ISSN 1413-9243

TEXTOS NEPO

73

CAMPINAS, MARÇO DE 2016

UMA PROPOSTA DE UTILIZAÇÃO DO *SOFTWARE R* PARA A CONSTRUÇÃO DE ALGORÍTMOS DE AVALIAÇÃO DA QUALIDADE DA DECLARAÇÃO DA IDADE

LUCIANA CORREIA ALVES (ORG.)
PEDRO GOMES ANDRADE
PIER FRANCESCO DE MARIA
ANA CAMILA RIBEIRO PEREIRA
RAFAEL LOPES MARINS
GUSTAVO PEDROSO DE LIMA BRUSSE
KELLY CRISTINA DE MORAES CAMARGO

FICHA CATALOGRÁFICA: Adriana Fernandes

Alves, LucianaCorreia (Org.) et al.

Uma proposta de utilização do *Software R*para a construção de algoritmos de avaliação da qualidade da declaração da idade /Luciana Correia Alves (Org.) et al. - Campinas, SP: Núcleo de Estudos de População “Elza Berquó” / Unicamp, 2016.

30p.

(Uma proposta de utilização do *Software R*para a construção de algoritmos de avaliação da qualidade da declaração da idade, TEXTOS NEPO 73).

1. Declaração de idade. 2. *Software R*. 3. Qualidade dos dados. 4. Algoritmos.7. Título. 8. Série.

SÉRIE TEXTOS NEPO

TEXTOS NEPO - publicação seriada do Núcleo de Estudos de População "Elza Berquó" da UNICAMP - foi criado em 1985 com a finalidade de divulgar pesquisas no âmbito deste Núcleo de Estudos e Teses defendidas dentro do Programa de Pós-Graduação em Demografia do IFCH/UNICAMP. Apresentando uma vocação de cadernos de pesquisa, até o presente momento foram publicados **setenta e trêsnúmeros**, contando com este, relatando trabalhos situados nas áreas temáticas correspondentes às linhas de pesquisa do NEPO.

Os exemplares que compõem a série vêm sendo distribuídos para instituições especializadas na área de Demografia, ou mesmo dedicadas a áreas afins, no País e no exterior, além de ser objeto de constante consulta no próprio Centro de Documentação do NEPO. Essa distribuição é ampla, abrangendo organismos governamentais ou não governamentais – acadêmicos, técnicos e/ou prestadores de serviços.

A Coleção **Textos NEPO** também está acessível na homepage do NEPO, em publicações, cujo acesso se dá através do endereço eletrônico: http://www.nepo.unicamp.br.

DrªMarta Maria do Amaral Azevedo
Coordenadora

Dr.Alberto Augusto Eichman Jakob
Coordenador Associado

SUMÁRIO

RESUMO

O objetivo do presente trabalhoconsiste em apresentar a aplicação de técnicas demográficas em análisesda qualidade de declaração da idade, por meio de algoritmos em *R*.A motivação do trabalho advém do que foi aprendido em disciplinas do Programa de Pós-Graduação em Demografia da UNICAMP. Ao longo dos cursos,notou-se a necessidade da construção de algoritmos que pudessem facilitar as análises, antes feitas predominantemente em *Excel*. A sistematização em *R* pode agir no sentido de colaborar com a reprodução e a promoção da verificação da qualidade de uma das principais variáveis demográficas, a idade. Além de empregarmétodos tradicionais aprendidos na disciplina,este trabalho procura aprofundar a discussão da aplicação dos próprios métodospor meio de aprimoramentos e/ou técnicas mais recentes. Os dados utilizados para exemplificar a utilização dos algoritmos foram da amostra dos Censos Demográficos brasileiros de 1980 e 2010, para oestado do Amapá. Verificou-se que a escolha dométodo usadodeve estar, sobretudo, associadaà disponibilidade dos dados e também: 1) aos dígitos analisados para verificar a atração; e 2) à faixa etária coberta por cada método.

Palavras-chave: Declaração de idade.S*oftware R*.Qualidade dos dados.Algoritmos.

ABSTRACT

The main purpose of this workis to present the application of demographic techniques for analysis of the age data accuracy through algorithms in *R*. The motivation came from the courses attended at the Graduate Program in Demography of UNICAMP. Throughout that, it was seen the need to build algorithms that could facilitate the analysis, prior made mainly on *Excel*. The systematization in *R* can act in the way to facilitate the replication and the promotion of the evaluation of quality of one of the main demographic variable, the age. In addition to traditional methods learned in the course, this work attempts to further the methods application discussion, seeking the latest and most techniques. The data used to exemplify the algorithms application were collected from the samples of Brazilian Demographic Census of 1980 and 2010, using the state of Amapá. It was found that the method choice to be used should be primarily associated to the availability of data and, thereafter, to two other aspects: 1) the digits analyzed to verify age heaping and digit preference; and 2) the age range of each method.

Keywords: Age heaping.S*oftware R*; Data quality; Algorithms.

UMA PROPOSTA DE UTILIZAÇÃO DO *SOFTWARE R*PARA A CONSTRUÇÃO DE ALGORÍTMOS DE AVALIAÇÃO DA QUALIDADE DA DECLARAÇÃO DA IDADE

Luciana Correia Alves (org.)[1]
Pedro Gomes Andrade♠
Pier Francesco De Maria•
Ana Camila Ribeiro Pereira♣
Rafael Lopes Marins▽
Gustavo Pedroso de Lima Brusse♦
Kelly Cristina de Moraes Camargo♥

1. APRESENTAÇÃO

A motivação para a elaboração desse trabalho vem do conhecimento das técnicas demográficas aprendidas na disciplina DM002 (Laboratório de Análise Demográfica I), oferecida pelo Departamento de Demografia do Instituto de Filosofia e Ciências Humanas (IFCH) da UNICAMP,durante o primeiro semestre de 2015, pela Profª. Drª. Luciana Alves. A disciplina, essencialmente prática, permitiu o nosso primeiro contato e a aplicação de um conjunto de técnicas utilizadas na demografia, cujo arcabouço teórico era fornecido pela disciplina DM001 (Análise Demográfica I), ministrada pelo Prof.Dr. José Marcos Pinto da Cunha. O presentetrabalho foi desenvolvido por um grupo de alunos do Programa dePós-Graduação em Demografia, sob supervisão da Profª. Drª. Luciana Alves.

O conhecimento adquirido a partir dessas disciplinasdespertou algumas reflexões como: 1)a necessidade de estudar a qualidade dos dados e; 2)a indispensabilidade do ajuste de possíveis distorções na declaração de idade, como acontece, por exemplo,no Censo Demográfico, do Instituto Brasileiro de Geografia e Estatística (IBGE). O uso de informações populacionais distribuídas por idade é essencial desde os métodos mais simples até os mais sofisticados.Ademais,diferenciais nessa questão podem resultar em alterações dos resultados levando a análises errôneas Portanto, o estudo da qualidade da declaração de idade torna-se relevante não só para a Demografia, mas para qualquer área que utiliza indicadores sociodemográficos distribuídos por idade.

[1]Professora do Departamento de Demografia e pesquisadora do Núcleo de Estudos de População "Elza Berquó" da Universidade Estadual de Campinas (UNICAMP). Email: luciana@nepo.unicamp.br.
♠Doutorando do Programa de Pós-Graduação em Demografia da UNICAMP. Email: pedrogandrade@yahoo.com.br.
•Doutorando do Programa de Pós-Graduação em Demografia da UNICAMP. Email: dpierf@gmail.com.
♣Doutoranda do Programa de Pós-Graduação em Demografia da UNICAMP. Email: anacamilarp@gmail.com.
▽Mestrando do Programa de Pós-Graduação em Demografia da UNICAMP. Email: rlmarins@gmail.com.
♦Mestrando do Programa de Pós-Graduação em Demografia da UNICAMP. Email: gustavo.brusse@gmail.com.
♥Mestranda do Programa de Pós-Graduação em Demografia da UNICAMP. Email: kee.cmc@gmail.com.

Após o término da disciplina, verificou-se que as técnicas aprendidas em *Excel* poderiam ser sistematizadasno *softwareR*, visando facilitar sua reprodução a partir de bases de dados diversas em menor tempo.Notou-se, assim, um potencial aproveitamento nas pesquisas de cunho demográfico. A disseminação de técnicas demográficas por meio de códigos em *R* pode facilitar e promover a utilização das mesmas, por exemplo, estimulando a avaliação da qualidade dos dados, antes que sejam aplicadas técnicas demográficas.

Este trabalho se divide em mais6 seções, além desta introdução. A primeira discute a importância da utilização do *software R* na Demografia. A segunda,a relevânciada discussão da qualidade da declaração de idade nos estudos demográficos. Em seguida, são retratadasas principais técnicas de detecção de erros de declaração da idade eos algoritmos propostos para sua captação. Após a apresentação teórica e dos algoritmos, é realizada a aplicação das técnicas em um estudo de caso (oestado do Amapá em 1980 e 2010, que teve avanços na qualidade da declaração de idade). Cabe ressaltar que todos os algoritmos utilizados são expostosao longo do trabalho e disponibilizados por meiodo site: Baixar Algoritmo. Por fim, são tecidas algumas breves considerações finais.

Desta forma, esse trabalho pretende realizar uma aplicação metodólogica das principais técnicas de avaliação da declaração de idade, promovendo a utilização do *software R* nademografia,nãotendo como objetivoinvestigar a fundo os fatores associadosà qualidade da declaração. Ao longo do trabalho, o leitor poderá compreender a utilização do *R*, possibilitando a implementação dos algoritmos propostos, mesmo aos usuários que não são familiarizados com o *software*.

2. A IMPORTÂNCIA DA UTILIZAÇÃO DO SOFTWARE *R* NA DEMOGRAFIA

O *softwareR* compreende uma linguagem de programação criada em 1996 por Ross Ihaka e Robert Gentleman na universidade de Auckland, Nova Zelândia. O *software* não se finalizou com sua criação, uma vez que é cotidianamente aprimorado a partir de um esforço colaborativo de indivíduos de todo o mundo. Apesar de não ser propriamente um*software* estatístico, o *R* é uma série integrada de instalações de pacotes que permite, por exemplo, a manipulação de dados, a realização de cálculos e a geração de gráficos (SILVA; DINIZ; BORTOLUZZI, 2009).

Ser intitulado como "*R*" decorreu da utilização da primeira letra do nome de seus principais criadores, e também em razão de um jogo que utiliza a linguagem "S". Dessa forma, observa-se que o *R* se assemelha a tal linguagem, desenvolvida pela *AT&T's Bell Laboratories*, mas com um diferencial importante: o *R*é um *software* livre (SILVA; DINIZ; BORTOLUZZI, 2009), ou seja, gratuito e de domínio público. Em decorrência disso, sua utilização tem se popularizado tanto em universidades

como em empresas, uma vez que *softwares* estatísticos semelhantes possuem preços de aquisição elevados.

Mesmo gratuito, trata-se de um *software* expansivo que possui alta capacidade de programação, podendo ser empregado nas mais distintas áreas do conhecimento. Em decorrência do fato de ser livre, é possível encontrar na internet vários pacotes que podem ser rodados em *R*, constantemente disponibilizados e atualizados pelos próprios usuários. Esses pacotes devem ser entendidos como bibliotecas para funções específicas; portanto, quando um indivíduo cria uma rotina,é possível que ele disponibilize sua elaboração na rede, para que outros usuáriospossam utilizá-la para a mesma atividade, ou adaptá-la.

Como o *R*conta com colaboração internacional, quaisquer defeitos ou possibilidades de melhorias de programação são detectados e corrigidos por indivíduos de várias partes do mundo, o que aumenta a confiabilidade do programa (PACHECO; CUNHA;ANDREOZZI, 2014). Sendo assim, existe a possibilidade de usar o *softwareR* para implementação dos estudos de técnicas demográficas.

Ao longo do tempo, a Demografia tem se caracterizando pelo desenvolvimento de técnicas e análises que buscam descrever as mudanças pelas quais passam as populações humanas.Desta maneira, o manuseio de técnicas diretas e indiretas, que corrigem e/ou ajustam dados, a construção e a implementação de modelos estatísticos e a posterior análise de seus resultados compreendem o que podemos chamar de fazer demográfico (HAKKERT, 1996). O demográfo aplica tais métodos quantitativos para planejamento, diagnóstico e avaliação do tamanho, da distribuição, da composição e da organização de uma população.

Portanto, o manuseio de microdados, de pesquisas comoo Censo Demográfico e a Pesquisa Nacional por Amostra de Domicílios (PNAD),ambas do IBGE, são afazeres diários que podem ser simplificados com a utilização do *R*. Contudo, nas plataformas de pesquisas acadêmicas nacionais, são poucos os trabalhos que desenvolvem pacotes para a aplicação dos modelos estatísticos com o *R*. Também é difícil encontrar pesquisas que informem, em sua metodologia, que os dados foram trabalhados em*R*e quais foram as rotinas utilizadas. Observa-se, então, certa relutância no uso do *software*, seja pela familiaridade com outros pacotes estatísticos, seja pela hesitação que a elaboração ou adaptação de uma rotina pode causar na primeira impressão.

3. A IMPORTÂNCIA DA QUALIDADE DA DECLARAÇÃO DE IDADE EM ESTUDOS DEMOGRÁFICOS

A confiança na estrutura etária da população em estudoé pressuposto de diversos métodos e técnicas, diretas e indiretas, relacionadas às componentes da dinâmica demográfica: fecundidade, mortalidade e migração (ONU, 1983; MOULTRIE et al., 2013). Todavia, adverte-se que nem sempre a

distribuição por idade simples possui boa qualidade (*cf.* ONU, 1955, p. 34) e isso pode dificultar os estudos demográficos quando os mesmos não consideram esse problema. Por isso, a qualidade da declaração da idade é uma preocupação constante dos demógrafos de várias partes do mundo. Inclusive, no Brasil se experimentaram diversos avanços nos últimos anos, sobretudono que tange a forma de coleta dos dados censitários, por exemplo.

Dessa forma, em 1955, dois anos após a publicação do seu primeiro manual, a Organização das Nações Unidas (ONU) lança o Manual II, "*Methods of appraisal of quality of basic data for population estimates*", com o intuito de propor métodos para avaliação da qualidade de dados vinculados a estimativas populacionais. Na época, existia grande interesse em propor métodos que possibilitassem avaliar a acurácia dos dados em regiões pobres do mundo, de modo a viabilizara realização de estudos e aplicações de métodos demográficos de mortalidade e fecundidade, sobretudo em regiões que registravam altas taxas de natalidade.

Nesse sentido, o Manual II foi a primeira publicação com visibilidade internacional a organizar estes métodos, sendo relevante até hoje. Nele, é possível verificar diversas recomendações, como, por exemplo, o recurso de se trabalhar com grupos etários quinquenais de, com o intuito de suavizar possíveis erros de declaração de idade (como a atração por dígitos). Este recurso é uma importante ferramenta, mas quando existe o interesse em trabalhar com idades simples, ou construir indicadores com grupos de idades que não sejam quinquenais, esta estratégia se torna pouco útil. Nestes casos, existem técnicas que podem ser utilizadas, como os métodos de interpolaçãode Sprague, Karup-King e Beers[2], cujo objetivo é suavizar a distribuição por idade, eliminando possíveis distorções em sua declaração.

Em geral, os erros na declaração na idade ocorrem por dois motivos: 1) a preferência por dígitos; e 2) o erro/viés de memória do respondente. Além destes, em pesquisas domiciliares, é comum um indivíduo, além de prestar informações sobre si próprio, fazer o mesmo a respeito de outros indivíduos, não presentes no domicílio no momento da pesquisa (sendo a este dado o nome de informante *proxy*). Neste caso, o erro de memória pode estar presente quando o respondente declara informações de si e,de forma aindamais acentuada, quando esteresponde por outras pessoas do domicílio.

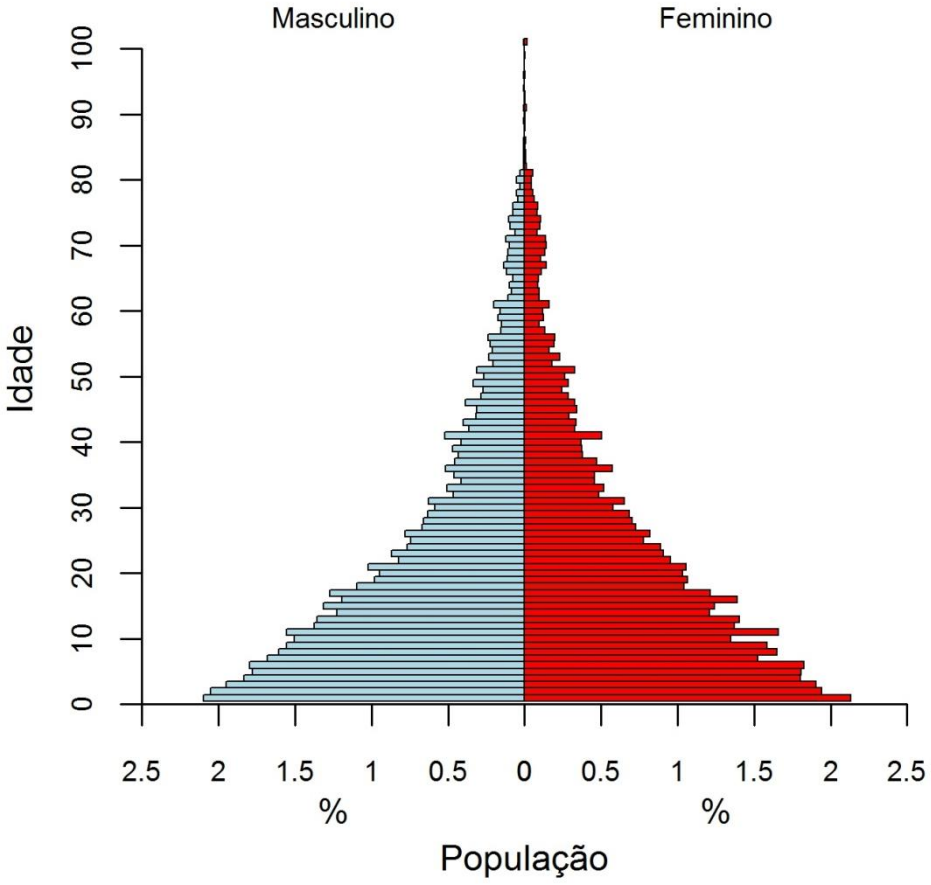
Paes e Albuquerque (1999) recomendam a avaliação de possíveis erros nos dados antes da geração de indicadores demográficos e socioeconômicos. A ONU (1955) sugere algumas técnicas para avaliar a qualidade da declaração da idade. O método mais simples, porém não menos eficaz, é realizar uma inspeção visual da pirâmide etária por idade simples, verificando distorções em torno de

[2] Para maiores informações a respeito destes métodos, ver Shryock e Siegel (1976).

dígitos específicos (como “0” e “5”, por exemplo), conforme observado no Amapá em 1980, para idades como 40, 45, 50, 60 anos (Gráfico 1).

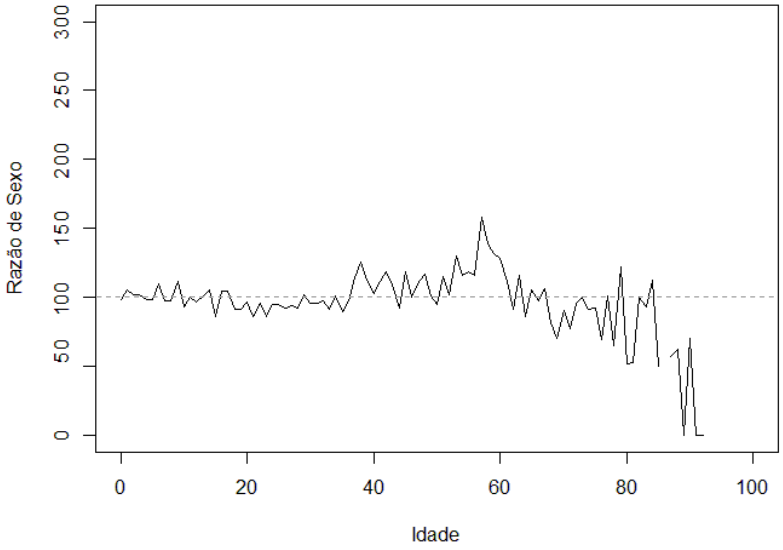
Nestas idades, é possível observar claramente pontos de inflexão na distribuição etária para ambos os sexos, sendo que, para esse tipo de estrutura em formato piramidal,o esperado é tergrupos etários subsequentes menores que os precedentes. Outra forma de avaliação, também simples, seria a análise das séries de razões de sexo distribuídas por idade simples, propiciando aindaanalisar a preferência por dígitos (Gráfico 2). Neste caso, observa-se que a distribuição da série não é suave, principalmente nas idades inferiores a 20 anos e naquelas acima de 35 anos, indicando a possível presença de distorções na declaração da idade.

GRÁFICO 1 -Distribuição relativa por sexo e idade simples, Amapá, 1980



Fonte: IBGE (1980).

GRÁFICO 2 -Razão de sexo por idade simples, Amapá, 1980



Fonte: IBGE (1980).
Notas: Descontinuidades causadas pela ausência de homens ou mulheres na amostra do censo.

Outras técnicas mais sofisticadas ajudam a identificar e comparar erros, como atração por dígitos específicos. Dentre elas, podemos destacar: o Índice de Myers; o Índice de Whipple e; o Método das Nações Unidas. Estas técnicas e variações, propostas mais recentemente, serão detalhadas na próxima seção.

4. TÉCNICAS DE DETECÇÃO DE ERROS DE DECLARAÇÃO DA IDADE

Os índices para avaliação da declaração da idade foram construídos a fim de fornecer uma medida mais objetiva, ajudando a complementar a inspeção visual realizada a partir dos gráficos, favorecendo a identificação dos dígitos nos quais ocorre a atração. De forma geral, todos os índices a serem apresentados alcançam esse objetivo; o que varia é a capacidade diagnóstica, uma vez que a variação consiste nos grupos etários utilizados no cálculo e os dígitos finais analisados.

Cabe ressaltar que, ao longo do tempo, foram feitas novas recomendações a partir da criação de outros índices ou pelas modificações dos originais, possibilitando avaliar a atração de todos os dígitos finais de “0” a “9” de mais de uma maneira. Em suma, os índices aqui apresentados(Myers, Whipple e o método das Nações Unidas) partem do pressuposto de que não existe atração por dígitos finais (ou seja, a distribuição seria uniforme) e, a partir disso, se propõem a identificar em qual ou quais dígitos existe a violação deste pressuposto.

4.1 Índice de Whipple

O método desenvolvido por Whipple (formulado em 1924) procura por sinais de atração por dígitos na declaração de idade. Conforme reportado pela ONU (1955), é um método muito simples de ser aplicado, com resultados efetivos, embora seja limitado ao ser aplicável, em sua origem,somente para os dígitos 0 e 5. Em compensação, como a própria ONU (1955) ressalta, esses dois dígitos são os mais comuns de serem preferidos pelos pesquisadores, ainda que existam diferenças culturais em todo o mundo. Ademais:"pode se assumir que esta preferência [pelos dígitos 0 e 5] está geralmente associada a outras fontes de inacurácia na declaração de idade, e os índices podem ser aceitos como uma medida geral de credibilidade da distribuição etária" (ONU, 1990, p. 20, tradução nossa).

Os dados utilizados para o cálculo do índice levam em conta a população de 23 a 62 anos. Esta decisão é de certo modo arbitrária, mas reflete o fato de que, tanto nas idades mais jovens quantonas mais avançadas, a qualidade da declaração de idade tende a ser pior, por conta de questões para além da atração digital (POSTON; MICKLIN, 2005). Ademais, o pressuposto de decremento linear na população com o avançar da idade geralmente não é válido para a população idosa (SHRYOCK; SIEGEL, 1976) e, em geral, é menos plausível para os mais jovens e os mais velhos (NOUMBISSI, 1992).

Para a aplicação do método é necessário ter a população por idade simples. A partir disto, obtém-se a soma da população com idades múltiplas de 5 e da população entre 23 e 62 anos, para depois computar a razão entres essas somas (equação 1). Primeiramente, o método foi proposto para verificar a atração no dígito "0" (equação 2) e nos dígitos "0 e 5" em conjunto (equação 1). Complementarmente, Roger *et al.*(1981)mostraram queé possível aplicar a mesma equação do dígito "0" para verificar a atração de todos os outros dígitos, obtendo um "*Whipple-type index*" que analisa a atração por qualquer dígito (equação 2).

$$IW_{0,5} = \frac{P_{25} + P_{30} + \cdots + P_{55} + P_{60}}{{}^{1}/_{5}\,(P_{23} + P_{24} + \cdots + P_{61} + P_{62})} \times 100, \qquad 0 \leq IW_{0,5} \leq 500 \qquad \textbf{(1)}$$

$$IW_i = \frac{\sum \text{idades terminadas em } i \text{ entre 23 e 62}}{{}^{1}/_{10}\,(P_{23} + P_{24} + \cdots + P_{61} + P_{62})} \times 100, \qquad \begin{cases} 0 \leq IW_i \leq 1.000 \\ i = 0, 1, \ldots, 9 \end{cases} \qquad \textbf{(2)}$$

Em ambas as equações, P_i indica a população de determinada idade,i (na equação 2) indica o dígito final analisado e IW_ise refere ao resultado do índice para o dígito analisado. Os valores do

índice de Whipple para os dígitos “0” e “5” conjuntamente(equação 1),variam de 0 (quando ninguém, na população, declarou qualquer idade com determinado dígito) a 500 (quando só há, na população, pessoas que declararam idades com dígito final 0 ou 5).

Geralmente, o índice varia de 100 (quando não há sinal de atração por dígito) a 500. Nos casos em que se analisa um dígito específico e não a atração pelos múltiplos de 5 (equação 2), o índice varia de 0 a 1.000, nos indicando que, quanto maior o resultado, maior a violação do pressuposto de distribuição uniforme dos dígitos finais. Os resultados obtidos a partir do índice de Whipple (independente de analisar a atração pelos dígitos 0 e 5 ou por qualquer dígito em separado) podem ser classificados conforme apresentado na tabela abaixo.

TABELA 1 -Classificação do Índice de Whipple

Classificação dos dados	**Índice de Whipple**
Precisos	De 99,0 a 104,9
Pouco precisos	De 105,0 a 109,9
Aproximados	De 110,0 a 124,9
Grosseiros	De 125,0 a 174,9
Muito grosseiros	175,0 ou mais

Fonte: Formiga; Ramos e Monteiro (2000).

A partir das modificações propostas por Roger;Waltisperger e Corbille-Guitton(1981); Noumbissi (1992) sugere um índice de Whipple modificado que permita a análise de atração para todos os dígitos. Isto seria mais apropriado, pois o pressuposto de linearidade é mais razoável para um intervalo etário menor (SPOORENBERG, 2007). Abaixo (equação 3), apresentamos o exemplo para a análise de atração pelos dígitos 0, 1 e 9: como se nota, a modificação põe no denominador a soma da população por grupos quinquenais nos quais a idade com o dígito a ser analisado é o ponto médio do grupo.

Por exemplo, em IWm_0(onde IWm representa o incide de Whipple modificado e o sub-índice 0 indica o dígito final para o qual está sendo calculado este índice),tem-se no numerador a soma da população (representada por P) com 30, 40, 50 e 60, multiplicado por cinco e o denominador a soma do volume populacional dos grupos etários nos quais a idade terminada em 0 é ponto médio: 28 a 32, 38 a 42, 48 a 52 e 58 a 62 anos.

$$IWm_0 = \frac{5 \times (P_{30} + P_{40} + P_{50} + P_{60})}{{}_5P_{28} + {}_5P_{38} + {}_5P_{48} + {}_5P_{58}}, \qquad 0 \leq IW_0 \leq 5 \qquad \textbf{(3)}$$

$$IWm_1 = \frac{5 \times (P_{31} + P_{41} + P_{51} + P_{61})}{{}_5P_{29} + {}_5P_{39} + {}_5P_{49} + {}_5P_{59}}, \qquad 0 \leq IW_1 \leq 5$$

⋮

$$IWm_9 = \frac{5 \times (P_{29} + P_{39} + P_{49} + P_{59})}{{}_5P_{27} + {}_5P_{37} + {}_5P_{47} + {}_5P_{57}}, \qquad 0 \leq IW_9 \leq 5$$

Como este não é um índice geral para análise da qualidade dos dados, não temos uma tabela que categorize o nível de acurácia. Ao invés disto, pode-se adaptar a Tabela 1, de modo que os valores obtidos reflitam o nível de atração por cada dígito.

TABELA 2 -Classificação da atração por dígito do índice de Whipple modificado

Nível de atração	Índice de Whipple modificado
Mínima	De 0,990 a 1,049
Pequena	De 1,050 a 1,099
Média	De 1,100 a 1,249
Grande	De 1,250 a 1,479
Muito grande	1,750 ou mais

Fonte: Elaboração dos autores.

Dado que ter um índice para cada um dos 10 dígitos não é prático em termos de comparação espaço-temporal, Spoorenberg (2007) propõe um índice de Whipple totalmente modificado ("*Total Modified Whipple's Index*", no original). Este serve como uma medida agregada que, de forma similar ao índice de Myers, dá um resultado global a respeito da preferência por dígito (equação 4). O índice assume:0 quando não há preferência observada; 13 quando há preferência perfeita por um dígito (por exemplo, todos declarando idades múltiplas de 5); e 16 quando há preferência por apenas dois dígitos (por exemplo, quandotodos declaram idades com dígito final igual a 0 ou a 5) (SPOORENBERG, 2007). Estes são os valores mais comuns que o índice de Whipple totalmente modificado assume.

$$W_{tot} = \sum_{i=0}^{9} |IWm_i - 1|, \qquad IWm_i \geq 0 \qquad \textbf{(4)}$$

Onde IWm_i é o valor do índice de Whipple modificado para um dos dez dígitos finais$i$ (calculado na equação 3).

4.2 Índice de Myers

Segundo Myers (1940), os erros de declaração de idade são provenientes da falta de acurácia na obtenção de estatísticas e, principalmente, pelas informações distorcidas dadas ao entrevistador, seja de forma intencional, seja por viéses ou erros de memória. Para detectar uma possível preferência de resposta para determinados dígitos na idade declarada, Myers desenvolveu um método apresentado no "*Transactions of the Actuarial Society of America*", no ano de 1940.

Para a implementação deste método, a ONU (1955) propõe a soma da população que possui o mesmo dígito final da idade declarada para o grupo de 10 a 99 anos (G_1) e para o grupo de 20 a 99 anos (G_2). A população acima de 100 anos pode ser dada como igual a zero, pois não tem indícios de afetar significativamente os resultados (MYERS, 1940, p. 413). Como a população tende a ser menor com o avanço dos dígitos finais (ou seja, para cada dígito sucessivo a população é mais velha e menor que a população anterior), essas populações são multiplicadas pelos coeficientes $1, 2, 3, \dots, 10$ no G_1 e pelos coeficientes complementares $9, 8, 7, \dots, 0$ no G_2 (equação 5). Somando os grupos G_1 e G_2, obtém-se a chamada "*blended population*" (BP) para cada dígito final (equação 6).

$$\begin{cases} G_1(i) = (i+1) \times \sum_{a=10}^{99} P_i\,, & \forall i \in \{0,1,\dots,9\} \\ G_2(i) = (9-i) \times \sum_{a=20}^{99} P_i\,, & \forall i \in \{0,1,\dots,9\} \end{cases} \tag{5}$$

$$BP_i = \left[(i+1) \times \sum_{a=10}^{99} P_i\right] + \left[(9-i) \times \sum_{a=20}^{99} P_i\right] \Longrightarrow f_i = BP_i \div \sum_{i=0}^{9} BP_i \tag{6}$$

Nas equações acima, i é o dígito final analisado, P_i é a população que declarou ter idade terminando no dígito i, BP é a "*blended population*" e f_i é a frequência (dada em valores de 0 a 1) que cada dígito i tem no total.

Para Myers (1940), a distribuição da população segundo o último dígito da idade declarada deveria obedecer uma distribuição uniforme, portanto, a proporção da população esperada em cada dígito seria de 10%. O índice de Myers para cada dígito final da idade declarada é calculado como o

desvio de $f_i$em relação aos 10% esperados (equação 7), enquanto o índice geral, que resume a qualidade dos dados para todos os dígitos, pode ser escrito como a soma dos índices de cada dígito (equação 8).

$$IM_i = |100 \times f_i - 10| \Longrightarrow 0 \leq IM_i \leq 90 \quad \textbf{(7)}$$

$$IM = \sum_{i=0}^{9} IM_i \Longrightarrow 0 \leq IM \leq 180 \quad \textbf{(8)}$$

Nas equações acima, IM_i é o índice de Myers para cada dígito final i, enquanto f_i é a frequência obtida da equação 6 e IM (na equação 8) é o índice de Myers geral.

Segundo a ONU (1955), o índice de Myers pode variar entre 0 a 180. Caso toda a população respondesse a idade declarada com o mesmo dígito final, o Índice de Myers seria igual a 180. Por sua vez, se todos os respondentes estiverem uniformemente distribuídos entre os 10 dígitos finais, o índice de Myers assume valor 0. Portanto, quanto menor for a preferência digital, mais próximo de 0 o índice será, indicando melhor qualidade da informação. Os resultados obtidos a partir do índice geral de Myers podem ser classificados conforme apresentado na tabela abaixo.

TABELA 3 - Classificação da atração geral pelo índice de Myers

Nível de atração	Índice de Myers
Baixo	Até 4,9
Mediano	De 5,0 a 14,9
Alto	De 15,0 a 29,9
Muito alto	30,0 a 180,0

Fonte: Formiga; Ramos e Monteiro (2000).

4.3 Método das Nações Unidas (AAI)

Quando a informação por idade simples não está disponível, mas se tem dados para grupos etários quinquenais até 70 anos e mais, as Nações Unidas sugerem o uso do *Age Accuracy Index* (AAI). Este método não é recomendado para populações com flutuações nas proporções de pessoas por idade, como no caso de pequenas áreas, depaíses afetados por guerras e deáreas de grandes fluxos migratórios. Entretanto, esse método difere dosíndicesde Myers e Whipple por, além de captar erros na declaração de idade, considerar o diferencial de cobertura por idade no Censo, refletindomelhor a exatidão da variável idade (ONU, 1955).

O método consiste em calcular as razões de sexo RS(equação 9) e as razões de idade RI para cada sexo s(equação 10), em cada faixa etária j (representando os grupos de 5-9 a 65-69 anos). Em seguida, mede-se o distanciamento destas razões em relação ao grupo anterior (no caso de RS) e em relação a 100 (no cas das RI), para se chegar a um índice geral sobre a acurácia dos dados (equação 11).

$$RS_j = \frac{P_j^m}{P_j^f} \times 100 \Longrightarrow D_S = \sum_{j=2}^{14} \left| RS_j - RS_{j-1} \right| \Longrightarrow MD_S = \frac{D_S}{13} \quad \textbf{(9)}$$

$$RI_j^s = \left[\frac{P_j^s \times 100}{1/3 \left(P_{j-1}^s + P_j^s + P_{j+1}^s \right)} \right] \Longrightarrow D_I^s = \sum_{j=2}^{14} \left| RI_j^s - 100 \right| \Longrightarrow {MD_I}^s = \frac{D_I^s}{13} \quad \textbf{(10)}$$

$$AAI = 3MD_S + {MD_I}^m + {MD_I}^f, \qquad AAI \geq 0 \quad \textbf{(11)}$$

Na equação 9, P_j é a população em cada um dos grupos etários quinquenais (iniciando em $j = 2$, que é o grupo de 5 a 9 anos), m e f indicam os dois sexos, RS_j é a razão de sexo para o grupo quinquenal j, D_S indica a diferença entre as razões de sexo dos grupos j e $j - 1$, enquando MD_S indica a média das diferenças. Na equação 10, s indica que o cálculo das razões de idade deve ser feito isoladamente - para homens e para mulheres-, RI_j^s representa a razão de idade do grupo etário j para o sexo s, D_I^S é o módulo do desvio, em relação a 100, de cada razão de idade e MD_I^S é o desvio médio para cada sexo.

Os resultados obtidos do índice de acurácia das Nações Unidas podem ser agrupados, conforme apresentado na tabela abaixo, em precisos, imprecisos e altamente imprecisos.

TABELA 4 -Classificação dos dados segundo o Método das Nações Unidas (AAI)

Classificação dos dados	***Age accuracy index***
Precisos	De 0,0 a 19,9
Imprecisos	De 20,0 a 39,9
Altamente imprecisos	40,0 ou mais

Fonte: Formiga; Ramos e Monteiro (2000).

5 ALGORITMOS PARA DETECÇÃO DE ERROS DE DECLARAÇÃO DA IDADE

O *R* possui diversas funções implementadas em seu módulo básico e em pacotes construídos para a execução de atividades específicas, disponíveis de forma gratuita para download[3]. Nos casos em que o usuário necessite de alguma função que ainda não foi criada ou disponibilizada, é possível construí-la por meio da elaboração de algoritmos em linguagem de programação *R*. A partir disso, é possível alocar estas funções em um pacote e acessá-las a partir do diretório de trabalho do computador. Para isto, as funções devem ser armazenadas dentro de um *script*, salvas com a extensão "*.r*", por exemplo "GEQD.r"(Grupo de Estudos em Qualidade dos Dados),e, quando for desejado acessá-las, basta utilizar o comando:

```
source("GEQD.r")
```

Conforme já apresentado, um dos objetivos deste trabalho é a construção de um pacote com funções que possibilitem avaliar a qualidade da declaração da idade. Os índices de Myers e Whipple (além do Whipple Modificado) necessitam dos mesmos dados de entrada, que é a distribuição por idade simples da população. Já no caso do Método das Nações Unidas,os dados devem estar distribuídos por sexo egrupos quinquenais de idade até 70 anos e mais. Esta simples diferença nos mostra que, enquanto o Método das Nações Unidas (AAI) só pode ser aplicado para a população total, os índices de Myers, Whipple e Whipple Modificado podem ser computados para o total e para ambos os sexos.

Desta forma, antes da construção dos algoritmos, é importante indicara forma como devem ser apresentados os dados de entrada. Nesse sentido, para Myers e Whipple, os dados devem estar distribuídos em um vetor linha de tamanho 101, contabilizando o volume de pessoas com menos de 1 ano até 100 anos ou mais de idade (Tabela 5). Já para o Método das Nações Unidas, os dados são distribuídos em uma matriz de 2 linhas e 15 colunas,sendo que as colunas representam os grupos etários (começando de "0 a 4 anos", até o grupo de "70 anos e mais") e cada linha indica o sexo (Tabela 6).

TABELA 5 - Exemplo de dados de entrada para cálculo dosíndice de Myers e Whipple

	Idade								
	<1	1	2	3	(...)	97	98	99	100+
População	369	348	374	418	(...)	1	0	0	1

Fonte: Elaboração dos autores com dados fictícios.

[3]Disponível em:<https://cran.r-project.org/>. Acesso em: 09 fev. 2016.

TABELA 6 -Exemplo de dados de entrada para cálculo do Método das Nações Unidas

	Idade								
	0-4	5-9	10-14	15-19	(...)	55-59	60-64	65-69	70+
Homens	493	364	378	311	(...)	202	150	102	113
Mulheres	385	346	357	305	(...)	205	183	125	131

Fonte: Elaboração dos autores com dados fictícios.

A partir da declaração dos dados de entrada para cada um dos métodos, são apresentados os algoritmos construídos em *R*,utilizando apenas funções do pacote básico, o que possibilita suaaplicação em qualquer versão do *R*, até a versão atual (*R* 3.2.3, de Dezembro/2015). Os Quadros 1 a 4 apresentam os algoritmos para o Índice de Myers, o Índice de Whipple, o Whipple Modificado e o Método das Nações Unidas, respectivamente.

QUADRO 1 -Algoritmo para o Índice de Myers (IM)

```
IM=function(data){
  IMy=matrix(NA,nrow=10,ncol=3)
  IM =matrix(NA,nrow=11,ncol=1)
for(i in1:10) {
IMy[i,1]=data[,10+i]+data[,20+i]+data[,30+i]+data[,40+i]+data[,50+i]+data[,
    60+i]+data[,70+i]+data[,80+i]+data[,90+i]
IMy[i,2]=data[,20+i]+data[,30+i]+data[,40+i]+data[,50+i]+data[,60+i]+data[,
    70+i]+data[,80+i]+data[,90+i]
IMy[i,3]=((i)*IMy[i,1])+((10-i)*IMy[i,2])
}
for(jin1:10) {
IM[j,]=abs(((IMy[j,3]/sum(IMy[,3]))*100)-10)
    }
  IM[11,]=abs(IM[1,]+ IM[2,]+ IM[3,]+ IM[4,]+ IM[5,]+ IM[6,]+ IM[7,]+
   IM[8,]+ IM[9,]+ IM[10,])
rownames(IM)=c("IM_0","IM_1","IM_2","IM_3","IM_4","IM_5","IM_6","IM_7","IM_
   8","IM_9","IM_GLOBAL")
return(round(IM,2))
}
```

Fonte: Elaboração dos autores.
Nota: Possíveis atualizações do algoritmo serão disponibilizadas no link:
http://www.nepo.unicamp.br/publicacoes/textos_nepo/script/download.php

QUADRO 2 -Algoritmo para o Índice de Whipple (IW)

```
IW=function(data){
IW=matrix(NA,nrow=11,ncol=1)
IWh=matrix(NA,nrow=10,ncol=3)
for (iin1:3) {
IWh[i,1] =data[,30+i] +data[,40+i] +data[,50+i] +data[,60+i]
    }
for (iin4:10) {
IWh[i,1] =data[,20+i] +data[,30+i] +data[,40+i] +data[,50+i]
    }
for (i in1:10) {
IWh[i,2] =IWh[i,1]/((1/10)*sum(data[,24:63]))*100
IWh[i,3] =abs(IWh[i,2]-1)
IW[i,] =IWh[i,2]
    }
IW[11,] = (data[,26] +data[,31] +data[,36] +data[,41] +data[,46] +data[,51]
   +data[,56] +data[,61])/((1/5)*sum(data[,24:63]))*100
rownames(IW)
  =c("IW_0","IW_1","IW_2","IW_3","IW_4","IW_5","IW_6","IW_7","IW_8","IW_9",
  "IW_0_5")
return(round(IW,2))
}
```

Fonte: Elaboração dos autores.
Nota: Possíveis atualizações do algoritmo serão disponibilizadas no link:
http://www.nepo.unicamp.br/publicacoes/textos_nepo/script/download.php

QUADRO 3 -Algoritmo para o Índice de Whipple Global (IWg)

```
IWg=function(data){
IWm=matrix(NA,nrow=10,ncol=4)
for (iin1:3) {
IWm[i,1] =data[,30+i] +data[,40+i] +data[,50+i] +data[,60+i]
IWm[i,2] =sum(data[,(28+i):(32+i)])+sum(data[,(38+i):(42+i)])
   +sum(data[,(48+i):(52+i)]) +sum(data[,(58+i):(62+i)])
    }
for (iin4:10) {
IWm[i,1] =data[,20+i] +data[,30+i] +data[,40+i] +data[,50+i]
    IWm[i,2] =sum(data[,(18+i):(22+i)]) +sum(data[,(28+i):(32+i)])
   +sum(data[,(38+i):(42+i)]) +sum(data[,(48+i):(52+i)])
    }
for (i in1:10) {
IWm[i,3] =5 * (IWm[i,1])/IWm[i,2]
IWm[i,4] =abs(IWm[i,3]-1)
    }
IWg=sum(IWm[,4])
return(round(IWg,2))
}
```

Fonte: Elaboração dos autores.
Nota: Possíveis atualizações do algoritmo serão disponibilizadas no link:
http://www.nepo.unicamp.br/publicacoes/textos_nepo/script/download.php

QUADRO 4 -Algoritmo para o Método das Nações Unidas (IN)

```
IN=function(data){
RS=matrix(NA,nrow=ncol(data),ncol=1)
D=matrix(NA,nrow=14, ncol=1)
RI=matrix(NA,nrow=13, ncol=4)
for (iin1:ncol(data)) {
RS[i,1] = (data[1,i]/data[2,i])*100
}
for (jin1:14) {
D[j,1] =abs(RS[j+1,1]-RS[j,1])
   }
for (kin2:14) {
RI[k-1,1] = (data[1,k]/((1/3)*(data[1,k-1] +data[1,k] +data[1,k+1])))*100
RI[k-1,2] = (data[2,k]/((1/3)*(data[2,k-1] +data[2,k] +data[2,k+1])))*100
RI[k-1,3] =abs(RI[k-1,1]-100)
RI[k-1,4] =abs(RI[k-1,2]-100)
    }
IN=(3*(sum(D[-14,])/13))+(sum(RI[,3])/13)+(sum(RI[,4])/13)
return(round(IN,2))
}
```

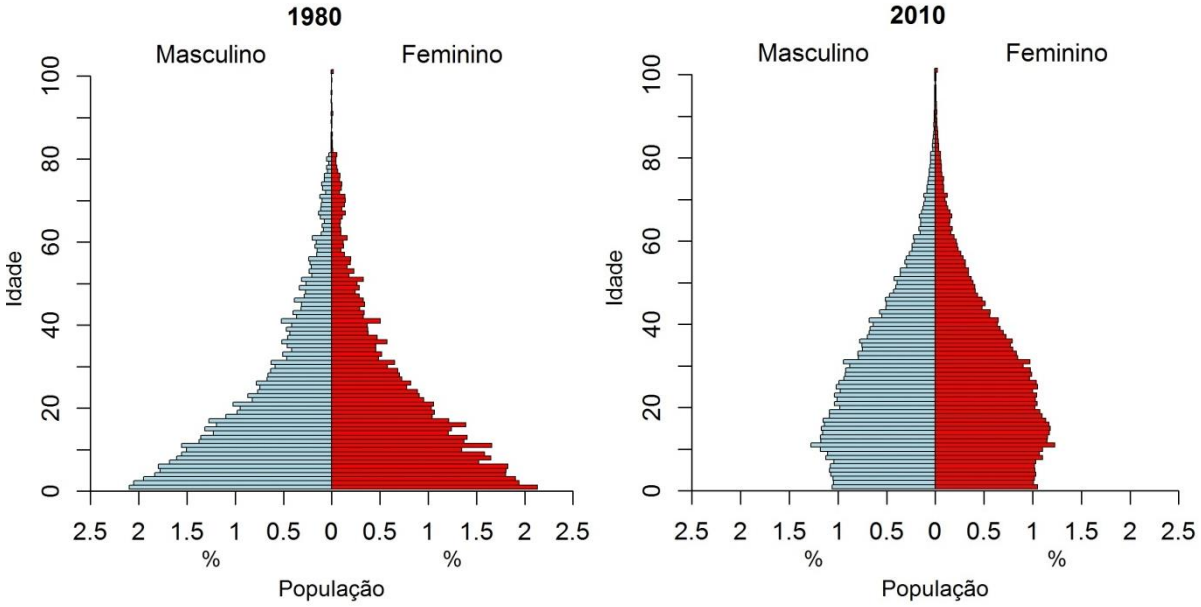
Fonte: Elaboração dos autores.
Nota: Possíveis atualizações do algoritmo serão disponibilizadas no link:
http://www.nepo.unicamp.br/publicacoes/textos_nepo/script/download.php

6 APLICAÇÃO DOS ALGORITMOS

Com o intuito de exemplificar a aplicação dos métodos apresentados neste trabalho, foram escolhidos os dados das amostras dos Censos Demográficos de 1980 e 2010 do estado do Amapá. A escolha deste estadose justifica pelasua notável evolução na qualidade da declaração da idade neste período, conforme será mostrado a seguir.Para a análise da qualidade da declaração da idade,se optou por primeiro realizar uma inspeção visual a partir dos gráficos (pirâmides etárias e séries de razões de sexo) e, posteriormente, aplicou-se cada um dos índices apresentados anteriormente.

Como já foi colocado, a ONU (1955)recomenda que,antes do cálculo dos índices,seja realizada uma inspeção visual das pirâmides etárias por idade simples (Gráfico 3). Pode-se notar, tanto na pirâmide de 1980 como, em menor proporção, na de 2010, que existia certa atração em torno dos dígitos finais“0” e“5”, principalmente nas idades 40, 45, 50 e 60 anos.Ou seja, há um aparente aumento inesperado no número de pessoas nessas idades, violando o pressuposto de distribuição uniforme entre os dígitos finais.Tais distorções são ocasionadas,provavelmente,por erros de declaração da idade, especificamente pela preferência digital. É importante salientar que o recurso de se formar grupos quinquenais poderia suavizar oproblema, mas é de grande relevânciaidentificar se a preferência digital realmente ocorre e em qual ou quais dígitos.

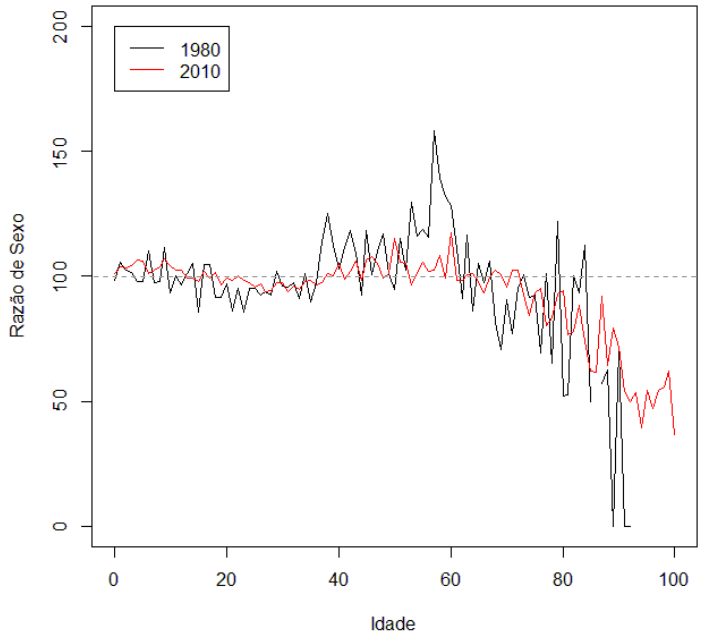
GRÁFICO 3 -Distribuição relativa por sexo e idade simples, Amapá, 1980-2010



Fonte: IBGE (1980;2010).

Outra maneira de detectar problemas na declaração da idade é por meio da análise da série de razão de sexo (Gráfico 4). Observa-se uma tendência de atração em certos dígitos, especialmente em 1980, devido a picos inesperados na série em torno de certasidades (menores que 20 e maiores que 40 anos).No ano de 2010, observa-se que a atração digitaldiminuiu, resultando em uma série de razão de sexo mais suave, fruto possivelmente de uma melhor qualidade na declaração de idade. Entretanto, como observado pela ONU (1955), outros fatores podem influenciar a razão de sexo, por exemplo, a seletividadena migração e na mortalidade, sobretudo em idades acima dos 40 anos– sendoque tais fatoresnão são estudados neste trabalho.

GRÁFICO 4 -Razão de sexo por idade simples, Amapá, 1980-2010



Fonte: IBGE (1980;2010).
Notas: Descontinuidades causadas pela ausência de homens ou mulheres na amostra do censo.

É importante ressaltar que a análise da pirâmide etária e da razão de sexo do estado do Amapá, ambas por idade simples, apenas apresenta um panorama geral sobre os erros de declaração de idade. Desta maneira, evidencia-se a necessidade de aplicar as técnicas de análise e detecção da qualidade na declaração de idade apresentadas na seção 4.

Para dar início à utilização dos algoritmos, é necessário fazer o download do pacote no site http://www.nepo.unicamp.br/publicacoes/textos_nepo/script/download.php, ou copiar as funções disponíveis nos Quadros 1 a 4.Os dados de entrada podem estar em qualquer formato, desde queseu conteúdo respeite os formatos apresentados nas Tabelas 5 e 6 e estejam localizados, juntamente com o pacote "GEQD.r", na mesma pasta de trabalho. Recomenda-se a utilização dos formatos ".csv" ou ".txt", para não ser necessária a utilização de pacotes de leitura de dados. Nesse trabalho, optou-se por usar um arquivo com extensão ".csv".Para dar inicío ao exemplo prático, foram utilizados os seguintes comandos:

```
setwd("C:\\Users\\Diretorio")#Aqui estão os dados e o pacote
source("GEQD.r") #Acessa o pacote para usar as funções implementadas
data<- read.csv("ExemploAP.csv",sep=";",header=T) #Importa os dados
em idade simples, para os índices de Myers e Whipple
data2 <- read.csv("ExemploAP2.csv",sep=";",header=T) #Importa os
dados em idade quinquenal e por sexo, para o método das Nações Unidas
```

Finalmente, cada índice pode ser chamado como uma função a partir dos seguintes comandos:

```
IM(data) #Retorna o Índice de Myers (equações 7 e 8)
IW(data) #Retorna o Índice de Whipple (equações 1 e 2)
IWg(data) #Retorna o Índice de Whipple Modificado Global (equação 4)
IN(data2) #Retorna o Age Accuracy Index (equação 11)
```

Após a execução dos comandos, os resultados podem ser analizados a partir de gráficos ou tabelas, no próprio *R* ou, por exemplo, sendo exportados para formatos como ".csv" eutilizados por programas como *Excel*.

Os resultados das funções IM, IW, IWg e IN, para os dados do Censo Demográfico do estado do Amapá em 1980 e 2010, são apresentados na Tabela 7. Segundo o apontado anteriormente, pela análise preliminar da pirâmide etária por idade simples e pela série da razão de sexo, as classificações dos resultados dos quatro métodos revelam que o nível geral da qualidade dos dados de idade para o estado do Amapá aumentou de forma significativade 1980 para 2010, independentemente do método analisado.

TABELA 7 -Índices e classificação para a verificação da qualidade da declaração de idade, Amapá, 1980-2010

Índice	1980		2010	
	Valor	Classificação	Valor	Classificação
Whipple 0 e 5	116,18	Aproximados	105,39	Pouco precisos
Whipple modificado global	0,71	Sem preferência	0,26	Sem preferência
Myers global	5,47	Mediano	2,35	Baixo
Método das Nações Unidas	28,46	Imprecisos	13,12	Precisos

Fonte: IBGE (1980;2010).

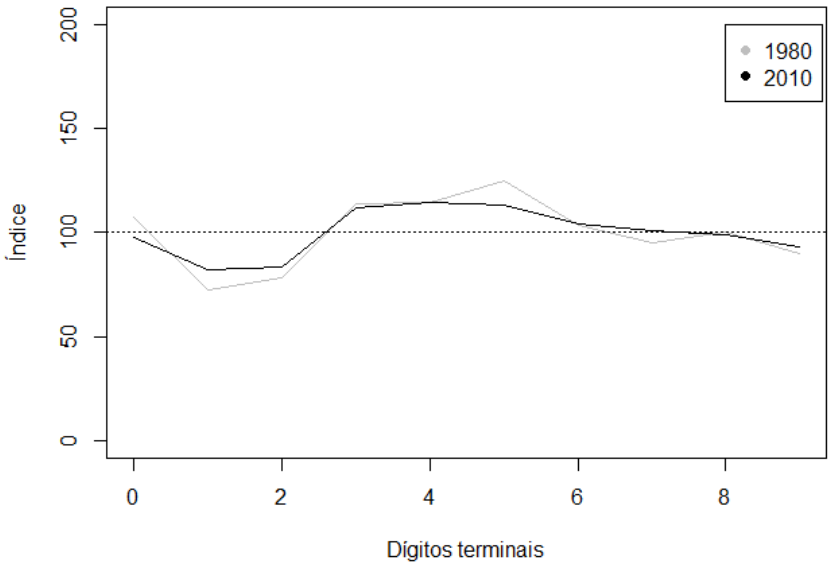
É importante explicitar que cada método tem uma amplitude de idade específica e uma escala que determina qual o nível de qualidade de dados existentes para aquela população, a partir de uma classificação a *posteriori*. Desta forma, não há possibilidade de uma comparação direta dos resultados de cada índice, embora os *scores* apontem para uma evolução generalizada de melhora na qualidade da declaração da idade.

Pode-se constatar que a distribuição etária da população do Amapá passou a ter grau de confiabilidade aceitável em 2010, de modo que, ao aplicar técnicas demográficas, não seria

necessário ajustes ou suavizações.Contudo, em 1980, foi possível constatar baixa qualidade da declaração de idade.Para contornar o problema de preferência digitalnos dados do Amapá de 1980, aconselha-se o ajuste ou a suavização: o uso de técnicas como o agrupamento das idades em grupos quinquenais, por exemplo, eliminaria a maioria das irregularidades devidas à preferência digital (PAES; ALBUQUERQUE, 1999, p. 37).

A partir doGráfico 5, é possível analisar a preferência por dígito pelo índice de Whipple (equação 2), de modo a identificar qual dígito final foi detectado como o preferencial em cada ano censitário. Em um cenário de ausência de atração de dígito final, esperia-se uma reta em 100, ou seja, não haveria violação do pressuposto de distribuição uniforme. Neste caso, observa-se preferência pelos dígitos“3”, “4” e “5” (em maior magnitude para o “5”), indicando dados aproximados nos três dígitos, caso utilizássemos a classificação da Tabela 1. Outro ponto é que em 2010 houve uma maior tendência linear do que em 1980, indicando melhora na qualidade da declaração da idade.

GRÁFICO 5 -*Whipple-type index* por dígito, Amapá, 1980-2010

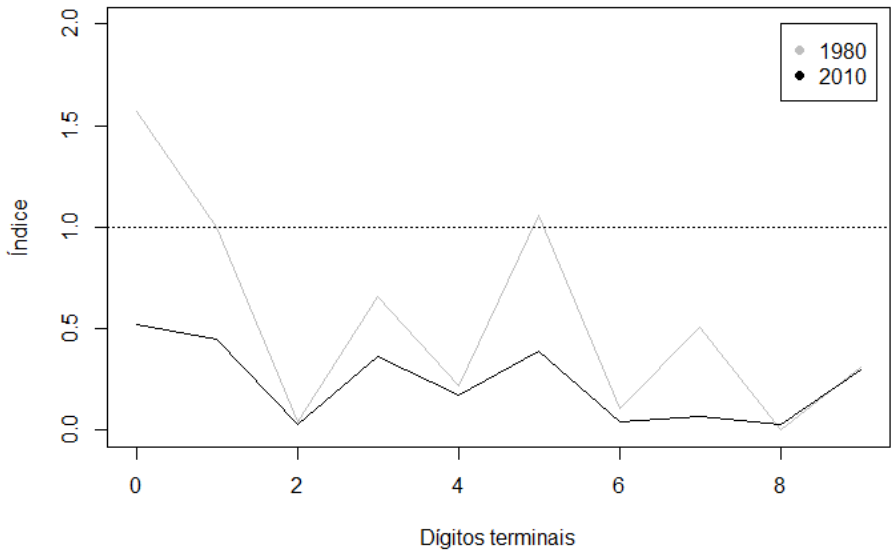


Fonte: IBGE (1980; 2010).

O Gráfico 6 pode ser utilizado para a mesma finalidade. Neste caso, espera-se uma reta próxima de 1, em um cenário de ausência de preferência por dígitos finais. É possível analisar que, em 1980, existia uma maior preferência pelos dígitos 0 e 5. A variação dos resultados encontrados nos Índices de Myers e Whipple para a atração de cada dígito finaldecorre da amplitude etária considerada em cada método. No caso do Índice de Whipple,são utilizadas as idades entre 23 e 62

anos; já no caso do Índice de Myers, as de 10 a 99 anos. Nesse sentido, o Índice de Myers considera grupo etários que são mais suscetíveis a erros de declaração de idade, como é o caso dos idosos.

GRÁFICO 6 -Índice de Myers por dígito, Amapá, 1980-2010



Fonte: IBGE. (1980;2010).

7 CONSIDERAÇÕES FINAIS

A idade é a principal variável demográfica,ponto de partida de diversos estudos, e a qualidade de sua declaração é pressuposto para a maior parte dos métodos demográficos. Com o constante avanço da qualidade dos dados demográficos brasileiros, a preocupação com a qualidade da declaração da idade foi sendo deixada de lado.Todavia,tal qualidade pode ser questionadaà medida que são necessários recortes específicos da população (como, por exemplo, estratificações sociais e/ou regionais).

Assim sendo, este trabalho buscou apresentar uma breve revisão bibliográfica dos principais métodos de análise da qualidade da declaração de idade e propor a construção de algorítimos, em linguagem *R*,que fomentem e facilitem a avaliação da qualidade de dados demográficos.Acredita-se também que os algoritmos sirvam como instrumentos que auxiliam e agilizam a estimação dos índices, possibilitando avaliar a qualidade da declaração da idade em várias localidades, recortes soioeconômicos e demográficos ao mesmo tempo.

Em razão dos diferentes métodos de cálculo para classificar a qualidade da declaração da idade, bem como as populações-alvo de cada técnica, a comparação entre os índices apresentados deve permanecerno âmbito teórico, analisando a potencialidade de cada método, seus dados de entrada e quando cada um pode ser aplicado. Neste sentido, nota-se que os métodos variam

basicamente em relação a dois aspectos: 1)dígitos analisados para verificar a atração; e 2)faixa etária utilizada emcada método de cálculo.

Cabe destacar que os índices de Myers e Whipple são utilizados para idades simples e podem também ser aplicados tanto para a população geral como por sexo, enquanto o Método das Nações Unidasapresenta-se como uma proposta de mensurar a atração de dígitos finais, em casos nos quaisse tem disponível apenas a distribuição etária por grupos quinquenais e sexo. Desta forma, nota-se que analisar a disponibilidade dos dadosé uma primeira regra de escolhado método; via de regra, é preferível ter em mãos dados por idade simples e sexo, o que permite um estudo mais acurado da qualidade da declaração de idade em um local e ano.

O Índice de Whipple, por exemplo, foi originalmente pensado para a detecção da preferência de idades terminadas em “0” e “5”, considerando somente o intervalo entre 23 e 62 anos.Todavia, é possível estender o método para o dígito “0” para todos os dígitos (fato que não foi proposto em sua construção inicial). Posteriormente, foi pensada uma versão “modificada”, proporcionando o cálculo de um índice global, não só considerando os dígitos “0” e “5”, como todos os outros. O Índice de Myers possui o maior intervalo etário no seu cálculo, contemplandoas idadesde 10 a 99 anos, e o método possibilita apresentar resultados para cada um dos dígitos finais e um índice global.Por fim, o MétododasNações Unidas(AAI) utiliza a distribuição etária por grupos quinquenais e por sexo, de 0-4 anos até 70 anos e mais, mas não possui um resultado acerca de cada dígito final.

Em relação à qualidade da informação da declaração da idade, a ONU (1955) afirma que o método AAI pode ser melhor que os índices de Myers e Whipple por levar em consideração: a preferência digital; a omissão de indivíduos no Censo (cobertura); e a falta de declaração de idade (não respostas). Além disso, ele é um importante recurso para estudar a declaração da idade em casos em que não se tem os volumes populacionais por idade simples. No caso dos índices de Whipple e de Myers, estes levam em consideração apenas a preferência digital, embora possamos obter estes índices, além da população geral, para homens e mulheres em separado.Entretanto, o AAI é um índice geral que não informa quais dígitos são preferidos e possui restrições quanto ao seu uso, não sendo aplicável tanto para populações nas quaishá grandes distorções na distribuição por sexo e idade (guerras, epidemias ou migrações, por exemplo), como em pequenas áreas.

Nesse sentido,a escolha do método a ser utilizado deve considerar: as características da população estudada, sobretudo no que se refereà distribuição etária; a disponibilidade de dados e sua especificidades;e a magnifitude e a necessidade de utilização dos idosos no cálculo. Conclui-se, portanto, que, paradefinir o melhor método, deve-se ponderarcada caso, e avaliar qual método melhor se adequa aos diversos dados de entrada e às classificaçãos.

A construção de algorítmos para os métodos apresentados, elaborados em linguagem *R*, pode ajudar na disseminação e na utilização destes métodos que, apesar de serem propostos na primeira metade do século XX, continuam relevantes até os dias de hoje. A proposta de um pacote de análise da qualidade da declaração da idade no *software R*pode incentivar que outros pacotes sejam criados ou que esta versão inicial de "GEQD.r" seja aprimorada a partir da colaboração de outros usuários.

Por fim, acredita-se que o trabalhopossa estimular a análise da qualidade da declaração da idade, a qual vem sendo deixada de lado em muitos trabalhos de análise demográfica.Especialmente em casos nos quais são utilizados recortes específicos – comopor renda e nível de instrução, dentre outros –, a aplicação de técnicas demográficas (diretas ou indiretas) demanda saber se as subpopulações em questão têm vieses na declaração de idade. Para além dos fatores elucidados, saber o grau de acurácia dosdados pode servir como uma ferramenta de análise comparativa entre subpopulações, de modo a constatar se há associação entre atração digital e variáveis socioeconômicas (como renda e escolaridade) ou espaciais (como situação do domicílio ou região de residência).

REFERÊNCIAS

FORMIGA, M. C. C.; RAMOS, P. C. F.; MONTEIRO, M. F. G. A qualidade dos dados censitários populacionais e sua associação com fatores socioeconômicos: um estudo para as mesorregiões do Estado do Rio Grande do Norte – Brasil. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS**,** 12., 2000, Caxambu, MG. **Anais…** Belo Horizonte, MG: ABEP, 2000.

HAKKERT, R. **Fontes de dados demográficos.** Belo Horizonte, MG: ABEP, 1996. 72p. (Série Textos Didáticos, n. 3).

IBGE. **Amostra do Censo Demográfico 2010.** Rio de Janeiro, RJ, 2010.

_____. **Amostra do Censo Demográfico 1980.** Rio de Janeiro, RJ, 1980.

MOULTRIE, T. et al.**Tools for Demographic Estimation.** Paris: IUSSP, 2013. Disponível em:<http://demographicestimation.iussp.org>.Acesso em:09jan. 2016.

MYERS, R. J. Errors and Bias in the Reporting of Ages in Census Data. **Transactions of the Actuarial Society of America,** Schaumburg, v. 41, part 2, n. 104, p. 395-415,1940.

NOUMBISSI, A. L'indice de Whipple modifié: une application aux données du Cameroun, de la Suède et de la Belgique. **Population,**Paris, v. 47, n. 4, p. 1038-1041, 1992.

ONU. **Demographic Yearbook 1988**. New York, NY: United Nations, 1990. 1319p.

_____.**Manuals on methods of estimating population:**manual x: indirect techniques for demographic estimation. New York, NY: Department of International Economic and Social Affairs – ONU, 1983.304p.

_____. **Manuals on methods of estimating population:**manual II:methods of appraisal of quality of basic data for population estimates. New York, NY: United Nations, 1955. 76p. (Population Studies, n. 23).

PACHECO, A. G. F.; CUNHA, G. M.; ANDREOZZI, V. L.**Aprendendo *R*.**Rio de Janeiro, RJ: Escola Nacional de Saúde Pública, 2014.

PAES, N. A; ALBUQUERQUE, M. E. E. Avaliação da qualidade dos dados populacionais e cobertura dos registros de óbitos para as regiões brasileiras. **Revista de Saúde Pública,**São Paulo, SP, v. 33, n. 1, p. 33-43,1999.

POSTON, D. L.; MICKLIN, M. **Handbook of population.** New York, NY: Kluwer Academic/Plenu Publishers, 2005.

ROGER, G.; WALTISPERGER, D.; CORBILLE-GUITTON, C. **Les structures par sexe et age en Afrique.** Paris: Groupe de Demographie Africaine (IDP-INED-INSEE-MICOOP-ORSTOM), 1981.

SHRYOCK, H. S.; SIEGEL, J. S. **The methods and materials of demography.**San Diego, A: Academic Press, 1976.

SILVA, B. F.; DINIZ, J.; BORTOLUZZI, M. A. **Minicurso de estatística básica:**introdução ao software *R*. Santa Maria, RS: Universidade Federal de Santa Maria, 2009.(Programa de Educação Tutorial).

SPOORENBERG, T. La qualité des déclarations par âge: extension et application de l'indice de Whipple modifié. **Population,** Paris, v. 62, n. 4, p. 847-859, 2007.