendo a mesma totalmente influenciada pela variação proveniente do sorteio de produtos.

No processo de definição de qual plano amostral usar em uma pesquisa, diversos quesitos sobre os planos amostrais são analisados além da precisão. Questões relacionadas com o custo de coleta das informações, existência de cadastros para a seleção das amostras e custos com a atualização desses cadastros têm um forte peso na escolha do plano amostral. Assim, algumas características do plano amostral bidimensional podem reduzir esses custos, e tornar o mesmo preferível ao plano amostral conglomerado em dois estágios.

Uma vantagem do plano amostral bidimensional é que não há a necessidade de um cadastro de celas (de produtos×locais), apenas um cadastro de locais de compra e um outro, independente, de produtos. Portanto, não haveria necessidade da construção de cadastros de produtos para cada local selecionado, uma vez que eles já foram selecionados de um cadastro universal de produtos. Assim, haveria uma economia operacional e consequentemente de custos.

Como no plano amostral bidimensional são retiradas amostras independentes de produtos e de locais de compra, é possível ter controle tanto sobre a amostra de locais como sobre a amostra de produtos, diferente do plano amostral conglomerado em dois estágios, no qual o controle é exercido apenas sobre a amostra de locais de compra. Na AB o tamanho da amostra de locais (que será o mesmo para cada produto) pode ser planejado e previamente conhecido, permitindo assim, a vantagem de produzir estatísticas a respeito dos produtos selecionados com uma precisão especificada.

Os resultados deste trabalho, apontaram uma maior magnitude das variâncias sob Amostragem Bidimensional Simples quando comparadas com as variâncias sob Amostragem Conglomerada Simples em 2 estágios. Uma consideração a fazer é que, dada a variabilidade de produtos no mercado, e consequentemente a alta variabilidade de preços de produtos, as pesquisas de variação de preços fazem uma estratificação dos produtos. Vos (1964) apresentou expressões para as variâncias de estimadores considerando estratificação. Segundo Norberg (2004), a Suécia adota a AB com estratificação em sua pesquisa de variação de preços e o IBGE (2016) agrupa os produtos conforme categorias de consumo das famílias brasileiras. Assim, uma análise comparativa entre ACS2 com um plano amostral bidimensional com estratificação por produtos é uma opção de trabalho futuro para verificar como a estratificação melhoraria a eficiência do plano amostral bidimensional.

Também como trabalho futuro, visto que essas análises foram feitas com conglomerados de mesmo tamanho, pode-se comparar as variâncias do estimador do IPC sob AB e ACS2 com conglomerados de tamanhos distintos. O método de seleção das UPA´s, das linhas e colunas (na AB) foi a AAS. Uma nova análise, como a realizada neste trabalho, poderia ser feita considerando a amostragem PPT em vez de AAS.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Albieri, S. (2006). Pesquisas por amostragem: política de divulgação de estimativas com baixa precisão amostral. CONFEST 2006. Disponível em: https://www.ibge.gov.br/confest_e_confege/pesquisa_trabalhos/CD/mesas_redondas/294-3.pdf.

Bethlehem, J.(2009). *Applied Survey Methods: a statistical perspective*. New Jersey: John Wiley & Sons.

Bolfarine, H. & Bussab, W. O. (2005). *Elementos de Amostragem*. (5a ed.). São Paulo: Blucher.

Cochran, W. G. (1977). *Sampling Techniques*. (3a ed.). New York: Wiley.

Dalén, J. & Ohlsson, E. (1995). Variance Estimation in the Swedish Consumer Price Index. Journal of Business & Economic Statistics, 13(3), 347–356.

Fava, V. L. A. (2007). Precisão dos Índices de Preços. EconomiA, 8(1), 39–63, Brasília.

Fisher, I. (1922). *The making of index numbers: a study of their varieties, tests and reliability*. Boston: Houghton.

IBGE. (2013). Sistema Nacional De Índices De Preços Ao Consumidor - Métodos de Cálculo. 7ª ed. Rio de Janeiro: IBGE.

IBGE (2016). Para compreender o INPC: um texto simplificado. 7ª ed. Rio de Janeiro: IBGE.

Juillard, H. (2016). Two-dimensional sampling in practice. Case Studies in Business, Industry and Government Statistics, 6(1), 36–49.

Juillard, H.; Chauvet, G. & Ruiz-Gazen, A. (2017). Estimation under cross-classified sampling with application to a childhood survey. Journal of the American Statistical Association, 112(518), 850-858.

Norberg, A. (2004). Comparison of Variance Estimators for the Consumer Price Index. Ottawa Group Meeting - Helsinki, 8(8), 1–29.

Ohlsson, E. (1996). Cross-Classified Sampling. Journal of Official Statistics, 12(3), 241–251.

Oliveira, L.D.S. (2018). Comparação de métodos de amostragem aplicáveis à estimação de índices de preços ao consumidor. (Dissertação de mestrado). Escola Nacional de Ciências Estatísticas (ENCE), Rio de Janeiro.

Pessoa, D. G. C., e Silva, P. L. do N. (1998). *Análise de dados amostrais complexos*. São Paulo: Associação Brasileira de Estatística.

R CORE TEAM. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Vienna, Áustria.

Vos, J. W. E. (1964). Sampling in Space and Time. Review of International Statistical Institute, 32(3), 226–241.

WFP. World Food Programme. Disponível em: <http://www1.wfp.org/>. Acesso em: 10/7/2017a.

WFP. World Food Programme: Dataset. Disponível em: <https://data.humdata.org/dataset/wfp-food-prices>. Acesso em: 24/7/2017b.

APÊNDICE A: Estimadores Total Amostral e suas Variâncias

Neste apêndice apresenta-se as expressões dos estimadores do total amostral para cada plano amostral considerado neste trabalho, bem como suas respectivas variâncias. Primeiramente apresentamos a notação para a população e para a amostra. UPA é a unidade primária de amostragem e USA é a unidade secundária de amostragem. Este artigo trabalha com conglomerados de mesmo tamanho, assim, seja:

$U = \{1,2,\dots,i,\dots,N\}$ o conjunto que representa a população de UPAs;

$s = \{i_1, i_2, \dots, i_n\}$ o conjunto que representa a amostra de UPAs;

$U_i = \{1,2,\dots,j,\dots,M\}$ o conjunto que representa a população de USAs dentro da UPA i;

$s_i = \{j_1, j_2, \dots, j_m\}$ o conjunto que representa a amostra de USAs dentro da UPA i;

y a variável de interesse;

y_{ij} o valor da variável y para a USA j da UPA i;

$Y_i = \sum_{j\in U_i} y_{ij}$ o valor total da variável y para a UPA i;

$M_0 = \sum_{i\in U} M = NM$ o tamanho total da população de USAs;

$m_0 = \sum_{i\in s} m = nm$ o tamanho da amostra de USAs;

$Y = \sum_{i\in U}\sum_{j\in U_i} y_{ij} = \sum_{i\in U} Y_i$ o total populacional;

$\bar{Y}_C = \frac{Y}{N}$ O total médio por UPA;

$\bar{Y}_i = \frac{Y_i}{M}$ a média por USA na UPA i;

$\bar{Y} = \frac{Y}{M_0}$ a média por unidade na população toda;

π_i a probabilidade de seleção da UPA i;

π_{ij} a probabilidade de seleção da USA j da UPA i;

$\pi_{j|i}$ a probabilidade condicional de seleção da USA j dentro da UPA i;

$f_1 = \frac{n}{N}$ é a fração amostral do primeiro estágio;

$f_2 = \frac{m}{M}$ é a fração amostral do segundo estágio.

A.1. ESTIMADOR DE TOTAL POPULACIONAL E SUA VARIÂNCIA – AMOSTRAGEM CONGLOMERADA SIMPLES EM 1 ESTÁGIO

Nesta seção apresenta-se o estimador de total populacional de Horvitz – Thompson[6] e sua variância para o plano amostral conglomerado em um estágio com amostragem aleatória simples no primeiro estágio (ACS1). Este trabalho considera os conglomerados com o mesmo tamanho, assim, o referido estimador é dado por:

[6]Para mais detalhes sobre este estimador consulte Cochran (1977, p. 259-261).

$$\hat{Y}_{HT} = \sum_{i\in U} \frac{Y_i}{\pi_i} = \sum_{i\in s} \sum_{j\in U_i} w_{ij} y_{ij} \quad (9)$$

onde: $w_{ij} = \frac{1}{\pi_{ij}} = \frac{1}{\pi_i}\frac{1}{\pi_{j|i}} = \frac{N}{n}$ é o peso amostral.

Logo, esse estimador pode ser reescrito como:

$$\hat{Y}_{HT} = \frac{N}{n} \sum_{i\in s} \sum_{j\in U_i} y_{ij} \quad (10)$$

Sua variância é dada por:

$$V_{ACS1}(\hat{Y}_{HT}) = N^2 \frac{1-f_1}{n} S_e^2 \quad (11)$$

onde: $S_e^2 = \frac{1}{N-1}\sum_{i\in U}(Y_i - \bar{Y}_C)^2$ é chamada de variância entre os conglomerados. Detalhes podem ser consultados, por exemplo, em Cochran (1977,p. 233-248).

A.2. ESTIMADOR DE TOTAL POPULACIONAL E SUA VARIÂNCIA – AMOSTRAGEM CONGLOMERADA SIMPLES EM 2 ESTÁGIOS

No planejamento amostral conglomerado simples em 2 estágios (ACS2), o estimador de Horvitz-Thompson para o total populacional é dado por:

$$\hat{Y}_{HT} = \sum_{i\in s} \sum_{j\in s_i} w_{ij} y_{ij} \quad (12)$$

onde: $w_{ij} = \frac{1}{\pi_{ij}} = \frac{1}{\pi_i}\frac{1}{\pi_{j|i}} = \frac{N}{n}\frac{M}{m}$

Logo, esse estimador pode ser reescrito como:

$$\hat{Y}_{HT} = \frac{N}{n}\sum_{i\in s}\frac{M}{m}\sum_{j\in s_i} y_{ij} = \frac{N}{n}\sum_{i\in s}\hat{Y}_i \quad (13)$$

com $\hat{Y}_i = \frac{M}{m}\sum_{j\in s_i} y_{ij}$.

A sua variância é dada por:

$$V_{ACS2}(\hat{Y}_{HT}) = \frac{N^2}{n}\frac{(1-f_1)}{N-1}\sum_{i\in U}(Y_i - \bar{Y}_c)^2 + \frac{N}{n}\frac{M^2}{m}(1-f_2)\sum_{i\in U} S_i^2 \quad (14)$$

onde $S_i^2 = \frac{1}{M-1}\sum_{j\in U_i}\left(y_{ij} - \bar{Y}_i\right)^2$ é a variância dentro do conglomerado U_i.

A.3. ESTIMADOR DE TOTAL POPULACIONAL E SUA VARIÂNCIA – AMOSTRAGEM BIDIMENSIONAL

Considerando uma população bidimensional disposta em uma matriz com N linhas e M colunas, totalizando $N \times M$ unidades elementares, uma amostra bidimensional simples S desta população é um cruzamento das amostras S^L e S^C, onde S^L e S^C são amostras aleatórias simples sem reposição (AASs) independentes das linhas e das colunas da matriz populacional, com tamanhos n e m respectivamente (Ohlsson, 1996).

O estimador total da variável y para a Amostragem Bidimensional é dado por:

$$\hat{Y}_{HT} = \frac{NM}{nm}\sum_{i\in s^L}\sum_{j\in s^C} y_{ij} \quad \text{(Ohlsson, 1996, eq. 1.7)} \quad (15)$$

Segundo Ohlsson (1996) e Juillard (2016), a variância de $\hat{Y}_{HT}$ sob o plano amostral bidimensional simples, pode ser decomposta em:

$$V_{ABS}(\hat{Y}_{HT}) = V^L + V^C + V^{LC} \quad \text{(Ohlsson, 1996, eq. 1.3)} \tag{16}$$

onde:

$$V^L = \frac{N^2M^2}{n(N-1)}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2 \tag{17}$$

$$V^C = \frac{N^2M^2}{m(M-1)}\left(1-\frac{m}{M}\right)\sum_{j=1}^{M}\left(\bar{Y}_j - \bar{Y}\right)^2 \tag{18}$$

$$V^{LC} = \frac{N^2M^2}{nm}\left(1-\frac{n}{N}\right)\left(1-\frac{m}{M}\right)\frac{1}{(N-1)(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}\right)^2 \tag{19}$$

APÊNDICE B: COMPARAÇÃO DE MÉTODOS DE AMOSTRAGEM

A existência de planos amostrais alternativos requer uma forma de saber qual dos planos amostrais é mais eficiente. Em termos estatísticos, o mais eficiente é aquele que produz estimativas mais precisas, ou seja, com menores variâncias para os estimadores.

Estimadores não viesados são aqueles que possuem valor esperado igual ao valor do parâmetro da população que buscam estimar. Quando dois estimadores são não viesados, uma forma de comparar dois planos amostrais é através do Coeficiente de Variação (CV) que é dado por:

$$CV = 100\frac{\sqrt[2]{V_p(\hat{Y}_{HT})}}{E_p(\hat{Y}_{HT})} = 100\frac{EP_p(\hat{Y}_{HT})}{E_p(\hat{Y}_{HT})} \tag{20}$$

onde $EP_p(\hat{Y}_{HT})$ é o erro padrão do estimador de Horvitz – Thompson. Apresentou-se a expressão do CV multiplicada por 100 de forma a tornar o resultado mais intuitivo.

Usamos o CV para comparar os planos amostrais dessa dissertação. O plano amostral com CV menor é considerado como o plano mais eficiente, do ponto de vista de precisão do estimador, em comparação com o plano amostral com CV mais elevado.

AGRADECIMENTOS e COLABORAÇÕES

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido na forma de bolsa para realização do Mestrado em População, Território e Estatísticas Públicas da Escola Nacional de Ciências Estatísticas (ENCE). Agradecemos também ao revisor pelos comentários que ajudaram a aperfeiçoar a apresentação do artigo.

Revista Brasileira de Estatística, V.79, N. 246, p.102-117, janeiro-junho 2021

COMPARAÇÃO DE PLANOS AMOSTRAIS PROBABILÍSTICOS ALTERNATIVOS COM UM PLANO AMOSTRAL NÃO PROBABILÍSTICO

João Gabriel Malaguti

joaogmalaguti@gmail.com

Escola Nacional de Ciências Estatísticas

Samuel Faria Candido

samuelfariacandido@gmail.com

Universidade Federal de Minas Gerais, Departamento de Estatística

Marcel de Toledo Vieira

marcel.vieira@ice.ufjf.br

Universidade Federal de Juiz de Fora, Departamento de Estatística

Resumo: Institutos de pesquisa privados adotam com frequência planos amostrais não probabilísticos para a realização de inferências estatísticas, ainda que utilizando pressupostos que necessitam amostragem probabilística. Neste artigo, criamos um algoritmo para imitar amostragem por cotas de modo a poder comparar a precisão desta com métodos probabilísticos num ambiente controlado, por simulações Monte Carlo. Criamos estudos de simulação para observar os efeitos do plano amostral e das diferentes fórmulas para estimativas de erro padrão sobre estas amostras. Resultados apontam que o nível de cobertura se encontra próximo de 40% para amostragem por cotas, o que implica em um nível de significância efetivo de aproximadamente 60%.
Palavras-chave: Amostragem; Simulação Monte Carlo; Nível de cobertura; Nível de significância efetivo

Abstract: Private research institutes frequently adopt non-probability sampling plans for statistical inferences, using assumptions that rely on probability sampling. In this paper, we created an algorithm to mimic quota sampling in order to compare its precision with probabilistic methods in a controlled environment, using Monte Carlo simulations. Simulations were developed to observe the effects of both the sampling plan and the effect of the different formulas for standard error estimation on those samples. Results show that the coverage level is close to 40% for quota sampling, which implies an effective significance level of 0.6.
Keywords: Sampling; Monte Carlo simulation; Confidence intervals coverage; Effective significance level

1.INTRODUÇÃO

O Professor Djalma Galvão Carneiro Pessoa foi responsável pela produção de inúmeras contribuições metodológicas à área de Amostragem, por exemplo, em: (i) métodos para a análise de dados amostrais complexos (Pessoa & Silva, 1998; Pessoa et al., 2002); (ii) tratamento de não resposta, crítica e imputação (Pessoa et al., 2000); (iii) pareamento probabilístico (Pessoa, 2013); e (iv) estimação em pequenas áreas (Brendolin et al., 2013); dentre outras. O terceiro autor do presente artigo foi academicamente muito influenciado pelo Professor Djalma, a partir da leitura de seus trabalhos, de encontros e conversas que se deram desde o ano 2000, em ambientes de trabalho no IBGE, inúmeros eventos científicos, apresentações de seminários e visitas acadêmicas. Os dois primeiros autores tiveram menos contato com o Professor, assistindo algumas palestras por ele ministradas, mas seus comentários sobre o presente trabalho, quando apresentado na V Escola de Amostragem e Metodologia de Pesquisa em 2017, foram muito relevantes e permitiram a evolução deste artigo.

A amostragem é a base de toda teoria estatística, sem uma amostra não se realizam inferências. Um dos focos de discussão dessa área é a amostragem não probabilística, em especial a amostragem por cotas (ApC), e como essa se compara a métodos probabilísticos como amostragem aleatória simples.

Institutos de pesquisa privados utilizam-se de fórmulas para amostragem probabilística mesmo quando sua pesquisa é feita com amostragem por cotas. Tal prática é falha, pois com a ausência da aleatoriedade, perde-se a capacidade de calcular o erro padrão da estatística (Moser, 1952). Apesar disto, são publicadas margens de erro e níveis de confiança das estimativas provindas desse método, sem base teórica e, portanto, sem legitimidade (Ferraz, 1996).

Nesse trabalho foi criada uma rotina no *software* R (R Core Team, 2021) para que, a partir de um banco de dados populacional, fossem retiradas amostras por oito métodos diferentes; sendo eles: amostragem aleatória simples com reposição (AASc), amostragem aleatória simples sem reposição (AASs), amostragem sistemática (ASis), amostragem estratificada simples com alocação uniforme, proporcional, e alocação ótima de Neyman (AESu, AESp e AESo), além da amostragem por cotas utilizando duas fórmulas distintas.

Para cada tipo de amostragem foram feitas 100.000 (cem mil) simulações, e, em cada uma delas se calculou a média amostral ($\bar{y}$) e o erro padrão (EP). A partir destes, o intervalo de confiança com nível de significância 95% e o nível de cobertura (NC).

Já para amostragem por cotas, foram utilizadas as fórmulas para as estatísticas como se fossem AASs e, em um segundo momento, como se fossem AES (ApCAAS e ApCAES, respectivamente). Apesar de os institutos de pesquisa utilizarem apenas as fórmulas AAS, utilizamos as fórmulas AES pelo fato das cotas se assemelharem superficialmente com estratos, segundo Kish (1965).

Para a realização do estudo de simulação foram consideradas como população alvo as observações do banco de dados fornecido pelo CISDESTE (Consórcio Intermunicipal de Saúde da Região Sudeste/MG), que contém informações sobre o atendimento ambulatorial na região sudeste mineira durante o período de fevereiro de 2014 e abril de 2017.

Este artigo está organizado da seguinte maneira: a seção seguinte contém a revisão da literatura, com a seção 3 apresentando a metodologia, definindo as estatísticas e métodos de amostragem utilizados. A seção 4 apresenta os resultados dos estudos de simulação e a última seção apresenta as considerações finais.

2.REVISÃO DA LITERATURA

A literatura sobre amostragem não probabilística iniciou-se relativamente próxima da amostragem como um todo (Neyman, 1934), mas conta com poucos estudos próprios. É possível perceber um padrão no interesse nessa área: quando alguma pesquisa que a emprega erra bastante na previsão (geralmente de resultados de eleições), renova-se o interesse.

O primeiro grande escândalo relacionado à amostragem por cotas foram os resultados das pesquisas eleitorais americanas de 1948, que previram a vitória de Dewey sobre Truman. O próprio governo americano comissionou relatórios sobre tais pesquisas. Por exemplo, Daniel Katz (1949), em seu relatório explicita que o raciocínio por trás do método de controle de cotas é criar um cruzamento que, em tese, representa a população proporcionalmente em termos de sexo, idade, status socioeconômico, urbanização e área geográfica (os chamados “controles”).

Também no relatório, se levanta que, se o cruzamento obtido pelo método de cotas fosse realmente capaz de representar a população, outras características (como religião e ocupação) também deveriam ser propriamente representadas, o que não aconteceu. Katz também diz que os ajustes de correção introduzidos, como ajuste dos pesos, se mostraram limitados pela amostragem.

Os próprios controles não são isentos de problemas. Os controles relacionados à classe social ou econômica inevitavelmente são definidos em termos vagos, cabendo ao entrevistador julgar a qual categoria o indivíduo pertence, tendo então um viés que não pode ser mensurado. O pouco controle dos agentes de campo também permite aos entrevistadores alocar respondentes nas células em que casos são difíceis de encontrar, ao invés daquelas células às quais eles realmente pertencem. Isso se aplica particularmente a controles vagos, como “classe social” (Katz, 1949).

Kish (1965) ressalta que idade, sexo e região geográfica são os controles mais usados, mesmo quando a relação desses com as variáveis pesquisadas são fracas. Quando se utilizam outros controles geralmente seu número é baixo, visto que quanto mais complicado o esquema de cotas, mais difícil se torna o trabalho do entrevistador e mais cara fica a pesquisa. E, mesmo que se pudesse conseguir variáveis sociodemográficas com precisão, estas não garantem controle sobre outras variáveis relacionadas com a pesquisa (Moser & Stuart, 1953).

A diferença fundamental entre a amostragem por cotas e a amostragem aleatória é que, depois que as cotas são definidas e a quantidade de entrevistas são alocadas para cada entrevistador, a escolha da amostra é deixada a critério do mesmo (Moser, 1952).

Viés também pode ser introduzido quando os entrevistadores não conseguirem uma amostra representativa de respondentes dentro das cotas: e.g. a amostra de pessoas com idade igual ou superior a 65 pode conter uma grande quantidade de pessoas com 65 e 66, de modo que a parcela mais velha da população não é bem representada.

Mesmo se a cota está corretamente coletada para os controles e para a maioria das variáveis não mensuradas, ainda pode ser não representativa com relação à variável alvo. Em 1953, Moser e Stuart confirmaram com sua pesquisa que a amostra por cotas é viesada, principalmente em questões ligadas à educação e à ocupação.

Além disso, existe também o fato de que tais amostras tendem a ser viesadas para aqueles dispostos a responder, facilmente acessíveis e interessados no tópico da pesquisa. E como aqueles que não respondem são substituídos, estes são sub-representados (Yang & Banamah, 2014).

Tal substituição é feita para tentar superar a não resposta e o não contato, mas Yates (Moser & Stuart, 1953) argumenta que a amostragem por cotas é sujeita a não contato de tipo diferente e indeterminado. Marsh &

Scarbrough (1990) mostram em sua pesquisa que existem diferenças significativas entre as estimativas do método por cotas e aleatório quando o nível de resposta é satisfatório. Quando o nível é baixo, no entanto, os resultados das amostras tendem a ser similares.

É importante reiterar que inexiste base teórica para calcular o erro padrão da estimativa provinda do método de cotas. Apesar disso, não é incomum que institutos utilizem a fórmula para cálculo supondo amostragem aleatória simples sem reposição para estimar o erro padrão. Essa prática resulta numa subestimação da variância total. Além do fato que amostras por cotas são frequentemente em múltiplos estágios e as fórmulas utilizadas não refletem isso. Mesmo assim, em geral a variância da amostra por cotas é maior do que a da amostra aleatória (Moser & Stuart, 1953).

O Brasil também teve (e tem) a mesma relação para com a amostragem por cotas: interesse cresce quando pesquisas erram. Como tentativa falha de solucionar certos problemas foi criada uma lei que obriga institutos privados a publicar margens de erro e intervalos de confiança junto com o resultado de suas pesquisas eleitorais (Brasil, 1997; Brasil, 2013; Brasil, 2019).

Em um boletim da Associação Brasileira de Estatística, Carvalho & Ferraz (2006) analisam os problemas que tal lei traz: as margens de erro declaradas, como admite o IBOPE, são baseadas em fórmulas de amostragem aleatória simples, mesmo que essas não se apliquem a qualquer método de amostragem não probabilística. No mesmo boletim, os autores discutem sobre o levantamento de resultados de pesquisas eleitorais documentado no livro Pesquisa Eleitoral: Críticas e Técnicas (Souza, 1990), que mostrou que a maioria das pesquisas errava muito além das margens de erro declaradas.

Apesar do foco em pesquisas eleitorais, elas não são o único tipo de pesquisa a empregar amostragem por cotas, sendo utilizadas por institutos de pesquisa privados em pesquisas de satisfação, de mercado e até mesmo de opinião. A facilidade e o baixo preço associados ao método provavelmente continuarão a ser atraentes, embora nem toda amostra não probabilística seja útil para inferência (Valliant, 2020).

Nos últimos anos, novos métodos têm sido criados e refinados para permitir análises de amostras não probabilísticas. Elliott & Valliant (2017) assim como Valliant (2020), apresentam a quasi-aleatorização (*quasi-randomization*) e a modelagem de superpopulação (*superpopulation modeling*) e discutem vantagens e desvantagens. No entanto, como o foco deste artigo é nos mecanismos da amostragem por cotas e não em inferência, não entraremos em

detalhes sobre estas novas abordagens para lidar com amostras não-probabilísticas.

3.METODOLOGIA

3.1 DADOS

A população alvo considerada para a realização dos estudos de simulação foram as observações sobre atendimentos ambulatoriais com deslocamento de ambulância ocorridos na mesorregião Sudeste do estado de Minas Gerais compreendidas entre o período de fevereiro de 2014 a abril de 2017, conforme o banco de dados fornecido pelo CISDESTE (Consórcio Intermunicipal de Saúde da Região Sudeste/MG).

As variáveis utilizadas presentes no banco de dados estão descritas no Quadro 1. O estudo considerou 127.120 observações válidas, sendo que 52,75% dos atendimentos foram para indivíduos do sexo masculino e 47,25% do sexo feminino.

Quadro 1 – Variáveis Existentes no Banco

Variável	Descrição
t2	Tempo entre saída da base e chegada no local.
sexo	Sexo do paciente (F e M).
fx_idade	Idade do paciente (0-20, 20-40, 40-60, 60+)
microrregiao	8 microrregiões do sudeste mineiro.
transporte	Tipo de transporte (pré-hospitalar e inter-hospitalar)

A variável de interesse nesse estudo é “t2”, o tempo em minutos entre a saída da ambulância da base e chegada no local de atendimento, também chamado de tempo de resposta ou tempo de ouro. A média populacional ($\bar{Y}$) da variável “t2” é 13,169 minutos.

As idades dos pacientes atendidos em cada ocorrência foram classificadas em quatro faixas. A maioria dos atendidos (34,11%) eram idosos (60 anos ou mais). Discriminando por sexo, observa-se que a classe com maior frequência foi a das mulheres com mais de 60 anos (18,31%). A distribuição por sexo e faixa etária foi utilizada como controle para a amostragem por cotas. Os dados relativos à faixa etária dos pacientes discriminada pelo sexo e suas porcentagens estão representados na Tabela 1 a seguir.

O atendimento ambulatorial possui dois tipos de transporte, o transporte inter-hospitalar, que consiste em 5,39% dos casos, e o transporte pré-hospitalar, que corresponde à grande maioria de 94,61% das situações.

Tabela 1 - Atendimentos por Sexo e Faixa Etária

Sexo	Faixa Etária	%
Masculino	1 a 20 anos	5,32
	20 a 40 anos	14,82
	40 a 60 anos	16,81
	Maior de 60 anos	15,80
Feminino	1 a 20 anos	5,30
	20 a 40 anos	11,53
	40 a 60 anos	12,10
	Maior de 60 anos	18,31

O CISDESTE abrange oito microrregiões do sudeste mineiro: Além Paraíba, Carangola, Leopoldina, Santos Dumont, Bicas, Muriaé, Ubá e Juiz de Fora, sendo esta última onde ocorre a maioria dos casos (51,36%). Os dados relativos aos atendimentos por microrregião e suas porcentagens estão representadas na Tabela 2:

Tabela 2 - Atendimentos por Microrregião

Microrregião	%
Além Paraíba	2,49
Carangola	6,40
Leopoldina	11,59
Santos Dumont	4,60
Bicas	3,97
Juiz de Fora	51,36
Muriaé	7,34
Ubá	12,25

Os atendimentos por microrregião e por tipo de transporte revelam que a maior parte dos atendimentos (48,89%) foram casos de transporte pré-hospitalar na microrregião de Juiz de Fora. Tal cruzamento é o utilizado para a

criação dos estratos para amostragem estratificada simples (AES). Com exceção da microrregião de Bicas (com 13,35%), a parcela dos atendimentos inter-hospitalares se encontra entre 3,13% e 8,8% do total da região.

3.2 NOÇÕES DE AMOSTRAGEM

Como neste trabalho a variância da variável de interesse "t2" (182,78 minutos quadrados) é conhecida para a unidade amostral (atendimento ambulatorial) podemos calcular o tamanho de amostra sem o uso de uma *proxy*. Fixando a margem de erro para estimação da média populacional de "t2" em 1 unidade e o nível de confiança em 95% obtivemos um tamanho de amostra de 703 para amostragem aleatória simples com reposição. Para simplificar os cálculos seguintes, adotou-se $n = 700$ como tamanho para amostra base.

Para a amostragem estratificada, utilizando o cruzamento entre microrregião e por tipo de transporte, consideramos as alocações uniforme, proporcional e ótima de Neyman. Neste artigo, dentro de cada estrato foi usada AASs. Para a alocação uniforme (AESu) atribui-se o mesmo tamanho de amostra para cada estrato, isto é, divide-se o valor base pelo número de estratos (16). Para amostragem estratificada simples com Alocação Proporcional (AESp) a amostra de tamanho n é distribuída proporcionalmente ao tamanho dos estratos, isto é, $n_h = n \cdot (N_h/N)$, onde n_h é o tamanho de amostra do estrato h, N_h é o tamanho populacional do estrato h e N é o tamanho da população como um todo.

A Alocação Ótima de Neyman, presumindo custo igual por estrato, tem o tamanho da amostra em cada estrato definido por $n_h = N_h S_h / \sum_{h=1}^{L} N_h S_h$, onde S_h é o desvio padrão da variável de interesse no estrato h (o equivalente para a população seria S) e L é o número de estratos. Devido a arredondamentos para cada estrato, os tamanhos de amostra são de 704, 706 e 710 para AESu, AESp e AESo, respectivamente. Tal arredondamento também afeta a amostragem por cotas (descrita na seção 4.1), resultando em tamanho de amostra de $n = 704$.

Definidos os métodos de amostragem probabilística, podemos agora definir as expressões para a estimação da média populacional e do correspondente erro padrão. Por Bolfarine & Bussab (2005) temos as fórmulas para os métodos de AAS e AES (Quadro 2). Inexiste estimador não viciado para o erro padrão do estimador da média sob amostragem sistemática, mas, por Wolter (1984), podemos utilizar o estimador para erro padrão sob amostragem aleatória simples sem reposição, dado que nossa população não apresenta tendência ou periodicidade.

Para o Quadro 2 temos que y_i é o valor da variável de interesse para a observação i, $\overline{y_h}$ é a média amostral para o estrato h, s^2 é a variância amostral

usada como estimativa para a variância populacional da variável de interesse, com s_h^2 sendo seu equivalente para o estrato h. Por conveniência chamamos de $\bar{y}$ a estimativa para a média populacional independentemente do tipo de amostragem e da fórmula utilizada.

Quadro 2 – Fórmulas para Estimação da Média Populacional e seu Erro Padrão

Método	Média	Erro Padrão
AASc	$\frac{\sum_{i=1}^{n} y_i}{n}$	$\sqrt{\frac{s^2}{n}}$
AASs	$\frac{\sum_{i=1}^{n} y_i}{n}$	$\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$
AES	$\sum_{i=1}^{L}\left(\frac{N_h}{N}\right)\cdot\overline{y_h}$	$\sqrt{\sum_{i=1}^{L}\left(\frac{N_h}{N}\right)^2\cdot\left(\frac{N_h-n_h}{N_h}\right)\cdot\frac{s_h^2}{n_h}}$

Fonte – Adaptado de Bolfarine & Bussab (2005).

4.SIMULAÇÃO MONTE CARLO

4.1 ALGORITMO

O algoritmo para amostragem por cotas descrito a seguir foi escrito na linguagem de programação R e se encontra disponível no repositório do primeiro autor[1]. Para os outros métodos se utilizam as funções já existentes no pacote *sampling* (Tillé & Matei, 2021). A função para amostragem por cotas precisa do banco de dados do qual será retirada a amostra, do tamanho da amostra e das variáveis categóricas do banco de dados escolhidas para serem os controles. Estas são usadas para a criação de um banco de dados auxiliar, que será pesquisado posteriormente.

Um valor qualquer entre 1 e o tamanho da população é sorteado para ser o ponto de largada (PdL), que garantirá amostras diferentes. Para garantir que a amostra seja completada mesmo quando este valor é alto (maior ou igual ao tamanho da população menos três vezes o tamanho da amostra), o banco de dados foi duplicado, dando origem a um banco de dados auxiliar que tem o dobro do tamanho do banco fornecido para seleção da ApC. O valor 3 foi escolhido de maneira empírica.

[1] https://github.com/JGMalaguti/ApC

Uma combinação dos níveis das variáveis de controle fornecidas é computada para determinar as cotas possíveis de modo a criar uma "matriz de teste", uma matriz contra a qual serão testados os elementos para se discernir a qual cota pertencem. O tamanho de cada cota é então calculado, de maneira idêntica à alocação proporcional para amostragem estratificada, e define o tamanho de cada elemento de uma lista onde serão armazenadas as identificações dos escolhidos para a amostra.

Finalmente, se inicia o loop que, a partir do ponto de largada, percorre o banco de dados auxiliar verificando a qual categoria pertence a observação. Se a lista de observações selecionadas nesta categoria ainda não foi totalmente preenchida, a observação é adicionada, caso contrário é descartada. O *loop* prossegue até encontrar o número desejado de observações para cada cota definida, a cada passo armazenando as identificações das observações de cada uma das cotas numa lista específica. Ao final do *loop*, a coleção de listas com as observações selecionadas para a ApC é transformada num vetor, que então é organizado e retornado pela função.

4.2 ESTUDOS DE SIMULAÇÃO

Simulação Monte Carlo (MC) é um método computacional para se estimar respostas para problemas probabilísticos a partir de amostras independentes (Brandimarte, 2014; Lewis & Orav, 1989). Sua precisão é relacionada ao número de réplicas ou rodadas do estudo, de modo que quanto maior o número de réplicas, maior a precisão.

O código para os estudos de simulação propriamente ditos é simples. Em cada réplica, retiram-se as amostras (por todos os métodos) e se estimam a média e o erro padrão. Depois de todas as iterações as médias de cada estatística são calculadas, assim como os intervalos de confiança e o nível de cobertura. Essas médias das estatísticas realizadas com os valores de todas as réplicas são chamadas médias de Monte Carlo e grafadas neste artigo como $\widehat{E(\bar{y})}$, $\widehat{E(EP)}$ e $\widehat{E(NC)}$.

O nível de cobertura é a média de uma variável indicadora que é igual a 1 se a média verdadeira se encontra no intervalo de confiança calculado a partir de cada amostra e 0, caso contrário. Para um nível de confiança de 95% espera-se um nível de cobertura nominal de 95%. A partir do nível de cobertura pode-se ter uma estimativa para o valor efetivo do nível de confiança.

Foram feitos dois estudos para poder entender melhor os efeitos relacionados a amostragem: para o primeiro estudo, o tipo de amostragem é fixo

(AASs) e o número de réplicas varia entre 10 e 100.000. O principal estudo fixa a quantidade de iterações em 100.000 e varia o tipo de amostragem.

5.RESULTADOS

5.1 EFEITO DA QUANTIDADE DE RÉPLICAS

O intuito deste estudo é analisar a influência do número de réplicas sobre as estatísticas, para isso selecionamos 10, 100, 1000, 10000 e 100000 amostras aleatória simples sem reposição de tamanho 700 da população e analisamos o comportamento dos estimadores em cada um desses casos (Tabela 3).

Tabela 3 - Tabela sobre a Influência da Quantidade de Simulações sobre as Estatísticas

Réplicas	$\left\|\widehat{E(\bar{y})} - \bar{Y}\right\|$	$\widehat{E(EP)}$	$\widehat{E(NC)}$ (%)
10	0,0945	0,495	100,000
100	0,0701	0,512	97,000
1000	0,0025	0,507	95,200
10000	0,0085	0,508	94,740
100000	0,0001	0,508	94,611

Com relação à estimação pontual da média, percebe-se que quanto mais réplicas são selecionadas, mais próximas as estimativas das médias de Monte Carlo ficam da média populacional. Por esses motivos decidiu-se fazer 100.000 iterações para o estudo principal pois com esse número de simulações garantimos uma estimativa pontual próxima do verdadeiro valor do parâmetro, além de nos permitir uma precisão maior na apresentação do nível de cobertura.

5.2 EFEITO DO TIPO DE AMOSTRAGEM

Temos então o principal estudo de simulação feito, comparando os tipos de amostragem (Tabela 4).

Estimou-se também o erro padrão por Monte Carlo (EP_{MC}), isto é, calculou-se o erro padrão utilizando os cem mil valores estimados para a média. Por Brandimarte (2014), sabemos que quando o número de réplicas é grande o suficiente, EP_{MC} tende ao erro padrão real das estimativas da média. Para os métodos probabilísticos, a estimação por MC tende a $\widehat{E(EP)}$, com a exceção da amostragem sistemática, que é mais distante da média de Monte Carlo da estimativa do erro padrão, demonstrando uma pequena sobre-estimação quando se utiliza a fórmula para AASs com as amostras sistemáticas.

Tabela 4 - Estatísticas Resumo da Simulação por Tipo de Amostragem

Tipo	$\left\|\widehat{E(\bar{y})} - \bar{Y}\right\|$	$\widehat{E(EP)}$	$\widehat{E(NC)}$ (%)	EP_{MC}
AASc	0,0006	0,509	94,458	0,512
AASs	0,0010	0,508	94,615	0,508
ASis	0,0010	0,508	96,670	0,453
AESu	0,0003	0,922	92,058	0,920
AESp	0,0012	0,482	94,636	0,484
AESo	0,0038	0,471	94,637	0,467
ApC_{AAS}	0,0787	0,462	40,777	4,075
ApC_{AES}	0,1256	0,513	43,324	4,064

Podemos corroborar a teoria da área (Cochran, 1977), tanto na proximidade dos resultados de AASs e AASc, quanto no fato que a amostragem sistemática possui nível de cobertura efetivo maior do que ambos os tipos de amostragem aleatória simples.

Focando apenas nas amostragens estratificadas, temos que AESu apresenta menor precisão e AESo a maior, como esperado, por usar mais informação sobre a população para alocar a amostra.

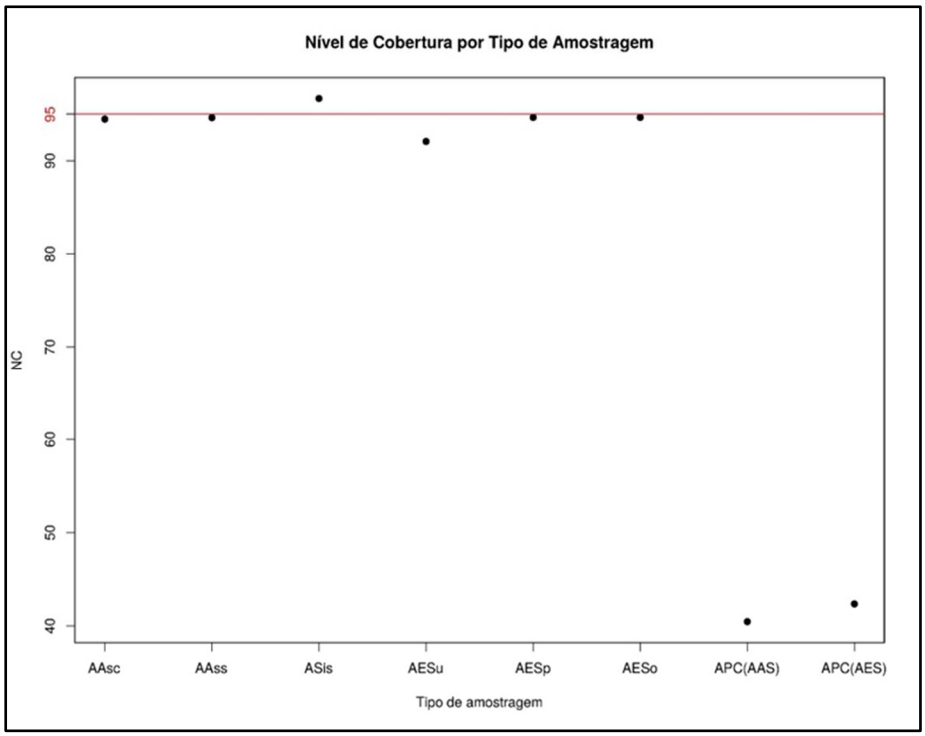
Os métodos probabilísticos produziram estimadores cujas distribuições tiveram médias que se aproximaram mais da média populacional, com nível de cobertura próximo do desejado (95%), enquanto que as duas variantes da amostragem por cotas, independente das fórmulas usadas, produziram estimativas cujas distribuições tiveram médias que ficaram mais distantes da média populacional. Além disso, para esse método de amostragem, a cobertura efetiva dos intervalos de confiança ficou próxima de 40%, como pode ser visto na Figura 1.

Comparando agora as amostras por cotas entre si, temos que aquela que usa as fórmulas AAS tem distribuição com média mais próxima do valor populacional e menor EP. Ambas as amostras apresentam um nível de cobertura baixo, mas próximo entre elas.

O uso das fórmulas AES na ApC_{AES} provocou um aumento do nível de cobertura, de aproximadamente 40% para aproximadamente 42%. Este aumento resultou do aumento das estimativas do erro padrão, que expandiu a amplitude dos intervalos de confiança. Como a média dos erros padrão se

aproximou do valor obtido sob AASs, o déficit de cobertura é explicado principalmente pelo viés nas estimativas pontuais.

Figura 1 - Nível de Cobertura por Tipo de Amostragem



Podemos notar também que o erro padrão por Monte Carlo é quase 8 vezes maior do que a média de Monte Carlo dos valores estimados utilizando as fórmulas para os métodos probabilísticos ($\widehat{E(EP)}$), mostrando que ao utilizar as fórmulas, se subestima o erro padrão para a média.

6.CONSIDERAÇÕES FINAIS

É importante notar que o algoritmo não consegue traduzir o viés dos pesquisadores ou as percepções dos mesmos em relação a possíveis entrevistados, tornando, portanto, as estimativas de erro padrão em subestimativas. Moser & Stuart (1953) citam tal tipo de viés como um dos maiores provindos da amostragem por cotas.

Outro viés não traduzido de maneira computacional neste estudo é o viés que provêm de uma amostragem em vários estágios, o que geralmente ocorre em pesquisas de grande porte, como pesquisas eleitorais de um país. Existe uma quantidade de erro amostral em cada estágio que é ignorada quando se utilizam as fórmulas AAS ou AES. Mesmo neste estudo em que se trabalha apenas com amostragem em um estágio (quando na prática geralmente se trabalham com múltiplos), o erro padrão é subestimado quando se utiliza as fórmulas que presumem aleatoriedade, sendo cerca de 8 vezes maior.

Mesmo assim, os institutos de pesquisa privados publicam margens de erro e nível de confiança (100-α) que presumem amostragem aleatória simples, apesar de não utilizarem esse método. Utilizando um α inicial de 5% (confiança de 95%) para o cálculo do tamanho da amostra, verificamos que o nível de cobertura efetivo de intervalos de confiança de nível (100-α) baseados nas amostras por cotas se encontra próximo de 40% para as estimativas obtidas usando fórmulas AAS e de 42% para as estimativas obtidas usando fórmulas AES.

Em suma, o método de amostragem por cotas não possui a precisão que é publicada. De fato, o método não pode apresentar estimativas de precisão, visto que não existe base teórica para esse cálculo, já que as fórmulas existentes presumem aleatoriedade. Um método probabilístico de amostragem, usando o mesmo tamanho total de amostra, é preferível pois permite, além da obtenção de estimativas pontuais sem viés, cálculo seguro do nível de confiança e da margem de erro.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Bolfarine, H. & Bussab, W. (2005). *Elementos de amostragem* (1ª ed.). São Paulo: Blucher.

Brandimarte, P. (2014). *Handbook in Monte Carlo Simulation – Applications in Financial Engineering, Risk Management, and Economics* (1a ed.). Hoboken: John Wiley & Sons.

Brasil. Lei nº 9.504, de 30 de setembro de 1997. Estabelece normas para as eleições. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9504.htm. Acesso em: 25 jan. 2021.

Brasil. Lei nº 12.891, de 11 de dezembro de 2013. Altera as Leis nºs 4.737, de 15 de julho de 1965, 9.096, de 19 de setembro de 1995, e 9.504, de 30 de setembro de 1997, para diminuir o custo das campanhas eleitorais, e revoga dispositivos das Leis nºs 4.737, de 15 de julho de 1965, e 9.504, de 30 de setembro de 1997. Disponível em: https://bit.ly/3jlukDG. Acesso em: 25 jan. 2021.

Brasil. Resolução nº 23.600, de 12 de dezembro de 2019. Dispõe sobre pesquisas eleitorais. Disponível em: https://bit.ly/39efLy9. Acesso em: 25 jan. 2021.

Brendolin, N.; Souza, D.; Pessoa, D. G. C.; Onel, S.; Quintaes, V. (2013) Indicadores de pobreza nos municípios de Minas Gerais: comparação de métodos de estimação em pequenas áreas. Rio de Janeiro: Seminário de Metodologia do IBGE de 2013. URL: https://eventos.ibge.gov.br/smi2013/atividades/sessoes-tematicas/st4-modelagem-para-indicadores-de-pobreza.

Carvalho, J. F. & Ferraz, C. (2006). A falsidade das margens de erro de pesquisas eleitorais baseadas em amostragens por quotas. Debate: A estatística na pesquisa eleitoral, Boletim da Associação Brasileira de Estatística, São Paulo.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.

Elliott, M. R. & Valliant, R. (2017). Inference for nonprobability samples. Statistical Science, 32(2), 249-264.

Ferraz, C. (1996). Crítica metodológica as pesquisas eleitorais no Brasil. (Dissertação de mestrado). Universidade Estadual de Campinas (UNICAMP), Campinas.

Katz, D. (1949). An analysis of the 1948 polling predictions. Journal of Applied Psychology, 33(1), 15.

Kish, L. (1965). *Survey sampling.* Wiley-Blackwell.

Lewis, P. A.W. & Orav, E. J. (1989). *Simulation Methodology for Statisticians, Operations Analysts and Engineers*, Volume 1. (1a ed.) Belmont: Wadsworth.

Marsh, C. & Scarbrough, E. (1990). Testing nine hypotheses about quota sampling, Journal of Market Research Society (JMRS), 32(4), 485-506.

Moser, C. A. (1952). Quota sampling. Journal of the Royal Statistical Society. Series A (General), 115(3), 411-423.

Moser, C. A., & Stuart, A. (1953). An experimental study of quota sampling. Journal of the Royal Statistical Society. Series A (General), 116(4), 349-405.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97(4), 558-625.

Pessoa, D. G. C. (2013). Pareamento Probabilístico na Pesquisa de Avaliação do Censo Demográfico de 2010. Rio de Janeiro: Seminário de Metodologia do IBGE de 2013. URL: https://eventos.ibge.gov.br/smi2013/atividades/sessoes-tematicas/djalma-galvao-carneiro-pessoa-ibge-pareamento-probabilistico-na-pesquisa-de-avaliacao-do-censo-demografico-de-2010.

Pessoa, D. G. C. e Silva, P. L. N. (1998). Análise de Dados Amostrais Complexos. São Paulo: Associação Brasileira de Estatística (ABE).

Pessoa, D. G. C.; Silva, P. L. N.; Lila, M. F. (2002) Análise Estatística de Dados da PNAD: Incorporando a Estrutura do Plano Amostral. Boletim ABRASCO, 7, 659-670.

Pessoa, D. G. C.; Silva, P. L. N.; Santos, A. R. (2000) Imputação para Não Resposta Parcial da Renda na Pesquisa Mensal de Emprego do IBGE. Relatório Técnico. Rio de Janeiro: Departamento de Metodologia, Diretoria de Pesquisa, IBGE. URL: https://biblioteca.ibge.gov.br/visualizacao/livros/liv83461.pdf.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Souza, J. (1990). *Pesquisa Eleitoral: Criticas e Técnicas*. Brasília: Editora do Senado.

Tillé, Y. & Matei, A. (2021). *sampling: Survey Sampling.* R package version 2.9. URL https://CRAN.R-project.org/package=sampling.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. Journal of Survey Statistics and Methodology, 8(2), 231-263.

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. Journal of the American Statistical Association, 79(388), 781-790.

Yang, K. & Banamah, A. (2014). Quota sampling as an alternative to probability sampling? An experimental study. Sociological Research Online, 19(1), 56-66.

AGRADECIMENTOS e COLABORAÇÕES

International Association of Survey Statisticians (IASS); CISDESTE; CNPq; FAPEMIG; Alfredo Chaobah.

Revista Brasileira de Estatística, V.79, N. 246, p.118-143, janeiro-junho 2021

COMPARANDO O DESEMPENHO DE MÉTODOS DE PAREAMENTO PARA INTEGRAR DADOS SOBRE A AGROPECUÁRIA

Andrea Diniz da Silva

andrea.diniz100@gmail.com

Escola Nacional de Ciências Estatísticas – ENCE/IBGE

Djalma Galvão Carneiro Pessoa

Escola Nacional de Ciências Estatísticas – ENCE/IBGE

José André de Moura Brito

jambrito@gmail.com

Escola Nacional de Ciências Estatísticas – ENCE/IBGE

Flavio Pinto Bolliger

fbolliger@gmail.com

Food and Agriculture Organization of the United Nations - FAO

Resumo: Importantes indicadores econômicos e sociais, tais como indicadores de desenvolvimento sustentável e produto interno bruto, dependem da produção contínua de estatísticas agropecuárias de boa qualidade. Porém, a despeito da importância do tema, não é raro que sistemas nacionais de estatísticas agropecuárias sejam baseados em censos decenais, cujos dados se tornam obsoletos rapidamente; e pesquisas por amostragem não probabilística, que não permitem estimar erro amostral ou medidas de precisão. Para melhorar a produção de tais estatísticas, métodos de pareamento têm sido utilizados com sucesso para integrar dados de registros e pesquisas, a custos relativamente baixos. O presente artigo relata um estudo empírico, replicado para três Estados brasileiros, conduzido para identificar um método de pareamento com resultados comparativamente melhores. Foram comparados oito métodos, utilizando nome e endereço de estabelecimentos agropecuários registrados no Cadastro Central de Empresas do IBGE ou nas Secretarias Estaduais de Fazenda dos estados do Maranhão, Paraíba e Santa Catarina. Mais especificamente, foram considerados: um método baseado no modelo de decisão de Fellegi-Sunter, dois métodos baseados no algoritmo K-Means, quatro métodos usando árvores de classificação e um método baseado no algoritmo SVM. Os resultados mostram superioridade dos métodos baseados em árvores de classificação e no SVM, que, em geral, resultaram em maiores proporções de comparações corretamente classificadas.

Palavras-chave: Pareamento de dados; integração de dados; Fellegi-Sunter; aprendizado de máquina.

Abstract: Important economic and social indicators, such as sustainable development indicators and gross domestic product, depend on the continuous production of good quality agricultural statistics. However, despite the importance of the topic, it is not uncommon for national systems of agricultural statistics to be based on decennial censuses, whose data quickly become obsolete; and non-probabilistic sample surveys, which do not allow estimating sampling error or precision measures. Record linkage methods have been used successfully to integrate data from registers and surveys and improve the production of agricultural statistics at relatively low cost. This article reports an empirical study, replicated for three Brazilian States, conducted to identify a record linkage method with comparatively better results. Eight methods were compared, using the name and address of agricultural establishments registered in the business register of IBGE or in the state tax administration systems in the states of Maranhão, Paraíba and Santa Catarina. The methods considered include: a method based on the Fellegi-Sunter decision model, two methods based on the K-Means algorithm, four methods using classification trees and one method based on the SVM algorithm. The general results show the superiority of the methods based on the classification trees and the SVM. In most cases, these methods resulted in higher proportions of correctly classified comparisons.

Keywords: record linkage; data integration; rural statistics; Fellegi-Sunter; machine learning.

1.INTRODUÇÃO

Há décadas a segurança alimentar se tornou um desafio e uma prioridade estratégica para a maioria dos países. Na atualidade, constitui um dos centros em torno dos quais se estrutura e se encaminha a agenda global orientada para o futuro da humanidade em bases sustentáveis. No caso da produção agropecuária, as políticas agrícolas têm empreendido esforços para equacionar a demanda crescente por alimentos suficientes, adequados, com preços acessíveis e a observância dos cuidados necessários com o meio ambiente, condição *sine qua non* para assegurar a capacidade produtiva ao longo do tempo.

A produção de estatísticas agropecuárias de qualidade constitui tarefa fundamental não apenas no que concerne à gestão de políticas agrícolas eficazes, mas também para elaboração de indicadores econômicos e sociais essenciais. Produto Interno Bruto, indicadores de desenvolvimento sustentável e para monitorar a erradicação da fome e da desnutrição, crescimento da produção e uso sustentável dos recursos naturais, são alguns exemplos cuja operacionalidade depende da produção contínua de estatísticas agropecuárias diversificadas, oportunas e de alta qualidade. Apesar de sua relevância, entretanto, não é raro que os sistemas nacionais de estatísticas agropecuárias sejam fundamentalmente baseados em censos de periodicidade decenal, cujos dados se tornam obsoletos rapidamente, e em pesquisas por amostragem não probabilística, que têm escopo limitado e não permitem que sejam feitas estimativas do erro amostral ou de medidas de precisão.

A despeito das limitações que se impõem à produção de estatísticas oficiais, a existência de fontes de dados públicas e privadas possibilita a estruturação de um sistema para a produção de estatísticas agropecuárias, baseado em pesquisas e registros administrativos. A integração de dados provenientes de tais fontes pode viabilizar a produção direta de informações sobre a agropecuária. Além disso, a base de dados integrada pode ser utilizada como cadastro de unidades para a seleção de amostras probabilísticas para a realização de pesquisas, sendo possível, neste caso, a realização de estimativas e o cálculo de medidas de precisão.

Porém, apesar da existência de fontes de dados, a simples agregação do seu conteúdo não possibilita a produção de estatísticas, considerando a necessidade de atender aos padrões de qualidade adotados na produção das estatísticas oficiais. Isto ocorre, principalmente, porque alguns registros estão incompletos e porque algumas unidades possuem mais de um registro, ou seja, estão duplicadas. Sendo assim, uma questão que precisa ser enfrentada é a necessidade de identificar os registros pertencentes a uma mesma unidade para, então, completar as informações com dados de mais de uma fonte e remover aquelas que estão duplicadas.

Se, por um lado, nem todas as fontes de dados dispõem de um identificador único, o que facilitaria sobremaneira a identificação dos registros pertencentes à mesma unidade, dentro e entre bases; por outro, há hoje uma variedade de métodos de pareamento estatístico que não dependem de tal informação. Não obstante, a despeito da variada gama de métodos, não há um cuja eficácia seja superior em relação aos demais, exceto em aplicações específicas. Isto ocorre porque o desempenho de cada método depende, pelo menos, da estrutura e da qualidade dos dados utilizados. Para contribuir com esta discussão, o trabalho realizado teve como foco principal comparar métodos de pareamento estatístico e identificar aquele que apresentasse melhores resultados para a integração de dados sobre a agropecuária brasileira.

O artigo está dividido em cinco seções, incluindo essa introdução. Na segunda seção é apresentado um panorama de métodos de pareamento de dados desenvolvidos nas últimas décadas. Na terceira parte é apresentada uma visão geral sobre pareamento, em particular, descrevendo o processo e as etapas de um método genérico. Na quarta seção são apresentadas as fontes dos dados utilizadas e os procedimentos adotados no pareamento dos dados, além dos resultados gerais do pareamento utilizando oito métodos. Por fim, na última seção são apontadas algumas possibilidades de extensão do presente trabalho.

O presente artigo mostra resultados da pesquisa desenvolvida durante a elaboração da tese de doutoramento da autora (Silva, 2018), que contou com a orientação do Prof. Djalma Galvão Carneiro Pessoa. Com sua costumeira generosidade e interesse teórico e acadêmico, Djalma deu contribuição decisiva para a escolha, aplicação e avaliação dos métodos comparados, bem como na análise dos resultados.

2.UM POUCO DE HISTÓRIA

Desde o surgimento dos primeiros registros, listas, arrolamentos, onde informações individuais são relacionadas, a realização de pareamento de registros tem sido uma estratégia para juntar informações de um mesmo indivíduo, quer seja ao longo do tempo ou quando provenientes de duas ou mais fontes distintas. Nesse sentido, é possível afirmar que pareamento de dados existe desde que existem registros (Fellegi, 1997) e, portanto, remonta pelo menos ao ano 2.238 a.C., quando o imperador da China, Yao mandou realizar um censo da população e das lavouras cultivadas (IBGE, 2012).

O primeiro registro sobre o uso do termo *record linkage*, traduzido aqui como pareamento de dados, ou simplesmente pareamento, é observado em 1946 por Halbert Dunn, que o definiu como um processo usado para reunir fatos sobre a vida de uma pessoa, com a finalidade de compor um livro da vida. Dunn estava tratando das ocorrências e eventos vividos por um indivíduo entre seu nascimento e sua morte. Desde então, muitas outras definições surgiram para descrever pareamento de dados, na maioria dos casos, convergindo para a ideia de reunir informações sobre o mesmo indivíduo ou evento, provenientes de fontes de dados diferentes.

Inicialmente, o pareamento era realizado por meio de inspeção visual de listas, ordenadas alfabeticamente pelo nome ou outro atributo individual. O primeiro experimento considerando pareamento automatizado é atribuído a Newcombe et al. (1959). Os autores registraram os primeiros passos do desenvolvimento de um método para parear dados de um mesmo indivíduo provenientes de duas bases de dados de forma automática, ou seja, com o uso de ferramentas computacionais. Além do pioneirismo do exercício feito, contribuições como a introdução de definições e a discussão de possibilidades e limites do uso de registros administrativos também foram propostas para a reflexão. Quando se trata de pareamento, não se pode deixar de reconhecer a contribuição dada por Newcombe e colegas, por terem reconhecido o pareamento como um problema estatístico: na presença de erro na informação decidir que dupla de registros deve ser classificada como par. (Fellegi, 1997).

A partir dos anos 1960, quatro fatores influenciaram a evolução do pareamento: oportunidade, meios, necessidades e restrições (Fellegi, 1997). O primeiro associado à criação de cadastros para atender a evolução do estado de bem-estar social e do sistema tributário, ocorrida pós Segunda Guerra; o segundo ligado ao rápido desenvolvimento tecnológico, resultando no aumento da capacidade de memória e de processamento dos computadores de grande porte e no surgimento dos computadores pessoais nos anos 70; o terceiro resultante do aumento da demanda por informações para a tomada de decisão, diante da descentralização político administrativa e da criação de esferas participativas populares; e, por fim, um quarto fator, com efeito restritivo, presente em alguns países, o aumento dos níveis de preocupação dos indivíduos com a privacidade e segurança dos dados.

Neste cenário, Fellegi & Sunter (1969) trataram de formular um modelo matemático para servir de arcabouço teórico de uma solução baseada no uso de recursos computacionais. Em sua formulação, os autores propuseram uma ampliação das categorias utilizadas por Newcombe e colegas para classificar duplas de registros que pertencem ou não pertencem ao mesmo indivíduo, na medida em que estavam interessados não apenas em saber se os dois registros comparados pertencem ao mesmo indivíduo ou não, mas também se há evidências suficientemente fortes para justificar a decisão com determinado nível de erro.

Sendo assim, propuseram a adição de uma categoria que pudesse ser utilizada para classificar as duplas sobre as quais não houvesse evidências suficientemente fortes sobre sua verdadeira categoria. Assim, passaram a trabalhar com três categorias ou classes: Par, Não-Par e Possível-Par. Graças à teoria de pareamento formulada pela dupla Fellegi-Sunter, a realização de pareamento estatístico se popularizou. Por fornecer os elementos principais para a realização do pareamento, ela é, ainda hoje, muito utilizada como base de vários métodos de pareamento. Porém, atualmente existe uma variada gama de aplicações reais onde são adotadas simplificações e modificações ad hoc para assegurar maior eficácia. Como resultado, as soluções adotadas não podem ser generalizadas e aplicadas a outros casos de pareamento, e nem sempre se aproveitam dos avanços tecnológicos e metodológicos já obtidos (Winkler, 1993).

Na tentativa de obter refinamentos que permitissem tanto melhorar os resultados do pareamento quanto a generalização da teoria base desenvolvida pelos autores, alternativas e extensões ao seu trabalho têm sido propostas ao longo dos últimos anos. Duas das principais contribuições registradas são o uso

de funções de comparação que permitem calcular valores de similaridade quando a variável comparada contém variações tipográficas, portanto de concordância parcial (Winkler, 1990), e o uso do algoritmo Expectation-Maximisation para obter melhores estimativas da probabilidade de que determinada dupla de registros pertença e de que não pertença ao mesmo indivíduo. (Winkler, 2000b).

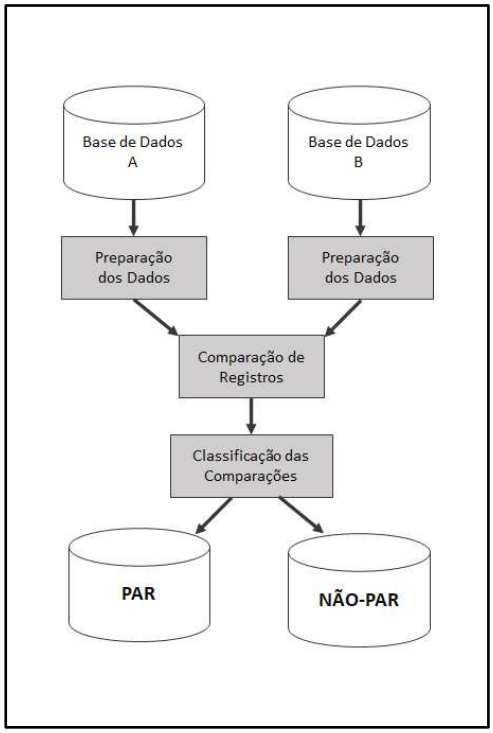
A despeito dos avanços obtidos, tal modelo de decisão tem uma fragilidade intrínseca, qual seja, se basear em pressuposto de independência entre os erros de registro das variáveis utilizadas no pareamento, portanto de que a probabilidade de concordância em uma variável, digamos primeiro nome, independe da probabilidade de concordância em outras variáveis de nome, endereço etc. Tal pressuposto não se verifica na prática; portanto, uma regra de decisão que maximize as probabilidades de classificação correta nem sempre pode ser especificada (Smith & Newcombe 1975; Winkler, 1989, 1995, 2000a e 2000b).

Uma abordagem diferenciada para enfrentar o problema de pareamento, livre da necessidade de pressuposto de independência, surgiu no fim da década de 90, com o início da adoção de técnicas já utilizadas em mineração de dados e recuperação de informação, como as de inteligência artificial, particularmente aprendizado de máquina (Christen, 2007). Trilha seguida nos anos iniciais da década de 2000, quando foram desenvolvidos métodos de pareamento baseados em algoritmos desenvolvidos a partir dos anos 70, como é o caso do K-means (Hartigan & Wong, 1979); da árvore de classificação (Breiman, 1998); e dos algoritmos baseados em redes neurais artificiais, reeditados nos anos 80 (Haykin, 1999); máquinas de vetores-suporte (Boser et al., 1992; Vapnik, 1998); e classificadores Bayesianos (James et al., 2015). Porém, ainda restou para a próxima geração o desafio de encontrar métodos que lidem bem com a integração de dados provenientes de fontes classificadas como *big data*.

3.MÉTODO EM TRÊS ETAPAS

O método de pareamento proposto é um método em três etapas, envolvendo preparação dos dados, comparação de todas as duplas possíveis de registros e classificação de cada comparação como Par ou Não-Par. Tal método pode ser mais bem compreendido a partir de um fluxo de processo (Figura 1) seguido por uma descrição de suas etapas, conforme se segue.

Figura 1 - Processo de pareamento em três etapas



O fluxo descreve um método que, além de contemplar todas as etapas de um processo completo, também se aplica aos casos em que são utilizados dados sintéticos e conjuntos de dados já limpos e padronizados.

A escolha das técnicas que serão utilizadas em cada etapa depende da disponibilidade de dados e recursos, mais especificamente, da estrutura de dados, completude e precisão, tamanho do conjunto de dados, recursos computacionais, equipe etc. Um dos principais problemas atuais é a distinção entre dados estruturados e não estruturados, introduzido no contexto do uso de fontes descritas como *big data*. Uma vez que as condições sejam conhecidas, o método poderá ser então desenhado, selecionando as técnicas que se adequam aos recursos do projeto. Várias técnicas para lidar com as questões relacionadas a cada etapa do processo de pareamento, que incluem procedimentos, funções e algoritmos, foram desenvolvidas nos últimos 60 anos. Uma lista não exaustiva é apresentada a seguir.

3.1. ETAPA UM: PREPARAÇÃO

A preparação de dados raramente é considerada uma componente-chave de um procedimento científico. Porém, a preparação inadequada, especialmente falhas no processo de limpeza e padronização dos dados, pode levar a uma redução de 20% a 30% no número de pares corretos (Winkler, 2014). Para mitigar o impacto dessa fonte de erro, um projeto de pareamento deve

dedicar uma parte significativa de seus recursos à preparação de dados. Além da limpeza e padronização dos dados, a avaliação da qualidade das variáveis que serão utilizadas no pareamento é fundamental para garantir a qualidade do resultado final.

Critérios como integridade, consistência e distribuição de frequência podem ser usados para avaliar e selecionar variáveis de pareamento. As contagens e as taxas de resposta do item podem ajudar a medir a integridade. No entanto, taxa de resposta próxima a um significa apenas que não há dados ausentes e não que todas as respostas são válidas ou aceitáveis. Assim, a consistência também deve ser considerada e pode ser avaliada usando contagens e taxas, mas levando em consideração apenas respostas aceitáveis. A distribuição das respostas deve ser considerada ao escolher variáveis de pareamento, uma vez que concentrar os valores em um ou em alguns valores específicos pode reduzir o poder explicativo da variável. Existem várias medidas de concentração que podem ser usadas para realizar essa avaliação. As mais populares, principalmente devido à sua interpretabilidade e comparabilidade, são o Índice de Entropia (Shannon, 1948) e o Índice de Gini (Gini, 1912; Ceriani & Verme, 2012). Para variáveis categóricas, métodos gráficos podem ser usados.

Identificar nível e padrão de erro das variáveis permite correção automática. A forma como as datas de nascimento, nomes e endereços são registrados pode variar consideravelmente entre as fontes, mas a variação pode seguir um ou mais padrões identificáveis em cada banco de dados. Levando isso em consideração, Winkler (2014) e Christen (2012) propõem um conjunto mínimo de ações para padronização de dados que devem ser consideradas em qualquer projeto de pareamento. Tal conjunto inclui a remoção de caracteres indesejados, como vírgulas, ponto e vírgula, ponto, barra, aspas etc.; expansão de abreviações e acrônimos; adoção do padrão DDMMAAAA para datas; adoção do padrão F/M para mulheres e homens, respectivamente; correção/supressão de valores de atributos numéricos, como ano de nascimento ou idade, fora dos intervalos válidos; decomposição de variáveis que geralmente são expressas por mais de um componente, como nome e endereço.

Atenção especial deve ser dada ao selecionar variáveis de blocagem, se aplicável, já que tais variáveis são utilizadas para dividir os registros em blocos dentro dos quais estes serão comparados dois a dois. Sua escolha deve levar em consideração dois efeitos esperados da blocagem: os registros dentro dos blocos devem ter mais chance de parear entre si que com registros de fora do bloco (O'Hare et al., 2019) e os tamanhos dos blocos, dados pelo número de comparações, devem ser adequados à capacidade de processamento dos

recursos computacionais disponíveis. Adicionalmente, além do padrão de qualidade aplicado às variáveis de pareamento, as variáveis de blocagem, idealmente, não devem ter valor faltante. Apesar de ser possível considerar os valores perdidos como um bloco, isso pode levar à comparação de registros em blocos diferentes.

3.2. ETAPA DOIS: COMPARAÇÃO

Comparar registros quando não há identificador único em ambos os conjuntos de dados não é uma tarefa trivial. O uso de quase identificadores como nome, endereço e a maioria dos atributos relacionados com características socioeconômicas dos indivíduos aumenta o risco de classificação incorreta porque suas informações podem não ser suficientemente completas ou precisas. Assim, quando esses quase identificadores são usados para medir o quão semelhantes os registros são, a medida de similaridade obtida também pode ser imprecisa. Dois dos problemas mais importantes enfrentados ao usar quase identificadores estão relacionados a erros no registro de informações sobre indivíduos e a presença de homônimos. Ambos são até certo ponto inevitáveis. Para resolver o primeiro problema, o uso de uma função de comparação que permita a atribuição de valores de similaridade quando há concordância parcial deve ser preferível ao uso de funções de comparação baseadas na similaridade exata. Para o segundo, a redução do impacto da presença de homônimos pode ser obtida atribuindo-se pesos inversamente proporcionais à sua frequência. Portanto, maior peso pode ser atribuído a nomes menos frequentes, conforme proposto por Newcombe e coautores (1959).

Nos últimos 30 anos, várias funções de comparação foram propostas em alternativa à classificação binária. As mais populares são baseadas no conceito de distância, como o número mínimo de operações necessárias para que duas variáveis se tornem iguais, além do número e a posição dos caracteres coincidentes. Este é o caso das funções Hamming, que contam o número de substituições necessárias para fazer duas variáveis textuais exatamente iguais, adicionando o custo de cada substituição ao valor de similaridade; Episode, que considera apenas inserções; Longest Common Subsequence, que conta o número de inserções e exclusões necessárias; e a distância de Levenshtein, que considera inserções, exclusões e substituições necessárias, ver Christen (2012). Essas funções de comparação podem ser aplicadas para comparar data, idade, hora, coordenadas geográficas, bem como nomes. Nas funções descritas, diferentes pesos são associados às operações de inserção, exclusão e substituição.

Outra abordagem considera o número de partes coincidentes de uma variável, como a função Q-gram, que usa o conceito de uma janela móvel contendo q caracteres, geralmente dois ou três. A janela começa com o primeiro caractere do texto, vai para o segundo e assim por diante. Cada conjunto de caracteres na janela é um q-grama. Depois de contar o número de q-gramas correspondentes, a similaridade é calculada, geralmente usando o coeficiente de Sobreposição, de Jaccard ou Dados (Christen, 2012).

Uma combinação de ambos os conceitos foi proposta por Winkler (1990), com base na função desenvolvida por Jaro (1989), dando origem à função Jaro-Winkler. O valor de similaridade é calculado com base no número de inserções, exclusões e transposições necessárias, sendo aumentado quando os primeiros caracteres concordarem. A aplicação da função de Jaro-Winkler resulta na atribuição de valores de similaridade entre zero e um. A similaridade máxima é igual a um e ocorre quando a variável comparada é totalmente igual em ambos os registros. Em pareamento, a função de comparação de Jaro-Winkler se mostrou mais eficiente que muitas funções propostas anteriormente. Em experimentos com grandes conjuntos de dados de empresas, Belin (1993) mostrou que o uso das funções de comparação Jaro e Jaro-Winkler melhorou o desempenho do pareamento quando comparadas às demais. Além disso, um estudo de Budzinsky (1991, appud Porter & Winkler, 1997) envolvendo a avaliação de vinte funções de comparação indicou que a função de Jaro-Winkler foi superior para comparar registros com algum nível de erro, geralmente na grafia.

3.3. ETAPA TRÊS: CLASSIFICAÇÃO

Decidir se dois registros, um representado em um conjunto de dados contendo N registros individuais e o outro em um segundo conjunto de dados contendo M registros individuais, correspondem ao mesmo indivíduo é basicamente um problema de classificação. Prever uma resposta qualitativa como “ambos os registros pertencem ao mesmo indivíduo” ou “os registros pertencem a indivíduos diferentes” é equivalente a atribuir uma categoria ou classe à comparação dos dois registros. Tal problema envolve um modelo de decisão que traduz a aplicação de uma regra para todas as comparações, envolvendo, assim, todos os registros de ambos os conjuntos de dados. O modelo de decisão pode ser determinístico quando é baseado em regras fixas; não determinístico quando baseado em regras flexíveis; ou ser híbrido combinando ambas as abordagens.

Os modelos de decisão determinísticos dependem de um conjunto de regras fixas, arbitradas a priori, que serão aplicadas a cada par de registros, independentemente de outros pares de registros; enquanto os não determinísticos são baseados em regras flexíveis que dependem do conjunto completo de pares de registros e geralmente são baseados em abordagens probabilísticas ou em um modelo preditivo construído em uma amostra de pares de registros previamente classificados.

Na classificação determinística binária, se dois registros estiverem em conformidade com determinada regra, eles são considerados um **Par**, caso contrário são considerados **Não-Par**. As regras mais comuns na classificação determinística envolvem igualdade total de nomes de pessoas, produtos, empresas etc. ou a escolha de margens de erro ou limites de tolerância para que a diferença entre dois valores, como a idade. Essas regras são estabelecidas antes do início do processo de classificação e não são alteradas durante o processo. Consequentemente, uma vez que o conjunto de regras é especificado, a classificação determinística é simples e direta. Um caso especial de classificação determinística é aquele que usa um identificador único, como CPF ou CNPJ.

A classificação não determinística envolve o desenvolvimento de um modelo para a construção de um classificador. Em pareamento, este modelo corresponde a um modelo estatístico para prever a classe à qual cada comparação pertence, digamos, par ou não-par, com base nos valores dos preditores, ou seja, variáveis de pareamento. Os classificadores não determinísticos podem ser de dois tipos: supervisionados e não supervisionados. Os classificadores supervisionados são construídos considerando o conhecimento prévio da classe a que pertence um subconjunto das comparações, ou seja, uma série de comparações previamente e corretamente classificadas como Par ou Não-Par. O conjunto desses pares de registros é chamado de conjunto de treinamento, dados de treinamento ou amostra de treinamento. Nesse caso, o classificador é construído com base nas informações obtidas do conjunto de treinamento e, em seguida, usado para classificar comparações para as quais a classe não é conhecida. Se tais informações não forem usadas para construir um classificador, mas apenas para “aprender” sobre a estrutura e o relacionamento entre as variáveis, o modelo de decisão é dito não supervisionado (James et al., 2015). Modelos de decisão com abordagens híbridas, que combinam o uso de classificadores supervisionados e não supervisionados, também foram testados com algum sucesso por Elfeky et al. (2002).

No pareamento de dados de pessoas, os classificadores não supervisionados mais populares são aqueles baseados no modelo de decisão proposto por Fellegi e Sunter (1969) e suas extensões, como o uso da função de comparação Jaro-Winkler (Winkler, 1990) para calcular os valores de similaridade das variáveis de pareamento e o algoritmo EM (Dempster et al., 1977) para estimar a probabilidade de uma comparação pertencer ao conjunto Par e Não-Par. A ideia básica dessa classe de modelos é calcular um peso que depende desses valores e, em seguida, classificar a comparação como Par se tal peso estiver acima de um valor arbitrado e como Não-Par caso contrário. Em teoria, os classificadores Fellegi-Sunter permitem que todos os verdadeiros pares sejam classificados como Par, simplesmente arbitrando esse valor tão pequeno quanto o peso mínimo. No entanto, isso também resulta na classificação incorreta de comparações cuja verdadeira classe é Não-Par.

Classificadores baseados no algoritmo de agrupamento K-means (Hartigan e Wong, 1979) são populares na literatura por sua interpretabilidade e facilidade de implementação. Como esses classificadores também não são supervisionados, eles não dependem do conhecimento prévio dos dados usados no pareamento, ou seja, de dados de treinamento. Um dos problemas dessa abordagem decorre do fato de que esses algoritmos geralmente produzem uma solução que corresponde a um ótimo local. Assim, mais de uma execução do algoritmo pode levar a diferentes configurações dos conjuntos Par e Não-Par. Como o algoritmo K-means é instável e execuções repetidas podem produzir soluções muito diferentes, o algoritmo bclust (Bagged Clustering) proposto por Leish (1999) é uma alternativa para obter resultados de classificação mais robustos. Os autores propõem combinar os resultados do K-means em um novo conjunto de dados e usá-los como entrada para um método hierárquico. A ideia geral é construir amostras tipo Bootstrap (Efron, 1979), executar o K-means em cada uma, combinar todos os k-centroides produzidos por esse algoritmo em um novo conjunto de dados e então executar um algoritmo de cluster hierárquico. Além de reduzir os efeitos da variação devido à característica do K-means citada acima, o uso de centroides em vez do conjunto completo de comparações reduz o tempo e a memória necessários para seu processamento.

Uma variedade de classificadores supervisionados surgiu nos últimos 30 anos. Dentre os mais populares estão os classificadores baseados em árvores de classificação, tanto em sua versão mais simples, com base em uma única árvore, quanto em variações que combinam várias árvores. Tais classificadores estão implementados em algoritmos que consideram partição binária recursiva, ou seja, divisão sucessiva dos elementos em dois grupos de acordo com os

valores de seus preditores. Em pareamento, usando classificação binária, as comparações são classificadas em Par ou Não-Par de acordo com os valores de similaridade das variáveis de pareamento. Inicialmente, o conjunto de comparações é dividido em dois grupos e, em seguida, cada grupo é dividido em dois e assim por diante. A classificação é feita buscando maximizar a homogeneidade dos valores de similaridade das variáveis de pareamento. A homogeneidade é geralmente medida pela variância, entropia ou índice de Gini (Breiman et al., 1984).

Classificadores baseados em árvore única são geralmente instáveis e com alta variação, pois pequenas mudanças nos dados de treinamento podem levar a mudanças substanciais no classificador, o que significa que o padrão de classificação pode variar muito, levando à composição de classes completamente diferentes (Breiman, 1998) quando o pareamento é repetido. No entanto, a combinação de vários classificadores baseados em uma única árvore pode reduzir a variação ou viés dos valores preditos. Alguns dos mais populares são Bagging (Bootstrap Aggregating), Bumping (Bootstrap Umbrella of Model Parameters) e ADA Boosting (Adaptive Boosting). Eles são desenvolvidos com base em amostras do tipo Bootstrap dos dados de treinamento. O classificador Bagging é a combinação linear de um número arbitrado de classificadores, o que na prática significa utilizar o valor médio ou o valor mais frequente da classe atribuída à comparação; o Bumping é o classificador que fornece um mínimo (local) para um critério de trabalho, geralmente complexidade medida pelo tamanho da árvore; e o classificador ADA é obtido pela combinação linear de todos os classificadores. Uma característica do ADA, que o distingue de outros classificadores genericamente chamados de boosting, é que ele considera o desempenho do classificador anterior para definir os dados de treinamento que serão usados para desenvolver o próximo classificador e permite obter um classificador baseado nas probabilidades dos vários classificadores desenvolvidos a partir das diferentes amostras, por isso é estocástico (Breiman, 1994 e 1998; Tibshirani & Knight, 1999; Friedman et al., 1998).

Outros classificadores supervisionados podem ser definidos a partir de algoritmos do tipo Support Vector Machine (SVM), conforme proposto por Cortes & Vapnik (1995). Os classificadores SVM são do tipo binário não probabilístico e funcionam como classificadores de margem máxima porque têm um mecanismo de busca que permite encontrar o limite de discriminação cuja distância dos valores mais próximos seja máxima. Em vez de trabalhar com limite de corte, como no modelo de decisão desenvolvido por Fellegi e Sunter, a ideia é encontrar um hiperplano ideal que separe as comparações em dois grupos (Par

e Não-Par). Na prática, vários hiperplanos podem ser encontrados, mas o hiperplano ótimo é aquele que garante a maior margem entre os dois grupos (Cortes & Vapnik, 1995; Vapnik, 1998).

Uma variedade de abordagens tem fornecido bons resultados quando utilizadas em métodos de pareamento. Vantagens associadas à interpretabilidade e eficácia da classificação são confrontadas com a necessidade de conhecer a distribuição dos pares verdadeiros e a má especificação do modelo com base na independência do erro, como é o caso dos classificadores bayesianos (James et al., 2015). Métodos de pareamento baseados nos algoritmos de classificação descritos foram testados com sucesso em muitas aplicações de dados reais e em uma variedade de trabalhos acadêmicos usando dados sintéticos ou simulações. Porém, até o momento, nenhum deles se mostrou superior aos demais, independentemente das condições materiais como recursos computacionais e tempo necessário, estrutura e qualidade dos dados.

Métodos de pareamento baseados no modelo de decisão Fellegi-Sunter têm sido amplamente usados em institutos nacionais de estatística para pareamento de dados de população. Embora com base em premissas de independência que na prática não se verificam, bons resultados têm sido alcançados. Por outro lado, métodos baseados em classificadores supervisionados têm apresentado melhores resultados em trabalhos acadêmicos. No entanto, tais métodos dependem do conhecimento prévio da verdadeira classe (Par ou Não-Par) associada a um subconjunto das comparações, ou seja, dados de treinamento, o que nem sempre é fácil de obter. Portanto, as escolhas que vão determinar o desenho do pareamento estão intrinsecamente associadas a fatores externos ao método.

4. PAREAMENTO USANDO NOME E ENDEREÇO DE ESTABELECIMENTOS AGROPECUÁRIOS

Nesta seção são apresentados os dados, os procedimentos e os resultados de uma aplicação com dados reais, replicada para três estados brasileiros, para a comparação do desempenho de oito métodos de pareamento para a integração de dados sobre a agropecuária, utilizando nome e endereço de estabelecimentos agropecuários como variáveis de pareamento.

4.1. DADOS E VARIÁVEIS

Os dados utilizados referem-se aos estabelecimentos agropecuários registrados em dois cadastros, quais sejam: Cadastro Central de Empresas do

Instituto Brasileiro de Geografia e Estatística (IBGE), doravante referido como CEMPRE, e da Secretaria Estadual de Fazenda, doravante referida como SEFAZ, dos estados do Maranhão (MA), Paraíba (PB) e Santa Catarina (SC). Vale registrar que tais estados foram os que, junto com Ceará e Mato Grosso do Sul, responderam positivamente ao pedido de acesso aos dados. Infelizmente, por limitações na capacidade computacional, estes dois últimos foram excluídos do estudo. O ano de referência do primeiro cadastro é 2014 e dos demais é 2016. Foram considerados, apenas, os dados de estabelecimentos não duplicados dentro de um mesmo conjunto e com informação de nome e identificador único. O total de registros finais em cada arquivo varia com o Estado (Tabela 1).

Tabela 1 - Número de estabelecimentos agropecuários envolvidos no pareamento, por fonte dos dados, segundo o Estado brasileiro

Estado	Estabelecimentos Agropecuários	
	CEMPRE	SEFAZ
Maranhão	820	750
Paraíba	334	348
Santa Catarina	2.173	235

Fonte: IBGE, Cadastro Central de Empresas (CEMPRE) - 2014 e Secretaria Estadual de Fazenda (SEFAZ) - 2016.

Informações básicas como nome, endereço e CNPJ foram as únicas informações disponibilizadas por todas as fontes de dados. Portanto, as primeiras foram utilizadas como variáveis de pareamento e a última como variável de verificação.

4.2. PAREAMENTO: PREPARANDO - COMPARANDO - CLASSIFICANDO

Oito métodos de pareamento não determinístico foram comparados, dentre os quais três são não supervisionados e cinco são supervisionados. Cada método consiste em três etapas: preparação, comparação e classificação, conforme descrito na seção 3. As duas primeiras etapas, preparação e comparação, foram realizadas estritamente da mesma forma para todos os métodos de pareamento, enquanto o algoritmo usado para classificar as comparações como Par e Não-Par variou em cada método. Assim, o que diferencia um método de pareamento de outro nesta aplicação é o algoritmo de classificação utilizado na terceira etapa.

A primeira etapa, preparação dos dados, incluiu a padronização do formato dos arquivos de dados recebidos, a criação de um código correspondente ao estado onde se encontra o estabelecimento, padronização

de nomes, transformação do CNPJ em variável categórica, padronização do código para dados faltantes, exclusão de registros sem nome ou CNPJ, conversão de maiúsculas para minúsculas, substituição de caracteres latinos especiais, padronização do tipo de rua e do nome do estabelecimento, separação de partes singulares dos nomes e remoção de pontuação, preposições e artigos, espaços duplos, no início e no final dos valores.

O conjunto final de variáveis de pareamento incluiu nome e endereço do estabelecimento, os quais foram divididos em partes singulares, conforme mencionado acima. Considerando a baixa frequência de nomes com mais de quatro partes, tanto para estabelecimento quanto para logradouro, foram utilizadas apenas as primeiras quatro partes dos nomes. Além disso, foi utilizado como variável de pareamento o número do endereço do estabelecimento. Isso resultou no uso de nove variáveis de pareamento: nome1, nome2, nome3, nome4, logradouro1, logradouro2, logradouro3, logradouro4 e número.

Na segunda etapa, para cada estado brasileiro, todos os estabelecimentos cadastrados no CEMPRE foram comparados com todos os estabelecimentos cadastrados na SEFAZ. O número de comparações varia conforme o estado, pois é resultado do produto cartesiano das quantidades de estabelecimentos em cada fonte (Tabela 2).

Tabela 2 - Número de pares de comparações, por estado

Estado	Comparações
Maranhão	615.000
Paraíba	116.232
Santa Catarina	510.655

Fonte - Proposta de Método de Pareamento para Integrar Dados sobre a Agropecuária. Tese de Silva, A. D. (2018).

Em cada comparação, um valor de similaridade entre 0 e 1 foi calculado para cada variável de pareamento, usando a função de comparação Jaro-Winkler (Winkler, 1990). Como resultado, n vetores de tamanho nove foram gerados, sendo n o número de comparações (Tabela 2), contendo os valores de similaridade das nove variáveis de pareamento utilizadas. A seguir, tal vetor será referido como vetor de comparação.

Oito classificadores foram utilizados para classificar cada comparação como Par ou Não-Par, com base nos respectivos vetores de comparação, completando assim a aplicação de oito métodos de pareamento. Conforme comentado, como o que diferencia os métodos de pareamento é o algoritmo de

classificação, doravante cada método de pareamento será referido pelo nome do algoritmo de classificação correspondente. Portanto, podemos dizer que os seguintes métodos de pareamento foram adotados: FS (Fellegi-Sunter), Kmeans (K-Means), Bclust (Bagged Clustering), Rpart (Recursive Partitioning Tree), Bagging (Bootstrap Aggregating), Bumping (Bootstrap Umbrella) de parâmetros do modelo), ADA (Adaptive Boosting) e SVM (Support Vector Machine).

Em todos os casos, foi adotada classificação binária, ou seja, cada comparação foi alocada em um de dois grupos: Par ou Não-Par. O uso de uma categoria para Possível-Par, originalmente introduzida por Fellegi e Sunter (1969) não foi considerado, visto que nem todos os algoritmos adotados fornecem resultados de boa qualidade para classificações em mais de duas classes. A adoção da classificação binária em todos os métodos objetivou melhorar as condições de comparabilidade dos métodos de pareamento.

Como no presente estudo o método FS está sendo comparado com métodos supervisionados, ou seja, métodos que utilizam alguma informação sobre a verdadeira classe de um subconjunto das comparações, a escolha desse limite também se beneficiou dessa informação, garantindo assim o melhor desempenho possível para Método FS. Em cada estado, foi escolhido o limite que fornece a melhor média harmônica entre precisão e sensibilidade, conhecida como Estatística-F. Em outras palavras, foi considerada a média harmônica entre a proporção de pares preditos corretamente em relação à quantidade de pares preditos e à proporção de pares preditos corretamente em relação ao total de pares existentes. A formalização dessas três medidas é mostrada adiante, na seção 4.3.

Para os algoritmos de classificação, cujos classificadores foram desenvolvidos a partir de amostras do tipo Bootstrap, o número de amostras utilizadas foi o padrão do pacote RecordLinkage. Portanto, nos algoritmos Bagging, Bumping e ADA, o número de amostras foi de 25, enquanto no método Bclust esse número foi 10. Isso significa que os três primeiros classificadores são o resultado da combinação de 25 classificadores simples, enquanto o último resulta da combinação de 10 classificadores simples. O número de iterações pode ser ajustado por conveniência.

Todos os métodos de pareamento geraram múltiplos pares para alguns estabelecimentos. Para reduzir o número de pares a um-para-um, um procedimento de redução foi aplicado. O procedimento consistiu em manter apenas o primeiro par formado por estabelecimento, respeitando a ordem das comparações no arquivo final, e descartar os demais pares envolvendo o mesmo estabelecimento. No método FS, as comparações são ordenadas pelo valor do

peso de comparação, do maior para o menor. Nesse caso, os pares excluídos são aqueles com o mesmo ou menor peso de comparação, consequentemente, com a mesma ou menor chance de ser um par verdadeiro. Nos outros métodos, a ordem das comparações depende da ordem original nos arquivos de entrada. Assim, se um arquivo de entrada for reordenado, o resultado da redução pode ser diferente, mesmo se o pareamento for executado usando o mesmo método. Para esses métodos de pareamento, o uso de um algoritmo mais sofisticado pode permitir ordenar de acordo com algum peso de comparação.

Os procedimentos adotados para preparar os dados tanto para o pareamento quanto para a redução foram realizados utilizando as funções do pacote dplyr versão 0.7.4 (Wickham et al., 2017), na linguagem de programação R. O pacote RecordLinkage (Borg & Sariyar, 2016) da linguagem de programação R foi utilizado para realizar as etapas de comparação e classificação. Para comparação e cálculo dos valores de similaridade correspondentes de variáveis de pareamento, foi usada a função compare.linkage com a função jarowinkler. A classificação foi realizada usando emClassify e classifyUnsup, funções para métodos não supervisionados e classifySup para métodos supervisionados. Os códigos em linguagem R estão disponíveis mediante solicitação encaminhada à primeira autora.

4.3. COMPARAÇÃO DE DESEMPENHO

O desempenho dos métodos de pareamento foi medido por meio de três estatísticas baseadas nas quantidades de verdadeiros e falsos positivos e verdadeiros e falsos negativos, a saber: precisão, sensibilidade e sua média harmônica, também conhecida como Estatística-F.

O desempenho dos métodos supervisionados, ou seja, Rpart, Bagging, Bumping, ADA e SVM, foi avaliado mediante aplicação da técnica de validação cruzada. O objetivo era determinar o desempenho do classificador em um conjunto de dados independente, reduzindo assim o viés do classificador devido ao superajuste (*overfitting*). Os arquivos de dados originais CEMPRE e SEFAZ, aqui denominados A e B, foram divididos aleatoriamente em 10 partes de tamanho aproximadamente igual, A_i e B_j, i, j = 1, 2,…, 10. A cada etapa, A_i e B_j foram excluídos dos arquivos A e B, restando as partes $A_{i'}$ e $B_{j'}$. Pares de comparação envolvendo registros de $A_{i'}$ e $B_{j'}$ foram então usados como dados de treinamento para desenvolver um classificador. Cada classificador foi usado para prever a classe das comparações realizadas com os estabelecimentos dos arquivos (partes) A_i e B_j. O procedimento tem a vantagem de garantir que todos os pares de comparação sejam utilizados, tanto como dados de treinamento

quanto de teste, uma vez e apenas uma vez. Os resultados das 100 classificações realizadas foram usados para calcular as medidas de desempenho supracitadas para os métodos comparados.

O CNPJ dos estabelecimentos agropecuários foi utilizado para determinar a verdadeira classe de cada comparação, ou seja, para construir um padrão ouro de classificação. Para cada comparação, o CNPJ registrado no CEMPRE foi comparado com o número registrado na SEFAZ. Foi assumido que os registros com exatamente o mesmo CNPJ pertencem ao mesmo estabelecimento, enquanto os demais a estabelecimentos distintos.

Foram computadas quatro grandezas: VP, FN, FP e VN, ou seja, o número de comparações envolvendo registros pertencentes ao mesmo estabelecimento que foram efetivamente classificados como Par, portanto verdadeiros positivos (VP); o número de comparações envolvendo registros pertencentes ao mesmo estabelecimento que foram classificados como Não-Par, portanto falsos negativos (FN); as envolvendo registros pertencentes a diferentes estabelecimentos que foram classificados como Par, os falsos positivos (FP); e o número de comparações pertencentes a diferentes estabelecimentos que foram classificados como Não-Par, portanto verdadeiros negativos (VN). Uma visão geral dos resultados para os oito métodos de pareamento pode ser obtida nas respectivas Matrizes de Confusão (Figuras 2a a 2c).

Figura 2a – Matrizes de confusão para o Estado do Maranhão

Método: FS

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	178 (VP)	208 (FN)
Não-Par	39 (FP)	614.575 (VN)

Método: Kmeans

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	256 (VP)	130 (FN)
Não-Par	208.646 (FP)	405.968 (VN)

Método: Bclust

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	65 (VP)	321 (FN)
Não-Par	71.337 (FP)	543.277 (VN)

Método: RPart

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	258 (VP)	128 (FN)
Não-Par	61 (FP)	614.553 (VN)

Método: Bagging

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	302 (VP)	84 (FN)
Não-Par	69 (FP)	614.545 (VN)

Método: Bumping

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	282 (VP)	104 (FN)
Não-Par	69 (FP)	614.545 (VN)

Método: ADA

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	294 (VP)	92 (FN)
Não-Par	42 (FP)	614.572 (VN)

Método: SVM

Classe Verdadeira \ Classe Predita	Par	Não-Par
Par	220 (VP)	166 (FN)
Não-Par	44 (FP)	614.570 (VN)

Figura 2b – Matrizes de confusão para o Estado da Paraíba

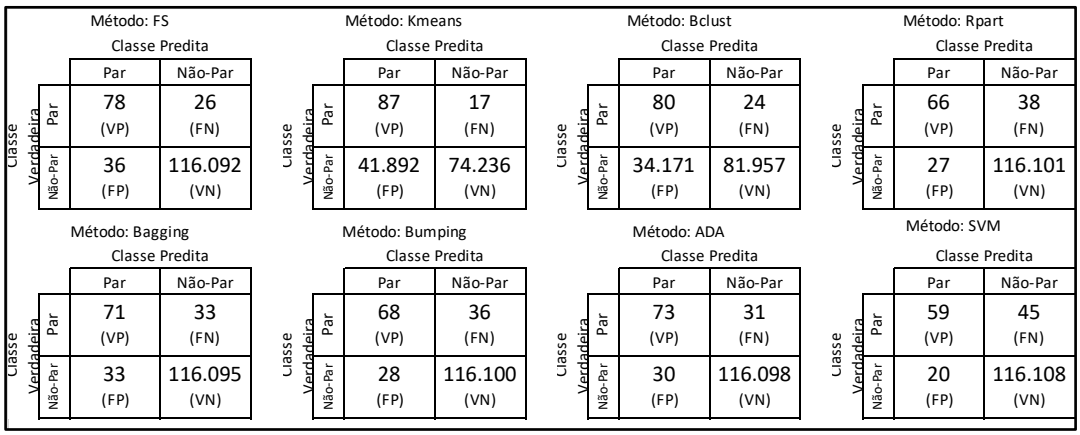
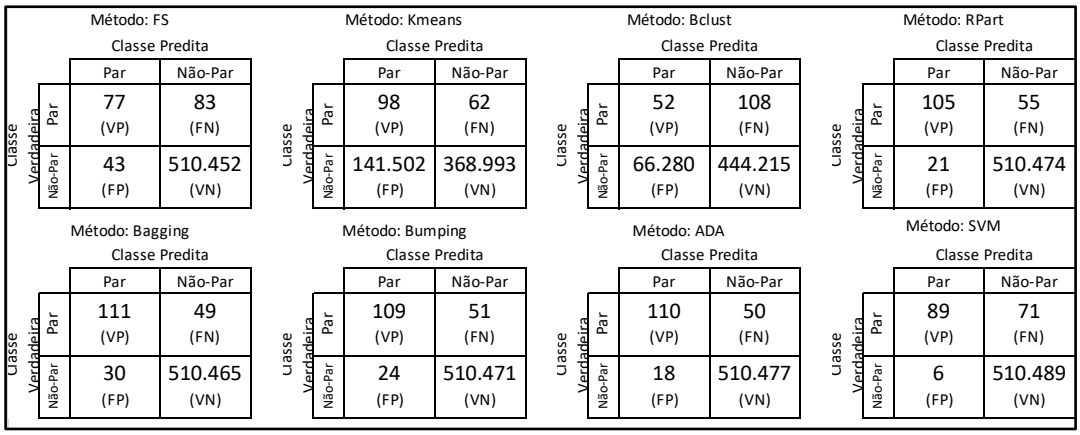


Figura 2c – Matrizes de confusão para o Estado de Santa Catarina



As quantidades acima fornecem uma visão geral da qualidade do pareamento, especialmente quando os resultados são consistentes em diferentes cenários, o que não é o caso no presente estudo. Assim, para uma visão mais precisa do desempenho dos métodos de pareamento, medidas resumo de cada método por estado foram utilizadas (Tabela 3). Tais medidas são **precisão**, $p=VP/(VP+FP)$ e **sensibilidade**, $s=VP/(VP+FN)$ e **Estatística-F**, sendo $F=2ps/(p+s)$. Como essas três medidas estão associadas a proporções, elas variam entre zero e um. O método é melhor avaliado quanto mais próximo do valor 1 as medidas associadas estiverem.

Tabela 3 - Medidas resumo do pareamento, por Estado, segundo o método

Método	Precisão			Sensibilidade			Estatística-F		
	MA	PB	SC	MA	PB	SC	MA	PB	SC
Não Supervisionado									
FS	0,82	0,68	0,64	0,46	0,75	0,48	0,59	0,72	0,55
Kmeans	0,00	0,00	0,00	0,66	0,84	0,61	0,00	0,00	0,00
Bclust	0,00	0,00	0,00	0,17	0,77	0,33	0,00	0,00	0,00
Supervisionado									
Rpart	0,81	0,71	0,83	0,67	0,63	0,66	0,73	0,67	0,73
Bagging	0,81	0,68	0,79	0,78	0,68	0,69	0,80	0,68	0,74
Bumping	0,80	0,71	0,82	0,73	0,65	0,68	0,77	0,68	0,74
ADA	0,88	0,71	0,86	0,76	0,70	0,69	0,81	0,71	0,76
SVM	0,83	0,75	0,94	0,57	0,57	0,56	0,68	0,64	0,70

Fonte: Proposta de Método de Pareamento para Integrar Dados sobre a Agropecuária. Tese de Silva, A. D. (2018).

(1) Todos os valores 0,00 correspondem a arredondamento de valor menor que 0,01.

No estado do Maranhão o método ADA é o único que se destaca pelo alto valor de sua precisão. Os outros métodos, exceto Kmeans e Bclust, apresentam precisão semelhante. Nos estados da Paraíba e Santa Catarina, o método SVM é o de maior precisão; além disso, os demais métodos supervisionados são mais precisos que os não supervisionados, embora a diferença apareça apenas na segunda casa decimal. Independentemente do estado, os métodos baseados no algoritmo de classificação K-Médias (Kmeans e Bclust) apresentam precisão mais baixa que os demais métodos.

Não é incomum haver relação inversa entre a precisão e a sensibilidade, ou seja, a maior precisão vir associada com menor sensibilidade e vice-versa. No limite, se todas as comparações forem classificadas como Par, o número de FN será zero e a sensibilidade será igual a 1. Porém, neste caso, o total de VP será dividido pelo total de comparações e a precisão ficará muito próxima de zero. Isto é bem ilustrado no presente estudo, visto que o método com maior precisão é também aquele com menor sensibilidade (SVM) e aqueles com menor precisão são os com maior sensibilidade (Kmeans e Bclust).

Os resultados empíricos não indicam um método com maior precisão e sensibilidade, concomitantemente. Assim sendo, usar uma medida sintética,

como a Estatística-F, para comparar a superioridade geral do método, torna-se ainda mais importante. Observe que a superioridade geral aqui se refere a uma medida de consenso, uma vez que a Estatística-F é a média harmônica das duas medidas apresentadas. Em casos específicos, pode ser preferível escolher maximizar a precisão ou a sensibilidade. Neste caso, pesos diferenciados podem ser associados com uma ou outra medida, quando do cálculo da média. Uma reflexão sobre o assunto é proposta por Hand e Christen (2018).

Nos estados do Maranhão e Santa Catarina, os métodos supervisionados (Rpart, Bagging, Bumping, ADA e SVM) resultaram em maior Estatística- F, com os maiores valores nominais associados com o método ADA. Na Paraíba, o método FS apresentou Estatística-F ligeiramente maior (0,01). Portanto, o método ADA demonstrou superioridade ou equivalência quando comparado aos outros sete métodos, na maioria das vezes competindo apenas com os demais métodos supervisionados. Em todos os estados, os métodos supervisionados competem entre si, apresentando pequena variação aparente no valor da respectiva Estatística-F.

Para entender o quanto a variação diferencia um método supervisionado do outro, foi realizado teste de Kruskal-Wallis de igualdade dos valores médios das Estatísticas-F baseados nas 100 classificações aplicando validação cruzada. Os testes revelaram que a diferença não é significativa, exceto para o método SVM quando utilizado com dados do estado do Maranhão. Os resultados gerais mostram superioridade dos métodos supervisionados, ou seja, Rpart, Bagging, Bumping, ADA e SVM. Na maioria dos casos esses métodos resultaram em maiores proporções de comparações corretamente classificadas, corroborando com a literatura especializada.

5. DISCUSSÃO E TRABALHOS FUTUROS

Os métodos de pareamento experimentados abrangem a aplicação de uma variedade de abordagens, tanto tradicionais - como as baseadas no modelo de decisão Fellegi-Sunter, quanto recentes - como as da área de aprendizado de máquina. Assim, são oferecidas alternativas para realizar pareamento em diferentes cenários metodológicos. No entanto, o presente trabalho não pretendeu ser exaustivo do conjunto de métodos de pareamento apresentados na literatura e já experimentados, ainda que utilizando apenas dados sintéticos. Métodos baseados em classificadores Bayesianos ou usando técnicas de Redes Neurais Artificiais, por exemplo, são populares na literatura e não foram utilizados no presente trabalho. Além disso, há uma série de versões dos

algoritmos usados como base para os métodos de pareamento apresentados que também podem ser experimentados. Este é o caso do Bumping Adaptativo, Random Forest, Bayesian SVM, entre outros.

Mesmo que se queira estender o presente estudo mantendo-se os mesmos métodos de pareamento utilizados, ainda existem inúmeras possibilidades de combinações a serem exploradas. Por exemplo, em todos os métodos de pareamento foi utilizada a mesma função de comparação para calcular valores de similaridade (Jaro-Winkler). Embora a função mais popular na literatura tenha sido usada, em casos específicos, outras funções podem funcionar melhor. Além disso, os algoritmos de classificação utilizados podem ter seus valores de parâmetros e funções modificados.

O uso de valores de parâmetro e funções padrão do pacote RecordLinkage não garante necessariamente o melhor desempenho possível de todos os métodos. Exemplos de ajustes possíveis incluem aumentar o número de amostras usadas para implementar os algoritmos Bclust, Bagging, Bumping e ADA; substituir a função linear usada na implementação do algoritmo SVM por uma função não linear; modificação dos valores de custo de alocação de comparação em cada classe. Da mesma forma, o limite que distingue Par de Não-Par no método baseado em modelo de decisão de Fellegi-Sunter também pode ser definido com base em uma série de amostras do tipo Bootstrap, usando ou não modelagem para encontrar o valor mais apropriado. Um bom começo pode ser a busca de valores para os parâmetros dos três métodos que apresentaram melhor desempenho nominal o maior número de vezes, portanto, Bagging, Bumping e ADA, o que pode permitir diferenciar significativamente seus desempenhos.

A extensão do presente trabalho a dados de pesquisas, censitárias ou amostrais, é certamente uma contribuição importante para a integração de dados sobre a agropecuária. O uso apenas de registros no presente trabalho deveu-se à limitação de tempo e recursos para construir um padrão ouro a partir de censos e pesquisas para permitir calcular medidas de desempenho dos métodos. Para que isso fosse possível, nas condições dadas, o trabalho restringiu-se a fontes com informação de um identificador único, neste caso o CNPJ. No entanto, o trabalho pode ser replicado, e um padrão ouro pode ser construído usando técnicas semiautomáticas ou assistidas, ao menos para uma subpopulação.

Por último, mas não menos importante, o desenvolvimento de ferramentas computacionais em linguagem R para realizar procedimentos de limpeza e padronização de dados para pareamento pode incentivar o uso de tais métodos para resolver ainda mais problemas, como imputação de dados, por

exemplo. Embora já exista um conjunto de comandos, especialmente nos pacotes stringr e plyr, um pacote para preparar os dados para RL ainda não foi encontrado no R.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Belin, T. R. (1993). Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment. Survey Methodology, 10(1), 13-29. Statistics Canada.

Borg, Andreas; Sariyar, Murat. (2016). Package 'RecordLinkage': Record Linkage in R. Disponível em: https://r-forge.r-project.org/projects/recordlinkage/. Acesso em: 15/01/2021.

Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classiers. EECS Department University of California e AT&T Bell Laboratories.

Breiman, L. (1998). Arcing Classifiers. The Annals of Statistics, 26(3), 801-849.

Breiman, L. Bagging Predictors. Technical Report No. 421. Department of Statistics University of California. 1994.

Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability series. 1st Edition. Wadswort.

Budzinsky, C. D. (1991). "Automated Spelling Correction," Statistics Canada.

Fellegi, I. P., and Sunter, A. B. (1969)., "A Theory for Record Linkage," Journal of the American Statistical Association, 64, 1183-1210.

Ceriani, L.; Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità by Corrado Gini. J Econ Inequal 10:421–443..

Christen, P. (2007). A Two-Step Classification Approach to Unsupervised Record Linkage. Department of Computer Science, The Australian National University.

Christen, P. (2012). *Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer.

Cortes, C.; Vapnik, V. (1995). *Support-Vector Networks in Machine Learning*, 20, 273-297. Kluwer Academic Publishers, Boston.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statist. Soc. B, 39, 1–38.

Dunn, H. L. (1946). Record Linkage. National Office of Vital Statistics, U. S. Public Health Service, Federal Security Agency. Washington, D. C.

Efron, Bradley. (1979). Bootstrap Methods: Another Look at The Jackknife. The Annals of Statistics, 7(1), 1-26.

Elfeky, M. G.; Verykios, V. S.; Elmagarmid, A. K. (2002). A Record Linkage Toolbox. Purdue University, Purdue e-Pubs, Indiana.

Fellegi, I. P. (1997). Record Linkage and Public Policy – A Dynamic Evolution. Record Linkage Techniques. Statistics. Canada.

Fellegi, I.; Sunter, A. (1969). A Theory for Record Linkage, Journal of the American Statistical Association, 64, 1183-1210.

Friedman, J.; Hastie, T.; Tibshirani, R. (1998). Additive Logistic Regression: a Statistical View of Boosting.

Gini, C. (1912). Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. C. Cuppini, Tipografia di Paolo Cuppin. Bologna.

Hand, David & Christen, Peter. (2018). A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing. 28. 10.1007/s11222-017-9746-6.

Hartigan, J. A; Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1. Blackwell Publishing for the Royal Statistical Society.

Haykin, Simon. (1999). Neural Networks: A Comprehensive Foundation. 2nd edition. Prentice Hall. Simon & Schuster. New Jersey.

IBGE. (2012). Sínteses históricas - Históricos dos censos - Panorama introdutório. Disponível em: https://memoria.ibge.gov.br/historia-do-ibge/historico-dos-censos/panorama-introdutorio.html. Acesso em: 12/05/2021.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2015). *An Introduction to Statistical Learning: with Applications in R*. 6th edition. Springer.

Jaro, Matthew A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84(406),.414-420.

Leish, F. Bagged Clustering. (1999). Working Paper No. 51 August. Working Paper Series of Adaptive Information Systems and Modelling in Economics and Management Science. Vienna University of Economics and Business Administration.

Newcombe, H. B.; Kennedy, J. M.; Axford S. J.; James A. P. (1959). Automatic linkage of vital records. Science, the American Association for the Advancement of Science, 130(3381), 954-959.

O'Hare, Kevin; Jurek-Loughrey, Anna; Campos, Cassio de. (2019). An unsupervised blocking technique for more efficient record linkage. Data & Knowledge Engineering. Volume 122, pp. 181-195. https://doi.org/10.1016/j.datak.2019.06.005.

Porter, E. H.; Winkler, W. E. (1997). Approximate String Comparison and its Effect on an Advanced Record Linkage System. Record Linkage Techniques. U.S. Bureau of the Census, Research Report.

Shannon, C. E. (1948). A Mathematical Theory of Communication. Reprinted with corrections from The Bell System Technical Journal, 27, pp. 379–423, 623–656.

Silva, Andrea Diniz da. (2018) Proposta de Método de Pareamento para Integração de Dados sobre Agropecuária, Tese de Doutoramento. Escola Nacional de Ciências Estatisticas. Ano de obtenção: 2018

Smith, M. E.; Newcombe, H. B. (1975). Methods of Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories. Methods of Information in Medicine, 14, 118-125.

Tibshirani, R.; Knight, K. (1999). Model Search by Bootstrap Bumping. Journal of Computational and Graphical Statistics, 8(4), 671-686.

Vapnik, V. N. (1998). Statistical Learning Theory. A Wiley-Interscience Publication. John Wiley & Sons, Inc.

Wickham, Hadley; FRANCOIS, Romain; HENRY, Lionel; MÜLLER, Kirill. (2017). Package 'dplyr': A Grammar of Data Manipulation. Disponível em:

https://github.com/tidyverse/dplyr. Acesso em: 15/01/2021.

Winkler, W. E. (1989). Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage. Survey Methodology, 15, 101-117.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. U.S. Bureau of the Census, Stat. Research Div.

Winkler, W. E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Survey Research Methods Section, American Statistical Association.

Winkler, W. E. (1995). Matching and Record Linkage. U.S. Bureau of the Census. Washington D. C.

Winkler, W. E. (2000a). Machine learning, information retrieval, and record linkage. American Statistical Association, Proceedings of the Section on Survey Research. Methods Section.

Winkler, W. E. (2000b). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Statistical Research Division. Methodology and Standards Directorate, U.S. Bureau of the Census. Washington D. C.

Winkler, W. E. (2014). Matching and Record Linkage. WIREs Comput Stat, 6, 313–325.

AGRADECIMENTOS e COLABORAÇÕES

A autora agradece o apoio do Instituto Brasileiro de Geografia e Estatística – IBGE e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pelo financiamento; e a todas e todos que contribuíram com dados e ideias para o desenvolvimento do projeto de tese que deu origem ao presente artigo. Em especial, agradece aos professores José André Brito, Djalma Pessoa e Ray Chambers, pela orientação repleta de generosidade, compromisso, solidariedade e bom humor; e aos professores Carlos Enrique Guanziroli, Gustavo da Silva Ferreira, Pedro Luis do Nascimento Silva e Raydonal Ospina pelo trabalho cuidadoso e dedicado de revisão da tese.

Revista Brasileira de Estatística, V.79, N. 246, p.144-162, janeiro-junho 2021

APLICAÇÕES DE ESTIMADORES LINEARES DE BAYES PARA PREVISÃO EM POPULAÇÃO FINITA

Kelly C. M. Gonçalves

kelly@dme.ufrj.br

Fernando A. S. Moura

fmoura@im.ufrj.br

Helio S. Migon

migon@im.ufrj.br

Departamento de Métodos Estatísticos/ Instituto de Matemática/ Universidade Federal do Rio de Janeiro

Resumo: Baseado em técnicas de mínimos quadrados e especificação apenas do primeiro e segundo momentos *a priori*, o método linear de Bayes permite conciliar no contexto de amostragem de população finita as abordagens baseada em desenhos e em modelos. O método é aplicado a um modelo de regressão geral e, hipóteses de permutabilidade de segunda ordem geram resultados análogos a desenhos amostrais comuns. Por meio de utilização do pacote *BayesSampling* do software R, o método é aplicado a um exemplo preliminar, com o objetivo de ilustrar a especificação *a priori*. Uma extensão do método para estimação de eficácia de vacina em ensaios clínicos é apresentada. Além disso, o método é introduzido no contexto de estimação em pequenos domínios via modelos hierárquicos.
Palavras-chave: permutabilidade, momentos, estimação em pequenas áreas

Abstract: Based on least squares techniques and specification of the first and second prior moments only, Bayes' linear method enables conciliating the design-based and model-based approaches in the context of finite population sampling. The method is applied to a general regression model and hypotheses about second order exchangeability generate results analogous to those obtained with common sample designs, under the design-based approach. Using the BayesSampling package of the R software, the method is applied to a preliminary example, with the aim of illustrating the prior specification. An extension of the method for estimating vaccine efficacy in clinical trials is presented. In addition, the method is introduced to the context of small domains estimation via hierarchical models.
Keywords: exchangebility; moments; small area estimation

1.INTRODUÇÃO

Pesquisas amostrais constituem um método importante de obter informações precisas de uma população finita. A teoria mais tradicional empregada no tratamento dessas amostras, com a finalidade de inferir quantidades não observadas na população, é baseada no uso da distribuição aleatória induzida pelo desenho amostral (Neyman, 1934). Horvitz e Thompson (1952) propuseram uma teoria geral de amostragem sob probabilidade desigual e o método de estimação ponderada, conhecido como "estimador de Horvitz e Thompson".

Por ser livre da hipótese de distribuição para os dados, a teoria de amostragem baseada no desenho amostral se mostra atraente para as agências oficiais de estatística em todo o mundo. No entanto, o uso desta abordagem, muito mais descritiva em pesquisas do que para fins de previsão, induziu a uma lacuna no desenvolvimento teórico associado ao uso analítico desta abordagem. Dessa forma, em algumas situações específicas, como na estimação em pequenos domínios e sob a presença da não resposta, esta teoria se mostrou ineficiente, fornecendo preditores imprecisos. Em ambos os casos é fundamental a suposição de um modelo probabilístico que relacione as unidades observadas com as não amostradas. Desta forma, de um lado podemos ter argumentos de que a inferência baseada em modelo depende muito das suposições do modelo, que podem ser irrealistas e, de outro, a estimativa intervalar para os parâmetros da população se baseia no Teorema Central do Limite, que não pode ser aplicado em muitas situações práticas, onde o tamanho da amostra não é grande o suficiente e / ou pressupostos de independência não são razoáveis. Uma revisão bem completa destas metodologias com o auxílio de pacotes computacionais é apresentada em Pessoa & Silva (2018).

No contexto de modelos de superpopulação, Zacks (2002) mostrou que alguns estimadores baseados no desenho amostral podem ainda ser recuperados usando uma abordagem de modelo de regressão geral. Segundo Little (2003), “a especificação cuidadosa do modelo sensível ao desenho da pesquisa pode ser uma ferramenta adequada para se obter um modelo melhor especificado, e as estatísticas Bayesianas fornecem um tratamento coerente e unificado da inferência descritiva e analítica da pesquisa”.

Sob o paradigma bayesiano, Smouse (1984) propôs uma forma de conciliar as abordagens baseadas no desenho amostral e em modelos. O método incorpora informações *a priori* em modelos para inferência em populações finitas, usando técnicas de mínimos quadrados e especificação

apenas do primeiro e segundo momentos das distribuições envolvidas para descreverem o conhecimento prévio sobre as estruturas presentes na população. O'Hagan (1985) apresentou os estimadores lineares de Bayes em alguns contextos específicos, lidando com várias estruturas populacionais, como estratificação e agrupamento. Além disso, O'Hagan (1985) mostrou como alguns estimadores baseados no desenho amostral podem ainda ser obtidos como um caso particular da abordagem apresentada. Silva (1992), por sua vez, empregou a abordagem de estimação linear de Bayes de forma inovadora no Brasil, para pesquisas repetidas no tempo.

Esta é sem dúvida uma área de pesquisa de grande interesse ao longo de décadas. Algumas referências interessantes incluem, por exemplo, Karunami & Zhang (2003) e Al-khasawneh (2019), que discutiram a estimação de média em amostras repetidas no tempo e desenvolveram estimadores ótimos bayesianos lineares e bayesiano empírico para o t-ésimo tempo baseados nas amostras repetidas obtidas anteriormente.

Cocchi & Mouchart (1990) exploraram aplicações de estimadores lineares bayesianos em populações finitas na presença de variáveis auxiliares categóricas. Mostraram ainda que a abordagem é de fácil implementação, resolvendo dificuldades referentes à marginalização de parâmetros em modelos de superpopulação. Em outro artigo, Cocchi & Mouchart (1996) usaram quasi-estimação linear de Bayes em amostragem estratificada de populações finitas. Um modelo hierárquico foi proposto produzindo soluções que se aproximaram daquelas obtidas com suposição de normalidade.

Gonçalves et al. (2014) apresentaram de forma alternativa o método linear de Bayes aplicado a um modelo de regressão linear geral para previsão em população finita e mostraram como obter alguns estimadores baseados nos respectivos desenhos amostrais como casos particulares. Gonçalves et al (2014) estenderam ainda a metodologia para o contexto de estimador de razão e dados categóricos.

Por outro lado, o desenvolvimento de pacotes computacionais pode facilitar a disseminação e uso da metodologia. No contexto de estimação em população finita, Djalma Pessoa e co-autores se tornaram uma referência no desenvolvimento de materiais didáticos e pacotes. Como por exemplo, o pacote *convey* (Pessoa et al., 2020), que estima medidas de pobreza, desigualdade e bem-estar com base em microdados coletados por uma pesquisa amostral complexa.

Com essa motivação, mais recentemente, o pacote *BayesSampling* (Figueiredo & Gonçalves, 2021) do CRAN do R (https://cran.r-project.org/) foi

desenvolvido com o objetivo de facilitar a implementação do método linear de Bayes apresentado em Gonçalves et al (2014). Existem outros pacotes em R na área de amostragem, tais como: *TeachingSampling* (Rojas, 2020), *survey* (Lumley, 2020) e *sampling* (Tillé & Matei, 2021). No entanto, poucos pacotes que usam a abordagem bayesiana estão disponíveis, sendo *BayesSampling* o primeiro a lidar com estimadores lineares de Bayes no contexto de população finita.

O objetivo deste trabalho é, portanto, fazer uma revisão de métodos de amostragem de população finita baseada em modelos sob o paradigma bayesiano. Mais especificamente, a metodologia é apresentada numa estrutura geral de modelos de regressão com hipóteses acerca apenas de primeiro e segundo momentos e, em seguida, a estimação dos parâmetros é obtida de forma aproximada com base no método linear de Bayes. A predição para as unidades não observadas, maior interesse neste contexto, é feita aplicando apenas propriedades de esperança e variância condicionais.

Na Seção 2 introduzimos o conceito de modelo de superpopulação geral, e discutimos suposições de permutabilidade na estrutura do modelo que produzem estimativas pontuais iguais às estimativas obtidas supondo a abordagem baseada no desenho amostral para alguns casos particulares. Na Seção 3 apresentamos os estimadores lineares bayesianos para um modelo geral, ilustrando com o caso particular de permutabilidade entre todas as unidades. O pacote *BayesSampling* também é apresentado de uma forma geral nesta seção. Na Seção 4 fazemos algumas aplicações da metodologia a dados reais. A Seção 5 apresenta uma extensão do método linear de Bayes para o contexto de pequenas áreas. Finalmente, nossas principais conclusões estão na Seção 6 do artigo.

É com grande prazer que contribuímos para esta homenagem ao Prof. Djalma Pessoa. Em sua trajetória acadêmica, ao retornar de Berkeley, Djalma ministrou na USP um excelente curso de Teoria da Decisão usando o livro do Ferguson. Djalma muito se empenhou pela criação da ABE e, mais tarde, na direção da ENCE, implementou diversas modificações estruturais significativas. O Prof. Djalma foi, sem dúvida, um "descobridor" de talentos e muito contribuiu para o desenvolvimento da Estatística Brasileira.

2.AMOSTRAGEM DE POPULAÇÃO FINITA BASEADA EM MODELOS

Considere $\boldsymbol{U}$={u_1,…,u_N} uma população finita com N unidades, sendo N conhecido, e seja $\boldsymbol{y} = (y_1, \dots, y_N)'$ o vetor de quantidades desconhecidas da variável de interesse, associado a $\boldsymbol{U}$. O vetor resposta $\boldsymbol{y}$ pode ser particionado em uma parte observada $\boldsymbol{y}_s$, baseada numa amostra de tamanho n, e numa parte não observada $\boldsymbol{y}_{\bar{s}}$ de tamanho $N-n$. O interesse principal é prever uma função de $\boldsymbol{y}$, como por exemplo, o total populacional $T = \sum_{i=1}^{N} y_i = \mathbf{1}_s'\boldsymbol{y}_s + \mathbf{1}_{\bar{s}}'\boldsymbol{y}_{\bar{s}}$, onde $\mathbf{1}_s'$ e $\mathbf{1}_{\bar{s}}'$ são vetores unitários de dimensões n e $N-n$, respectivamente.

Enquanto na teoria convencional de amostragem as unidades da população são tratadas como constantes (parâmetros), não expressando nenhuma relação entre as unidades da amostra e as unidades não amostradas, sob o enfoque de modelos de superpopulação, os valores das características de interesse são considerados realizações de variáveis aleatórias. Nesta abordagem, um modelo paramétrico é assumido para os valores na população $y_i's$, e então o "Melhor Preditor Linear Não Viciado" (do inglês "EBLUP") é obtido para a parte desconhecida $\boldsymbol{y}_{\bar{s}}$.

Um modelo de superpopulação é construído assumindo que o valor da variável de interesse associado à i-ésima unidade da população, y_i, $i = 1, \dots, N$, é constituído pela soma de um elemento determinístico m_i e um elemento aleatório ε_i, isto é,

$$y_i = m_i + \varepsilon_i. \qquad (1)$$

O vetor aleatório $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)$ tem distribuição parcialmente especificada com vetor de médias nulo e matriz de covariância $\boldsymbol{V}$ positiva definida com elementos da diagonal v_i e fora da diagonal c_{ij} , para $i, j = 1, \dots, N$, tal que $i \neq j$. Dessa forma, $E(y_i) = m_i, Var(y_i) = v_i$ e $Cov(y_i, y_j) = c_{ij}$. A modelagem explícita de estruturas na população por meio de modelos de superpopulação resulta, para alguns casos especiais, nas mesmas inferências pontuais para parâmetros populacionais, realizadas sob a amostragem baseada no desenho. A seguir descrevemos alguns exemplos.

M1) Se $m_i = m, v_i = v$ e $c_{ij} = c$ para todo $i, j = 1, \dots, N, i \neq j$, estamos assumindo implicitamente que a população não tem estruturas relevantes e unidades populacionais são permutáveis com relação à variável de interesse. Os preditores obtidos neste caso são equivalentes aos estimadores obtidos via amostragem aleatória simples na abordagem baseada no desenho. A

correlação introduzida no modelo pode ser justificada para simular uma amostragem aleatória simples sem reposição. No caso com reposição bastaria assumir que $c_{ij} = 0$ para todo $i, j = 1, \dots, N$.

M2) Se $m_i = mx_i, v_i = vx_i^2$ e $c_{ij} = cx_ix_j$ para todo $i, j = 1, \dots, N, i \neq j$, onde x_i é variável auxiliar associada à unidade populacional i, isto é equivalente a assumir que $E(y_i/x_i) = m, Var(y_i/x_i) = v$ e $Cov(y_i/x_i, y_j/x_j) = c$. A hipótese de permutabilidade neste caso é adotada para a razão $y_i/\, x_i$ e os preditores obtidos neste caso são equivalentes ao estimador razão obtido via amostragem aleatória simples na abordagem baseada no desenho.

M3) Para o caso que $\mathrm{y_i}$ for uma variável binária, ou seja, $\mathrm{y_{ik}} = 1$ se a unidade i pertence à categoria k e $\mathrm{y_{ik}} = 0$, caso contrário, temos um caso particular de M1). Supondo permutabilidade entre unidades de uma mesma categoria, também obtemos preditores equivalentes aos estimadores das proporções sob a abordagem baseada no desenho. Nesse caso, teríamos que $\mathrm{m_i} = \mathrm{m_k}$ e $\mathrm{v_i} = \mathrm{v_k} = \mathrm{m_k}(1 - \mathrm{m_k})$, se a unidade i pertence à categoria k. Além disso, $\mathrm{c_{ij}} = \mathrm{c_k} = \mathrm{m_k}(\mathrm{m_{kk}} - \mathrm{m_k})$ se i e j são unidades pertencentes à mesma categoria k, para $\mathrm{m_{kk}} = \mathrm{P}(\mathrm{y_{jk}} = 1 | \mathrm{y_{ik}} = 1)$ e, $\mathrm{c_{ij}} = \mathrm{c_{kk'}} = \begin{cases} \mathrm{m_k}(\mathrm{m_{k'k}} - \mathrm{m_{k'}}), & \text{se } \mathrm{i} \neq \mathrm{j} \\ -\mathrm{m_k m_{k'}}, & \text{se } \mathrm{i} = \mathrm{j} \end{cases}$, para o caso em que a unidade i pertence à categoria k e j pertence à categoria $\mathrm{k'}$.

M4) Se a população é dividida em H estratos, cada um de tamanho $N_h, h = 1, \dots, H$, tal que $N = \sum_{h=1}^{H} N_h$ e denotando por y_{hi} a variável de interesse associada à i-ésima unidade no h-ésimo estrato, para $i = 1, \dots, N_h, h = 1, \dots, H$, podemos assumir que $m_i = m_h\,, v_i = v_h, c_{ij} = c_h$, se i e j pertencem ao estrato h e $c_{ij} = c_{hh'}$, se i pertence a um estrato h e j pertence a um estrato h', tal que $h \neq h'$. A hipótese de permutabilidade neste caso é adotada para unidades pertencentes a um mesmo estrato. A correlação $c_{hh'}$ entre diferentes estratos permite que informações obtidas sobre um estrato possam mudar as crenças sobre outros estratos em algumas aplicações especiais. No entanto, os preditores obtidos neste caso só são equivalentes aos estimadores obtidos via amostragem aleatória simples estratificada se assumimos que observações em diferentes estratos não estão correlacionadas, ou seja, $c_{hh'} = 0$.

M5) Considere uma população dividida em H distintas subpopulações, os conglomerados, cada um de tamanho $N_h, h = 1, \dots, H$, tal que $N = \sum_{h=1}^{H} N_h$. Denote a variável de interesse associada à i-ésima unidade no h-ésimo conglomerado por y_{hi}, para $i = 1, \dots, N_h, h = 1, \dots, H$. Cada conglomerado é supostamente uma micro representação do universo, sendo esperado portanto heterogeneidade dentro dos conglomerados, no entanto homogeneidade entre eles. Tais estruturas refletem num modelo de superpopulação com hipóteses de permutabilidade de primeira e de segunda ordem da forma: $m_i = m$, $v_i = v_h + c + \gamma$, se i pertence ao conglomerado $h, c_{ij} = c$, se i e j pertencem ao mesmo conglomerado h e $c_{ij} = \gamma$, caso contrário. A variância dentro do conglomerado h é explicada por uma variabilidade particular de cada subgrupo, uma variabilidade constante entre unidades distintas no mesmo subgrupo e outra entre subgrupos distintos. Além disso, unidades distintas num mesmo conglomerado ou em conglomerados distintos são supostamente correlacionadas. Contudo restringir $\gamma = 0$ é razoável em diversos casos e importante na obtenção de preditores numericamente iguais aos estimadores obtidos via desenho amostral.

3.ESTIMADORES LINEARES BAYESIANOS

De uma forma mais geral, o modelo de superpopulação da equação (1) pode ser reescrito para o vetor completo $\boldsymbol{y}$ na forma de um modelo de regressão com hipóteses também apenas acerca do primeiro e segundo momentos da distribuição do erro aleatório. Sob o enfoque bayesiano temos o modelo com distribuição *a priori* para o vetor paramétrico descrito da seguinte forma:

$$\mathbf{y}|\boldsymbol{\beta} \sim [\boldsymbol{X\beta}, \boldsymbol{V}] \qquad (2)$$
$$\boldsymbol{\beta} \sim [\boldsymbol{a}, \boldsymbol{R}],$$

onde: $\boldsymbol{X}$ é uma matriz de covariáveis com dimensão $N \times p$, supostamente conhecida para todas as unidades de U; $\boldsymbol{\beta}$ é um vetor $p \times 1$ de coeficientes e $\boldsymbol{V}$ é a matriz de covariância $N \times N$. Particionando o vetor $\boldsymbol{y}$ em uma parte que depende da amostra e outra não observada, a primeira equação do modelo em (2) pode ser reescrito da forma:

$$\begin{pmatrix} \boldsymbol{y}_{\bar{s}} \\ \boldsymbol{y}_{s} \end{pmatrix} |\boldsymbol{\beta} \left[\begin{pmatrix} \boldsymbol{X}_{\bar{s}}\boldsymbol{\beta} \\ \boldsymbol{X}_{\boldsymbol{s}}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V}_{\bar{s}} & \boldsymbol{V}_{\bar{s}s} \\ \boldsymbol{V}_{s\bar{s}} & \boldsymbol{V}_{s} \end{pmatrix} \right]. \qquad (3)$$

Considerando a abordagem bayesiana neste caso, estamos num cenário de distribuição parcialmente especificada, facilitando assim o processo de definição da distribuição *a priori*, questionado em algumas situações pelo caráter subjetivo. A inferência neste caso baseia-se fundamentalmente no método linear de Bayes. Os estimadores têm a propriedade de minimizar a perda quadrática *a posteriori* dentre todos os estimadores lineares e, podem ser vistos como aproximações das médias *a posteriori*. Além disso, se a normalidade é assumida, o estimador linear de Bayes e sua variância coincidem com a média e variância *a posteriori,* respectivamente. O estimador linear de Bayes para $\boldsymbol{y}_{\bar{s}}$, condicional a $\boldsymbol{\beta}$, e sua variância associada (Gonçalves et al., 2014) são dados por:

$E(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s,\boldsymbol{\beta}) \approx \boldsymbol{X_s\beta} + \boldsymbol{V}_{\bar{s}s}\boldsymbol{V_s}^{-1}(\boldsymbol{y}_s - \boldsymbol{X_s\beta})$ e $V(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s,\boldsymbol{\beta}) \approx \boldsymbol{V}_{\bar{s}} - \boldsymbol{V}_{\bar{s}s}\boldsymbol{V_s}^{-1}\boldsymbol{V}_{s\bar{s}}$.

Além disso, como

$$\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{y}_s\end{pmatrix}\left[\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{X_s\beta}\end{pmatrix},\begin{pmatrix}\boldsymbol{R} & \boldsymbol{X_s'\beta X_s}\\ \boldsymbol{X_s\beta X_s'} & \boldsymbol{V_s}\end{pmatrix}\right],$$

podemos obter de forma análoga o estimador linear de Bayes para $\boldsymbol{\beta}$ da forma:

$E(\boldsymbol{\beta}|\boldsymbol{y}_s) = \widehat{\boldsymbol{\beta}} \approx \boldsymbol{C}(\boldsymbol{X_s'V_s^{-1}y_s} + \boldsymbol{R^{-1}a})^{-1}$ e $V(\boldsymbol{\beta}|\boldsymbol{y}_s) = \boldsymbol{C} \approx (\boldsymbol{R^{-1}} + \boldsymbol{X_s'V_s^{-1}X_s})^{-1}$.

Note que, sob distribuição *a priori* não informativa (fazendo $\boldsymbol{R^{-1}} \to \boldsymbol{0}$), o estimador linear de Bayes para $\boldsymbol{\beta}$ coincide com o estimador de mínimos quadrados para $\boldsymbol{\beta}$.

Finalmente, usando propriedades de esperança e variância condicionais, obtemos o estimador linear bayesiano para $\boldsymbol{y}_{\bar{s}}$, da forma:

$E(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s) \approx \boldsymbol{X_s}\widehat{\boldsymbol{\beta}} + \boldsymbol{V}_{\bar{s}s}\boldsymbol{V_s}^{-1}(\boldsymbol{y}_s - \boldsymbol{X_s\beta})$ e $V(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s) \approx \boldsymbol{X_sCX_s'} + \boldsymbol{V}_{\bar{s}}$. (4)

Com isso, o estimador linear bayesiano para o total populacional e sua variância são dados por:

$$\hat{T} = \boldsymbol{1_s'y_s} + \boldsymbol{1_{\bar{s}}'}E(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s) \text{ e } V\left(\hat{T}\middle|\boldsymbol{y}_s\right) = \boldsymbol{1_{\bar{s}}'}V(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s)\boldsymbol{1}_{\bar{s}},$$

para $E(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s)$ e $V(\boldsymbol{y}_{\bar{s}}|\boldsymbol{y}_s)$ dados pela Equação (4).

Gonçalves et al (2014) revisitam alguns dos desenhos mais comuns, como casos particulares do modelo de regressão em (2). Todos partem de premissas de permutabilidade que refletem estruturas da população, como descrito na Seção 2. Por exemplo, no caso do modelo M1), bastaria assumir no modelo de regressão em (2) que $\boldsymbol{\beta}$ é um parâmetro escalar, $\boldsymbol{X} = \boldsymbol{1}_N$, $\boldsymbol{a} = m$, $\boldsymbol{R} = c$ e $\boldsymbol{V} = \sigma^2\boldsymbol{I}_N$, para $\sigma^2 = v - c$ e $\boldsymbol{I}_N$ uma matriz identidade $N \times N$. Com isso, o estimador linear de Bayes para este caso é dado por:

$$\hat{T} = n\bar{y}_s + (N-n)\hat{\mu}, \quad (5)$$

onde $\bar{y}_s$ é a média amostral e $\hat{\mu} = w\bar{y}_s + (1-w)m$, para $w = \frac{n\sigma^{-2}}{c^{-1}+n\sigma^{-2}}$.

Note que $\hat{\mu}$ é uma média ponderada entre a média *a priori* m e a média amostral, com peso w. Além disso, sob distribuição *a priori* vaga, ou seja, se $v \to \infty$ e $c \to \infty$, para σ^2 fixo, o estimador da equação (5) coincide numericamente com o estimador baseado no desenho amostral. O mesmo ocorre com a variância do estimador. Gonçalves et al (2014) descrevem os estimadores lineares bayesianos para o caso do estimador razão e de permutabilidade em estratos, expandindo a metodologia para dados categóricos.

3.1 PACOTE *BayesSampling* DO R

O pacote *BayesSampling* do R (Figueiredo & Gonçalves, 2021) apresenta funções para os estimadores descritos em Gonçalves et al (2014). O objetivo principal do pacote é viabilizar ao usuário a aplicação da metodologia em seus problemas práticos.

A função "BLE_Reg" do pacote apresenta o preditor linear bayesiano para o total e média populacionais, com base no modelo de regressão geral em (3), o que permite ao usuário estender o método para outros cenários, necessitando apenas que o mesmo indique que hipóteses de permutabilidade devem ser consideradas na forma de $\boldsymbol{X}$, $\boldsymbol{a}$, $\boldsymbol{R}$ e $\boldsymbol{V}$. Outras funções do pacote calculam os preditores bayesianos para total e média sob casos particulares de permutabilidade e dependem da função "BLE_Reg", sendo esta portanto a principal função do pacote. O pacote apresenta os preditores, simples e razão, para o contexto de permutabilidade entre todas as unidades e permutabilidade dentro de estratos. Uma versão atualizada do pacote incluindo o caso de dados categóricos foi recentemente publicada no CRAN do R.

Embora nos argumentos das funções se espere que o usuário defina as quantidades *a priori* que deseja, no caso do usuário não informá-las um aviso é demonstrado na tela e o caso de distribuição *a priori* vaga é assumido. Portanto os resultados esperados neste caso são numericamente iguais aos obtidos sob a abordagem clássica, baseada no desenho amostral.

4.APLICAÇÕES

As aplicações a seguir foram feitas com o pacote *BayesSampling* do R (Figueiredo & Gonçalves, 2021) e têm como objetivo ilustrar os estimadores lineares de Bayes em problemas práticos de predição em população finita.

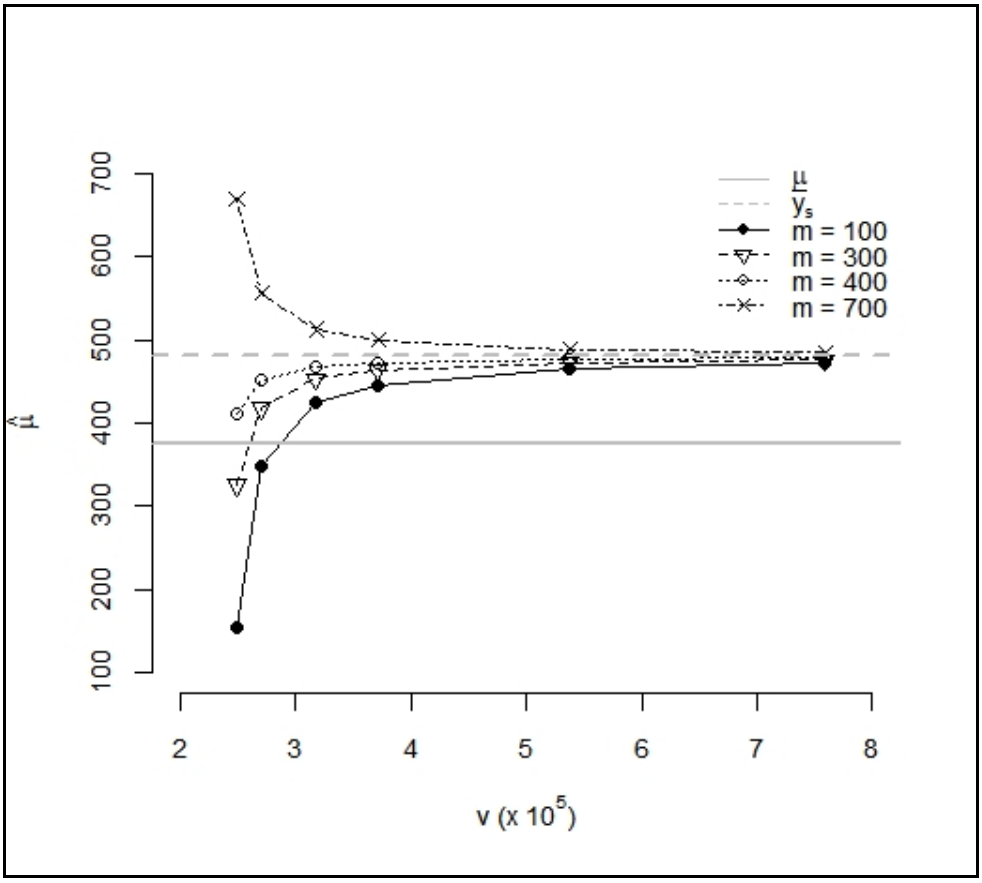
4.1 EXEMPLO INTRODUTÓRIO

O conjunto de dados *BigCity* está disponível no pacote *BayesSampling* (Figueiredo & Gonçalves, 2021) e é usado também em outros pacotes de amostragem do R, como o *TeachingSampling* (Rojas, 2020). A base de dados corresponde a variáveis socio-eoconômicas para uma população de $N = 150.266$ habitantes de uma cidade particular para um ano específico.

Tomamos uma amostra aleatória simples de 10% dessa população para diminuição do custo computacional e consideramos essa como a população-alvo. Com o objetivo de estimar a renda média (μ) desta população, selecionamos uma amostra aleatória simples de $n = 20$ habitantes. O principal interesse neste estudo é avaliar o impacto de especificação dos momentos *a priori* nas estimativas da renda média populacional.

A Figura 1 apresenta as estimativas pontuais para diferentes valores de m e v, além da média amostral (linha tracejada cinza) e a verdadeira média populacional (linha cheia cinza). Podemos ver que à medida que a variância *a priori* aumenta as estimativas com base no método linear de Bayes tendem à média amostral e a velocidade com que isso ocorre depende do valor escolhido para a média *a priori* m. Por outro lado, para que o valor verdadeiro da renda média populacional seja melhor estimado faz-se necessário definir um valor menor para m atrelado a uma variância *a priori* menor. A estimativa clássica por si só traria uma superestimação da verdadeira renda média sem possibilidades de introdução de um conhecimento *a priori* que permitisse uma redução desta estimativa.

Figura 1 – Estimação da renda média de uma população (conjunto de dados *BigCity)* via método linear de Bayes variando a média e a variância *a priori.* A média amostral é representada pela linha tracejada em cinza e a verdadeira média populacional pela linha cheia em cinza.



4.2 ENSAIOS CLÍNICOS

O objetivo de um ensaio clinico de uma vacina é avaliar sua eficácia comparativamente com um placebo ou mesmo com alguma outra vacina já existente. As agências reguladoras (Anvisa, FDA etc.) têm por missão certificar a segurança e eficácia do produto antes de liberá-lo para uso na população. Esta é uma atividade de alta relevância e de fundamentação cientifica sólida, exigindo tratamento estatístico esmerado para validar as hipóteses científicas estabelecidas. Um ensaio clínico se desenvolve em diversas etapas. Após os estudos pré-clínicos passam-se as chamadas fases I,II e III. As duas primeiras fases, em geral laboratoriais, visam avaliar a segurança das vacinas, verificando a produção de anticorpos e a imunidade celular. A fase III dedica-se a determinar a eficácia da vacina, contrastando os resultados dos grupos: placebo e vacinados.

O planejamento desses ensaios clínicos deve ser cuidadosamente elaborado. Inicia-se por estabelecer, com clareza, quais são as questões ou

hipóteses cientificas que se deseja responder. Por exemplo, qual a proteção de múltiplas doses para evitar a doença, qual o intervalo entre doses, etc. A seguir define-se, cuidadosamente a população alvo, como: idosos, indivíduos com comorbidades, etnias, grupos de maior exposição à doença ou de incidência da doença, etc. Pode-se considerar dois braços, por exemplo vacina e placebo, ou múltiplos braços, por exemplo placebo e doses múltiplas, a intervalos fixos de tempo. Outro ponto relevante diz respeito aos desfechos esperados. Estes implicarão, eventualmente, em amostras maiores posto que os eventos podem se tornar mais raros. Por exemplo, estudar a duração da imunidade exigirá estudos mais longos do que simplesmente avaliar a proteção geral ou a transmissão da doença. Os ensaios clínicos de vacinas são ensaios aleatorizados e duplo cego. No caso particular da vacina do Butantan para a Covid-19 define-se como caso os indivíduos que apresentem uma lista de sintomas clínicos, por exemplo tosse, febre, etc, seguindo os critérios estabelecidos pela FDA.

Nos estudos de eficácia da Coronavac, vacina produzida pela Sinovac e pelo Instituto Butantan, divulgado em 12 de janeiro de 2021 pelo Governo de São Paulo (fonte: https://www.cnnbrasil.com.br/saude/2021/01/12/eficacia-de-78-da-coronavac-era-uma-meia-verdade-diz-medico-sanitarista), a população foi dividida em $H = 2$ estratos, sendo o grupo 1 composto pelos indivíduos que foram vacinados e o grupo 2 pelos que receberam placebo. Dentro destes estratos são analisados em que categoria dentre $J = 4$ possíveis cada indivíduo se encaixa. São elas: casos graves e moderados ($j = 1$), casos leves ($j = 2$), casos muito leves ($j = 3$) e assintomáticos ($j = 4$).

Defina a variável de interesse como sendo

$$y_{ijh} = \begin{cases} 1, se\ i-ésimo\ indivíduo\ do\ estrato\ h\ está\ na\ categoria\ j \\ 0, caso\ contrário, \end{cases}$$

para $j = 1, \dots, 4$, $i = 1, \dots, n_h$, $h = 1,2$. Nesse caso, é razoável supor permutabilidade para indivíduos dentro de uma mesma categoria e mesmo estrato. Sendo assim teríamos as seguintes hipóteses de permutabilidade no modelo (1):

$m_{jh} = E(y_{ijh}) = P(y_{ijh} = 1)$, $v_{jh} = V(y_{ijh}) = m_{jh}(1 - m_{jh})$, $c_{jh} = Cov(y_{ijh}, y_{i'jh}) = m_{jh}(m_{jjh} - m_{jh})$, para $i \neq i'$ e

$\sigma^2_{jh} = v_{jh} - c_{jh} = m_{jh}(1 - m_{jjh})$, onde $m_{jjh} = P(y_{i'jh} = 1 | y_{ijh} = 1)$.

Para $j \neq j'$, obtemos ainda que $Cov(y_{ijh}, y_{i'jh}) = \begin{cases} m_{jh}(m_{j'jh} - m_{j'h}), se\ i \neq i' \\ -m_{jh}m_{j'h}, se\ i = i' \end{cases}$,

onde $m_{j'jh} = P(y_{i'j'h} = 1 | y_{ijh} = 1)$.

Além disso, para todo $h \neq h'$, podemos assumir covariância igual a zero. O modelo neste caso é uma extensão do modelo apresentado em Gonçalves et al. (2014) para dados categóricos, com o objetivo principal sendo estimar a eficácia global da vacina Coronavac, a qual é dada por:

$$Ef = \frac{\sum_{j=1}^{3} p_{j1}}{\sum_{j=1}^{3} p_{j2}},$$

onde $p_{jh} = \frac{1}{N}\sum_{i=1}^{N_h} y_{ijh}$ é a proporção de indivíduos na categoria j no estrato h.

Portanto, para estimar a eficácia global devemos usar o estimador linear de Bayes de cada componente p_{jh} do numerador e denominador. O estimador linear de Bayes para este caso é uma extensão do obtido para dados categóricos em Gonçalves et al. (2014). Neste caso usaríamos nas equações (2) e (3) as seguintes suposições: $\boldsymbol{\beta}$ é um vetor de dimensão $H \times (J-1) = 6$;

$\boldsymbol{X} = \boldsymbol{I}_N$, $\boldsymbol{a} = (m_{11}, \dots, m_{31}, m_{12}, \dots, m_{32})$, $\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 \end{pmatrix}$, para $\boldsymbol{R}_h = \{r_{hjj'}\}$ com $r_{hjj} = c_{jh}$ e $r_{hjj'} = m_{jh}(m_{j'jh} - m_{j'h})$, para $h = 1,2$ e $j = 1, \dots, 3$, $\boldsymbol{V}_s = \begin{pmatrix} \boldsymbol{V}_{1s} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{2s} \end{pmatrix}$, para $\boldsymbol{V}_{hs} = \frac{1}{n_h}\{v_{hjj'}\}$ com $v_{hjj} = \sigma_{jh}^2$ e $v_{hjj'} = -m_{jh}m_{j'jh}$, para $h = 1,2$ e $j = 1, \dots, 3$. De forma análoga obtemos que $\boldsymbol{V}_{\bar{s}} = \begin{pmatrix} \boldsymbol{V}_{1\bar{s}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{2\bar{s}} \end{pmatrix}$, para $\boldsymbol{V}_{h\bar{s}} = \frac{n_h}{N_h - n_h}\boldsymbol{V}_{hs}$. Vale ressaltar que para estimar a variância do estimador, poderíamos aproximar usando a fórmula do estimador razão, haja visto que a obtenção do estimador linear bayesiano neste caso seria analiticamente complicada.

Um estudo feito com uma amostra de $n = 9.252$ indivíduos, sendo $n_1 = 4653$ no grupo dos vacinados e $n_2 = 4.599$ indivíduos no grupo dos que receberam placebo, observou-se $\hat{p}_{11} = 0$, $\hat{p}_{21} = 0{,}15\%$, $\hat{p}_{31} = 1{,}68\%$, $\hat{p}_{12} = 0{,}15\%$, $\hat{p}_{22} = 0{,}52\%$, $\hat{p}_{32} = 2{,}96\%$, obtendo assim a eficácia global estimada de $\widehat{Ef} = 50{,}4\%$. A aplicação de um método bayesiano neste caso, em particular o estimador linear bayesiano permitiria a introdução de um conhecimento prévio de especialista, ou no caso de pouca informação uma distribuição *a priori* vaga poderia ser utilizada. No caso do grupo 2 (placebo) ainda é possível usar na

especificação *a priori* dados observados durante a pandemia na população como um todo, já que não havia garantia de imunização nos indivíduos. A dificuldade em especificar os momentos *a priori* dados neste caso particular por probabilidades marginais e condicionais poderia ser contornada, assim como proposto em Gonçalves et al. (2014), por especificação de correlações *a priori*.

5.EXTENSÃO PARA PEQUENAS ÁREAS

Estimação em pequenas áreas é um problema que vem crescendo em importância, devido à grande demanda por informações estatísticas cada vez mais detalhadas e precisas para os setores público e privado. Uma das dificuldades neste campo é produzir estimativas confiáveis das características de interesse investigadas, baseando-se na informação proveniente de amostras muito pequenas obtidas nas pequenas áreas. Há casos também, em que os domínios de interesse são especificados somente após a pesquisa ter sido planejada, tendo assim uma amostra pequena para uma particular área. Uma solução possível para o problema de estimação é emprestar informações de outras áreas por meio de um modelo.

Por estas e outras razões, abordagens para estimação em pequenas áreas usando modelos hierárquicos têm sido apresentadas na literatura, incluindo o enfoque bayesiano. A inferência nesta abordagem é realizada encontrando a distribuição *a posteriori* de todos os parâmetros do modelo. Como é sabido, uma vantagem de usar esta abordagem é que são levadas em consideração as incertezas associadas aos parâmetros do modelo. No entanto, em geral, o núcleo da distribuição *a posteriori* não apresenta forma analítica conhecida e, portanto, torna-se necessária a utilização de métodos numéricos que aproximem esta distribuição, como métodos de Monte Carlo via Cadeias de Markov [MCMC] (Gamerman & Lopes, 2006). Dentre estes métodos destacam-se o algoritmo de Metropolis-Hastings e o Amostrador de Gibbs. O método linear de Bayes pode ser aplicado também neste contexto. Em particular, como o algoritmo de Metropolis-Hastings se baseia na escolha de uma distribuição auxiliar para sortear um valor para a cadeia de Markov, nossa proposta consiste em utilizar o método de estimação linear de Bayes para propor estas distribuições auxiliares.

5.1 MÉTODO LINEAR DE BAYES EM MODELOS HIERÁRQUICOS

Um modelo linear generalizado hierárquico pode ser descrito da seguinte forma: para $i = 1, \dots, n$

$$\begin{aligned} &y_i|\eta_i \sim FE(\eta_i) \\ &\lambda_i = g(\eta_i) = \boldsymbol{X_i\beta_i} \\ &\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\omega}_i, \boldsymbol{\omega}_i \sim [\mathbf{0}_p, \boldsymbol{\Sigma}], \end{aligned} \tag{6}$$

em que y_i é a variável de interesse associada à pequena área i; $\boldsymbol{\beta}$ é um vetor $p \times 1$ de coeficientes da regressão; $\boldsymbol{X_i}$ é o vetor de covariáveis com dimensão $1 \times p$; η_i é o parâmetro natural da distribuição, atualizado via função de ligação; λ_i's são funções lineares dos coeficientes; $\boldsymbol{\omega}_i$ é o vetor $p \times 1$ de efeitos aleatórios correspondente à área i e tem distribuição parcialmente especificada com p-vetor de médias nulo e matriz de covariância $\boldsymbol{\Sigma}$. Além disso, FE indica que a distribuição de $y_i|\eta_i$ pertence à família exponencial, ou seja, pode ser escrita da forma:

$$p(y_i|\eta_i) = \exp\{\phi_i[\psi_i(y_i)\eta_i - a(\eta_i)]\}\, b(y_i, \phi_i),$$

em que $\psi_i(.)$, $a(.)$ e $b(.,.)$ são funções conhecidas e ϕ_i é um parâmetro de escala.

Sob o enfoque Bayesiano o modelo ainda deve ser completado com a distribuição *a priori* do vetor paramétrico. Sob a hipótese de independência *a priori*, é comum assumir que $\boldsymbol{\beta} \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0)$ e $\boldsymbol{\Sigma^{-1}} \sim Wishart(\boldsymbol{W}, \boldsymbol{\nu})$.

A distribuição *a posteriori* do vetor paramétrico é obtida via Teorema de Bayes e, em geral, não apresenta forma conhecida, o que torna necessária a utilização de métodos numéricos que aproximem esta distribuição, como os métodos de MCMC. Estes consistem da obtenção de distribuições condicionais completas e amostras da distribuição *a posteriori* com base nestas. Neste caso, algumas condicionais completas não apresentam forma conhecida e portanto o método amostrador de Gibbs com passos de Metropolis-Hastings são utilizados. No passo de Metropolis-Hastings uma distribuição auxiliar é utilizada para sortear um valor proposto para a cadeia. Nossa proposta constitui em utilizar o método linear de Bayes para propor estas distribuições auxiliares. O método proposto baseia-se numa análise sequencial para séries temporais (Migon et al., 2013), mas em um contexto em que o parâmetro não evolui dinamicamente e que pode ser aplicado para estimação e previsão em população finita, ou, em particular, em pequenos domínios. Para isso agregamos ao modelo (6) uma distribuição *a priori* conjugada para o parâmetro natural da distribuição, da forma $\eta_i \mid \boldsymbol{y}^{i-1} \sim CExp(r_i, s_i)$, ou seja:

$p(\eta_i \mid \boldsymbol{y}^{i-1}) = k(r_i, s_i)\exp\,[r_i\eta_i - s_i\,a(\eta_i)],$

em que r_i e s_i são funções de $\boldsymbol{y}^{i-1}$; e $\boldsymbol{y}^{i-1}$ representa a informação da amostra até a $(i-1)$-ésima observação. Neste caso a distribuição *a posteriori de* η_i é dada por:

$p(\eta_i \mid \boldsymbol{y}^{i}) = k(r_i^*, s_i^*)\exp\,[r_i^*\eta_i - s_i^* a(\eta_i)],$

em que $r_i^* = r_i + \phi_i y_i$ e $s_i^* = s_i + \phi_i$. Logo, $\eta_i \mid \boldsymbol{y}^{\mathrm{i}} \sim CExp(r_i^*, s_i^*)$.

A análise sequencial inicia com a distribuição *a priori* parcialmente especificada $\boldsymbol{\beta}|\boldsymbol{y}^{i-1} \sim [\boldsymbol{m}_{i-1}, \boldsymbol{C_{i-1}}]$. Os momentos $\boldsymbol{m}_{i-1}$ e $\boldsymbol{C}_{i-1}$ são utilizados para determinar, através de um sistema de equações, r_i e s_i. Em seguida a distribuição de $\boldsymbol{\beta}$ deve ser atualizada, mas como esta não é conhecida, utilizamos uma aproximação via método linear de Bayes da seguinte forma:

$$p(\boldsymbol{\beta}|\boldsymbol{y}^{i}) \propto p(\boldsymbol{\beta}|\boldsymbol{y}^{i-1})p(y_i|\boldsymbol{\beta})$$

$$= \int p(\boldsymbol{\beta}|\eta_i, \boldsymbol{y}^{i-1})\, p(\eta_i \mid \boldsymbol{y}^{i-1})p(y_i|\boldsymbol{\beta})d\eta_i$$
$$= \int p(\boldsymbol{\beta}|\eta_i, \boldsymbol{y}^{i-1})\, p(\eta_i \mid \boldsymbol{y}^{i})d\eta_i$$
$$\approx [\boldsymbol{m}_i, \boldsymbol{C}_i],$$

em que $\boldsymbol{m}_i = \boldsymbol{m}_{i-1} + \boldsymbol{C}_{i-1}\boldsymbol{X_i'}(f_i^* - f_i)/q_i$ e $\boldsymbol{C}_i = \boldsymbol{C}_{i-1} - \boldsymbol{C}_{i-1}\boldsymbol{X_i'}\boldsymbol{X_i}\boldsymbol{C}_{i-1}(1 - q_i^*/q_i)/q_i$, para f_i e q_i média e variância *a priori* de λ_i, respectivamente, os quais são obtidos com base em (6), e f_i^* e q_i^* denotam, respectivamente, a média e variância *a posteriori* de λ_i, os quais são obtidos como função de r_i^* e s_i^*. A seguir está um passo-a-passo da análise sequencial proposta para obtenção de uma aproximação para a distribuição *a posteriori* de $\boldsymbol{\beta}$: faça $i = 1$ e

(i) $\boldsymbol{\beta}|\boldsymbol{y}^{i-1} \sim [\boldsymbol{m}_{i-1}, \boldsymbol{C}_{i-1}]$;

(ii) $\lambda_i|\boldsymbol{y}^{i-1} \sim [f_i, q_i]$ para $f_i = \boldsymbol{X_i}\,\boldsymbol{m}_{i-1}$ e $q_i = \boldsymbol{X_i}(\boldsymbol{C}_{i-1} + \boldsymbol{\Sigma})\boldsymbol{X_i'}$;

(iii) $\eta_i \mid \boldsymbol{y}^{\mathrm{i}-1} \sim CExp(r_i, s_i)$ e determine r_i e s_i em correspondência com f_i e q_i, por meio de compatibilização da média e variância, ou seja $f_i = E\big(g(\eta_i)\big|\boldsymbol{y}^{i-1}\big) = \alpha(r_i, s_i)$ e $q_i = V\big(g(\eta_i)\big|\boldsymbol{y}^{i-1}\big) = \gamma(r_i, s_i)$;

(iv) $\eta_i \mid \boldsymbol{y}^{\mathrm{i}} \sim CExp(r_i^*, s_i^*)$;

(v) $\lambda_i|\boldsymbol{y}^{i} \sim [f_i^*{}_i, q_i^*]$, onde $f_i^* = E\big(g(\eta_i)\big|\boldsymbol{y}^{i}\big) = \alpha(r_i^*, s_i^*)$ e $q_i^* = V\big(g(\eta_i)\big|\boldsymbol{y}^{i}\big) = \gamma(r_i^*, s_i^*)$;

(vi) $\boldsymbol{\beta}\big|\boldsymbol{y}^{i} \sim [\boldsymbol{m}_i, \boldsymbol{C}_i]$ e, faça $i = i + 1$ e repita todo o processo para $i = 1, \ldots, n$.

No entanto, além de $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ e $\boldsymbol{\omega_i}$ devem ser estimados. Portanto, o método de estimação proposto para essa classe de modelos via linear Bayes segue os seguintes passos:

(i) inicie o contador fazendo $k = 0$ e atribua valores iniciais para os parâmetros, i.e., $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $\boldsymbol{\omega}_i^{(0)}$, para $i = 1, \ldots, n$;

(ii) faça $k = k + 1$ e obtenha pelo método linear de Bayes os momentos *a posteriori* de $\boldsymbol{\beta}^{(k)}$ da forma $\boldsymbol{\beta}^{(k)}|\boldsymbol{y}_s, \boldsymbol{\Sigma}^{(k-1)}, \boldsymbol{\omega}^{(k-1)} \sim [\boldsymbol{m}_n, \boldsymbol{C}_n]$;

(iii) amostre $\boldsymbol{\beta}^*$ de uma distribuição proposta $N(\boldsymbol{m}_n, \boldsymbol{C}_n)$ e faça $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^*$ com probabilidade p e $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)}$ com probabilidade $1 - p$, para $p = \min(1, A)$, sendo $A = \frac{p(\boldsymbol{\beta}^*|\boldsymbol{y}_s, \boldsymbol{\Sigma}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})q(\boldsymbol{\beta}|\boldsymbol{m}_n, \boldsymbol{C}_n)}{p(\boldsymbol{\beta}|\boldsymbol{y}_s, \boldsymbol{\Sigma}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})q(\boldsymbol{\beta}^*|\boldsymbol{m}_n, \boldsymbol{C}_n)}$, onde q é a densidade proposta com momentos $\boldsymbol{m}_n$ e $\boldsymbol{C}_n$;

(iv) amostre de $\boldsymbol{\omega}_i^{(k)}$, para $i = 1, \ldots, n$, usando um passo de Metropolis-Hastings;

(v) amostre de $\boldsymbol{\Sigma}^{(k)}$, usando o amostrador de Gibbs e faça $k = k + 1$;

(vi) repita dos passos (ii) ao (v) até que a convergência seja obtida.

A distribuição proposta obtida via método linear de Bayes fornece uma boa aproximação para a distribuição objetivo $p(\boldsymbol{\beta}\,|\,.)$, portanto as taxas de aceitação são em geral bastante razoáveis. O método pode ainda ser aplicado a uma extensão do modelo (6) para dois níveis, quando se tem informações a respeito das unidades dentro de cada pequena área.

6.CONCLUSÕES

A adoção do paradigma Bayesiano de inferência é algumas vezes questionada pela necessidade de suposição de uma distribuição *a priori* conjunta para o vetor paramétrico. Neste trabalho, o método linear de Bayes foi apresentado como uma alternativa no contexto de população finita, já que requer apenas a determinação *a priori* de médias, variâncias e covariâncias para os parâmetros. Em trabalhos práticos especialistas da área podem ajudar na tarefa de escolha dessa informação *a priori* na forma de momentos.

Neste trabalho adotamos a abordagem de inferência em populações finitas baseada em modelos e mostramos como hipóteses de permutabilidade devem ser adotadas de forma a gerar resultados análogos aos estimadores via abordagem baseada em desenho amostral. Em seguida apresentamos o método linear de Bayes neste contexto aplicado a uma abordagem de modelo de regressão geral, com o pacote *BayesSampling*, feito para aplicação desta

metodologia em previsão em populações finitas. Apresentamos duas aplicações inéditas da metodologia. A primeira aplicação é feita numa base de dados muito comum em pacotes de amostragem, com a finalidade de ilustrar o impacto das especificações *a priori* nos preditores. Na segunda aplicação estendemos preditores obtidos em Gonçalves et al. (2014) com o objetivo de estimar a eficácia da vacina Coronavac para os casos de Covid-19.

Finalmente, apresentamos uma extensão da metodologia para modelos hierárquicos em pequenos domínios. Propomos um procedimento de inferência que utiliza métodos de simulação estocástica, em particular o MCMC, com um detalhe adicional de elaboração de uma distribuição proposta no passo de Metropolis-Hastings via método linear de Bayes.

7.REFERÊNCIAS BIBLIOGRÁFICAS

Al-khasawneh, M. F. (2019). Empirical Bayes sequential estimation of a finite population mean, Journal of Statistical Computation and Simulation, 89(12), 2175-2186.

D Cocchi & M Motjchart (1990). Linear bayes estimation in finite populations with a categorical auxiliary variable. Statistics, 21(3), 437-454.

D Cocchi & M Motjchart (1990). Quasi-linear Bayes Estimation in Stratified Finite Populations. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 293-300.

Figueiredo, P. S. & Gonçalves, K. C. M (2021). BayesSampling: Bayes Linear Estimators for Finite Population. R package version 1.1.0. https://CRAN.R-project.org/package=BayesSampling.

Gamerman, D. & Lopes, H. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Gonçalves, K.C.M., Moura, F.A.S & Migon, H.S. (2014). Bayes linear estimation for finite population with emphasis on categorical data. Survey Methodology, 40(1), 15-28.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

Karunamunia, R. J. & Zhang, S. (2003) Optimal linear Bayes and empirical Bayes estimation and prediction of the finite population mean. Journal of Statistical Planning and Inference, 113, 505-525.

Little, R.J. (2003). The Bayesian approch to sample survey inference. In Analysis of Survey Data, (Eds., R.L. Chambers and C.J. Skinner), chapter 4, 49-52. New York: John Wiley & Sons Inc.

Lumley, T. (2020). survey: analysis of complex survey samples. R package version 4.0.

Migon, H. S., Schmidt, A. M., Ravines, R. E. R. e Pereira, J. B. M. (2013) An efficient sampling scheme for dynamic generalized models. Computational Statistics, 28, 2267-2293.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97, 558-625.

O'Hagan, A. (1985). Bayes linear estimators for finite populations. Relatório Técnico 58, Department of Statistics - University of Warwick.

Pessoa, D., Damico, A. & Jacob, G. (2020). convey: Income Concentration Analysis with Complex Survey Samples. R package version 0.2.2.https://CRAN.R-project.org/package=convey.

Pessoa, D.; Silva, P. L. N. (2018) *Análise de dados amostrais complexos*. (https://djalmapessoa.github.io/adac/index.html)

Rojas, H. A. G. (2020). TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population. R package version 4.1.1.https://CRAN.R-project.org/package=TeachingSampling.

Silva, D. B. N. (1992) 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.

Smouse, E. (1984). A note on bayesian least squares inference for finite population models. Journal of the American Statistical Association, 79, 390-392.

Tillé, Y. & Matei, A. (2021). sampling: Survey Sampling. R package version 2.9. https://CRAN.R-project.org/package=sampling.

Zacks, S. (2002). In the footsteps of Basu: The predictive modelling approach to sampling from finite population. Sankhyā: The Indian Journal of Statistics, Series A, 64, 532-544.

AGRADECIMENTOS e COLABORAÇÕES

Kelly C. M. Gonçalves recebe financiamento da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ, ARC/210.047/2018), Brasil.

Helio S. Migon recebe financiamento da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ, E-26/007/10667/2019), Brasil.

ISSN 2675-3243

Volume 79

Número 246

Janeiro/Junho 2021