



**INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA**

**ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS**

**PROGRAMA DE PÓS-GRADUAÇÃO POPULAÇÃO, TERRITÓRIO E ESTATÍSTICAS PÚBLICAS**

**TESE DE DOUTORADO**

# Controle Estatístico de Confidencialidade em microdados de pesquisas amostrais domiciliares

**Bruno Freitas Cortez**

Rio de Janeiro, RJ
Fevereiro de 2023

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

# Controle Estatístico de Confidencialidade em microdados de pesquisas amostrais domiciliares

**Bruno Freitas Cortez**

**Tese**

Apresentada ao Programa de Pós-Graduação em População, Território e Estatísticas Públicas da Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística como requisito parcial para obtenção do título de

**Doutor em População, Território e Estatísticas Públicas**

Rio de Janeiro, RJ
Fevereiro de 2023

C828c Cortez, Bruno Freitas
Controle estatístico de confidencialidade em microdados de pesquisas amostrais domiciliares / Bruno Freitas Cortez. - Rio de Janeiro, 2023.
204 f.
Inclui referências, apêndice e anexo.
Orientador: Prof. Dr. Maysa Sacramento de Magalhães.

Tese (Doutorado em População, Território e Estatísticas Públicas) – Escola Nacional de Ciências Estatísticas.

1. Pesquisa por amostra de domicílios – Controle estatístico – Brasil - Teses. I. Magalhães, Maysa Sacramento de. II. Escola Nacional de Ciências Estatísticas. III. IBGE. IV. Título.

CDU: 314.6(81):519.248

# INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

# ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS

**Bruno Freitas Cortez**


## Controle Estatístico de Confidencialidade em microdados de pesquisas amostrais domiciliares

Tese apresentada ao Programa de Pós-Graduação em População, Território e Estatísticas Públicas da Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística, como requisito parcial para a obtenção do título de Doutor.

**Banca Examinadora:**

**Maysa Sacramento de Magalhães**
Orientador - ENCE/IBGE

**José André de Moura Brito**
ENCE/IBGE

**Maurício Teixeira Leite de Vasconcellos**
ENCE/IBGE

**Marcel de Toledo Vieira**
UFJF

**Thaís Paiva Galletti**
UFMG

**Rio de Janeiro, 27 de fevereiro de 2023**

# DEDICATÓRIA

*À minha filha Ingrid*

# AGRADECIMENTOS

À minha família, principalmente à minha mãe por todo seu esforço e incentivo ao estudo que formou minha base, e à minha esposa Juliana pelo companheirismo nestes últimos anos.

À minha orientadora Maysa, pela confiança e orientação recebida neste trabalho.

À banca examinadora desta tese, pelas contribuições para o aprimoramento deste trabalho.

Ao Instituto Brasileiro de Geografia e Estatística, meus chefes e coordenadores, atuais e antigos, pela oportunidade concedida.

Aos meus colegas de curso, especialmente ao pessoal do "grupo de *Whatsapp*" da turma que estava lá 24 horas por dia para auxiliar em quaisquer dúvidas que tivéssemos. Nas questões burocráticas, que eu muitas vezes esquecia, sempre me salvaram!

E, por fim, à Escola Nacional de Ciências Estatísticas pela estrutura, professores e funcionários, local onde eu também fiz minha graduação e mestrado, ou seja, minha segunda casa.

Muito Obrigado!

# RESUMO

## Controle Estatístico de Confidencialidade em microdados de pesquisas amostrais domiciliares

Bruno Freitas Cortez

Escola Nacional de Ciências Estatísticas, IBGE, 2023

Orientadora: Maysa Sacramento de Magalhães

Dados e informações são, cada vez mais, produzidos, demandados e valorizados na sociedade atual. Entretanto, quando se trata de estatísticas oficiais há que se considerar e garantir a confidencialidade das informações dos respondentes, afinal, a obtenção de dados confiáveis depende de sua cooperação e boa vontade. Desta forma, garantir a confiança e aceitação do público necessárias aos Institutos Nacionais de Estatística (INE) é crucial, e uma das formas para tal é assegurar que os respondentes não possam ser identificados a partir dos dados publicados. O campo de estudo de Controle Estatístico de Confidencialidade (CEC) foi desenvolvido justamente em resposta a desafios em relação à confidencialidade das informações que os profissionais dos INE enfrentavam na prática, resultado principalmente de inovações tecnológicas e novas formas de disseminação de dados. As técnicas de CEC visam tratar os dados para que possam ser divulgados sem revelar informações confidenciais neles contidas, ao mesmo tempo que tentam reduzir a perda de informação advinda deste processo. Assim, esta tese tem por objetivo propor uma abordagem para a implantação de métodos de CEC, para os microdados de uso público, em pesquisas amostrais domiciliares. A proposta engloba a seleção de cenários de revelação e variáveis-chave, a estimação do risco de revelação dos registros da pesquisa, a aplicação de métodos de mascaramento para os registros com risco acima de um limite considerado tolerável, a mensuração da perda de informação resultante da etapa de mascaramento, e explora as similaridades e sinergia com os processos de crítica e imputação. Para ilustrar o desenvolvimento da abordagem proposta, foi utilizada a PNAD Contínua, por ser a principal pesquisa amostral domiciliar do IBGE, que abrange todo o território nacional. Adicionalmente, foi desenvolvido um escore que leva em conta conjuntamente as medidas de risco e de utilidade das informações para auxiliar na decisão entre distintas configurações de mascaramento. Os resultados obtidos, a partir do que foi proposto e desenvolvido neste trabalho, indicam que técnicas de CEC são ferramentas fundamentais para garantir a confidencialidade das informações.

**Palavras-chave**: controle estatístico de confidencialidade, microdados, risco de revelação, pesquisas amostrais domiciliares.

# ABSTRACT

## Statistical Disclosure Control in microdata from household sample surveys

Bruno Freitas Cortez

Escola Nacional de Ciências Estatísticas, IBGE, 2023

Advisor: Maysa Sacramento de Magalhães

Data and information are increasingly produced, demanded, and valued in current society. However, when it comes to official statistics, it is necessary to consider and guarantee the confidentiality of respondents' information, after all, reliable data depends on their cooperation and goodwill. In this way, guaranteeing public trust and acceptance of the National Statistics Institutes (NSI) is crucial, and one of the ways to do this is to ensure that respondents cannot be identified from the published data. Statistical Disclosure Control (SDC) field of study was developed precisely in response to challenges regarding the confidentiality of information that NSI professionals faced in practice, mainly because of technological innovations and new forms of data dissemination. SDC techniques aim to treat the data so that it can be released without revealing confidential information contained therein, while trying to reduce the information the implementation of loss arising from this process. Thus, this thesis aims to propose an approach for the implementation of SDC methods, for public use microdata, in household sample surveys. The proposal encompasses the selection of disclosure scenarios and key variables, the disclosure risk estimation of the survey records, the application of masking methods for records with risk above a limit considered tolerable, the measurement the information loss resulting from the masking step, and explores the similarities and the synergy with the edit and imputation processes. To illustrate the development of the proposed approach, the Continuous National Household Sample Survey was used, as it is the main household sample survey by IBGE, covering the entire national territory. Additionally, a score was developed that jointly takes into account the measures of risk and information utility to assist in the decision between different masking configurations. The results obtained, based on what was proposed and developed in this work, indicate that CEC techniques are fundamental tools to guarantee the confidentiality of information.

**Keywords**: Statistical Disclosure Control. Microdata. Household sample surveys.

# SUMÁRIO

# LISTA DE GRÁFICOS

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE QUADROS

## LISTA DE ABREVIATURAS E SIGLAS

ABS: *Australian Bureau of Statistics*
ACS: *American Community Survey*
CadÚnico: Cadastro Único
CAGED: Cadastro Geral de Empregados e Desempregados
CANCEIS: *Canadian Census Edit and Imputation System*
CBS: *Centraal Bureau Voor De Statistiek*
CEC: Controle Estatístico de Confidencialidade
Censo Superior: Censo da Educação Superior
CoE SDC: *Centre of Excellence on Statistical Disclosure Control*
CPS: *Current Population Survey*
CTPS: Carteira de Trabalho da Previdência Social
DANE: *Departamento Administrativo Nacional de Estadística*
DESTATIS: *Statistisches Bundesamt*
DIEESE: Departamento Intersindical de Estatística e Estudos Socioeconômicos
DIS: *Data Intrusion Simulation*
ECINF: Economia Informal Urbana
ENAP: Escola Nacional de Administração Pública
ENCE: Escola Nacional de Ciências Estatísticas
G-Confid: *Disclosure Avoidance - Generalized System*
GSS: *Government Statistical Service*
IBGE: Instituto Brasileiro de Geografia e Estatística
INDEC: *Instituto Nacional de Estadística y Censos*
INE: Instituto Nacional de Estatística
INE, IP: Instituto Nacional de Estatística de Portugal
INEGI: *Instituto Nacional de Estadística y Geografía*
INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
INSEE: *Institut National de la Statistique et des Études Économiques*
ISTAT: Instituto Nazionale di Statistica
LAI: Lei de Acesso à Informação
MSU: *Minimal sample uniques*
NDC: *National Data Center*
NSI: *National Statistics Institutes*
ONS: *Office for National Statistics*
ONU: Organização das Nações Unidas
PED: Pesquisa do Emprego e Desemprego
PeNSE: Pesquisa Nacional de Saúde do Escolar
PINTEC: Pesquisa de Inovação Tecnológica
PME: Pesquisa Mensal de Emprego
PNAD: Pesquisa Nacional por Amostra de Domicílios
PNAD Contínua: Pesquisa Nacional por Amostra de Domicílios Contínua
PNS: Pesquisa Nacional de Saúde
POF: Pesquisa de Orçamentos Familiares
POSIC: Política de Segurança da Informação e Comunicações

PRAM: *Post-randomization Method*
RAIS: Relação Anual de Informações Sociais
R.CD-IBGE: Resolução do Conselho Diretor do IBGE
RIDE: Região Administrativa Integrada de Desenvolvimento
RM: Região Metropolitana
SAR: Sala de Acesso a Dados Restritos
SIC: Serviço de Informações ao Cidadão
SDC: *Statistical Disclosure Control*
SIDRA: Sistema IBGE de Recuperação Automática
SIPD: Sistema Integrado de Pesquisas Domiciliares
STATCAN: *Statistics Canada*
SUDA: *Special Uniques Detection Algorithm*
TI: tecnologia da informação
UF: Unidade da Federação
UNECE: Comissão Econômica das Nações Unidas para a Europa
UNESCO: Organização das Nações Unidas para a Educação, a Ciência e a Cultura
UPA: Unidade Primária de Amostragem

# LISTA DE SÍMBOLOS

$D_j$ ….. j-ésimo domicílio nos microdados da pesquisa.

$I_i$ ….. i-ésima pessoa/indivíduos nos microdados da pesquisa.

k ….. valor de chave

i ….. registro da amostra

i* ….. unidade da população de conhecimento prévio do intruso

$E_{cj}$ ….. função de entropia da j-ésima variável categórica que contém um conjunto de *c* categorias.

$f_{cj}$ ….. frequência da categoria $c_j$ da j-ésima variável

$f_k$ ….. número de registros na amostra com valor de chave *k*

$F_k$ ….. número de registros na população com valor de chave *k*

$\pi_k$ ….. probabilidade de uma unidade da população possuir o valor de chave *k*

$p_k$ ….. probabilidade de uma unidade da população com valor de chave *k* ser selecionada na amostra.

$\hat{p}_k^d$ …. estimador de $p_k$ com base no desenho amostral

$\hat{p}_k$…. forma mais simples de designar $\hat{p}_k^d$

$r_i$ ….. risco de revelação do i-ésimo registro na amostra

$r_k$ ….. risco de revelação de um registro na amostra com valor de chave *k*

$\tilde{r}$ …. mediana dos riscos de revelação individuais estimados

$r_{lo}$ …. l-ésima medida de risco calculada para os dados originais

$r_{lu}$ …. l-ésima medida de risco calculada para os dados mascarados pela u-ésima configuração proposta

$\lambda$ …. peso referente ao conjunto das medidas de risco do escore

$1-\lambda$ …. peso referente ao conjunto das medidas de utilidade do escore

s ..... função de escore que leva em conta medidas de risco de revelação e utilidade da informação.

$MAD(r)$ ..... desvio absoluto mediano dos riscos de revelação individuais estimados

$w_i$ ..... peso amostral do registro *i*

$\alpha_l$ ..... peso, no escore, da l-ésima medida de risco

$\beta_m$ ..... peso, no escore, da m-ésima medida de utilidade

# CAPÍTULO 1: INTRODUÇÃO

Dados e informações são, cada vez mais, produzidos, demandados e valorizados na sociedade atual. No fim da década de 2010, estimava-se que 2,5 quintilhões de bytes, ou 2,2 milhões de terabytes, eram gerados mundialmente por dia, de forma ininterrupta, por intermédio dos mais variados dispositivos eletrônicos e dos dados disponibilizados na Internet (SANTOS e KOWATA, 2018). Esta produção segue aumentando exponencialmente, dobrando a cada dois anos (CASTANHA, 2021). Nas empresas privadas muitas vezes suas bases de dados são classificadas contabilmente como ativos, na categoria ativos intangíveis (MACHADO e FAMÁ, 2011), e sua qualidade é fator de relevante vantagem competitiva (ROCHA, 2019).

Os dados também são vitais para a manutenção das funções de Estado, assim como são utilizados na sua interação com a sociedade como, por exemplo, para a: promoção da transparência, o controle social, a melhoria da eficiência dos serviços públicos, a implementação de políticas públicas. Adicionalmente, o Estado brasileiro é um grande produtor de dados e muitos dos quais são gerados nas várias fases do ciclo de políticas públicas: da elaboração até a avaliação. Contudo, tão importante quanto a produção é a disseminação destes dados. Neste sentido, ressalta-se que o governo brasileiro junto com outros sete países (África do Sul, Estados Unidos, Filipinas, Indonésia, México, Noruega e Reino Unido) e organizações da sociedade civil, anunciaram a criação da Parceria para Governo Aberto (*Open Government Partnership*) em 2011. Esta parceria, que hoje já conta com mais de 70 países, tem como um dos temas basilares a transparência, e os chamados “dados abertos” representam uma das dimensões fomentadas do “governo aberto” que visa a transformação da governança do século XXI (ENAP, 2017).

Entretanto, quando se trata de estatísticas oficiais também há que se considerar e garantir a privacidade das unidades fornecedoras da informação, sejam pessoas físicas ou jurídicas. Dados confiáveis dependem da boa vontade e cooperação dos respondentes, não importando se a participação na pesquisa é opcional ou obrigatória por lei. Assim, garantir a confiança e aceitação do público frente ao Instituto Nacional de Estatística (INE) é crucial, e uma das formas para tal é assegurar que os respondentes

não possam ser identificados a partir dos dados publicados (ONU, 2015). Desta forma os institutos tomam algumas medidas como, por exemplo, a remoção de quaisquer informações que possam levar a uma identificação direta do respondente (nomes, telefones, endereços, entre outros) nos dados disponibilizados. Contudo, estes procedimentos mais simples podem não ser suficientes e cada vez mais vem sendo empregado o uso de técnicas de Controle Estatístico de Confidencialidade (CEC).

O CEC é definido por Willenborg e De Waal (2001) como área de estudo relacionada com a modificação dos dados utilizados para fins estatísticos, que contêm informações confidenciais sobre alguma entidade individual (indivíduo, família, empresa, etc), a fim de evitar que terceiros utilizem esses dados para identificar ao menos uma destas entidades, revelando, assim, informações confidenciais sobre as mesmas. Hundepool *et al.* (2012), por sua vez, definem o CEC como um conjunto de métodos que visa mitigar o risco de revelação de indivíduos, empresas ou outras organizações. Este risco é minimizado para um determinado nível considerado aceitável, buscando maximizar a quantidade de informação divulgada. Benschop *et al*. (2019) utilizam elementos das duas definições anteriores ao afirmarem que o CEC procura tratar o dado para que o mesmo possa ser publicado ou divulgado sem revelar qualquer informação confidencial que ele possa conter, ao mesmo tempo que limita a perda de informação advinda do processo de anonimização do dado.

Em comum, estas definições apontam três elementos que serão discutidos neste trabalho, quais sejam: a necessidade, mais do que somente uma determinação legal e sim por demanda da sociedade, dos INE divulgarem informações; a obrigação destes produtores de informação respeitarem a confidencialidade das entidades respondentes; e como o CEC pode tentar conciliar estes dois objetivos. Antes, porém, deve-se pontuar por qual razão o CEC é uma área de estudo razoavelmente recente, despontando em fins do século passado (como será visto em mais detalhes no Capítulo 3), se a obrigação de divulgar informações e proteger a confidencialidade dos respondentes é algo inerente aos INE há muito mais tempo.

Elliot e Domingo-Ferrer (2018) argumentam que o campo de estudo do CEC não foi criação de um artigo revolucionário ou da invenção de alguma nova técnica matemática. Na verdade, seria uma área desenvolvida de forma lenta e gradual em

resposta a desafios que os profissionais dos INE enfrentavam na prática. Mais precisamente, seria o resultado de três mudanças técnico-sociais interrelacionadas: o avanço da informática que gerou a facilidade de processar grandes quantidades de dados, acompanhado do aumento da demanda da sociedade por eles; a possibilidade de processo e disseminação de dados detalhados em arquivos digitais; e a profusão do número de organizações detentoras de dados sobre unidades individuais. Concomitantemente estes fatores fazem com o que o número potencial de intrusos com recursos para utilizar indevidamente os bancos de dados aumente exponencialmente.

Hundepool *et al.* (2012) detalham também mudanças drásticas observadas no trabalho dos INE, geradas pelas novas possibilidades associadas ao desenvolvimento da tecnologia da informação ocorridas nas últimas décadas. Elencam, por exemplo, a transição da publicação em papel, que consistia basicamente em um conjunto de tabelas, limitadas em tamanho por um número de páginas pré-definido, para uma disponibilidade de informação muito mais detalhada e em maior volume em meio digital; a disponibilidade de computadores dotados de grande capacidade de processamento e *softwares* estatísticos que abarcam variadas técnicas das mais simples até as mais avançadas para a maioria dos pesquisadores; e a análise de dados agregados tabulados seja, cada vez mais, substituída pela análise dos dados individuais das unidades de pesquisa a partir de microdados.

Esta crescente demanda de microdados por pesquisadores é justificada, uma vez que para tecer conclusões sobre nossa sociedade, utilizando-se análises com base empírica, muitas vezes torna-se possível apenas ao se investigar dados com o maior detalhamento possível (TEMPL, 2017). Desta forma, há uma crescente pressão por parte do Estado e da sociedade para os INE divulgarem microdados que possam promover novas pesquisas, descobertas, assim como subsidiar políticas públicas. Neste sentido, muitos órgãos governamentais veem como parte de sua missão compartilhar dados detalhados no nível do registro individual (TAYLOR *et al.*, 2018).

Isso, no entanto, vem acompanhado de uma série de desafios legais, éticos e técnicos. Os princípios e regulamentos de proteção da confidencialidade impõem restrições ao acesso e uso de dados individuais. Não obstante, os produtores de estatísticas enfrentam o desafio de garantir a confidencialidade dos entrevistados ao

mesmo tempo em que devem tornar os arquivos de microdados acessíveis. Ressalta-se que estes produtores não são apenas obrigados a proteger a confidencialidade das informações recebidas, mas ela também é crucial para manter a confiança dos entrevistados pelos INE e garantir a fidedignidade de suas respostas (TEMPL, 2017).

Esta questão está contemplada nos Princípios Fundamentais das Estatísticas Oficiais, aprovados pela Assembleia Geral das Nações Unidas, em 2014. Ali se firmaram os dez princípios que refletem os valores com os quais as estatísticas devem ser produzidas e analisadas. Especificamente, o Princípio Fundamental de número 6, que trata da Confidencialidade das Informações, estabelece que: "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" (IBGE, 2018a).

No documento em questão, é enfatizado que existem várias dimensões do tema "confidencialidade" aos quais os INE devem levar em consideração. Por exemplo: a presença de um arcabouço jurídico com disposições claras estabelecidas em lei para garantir o sigilo estatístico; estabelecer sistemas eficientes de segurança em tecnologia da informação (TI) para garantir a confidencialidade dos dados contra ameaças externas; desenvolver uma política interna de confidencialidade abrangente que estabeleça princípios para uma cooperação com os respondentes, aliada a mecanismos de suporte que garantam que esta política funcione na prática; e, dentre outras, fornecer dados para terceiros – que é uma contribuição essencial dos INE – mas que estes sejam de forma anonimizada (ONU, 2015).

Trazendo este conceito multidimensional para o contexto do IBGE, observa-se que, no âmbito jurídico, a atividade do IBGE é regida pela Lei nº 5.534, de 14 de novembro de 1968, que dispõe sobre a obrigatoriedade da prestação de informações estatísticas. Seu artigo primeiro trata da obrigatoriedade de toda pessoa natural ou jurídica de direito público ou privado em prestar as informações solicitadas pelo Instituto. Contudo, o parágrafo único deste artigo estabelece a contrapartida desta obrigação:

> As informações prestadas terão caráter sigiloso, serão usadas exclusivamente para fins estatísticos, e não poderão ser objeto de certidão, nem, em hipótese alguma, servirão de prova em processo

> administrativo, fiscal ou judicial, excetuado apenas, no que resultar de infração a dispositivos desta lei. (IBGE, 2018a, p.13).

Adicionalmente, a Lei nº 5.878, de 11 de maio de 1973, dispõe sobre o IBGE e dá outras providências, em seu artigo 6º diz que:

> As informações necessárias ao Plano Geral de Informações Estatísticas e Geográficas serão prestadas obrigatoriamente pelas pessoas naturais e pelas pessoas jurídicas de direito público e privado e utilizadas exclusivamente para os fins que se destinam, não podendo servir de instrumento para qualquer procedimento fiscal ou legal contra os informantes, salvo para efeito do cumprimento da presente Lei. (IBGE, 2018a, p. 14).

É preciso destacar que o IBGE e todos os demais órgãos da administração pública também estão subordinados à Lei nº 12.527, de 18 de novembro de 2011, conhecida como Lei de Acesso à Informação (LAI), que regula o acesso a informações conforme previsto na Constituição Federal. O Instituto possui o Serviço de Informações ao Cidadão (SIC), previsto no artigo 9º desta Lei que atende e orienta o público quanto ao acesso a informações, informa sobre a tramitação de documentos nas suas respectivas unidades e protocola documentos e requerimentos de acesso a informações. Entretanto o sigilo referente a informações individuais é garantido no artigo 22 da referida Lei (IBGE, 2018a).

No âmbito da política interna, a Instituição lançou em 2013 a publicação denominada “Código de Boas Práticas das Estatísticas do IBGE”, com o intuito de promover a padronização da conduta na aplicação das “melhores práticas estatísticas, fundamentais para a credibilidade institucional e, portanto, para o reconhecimento e a confiança da sociedade nas informações que a Instituição produz” (IBGE, 2013, p. 4). A publicação elenca 17 princípios norteadores, divididos em três seções, quais sejam: ambiente institucional e coordenação; processos estatísticos; e produtos estatísticos. O princípio número 4, contido na primeira seção, é justamente o da confidencialidade estatística, segundo o qual “O IBGE deve garantir a proteção e a confidencialidade das informações individualizadas com as quais são produzidas as estatísticas oficiais” (IBGE, 2013, p. 21).

Sobre este princípio específico, o IBGE lançou no ano de 2018 a publicação “Confidencialidade no IBGE: Procedimentos adotados na preservação do sigilo das

informações individuais nas divulgações de resultados das operações estatísticas". Em consonância com os já mencionados Princípios Fundamentais das Estatísticas Oficiais e o Código de Boas Práticas das Estatísticas do IBGE, esta publicação detalha os mecanismos de suporte internos, como os comitês específicos para tratar das questões de confidencialidade e segurança das informações, assim como os procedimentos adotados para anonimização dos respondentes em microdados das diversas pesquisas da Instituição.

Com relação aos comitês mencionados, ao todo três deles foram criados, conforme publicação do IBGE (IBGE, 2018a). O mais antigo é o Comitê de Sigilo, criado em 2001. Suas principais atribuições são: propor soluções para questões ligadas à confidencialidade das informações de natureza estatística e geocientífica relativas ao IBGE; apreciar e dar o parecer sobre solicitações de acesso às informações confidenciais; e avaliar rotineiramente o cumprimento do princípio da confidencialidade estatística. O segundo, criado em 2003, é o Comitê de Avaliação de Acesso a Microdados não Desidentificados. Sua atribuição é avaliar os projetos de solicitação de microdados, em relação à finalidade e objetivo do projeto, assim como seu produto final, para decisão sobre o seu acesso. Por fim, o terceiro é o Comitê de Segurança da Informação e Comunicações do IBGE, criado em 2015. Tem como objetivo principal nortear as ações e investimentos necessários para implantar os mecanismos de proteção definidos pela Política de Segurança da Informação e Comunicações (POSIC) do IBGE.

Cabe destacar que a Presidência da República emitiu o Decreto nº 9.759, de 11 de abril de 2019, que posteriormente foi alterado pelo Decreto nº 9.812, de 30 de maio de 2019, que extinguiu e estabeleceu diretrizes, regras e limitações para colegiados da administração pública federal. Desta forma, os colegiados existentes no IBGE foram afetados, incluindo os comitês mencionados que foram revogados no ano de 2020. Entretanto, internamente nada mudou com relação às práticas e mecanismos de suporte relativos à questão da confidencialidade dos dados. Ademais, seguindo as normas dos Decretos nº 9.759 e nº 9.812, o Comitê de Sigilo foi criado novamente pela Resolução do Conselho Diretor do IBGE (R.CD-IBGE) n° 6/2021, de 11 de março de 2021. O Comitê de Avaliação de Acesso a Microdados não Desidentificados, por sua vez, foi recriado pela R.CD-IBGE n° 9/2021, de 18 de maio de 2021.

Todo este esforço institucional demostra a atenção que o IBGE dispensa tanto ao tema da confidencialidade no sentido mais amplo, quanto a um de seus aspectos essenciais que é a confidencialidade das informações individualizadas dos respondentes.

É importante frisar que tanto em disposições legais, quanto em normas internas no IBGE aqui citadas, está presente o termo "sigilo" que contém o mesmo significado do termo "confidencialidade" para a literatura de CEC. Em outras palavras, nesse sentido estrito, os termos podem ser considerados intercambiáveis. Isto será tratado com mais detalhes na Seção 2.1, mas resumidamente refere-se ao status concedido às informações sobre um respondente, pelo seu receptor e atual detentor.

Todavia, o ponto central da questão é que a segurança de que os dados prestados pelos informantes aos INE serão utilizados somente para fins estatísticos, influenciam tanto a decisão de responder a uma pesquisa, quanto a qualidade dessa resposta, fato contemplado em uma extensa literatura sobre este tema. Singer (1995), por exemplo, afirma que a garantia da confidencialidade afeta tanto a taxa de resposta, quanto a qualidade das respostas quando se trata de informações sensíveis. Em similar conclusão chegaram Couper *et al.* (2008) ao relacionar a redução de disposição de participar de uma pesquisa não necessariamente ao risco real de revelação de informação, e sim ao dano que a revelação pode causar. Couper *et al.* (2010) em pesquisa posterior confirmam que o possível dano causado pela revelação é o ponto central levado em conta pelos respondentes. Mas ponderam que se o risco de revelação isoladamente aparentemente não importa na decisão de participar de uma pesquisa, o mesmo não ocorre quando ele está associado a temas sensíveis, ou seja, a combinação de risco *versus* dano é relevante para o respondente. Com relação à qualidade das respostas, Connors *et al.* (2019), por exemplo, encontraram evidências que parte dos entrevistados alteram as informações que originalmente dariam em uma pesquisa, ao serem informados que estas poderão ser divulgadas de forma pública.

Cabe salientar, como já destacado neste capítulo, que o CEC é uma área de conhecimento que foi desenvolvida em resposta a desafios que os profissionais dos INE enfrentavam (ELLIOT e DOMINGO-FERRER, 2018). Entretanto, o debate sobre o tema atualmente deve ser mais amplo, pois a informação é considerada um ativo de grande

valor e cada vez mais empresas privadas possuem bases de dados individualizadas. Estas bases podem ser desde simples listas de clientes até arquivos com detalhadas informações e preferências pessoais, muito utilizada em redes sociais, por exemplo, como insumo para algoritmos que visem entregar determinada sugestão (produtos, serviços, etc) a cada indivíduo específico. Não é raro ver as informações desses arquivos, que deveriam ser confidenciais, serem divulgadas ou utilizadas de forma indevida.

Portanto, fica evidente a importância de garantir a confidencialidade das informações prestadas aos órgãos produtores de informação, contudo há que se levar em conta que o maior detalhamento dos dados disseminados aliado a um ferramental técnico-metodológico cada vez mais sofisticado disponível a um número crescente de usuários é uma realidade. Desta forma, torna-se necessário que os procedimentos de garantia de proteção dos dados individualizados também se sofistiquem, e daí ressalta-se a imprescindibilidade do estudo e utilização de técnicas de CEC nos órgãos produtores de estatísticas oficiais.

A utilização de técnicas mais robustas de CEC materializa-se em questões relacionadas à avaliação do risco de revelação em uma pesquisa, aos procedimentos de mascaramento de dados, ao cálculo da perda de informação, por exemplo, levando em consideração as particularidades de cada pesquisa.

A seguir, na Seção 1.1 apresenta-se a motivação, a justificativa para o que será desenvolvido na tese. Na Seção 1.2 são apresentados o objetivo geral e os objetivos específicos deste trabalho. E, por fim, na Seção 1.3 apresenta-se a descrição do que será tratado em cada capítulo da tese.

## 1.1. Motivação e justificativa

O IBGE, como órgão oficial de produção de estatísticas do Brasil, adota uma série de medidas para garantir a confidencialidade das informações nos microdados disseminados de suas pesquisas. Nas pesquisas amostrais domiciliares os dados são desidentificados, ou seja, eliminam-se variáveis de identificação direta dos informantes, como, por exemplo, nome e endereço. Além disso, a ordenação dos registros é feita de forma aleatória dentro da menor unidade de divulgação (Unidade da Federação,

município, etc) e a própria amostragem é técnica válida para a redução do risco de revelação de dados individuais. No Censo Demográfico, não é feita a divulgação em microdados do que se convencionou chamar de "resultados do universo"[1]. Neste caso, os dados disponibilizados referem-se ao agregado das variáveis investigadas no questionário básico do censo, por setor censitário[2]. Como medida adicional de proteção, nos setores com menos de cinco domicílios particulares permanentes apenas as variáveis estruturais tais como: subdivisões geográficas, número de domicílios e a população por sexo, são mantidas. Por fim, no caso das pesquisas econômicas, não são divulgados microdados, sendo a divulgação de tais informações feita apenas por tabelas e, ainda assim, estas passam por procedimentos para garantir a desidentificação de seus resultados, como a exigência de pelo menos três respondentes por célula da tabela, dentre outros (IBGE, 2018a).

Entretanto, não é calculada nenhuma medida formal para o risco de revelação nos microdados de uso público, seja dos registros individualizados, seja uma medida do risco global. Desta forma, não há como se mensurar se as medidas adotadas foram suficientes para garantir um determinado nível de proteção mínimo para todos os registros. Igualmente não é possível apontar quais registros, caso existam, apresentam um patamar de risco acima do considerado aceitável e, sendo assim, também não há como adotar métodos específicos de anonimização para contornar este problema.

Deve ser destacado também que os dados divulgados das pesquisas domiciliares possuem, na maioria das vezes, uma estrutura hierárquica, ou seja, registros de pessoas estão contidos e associados a um registro de domicílio. Duncan *et al.* (2011) afirmam que a informação do domicílio aumenta o risco de revelação dos registros dos moradores contidos nele, e que este incremento pode ser particularmente grande para domicílios com muitos moradores. Silva (2020) fez este tipo de cálculo para os microdados da amostra do Censo 2010, confirmando que a estrutura hierárquica tem

[1] Existem dois tipos de questionários no Censo: o "básico" e o "da amostra". Todas as perguntas do questionário básico estão contidas no questionário da amostra, que é mais amplo, de forma que essas variáveis comuns são investigadas censitariamente, ou seja, para todos os domicílios e pessoas. Isto permite que os registros dos dois tipos de questionários formem o conjunto Universo, ou seja, informações básicas para a totalidade da população recenseada.

[2] é a menor porção de área utilizadas pelo IBGE para planejar, coletar e disseminar os resultados dos Censos e Pesquisas Estatísticas.

caráter de ampliar o risco de revelação. Neste trabalho a autora, inclusive, recomendou que se utilizasse algum tipo de método que suprimisse a possibilidade de relacionamento entre domicílios e pessoas, na disseminação dos arquivos de uso público.

O cálculo dos riscos individuais e global é dependente de dois elementos que são assumidos: os cenários de revelação e as variáveis-chave. De forma resumida, o primeiro diz respeito a estabelecer quem tentará vincular um respondente a um registro específico e como isso será feito; já o segundo se refere a um conjunto de variáveis que, em combinação, pode resultar em identificação do respondente. Sendo assim, as particularidades de cada pesquisa devem ser levadas em conta na definição destes dois elementos.

Após o cálculo dos riscos individuais e global, pode ser que haja a necessidade de aplicação de métodos de mascaramento de dados, tendo em vista que estes métodos geram alguma perda de informação. Logo, busca-se minimizar o risco de revelação, minimizando também a perda de informação decorrente das possíveis alterações feitas nos dados originais. Assim, há que se calcular uma medida de perda da informação advinda dos métodos de mascaramento propostos. Como são possíveis várias abordagens distintas de mascaramento, privilegia-se aquela que, fixado um nível máximo de risco de revelação aceitável, traga a menor perda de informação.

Resta ainda uma questão prática a ser avaliada: postos os vários passos que integram o CEC, como se dará a sua incorporação no processo de produção da pesquisa? Do planejamento à disseminação do produto final, existem várias etapas que compõem uma pesquisa em um INE. Assim, ao se adicionar uma nova etapa, é preciso avaliar onde esta se encaixará no fluxograma atual dos processos de produção das pesquisas e possíveis sinergias entre eles, para otimizar a carga de trabalho adicional que será demandada e, também, possivelmente, a melhoria nos processos.

## 1.2. Objetivos

O objetivo geral desta tese é propor/ uma abordagem para a implantação de métodos de Controle Estatístico de Confidencialidade, em pesquisas amostrais

domiciliares. Para tanto, cenários de revelação de informações confidenciais e variáveis-chave para estes cenários serão propostos, para o cálculo das estimativas do risco de revelação. Nestas estimativas será considerada a estrutura hierárquica dos registros (domicílios / pessoas), levando em conta ainda seu efeito sobre os diferentes tamanhos de domicílios, em número de moradores.

Além disso, questões relativas aos procedimentos de mascaramento de dados e posterior cálculo da perda de informação decorrente deste processo serão abordadas. As etapas serão guiadas tendo como norte a parte prática, ou seja, visando como incorporá-las no processo de produção da pesquisa. Neste ponto, atenção especial se dará à sinergia existente entre os métodos de CEC e aos de crítica e imputação. A abordagem proposta será complementada com a criação de um escore que combine as informações de risco de revelação e perda de informação, com o intuito de otimizar esse *trade-off*.

Para ilustrar a abordagem proposta, escolheu-se trabalhar com os dados da PNAD Contínua. Esta é a principal pesquisa amostral domiciliar do IBGE, abrangendo todo o território nacional, e produz informações continuamente de vários temas sobre a população como, por exemplo, características demográficas, inserção no mercado de trabalho e educação. Os resultados são divulgados para diversos recortes geográficos, variando desde o total do país, até para os municípios das capitais em seu nível mais desagregado. A periodicidade da divulgação de suas informações, depende do indicador em questão, podendo ser mensal, trimestral, anual ou até mesmo variável.

**Objetivos específicos**

- Propor e apresentar procedimentos para a definição de cenários de revelação e variáveis-chave para os microdados de uso público.
- Comparar métodos para a avaliação de risco de revelação disponíveis na literatura e sua aplicabilidade na PNADC.
- Estimar os riscos de revelação individual para os registros da PNADC e o risco global da pesquisa, considerando a hierarquia domiciliar.
- Propor abordagens de anonimização para os registros com alto risco de revelação, visando minimizar a perda de informação.

- Avaliar a adequação das medidas de risco de revelação e utilidade dos dados, assim como as suas possíveis ponderações, para a elaboração do escore.
- Avaliar as implicações dos métodos de mascaramento dos dados no processo de crítica e imputação na pesquisa.

## 1.3. Descrição dos capítulos

Neste capítulo foram apresentados a introdução, com um histórico sobre a origem do Controle Estatístico de Confidencialidade, sua relação com os INE, e as práticas atuais do IBGE no que diz respeito à confidencialidade dos dados produzidos pela Instituição. Também se apresentou a motivação, a justificativa e a contribuição da tese, bem com os objetivos da mesma.

No Capítulo 2 são descritos e discutidos os principais conceitos sobre Controle Estatístico de Confidencialidade na disseminação de microdados.

O Capítulo 3 traz uma revisão da literatura tanto do tema Controle Estatístico de Confidencialidade aplicado a microdados, quanto dos procedimentos de CEC que os INE atualmente utilizam em seus microdados disseminados.

No Capítulo 4 são tratadas as questões relativas à fonte de dados escolhida na tese. É apresentada a Pesquisa Nacional por Amostra de Domicílios Contínua e justificada a sua escolha, assim como é descrito o que o IBGE faz atualmente para garantir a confidencialidade das informações individualizadas desta pesquisa.

No Capítulo 5 são apresentadas as questões ligadas ao risco de revelação. São propostos procedimentos para definição dos cenários de revelação e variáveis-chave para estes cenários. São estimados, então, os riscos individuais e globais para os cenários definidos, considerando-se a estrutura hierárquica e seu efeito diferencial em relação ao tamanho do domicílio. A estimação dos riscos é realizada pelo modelo probablístico originalmente proposto por Benedetti e Franconi (1998), coloquialmente chamado de “abordagem italiana”.

No Capítulo 6 são tratadas as questões relativas aos procedimentos de mascaramento de dados e posterior cálculo da perda de informação decorrente deste processo. Especial ênfase é dada para a combinação dos métodos de recodificação

global e supressão local para o mascaramento, e o uso de medida de entropia para perda de informação. É proposto um escore que combine as informações de risco de revelação e perda de informação, com o intuito de otimizar esse *trade-off*. Também, neste capítulo, se aborda a incorporação das etapas do CEC no processo de produção de uma pesquisa domiciliar, especialmente destacando a sua estreita ligação com os processos de crítica e imputação já existentes no IBGE.

O Capítulo 7, por fim, traz as considerações finais e possibilidade de trabalhos futuros.

# CAPÍTULO 2: PRINCIPAIS CONCEITOS DE CEC EM SUA APLICAÇÃO PARA MICRODADOS

Skinner (2009) argumenta que os INE normalmente asseguram aos respondentes de suas pesquisas, que as informações ali prestadas serão tratadas de forma confidencial. Este tipo de garantia se relaciona tanto ao modo de como estes dados serão utilizados dentro da instituição que está conduzindo a pesquisa, como também em relação aos produtos que venham a ser por ela disseminados. Para o primeiro caso é possível adotar medidas como, por exemplo, segurança em TI, protocolos internos para os funcionários, dentre outras. Com relação ao segundo caso, igualmente existem opções de abordagens a serem escolhidas. Neste sentido, Templ (2008) cita alguns exemplos, como o acesso remoto, tal que o pesquisador pode acessar os dados de uma conexão segura, fazer suas análises, mas não pode baixá-los. Outra é a chamada execução remota em que é geralmente necessária a geração de dados sintéticos. Neste caso, os pesquisadores podem construir e testar suas abordagens nos dados sintéticos e apenas na etapa final elas são aplicadas aos dados reais, com o resultado sendo inspecionado pelo INE para evitar algum tipo de revelação.

Estas opções, contudo, possuem limitações, tais como: questões de impossibilidade total ou parcial de acesso remoto; carga pesada de trabalho de pessoal para se gerar dados sintéticos ou para checar os diferentes resultados produzidos; presença de pessoal capacitado que conheça todos os métodos que foram aplicados aos dados para analisar os resultados com propriedade; impossibilidade de implementar determinados aplicativos ou métodos de análise no ambiente remoto; entre outras.

Desta forma, para se contornar estas soluções que demandam muitos recursos e tempo ou questões que possam esbarrar em restrições legais, uma opção viável é a utilização de técnicas de CEC nos microdados a serem disseminados. Com isto seria possível o pesquisador analisar os dados com seus próprios métodos utilizando os aplicativos de sua preferência (TEMPL, 2008). Skinner (2009) acrescenta que esta é uma solução interessante para os INE, uma vez que o CEC abarca um conjunto de importantes ferramentas para a divulgação de produtos estatísticos – também chamados na

literatura de *outputs* – seguros em termos da confidencialidade das respostas prestadas pelas unidades pesquisadas.

Os conceitos descritos neste capítulo têm como foco o CEC em sua aplicação a microdados, que é o escopo desta tese. Entretanto, cabe ressaltar que existem técnicas de CEC para todos os tipos de produtos disseminados que possam conter risco de revelação do informante. Duncan *et al.* (2011) citam, por exemplo, que mesmo na atual era da disseminação por meios eletrônicos, a divulgação de dados em forma de tabelas – que são produtos dos microdados – continua relevante nos INE. A diferença é que as publicações em papel estão sendo substituídas por ferramentas *online* nas quais o usuário pode gerar seu próprio plano tabular. Hundepool *et al.* (2012) acrescentam que estes conjuntos de tabelas podem conter, inclusive, casos mais complexos como as chamadas tabelas hierárquicas (que contém sub tabelas dentro dela) ou as vinculadas (que possuem células em comum com outras), que possivelmente tornem maior o risco de revelação. Há ainda outros tipos de produtos oriundos dos microdados que podem ser submetidos a técnicas de CEC. Griffiths *et al.* (2019), por exemplo, fizeram um extenso trabalho sobre quais os riscos e como proteger a confidencialidade de uma vasta gama de *outputs*, como gráficos do tipo caixa, histogramas, estatísticas descritivas, dentre muitos outros.

O intuito desta seção é, então, explicar de forma sucinta os conceitos referentes ao CEC, voltados para a questão dos microdados, para que eles possam ser melhor compreendidos na leitura do presente trabalho. Alguns deles serão explorados com mais detalhes em capítulos posteriores aos quais os temas em que eles estão inseridos serão abordados.

## 2.1. Privacidade, confidencialidade e sigilo

Segundo o dicionário Houaiss os adjetivos “confidencial” e “sigiloso” são considerados sinônimos (HOUAISS, 2008). Isto é um exemplo de que alguns conceitos utilizados em CEC e que, por conseguinte, serão utilizados nesta tese, devem ser apresentados de forma precisa para evitar possíveis ambiguidades. Não se propõe aqui levantar uma discussão filosófica ou mesmo mais aprofundada sobre eles. Ademais,

dependendo do campo de estudo em que são empregados, estes conceitos podem ter outros significados mais específicos. O objetivo desta seção é, então, apurar como a literatura de CEC majoritariamente os emprega e, a partir daí, uniformizar o que se entende por eles no contexto deste presente trabalho.

Duncan *et al*. (2011) considera a confidencialidade, no âmbito do CEC, como sendo um status concedido às informações sobre um respondente. Seria um compromisso assumido ao provedor da informação pelo receptor e atual detentor da informação. Hundepool *et al.* (2012) seguindo definição similar, apontam ainda que enquanto a confidencialidade é um conceito que diz respeito aos dados, a privacidade diz respeito às unidades respondentes. Argumentam que a violação da confidencialidade pode resultar na divulgação de dados que prejudiquem o respondente. Configurar-se-ia em um ataque à privacidade por ser uma intromissão na autonomia do respondente sobre a forma como seus dados são utilizados. Skinner (2009) corrobora estas definições quando estabelece que o termo "confidencialidade" se aplica às respostas fornecidas pela entidade pesquisada.

Um terceiro termo, muito utilizado pelo IBGE em suas publicações, é o de "sigilo". Entretanto, ele geralmente possui o mesmo significado do termo "confidencialidade" descrito anteriormente. O Código de Boas Práticas das Estatísticas da Instituição menciona, por exemplo, o "sigilo estatístico" (IBGE, 2013). De forma análoga à publicação que aborda os procedimentos de confidencialidade do IBGE, utiliza muitas vezes a expressão "sigilo das informações", descreve o "Comitê de sigilo", e assim por diante (IBGE, 2018a).

Esta tese utilizará os conceitos de confidencialidade e privacidade, com o mesmo significado da literatura, tal que o primeiro diz respeito aos dados e o segundo aos respondentes. Adicionalmente, ao se referir mais especificamente a questões do IBGE, pode se fazer necessário utilizar a expressão "sigilo", mas esta poderá ser considerada como um sinônimo, ou tradução alternativa de "confidencialidade".

## 2.2. Microdados de uso público e de acesso restrito

A presente tese possui foco nos chamados microdados de uso público, isto é, aqueles que um INE divulga sem restrições em sua página na Internet ou em outros meios. Exatamente por isso, neste tipo de arquivo, é essencial garantir a confidencialidade das informações prestadas pelos respondentes.

Entretanto, deve-se destacar que os INE também podem disponibilizar arquivos de microdados com acesso restrito. A informação contida neles geralmente é mais vasta em relação aos de uso público. Pode haver um maior número de variáveis identificadoras, assim como determinadas variáveis que são disseminadas publicamente com algum método de anonimização, e aqui se encontrarem inalteradas. Devido ao maior risco de revelação, seu acesso por pesquisadores é condicionado, geralmente, à assinatura de algum termo para sua utilização apenas visando fins estatísticos (DE WAAL e WILLENBORG, 1996). Adicionalmente, na maioria dos casos, o acesso se dá em ambiente seguro de informática controlado pelo INE (ELLIOT e DOMINGO FERRER, 2018).

Neste sentido, o IBGE dispõe de uma Sala de Acesso a Dados Restritos (SAR), com vistas a atender a demanda por informações individualizadas, que é utilizada principalmente pela comunidade científica e por analistas de políticas públicas. Mediante solicitação do usuário externo, seu projeto de pesquisa é avaliado pelo Comitê de Avaliação de Acesso a Microdados não Desidentificados. Caso aprovado, o requerente assina um termo de compromisso de que preservará o sigilo das informações estatísticas ao acessar os microdados. O acesso aos dados não desidentificados são efetuados em ambientes especialmente criadas no IBGE para esta finalidade, homologadas pela Diretoria de Informática quanto ao aspecto de segurança. Adicionalmente, os arquivos gerados pelo usuário são liberados somente após a verificação, em relação à preservação de sua confidencialidade, pela área responsável pela produção dos microdados não desidentificados (IBGE, 2018a).

## 2.3. Intruso

Intruso pode ser definido como alguém que tentará, a partir dos dados disseminados, descobrir informação que ele previamente desconhecia de um ou mais respondentes. Willenborg e De Waal (2001) afirmam que o intruso pode ser desde uma pessoa até uma organização com as intenções das mais variadas possíveis como, por exemplo, obter algum tipo de ganho com a informação ou tentar provar que a segurança dos dados é falha. É importante frisar que enquanto um pesquisador utiliza o banco de dados para fazer inferências sobre um grupo grande de unidades de pesquisa ou traçar seu perfil "médio", o intruso tem o foco na unidade isoladamente, seja ela apenas uma em particular ou o máximo de unidades que conseguir identificar.

## 2.4. Tipos de variáveis

As variáveis presentes em um banco de dados possuem uma classificação própria sob o ponto de vista do CEC. Algumas podem ter um conteúdo que o respondente não gostaria que fosse revelado publicamente, outras podem trazer informações relevantes que permitam identificá-lo e assim por diante. Esta seção apresenta uma descrição de como a literatura classifica as variáveis no âmbito do CEC, exemplificando-as

### 2.4.1. Variáveis identificadoras e sensíveis

De Waal e Willenborg (1996) definem variável identificadora como sendo aquela em que sozinha, ou em combinação com outras variáveis pode ser usada para identificar algumas entidades, por um intruso. Exemplos mais usuais são: alguma variável geográfica, sexo, nacionalidade, idade, ocupação, grau de instrução, entre outras. Definem ainda um subconjunto específico como sendo identificadoras diretas, tais como: nome, telefone, endereço, número de documentos, etc. Estas últimas devem ser removidas de quaisquer bancos de dados antes que ele seja disseminado para o público, pois, de outra forma, a identificação da entidade seria muito simples. Esta será a

terminologia adotada nesta tese, mas deve ser ressaltado que na literatura alguns autores podem usar nomes idênticos para coisas diferentes. Por exemplo, Elliot e Domingo-Ferrer (2018) chamam de variáveis “identificadoras” o que foi definido aqui como “identificadoras diretas”, e de “quase identificadoras” as aqui definidas como “identificadoras”.

As variáveis sensíveis são definidas por De Waal e Willenborg (1996) como aquelas que representam algum tipo de característica que um respondente não gostaria que fosse revelado sobre ele. Os autores argumentam que determinar quais variáveis são identificadoras ou quais são sensíveis é uma questão a ser resolvida com bom senso, geralmente pelos produtores dos dados, uma vez que não há uma regra exata para determiná-las. Inclusive questões culturais de cada país podem influenciar essa decisão. Como exemplo, é apontado para o fato que na Holanda a variável de rendimento seria considerada sensível, mas já na Suécia não. Para efeitos práticos, os autores apontam que no Statistics Netherlands, todas as variáveis não classificadas como identificadoras são tratadas, a princípio, como sensíveis, mas sabendo-se que algumas são mais sensíveis do que outras. Ressaltam ainda que nada impede que uma variável possa ser considerada ao mesmo tempo identificadora e sensível.

### 2.4.2. Variável-chave: o conceito de chave e valor da chave

Templ (2017) define variáveis-chave como um conjunto de variáveis que, se usadas em combinação, podem ser utilizadas para identificar respondentes de um banco de dados. Esta definição é muito parecida com a das variáveis “identificadoras” de De Waal e Willenborg (1996) e das “quase identificadoras” de Elliot e Domingo-Ferrer (2018), também chamadas de “atributos chave” por estes últimos. Para pontuar melhor a diferença deste termo entre os autores, é preciso introduzir o conceito de chave e valor de chave.

Chave pode ser definida como uma combinação de variáveis identificadoras. Assim, as variáveis identificadoras que constituem uma chave seriam também chamadas

de variáveis-chave. Um valor de chave seria, então, os valores assumidos pelas variáveis que constituem a chave (DE WAAL e WILLENBORG, 1996). A título de ilustração, é possível supor um cenário de revelação em que uma chave pudesse ser definida pelas variáveis “idade em anos completos”, “sexo”, “grau de instrução” e “ocupação”. Estas seriam as variáveis-chave. Um possível valor da chave poderia ser 50 (a idade), e os valores das categorias referentes a sexo feminino, nível superior completo e ocupação de “contador”.

### 2.4.3. Variáveis de peso e de plano amostral

A amostragem tem efeito sobre o risco de revelação e deve ser levada em consideração em sua estimativa. Isto ocorre porque ela introduz uma incerteza adicional uma vez que geralmente não se sabe se uma entidade em particular participou da pesquisa (TEMPL, 2017). Em outras palavras, um valor de chave único na amostra não é necessariamente único na população, de tal modo que, num primeiro momento, o intruso não terá certeza se fez a identificação correta. No entanto, para os usuários utilizarem corretamente os dados disseminados das pesquisas amostrais, os INE precisam divulgar os pesos amostrais de cada registro, assim como informações sobre o plano amostral, tais como conglomerados ou estratos, dependendo do tipo de amostragem realizada. Contudo, é preciso tomar alguns cuidados, pois estas informações podem também ser utilizadas para uma tentativa de revelação.

No caso dos pesos amostrais, por exemplo, não é raro que determinados domínios tenham distintas frações amostrais em uma pesquisa. Assim, essas subpopulações correspondentes podem ser reconhecidas em função de seus diferentes pesos amostrais. Idealmente os pesos devem ser disponibilizados apenas quando eles não forneçam informação adicional que possa ser utilizada para auxílio em uma tentativa de revelação. Caso isto não ocorra medidas podem ser tomadas como, por exemplo, subamostrar unidades com pesos baixos de forma que fiquem com pesos

semelhantes aos demais (DE WAAL e WILLENBORG, 1996), ou mesmo a utilização de métodos de mascaramento nos pesos amostrais (TEMPL, 2017).

Raciocínio análogo deve ser feito para variáveis relativas ao plano amostral. É preciso levar em conta se variáveis como os conglomerados ou os estratos revelam informações adicionais sobre as entidades dos domínios utilizados para tais fins. Templ (2017) argumenta ainda que a variável de peso deve ser analisada conjuntamente com as demais do plano amostral. Ilustra essa afirmação supondo que um método de anonimização utilizado numa variável de estrato seria inútil se, por exemplo, os pesos dentro dos estratos fossem iguais. Neste caso, não seria difícil um intruso recompor a informação anonimizada.

## 2.5. Tipos de riscos de revelação

Revelação é definida, segundo Hundepool *et al.* (2012), quando alguma entidade (pessoa, organização, etc) reconhece ou descobre algo que até então não sabia sobre outra entidade, por meio de dados disseminados. A literatura geralmente divide o risco de revelação em dois tipos: revelação de identidade ou de atributo.

A revelação de identidade ocorre quando um intruso consegue determinar uma relação de um para um entre um registro nos microdados e uma entidade alvo (de seu conhecimento próprio ou que ele tenha informação externa) com um determinado grau suficiente de confiança (DE WAAL e WILLENBORG, 1996). Este risco existe mesmo após a exclusão das variáveis identificadoras diretas do banco de dados a ser disseminado. Isto ocorre porque determinadas variáveis podem ter valores extremos como, por exemplo, a idade ou a renda mensal que em conjunto com alguma outra (identificação geográfica, por exemplo) ou até mesmo sozinhas facilitem bastante a identificação. Uma variável categórica que apresente uma frequência muito rara em uma categoria poderia ter o mesmo efeito, como a ocupação de “senador” ou, no limite, “presidente da república”. Outra possibilidade é uma combinação rara de respostas a certas variáveis identificadoras perguntadas. Não é difícil imaginar, por exemplo, que uma pessoa com

exatos 53 anos, de nacionalidade ucraniana, com a ocupação de “arquiteto”, habitando no município do Rio de Janeiro e com determinada estrutura domiciliar (número de moradores, sexo e idade deles, relação de parentesco, etc) que seja única na amostra da pesquisa também possa ser única na população.

Elliot e Domingo-Ferrer (2018) definem a revelação de atributo como sendo a que ocorre quando um intruso consegue estimar o valor de determinada variável de uma entidade com base nos dados disseminados. Os autores apontam que uma parte da literatura faz uma distinção entre o que seria revelação por atributo e inferencial, tal que a primeira seria uma estimação com absoluta certeza e a outra dada um nível de confiança determinado. Entretanto, os autores ponderam que esta seria uma distinção falsa, pois o processo de tentativa de revelação sempre estará sujeito a incertezas como, por exemplo, erros na própria coleta dos dados da pesquisa.

Templ (2017) é um autor que faz a diferença entre estes dois tipos de revelação, apresentando exemplos práticos. Para a revelação de atributo imagina um hospital disseminando dados tal que todas as pacientes do sexo feminino entre 56 e 60 anos tivessem câncer. Desta forma o intruso saberia, mesmo sem conseguir identificar nenhum registro específico, a condição médica de todas as mulheres daquele grupo etário internadas no hospital. Para a revelação inferencial, o autor supõe um modelo estatístico de alto poder de predição para uma variável sensível, que o intruso poderia obter a partir dos atributos disponíveis no dado disseminado. Retomando o argumento de Elliot e Domingo-Ferrer (2018), seria possível sugerir que no primeiro caso os dados disponibilizados pelo hospital poderiam conter erros. Estes poderiam estar presentes tanto na variável sexo, idade, ou condição médica, poderia também haver omissão de registros, dentre outros. Desta forma, seria impossível garantir a certeza na revelação do atributo em ambos os casos.

## 2.6. Risco individual e global

Existem duas abordagens possíveis para se calcular o risco de revelação em um conjunto de dados: o cálculo do risco para cada registro individualmente, ou obter uma medida de risco para o arquivo como um todo. Este segundo tipo de risco poderia ser definido como a probabilidade de qualquer registro individual ser identificado, ou por algum tipo de agregação dos riscos individuais, levando em conta a média ou o máximo destes valores, por exemplo (WILLENBORG e DE WAAL, 2001).

De forma resumida, é possível dizer que os riscos individuais podem ser usados para apontar quais registros não são seguros para disseminação e, por isso, necessitam de métodos de CEC voltados a uma unidade específica como, por exemplo, a supressão local. Por outro lado, o risco global pode ser usado para saber se o arquivo como um todo seria considerado seguro e, caso contrário, seria preciso aplicar regras que atinjam o arquivo inteiro como, por exemplo, a recodificação global (WILLENBORG e DE WAAL, 2001). Este assunto será abordado com maior detalhamento no Capítulo 5.

## 2.7. Cenários de revelação

Willenborg e De Waal (2001) apontam que o risco de revelação de um microdado não depende apenas de seu conteúdo. Se isto fosse verdade, o risco não iria variar caso esses dados fossem analisados por um potencial intruso com informações disponíveis sobre a população alvo pesquisada, ou por um outro sem nenhum tipo de conhecimento ao seu dispor. Desta forma, para estimar o risco de revelação de determinado microdado, é preciso levar em consideração também o intruso e o tipo de informação que ele já tenha previamente. É claro que na prática não há como se obter esta informação exata, entretanto, em muitos casos, é possível fazer suposições razoáveis. Os autores definem, então, o cenário de revelação como sendo o modelo sobre quais informações o intruso supostamente tem à sua disposição e como ele irá utilizá-las para obter informação extra de um determinado conjunto de dados.

Hundepool *et al.* (2012) argumentam que, em relação aos microdados, definir mais de um cenário geralmente é indicado, uma vez que diferentes fontes de informação podem estar ao alcance do intruso. É preciso checar, de acordo com a pesquisa em questão, que variáveis estão disponíveis livremente como, por exemplo, em registros administrativos ou cadastros públicos ao nível individual da unidade de pesquisa, ao qual o intruso possa fazer uso para uma tentativa de revelação. Após estabelecida a lista de variáveis que potencialmente seriam usadas como chave, um outro importante fator a se definir seria a estratégia de procura empregada pelo intruso. Esta pode variar desde um simples reconhecimento espontâneo, ou seja, uma revelação não intencional por parte de um usuário dos dados, até uma tentativa de um intruso tentar a revelação utilizando técnicas estatísticas mais sofisticadas como, por exemplo, de pareamento. Este assunto será abordado com maior detalhamento no Capítulo 5.

## 2.8. Anonimidade e anonimização

Anonimização, no contexto de CEC, pode ser entendido como um processo que faz com que seja praticamente impossível a identificação de um respondente em uma base de dados. Em outras palavras, um processo que torna insignificante o risco de revelação de identidade (DUNCAN et al, 2011). Templ (2017), embora não defina formalmente este conceito, o emprega com o mesmo sentido. Para estes autores, a anonimidade dos registros, resultante deste processo de anonimização seria, então, o que se deve buscar ao se aplicar as técnicas de CEC.

Hundepool *et al.* (2012), por outro lado, consideram um registo anonimizado aquele em que seus identificadores diretos foram removidos, sendo dados anonimizados os que contenham apenas registros deste tipo. Exatamente por isso, os autores argumentam que a anonimidade, por si só, não seria garantia de confidencialidade, dado que uma particular configuração de outros atributos – variáveis indicadoras – poderia identificar o respondente tal qual uma impressão digital.

O IBGE utiliza este último sentido para este conceito. Em sua publicação sobre as questões de confidencialidade na Instituição, argumenta que não basta disseminar arquivos anônimos e, por isso, apresenta outros procedimentos adotados na divulgação de dados de suas pesquisas (IBGE, 2018a). De fato, como a Instituição ainda não faz o cálculo de riscos de revelação dos registros contidos nos microdados, seria inviável empregar o termo "anonimização" com o sentido utilizado por Duncan *et al.* (2011) e Templ (2017). Entretanto, devido ao escopo desta tese, anonimização aqui será entendido como o processo que visa minimizar o risco de revelação, tal como conceituaram Duncan *et al.* (2011).

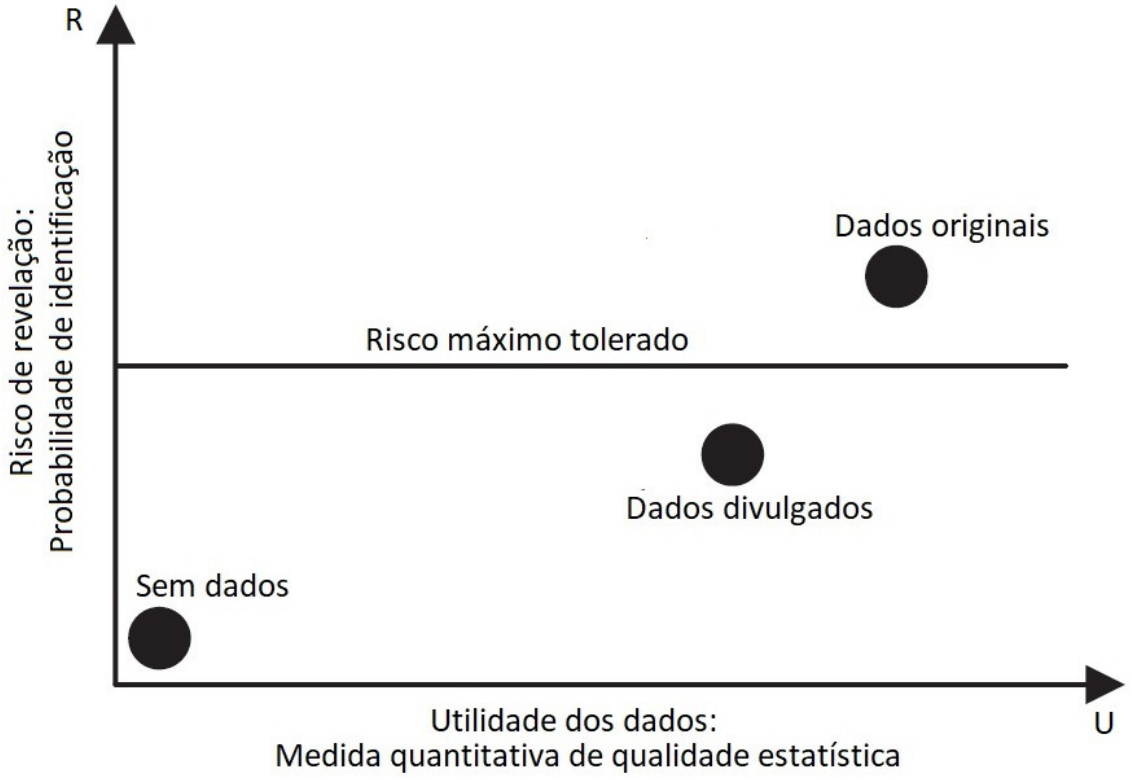
## 2.9. Risco de revelação *versus* utilidade da informação

Os métodos de CEC devem ser empregados de forma que tentem otimizar a relação entre risco de revelação e utilidade da informação. Assim, enquanto o risco deve ser mantido abaixo de um limite máximo aceitável, seja por lei ou por boas práticas, a utilidade deve estar acima de um limite mínimo admissível para os usuários do dado (ELLIOT e DOMINGO FERRER, 2018). É importante ter em mente que remover as variáveis identificadoras diretas de um microdado ou a realização de posteriores métodos de anonimização no mesmo, embora reduzam o seu risco de revelação, não o tornam nulo. A única forma de se obter risco nulo é a não divulgação da informação e, exatamente por isso, esta é uma opção válida. O IBGE, por exemplo, não divulga microdados de suas pesquisas econômicas, como já mencionado no Capítulo 1. Entretanto, se o usuário não pode ter a informação, a sua utilidade para ele também será nula. Por outro lado, se os dados originais fossem divulgados sem nenhum tipo intervenção, o risco de revelação poderia ser muito alto.

Para ilustrar o processo de otimização entre risco e utilidade, Duncan *et al.* (2001) desenvolveram o chamado mapa de confidencialidade R-U, tal que R é uma medida quantitativa do risco de revelação e U uma medida quantitativa de utilidade, como mostra a Ilustração 2.1. Na mesma estão representados os dados originais da

pesquisa que possuem grande utilidade, mas que podem ter um grande risco de revelação associado, em contraposição a não disseminação dos dados que não apresenta riscos, porém também sem utilidade. Hundepool *et al.* (2012) afirmam que os INE devem definir um risco máximo tolerável, a partir de seus padrões, políticas e diretrizes, com o intuito de manter a utilidade dos dados que vierem a ser divulgados, mas que estejam abaixo deste limite de risco.

Ilustração 2.1: - Mapa de confidencialidade R-U.




Fontes: Adaptado de Duncan *et al*. (2011).

## 2.10. Métodos de mascaramento de dados

Após o cálculo dos riscos de revelação, seja individual ou global, caso pelo menos um valor ultrapasse o limite máximo aceitável que foi definido, é preciso tomar providências para contornar este problema. Neste sentido, há uma série de métodos de mascaramento de dados, que possibilitam a redução do risco de revelação, ao passo que se tente alterar o mínimo possível os dados originais.

Estes métodos diferem dependendo se a variável em questão é categórica ou contínua. Além disso, os métodos podem ser divididos em perturbativos e não perturbativos. Existe também a hipótese de não disseminar os dados originais, substituindo-os por dados sintéticos que preservem determinadas características suas. Este assunto será tratado com mais detalhe no capítulo 6.

### 2.10.1. Métodos de mascaramento não perturbativos

Estes métodos partem do pressuposto de não distorcer os dados originais, mais especificamente, são baseados em supressões parciais ou reduções de detalhes. Alguns métodos podem ser usados tanto em variáveis contínuas ou categóricas, entretanto alguns são adequados apenas para variáveis categóricas (HUNDEPOOL *et a*l, 2012). A seguir estão listados os principais métodos disponíveis na literatura, com uma breve descrição deles.

- Amostragem: em vez de publicar os microdados originais, publica-se uma amostra de seu conjunto de registros. É um método adequado para variáveis categóricas, mas pode não ser útil para variáveis contínuas. Isto ocorre já que possíveis valores extremos dessa variável podem continuar no banco de dados e, mesmo com uma fração amostral maior, ainda haja grande possibilidade de revelação (HUNDEPOOL *et al*, 2012).
- Recodificação global: ocorre uma recodificação na variável em todo o arquivo de dados. Para uma variável categórica, duas ou mais categorias são combinadas para formar uma outra nova e menos específica. Esta técnica pode ser usada também em variáveis contínuas, tal que seus valores sejam substituídos por um intervalo. Neste caso a variável passaria a ser discreta ou ordinal (ELLIOT e DOMINGO FERRER, 2018).
- Recodificação de topo e fundo: pode ser considerado um caso especial de recodificação global a ser usada em variáveis passíveis de ordenação (contínuas, discretas ou ordinais). A ideia é que valores acima ou abaixo de limites especificados, sejam agregados formando uma nova categoria (HUNDEPOOL *et al*, 2012). É uma técnica adequada para a remoção de valores extremos.

- Supressão local: quando uma combinação de valores de chave está presente em pouquíssimos registros, ela é chamada de uma combinação não segura, porque pode levar a identificação do respondente. Assim, um ou mais valores de um registro individual são substituídos por valores faltantes (*missing*) para eliminar esta combinação não segura (HUNDEPOOL *et al*, 2012). É uma técnica muito utilizada após a recodificação global e de topo/fundo para lidar com casos pontuais restantes (TEMPL, 2017).

### 2.10.2. Métodos de mascaramento perturbativos

Estes métodos visam substituir valores originais dos microdados por outros com algum tipo de perturbação. A maior parte destes métodos foram desenvolvidos para variáveis contínuas, mas existem alguns pensados para variáveis categóricas e outros que podem ser utilizados em ambos os casos (HUNDEPOOL *et al*, 2012). A seguir estão listados os principais métodos disponíveis na literatura, com uma breve descrição deles.

- Adição de ruído: utilizado apenas em dados numéricos[3]. O método mais utilizado consiste em adicionar, em cada registro da variável, um ruído com distribuição $N(0,\alpha\Sigma)$, tal que $\Sigma$ corresponde à matriz de covariância dos dados originais. Médias e correlações dos dados originais podem ser preservados, nos dados mascarados, com a escolha apropriada do valor $\alpha$. Transformações lineares adicionais nos dados mascarados podem ser feitas para assegurar que a matriz de covariância de amostra das variáveis mascaradas seja um estimador não viesado para a matriz de covariância das variáveis originais (ELLIOT e DOMINGO FERRER, 2018). Na literatura, ainda constam métodos de adição de ruídos sem correlação, ou seja, ruído branco, assim como com adição de ruídos com posterior transformações não lineares. Este último teria a vantagem de ser aplicado em variáveis discretas e suas distribuições univariadas seriam preservadas, mas na prática ele é muito pouco utilizado (HUNDEPOOL *et al*, 2012).

[3] Hundepool e De Wolf (2012) acrescentam que existe uma exceção para esta regra. Seria um método não muito prático desenvolvido por Sullivan (1989) para adição de ruído em variáveis categóricas.

- Ruído multiplicativo: a ideia é análoga ao método anteriormente apresentado, mas se multiplica o ruído, ao invés de sua adição. Este procedimento tentaria contornar um problema do ruído adicionado que é a sua variância constante, uma vez que valores pequenos teriam uma perturbação relativamente grande, ao passo que os maiores valores seriam pouco perturbados. Em pesquisas econômicas, por exemplo, as grandes empresas são justamente as com maior risco de revelação e uma perturbação pequena teria pouco impacto na redução deste risco (HUNDEPOOL *et al*, 2012).
- Microagregação: é um método aplicado tipicamente para dados contínuos. Primeiramente se particiona os registros entre grupos e depois se atribui um valor agregado (geralmente, mas não necessariamente, a média aritmética), para cada variável no grupo. Em outras palavras, os valores individuais das variáveis dos registros são substituídos pelo valor agregado do grupo (TEMPL, 2017). Quanto mais similares são os registros dentro de um grupo, espera-se que menos informação seja perdida ao substituir sua informação pelo valor agregado (ELLIOT e DOMINGO FERRER, 2018). Existem vários métodos para a criação desses grupos como, por exemplo, o de Fadel *et al.* (2021).
- Troca de dados (*Data swapping*): este método troca o valor de certas variáveis de um registro com os valores correspondentes da mesma variável de outro registro. Isto é feito de modo a preservar certas estatísticas resumo dos dados. Embora o método tenha sido proposto inicialmente para variáveis categóricas, uma variante chamada *rank swapping* é também aplicável para variáveis numéricas (ELLIOT e DOMINGO FERRER, 2018).
- Arredondamento: este método altera os valores das variáveis por valores arredondados escolhidos em um conjunto de "pontos de arredondamento". Assim, valores que se encontram entre dois pontos são atraídos para o ponto mais próximo, em valores absolutos. É um método adequado para variáveis numéricas, principalmente contínuas. Dado o conjunto de dados, geralmente o arredondamento é feito em uma variável por vez, mas existem métodos de se proceder um arredondamento multivariado (HUNDEPOOL *et al*, 2012).

- Pós aleatorização (PRAM): o PRAM (*Post-randomization Method*) é um método designado para variáveis categóricas. Neste caso, valores de uma variável são estocasticamente alterados por outros de acordo com um mecanismo de probabilidade markoviana (ELLIOT e DOMINGO FERRER, 2018). Se há um grande número de variáveis-chave categóricas, a recodificação pode não ser suficiente para reduzir o risco de revelação, ou a supressão local pode levar a uma perda de informação grande. Neste caso, o PRAM pode ser uma alternativa eficiente (TEMPL, 2017). Maior detalhamento deste método pode ser visto em Gouweleeuw et al. (1998).

### 2.10.3. Método de mascaramento determinístico ou probabilístico

Um método de mascaramento, também pode ser classificado em determinístico ou probabilístico. Segundo, Benschop *et al.* (2019) os métodos probabilísticos dependem de algum mecanismo probabilístico ou de geração de número aleatório. Assim, toda a vez que este tipo de método é utilizado, o resultado pode ser diferente. Recomenda-se, então, que se guarde a semente geradora do número aleatório caso haja a intenção de produzir resultados replicáveis. Os métodos determinísticos, por sua vez, seguem um determinado algoritmo e produzem os mesmos resultados, se aplicados repetidamente aos dados com o mesmo conjunto de parâmetros. O PRAM é um exemplo de método probabilístico, já a recodificação global é um exemplo de método determinístico.

O Quadro 2.1 traz uma visão geral de todos os métodos de mascaramento apresentados na Seção 2.10, em função de como podem ser classificados e para quais tipos de variáveis são recomendáveis.

Quadro 2.1: Classificação dos métodos de mascaramento e tipos de variáveis para quais são recomendáveis.

| Método | Classificação do método | Tipo de variável |
|---|---|---|
| Amostragem | Não perturbativo, probabilístico | Categórica |
| Recodificação global | Não perturbativo, determinístico | Categórica / contínua |
| Rec. de topo ou fundo | Não perturbativo, determinístico | Categórica / contínua |
| Supressão Local | Não perturbativo, determinístico | Categórica |
| Adição de ruído | Perturbativo, probabilístico | Contínua |
| Microagregação | Perturbativo, probabilístico | Contínua |
| Troca de dados | Perturbativo, probabilístico | Categórica / contínua[4] |
| Arredondamento | Perturbativo, determinístico | Contínua |
| Pós aleatorização | Perturbativo, probabilístico | Categórica |

Fonte: Elaborado pelo autor

### 2.10.4. Dados sintéticos

Os métodos vistos até aqui têm por objetivo mascarar os dados originais. Uma outra abordagem possível é considerar estes dados originais como a realização de algum modelo estatístico. A partir daí seria possível substituí-los por amostras – os dados sintéticos – geradas de acordo com esse modelo. Os dados sintéticos consistem, então, em um conjunto de registros individuais sintéticos a serem disseminados no lugar dos registros originais (DUNCAN et al, 2011).

Hundepool et al. (2012) apontam que existem três possibilidades de utilização deste tipo de dado. A primeira seria um banco de dados totalmente sintético em substituição às informações originais. A segunda, utilizar dados parcialmente sintéticos com apenas as informações mais sensíveis – sejam registros ou variáveis – substituídas. Em outras palavras, apenas estes valores previamente selecionados seriam sintéticos, para os demais seriam divulgados os dados originais. Por fim, outra opção é um dado híbrido combinando informação original e sintética, podendo esta última parte ser maioria ou minoria das informações disseminadas. Neste caso, um método seria

[4] Considerando-se a variante *rank swapping*.

utilizado para criar o banco de dados que combinasse ambas as informações, segundo parâmetros específicos selecionados. Exemplos de abordagens para a criação deste tipo de dados podem ser vistos em Dandekar et al. (2002), Sebé *et al*. (2002) e Muralidhar e Sarathy (2006).

Elliot e Domingo Ferrer (2018) destacam que este tipo de dado, em algumas vezes, é visto na literatura não como um método de CEC e sim como uma alternativa a ele. Entretanto, os autores afirmam que isto é basicamente uma preocupação semântica, pois o objetivo deste método é o mesmo dos outros de CEC: disseminar microdados que sejam úteis ao passo que mantenham a confidencialidade dos respondentes.

Idealmente os dados sintéticos devem ser estatisticamente equivalentes aos da população de interesse a qual eles substituirão. Entretanto, apesar deste método gerar bases de dados com risco de revelação menor do que se fossem utilizados os métodos tradicionais, eles possuem menor valor analítico (TEMPL, 2017). Isto ocorre porque é geralmente muito difícil controlar todos os tipos de análises que os usuários possam ter interesse. Os pesquisadores podem, por exemplo, querer examinar uma variedade de subpopulações e é difícil para o modelo que gerou os dados abarcar todas as possíveis relações condicionais entre as variáveis e possíveis domínios de interesse (WILLENBORG e DE WAAL, 2001). De qualquer modo, eles podem servir muito bem para arquivos de uso público, aos quais o pesquisador pode testar métodos ou com propósitos didáticos (TEMPL, 2017). A teoria sobre este método e sua implementação, podem ser vistas com grande detalhamento em Drechsler (2011).

## 2.11. *K*-anonimidade

Registros com valores de chave únicos ou raros em uma amostra são mais propensos a serem identificados. Assim, uma forma de proteger a confidencialidade dos dados é assegurar que cada valor de chave seja compartilhado por, pelo menos, um número *k* de registros na amostra (TEMPL, 2017). Neste caso, o produtor dos dados após escolher o valor de *k* que julgar adequado, pode aplicar métodos de mascaramento nas variáveis-chave com vistas à obtenção da respectiva *k*-anonimidade. Um arquivo de

dados com 2-anonimidade, por exemplo, é um arquivo tal que não há registros com valores de chave únicos, e assim por diante. Este conceito será utilizado e abordado com mais detalhes no Capítulo 6.

# CAPÍTULO 3: O HISTÓRICO DO CONTROLE ESTATÍSTICO DE CONFIDENCIALIDADE PARA MICRODADOS

Na introdução desta tese foi apresentado o argumento de Elliot e Domingo-Ferrer (2018) de que o campo de estudo do CEC seria o resultado de três mudanças técnico-sociais interrelacionadas, sendo uma delas o avanço da informática. Sob essa perspectiva, o CEC pode ser encarado como uma área resultante de discussões anteriores sobre questões da privacidade do indivíduo frente às novas tecnologias que despontavam. De fato, Holvast (2009) afirma que nos Estados Unidos poucos anos após a Segunda Guerra Mundial, numerosas publicações começaram a surgir abordando o conceito de privacidade em relação aos desenvolvimentos tecnológicos que poderiam colocá-la em risco. Em particular, o computador era visto como o principal meio de uma possível invasão de privacidade. Essas publicações culminaram na fundação em 1962 do Projeto "O Impacto da Ciência e Tecnologia na Privacidade", pela Comissão Especial de Ciência e Direito da Ordem dos Advogados da Cidade de Nova Iorque. O diretor de pesquisas deste projeto, Alan Westin, publicou resultados detalhados deste projeto em seu livro "*Privacy and Freedom*", de 1967, obra que estabeleceu uma base profunda para as discussões posteriores.

Nesta obra o autor define privacidade como "a reivindicação de indivíduos, grupos ou instituições para determinar por si próprios quando, como e em que medida as informações sobre eles são comunicadas a terceiros" (WESTIN, 1967, p.7). Holvast (2009) afirma que desde então praticamente não há publicações sobre este tema em que esta definição não esteja presente. Ao se comparar tal definição de privacidade, com a definida por Hundepool *et al*. (2012) que foi apresentada no Capítulo 2 desta tese vê-se que, de fato, ambas são, em essência, idênticas.

As estatísticas oficiais estão no centro do contexto desta discussão, uma vez que, o *Bureau* do Censo dos Estados Unidos, exatamente nesta época, foi o primeiro órgão de estatística a disponibilizar microdados. As informações eram referentes ao Censo Demográfico de 1960 daquele país, divulgadas em 1962, que representavam uma amostra de 1% da população. Já naquela época observou-se cuidados com a

confidencialidade dos dados, ao se remover identificadores diretos e divulgar recortes geográficos que possuíssem, no mínimo, 250 mil habitantes (McKENNA, 2019d).

Apesar dessas medidas do *Bureau* do Censo, a preocupação com os aspectos ligados à privacidade era crescente, afinal este era apenas o ponto inicial no que dizia respeito à disseminação de microdados. Em paralelo, discutia-se o projeto de criação do Centro Nacional de Dados (*National Data Center* - NDC) que abrigaria dados não apenas do *Bureau* do Censo, mas também de outros órgãos federais como a Receita Federal e a Administração da Seguridade Social, por exemplo (SAWYER e SCHECHTER, 1968). Artigos como os de Chartrand (1967), Dunn (1967) e Miller (1967) abordavam as implicações advindas com a criação do NDC, dando ênfase à questão da privacidade. Entretanto, talvez a melhor definição da dualidade entre oportunidades e riscos que estariam por vir, segundo o espírito da época, tenha sido dada por Sawyer e Schechter (1968) quando afirmou que o NDC "pode ser o recurso de informação mais importante que a ciência social já conheceu, e também pode ser o passo mais significativo no vigoroso ataque da tecnologia à privacidade pessoal" (SAWYER e SCHECHTER, 1968, p.1). É interessante notar que as definições de CEC tanto de Hundepool *et al*. (2012), quanto de Benschop *et al*. (2019) apresentadas no Capítulo 1, podem servir de resposta ao conflito exposto por Sawyer e Schechter (1968), ou seja, tentar aproveitar a máxima utilidade que os microdados possam oferecer, reduzindo o risco de revelação de informações confidenciais a um mínimo aceitável. Em outras palavras, ali já estava plantada a semente deste campo de estudo.

Entretanto, é na década de 1970 que a literatura aponta os primeiros trabalhos do que pode ser formalizado como CEC para microdados. Willenborg e De Waal (2001), por exemplo, apontam as importantes contribuições de Fellegi (1972) e Dalenius (1977). Na primeira obra, o autor aborda a questão da confidencialidade estatística tanto de forma geral, quanto no contexto de cada tipo de produto divulgado. Um destes produtos eram os microdados – chamados então de "amostra de indivíduos 'não identificados' em fita" – e tratados como uma recente inovação (FELLEGI, 1972). Já na segunda obra, o autor trabalha formalmente com a construção de uma metodologia para o CEC e introduz o conceito de que, idealmente, nada sobre um indivíduo deve ser revelado pelo banco de dados, que não poderia ser descoberto por outro meio (DALENIUS, 1977). Esta

é a ideia utilizada por Hundepool *et al*. (2012) para definir o que é revelação, como apresentado no Capítulo 2.

A partir de então, a produção acadêmica sobre o CEC em microdados começa a se intensificar. Observa-se a proposição de métodos de mascaramento, como, por exemplo, o da troca de dados para um banco de dados composto por variáveis categóricas (DALENIUS e REISS, 1978) e posteriormente abrangendo também variáveis contínuas (REISS *et al*. 1982). Hundepool *et al*. (2012), por sua vez, destacam a contribuição de Duncan e Lambert (1986) com a abordagem usual para a definição de risco de revelação de identidade, que foi definida do Capítulo 2, e, na já década seguinte, a introdução de novos métodos de mascaramento, como o da microagregação proposto por Defays e Nanopoulos (1993), ou mesmo formas de combiná-los como o caso da supressão local em conjunto com a recodificação global debatida por De Waal e Willenborg (1995). Essa é apenas uma pequena amostra dentre inúmeras outras contribuições importantes de diversos autores que poderiam ser citadas sobre o tema, tendo este acúmulo de conhecimento culminando, ainda na década de 1990, na publicação de um livro inteiramente dedicado a CEC (WILLENBORG e DE WAAL, 1996).

Já no início do século XXI começam a ocorrer congressos específicos que tratam nomeadamente da questão de confidencialidade de dados. Em 2001, há o primeiro encontro do "*Work Session on Statistical Data Confidentiality*", sediado naquele ano na cidade de Skopje, na Macedônia do Norte. A partir de então, este evento acontece a cada 2 anos, sendo organizado pela Comissão Econômica das Nações Unidas para a Europa (UNECE) e pelo Eurostat. Na última edição ocorrida em 2021 na cidade de Poznań, na Polônia, o evento passou a se chamar "*Expert Meeting on Statistical Data Confidentiality*". Paralelamente, no ano de 2004 em Barcelona, na Espanha, deu-se início a conferência denominada "*Privacy in statistical databases*" organizada pela divisão de privacidade de dados da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), também com periodicidade bienal. Desde então vários outros eventos foram criados como, por exemplo, a Conferência Anual Europeia de Proteção de Dados e Privacidade, iniciada em 2010, com sua décima primeira edição programada para ocorrer em Bruxelas, na Bélgica, em dezembro de 2021.

Ainda neste século, também foram publicados outros livros tanto abrangendo o CEC como um todo, citando-se como exemplos: Willenborg e De Waal (2001), Duncan *et al.* (2011), Hundepool *et al.* (2012) e Templ (2017), quanto de determinados aspectos do tema, como o livro de Drechsler (2011) específico sobre dados sintéticos. Além disso, foram criadas revistas científicas como o "*Journal of Privacy and Confidentiality*" que teve sua chamada inicial de trabalhos em 2008 e sua primeira edição no ano seguinte (ABOWD *et al*., 2009). Todos estes fatos ilustram o contínuo aumento de interesse para as questões de confidencialidade dos dados e, por conseguinte, pelos métodos de CEC, observados nas duas últimas décadas.

Este cenário não deve se alterar tão cedo. Shlomo (2018) argumenta que os contínuos avanços tecnológicos e a pressão cada vez maior dos governos por dados abertos e acessíveis, fazem com que novas formas de disseminação de dados, incluindo aplicativos baseados na *web*, sejam exploradas pelos INE. Por outro lado, a informatização da sociedade com os dados dos indivíduos muitas vezes facilmente obtidas na internet, aumentam os riscos de revelação. Isso muda o panorama de como estes riscos precisam ser definidos e os tipos de métodos de CEC que devem ser aplicados para proteger os dados.

É possível argumentar que embora nossa realidade atual seja muito diferente em relação àquela da década de 1960, esta observação de Shlomo (2018) é análoga e traz os mesmos elementos dos autores americanos daquele período, citados neste capítulo: inovações tecnológicas, novas formas de disseminação de dados e, consequentemente, novos riscos e desafios à confidencialidade das informações. Tudo leva a crer que estes elementos continuem presentes nas próximas décadas, fazendo com que a área de estudo do CEC seja cada vez mais relevante não só no âmbito dos INE, mas para a sociedade como um todo.

## 3.1. O Controle Estatístico de Confidencialidade no âmbito dos INE

Nesta seção será descrito como os principais INE tratam do tema de CEC em suas pesquisas. De forma geral, os INE do continente europeu são os que mais disponibilizam esse tipo de informação publicamente.

A confidencialidade estatística é um princípio fundamental das estatísticas oficiais consagrado no Código de Prática das Estatísticas Europeias. A harmonização de princípios e orientações em matéria de proteção de dados confidenciais é uma obrigação do Eurostat e das autoridades estatísticas nacionais do Sistema Estatístico Europeu (EUROSTAT, 2017). O Eurostat é o órgão de estatística da União Europeia responsável pela publicação de estatísticas e indicadores para a região. Em sua página da internet existe uma seção dedicada exclusivamente ao CEC. Além de várias informações e manuais, é possível saber a respeito do centro de excelência sobre o tema (*Centre of Excellence on Statistical Disclosure Control* – CoE SDC), que tem como objetivo reunir autoridades estatísticas interessadas na colaboração em projetos de CEC. Há projetos em andamento para o desenvolvimento de banco de dados de uso público para o Eurostat, assim como para garantir suporte, manutenção e desenvolvimento de aplicativos para CEC. Atualmente fazem parte deste segundo projeto os *softwares* μ-Argus para microdados, τ-argus para tabelas, e pacotes para o *software* R (sdcMicro, sdcMicroGUI, sdcTable e SimPop). O coordenador de ambos os projetos é o órgão oficial de estatística da Holanda, *Centraal Bureau Voor De Statistiek* (CBS), também conhecido como *Statistics Netherlands*.

Apesar do Eurostat promover a harmonização dos métodos estatísticos entre os INE da região, cada país tem a autonomia no que diz respeito à produção de suas estatísticas oficiais. Afinal cada lugar tem suas legislações específicas, como também demandas distintas por informações de acordo com as suas realidades. Desta forma, será descrito, nesta seção, como alguns INE estão tratando atualmente a questão do CEC na produção de suas pesquisas, o que possibilita ter uma ideia das similaridades e diferenças de abordagem que dispensam ao tema.

A Holanda é, com relação ao estudo do CEC, uma das principais referências tanto na Europa quanto no mundo, com considerável parte desta pesquisa feita dentro de seu INE. O *Statistics Netherlands* disponibiliza, em sua página da internet, uma publicação (HUNDEPOOL e DE WOLF, 2012) que detalha os procedimentos de CEC que são ali utilizados. Nesta publicação é feita uma divisão de subtemas entre os métodos utilizados para microdados, tabelas de frequências, tabelas quantitativas (de magnitude) e resultados de análises estatísticas (*outputs*). Os autores mencionam que embora o

conceito de CEC deva ser levado em conta durante todo o processo estatístico que envolve a pesquisa, tradicionalmente ele ocorre no fim do processo, imediatamente antes da publicação das informações.

No que diz respeito especificamente aos microdados, os autores apontam que o departamento responsável pela construção dos dados é também o responsável pela proteção estatística adequada dos mesmos. Ao selecionar os métodos de CEC a serem empregados, levam-se em conta tanto o risco de revelação quanto a perda de informação, sendo que diferentes níveis de proteção podem ser empregados caso os arquivos de dados sejam para uso público ou para uso sob contrato. Com relação a este último, ou para casos de acesso remoto, os microdados podem gerar *outputs* que não necessariamente satisfazem a política de confidencialidade da instituição. Nestes casos, outros métodos de CEC devem ser utilizados para proteção do tipo de *output* em questão (HUNDEPOOL e DE WOLF, 2012).

Os autores ainda descrevem quais os métodos de mascaramento – contidos dentre os apresentados na Seção 2.10 desta tese – e como eles são utilizados nos microdados do *Statistics Netherlands*. De forma resumida, o uso da recodificação global geralmente se dá em variáveis identificadoras, mas pode ser aplicada em outras variáveis desde que sejam categóricas. A supressão local, por sua vez, é usada geralmente como o último método de mascaramento para casos específicos, após a maior parte da proteção já ter sido fornecida por outros métodos, e seu uso não é limitado às variáveis identificadoras. Já a recodificação de topo, é usada como proteção adicional em situações que valores extremos de variáveis numéricas são considerados como tendo grande risco de revelação. Há também a possibilidade de aplicação de ruídos aleatórios nos pesos amostrais, para casos em que estes possam revelar informação previamente removida total ou parcialmente como, por exemplo, determinado recorte geográfico a eles atrelada. Por fim, o método do PRAM é utilizado nos arquivos de microdados sob contrato, para ser utilizado nas variáveis identificadoras categóricas, o que introduz um grau de incerteza nas mesmas, permitindo ao pesquisador ter acesso a níveis geográficos mais desagregados para suas análises. Isto ocorre uma vez que ao não se saber se todas as variáveis identificadoras de um registro estão inalteradas, não há certeza de uma possível revelação. Entretanto, o usuário deste

tipo de dado deve ter conhecimento estatístico suficiente para lidar de forma metodologicamente adequada com essas modificações. Desta forma, esses arquivos são destinados geralmente para pessoas com experiência na área de estatística ou como arquivos de teste de programas, visando a uma posterior análise com os dados originais por execução remota ou em ambiente controlado na instituição. Todos os métodos de mascaramento descritos são realizados por meio do *software* μ-Argus (HUNDEPOOL e DE WOLF, 2012).

No caso de Reino Unido, o *Office for National Statistics* (ONS), que é o órgão produtor das estatísticas oficiais daquele local, também disponibiliza em sua página da internet uma seção em que detalha os procedimentos de CEC que emprega. Ali é possível ter acesso a uma publicação (GSS, 2014) que relata minuciosamente os cuidados com a confidencialidade no âmbito de microdados de pesquisas sociais. De forma sucinta, e focando-se apenas nos dados de uso público, que é o escopo desta tese, neste documento é descrito como se deve proceder para a formação de cenários de revelação e definição de variáveis-chave; faz considerações sobre tamanho da amostra *versus* risco de revelação; aborda a questão de hierarquia de registros em pesquisas domiciliares, enfatizando o problema dos domicílios com muitos moradores que são de mais fácil identificação; e os cuidados para os casos de dados longitudinais. Entre os métodos de mascaramentos recomendados menciona-se desde os mais usuais como recodificações de topo, arredondamento ou recodificações globais, até medidas mais extremas como, por exemplo, a supressão do registro inteiro nos casos dos domicílios com muitos moradores (GSS, 2014).

É interessante destacar que a instituição também disponibiliza publicações que tratam da questão da confidencialidade de dados para áreas temáticas como, por exemplo, para estatísticas de saúde (ONS, 2006) e estatísticas vitais (ONS, 2014). Estas publicações, no entanto, se concentram em métodos de CEC para dados tabulares.

O órgão de estatísticas oficiais da França - *Institut National de la Statistique et des Études Économiques* (INSEE) – possui uma publicação voltada para o gerenciamento da confidencialidade de dados individualizados (BERGEAT, 2016). Estão ali contempladas as questões como: medidas de risco de divulgação, métodos de mascaramento de dados, medidas de perda de informação e alguns exemplos de

questões tratadas pelo INSEE. Já o *Instituto Nazionale di Statistica* (ISTAT) da Itália tem disponível uma publicação que sumariza todos os métodos e ferramentas de TI para a produção de suas estatísticas (ISTAT, 2018). Na parte que trata dos processos de análise, é dedicada uma parte específica para o CEC. Ali são descritos os métodos utilizados tanto para tabelas quanto para microdados. Para o caso do *Instituto Nacional de Estadística* (INE) da Espanha, Snorrason *et al.* (2015) em um relatório revisado por pares de outros INE, promovido pelo EUROSTAT sobre o Instituto, dedica uma seção para relatar o que é feito no órgão a respeito da confidencialidade de dados tanto em tabelas quanto em microdados.

O Instituto Nacional de Estatística de Portugal (INE, IP), por sua vez, disponibiliza um documento que detalha a sua "Política de Confidencialidade Estatística". Estão ali elencadas as bases jurídicas tanto nacionais, quanto da União Europeia, no que diz respeito ao "Princípio do Segredo Estatístico" utilizado pelo Instituto; as regras gerais de aplicabilidade da confidencialidade estatística; as regras de disponibilização de dados individualizados; e os processos estatísticos visando a proteção da confidencialidade. Sobre este último quesito é reforçado que a privacidade das entidades respondentes e a confidencialidade das suas informações prestadas são inerentes a todo o processo de produção das estatísticas oficiais, desde a elaboração da pesquisa até as avaliações após disseminação de seus resultados (INE,IP, 2019). Especificamente com relação à divulgação dos resultados, a Instituição afirma que são utilizados os métodos de CEC mais adequados que "consistem na análise e modificação/perturbação dos dados originais no sentido de eliminar a possibilidade de identificação direta ou minimizar a identificação indireta dos titulares da informação" (INE,IP, 2019, p.11), embora tais métodos não sejam especificados.

Por fim, é possível citar ainda que alguns órgãos divulgam com antecedência as inovações metodológicas no que diz respeito ao tratamento da questão de confidencialidade dos dados. É o caso, por exemplo, do *Statistisches Bundesamt* (DESTATIS) da Alemanha que publicou, ainda em 2018, como seriam tratadas estas questões no Censo daquele país em 2021 (DESTATIS, 2018).

Fora do continente europeu, os Estados Unidos são outra grande referência mundial nos estudos relativos ao CEC. Como foi relatado no início deste capítulo, desde

a década de 1960 inúmeros artigos ali surgiram apontando a preocupação com a privacidade individual frente a possíveis revelações que os dados estatísticos pudessem produzir na era da informática. Neste país, o órgão produtor das estatísticas oficiais - *United States Census Bureau* - disponibiliza em sua página da internet uma seção exclusiva sobre os temas de confidencialidade e CEC. Ali estão informações de como este tema é tratado internamente como, por exemplo, os comitês existentes; aspectos de TI relacionados à confidencialidade; os métodos de CEC utilizados em seus produtos e artigos de pesquisas sobre o tema feitos dentro da instituição.

Mais especificamente, em relação ao tratamento dos microdados disseminados pela instituição, existem artigos com bastante detalhamento dos procedimentos utilizados. De forma resumida, é possível citar, por exemplo, que a pesquisa mensal que mede a força de trabalho e outros temas em suplementos (*Current Population Survey* - CPS), assim como a pesquisa mensal que mede a renda e situação econômica da população (*Survey of Income and Program Participation* - SIPP), fazem uso das mesmas técnicas de mascaramento. Além da remoção de identificadores diretos e divulgação de áreas com limite mínimo de habitantes (100.000 para a CPS e 250.000 para a SIPP), se faz uso da recodificação de topo e de fundo para variáveis contínuas, recodificação global para variáveis categóricas para que cada categoria contenha no mínimo 10.000 pessoas estimadas, arredondamento para valores monetários e adição de ruído em casos específicos de pessoas ou domicílios com características incomuns (McKENNA, 2019a; McKENNA, 2019c).

O Censo dos Estados Unidos, por produzir uma base de dados bem maior do que as pesquisas previamente citadas, emprega outros métodos de mascaramento adicionais em seus dados, como a troca de dados, ou mesmo a utilização de dados parcialmente sintéticos. É importante destacar que algumas técnicas, como a recodificação de topo e de fundo, já eram utilizadas desde o Censo de 1990 (McKENNA, 2019d). Por outro lado, não há divulgação de microdados do Censo Econômico, que pesquisa as empresas daquele país. Os resultados são disseminados por meio de tabelas e utilizam-se métodos de CEC para dados tabulares (McKENNA, 2019b). Porém, o caso de maior destaque é o da *American Community Survey* (ACS), onde, na Conferência dos usuários de dados da ACS em maio de 2021, foi anunciado que os dados divulgados da

pesquisa serão totalmente sintéticos a partir de 2024 (RODRIGUEZ, 2021). A ACS é uma pesquisa nacional que coleta e produz informações das características sociais, econômicas, de habitação, e demográficas da população. Todo o ano são contactadas mais de 3,5 milhões de domicílios, que formam a amostra do levantamento (US CENSUS BUREAU, 2017).

Ainda na América do Norte, o *Statistics Canada* (STATCAN) possui uma publicação sobre as diretrizes de qualidade da Instituição (STATCAN, 2019), tal que dentre as questões abordadas, está contido o tema de confidencialidade. Um ponto interessante é que o órgão utiliza um aplicativo de desenvolvido internamente, chamado G-Confid (*Disclosure Avoidance - Generalized System)* indicado para tratar especialmente do CEC. O *Australian Bureau of Statistics* (ABS), por sua vez, disponibiliza em sua página na internet uma publicação (ABS, 2017) específica sobre como é tratada a questão da confidencialidade dos dados no órgão. Ali é mostrado com detalhes e exemplos inclusos, todas as etapas do processo de CEC, desde o cálculo do risco de revelação, passando pelos métodos de mascaramento aplicados até a mensuração do impacto destes métodos, nos dados disseminados.

Ao se pesquisar os INE dos países da América Latina, verifica-se que, em suas páginas da internet, é afirmado ao público que a confidencialidade das informações prestadas é assegurada. Em alguns casos apresentam-se adicionalmente os dispositivos legais garantidores referentes ao tema, como faz a Argentina (INDEC, 1999) por exemplo, ou, como outros países, é exibido o código de boas práticas estatísticas. Para esta última alternativa, há países que disponibilizam tanto o seu próprio código, tendo como o exemplo o Peru (INEI, 2012), quanto o código para países da América Latina e Caribe, caso do Uruguai (DANE e CEPAL, 2011).

Entretanto, é muito difícil localizar, nestes INE, publicações pormenorizando o processo de anonimização que utilizam em suas bases de dados divulgadas. O *Departamento Administrativo Nacional de Estadística* (DANE) da Colômbia, por exemplo, disponibiliza em sua página da internet um guia para anonimização das bases de dados do Sistema Estatístico Nacional daquele país (DANE, 2018). Esta publicação funciona como um manual para outros órgãos, com os principais tópicos referentes ao CEC. Todavia, apesar de ali constar que o documento “baseia-se na experiência do DANE

na implementação de seus próprios processos de anonimização em diferentes bancos de dados" (DANE, 2018, p.4), não se especifica quais são eles. Em outros países, todavia, é possível encontrar somente os dispositivos legais acerca da confidencialidade das informações ou o código de boas práticas estatísticas.

O *Instituto Nacional de Estadística y Geografía* (INEGI) do México, por outro lado, disponibiliza uma publicação sobre os aspectos relevantes da gestão de confidencialidade e proteção de dados individualizados na instituição (INEGI, 2019). Neste documento há um capítulo que trata de propostas para o futuro, onde há uma sessão específica que trata do CEC. Ali, afirma-se que "não é suficiente remover elementos que podem ser usados para identificar diretamente um indivíduo para garantir que esse indivíduo não possa mais ser identificado" (INEGI, 2019, p.28), alertando para os riscos de revelação apresentados no Capítulo 2 desta tese. É recomendado inclusive que se leve em consideração um manual (GT ART.29, 2014) produzido por um grupo de trabalho da União Europeia sobre CEC para microdados. Desta forma, é possível imaginar que o INEGI atualmente esteja trabalhando na implantação destes métodos em suas bases de dados.

Por fim, com relação ao IBGE, Silva (1988) já apontava, em fins da década de 1980, para importância da apreensão e utilização, no órgão, de métodos e técnicas para minimizar os riscos de revelação de informações. A ideia do texto era subsidiar o debate interno sobre a questão da confidencialidade das informações estatísticas. O autor destacou como recomendação principal, a criação de um grupo de trabalho responsável para propor normas e procedimentos para a manutenção da confidencialidade das informações coletadas, produzidas, disseminadas e armazenadas pela Instituição.

Na década seguinte, Bianchini (1994) retomava os argumentos de Silva (1988) apontando que muito pouco tinha sido feito no sentido de uma formulação de política de disseminação de informações. Apresentava também algumas considerações sobre o tratamento da questão da confidencialidade por INE internacionais, mencionando uma já vasta bibliografia sobre o assunto na época. Em sua conclusão, reforçava a necessidade de criação de um grupo de trabalho para se encarregar do tema relativo à confidencialidade das informações no IBGE.

Finalmente em 1999 foi criado o “Grupo de Sigilo” – que daria origem, em 2001, ao “Comitê de Sigilo” mencionado no Capítulo 1 – com o intuito de criar normas e procedimentos destinados à manutenção da confidencialidade das informações no IBGE, visando reduzir os riscos de revelação e o cumprimento da promessa de assegurar a privacidade dos respondentes. Naquele mesmo ano, o grupo lançou uma publicação, onde já apresentava as primeiras atividades realizadas, e apontava os desafios e perspectivas para o futuro (BIANCHINI *et al.*, 1999).

A partir deste ponto, o IBGE começou a voltar mais atenção ao tema. As atuais medidas visando garantir a confidencialidade nos microdados disseminados de suas pesquisas, foram brevemente descritas no Capítulo 1, e podem ser vistas com mais detalhes em IBGE (2018a). Deve ser ressaltado que, após essa publicação, a pesquisa sobre o tema segue em andamento na Instituição. Um exemplo é o trabalho de Fadel (2021) que investiga questões relativas à confidencialidade nos microdados da PeNSE, focando-se em domínios com frações amostrais mais altas, às vezes muito próximas a um censo, como é o caso de escolas particulares localizadas fora da capital em UF da Região Norte. No âmbito da Escola Nacional de Ciências Estatísticas (ENCE) do IBGE também há pesquisas realizadas sobre o tema, como o já citado trabalho de Silva (2020) acerca da avaliação do risco de revelação nos microdados da Amostra do Censo Demográfico 2010, assim como as contribuições de Arantes e Magalhães (2019) e Arantes *et al.* (2021) na aplicação de métodos de CEC para dados tabulares de uma pesquisa econômica – a Pesquisa de Inovação Tecnológica (PINTEC) – do IBGE.

Cabe ainda mencionar que a Instituição recentemente lançou a segunda edição do “Código de Boas Práticas das Estatísticas do IBGE” (IBGE, 2021). Embora tenha ocorrido uma alteração na nomenclatura do Princípio 4 de “confidencialidade estatística” para “sigilo estatístico” não há alterações em seu conteúdo fundamental. A ideia desta segunda edição era refletir as mudanças e inovações no contexto das estatísticas oficiais do Instituto, das tecnologias, fontes de dados emergentes e do arcabouço jurídico, assim como os resultados de uma auditoria externa realizada em 2016.

# CAPÍTULO 4: OS MICRODADOS E A PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS CONTÍNUA

Na década de 1960, como já mencionado no Capítulo 3, inicia-se a divulgação de microdados das pesquisas realizadas pelos INE. Já naquela época era possível vislumbrar tanto os ganhos em termos de recurso de informação, quanto os perigos em termos de privacidade dos indivíduos – decorrente da confidencialidade destas informações – que estavam por vir (SAWYER e SCHECHTER, 1968). Passadas pouco mais de cinco décadas, essa forma de disseminação da informação se tornou cada vez mais usual, mas os desafios no que diz respeito à questão da confidencialidade continuam presentes. Taylor *et al.* (2018) argumentam que ainda hoje é comum que uma etapa fundamental do CEC, correspondente ao cálculo do risco de revelação dos registros, ser negligenciada. Embora já existam vários métodos estatísticos bem estudados para tal fim na literatura, poucos órgãos de estatística os colocam em prática. Em vez disso, muitas vezes utilizam julgamentos subjetivos, experiências passadas, regras *ad hoc* ou normas genéricas a serem aplicadas conforme o tipo do dado a ser divulgado.

É preciso, todavia, fazer uma distinção entre os microdados provenientes de uma pesquisa de empresas e dos que têm como população alvo o indivíduo, como as pesquisas domiciliares, por exemplo. O'Keefe e Shlomo (2012), argumentam que os dados do primeiro tipo possuem características bem particulares, tais como: as maiores empresas sempre fazem parte da amostra, por outro lado quanto menor a empresa, menor a chance de sua inclusão na pesquisa; a maioria das variáveis são contínuas ao invés de discretas; a distribuição de muitas variáveis apresenta forte assimetria; os dados contêm muitos *outliers* que, na maioria das vezes, pertencem às maiores empresas da amostra; entre outros. Adicionalmente os dados deste tipo de pesquisas podem ser altamente estratégicos sob o ponto de vista comercial, tendo as empresas grande interesse na confidencialidade de suas informações.

Por todas estas características, muitos INE – e o IBGE é um deles – não disponibilizam os microdados deste tipo de pesquisas, tendo a divulgação de suas informações feitas por outros meios como, por exemplo, conjuntos de tabelas. O'Keefe e Shlomo (2012), apontam ainda outras estratégias que são utilizadas, como a

divulgação de microdados sintéticos em vez dos originais, o uso de um sistema de análise remota no qual o pesquisador não tem acesso direto aos microdados, dentre outros.

O IBGE divulga os dados deste tipo de pesquisa em formas de tabelas e, ainda assim, aplica procedimentos específicos nas mesmas visando não permitir a revelação de informação individualizada das unidades respondentes. Há procedimentos genéricos, como, por exemplo, exigir pelo menos três respondentes por célula da tabela, e outros específicos em função de particularidades temáticas de cada investigação. Em outras palavras, cada pesquisa terá procedimentos adicionais para a proteção da informação (IBGE, 2018a).

Por outro lado, o IBGE dissemina publicamente microdados desidentificados das pesquisas amostrais domiciliares, além da parte investigada por amostragem dos censos demográficos e da Pesquisa Nacional de Saúde do Escolar (PeNSE). Em todos estes casos, a unidade respondente é o indivíduo. As pesquisas deste tipo que possuem arquivos de microdados com acesso público na página da internet da Instituição, além da parte amostral dos Censos e da PeNSE, são: Economia Informal Urbana (ECINF), Pesquisa de Orçamentos Familiares (POF), Pesquisa Mensal de Emprego (PME), Pesquisa Nacional de Saúde (PNS), Pesquisa Nacional por Amostra de Domicílios (PNAD) e a Pesquisa Nacional por Amostra de Domicílios Contínua (IBGE, 2018a). Vale destacar que a PNAD e PME, embora ainda tenham microdados disponíveis, foram encerradas em 2015 e 2016, respectivamente, sendo substituídas, com metodologia atualizada, pela PNAD Contínua.

Como já descrito no Capítulo 1, nestes arquivos de uso público são eliminadas as variáveis de identificação direta dos informantes, como, por exemplo, nome e endereço. Há também um processo de ordenação aleatória dos domicílios dentro da área de divulgação, e o próprio processo de amostragem já contribui para a redução do risco de revelação de dados individuais. Contudo, não há institucionalizada uma medida formal de risco de revelação calculada tanto para o arquivo como um todo, quanto para os registros individuais que o compõe. Silva (2020), no entanto, fez estes cálculos para os microdados da parte amostral do Censo Demográfico 2010 e julgou que, em todos os

cenários de revelação propostos, os riscos seriam considerados muito altos para a sua disseminação pública.

É verdade que esta pesquisa analisada possui frações amostrais bem superiores às demais, ou seja, é o caso em que o processo de amostragem traz, em teoria, menores ganhos na proteção aos riscos de revelação. Ainda assim, é preciso se certificar que não haja ameaça à confidencialidade das informações em quaisquer arquivos a serem disponibilizados ao público. Por isso, é importante também estabelecer critérios em relação aos procedimentos a serem tomados, caso seja encontrado risco acima do limite tolerável (global ou individual) antes da divulgação dos dados.

Desta forma, a presente tese tem como foco principal propor uma abordagem para a implantação de métodos de CEC, tendo por base a avaliação do risco de revelação dos microdados das pesquisas domiciliares, disponibilizados publicamente pelo IBGE. Uma vez que é preciso tomar medidas para proteger a confidencialidade das informações, caso o risco calculado esteja acima de um limite tolerável definido, também serão abordadas as questões referentes aos procedimentos de mascaramento de dados e cálculo da perda de informação decorrente deste processo. Para ilustrar o processo será utilizada a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), referente ao segundo trimestre de 2019. Esta pesquisa e a justificativa de sua escolha e de seu recorte temporal serão detalhados nas seções seguintes.

## 4.1. O Sistema Integrado de Pesquisas Domiciliares e a PNAD Contínua

A PNAD Contínua faz parte do Sistema Integrado de Pesquisas Domiciliares (SIPD) que o IBGE começou a implantar no ano de 2006. Este sistema "constitui um modelo de produção de pesquisas amostrais domiciliares no qual o planejamento, a execução, a análise e a disseminação dos resultados são conduzidas de forma coordenada..." (IBGE, 2021b, p.6). Esta integração das pesquisas é uma ferramenta importante tanto para a harmonização dos conceitos e processos, quanto para a otimização dos recursos da Instituição.

Um ponto importante a ser destacado no âmbito do SIPD é a construção de uma estrutura amostral, chamada de "amostra mestra", que tem a finalidade de atender a

todas as pesquisas domiciliares do IBGE. De forma mais detalhada: selecionam-se unidades de área, geralmente um conjunto de setores censitários, que constituem as unidades primárias de amostragem (UPA) da amostra mestra, e as pesquisas domiciliares da Instituição podem utilizar uma subamostra destas UPA, como é o caso da POF, ou mesmo a amostra inteira, como é o caso da PNAD Contínua. Uma vez que distintas pesquisas podem utilizar uma mesma UPA, é preciso ter controle nas demais etapas da amostragem para que não haja domicílios comuns a mais de uma pesquisa (IBGE, 2021b).

No que concerne às motivações temáticas, o SIPD também foi pensado para suprir uma importante lacuna nas estatísticas oficiais do país, que era a produção, com abrangência nacional, de indicadores de curto prazo sobre trabalho e rendimento. Até então, tais indicadores estavam disponíveis apenas para as seis principais regiões metropolitanas investigadas pela Pesquisa Mensal de Emprego[5]: Rio de Janeiro, São Paulo, Belo Horizonte, Recife, Salvador e Porto Alegre. Adicionalmente, a produção regular de dados sobre consumo e orçamentos familiares, que fornece importante retrato sobre a qualidade de vida da população, era outra demanda a ser atendida pelo sistema de pesquisas. Por fim, esta integração também permite a possibilidade de investigações suplementares, possibilitando explorar novos temas nas pesquisas já existentes, com prazos de planejamento, execução e publicação mais reduzidos (IBGE, 2007). Atualmente o SIPD é composto por três pesquisas: a PNS, a PNAD Contínua e a POF (IBGE, 2021b).

Com relação à PNAD Contínua, a base de dados utilizada nesta tese, ela tem como objetivo a produção de "informações contínuas sobre a inserção da população no mercado de trabalho associada a características demográficas e de educação, e, também, para o estudo do desenvolvimento socioeconômico do País..." (IBGE, 2021b, p.5). Além dos temas permanentes, ainda é possível fazer investigação de temas suplementares de acordo com necessidades específicas. Um exemplo foi a investigação sobre turismo realizada na pesquisa referente ao terceiro trimestre de 2019. A abrangência da pesquisa é nacional, sendo seus principais resultados divulgados

[5] Esta pesquisa, realizada mensalmente, produziu indicadores até fevereiro de 2016.

anualmente não apenas para a totalidade do país, mas também para recortes geográficos menores como Grandes Regiões, Unidades da Federação, determinadas Regiões Metropolitanas que contêm municípios das capitais, assim como para todos os municípios das capitais. Com relação aos resultados conjunturais de mercado de trabalho, sua divulgação tem periodicidade trimestral para estes mesmos recortes geográficos, e mensal, por trimestres móveis, para o Brasil (IBGE, 2021b).

A PNAD Contínua foi implantada, inicialmente de forma experimental, em outubro de 2011, ainda sem abrangência nacional, com o intuito de realizar os ajustes necessários nos processos da pesquisa. Finalmente, a partir de janeiro de 2012, ela começou a cobrir todo o país fazendo parte, definitivamente, do conjunto de pesquisas do IBGE. Atualmente investiga-se, a cada trimestre, por volta de 211 mil domicílios, em aproximadamente 16 mil setores censitários. Cada domicílio é visitado em cinco ocasiões, sendo composta por um conjunto de perguntas conjunturais sobre a força de trabalho investigada em todas as entrevistas e divulgadas a cada trimestre, e outro grupo de perguntas adicionais, algumas delas investigadas apenas na primeira visita, com divulgação anual (IBGE, 2021b).

O plano amostral da PNAD Contínua é o de conglomerados em dois estágios de seleção, com estratificação das UPA, que podem ser compostas por um setor censitário ou um conjunto deles. As UPA são selecionadas com probabilidade proporcional ao número de seus domicílios. O segundo estágio consiste na seleção de 14 domicílios particulares permanentes ocupados dentro de cada UPA, feita por amostragem aleatória simples. A amostra de UPA e de domicílios é dividida pelos três meses de um trimestre. Entretanto, por se tratar de uma pesquisa contínua que acompanha mercado de trabalho, planeja-se a amostra para que haja rotação dos domicílios selecionados. O esquema de rotação da amostra que se adotou foi o 1-2(5), ou seja, domicílio é entrevistado um mês e sai da amostra nos dois meses seguintes, sendo este processo repetido num total de cinco vezes. Para tal, definiu-se 15 grupos de rotação de domicílios, dividindo a amostra de UPA nesses grupos, com cinco grupos pesquisados por mês. Ao final do trimestre, acumula-se a amostra para produção dos indicadores (IBGE, 2021b).

A Ilustração 4.1 traz o esquema de rotação da PNAD Contínua. Os números de 1 a 5 dentro dos quadrados em cinza indicam qual entrevista (primeira a quinta) está sendo realizada no determinado mês. Desta forma, é possível visualizar que, em cada mês, cinco grupos são pesquisados, totalizando os quinze grupos no trimestre. Nas colunas estão representados os painéis. Tomando-se como exemplo o painel "1M", observa-se o grupo que entrou na amostra com sua primeira entrevista no mês de outubro de 2011, fez a segunda entrevista em janeiro de 2012 e assim por diante até a última e quinta entrevista em outubro de 2012.

Ilustração 4.1: - Esquema de rotação da PNAD Contínua.

| MÊS | PAINEL | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1A | 1B | 1C | 1D | 1E | 1F | 1G | 1H | 1I | 1J | 1K | 1L | 1M | 1N | 1O | 2A | 2B | 2C | 2D | 2E | 2F | 2G | 2H | 2I | 2J | 2K | 2L | 2M | 2N | 2O | 3A | 3B | 3C | 3D | 3E | 3F | 3G | 3H | 3I |
| out-11 | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| nov-11 | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| dez-11 | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| jan-12 | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| fev-12 | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| mar-12 | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| abr-12 | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| mai-12 | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | | |
| jun-12 | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | | |
| jul-12 | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | | | |
| ago-12 | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | | |
| set-12 | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | | |
| out-12 | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | | |
| nov-12 | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | | |
| dez-12 | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | | |
| jan-13 | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | | |
| fev-13 | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | | |
| mar-13 | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | | |
| abr-13 | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | | |
| mai-13 | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | | |
| jun-13 | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | | |
| jul-13 | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | | |
| ago-13 | | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | | |
| set-13 | | | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | | |
| out-13 | | | | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | | |
| nov-13 | | | | | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | | | |
| dez-13 | | | | | | | | | | | | | | | | | | | | | | | | | | | 5 | | | 4 | | | 3 | | | 2 | | | 1 | |

Fonte: IBGE (2014).

## 4.2. Microdados da PNAD Contínua: disseminação e tema tratados

O IBGE atualmente disponibiliza de forma pública em sua página da internet um conjunto de seis arquivos de microdados por ano, da PNAD Contínua. Quatro destes arquivos são os da chamada "divulgação trimestral", ou seja, os referentes às informações colhidas em cada um dos quatro trimestres do ano calendário. Os outros dois são os da chamada "divulgação anual", que contêm informações sobre os temas e

tópicos suplementares investigados em uma visita específica ao longo dos trimestres civis do ano. Até o presente momento, estes temas são pesquisados apenas na primeira ou na quinta visita, então são os arquivos referentes a estas visitas que estão disponíveis. Estes arquivos são compostos pelo acumulado dos domicílios que estavam naquela determinada visita em cada um dos trimestres do ano. Desta forma, os dois arquivos de divulgação anual possuem em torno de quatro quintos dos registros dos microdados de divulgação trimestral. Isto se dá porque se acumula aproximadamente um quinto da amostra em cada um dos quatro trimestres.

Em um trimestre são visitadas 15.096 UPA espalhadas por todo o Brasil e, em cada uma delas, há entrevistas em 14 domicílios, o que totaliza 211.344 domicílios no período (IBGE, 2021b). Assim, os microdados da divulgação trimestral possuem atualmente as informações de cerca de 550 mil pessoas entrevistadas, enquanto nos de divulgação anual este montante é de, aproximadamente, 440 mil.

Deve ser destacado que algumas perguntas feitas aos respondentes podem variar tanto de acordo com o trimestre da pesquisa, quanto em relação a qual das cinco visitas está sendo feita do domicílio. Usualmente há um tópico suplementar de educação no segundo trimestre e outro sobre acesso à televisão, Internet e posse de telefone móvel celular para uso pessoal no quarto trimestre. Especificamente no ano de 2019 houve um questionário adicional sobre turismo no terceiro trimestre. Por outro lado, alguns temas como habitação, características gerais dos moradores e informações adicionais da força de trabalho são investigados na primeira visita. Já temas como trabalho de crianças e adolescentes, e outras formas de trabalho são pesquisados na quinta visita. Há ainda o tema relativo aos rendimentos de outras fontes que é investigado tanto na primeira quanto na quinta visita. A Ilustração 4.2 detalha as pesquisas suplementares anuais da PNAD Contínua em função do trimestre e da visita do ano de 2012 até 2020.

Outra característica importante de ser ressaltada sobre os microdados da PNAD Contínua, é que os domicílios são identificados por uma chave composta por três variáveis: a UPA a qual ele está inserido, o número de seleção do domicílio (dentre os 14 entrevistados) dentro da UPA e o painel a qual ele pertence. Esta chave permite que o domicílio possa ser acompanhado durante suas entrevistas nos diferentes trimestres.

Existe também uma chave que caracteriza o indivíduo que, além destas três variáveis mencionadas anteriormente, leva em conta também o número de ordem do morador dentro do domicílio. Entretanto, ela é útil apenas na identificação da pessoa no arquivo em que ela se encontra, não servindo para acompanhar o indivíduo entre diferentes trimestres.

Ilustração 4.2: - Temas e tópicos suplementares da PNAD Contínua de 2012 a 2020.

| Temas e tópicos suplementares | Ano | Forma de investigação anual | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Visita acumulada no ano (Visita) | | | | | Concentrada no trimestre (Trimestre) | | | |
| | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 |
| Características adicionais do mercado de trabalho | 2012 | X | | | | | | | | |
| | 2013 | X | | | | | | | | |
| | 2014 | X | | | | | | | | |
| | 2015 | X | | | | | | | | |
| | 2016 | X | | | | | | | | |
| | 2017 | X | | | | | | | | |
| | 2018 | X | | | | | | | | |
| | 2019 | X | | | | | | | | |
| Rendimento de outras fontes | 2012 | X | | | | | | | | |
| | 2013 | X | | | | | | | | |
| | 2014 | X | | | | | | | | |
| | 2015 | X | | | | | | | | |
| | 2016 | X | | | | X | | | | |
| | 2017 | X | | | | X | | | | |
| | 2018 | X | | | | X | | | | |
| | 2019 | X | | | | X | | | | |
| | 2020[1] | X | | | | X | | | | |
| Habitação | 2016 | X | | | | | | | | |
| | 2017 | X | | | | | | | | |
| | 2018 | X | | | | | | | | |
| | 2019 | X | | | | | | | | |
| | 2020[1] | X | | | | | | | | |
| Outras formas de trabalho | 2016 | | | | | X | | | | |
| | 2017 | | | | | X | | | | |
| | 2018 | | | | | X | | | | |
| | 2019 | | | | | X | | | | |
| Trabalho de crianças e adolescentes | 2016 | | | | | X | | | | |
| | 2017 | | | | | X | | | | |
| | 2018 | | | | | X | | | | |
| | 2019 | | | | | X | | | | |
| Educação | 2016 | | | | | | | X | | |
| | 2017 | | | | | | | X | | |
| | 2018 | | | | | | | X | | |
| | 2019 | | | | | | | X | | |
| TIC | 2016 | | | | | | | | | X |
| | 2017 | | | | | | | | | X |
| | 2018 | | | | | | | | | X |
| | 2019 | | | | | | | | | X |
| Turismo | 2019 | | | | | | | | X | |

Fonte: Extraído de http://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por__Amostra_de_Domicilios_continua /Trimestral /Microdados /PNADC_Pesquisas_Suplementares _Anuais_20210414.pdf em agosto de 2021.

## 4.3. Escolha do arquivo de dados a ser utilizado

No momento de elaboração da presente tese, existiam microdados da PNAD Contínua de disseminação pública na página da internet do IBGE até o ano de referência de 2020. Entretanto, é preciso levar em conta que este foi um ano atípico devido ao início da pandemia do SARS-COV-2, com as entrevistas da pesquisa passando a serem feitas por telefone e não mais presencialmente. Em nota técnica, a Instituição relatou que, de forma geral, as pesquisas amostrais por domicílio, por ela conduzidas, apresentaram percentuais mais elevados de não respostas, devido a entrevistas não realizadas, em relação aos anos anteriores. Especificamente em relação à PNAD Contínua, tal comportamento não acarretou prejuízos à qualidade de seus indicadores-chave no âmbito da divulgação trimestral. Entretanto, avaliou-se a necessidade de se estudar os padrões de não-resposta devido às mudanças no comportamento dos entrevistados e em relação à nova abordagem ao informante por meio de telefone. Avaliou-se também que deveriam ser divulgados apenas os domínios em que não houvesse prejuízo na qualidade da informação, suprimindo-se os demais (IBGE, 2020). Desta forma, devido à excepcionalidade na coleta das informações causada pela pandemia, que se estendeu, inclusive, no ano de 2021, e pelo propósito deste trabalho não necessitar da utilização de dados de divulgação mais recente possível, optou-se por utilizar microdados referentes ao ano de 2019.

Para o ano de 2019 existem seis arquivos de microdados disponíveis para o acesso público na página do IBGE: os quatros arquivos de divulgação trimestral e os dois de divulgação anual já descritos na seção anterior. Igualmente foi mencionado que os arquivos de divulgação trimestral possuem mais registros de indivíduos. A Tabela 4.1 apresenta o total de pessoas em cada arquivo divulgado publicamente para o ano escolhido para o estudo. A utilização de um arquivo com o maior número de registros possível é preferível, seja pela maior quantidade de indivíduos passíveis de identificação pelo intruso, seja por apresentar uma maior fração amostral, ou seja, menor proteção que a amostra naturalmente confere à redução do risco de revelação de uma entidade. Este fator, inclusive, foi determinante na escolha da PNAD Contínua, uma vez que esta pesquisa utiliza a amostra mestra inteira, como descrito na seção 4.1, possuindo, então,

o maior número de registros e fração amostral entre as pesquisas domiciliares regulares do IBGE. Ainda assim, deve-se destacar que o total de pessoas na amostra em cada arquivo de divulgação trimestral corresponde a pouco menos de 0,3% da população brasileira naquele ano. Este valor, por exemplo, é muito menor se comparado à população contida na pesquisa amostral do Censo Demográfico, onde Silva (2020) encontrou riscos de revelação muito altos para a disseminação pública.

Tabela 4.1: Total de pessoas na amostra nos microdados de divulgação anual e trimestral da PNAD Contínua, referentes ao ano de 2019.

| Microdados da PNAD Contínua 2019 | Pessoas na amostra |
|---|---|
| 1º trimestre | 553.308 |
| 2º trimestre | 551.348 |
| 3º trimestre | 550.227 |
| 4º trimestre | 542.802 |
| 1ª visita | 443.790 |
| 5ª visita | 433.535 |

Fonte: PNAD Contínua 2019 - microdados de divulgação anual e trimestral.

Definido que seriam usados microdados de divulgação trimestral para o ano de 2019, a etapa seguinte seria definir qual trimestre utilizar. Como já descrito na Seção 4.1, dependendo do trimestre pode haver algum tópico suplementar na pesquisa. Foi mostrado na Ilustração 4.2 que, para o ano escolhido, todos os trimestres apresentavam módulos adicionais, com exceção do primeiro. Uma vez que, para um intruso, geralmente quanto mais informação no banco de dados melhor, pela possibilidade de mais variáveis indicadoras para facilitar a revelação, e de mais variáveis sensíveis em caso de sucesso na identificação da entidade, optou-se por um dos trimestres com tópicos suplementares. Dentre os três temas disponíveis, observa-se que perguntas sobre educação possuem um potencial maior para conter variáveis que possam ser consideradas identificadoras em relação aos demais. Este também é um tema, ao contrário do suplemento de turismo que foi específico para o ano de 2019, que vem

sendo recorrente desde 2016 na pesquisa. Sendo assim, definiu-se por se utilizar nesta tese os microdados referentes à divulgação do segundo trimestre do ano escolhido.

A pesquisa escolhida é dividida em cinco partes temáticas. A primeira é chamada de "identificação e controle", que contém informações sobre a localização geográfica do domicílio, e variáveis sobre os seus aspectos amostrais (UPA, peso, painel, por exemplo). A segunda é referente às "características gerais dos moradores" que são as informações básicas de cada pessoa como idade, sexo, cor/raça, entre outros. A terceira é o suplemento específico do trimestre chamado de "características de educação para os moradores de 5 anos ou mais de idade", que contém uma série de perguntas sobre este tema para os entrevistados. A quarta parte é sobre as "características de trabalho das pessoas de 14 anos ou mais de idade", que é a mais extensa e, geralmente, a que mais interessa aos usuários. A quinta parte, por sua vez, compreende os "rendimentos de outras fontes", incluídos ali os rendimentos de programas sociais. Além destas informações, o banco de dados também contém uma série de variáveis derivadas das contidas em algumas das cinco partes anteriormente mencionadas. Um exemplo é o número de anos de estudo a partir das informações do curso que frequenta ou que frequentou. A Tabela 4.2 apresenta o total de variáveis no banco de dados escolhido, em cada tema pesquisado. O dicionário da pesquisa, que especifica cada variável assim como suas categorias é disponibilizado publicamente na página da internet do IBGE.

Este será, então, o banco de dados utilizado nos capítulos seguintes, nos quais serão apresentados os procedimentos de Controle Estatístico de Confidencialidade, assim como uma proposta de como incorporá-los ao processo de produção das pesquisas do IBGE. Devido a algumas características mais complexas da PNAD Contínua em relação à maioria das outras pesquisas domiciliares da Instituição, como a rotação da amostra, distintas informações colhidas dependendo da visita e do trimestre, dentre outras, faz com que essa aplicação seja um bom exemplo, com vistas à construção de uma abordagem de CEC para os microdados de uso público, deste tipo de pesquisas, disponibilizados pelo IBGE.

Tabela 4.2: Total de variáveis na amostra nos microdados de divulgação anual e trimestral da PNAD Contínua, referentes ao ano de 2019.

| PNAD Contínua 2019 – 2º trimestre | Número de variáveis |
|---|---|
| Parte 1 - Identificação e controle | 16 |
| Parte 2 - Características gerais dos moradores | 9 |
| Parte 3 - Características de educação | 60 |
| Parte 4 - Características de trabalho | 135 |
| Parte 5 - Rendimentos de outras fontes | 16 |
| Variáveis derivadas | 51 |
| Total geral | 287 |

Fonte: PNAD Contínua 2019 – 2º trimestre.

# CAPÍTULO 5: ESTIMAÇÃO DO RISCO DE REVELAÇÃO

Este capítulo tem como foco as questões relativas à estimação do risco de revelação para microdados de uma pesquisa amostral domiciliar. Esta estimação compreenderá tanto o risco de revelação individual de cada registro, quanto o risco global da pesquisa. Hundepool *et al.* (2012) argumentam que em um microdado de uma pesquisa censitária, o cálculo do risco de revelação pode ser feito de forma direta como uma função dos valores das variáveis indicadoras, assumindo-se que tais dados representem toda a população. Entretanto, nas pesquisas amostrais, raramente estes valores populacionais são conhecidos, sendo a estimação do risco de revelação feita por modelagem probabilística ou na aplicação de heurísticas[6] com base nas informações amostrais.

Como já descrito no Capítulo 1, atualmente o IBGE adota algumas medidas visando garantir a confidencialidade dos microdados disseminados de suas pesquisas amostrais domiciliares, porém não faz o cálculo do risco de revelação, seja individual ou global, destas informações. Desta forma, a estimação de uma medida objetiva do risco é fundamental para avaliar se as atuais medidas adotadas são suficientes e, consequentemente, asseverar a pretendida confidencialidade dos dados divulgados.

É preciso ressaltar que o processo de estimação do risco de revelação envolve etapas, que são tratadas neste capítulo, tais como a determinação de um cenário de revelação e das variáveis-chave em que serão baseados tais cálculos, o mapeamento de fontes externas de dados que possam auxiliar possíveis intrusos, dentre outras. Desta forma, o presente capítulo também tem por objetivo propor uma abordagem para estas etapas, tendo a PNAD Contínua referente ao 2º trimestre de 2019 como exemplo.

Deve-se ainda levar em conta que a estrutura hierárquica dos microdados da PNAD Contínua, mais especificamente, as pessoas alocadas em domicílios, deve ser considerada na estimação do risco de revelação. Esta estrutura aumenta o risco de revelação e seu efeito tende a ser ampliado quanto maior o tamanho do domicílio, em

[6] Algoritmos heurísticos não garantem encontrar a solução ótima de um problema, mas são capazes de retornar uma solução de qualidade em um tempo adequado para as necessidades da aplicação. O objetivo de uma heurística é tentar encontrar uma solução "boa" de maneira simples e rápida.

número de moradores. Assim, também objetiva-se propor uma medida para quantificar o aumento do risco em função da hierarquia e seu comportamento conforme o aumento do tamanho dos domicílios. Como mencionado no Capítulo 3, existem INE como o ONS do Reino Unido, por exemplo, que podem fazer a supressão de registros domiciliares que possuam número de moradores acima de um limite determinado.

Por fim, as estimativas do risco de revelação servirão ainda de insumos para o Capítulo 6 da tese, uma vez que elas indicam os possíveis registros alvo dos procedimentos de mascaramento de dados.

## 5.1. Risco de revelação

Revelação no contexto dos microdados está relacionada à possibilidade de identificação de um respondente dos microdados divulgados e, consequentemente, à possibilidade de se obter informações confidenciais sobre ele (HUNDEPOOL *et al*., 2012). Desta forma, a estimação do risco de revelação das unidades respondentes, e também do risco de revelação global para todo o conjunto de dados constituem importantes elementos do CEC (TEMPL, 2017).

Duncan *et al.* (2011) argumentam que antes de um INE disseminar qualquer microdado para uso público, é necessário avaliar o risco de um intruso comprometer a sua confidencialidade. Geralmente, se não é feito qualquer tipo de intervenção, traduzido na aplicação de certos procedimentos, os dados apresentam um risco de revelação inaceitavelmente alto. Por isso, é necessário tomar medidas para diminuir este risco para um determinado nível que seja considerado aceitável.

Um elemento foi elencado no parágrafo anterior ao se mencionar o risco de revelação do microdado: o intruso. Isto retoma a ideia de Willenborg e De Waal (2001), apresentada no Capítulo 2, de que o risco de revelação de um microdado não depende apenas de seu conteúdo, pois é preciso levar em consideração também o intruso e o tipo de informação que ele possa ter. Contudo, estes dois últimos são desconhecidos sendo, então, necessário fazer suposições sobre eles. Em outras palavras, é preciso definir o cenário de revelação e as variáveis-chave que serão utilizadas para a estimação do risco. Desta forma, "o risco de revelação é definido com base nas suposições do

cenário de revelação, ou seja, como o intruso pode se aproveitar dos dados divulgados para revelar informações sobre um entrevistado" (TEMPL, 2017, p.49).

Entretanto, vale ressaltar que alguns INE tomam medidas muito restritivas quanto ao conteúdo dos dados a serem disseminados, antes mesmo de qualquer avaliação de risco ou cenário de revelação. Hundepool *et al*. (2012), por exemplo, elencam as restrições à divulgação de microdados disponibilizados pelo *Statistics Netherlands*. Especificamente com relação aos microdados de uso público devem ser respeitadas um total de dez regras, quais sejam:

a) deve-se esperar ao menos um ano para divulgar os dados;
b) identificadores diretos, variáveis que indiquem qualquer localização geográfica diretamente (províncias da Holanda, por exemplo), nacionalidade, país de nascimento e etnia não podem ser divulgados;
c) somente uma variável que indique, de forma indireta, a localização geográfica (por exemplo, categorias de tamanho populacional do local de residência) pode ser divulgada. Ainda assim, as categorias dessa variável devem seguir algumas regras, como mínimo populacional em cada categoria, ente outras;
d) o número de variáveis identificadoras nos microdados podem ser no máximo quinze;
e) variáveis sensíveis não podem ser divulgadas;
f) não pode ser possível obter quaisquer informações adicionais que possam ajudar em algum tipo de identificação, a partir dos pesos amostrais divulgados;
g) ao menos duzentas mil pessoas na população devem estar contidas em cada categoria de uma variável identificadora;
h) ao menos mil pessoas na população devem estar contidas em qualquer cruzamento de categorias entre duas variáveis identificadoras;
i) deve haver pelo menos um total de cinco domicílios para todas as combinações de categorias das variáveis do domicílio. Essa regra se aplica aos domicílios com mais de um morador;
j) os registros devem ser divulgados com ordenação aleatória nos microdados.

Este é um exemplo de um conjunto muito restritivo de regras que certamente têm forte impacto na utilidade dos dados, mas reduzirão drasticamente o risco de revelação *a priori*.

Para outros tipos de conjuntos de dados como, por exemplo, aqueles de acesso restrito para pesquisadores, os autores elencam normas utilizadas pelo *Statistics Netherlands* muito mais brandas. De forma análoga, outros órgãos produtores de dados podem criar um conjunto de regras que visem garantir maior utilidade para as suas informações disseminadas publicamente. O importante deste exemplo é evidenciar que o planejamento para mitigar o risco de revelação se inicia bem antes da etapa de sua estimação. No caso desta tese, o arquivo de microdados já está disseminado, então esta etapa anterior não será abordada. Desta forma, a partir dos dados selecionados para o estudo, serão iniciadas as fases de seleção de cenários de revelação e variáveis-chave para a estimação do risco de revelação, o que não impede que os resultados encontrados sirvam de insumo para a posterior criação de regras para a disseminação de dados de uso público.

## 5.2. Cenários de revelação e variáveis-chave

Na seção anterior argumentou-se que, para avaliar o risco de revelação, é preciso assumir que a tentativa de tal revelação ocorre de acordo com um cenário especificado. Supõe-se, então, a existência de um intruso que tenta, de alguma forma, usar os microdados para revelar informações sobre os respondentes da pesquisa. Isto significa que normalmente avalia-se o risco de revelação condicionado a um cenário hipotético particular de ataque do intruso. Por isto, é sensato imaginar diferentes tipos de intrusos, que podem ter diferentes objetivos e tipos de informações (WILLENBORG e DE WAAL, 2001).

Benschop *et al.* (2019) acrescentam que os cenários de revelação também diferem dependendo de como será feita a disseminação dos dados. Há que se levar em conta, por exemplo, se os dados serão disponibilizados publicamente ou estarão restritos ao acesso por pesquisadores sob algum tipo de contrato de confidencialidade. O primeiro tipo demandará muito mais proteção, uma vez que o número de intrusos e

o tipo de informação que, pelo menos alguns destes, possam ter à disposição para uma possível tentativa de revelação será muito maior.

As análises e avaliações da abordagem apresentada nesta tese terão, por base, dados de uso público, justamente por sua maior vulnerabilidade ao ataque de intrusos. Entretanto, para exemplificar o caso de microdados de uso restrito, o potencial intruso seria o pesquisador que estivesse utilizando os dados sob licença, e o cenário de revelação mais plausível seria uma revelação espontânea por parte deste pesquisador. Benschop *et al.* (2019) definem como revelação espontânea, a possibilidade do intruso poder, de forma não intencional, reconhecer pelo menos um respondente enquanto estiver examinando os dados. Este é um risco ainda maior na presença de *outliers* ou combinações raras de respostas de variáveis-chave, como definidas na Seção 2.4.2. É preciso destacar que, embora este seja o cenário mais importante, por questão de segurança é possível considerar um cenário adicional em que o intruso tenha sob sua posse outros conjuntos de dados disponíveis publicamente e tente fazer alguma revelação utilizando-os.

No que diz respeito aos dados de uso público, é prudente estabelecer mais de um cenário de revelação, devido à sua maior exposição a potenciais intrusos. A literatura aponta dois caminhos principais a serem seguidos: supor que o intruso vai usar de alguma informação externa, como, por exemplo, outros conjuntos de dados previamente divulgados à procura de respondentes em comum; ou o conhecimento próprio sobre características das unidades respondentes que ele conheça para tentar fazer a revelação (TEMPL, 2017; BENSCHOP *et al*., 2019).

Duncan *et al.* (2011) argumentam que os cenários de revelação levarão a escolha das variáveis-chave a serem utilizadas. Em outras palavras, qualquer tentativa de revelação é feita com base em um conjunto de variáveis que estará disponível tanto ao intruso, quanto aquelas presentes no conjunto de dados alvo, permitindo que as unidades respondentes possam ser identificadas. Assim, de forma análoga ao que foi dito para os cenários de revelação, as variáveis-chave geralmente são oriundas de dois tipos de fontes: conjuntos de dados formais que contenham informações sobre a população alvo, ou conhecimento informal de ordem pessoal.

Benschop *et al.* (2019) acrescentam que o risco de revelação dependerá da inclusão ou exclusão das variáveis-chave escolhidas. Desta forma, o processo de seleção do conjunto destas variáveis deve, portanto, ser abordado com grande atenção e cuidado. Os autores sugerem, inclusive, que o primeiro passo que um órgão produtor de dados deveria fazer neste processo, seria o de realizar um inventário dos conjuntos de dados disponíveis no país. De posse desta lista, deveriam ser analisadas quais variáveis estariam contidas nestes bancos de dados e que poderiam ser utilizadas por potenciais intrusos. Estas informações são fundamentais para a decisão não só das variáveis-chave, mas também para o processo de CEC como um todo. É possível, por exemplo, supor que um cenário em que haja muita informação externa disponível para a utilização de um intruso em relação a um determinado arquivo de microdados alvo, faça com que os procedimentos usuais de CEC adotados por um órgão sejam aprimorados, visando garantir a confidencialidade das informações contidas em tais microdados.

### 5.2.1. Considerações para a definição de cenários de revelação e variáveis-chave

Para o banco de dados da PNAD Contínua referente ao 2º trimestre de 2019, serão utilizados os dois cenários de revelação, a partir das duas hipóteses destacadas na seção anterior: que o intruso se utilize de informações externas ou que ele faça uso de seu conhecimento próprio das unidades respondentes, neste caso as pessoas, para a tentativa de revelação. O intuito é que os procedimentos realizados, neste e nos próximos capítulos, sejam facilmente adaptáveis para as demais pesquisas amostrais domiciliares da Instituição.

Hundepool *et al*. (2012), dividem os processos do CEC para microdados em cinco principais etapas, a saber:

1. definição da necessidade de proteção à confidencialidade;
2. determinação das principais características e uso dos dados;
3. estimação do risco de revelação;
4. definição dos métodos de mascaramento a serem utilizados;
5. implementação de todo o processo.

Assim, antes de introduzir os dois cenários de revelação definidos – o que seria o processo inicial da terceira etapa – é preciso voltar a atenção para as duas etapas anteriores.

Na primeira etapa deve-se analisar quais são as unidades estatísticas e as variáveis presentes nos microdados a serem disseminados. No caso da PNAD Contínua, as unidades são os domicílios e os indivíduos, estes últimos são também as entidades respondentes. No Capítulo 1, foi exposto que o IBGE tem a obrigação legal de proteger a confidencialidade das informações dadas pelas pessoas naturais e jurídicas e utilizá-las exclusivamente para fins estatísticos. Já em relação à análise das variáveis dos microdados, é possível observar que a pesquisa, descrita no Capítulo 4, contém variáveis que podem ser consideradas sensíveis, especialmente pela grande quantidade de questões ligadas ao rendimento dos entrevistados.

Hundepool *et al*. (2012) sustentam que esta etapa serve para definir não apenas a imprescindibilidade da proteção à confidencialidade dos dados, mas também se é evidenciada a necessidade de algum tipo de tratamento nas informações. No caso da PNAD Contínua, diante do exposto no parágrafo anterior, fica claro a importância quanto à proteção da confidencialidade dos dados. Por outro lado, como o arquivo de dados a ser trabalhado já está disseminado, não cabe mais qualquer tipo de tratamento nas informações. Contudo, a título de ilustração, seria possível argumentar que as variáveis de rendimento muitas vezes apresentam *outliers* que podem aumentar o risco de revelação da unidade respondente. Esse problema é ainda maior quando tais variáveis estão conjugadas a uma variável de localização geográfica mais desagregada. Na pesquisa que será trabalhada, o menor nível geográfico de divulgação são os municípios das capitais das UF, que contêm um contingente populacional de tamanho razoavelmente grande. Ainda assim, uma análise dos *outliers* neste menor nível de desagregação, poderia indicar casos específicos de registros que necessitassem de algum tipo de tratamento.

A segunda etapa, segundo Hundepool et al. (2012), definida como a determinação das principais características e uso dos dados, pode ser subdividida em quatro processos. O primeiro seria referente a uma análise da metodologia da pesquisa e do questionário. O objetivo seria ter um entendimento da estrutura dos dados, para

já começar a dimensionar o nível de proteção requerido (dados censitários podem exigir medidas adicionais em relação a dados de pesquisas amostrais, por exemplo), uma possível lista de variáveis indicadoras a serem removidas (nesta fase excluem-se as indicadoras diretas e dependendo do tipo de disseminação do arquivo, outras poderiam ser consideradas, por exemplo), e análise das relações entre variáveis (limites de idade para resposta a itens de trabalho, por exemplo, ou mesmo relações diretas por meio das variáveis derivadas divulgadas, tais como anos de estudo, rendimento domiciliar *per capita*, dentre outras) a fim que possíveis perturbações posteriores nos dados não gerem inconsistências.

O segundo processo seria uma análise da necessidade dos usuários. De forma sintética, uma lista de prioridades de quais variáveis e seus níveis de detalhamento que deveriam ser incluídos na divulgação, tipos de análise estatísticas que geralmente são feitas nos dados e níveis de disseminação necessários, sob o ponto de vista dos usuários. O IBGE atualmente já faz consultas aos usuários externos acerca de questões temáticas em algumas de suas pesquisas como, por exemplo, no Censo Demográfico (IBGE, 2018b).

O terceiro processo seria utilizar as informações da etapa anterior, em conjunto com a política de disseminação de informações do órgão produtor, para definir como será feita esta disseminação. Para atender às necessidades de diferentes usuários, muitos órgãos adotam a política de produzir diferentes tipos de produtos a partir dos mesmos microdados de pesquisa. Por fim, o quarto processo refere-se à análise do plano e dos tipos de disseminação dos dados. Os autores argumentam que, geralmente antes da liberação dos microdados, resultados da pesquisa já estão disponíveis sob forma de tabelas ou outro tipo de informações agregadas. É preciso que os microdados – apesar de possíveis restrições que venham a sofrer – ao serem liberados, sejam coerentes com as informações já disponíveis previamente. Em outras palavras, que seja possível recalcular estas informações prévias, o tanto quanto possível, a partir dos microdados que vierem a ser disseminados (HUNDEPOOL *et al*., 2012).

Definidas estas questões, a próxima etapa, como já mencionado anteriormente, corresponderia à estimação do risco de revelação que, por sua vez, depende da definição dos cenários de revelação e variáveis-chave a serem selecionadas. Estas

questões serão abordadas utilizando a PNAD Contínua referente ao 2º trimestre de 2019 para ilustrar.

### 5.2.2. Cenários de revelação e variáveis-chave propostos para a PNAD Contínua

Para a estimação do risco de revelação do arquivo de dados escolhido, tendo em vista que se trata de microdados de uso público, optou-se por utilizar dois cenários de revelação, assim como mencionado na Seção 5.2.1. No primeiro caso, o intruso se utiliza de informações externas e no segundo ele faz uso de seu conhecimento próprio sobre as unidades respondentes.

**Cenário de Revelação 1**: com relação ao primeiro cenário (que será chamado de cenário 1), supõe-se que o intruso utilizará uma base de dados em sua posse. Esta base de dados poderá tanto ser oriunda de dados disponibilizados publicamente por quaisquer órgãos, quanto de dados coletados de forma privada por ele, ou a que ele tenha acesso (BENSCHOP *et al*., 2019). No que diz respeito aos dados públicos, é teoricamente possível mapear todas as informações disponíveis. Já para dados particulares, essa tarefa dificilmente será exequível, mas algumas hipóteses podem ser assumidas e serão exemplificadas oportunamente.

É importante ressaltar que alguns autores sugerem que se considere como dados públicos disponíveis ao intruso, os que possuam identificadores diretos como, por exemplo, nomes ou endereços (HUNDEPOOL *et al*., 2012). Outros argumentem que "tipicamente" os dados terão estes tipos de identificadores, mas não é uma condição necessária (BENSCHOP *et al*., 2019). Há que se ter em conta que o intruso pode ter à sua disposição uma combinação de dados particulares com esses identificadores diretos já combinados com informações disponíveis publicamente complementares, para a realizar sua tentativa de revelação.

Na literatura, dados públicos com identificadores diretos usualmente citados são, por exemplo, alguns registros administrativos, listas eleitorais e registros cadastrais de proprietários de terra (BENSCHOP *et al*., 2019). É claro que a disponibilidade dos mesmos varia de acordo com o país. No Brasil, geralmente os dados disseminados publicamente têm seus identificadores diretos removidos. Um exemplo é a base de

dados da Relação Anual de Informações Sociais (RAIS), na qual está presente o nome e CPF do trabalhador, mas estas informações são omitidas nos microdados disponíveis ao público. Curiosamente há, no entanto, informações divulgadas com identificadores diretos acompanhadas de variáveis sensíveis, como é o caso do "Portal da Transparência" do governo federal, no qual é possível consultar a remuneração de servidores do executivo federal na ativa e aposentados, assim como pensionistas. Neste caso, embora a revelação da variável de rendimento já esteja dada, o banco de dados não contém outras variáveis indicadoras (idade, raça/cor, etc) para utilização em possíveis tentativas de revelação futuras.

Em relação aos dados privados, observa-se que com o grande avanço tecnológico observado nas últimas duas décadas – principalmente com o advento da internet – as grandes corporações detêm cada vez mais informações pessoais de seus usuários e consumidores. Informações de todos os tipos podem estar contidos nestas bases de dados: documentos como RG e CPF, endereço residencial e de trabalho, informações financeiras, hábitos de consumo, entre muitos outros. Embora o governo brasileiro tenha sancionado a Lei 13.709/2018, que dispõe sobre a proteção de dados pessoais, vazamentos de informações podem acontecer. É possível citar os exemplos da base de dados dos usuários da rede de videogames *Playstation Network*, propriedade da empresa Sony em 2011, ou da base de usuários do Yahoo em 2016, ambas disseminadas por *hackers* (SANSANA, 2018). Desta forma, não há como descartar que existam inúmeras bases de dados privadas contendo variáveis identificadoras diretas com, pelo menos, algumas outras variáveis indicadoras mais comuns em quaisquer cadastros pessoais (localização geográfica/endereço, sexo, etc) com potencial risco de utilização por intrusos.

Antes, porém, de analisar as bases de dados que possam servir de informações externas para um intruso vincular suas informações com a PNAD Contínua, é preciso observar primeiro quais variáveis identificadoras esta última possui que são passíveis de tal vínculo.

Como descrito no Capítulo 4, o arquivo de dados a ser trabalhado é dividido em cinco partes mais uma seção de variáveis derivadas. Na parte 1 (identificação e controle) é possível obter variáveis de localização geográfica do registro, na parte 3

(características de educação) o próprio nível de instrução da pessoa pode ser utilizado como identificador, as partes 4 e 5 (características de trabalho e rendimentos de outras fontes, respectivamente) concentrariam as variáveis sensíveis. Assim é na parte 2 (características gerais dos moradores) que estariam incluídas a maioria das variáveis identificadoras. Entretanto essa parte é pequena, com apenas 9 variáveis como apresentado na Tabela 4.2, e nem todas as variáveis trazem informações relevantes. Por exemplo, uma delas se refere ao "número de ordem da pessoa no domicílio" e quatro delas compõem a informação de idade do indivíduo (dia, mês e ano de nascimento e a própria idade declarada). A Tabela 5.1 apresenta as variáveis contidas na parte 2 da PNAD Contínua do 2º trimestre de 2019.

Tabela 5.1: Variáveis contidas na parte 2 – características gerais dos moradores – da PNAD contínua do 2º trimestre de 2019, presentes nos microdados de uso público.

| Código da variável | Descrição do quesito |
|---|---|
| V2001 | Número de pessoas no domicílio |
| V2003 | Número de ordem |
| V2005 | Condição no domicílio |
| V2007 | Sexo |
| V2008 | Dia de nascimento |
| V20081 | Mês de nascimento |
| V20082 | Ano de nascimento |
| V2009 | Idade do morador na data de referência |
| V2010 | Cor ou raça |

Fonte: PNAD Contínua 2019 – 2º trimestre.

A busca de bases de dados públicas que possam ser utilizadas por um intruso, como já mencionado no início deste capítulo, é uma etapa integrante do processo de estimação do risco de revelação. É proposto, nesta tese, que esta busca leve em consideração aspectos de escopo, atualização e abrangência geográficas dos dados, detalhadas a seguir, e que podem ser utilizadas para quaisquer outras pesquisas.

A primeira questão é relativa ao escopo dos dados. Por exemplo, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza os microdados sobre o Censo Escolar da Educação Básica anualmente. Embora sejam perguntadas questões como idade, sexo, cor ou raça, entre outras que possuem categorias iguais ou que possam ser harmonizáveis com a PNAD Contínua, a população alvo da pesquisa é referente aos alunos que cursam até o ensino médio ou técnico. Na maioria dos casos, estas pessoas não terão trabalho ou rendimento, sendo possível supor que um intruso teria interesse muito reduzido em utilizar esta base de dados. Por outro lado, a já mencionada RAIS que é "uma das fontes estatísticas mais confiáveis sobre o mercado de trabalho formal... constitui referência nacional e internacional sendo considerado um verdadeiro Censo" (MINISTÉRIO DA ECONOMIA, 2021, p.5), teria muito mais atratividade para um intruso dado o seu escopo.

Uma segunda questão importante é referente à atualização dos dados. Um intruso poderia, por exemplo, se utilizar do Censo Demográfico como fonte auxiliar, porém como sua periodicidade é decenal, seu uso seria muito restrito. A já citada RAIS tem periodicidade anual, e existem fontes lançadas em intervalos de tempo ainda menores como o Cadastro Geral de Empregados e Desempregados (CAGED) que é mensal. Este último, por conter informações essenciais "para elaboração de Políticas de Emprego e Salário, bem como pesquisas e estudos sobre mercado de trabalho" (MINISTÉRIO DO TRABALHO E EMPREGO, 2021, p.6), também possui um escopo temático similar à PNAD Contínua.

Por fim, uma terceira questão levantada foi a abrangência geográfica dos dados disponíveis. Um exemplo é a Pesquisa do Emprego e Desemprego (PED) feita em conjunto pelo Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE) e a Fundação Seade. Esta pesquisa teve início em 1984 abrangendo apenas a Região Metropolitana de São Paulo, com periodicidade mensal. A partir de 1987, a PED começou a englobar também outras regiões como o Distrito Federal, e as Regiões Metropolitanas Porto Alegre, Belo Horizonte, Salvador, Recife e Fortaleza. Em alguns momentos a pesquisa nessas localidades foi descontinuada por falta de financiamento, culminando com o levantamento sendo apenas realizado no Distrito Federal em 2020 (DIEESE, 2021). Desta forma, seu uso para tentativa de revelação não seria inviabilizado,

mas limitado à sua abrangência geográfica, ao contrário dos dados da RAIS e CAGED, por exemplo. Caso seja importante considerar uma base com este tipo de limitação, é possível contornar esse problema uma vez que a estimação do risco de revelação pode ser feita considerando os domínios geográficos divulgados na pesquisa.

Levou-se em conta as questões definidas para a análise das bases de dados públicas disponíveis, contudo a quantidade limitada de variáveis indicadoras presentes na parte 2 da PNAD Contínua – e que são comuns à maioria das bases externas encontradas – faz com que a definição das variáveis-chave seja feita de forma quase direta. Entretanto, as categorias presentes nestas variáveis podem diferir entre pesquisas e há a necessidade de verificar se é necessário algum tipo de compatibilização. Assim, levando-se em conta as questões de escopo, atualização e abrangência geográfica, a escolha das variáveis-chave tiveram como base os seguintes bancos de dados externos: os já mencionados CAGED e RAIS[7], o Cadastro Único (CadÚnico) e o Censo da Educação Superior (Censo Superior).

O CadÚnico é um instrumento para identificação, caraterização e inclusão de famílias de baixa renda em programas sociais do governo federal. Atualmente, o CadÚnico contém informações de mais de 80 milhões de pessoas (SAMBIASE, 2020) e sua periodicidade de divulgação é anual. Já o Censo Superior é uma pesquisa estatística operacionalizada pelo INEP, tendo sua declaração realizada em todo território nacional, via internet, por meio de coleta de dados descentralizada. Os respondentes são todas as instituições de ensino superior públicas e privadas do país contendo, dentre outras informações, dados sobre todos os alunos nelas matriculados (INEP, 2021).

Analisando estas bases de dados citadas, conjuntamente com a PNAD Contínua selecionada, observou-se total compatibilidade para as categorias das variáveis “sexo”, “idade” e “raça/cor”. Com relação à variável educacional, há 11 classes possíveis tanto para a RAIS quanto para o CAGED. Neste caso seria uma informação mais agregada se comparada à PNAD Contínua, dado que a variável derivada “anos de estudo” possui 17 classes (0 a 16 anos de estudo). Entretanto, há um conjunto de perguntas sobre

[7] Existe uma outra base de dados pública harmonizada com informações conjuntas do CAGED, RAIS e da Carteira de Trabalho da Previdência Social (CTPS). Entretanto, justamente por conta da harmonização, as categorias de algumas variáveis são agregadas, ou seja, com menos informação.

instrução no CadÚnico, que torna possível replicar a variável mencionada da PNAD Contínua com as mesmas categorias. Processo análogo pode ser feito no Censo Superior, com a diferença que, neste caso, o número de categorias efetivamente observadas seria restrito em função da população alvo da pesquisa. Assim, partindo do pressuposto que o intruso tenha a informação mais desagregada possível, o Quadro 5.1 apresenta as variáveis-chave e suas respectivas categorias selecionadas para o cenário de revelação 1.

Quadro 5.1: Variáveis-chave e suas categorias selecionadas para o cenário de revelação 1.

| Código da variável | Descrição | Categorias | Número de categorias |
|---|---|---|---|
| V2007 | Sexo | Homem; Mulher | 2 |
| V2009 | Idade | 0 a 130 | 131 |
| V2010 | Cor/raça | Branca; Preta; Amarela; Parda; Indígena; Ignorado | 6 |
| VD3005 | Anos de estudo | 0 a 16 + não aplicável | 18 |

Fonte: PNAD Contínua 2019 – 2º trimestre.

De acordo com o Quadro 5.1, observa-se que há uma categoria “ignorado” na variável cor/raça. Isso implica que o intruso não conseguiria fazer a revelação do registro com base nessa variável, ou seja, seria equivalente a uma supressão local prévia. Ainda assim a revelação pode ser feita com base nas demais variáveis-chave. Já a categoria “não aplicável” para a variável anos de estudo refere-se às pessoas com menos de 5 anos de idade. É possível assumir que este grupo possui zero anos de estudo, mas, na prática, dificilmente um intruso teria interesse em fazer a revelação destas pessoas utilizando uma pesquisa focada nos temas de trabalho e rendimento. Outro aspecto a ser mencionado é que as variáveis relativas à idade e anos de estudos são discretas, mas para fins deste trabalho terão seus valores considerados como categorias (por exemplo: a categoria “0” para “zero anos de estudo”, “1” para “um ano de estudo, e assim por diante).

**Cenário de Revelação 2**: Com relação ao segundo cenário de revelação (que será chamado de cenário 2), o intruso faz uso de informação própria, tentando buscar no banco de dados da PNAD Contínua alguém de seu círculo de conhecidos. Neste caso, é preciso supor que ele conheça minimamente a estrutura domiciliar, ou seja, ter ideia das pessoas que compõem o(s) domicílio(s) alvos de revelação. O Quadro 5.2 apresenta as variáveis-chave e suas respectivas categorias selecionadas para o cenário de revelação 2.

Quadro 5.2: Variáveis-chave e suas categorias selecionadas para o cenário de revelação 2.

| Código da variável | Descrição | Categorias | Número de categorias |
|---|---|---|---|
| V2007 | Sexo | Homem; Mulher | 2 |
| V2009 | Idade | 0 a 130 | 131 |
| V2010 | Cor/raça | Branca; Preta; Amarela; Parda; Indígena; Ignorado | 6 |
| VD3004 | Nível de instrução | Menos de 1 ano de estudo; fundamental completo; fundamental incompleto; médio incompleto; médio completo; superior incompleto; superior completo; não aplicável | 8 |
| V2001 | Tamanho do domicílio | 1 a 30 | 30 |

Fonte: PNAD Contínua 2019 – 2º trimestre.

Como pode ser observado no Quadro 5.2, partiu-se da suposição que o intruso conhecesse as informações mais elementares de seu círculo de conhecidos como sexo, idade e cor/raça. Na questão da instrução, optou-se por utilizar a variável derivada mais resumida na PNAD Contínua que contém apenas sete categorias. Ao contrário do Cenário 1 em que o intruso pode obter a informação exata da escolaridade, em termos

de anos de estudo, por fontes externas, aqui supõe-se que ele tenha conhecimento desta informação, mas não necessariamente com total exatidão.

As estimativas do risco de revelação baseadas nas variáveis-chave para ambos os cenários foram calculadas para todos os recortes geográficos divulgados pela PNAD Contínua: Brasil, Grandes Regiões, Unidades da Federação, Regiões Metropolitanas que contém municípios das capitais e todos os municípios de capitais. Ressalta-se ainda que, seguindo a recomendação de Hundepool et al. (2012), qualquer suposição deve ser conservadora, implicando presumir o pior cenário possível para o órgão produtor dos dados. Então, considera-se que tanto a base da PNAD Contínua, quanto a base externa do intruso são livres de quaisquer tipos de erros, ou seja, não é possível fazer uma falsa revelação. Assim, o risco de revelação estimado será equivalente ao da situação mais adversa possível.

## 5.3. Estrutura hierárquica da PNAD Contínua

Benschop *et al*. (2019) argumentam que geralmente os microdados possuem uma estrutura hierárquica de forma que unidades individuais pertençam a algum grupo. O tipo mais comum de estrutura observada nas pesquisas amostrais domiciliares é o de pessoas que pertencem a determinado domicílio. Desta forma, é muito importante levar em conta essa estrutura hierárquica na estimativa do risco de revelação. Supõe-se que se um indivíduo for identificado, a estrutura domiciliar vai permitir a identificação dos outros moradores daquele domicílio.

Com base nesta premissa, Templ (2017) aponta que o risco de revelação para o domicílio – ou risco hierárquico – pode ser calculado subtraindo-se de 1, a probabilidade de nenhuma pessoa ter sido identificada no domicílio. Assim, pessoas na mesma unidade hierárquica terão o mesmo risco hierárquico (ou, equivalentemente, o mesmo risco de revelação do domicílio).

A Tabela 5.2 apresenta um exemplo considerando-se um domicílio de quatro moradores. Na coluna 4, o risco da pessoa (ou indivíduo) $I_j$ do domicílio $D_j$ não ser identificada é dado por: $(1 – R_j)$, com j variando de 1 a 4. Portanto o risco de revelação

do domicílio ou risco hierárquico (coluna 5) é dado por: $(1 – [(1-R_1). (1-R_2). (1-R_3). (1-R_4)])$.

Duncan *et al.* (2011) acrescentam ainda que este aumento no risco de revelação individual causado pela estrutura hierárquica pode ser particularmente grande para domicílios com muitos moradores. Como a PNAD Contínua possui esta estrutura hierárquica e aceita, teoricamente, um tamanho domiciliar de até 30 moradores, este aspecto será levado em conta na estimativa dos riscos de revelação.

Tabela 5.2: Risco de revelação hierárquico estimado, a partir dos riscos pessoais considerando-se um domicílio hipotético de quatro pessoas.

| Pessoa ($I_i$) | Domicílio ($D_j$) | Risco da pessoa ser identificada ($R_i$) | $(1 - R_i)$ | Risco Hierárquico ($R_h$) |
|---|---|---|---|---|
| $I_1$ | $D_1$ | 0,0016639 | 0,9983361 | 0,0043309 |
| $I_2$ | $D_1$ | 0,0016639 | 0,9983361 | 0,0043309 |
| $I_3$ | $D_1$ | 0,0005552 | 0,9994448 | 0,0043309 |
| $I_4$ | $D_1$ | 0,0004543 | 0,9995457 | 0,0043309 |

Fonte: Adaptado de Templ (2017).

## 5.4. Domínios de análise dos resultados

Como previamente comentado na Seção 5.2, os resultados apresentados nesta tese, para ambos os cenários de revelação, abrangem todos os recortes geográficos divulgados pela PNAD Contínua. Além do total nacional, estão contempladas as Grandes Regiões, as Unidades da Federação, as Regiões Metropolitanas que contém municípios das capitais e todos os municípios de capitais. Estas informações estão dispostas em três variáveis do banco de dados, como mostra o Quadro 5.3

Os Quadros A1 até A3 do Anexo A contém os rótulos das categorias de todas as variáveis contidas no Quadro 5.3. A variável UF indica em qual das vinte e sete Unidades da Federação reside o respondente. Desta forma, a Grande Região (Norte, Nordeste, Sudeste, Sul ou Centro-Oeste) pode ser obtida diretamente com base no primeiro dígito desta variável.

Quadro 5.3: Variáveis-chave e suas categorias selecionadas para o cenário de revelação 1.

| Código da variável | Descrição | Categorias | Número de categorias |
|---|---|---|---|
| UF | Unidade da Federação | 11; 12; 13; 14; 15; 16; 17; 21; 22; 23; 24; 25; 26; 27; 28; 29; 31; 32; 33; 35; 41; 42; 43; 50; 51; 52; 53 | 27 |
| Capital | Município da Capital | 11; 12; 13; 14; 15; 16; 17; 21; 22; 23; 24; 25; 26; 27; 28; 29; 31; 32; 33; 35; 41; 42; 43; 50; 51; 52; 53; *missing* | 28 |
| RM_RIDE | Região Metropolitana e Região Administrativa Integrada de Desenvolvimento | 13; 15; 16; 21; 22; 23; 24; 25; 26; 27; 28; 29; 31; 32; 33; 35; 41; 42; 43; 51; 52; *missing* | 22 |

Fonte: PNAD Contínua 2019 – 2º trimestre

A variável Capital, por sua vez, é indicadora se o registro pertence a um respondente residente no município da capital da UF. Em caso afirmativo, a variável tem o mesmo valor da UF, caso contrário não há informação (*missing*). Por exemplo, um registro proveniente de um residente do município de Porto Velho-RO tem a categoria "11" presente nesta variável, para um residente do município de Rio Branco-AC o valor é "12" e assim por diante. Desta maneira, um usuário pode identificar o município da capital dentro da UF correspondente, possibilitando a análise deste domínio.

Entretanto, sob a ótica do intruso a não informação (*missing*) também é de valia, uma vez que ela indica, por exclusão, uma informação extra: a de que o registro pertence a um respondente residente em um município fora da capital. Dado que o intruso pode facilmente recompor este recorte geográfico, ele também terá que ser levado em conta na estimação do risco de revelação, mesmo que a PNAD Contínua não divulgue resultados neste nível.

Situação semelhante verifica-se na variável RM_RIDE que é indicadora do registro pertencer a um respondente residente em município de Região Metropolitana (RM) ou Região Administrativa Integrada de Desenvolvimento (RIDE). Em caso afirmativo, a variável tem o mesmo valor da UF, caso contrário não há informação (*missing*), de forma análoga ao que foi descrito para a variável Capital. A diferença é que seis UF não possuem RM ou RIDE. Assim para as UF 11 (Rondônia), 12 (Acre), 14 (Roraima), 17 (Tocantins), 50 (Mato Grosso do Sul) e 53 (Distrito Federal), todos os valores da variável RM_RIDE serão *missing*.

Contudo, para as outras vinte e uma UF, o intruso pode se utilizar da não informação e criar um recorte geográfico adicional, para localizar o respondente como residente fora da RM ou RIDE. Então, mesmo que este novo recorte não seja divulgado na PNAD Contínua também tem que ser levado em consideração na análise dos resultados desta tese.

Estes dois casos ilustram um ponto que deve ser levado em conta na hora da disseminação de dados de uso público: ao se colocar no banco de dados alguma variável que possa indicar determinado recorte geográfico previamente planejado para divulgação, é possível que outro recorte não planejado seja também divulgado indiretamente? Em caso afirmativo, há problemas quanto à confidencialidade das informações com essa divulgação?

Logo, com a incorporação destes dois novos recortes geográficos, os resultados desta tese contemplam sete domínios de análise distintos, quais sejam:

- Brasil;
- Grandes Regiões;
- Unidades da Federação;
- RM ou RIDE que contém municípios das capitais;
- Municípios fora destas RM ou RIDEu;
- Municípios das capitais;
- Municípios fora das capitais.

## 5.5. Unicidades Amostrais

Nesta seção será trabalhado o conceito de unicidade, que pode ser entendido como um registro com valor de chave, definido na Seção 2.4.2, que seja único no arquivo de dados. A presença de unicidade em um arquivo de dados proveniente de um Censo deve sempre ser evitada, uma vez que um intruso saberá que se trata de um valor de chave único na população e, consequentemente, poderá identificar a entidade. No caso de dados oriundos de pesquisas amostrais, um valor de chave único pode não ser problema, pois não necessariamente ele também será único na população. No entanto, Duncan *et al.* (2011), alertam que a unicidade amostral é uma pré-condição necessária para a unicidade populacional. Os autores ainda destacam que a literatura é rica a respeito de métodos que utilizam informações amostrais para fazer inferências sobre unicidades populacionais. Estes métodos serão abordados com maior detalhamento na Seção 5.6.

Dado que um subconjunto das unicidades amostrais pode também ser unicidade populacional, é intuitivo perceber que, como argumenta Templ (2017), um registro com valor de chave único é considerado mais propenso a ser identificado por um intruso. É preciso destacar que quanto menor o tamanho da amostra, maior tenderá a ser o número de unicidades amostrais encontradas. Adicionalmente, quanto maior a fração amostral maior tenderá a ser a probabilidade que uma unicidade amostral seja também unicidade populacional. A Tabela 5.3 apresenta o número de unicidades amostrais, em registros de pessoas, da PNAD 2019 – 2º trimestre, para os cenários de revelação e recortes geográficos previamente especificados na Seção 5.2.

Com base na Tabela 5.3, é possível observar que o número de registros de pessoas com valores de chave únicos é maior quando se considera o Cenário 2. Isto é um indicativo que este cenário de revelação pode apresentar mais registros considerados não seguros para disseminação, situação esta que será avaliada na Seção 5.6. Em relação aos recortes geográficos verifica-se, como esperado, que as unicidades amostrais aumentam à medida que há a desagregação no recorte a ser disseminado.

Tabela 5.3: Unicidades amostrais nos registros de pessoas da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e recorte geográfico.

| Recorte Geográfico | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| Brasil | 551.348 | 1.768 | 0,3 | 7.480 | 1,4 |
| Grandes Regiões | 551.348 | 8.516 | 1,5 | 27.438 | 5,0 |
| Unidades da Federação | 551.348 | 43.796 | 7,9 | 96.902 | 17,6 |
| RM ou RIDE | 170.401 | 29.629 | 17,4 | 56.219 | 33,0 |
| Municípios fora de RM ou RIDE | 323.797 | 31.418 | 9,7 | 63.483 | 19,6 |
| Municípios de Capitais | 126.336 | 30.045 | 23,8 | 55.540 | 44,0 |
| Municípios fora de capitais | 425.012 | 39.165 | 9,2 | 80.240 | 18,9 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre.

As Tabelas A1, A3, A5, A7, A9 e A11 do Anexo A, trazem um maior detalhamento destes números ao discriminar as unicidades amostrais para cada unidade territorial dentro de cada recorte geográfico considerado. É possível observar uma grande heterogeneidade nos valores dentro dos recortes geográficos. Tomando-se como exemplo a Tabela A3, verifica-se que embora 17,6% dos registros de pessoas tenham valor de chave único se considerarmos a divulgação por UF no Cenário 2, esse valor dentro de cada UF varia de 10,0% em Minas Gerais, até 48,6% no Amapá.

Outra análise pertinente é comparar comportamentos distintos entre recortes geográficos diferentes. A Tabela 5.3 mostra que, para o Cenário 2, 33,0% dos registros têm valores de chave únicos considerando a disseminação da RM/RIDE contra apenas 19,6% quando esse registros se localizam fora da RM/RIDE (recorte que pode ser deduzido pelo intruso). Entretanto, ao se comparar as Tabelas A5 e A7 do Anexo A, essa situação se inverte em algumas UF. Tomando-se como exemplo a UF Amapá, observa-se que o percentual de unicidades amostrais é de 53,9% dentro da RM/RIDE e de 69,4% fora dela. Isto ocorre, pois, a RM Macapá contém a maior parte da população da UF, assim como a maioria dos registros na amostra da PNAD Contínua no Amapá. Isto

reforça o que foi mencionado na Seção 5.4, sobre a possibilidade da divulgação de determinado recorte geográfico, implicar que outro recorte não planejado seja de possível dedução por um intruso. Neste caso um domínio menor (municípios no Amapá fora da RM Macapá) possa ser deduzido ao se divulgar um domínio maior (municípios no Amapá dentro da RM Macapá).

Outro ponto a se destacar é que embora as Tabelas A5, A7, A9 e A11 apresentem unidades territoriais com unicidades amostrais em mais de metade dos registros de pessoas consideradas pelo Cenário 2, este resultado ainda não leva em conta a estrutura hierárquica da pesquisa. A partir da hipótese descrita na Seção 5.3 de que se um indivíduo for identificado, a estrutura domiciliar vai permitir a identificação dos outros moradores daquele domicílio, então, idealmente, deve-se utilizar os registros de domicílio como unidade de análise. Assim, neste nível hierárquico se considera unicidade como sendo os domicílios que possuam pelo menos um registro de pessoa com valor de chave único na amostra. A Tabela 5.4 apresenta o número de unicidades amostrais, tendo como base os registros de domicílios, para os cenários de revelação e recortes geográficos especificados nas Seções 5.2 e 5.4, respectivamente.

Com base na Tabela 5.4, é possível observar que o número percentual de unicidades aumenta em comparação aos da Tabela 5.3, como esperado, para todos os recortes geográficos. Para o Cenário 2, em dois destes recortes – municípios em RM/RIDE e municípios de capitais – verifica-se que a maior parte dos domicílios (56,9% e 70,9%, respectivamente) contém pelo menos uma unicidade amostral dentro dele. De forma análoga à análise dos registros de pessoas, as Tabelas A2, A4, A6, A8, A10 e A12 do Anexo A, detalham as informações da Tabela 5.4, ao discriminar o quantitativo de domicílios com unicidades amostrais para cada unidade territorial, dentro de cada recorte geográfico considerado.

Os registros de domicílios, de acordo com as tabelas do Anexo A, também apresentam grande heterogeneidade nos valores da proporção de unicidades dentro dos recortes geográficos. Por exemplo: para o Cenário 2, a Tabela 5.4 mostrou que 34,3% dos domicílios apresentam pelo menos um morador com valor de chave único ao se considerar a divulgação por UF. Entretanto, como pode ser observado pela Tabela A4, estes valores variam entre 19,5% para Santa Catarina, até 84,5% para o Amapá. Desta

forma, mesmo que este recorte geográfico divulgado pela pesquisa ainda não seja o mais desagregado, verificam-se a existência de UF em que a maior parte dos domicílios na amostra contém pelo menos uma unicidade. Ao nível de divulgação mais desagregado – municípios da capital – apenas o Rio de Janeiro possui menos da metade dos domicílios (45,4%) com pelo menos uma unicidade. Dentre as outras vinte e seis capitais, esse valor é superior a 80% para quinze delas e maior que 90% para quatro municípios, como mostra a Tabela A10. Entretanto, o maior valor encontrado foi para outro recorte geográfico, uma vez que 97,3% dos registros de domicílios em municípios fora de RM/RIDE no Amapá possuem pelo menos um morador com valor de chave único na amostra, como se observa na Tabela A8. Este recorte, como já mencionado anteriormente, embora não tenha resultados disseminados na PNAD Contínua, pode ser facilmente deduzido por um intruso.

Tabela 5.4: Unicidades amostrais nos registros de domicílios da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e recorte geográfico.

| Recorte Geográfico | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| Brasil | 187.061 | 1.635 | 0,9 | 5.330 | 2,9 |
| Grandes Regiões | 187.061 | 7.578 | 4,1 | 19.550 | 10,5 |
| Unidades da Federação | 187.061 | 36.252 | 19,4 | 64.231 | 34,3 |
| RM ou RIDE | 59.691 | 22.274 | 37,3 | 33.965 | 56,9 |
| Municípios fora de RM ou RIDE | 108.783 | 25.538 | 23,5 | 41.864 | 38,5 |
| Municípios de Capitais | 43.994 | 21.270 | 48,3 | 31.180 | 70,9 |
| Municípios fora de capitais | 143.067 | 31.691 | 22,2 | 52.681 | 36,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre.

Deve ser ressaltado que é provável que a maioria das unicidades amostrais não sejam unicidades populacionais, ao se levar em conta que a fração amostral da pesquisa é menor que 0,3% como apresentado na Seção 4.3. Entretanto, a análise das unicidades

amostrais mostra, especialmente para o Cenário 2 e recortes geográficos mais desagregados, um grande número, tanto absoluto quanto relativo, de valores de chaves único da amostra da pesquisa, o que requer atenção especial.

Este tipo de análise inicial pode ser especialmente importante em pesquisas que estão incorporando pela primeira vez técnicas de CEC, ou que planejem implementar alguma mudança (divulgação de novo recorte geográfico, de novas variáveis que possam ser usadas como indicadoras etc.). Outra possibilidade é conjugar a análise das unicidades amostrais com as de *outliers* em variáveis sensíveis. Neste caso, um valor atípico em um registo com valor de chave único, pode ser considerado não seguro para divulgação e necessitar de algum tipo de tratamento prévio.

Em termos operacionais, visando a implementação em um processo de produção de pesquisa, estes resultados foram obtidos utilizando o *software* R, por meio da função ”freqCalc” do pacote “sdcMicro”, em poucos segundos de processamento utilizando-se um computador pessoal (Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz, com 8GB de memória RAM). A função ”freqCalc” retorna, para cada registro, o número de valores de chaves idênticos encontrados no arquivo de dados. As unicidades representam os valores iguais a 1 encontrados, mas ao se trabalhar com Censos ou pesquisas com frações amostrais maiores, pode ser interessante mapear outros valores de chave com frequência pequena (abaixo de um limite k pré-determinado). Os programas em R utilizados nesta tese estão dispostos no Apêndice.

## 5.6. Estimação do risco de revelação

A partir dos cenários de revelação e variáveis-chave definidos na Seção 5.2 e levando-se em consideração a estrutura hierárquica descrita na Seção 5.3, o próximo passo é estimar o risco de revelação para os microdados da PNAD Contínua referente ao 2º trimestre de 2019. Como previamente descrito na Seção 2.6, existem duas abordagens para se calcular o risco de revelação: a individual para cada registro, ou a global para o arquivo como um todo. Nesta tese, a ênfase se dará na questão do risco individual, uma vez que se pretende localizar registros com probabilidade de identificação acima de um nível máximo a ser especificado. Duncan *et al.* (2011)

classificam os métodos de estimação do risco de revelação individual em dois grandes grupos: abordagens por modelos probabilísticos e abordagens que utilizem medidas heurísticas. Ambos serão vistos com maiores detalhes neste capítulo. O risco global do arquivo, por sua vez, pode ser estimado por métodos que utilizem os riscos individuais.

### 5.6.1. Estimação do risco de revelação por modelos probabilísticos.

Duncan *et al.* (2011) argumentam que registros que apresentam valores de chave incomuns ou raros na população têm alto risco de revelação, mas valores raros ou mesmo únicos na amostra não correspondem necessariamente a registros de elevado risco. A questão a ser respondida é então: como estimar os riscos de revelação com base nas observações amostrais? Para formular uma resposta a esta pergunta, os autores consideram $F_k$ como o número de registros na população com valor de chave $k$, $f_k$ como o número de registros na amostra com esta mesma chave e $1/F_k$ sendo a probabilidade de identificação de um registro com esta chave. Assim, é preciso estimar a frequência populacional $F_k$ a partir da frequência amostral $f_k$ (que será denominada como $F_k|f_k$). Os autores ainda destacam que a literatura aponta dois métodos principais para esta estimação. O primeiro é supor que $F_k|f_k$ tem uma distribuição de Poisson, abordagem proposta por Skinner e Holmes (1998) e aprimorada por Elamir e Skinner (2006). O segundo se baseia na suposição de que $F_k|f_k$ tem uma distribuição Binomial Negativa, ideia originalmente por proposta Benedetti e Franconi (1998) e posteriormente desenvolvida por Polettini e Stander (2004). Em ambos os métodos, a medida do risco individual ($r$) é estimada a partir desta estimativa de $F_k$ e pode ser agregada para a obtenção da medida de risco global.

Templ (2017) ainda elenca abordagens adicionais para a estimação do risco de revelação individual, tais como: por modelo multinomial-Dirichlet (HOSHINO e TAKEMURA, 1998), por modelos Poisson Gaussiana Inversa (CARLSON, 2002), entre outros. Contudo, sob o ponto de vista operacional, destaca-se o já mencionado modelo originalmente proposto por Benedetti e Franconi (1998), coloquialmente chamado de "abordagem italiana", uma vez que este se encontra incorporado tanto no *software* μ-Argus, quanto no pacote "sdcMicro" do *software* R. De fato, Templ (2017) aponta que

esta abordagem é popular na literatura. Adicionalmente, a sua implementação no processo de produção das pesquisas seria relativamente simples, uma vez que o *software* R é livre e já está em uso no IBGE. Assim, optou-se por utilizar a abordagem por modelagem probabilística na estimação dos riscos de revelação da PNAD Contínua em estudo, nesta tese.

### 5.6.2. A abordagem italiana: o modelo de Benedetti-Franconi.

A abordagem italiana pode ser definida como um método para estimar o risco de revelação de cada registro no arquivo a ser divulgado, utilizando variáveis-chave discretas. O modelo formaliza a relação entre as frequências da amostra e da população por meio de pesos amostrais, mostrando o impacto que o desenho da pesquisa tem na limitação da divulgação (BENEDETTI *et al*., 1998).

Para o desenvolvimento do modelo, Benedetti *et al*. (1998) inicialmente fazem suposições sobre o comportamento do intruso, assim como definem o que se entende por identificação. Com relação ao intruso considera-se que:

- ele tem previamente, por algum outro meio que não os dados disseminados, o conhecimento de indicadores diretos (nomes, endereços, por exemplo) e das variáveis-chave;
- o intruso tenta identificar o registro *i* na amostra comparando o valor de chave para tal unidade na amostra disseminada, com as unidades *i** de seu conhecimento prévio na população, que tenham esses mesmos valores de chave.

Estas duas considerações, por sua vez, são baseadas na premissa que não há divergência ou erros de mensuração nas informações das variáveis-chave. No que diz respeito à identificação, esta ocorre para um registro *i* na amostra disseminada quando duas condições são atendidas, quais sejam:

- há uma correspondência dos valores de chave (tais como definidos na Seção 2.4.2) entre registro *i* e uma unidade *i** de conhecimento prévio, disponível ao intruso;
- *i** é a unidade da população da qual o registro *i* é derivado.

Os autores argumentam que se apenas um registro *i* e uma unidade *i** de conhecimento prévio do intruso compartilham do mesmo valor de chave há uma correspondência de um para um. Caso o intruso tenha um arquivo auxiliar que contenha toda a população ou saiba, de outra forma, que *i** é único nesta população, ocorre a identificação. Assim, nesta abordagem, considera-se o pior cenário supondo que o intruso tenha conhecimento prévio de toda a população (BENEDETTI *et al*., 1998).

Para a definição do risco de revelação individual Benedetti *et al.* (1998) consideram uma amostra *S* de tamanho *n* selecionada de uma população finita *P* de *N* indivíduos de acordo com um desenho amostral *D*. Para cada registro *i* define-se o risco de revelação $r_i$, como a probabilidade de identificar tal registro diante das informações contidas na amostra. Em outras palavras, é a probabilidade de associar o registro *i* à unidade *i**, dada a amostra observada, denotada por:

$$r_i = \text{P(associar o registro } i \text{ à unidade } i^* \mid S)$$

Para a estimação de $r_i$, os autores definem $f_k$ e $F_k$ como sendo, respectivamente, o número de registros na amostra e o número de unidades na população com valor de chave *k*. Na amostra disseminada apenas um subconjunto dos valores possíveis de *k* é observado ($f_k > 0$), que serão os valores de interesse para a estimação do risco de revelação. Os autores argumentam ainda que o risco de revelação é definido tal que registros na amostra que possuam o mesmo *k* são idênticos para o intruso, em termos deste risco. Desta forma, denota-se $r_k$ como o risco de um registro que possui o valor de chave *k* (BENEDETTI *et al*., 1998).

Franconi e Polettini (2004), no entanto, argumentam que os valores de $F_k$ são geralmente desconhecidos, logo uma etapa inicial de inferência deve ser feita. As autoras apontam que Benedetti *et al.* (1998) inferiram sobre os valores desconhecidos de $F_k$ por meio de uma abordagem Bayesiana a partir das frequências amostrais $f_k$. O risco de revelação é estimado então como a média (a posteriori) de $1/F_k$ proveniente de uma distribuição de $F_k \mid f_k$, tal que:

$$r_k = E\left(\frac{1}{F_k} \mid f_k\right) = \sum_{h \geq f_k} \frac{1}{h} P\left(F_k = h \mid f_k\right) \qquad (5.1)$$

Para determinar a função de densidade de $F_k|f_k$, uma abordagem de superpopulação é introduzida (BETHLEHEM et al., 1990) (RINOTT, 2003) (POLETTINI, 2003), em que é pressuposto:

$$\pi_k \sim [\overline{\pi_k}] \propto {}^{1}/_{\pi_k}, \pi_k > 0, \qquad k = 1, \dots, K, independentemente,$$
$$F_k \mid \pi_k \sim Poisson(N\pi_k),\ F_k = 0,1, \dots, independentemente, \tag{5.2}$$
$$f_k \mid F_k, \pi_k, p_k \sim binomial(F_k, p_k),\ f_k = 0,1, \dots, F_k, independentemente$$

tal que $\pi_k$ é a probabilidade de uma unidade da população possuir o valor de chave *k*, enquanto $p_k$ é a probabilidade de uma unidade populacional com este valor de chave *k* ser selecionado na amostra.

Franconi e Polettini (2004) argumentam que sob estas hipóteses, a distribuição a posteriori de $F_k|f_k$ é binomial negativa com probabilidade de sucesso $p_k$ e número de sucessos $f_k$. A função de densidade de uma variável binomial negativa que conta o número de tentativas antes do j-ésimo sucesso, cada um como probabilidade $p_k$, é:

$$P[F_k = h \mid f_k = j] = \binom{h-1}{j-1} . p_k^j (1 - p_k)^{h-j}, h \geq j,\ j > 0 \tag{5.3}$$

Benedetti *et al.* (1998) demonstram que sob a hipótese de distribuição binomial negativa, a equação (5.1) pode ser expressa como:

$$r_k = E(F_k^{-1}/f_k) = \int_0^\infty \left\{ \frac{p_k \exp(-t)}{1 - q_k \exp(-t)} \right\}^{f_k} dt \tag{5.4}$$

tal que $q_k = 1 - p_k$. Utilizando a transformação $y = \{1 - q_k \exp(-t)\}^{-1}$, a equação (5.4) pode ser reescrita como:

$$r_k = \left(\frac{p_k}{q_k}\right)^{f_k} \int_1^{\frac{1}{p_k}} \frac{(y-1)^{f_k - 1}}{y} dy \tag{5.5}$$

que é uma função monotonicamente crescente em $p_k$ e monotonicamente decrescente em $f_k$ e $F_k$. Para estimar o risco, é necessário estimar o parâmetro $p_k$ para cada valor de chave *k*. A ideia dos autores é utilizar a informação do plano amostral da pesquisa. Desta forma, no estimador de superpopulação de máxima-verossimilhança de $p_k$, que é $\hat{p}_k^{MLE} = \frac{f_k}{F_k}$, é introduzida a informação dada pelo desenho amostral *D*. Tal informação pode ser resumida pelo estimador de Horvitz-Thompson de $F_k$:

$$\hat{F}_k = \sum_{i:k(i)=k} w_i \tag{5.6}$$

tal que $w_i$ é o peso amostral do registro *i* com valor de chave *k*. Por fim, obtém-se o estimador de $p_k$ com base no desenho amostral:

$$\hat{p}_k^{D} = \frac{f_k}{\sum_{i:k(i)=k} w_i} \tag{5.7}$$

Com vistas à criação de um algoritmo para a implementação computacional do modelo de Benedetti-Franconi, Capobianchi *et al.* (2001), utilizando o estimador da equação (5.7), daqui para frente chamado apenas de $\hat{p}_k$ por questão de simplicidade, estimam o risco de revelação como sendo:

$$r_k = \frac{\hat{p}_k}{1-\hat{p}_k} \log\left(\frac{1}{\hat{p}_k}\right), \text{ para } f_k=1, \tag{5.8}$$

$$r_k = \left(\frac{\hat{p}_k}{1-\hat{p}_k}\right) - \left[\left(\frac{\hat{p}_k}{1-\hat{p}_k}\right)^2 log\left(\frac{1}{\hat{p}_k}\right)\right], \text{ para } f_k=2, \tag{5.9}$$

Já para $f_k \geq 3$, o estimador é dado por:

$$r_k = \left(\frac{\hat{p}_k}{1-\hat{p}_k}\right)^{f_k} \left[A_0\left(1 + \sum_{j=o}^{f_k-3}(-1)^{j+1} \prod_{l=0}^{j} B_l\right) + (-1)^{f_k}\log(\hat{p}_k)\right] \tag{5.10}$$

tal que $A_0 = \frac{\hat{p}^{(1-f_k)}-1}{(f_k-1)}$ e $B_l = \frac{(f_k-1-l)^2}{(l+1)(f_k-2-l)} \frac{\hat{p}^{(l+2-f_k)}-1}{\hat{p}^{(l+1-f_k)}-1}$

Capobianchi *et al.* (2001) alertam que à medida que $f_k$ aumenta – e Templ (2017) acrescenta que o mesmo vale para o tamanho da amostra – a equação (5.10) pode tornar-se computacionalmente inviável devido à quantidade de cálculos[8], assim os autores lançam mão da seguinte aproximação que traz resultados satisfatórios:

$$r_k = \frac{\hat{p}_k}{f_k-(1-\hat{p}_k)} \tag{5.11}$$

### 5.6.3. Estimação do risco de revelação por heurísticas.

Além da abordagem tradicional da estimação dos riscos de revelação por modelos probabilísticos, há também uma abordagem alternativa proposta por Elliot *et al*. (2002) que usa métodos heurísticos baseados no conceito de "unicidades especiais". Se um registro com valor de chave *k*, proveniente de um conjunto Q de variáveis-chave é único na amostra, e também é único para uma chave de um subconjunto deste mesmo Q, ele é considerado uma unicidade especial.

Na Tabela 5.5, observa-se que o registro número 4 possui um valor de chave ("masculino", "branca", "não", "40 a 49") único para o conjunto de variáveis-chave escolhida ('sexo", "cor/raça", "sabe ler/escrever", e "grupo de idade"). Entretanto esse mesmo registro possui chaves únicas para conjuntos menores dessas variáveis-chave

[8] À época do artigo original dos autores, por conta da limitação de *software*, $r_k$ era calculado por meio da equação (5.10) para valores de $f_k$ até 40, no máximo.

(por exemplo, “cor/raça” e “grupo de idade”) sendo, então, considerado uma unicidade especial.

Elliot *et al*. (2002) argumentam que uma unicidade especial tem maior probabilidade de ser uma unicidade populacional, do que outra unicidade amostral sem essas características – denominada de unicidade aleatória. As unicidades especiais são classificadas pelo número e pelo tamanho do(s) menor(es) subconjunto(s) de variáveis-chave que fazem o registro ser único, chamadas de unicidades amostrais mínimas (*minimal sample uniques* - MSU). No exemplo da Tabela 5.5, o registro 4 é único com duas MSU de tamanho 2 (a combinação das variáveis “cor/raça” e “sabe ler/escrever” ou “cor/raça” e “grupo de idade” são únicas).

Tabela 5.5: Exemplo de unicidade especial em uma amostra.

| Registro | Sexo | Cor/Raça | Sabe ler/ escrever | Grupo de Idade |
|---|---|---|---|---|
| 1 | Masculino | Negra | Sim | 20 a 29 |
| 2 | Feminino | Branca | Sim | 20 a 29 |
| 3 | Masculino | Branca | Sim | 30 a 39 |
| 4 | Masculino | Branca | Não | 40 a 49 |
| 5 | Masculino | Negra | Sim | 20 a 29 |
| 6 | Feminino | Branca | Sim | 20 a 29 |
| 7 | Masculino | Branca | Sim | 30 a 39 |

Fonte: elaborado pelo autor.

Os autores apontam que conforme o tamanho das MSU diminui e o número de MSU aumenta em um registro, maior é a probabilidade que este registro corresponda a uma unicidade populacional. Desta forma, uma métrica que levasse em conta o tamanho e o número de MSU para um registro, dado um conjunto de variáveis-chave, seria um caminho para a estimação do risco de revelação dessa unidade. Este método foi denominado de Algoritmo de Detecção de Unicidades Especiais (*Special Uniques Detection Algorithm* - SUDA). Em seu produto final é calculado o escore SUDA, que leva em conta o tamanho e número de MSU, como descrito, tal que quanto o maior o escore, maior o risco da unicidade amostral (ELLIOT *et al.*, 2002).

Benschop *et al.* (2019) acrescentam que para a estimação do risco de revelação dos registros, os escores SUDA podem ser utilizados em combinação com a métrica DIS (*Data Intrusion Simulation*), introduzida por Elliot (2000), que é um método de estimação do risco de revelação global do arquivo de dados. Este método combinado, DIS-SUDA, de forma sintética, distribuiria a medida do risco global gerado pela métrica DIS, entre os registros, de acordo com o escore SUDA de cada registro individual.

Hundepool *et al.* (2012) destacam que este método possui algumas vantagens como, por exemplo, identificar as variáveis e categorias de variáveis que mais contribuem para o risco de revelação do registro. Adicionalmente, apontam que ele já foi exaustivamente testado e foi utilizado para detectar registros com alto risco de revelação no arquivo da amostra do Censo do Reino Unido de 2001. Em contrapartida, apontam como desvantagem ser um método heurístico, tal que é impossível especificar as condições em que os resultados permanecem robustos, o que é corroborado por Duncan *et al.* (2011). Sobre este último aspecto Benschop *et al.* (2019) argumentam que o escore DIS-SUDA não considera totalmente os pesos amostrais. Desta forma, quando a fração amostral é muito pequena (grandes pesos amostrais), as estimativas deste método começam a diferir sensivelmente em relação às realizadas por modelos probabilísticos.

Sob o ponto de vista operacional os escores SUDA e DIS-SUDA estão implementados no pacote "sdcMicro", do *software* R. O referido pacote, inclusive, contém uma interface gráfica denominada "sdcApp" em que estas medidas podem ser obtidas de forma direta. Entretanto, Templ (2017) faz a mesma observação de Benschop *et al.* (2019) sobre o método não considerar totalmente os pesos amostrais e a diferença observada quando estes pesos são grandes. Assim, a interface gráfica admite como parâmetro mínimo de fração amostral um valor de 1%. Uma vez que, como já observado no Capítulo 4, a fração amostral da PNAD Contínua 2019 – 2º trimestre é inferior a 0,3%, optou-se por não calcular as estimativas de risco de revelação por este método, nesta tese. Por este motivo também optou-se por não descrever em minúcias os escores SUDA e DIS-SUDA, que podem ser vistos com detalhes em Elliot *et al.* (2002) e Elliot e Manning (2003). Contudo, estudos com vistas para implementação deste método em pesquisas

do IBGE com frações amostrais maiores como, por exemplo, a parte amostral do Censo Demográfico, não podem ser descartados.

## 5.7. Estimativas do risco de revelação pela abordagem italiana.

Empregando o modelo de Benedetti-Franconi calculou-se estimativas do risco de revelação para cada registro de pessoa da PNAD Contínua 2019 do 2º trimestre e para o nível hierárquico superior – registro de domicílio – da forma exposta na Seção 5.3. A partir destes resultados, é possível mensurar o efeito da estrutura hierárquica, em relação ao tamanho do domicílio, assim como calcular medidas para o risco global da pesquisa.

### 5.7.1. Risco de revelação individual.

Uma vez que os dados da PNAD Contínua possuem estrutura hierárquica, os riscos de revelação individual que devem ser analisados são os referentes aos registros de domicílio. Para cada um destes registros, o modelo atribui uma estimativa do referido risco. Mencionou-se na Seção 2.9, que a única forma de se obter risco nulo é a não divulgação de qualquer informação e que um dos objetivos do CEC é que o risco deve ser mantido abaixo de um limite máximo aceitável. Desta forma, com relação à primeira afirmação, há associada uma estimativa não nula do risco para todos os registros de pessoas e, por conseguinte, de domicílios. Entretanto, para a segunda afirmação é preciso definir primeiro o que é este limite aceitável.

Não há indicado na literatura um valor específico para o que seja risco de revelação aceitável. Geralmente isto é determinado pelo produtor das informações e, além de possíveis diretrizes legais ou institucionais a serem seguidas, pode variar bastante de acordo com as características dos dados a serem disseminados. Um arquivo de uso público de uma pesquisa deve ter um limite máximo de risco muito mais baixo do que um arquivo de acesso restrito daquela mesma pesquisa, por exemplo. Por outro lado, arquivos de uso público de duas pesquisas distintas não necessariamente terão limites máximos iguais, pois uma pode conter mais informações consideradas sensíveis do que a outra, e assim por diante.

Assim, não faz parte do escopo desta tese definir qual deveria ser o limite máximo aceitável de risco de revelação que o IBGE deveria atribuir para o arquivo de uso público da PNAD Contínua. Entretanto, com o intuito de enriquecer a análise, fornecendo um panorama mais amplo da distribuição do risco individual dos registros, inicialmente arbitrou-se três limiares distintos. O primeiro de 1%, tal que registros abaixo desse limite poderiam ser considerados seguros, e o segundo e terceiro em 10% e 20% respectivamente, tais que registros acima deste nível seriam candidatos a ter uma atenção demandada por parte da Instituição. As Tabelas 5.6 e 5.7 apresentam estes resultados de acordo com os dois cenários de revelação propostos na Seção 5.2.

Tabela 5.6: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por recorte geográfico de divulgação, segundo o cenário de revelação 1.

| Recorte Geográfico | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| Brasil | 187.061 | 1.492 | 0,8 | 61 | <0,1 | 13 | <0,1 |
| Grandes Regiões | 187.061 | 7.438 | 4,0 | 189 | 0,1 | 31 | <0,1 |
| Unidades da Federação | 187.061 | 38.095 | 20,4 | 1.383 | 0,7 | 220 | 0,1 |
| RM ou RIDE | 59.691 | 20.963 | 35,1 | 545 | 0,9 | 70 | 0,1 |
| Municípios fora de RM ou RIDE | 108.783 | 25.906 | 23,8 | 632 | 0,6 | 61 | 0,1 |
| Municípios de Capitais | 43.994 | 21.149 | 48,1 | 920 | 2,1 | 95 | 0,2 |
| Municípios fora de capitais | 143.067 | 32.937 | 23,0 | 1.431 | 1,0 | 274 | 0,2 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

De acordo com as Tabelas 5.6 e 5.7, observa-se, como já esperado, a partir que o recorte geográfico de divulgação torna-se mais desagregado, a proporção de registros de domicílios acima dos limiares propostos tende a aumentar. É possível, no entanto, verificar que o cenário 2 possui mais registros acima dos limiares propostos, o que vai

ao encontro da análise feita na Seção 5.5, que mostrou um maior número de unicidades amostrais neste cenário. Em outras palavras, quanto mais desagregado é o recorte, maior tende a ser o risco de revelação do registro, e este também tende a ser maior no segundo cenário.

Logo, as maiores proporções encontradas se referem ao recorte geográfico de município da capital no cenário 2, tal que 6,1% dos domicílios na amostra deste recorte possuem risco estimado superior a 10%, e 1% possuem um risco maior que 20%, como mostra a Tabela 5.7. Entretanto, ainda que nenhum recorte geográfico fosse divulgado, ou seja, que os dados se refiram apenas para o Brasil, ainda existiriam domicílios com risco acima do limiar de 20% para ambos os cenários considerados, como pode ser observado em ambas as tabelas.

Tabela 5.7: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por recorte geográfico de divulgação, segundo o cenário de revelação 2.

| Recorte Geográfico | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| Brasil | 187.061 | 5.143 | 2,7 | 353 | 0,2 | 95 | 0,1 |
| Grandes Regiões | 187.061 | 19.271 | 10,3 | 1.079 | 0,6 | 255 | 0,1 |
| Unidades da Federação | 187.061 | 65.004 | 34,8 | 4.703 | 2,5 | 992 | 0,5 |
| RM ou RIDE | 59.691 | 31.936 | 53,5 | 1.699 | 2,8 | 285 | 0,5 |
| Municípios fora de RM ou RIDE | 108.783 | 41.621 | 38,3 | 2.430 | 2,2 | 422 | 0,4 |
| Municípios de Capitais | 43.994 | 30.367 | 69,0 | 2.669 | 6,1 | 452 | 1,0 |
| Municípios fora de capitais | 143.067 | 52.953 | 37,0 | 3.988 | 2,8 | 881 | 0,6 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

As Tabelas A13 até A24 do Anexo A trazem um maior detalhamento destes números ao discriminar os riscos de revelação para cada unidade territorial dentro de

cada recorte geográfico considerado. Com base nessas tabelas é possível observar uma grande heterogeneidade nos valores dentro dos recortes geográficos, de forma análoga ao que foi mostrado na análise das unicidades amostrais. Tomando-se como exemplo o nível mais desagregado que abrange todos os 187.061 registros – a UF – verifica-se na Tabela 5.7 que o percentual de registros acima dos limiares 10% e 20% são de 2,5% e 0,5% respectivamente, para o cenário 2. Entretanto, de acordo com a Tabela A16 estes valores podem chegar a 42,1% e 12,8%, que são referentes à UF Roraima. Ainda neste cenário, outras UF se destacam como Acre e Amapá com, respectivamente 22,1% e 22,7% de registros de domicílio acima do limiar de 10%, enquanto o percentual de registros acima do limiar de 20% corresponde a 5,4% e 6,9%, respectivamente.

Para os outros quatro níveis de desagregação que representam subdivisões da UF, o panorama de heterogeneidade entre as unidades territoriais dentro dos recortes geográficos permanece. Ao se considerar, por exemplo, o recorte geográfico de RM ou RIDE, observa-se que o percentual de registros acima dos limiares 10% e 20% são de 2,8% e 0,5% respectivamente, para o cenário 2, de acordo com a Tabela 5.7. Contudo, a Tabela A18 mostra que estes valores podem atingir 28,7% e 9,0% respectivamente para a RM Macapá neste cenário.

Outro aspecto que pode ser destacado, de forma análoga ao observado na análise das unicidades amostrais, é que algumas unidades territoriais específicas têm comportamento distinto em relação ao total do recorte geográfico. Por exemplo, observa-se, de acordo com a Tabela 5.7, que o recorte dos municípios da capital é o que possui maior percentual de registros acima dos limiares considerados. Entretanto, para a UF Roraima, o recorte correspondente aos municípios fora da capital, foi o que apresentou os maiores valores, tais que 69,7% dos registros estão acima do limiar de 10%, enquanto 29,0% se encontram acima do limiar de 20%, como mostra a Tabela A24. Situação semelhante pode ser verificada na comparação dos recortes geográficos dos municípios dentro e fora de RM ou RIDE. Embora a Tabela 5.7 apresente valores superiores para o primeiro nos limiares propostos, a situação inverte-se para a UF Amazonas, como mostra a comparação das Tabelas A18 e A20. Neste caso a proporção de registros acima dos limiares de 10% e 20% é muito superior para os domicílios localizados fora da RM, atingindo percentuais de 16,7% e 5,2%, respectivamente.

Como mencionado na Seção 5.5, é preciso ter em conta que em algumas UF da região Norte, a maior parte da população encontra-se dentro do município da capital ou da RM. Assim, quando a pesquisa dissemina este recorte geográfico nos microdados, o seu complementar – municípios fora da capital ou da RM – que pode ser mais rarefeito em termos populacionais, tornando-se identificável para o intruso. Isto aumenta o risco de revelação para os registros localizados nessas áreas.

De forma geral, em uma análise da Tabela 5.7 e das Tabelas A13 até A24 do Anexo A, observa-se que há um volume não desprezível de registros que apresentam risco de revelação acima do maior limiar proposto de 20%. Para todos os recortes geográficos considerados, este número é maior para o cenário 2, o que vai de acordo com as análises das unicidades apresentadas na Seção 5.5. Adicionalmente, verifica-se que nos recortes geográficos mais desagregados existem algumas unidades territoriais com alta concentração dos registros com risco de revelação acima dos limiares propostos. Entretanto, deve ser ressaltada a presença de registros de domicílios acima do limiar de 20%, mesmo para o recorte mais agregado possível.

Outra análise relevante é a da distribuição das estimativas dos riscos de revelação, dado o recorte geográfico e os cenários de revelação propostos. Atenção especial deve ser dada para os valores mais altos da distribuição, pois eles são os que representam os registros de maior risco de revelação nos microdados. É possível observar, com base nas Tabelas 5.8 e 5.9 que as distribuições das estimativas dos riscos de revelação em todos os recortes geográficos e cenários são assimétricas positivas. Corroborando com as análises anteriores, verifica-se também que os valores destas estimativas, nas medidas de posição consideradas, são maiores para o cenário 2. Da mesma forma, à medida que os recortes se tornam mais desagregados, há um aumento do risco de revelação, como já esperado.

Uma medida importante apresentada nas Tabelas 5.8 e 5.9 é referente ao risco máximo, ou seja, o registro de domicílio que apresentou o maior risco de revelação estimado em determinado recorte geográfico e cenário de revelação. Neste item é possível observar valores muito altos, indicando que existem registros de domicílios extremamente suscetíveis à revelação por um intruso. Ao se considerar o cenário 2,

constata-se que mesmo no recorte mais agregado possível existe um registro com risco de revelação superior a 70%.

Tabela 5.8: Estimativas dos riscos de revelação nas medidas de posição para os domicílios da PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação, segundo o cenário de revelação 1.

| Medidas de posição | Estimativa dos riscos de revelação para o recorte geográfico | | | | | | |
|---|---|---|---|---|---|---|---|
| | Brasil | Região | UF | RM/RIDE | Fora da RM/RIDE | Capital | Fora da Capital |
| Máximo | 0,305 | 0,430 | 0,591 | 0,591 | 0,529 | 0,493 | 0,725 |
| P99 | 0,006 | 0,036 | 0,089 | 0,097 | 0,084 | 0,131 | 0,100 |
| P95 | <0,001 | 0,006 | 0,042 | 0,052 | 0,044 | 0,071 | 0,047 |
| P90 | <0,001 | 0,002 | 0,027 | 0,036 | 0,030 | 0,051 | 0,030 |
| Q3 | <0,001 | <0,001 | 0,006 | 0,018 | 0,009 | 0,027 | 0,008 |
| Mediana | <0,001 | <0,001 | 0,002 | 0,003 | 0,002 | 0,009 | 0,002 |
| Q1 | <0,001 | <0,001 | <0,001 | <0,001 | <0,001 | 0,002 | <0,001 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela 5.9: Estimativas dos riscos de revelação nas medidas de posição para os domicílios da PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação, segundo o cenário de revelação 2.

| Medidas de posição | Estimativa dos riscos de revelação para o recorte geográfico | | | | | | |
|---|---|---|---|---|---|---|---|
| | Brasil | Região | UF | RM/RIDE | Fora da RM/RIDE | Capital | Fora da Capital |
| Máximo | 0,714 | 0,714 | 0,900 | 0,900 | 0,719 | 0,898 | 0,923 |
| P99 | 0,034 | 0,077 | 0,154 | 0,154 | 0,140 | 0,201 | 0,163 |
| P95 | 0,003 | 0,027 | 0,069 | 0,079 | 0,068 | 0,110 | 0,073 |
| P90 | <0,001 | 0,011 | 0,045 | 0,055 | 0,470 | 0,078 | 0,048 |
| Q3 | <0,001 | 0,002 | 0,020 | 0,029 | 0,220 | 0,043 | 0,022 |
| Mediana | <0,001 | <0,001 | 0,004 | 0,012 | 0,005 | 0,021 | 0,004 |
| Q1 | <0,001 | <0,001 | 0,001 | 0,003 | 0,001 | 0,006 | 0,001 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Se as informações das Tabelas 5.6 e 5.7 quantificaram o número de registros de domicílio acima de um limiar de risco que pode ser considerado não desejável para divulgação, os dados das Tabelas 5.8 e 5.9 mostram que, dentre estes registros, existem alguns especialmente vulneráveis, muito acima dos limiares propostos. Estes resultados mostram a importância da incorporação, no processo de pesquisa, da utilização de uma medida que estime o risco de revelação individual dos registros.

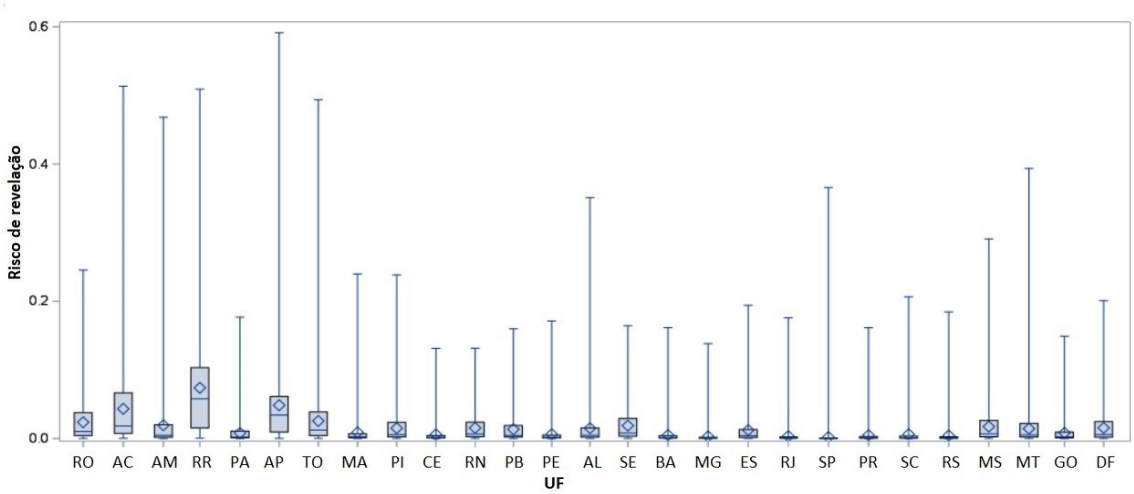
De forma análoga ao que foi mostrado na análise da proporção de registros não seguros por limiar, as distribuições apresentadas nas Tabelas 5.8 e 5.9 são muito heterogêneas quando considerada cada unidade territorial dentro dos recortes geográficos propostos. Tomando como exemplo os recortes geográficos mais desagregados que consideram todos os registros (UF) ou apenas um subgrupo destes (municípios fora da capital da UF), os Gráficos 5.1 a 5.4 apresentam os *boxplots* com a distribuição das estimativas de risco de cada unidade territorial, para ambos os cenários de revelação considerados.

É importante frisar que qualquer análise pode ser feita tanto para os registros de domicílio, quanto para o de pessoas. Contudo, neste último caso, deve-se atentar para a questão da estrutura hierárquica apresentada na Seção 5.3, tal que pessoas no mesmo domicílio terão o mesmo risco de revelação. Em outras palavras, o risco domiciliar será replicado para todos os seus moradores. Para ilustrar esta questão, optou-se por considerar nos Gráficos 5.1 a 5.4 a distribuição do risco relativo aos registros de pessoas. A próxima seção desta tese (5.7.2) aborda exclusivamente o tema do efeito da hierarquia sobre o risco de revelação.

Os Gráficos 5.1 e 5.2 mostram, como já mencionado, que há uma grande heterogeneidade para o risco de revelação dos registros de pessoas, entre as unidades territoriais, do recorte geográfico de divulgação referente à UF. Enquanto o risco tende a ser maior para as UF da Região Norte, as UF das Regiões Sul e Sudeste apresentam menores valores, em média. Contudo, podem existir pontos discrepantes com alto risco de revelação em quaisquer dessas unidades territoriais. Por exemplo, o valor máximo apresentado para este recorte geográfico, de acordo com os cenários 1 e 2 (0,591 e 0,900 apresentados nas Tabelas 5.8 e 5.9, respectivamente), se refere a pessoas em um domicílio da UF 16 (Amapá), embora a distribuição que apresentou, em média, os

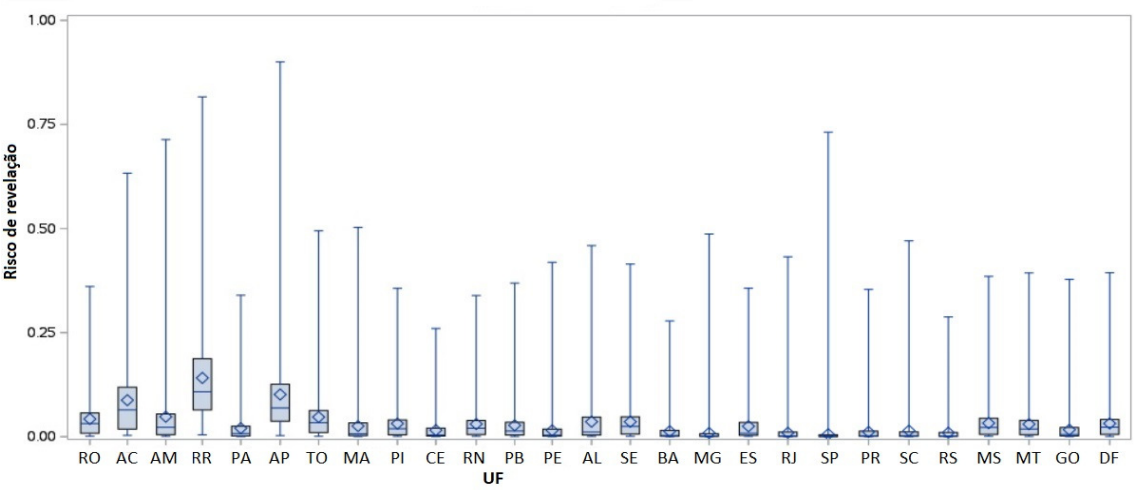
maiores valores de risco seja referente à UF 14 (Roraima). De modo análogo, a UF 35 (São Paulo) apresentou, de modo geral, as menores estimativas para o risco de revelação, apesar de ali ter sido observado um valor extremo superior ao da maioria das outras UF, especialmente com relação ao cenário 2.

Gráfico 5.1: Distribuição dos riscos de revelação do registro de pessoas, para o recorte geográfico de divulgação de Unidades da Federação, segundo o Cenário 1, em cada uma das UF.




Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

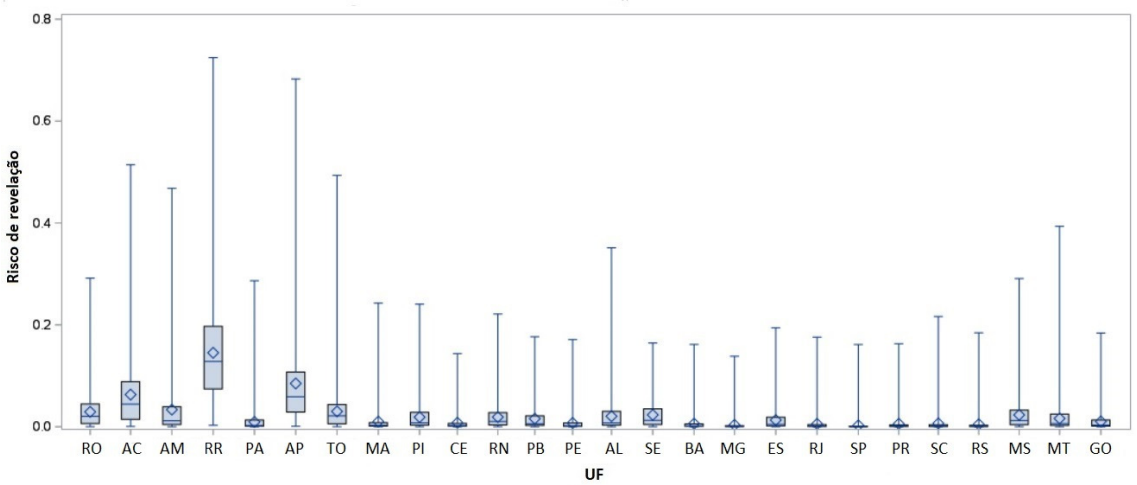
Gráfico 5.2: Distribuição dos riscos de revelação do registro de pessoas, para o recorte geográfico de divulgação de Unidades da Federação, segundo o Cenário 2, em cada uma das UF.




Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

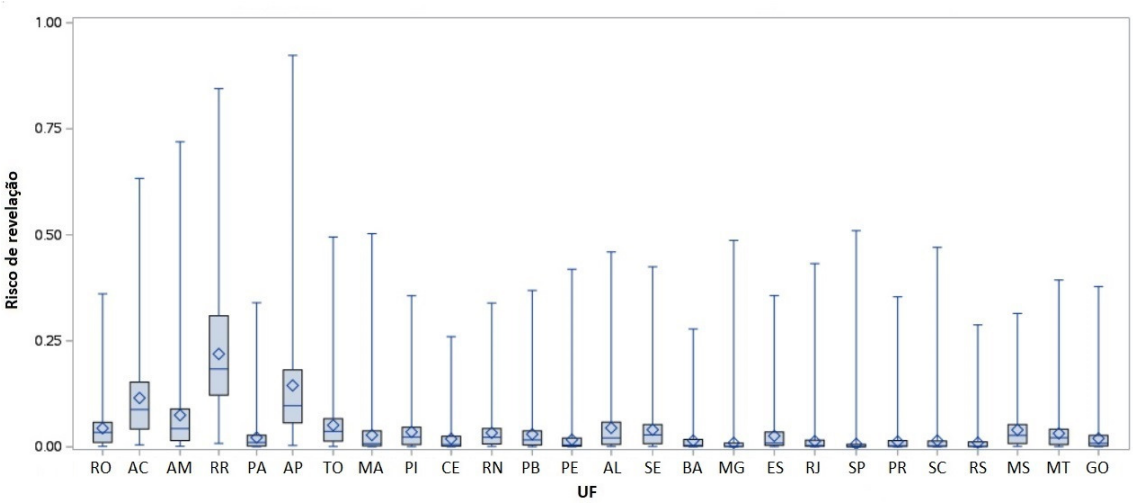
Panorama similar pode ser observado nos Gráficos 5.3 e 5.4, que evidenciam uma grande heterogeneidade para o risco de revelação dos registros de pessoas, ao se considerar o recorte geográfico de municípios fora da capital de cada UF. A distribuição das estimativas de risco para moradores em domicílios fora da capital da UF 14 (Roraima) apresentou, em média, os maiores valores. Ali também é que se observou o valor máximo observado com relação ao cenário 1, já em relação ao cenário 2, a maior estimativa pertencia a moradores em um domicílio da UF 16 (Amapá). Por outro lado, seguindo a tendência vista nos Gráficos 5.1 e 5.2, a distribuição das estimativas de risco referentes a moradores em domicílios fora da capital de UF das Regiões Sul e Sudeste foram as que apresentaram menores valores, de modo geral, para ambos os cenários considerados.

Gráfico 5.3: Distribuição dos riscos de revelação do registro de pessoas, para o recorte geográfico de municípios fora da capital da UF, segundo o Cenário 1, em cada uma das UF.




Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Gráfico 5.4: Distribuição dos riscos de revelação do registro de pessoas, para o recorte geográfico de municípios fora da capital da UF, segundo o Cenário 2, em cada uma das UF.




Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

### 5.7.2. Efeito da estrutura hierárquica por tamanho do domicílio

Como mencionado na Seção 5.3, de acordo com a literatura é importante levar em consideração a estrutura hierárquica na estimativa do risco de revelação e, como a PNAD Contínua possui uma estrutura hierárquica, o risco de revelação para o domicílio ou risco hierárquico (pelo menos uma pessoa dentro do domicílio ser identificada) será calculado. Desta forma, é intuitivo perceber que quanto maior o tamanho do domicílio, em número de moradores, maior será o efeito da estrutura hierárquica. Para estudar este efeito, os registros foram divididos em oito estratos, ou seja, desde domicílios com um até oito ou mais moradores. O critério de parada foi tal que o último estrato – o mais rarefeito – contivesse em torno de 1% dos registros da amostra. Neste caso, o valor ficou em 0,9% como mostram as Tabelas 5.10 e 5.11.

O foco da análise se dá nos registros que apresentaram os maiores riscos de revelação. Em outras palavras, o intuito é mensurar o impacto da estrutura hierárquica neste grupo, que é onde o INE deve concentrar maior atenção. Inicialmente, observou-se como se dá este comportamento nos registros com risco acima do último limiar

considerado (20%) da seção anterior, segundo os recortes geográficos e cenários de revelação propostos.

As Tabelas 5.10 e 5.11 apresentam, para ambos os cenários de revelação, a distribuição dos registros de domicílios com risco de revelação estimado acima de 20%, de acordo com seu tamanho, para os três recortes geográficos mais agregados que contém todos os 187.061 registros contidos na pesquisa. É possível observar que embora os domicílios com oito ou mais moradores representassem menos de 1% da amostra, em todos os casos estavam ali contidos o maior número absoluto dos casos não seguros dado o limiar proposto. Em números relativos, este montante é maior para o cenário 2 e para quanto mais agregado é o recorte geográfico. Neste segundo cenário, considerando uma disseminação apenas para o Brasil, 78 dentre os 95 casos de registros não seguros se encontram neste estrato.

Tabela 5.10: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado o limiar de risco de revelação de 20%, por tamanho do domicílio e recorte geográfico, segundo o cenário de revelação 1.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Brasil | | Região | | UF | |
| | Total | % | Total | % | Total | % | Total | % |
| Total | 187.061 | 100,0 | 13 | 100,0 | 31 | 100,0 | 220 | 100,0 |
| 1 | 29.507 | 15,8 | 1 | 7,7 | 3 | 9,7 | 6 | 2,7 |
| 2 | 51.234 | 27,4 | 2 | 15,4 | 5 | 16,1 | 30 | 13,6 |
| 3 | 47.043 | 25,1 | 0 | 0,0 | 1 | 3,2 | 22 | 10,0 |
| 4 | 35.023 | 18,7 | 0 | 0,0 | 3 | 9,7 | 39 | 17,7 |
| 5 | 14.766 | 7,9 | 1 | 7,7 | 2 | 6,5 | 30 | 13,6 |
| 6 | 5.473 | 2,9 | 1 | 7,7 | 3 | 9,7 | 31 | 14,1 |
| 7 | 2.262 | 1,2 | 0 | 0,0 | 1 | 3,2 | 19 | 8,6 |
| 8 ou mais | 1.753 | 0,9 | 8 | 61,5 | 13 | 41,9 | 43 | 19,5 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Isto explica, em parte, o que foi verificado na análise das Tabelas 5.6 e 5.7, quando mesmo no recorte geográfico mais agregado, era possível encontrar registros com risco de revelação estimado acima dos limiares propostos. Em outras palavras, domicílios com muitos moradores podem ser tão singulares, que mesmo com a agregação de áreas para a divulgação, eles continuariam tendo características únicas, que os deixariam vulneráveis à revelação.

Tabela 5.11: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado o limiar de risco de revelação de 20%, por tamanho do domicílio e recorte geográfico, segundo o cenário de revelação 2.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Brasil | | Região | | UF | |
| | Total | % | Total | % | Total | % | Total | % |
| Total | 187.061 | 100,0 | 95 | 100,0 | 255 | 100,0 | 993 | 100,0 |
| 1 | 29.507 | 15,8 | 1 | 1,1 | 5 | 2,0 | 8 | 0,8 |
| 2 | 51.234 | 27,4 | 3 | 3,2 | 7 | 2,7 | 31 | 3,1 |
| 3 | 47.043 | 25,1 | 1 | 1,1 | 3 | 1,2 | 35 | 3,5 |
| 4 | 35.023 | 18,7 | 3 | 3,2 | 11 | 4,3 | 76 | 7,7 |
| 5 | 14.766 | 7,9 | 1 | 1,1 | 9 | 3,5 | 76 | 7,7 |
| 6 | 5.473 | 2,9 | 4 | 4,2 | 21 | 8,2 | 154 | 15,5 |
| 7 | 2.262 | 1,2 | 4 | 4,2 | 24 | 9,4 | 156 | 15,7 |
| 8 ou mais | 1.753 | 0,9 | 78 | 82,1 | 175 | 68,6 | 457 | 46,0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

As Tabelas A25 até A28 do Anexo A, mostram estes resultados para os quatro recortes geográficos mais desagregados. Estes estão em tabelas separadas, para melhor visualização, pois o total de registros nestes recortes são distintos. De forma análoga, ao que é visto nas Tabelas 5.10 e 5.11, observa-se que os registros de domicílios com oito ou mais moradores, sempre em torno de 1% do total da amostra, concentram uma proporção comparativamente bem maior dos registros não seguros dado o limiar de

20%. Este número varia de 15,7% (domicílios em RM ou RIDE para o cenário 1) a 55,2% (domicílios fora de RM ou RIDE para o cenário 2).

Na seção anterior, analisou-se a distribuição das estimativas do risco de revelação a partir de suas medidas de posição. A ênfase se deu nos valores mais altos, por representarem os registros com maior risco de revelação. O mesmo pode ser feito para cada estrato de tamanho do domicílio. Para melhor visualização e comparação optou-se por tomar uma medida resumo de interesse, neste caso o percentil 99, da distribuição do risco. Assim, as Tabelas 5.12 e 5.13 apresentam estes valores para cada recorte geográfico e cenário de revelação propostos.

Tabela 5.12: Percentil 99 da distribuição do risco de revelação dos domicílios estimados da PNAD Contínua 2019 – 2º trimestre, por tamanho do domicílio e recorte geográfico, segundo o cenário de revelação 1.

| Tamanho do domicílio | P99 da distribuição do risco de revelação para recorte o geográfico | | | | | | |
|---|---|---|---|---|---|---|---|
| | Brasil | Região | UF | RM/RIDE | Fora da RM/RIDE | Capital | Fora da Capital |
| 1 | 0,002 | 0,025 | 0,055 | 0,048 | 0,050 | 0,066 | 0,058 |
| 2 | 0,004 | 0,030 | 0,074 | 0,075 | 0,071 | 0,094 | 0,082 |
| 3 | 0,004 | 0,033 | 0,079 | 0,086 | 0,077 | 0,107 | 0,091 |
| 4 | 0,007 | 0,038 | 0,096 | 0,103 | 0,088 | 0,147 | 0,106 |
| 5 | 0,018 | 0,047 | 0,123 | 0,129 | 0,109 | 0,171 | 0,133 |
| 6 | 0,025 | 0,058 | 0,163 | 0,193 | 0,138 | 0,233 | 0,172 |
| 7 | 0,033 | 0,071 | 0,196 | 0,175 | 0,162 | 0,232 | 0,257 |
| 8 ou + | 0,081 | 0,157 | 0,283 | 0,249 | 0,311 | 0,330 | 0,364 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

De acordo com as Tabelas 5.12 e 5.13, é possível observar que o valor do P99 da distribuição do risco de revelação tende a aumentar nos estratos que contemplam os registros com maior tamanho de domicílio. Esse incremento tende a ser maior tanto para o cenário 2, quanto para os recortes mais desagregados. Estes resultados corroboram com a percepção, apresentada na análise das Tabelas 5.10 e 5.11 que, devido à estrutura hierárquica presente nos dados, os domicílios com muitos moradores

podem se tornar vulneráveis à revelação, devido às suas características únicas. A Tabela 5.13 mostra, também como já mencionado anteriormente, que uma possível agregação de áreas para disseminação tem efeito limitado, no que diz respeito ao risco de revelação destes domicílios. Neste caso o valor do P99, no estrado com oito ou mais moradores, por exemplo, varia apenas de 0,549 a 0,360 entre os recortes mais e menos desagregados possíveis.

Estes resultados mostram que é importante que seja avaliada a forma de divulgação dos domicílios a partir de um determinado número de moradores. No Capítulo 3 foi mencionado, por exemplo, que a ONS pode até suprimir um registro inteiro de domicílio com muitos moradores. Certamente existem opções menos radicais que podem ser avaliadas, com o intuito de maximizar a utilidade dos dados e que podem ser estudadas e consideradas. Um exemplo é manter os registros de pessoas destes domicílios, mas impossibilitando que o usuário possa saber a qual domicílio elas pertencem. Diferentes soluções podem ser tomadas para distintas pesquisas, dependendo das variáveis disponíveis.

Tabela 5.13: Percentil 99 da distribuição do risco de revelação dos domicílios estimados da PNAD Contínua 2019 – 2º trimestre, por tamanho do domicílio e recorte geográfico, segundo o cenário de revelação 2.

| Tamanho do domicílio | P99 da distribuição do risco de revelação para recorte o geográfico | | | | | | |
|---|---|---|---|---|---|---|---|
| | Brasil | Região | UF | RM/RIDE | Fora da RM/RIDE | Capital | Fora da Capital |
| 1 | 0,018 | 0,046 | 0,069 | 0,053 | 0,064 | 0,073 | 0,072 |
| 2 | 0,012 | 0,043 | 0,090 | 0,086 | 0,080 | 0,120 | 0,095 |
| 3 | 0,015 | 0,046 | 0,102 | 0,102 | 0,091 | 0,146 | 0,110 |
| 4 | 0,021 | 0,054 | 0,124 | 0,122 | 0,105 | 0,181 | 0,132 |
| 5 | 0,036 | 0,077 | 0,173 | 0,179 | 0,154 | 0,224 | 0,193 |
| 6 | 0,073 | 0,136 | 0,264 | 0,256 | 0,232 | 0,307 | 0,287 |
| 7 | 0,103 | 0,209 | 0,356 | 0,309 | 0,280 | 0,380 | 0,382 |
| 8 ou + | 0,360 | 0,424 | 0,523 | 0,482 | 0,514 | 0,523 | 0,549 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

### 5.7.3. Risco de revelação global

Templ (2017) argumenta que além das medidas de risco de revelação individuais, uma medida de risco para o arquivo como um todo, chamado de risco global pode ser de interesse. O autor aponta três abordagens mais comuns para esta estimação: pelo número esperado de identificações, pela utilização de algum *benchmark* e por meio de modelos log-lineares. As duas primeiras se utilizam, de forma direta, dos resultados obtidos na estimação dos riscos individuais.

O número esperado de identificações, é considerada a abordagem mais simples, consistindo na soma dos riscos individuais estimados (TEMPL, 2017). Entretanto, dado que seu resultado produz um número absoluto – número de registros – um outro modo de apresentar o resultado, para ser melhor comparável para diferentes recortes geográficos, é utilizar a proporção esperada de registros identificados. Em outras palavras, dividir o valor encontrado pelo total de registros. Benschop *et al.* (2019) utilizam esta segunda forma, atribuindo-a uma outra intepretação, uma vez que a soma dos riscos, dividida pelo total de registros pode ser entendida como a média dos riscos. Ambas medidas estão dispostas nas Tabelas 5.14 e 5.15.

A segunda abordagem descrita, por *benchmark*, consiste em contar o número de observações com risco individual maior que um determinado limite. Benschop *et al.* (2019) diz que este limite pode ser absoluto ou relativo. O primeiro caso consistiria na contagem de registros com risco maior que um valor fixado, ou seja, as Tabelas 5.6 e 5.7 já trariam medidas do risco global, uma vez que indicam registros acima de determinados limiares absolutos. Os limites relativos, por sua vez, poderiam levar em conta algum tipo de medida de posição ou dispersão dos dados. Um exemplo é uma medida utilizada por Templ (2017) que conta o número de registros em que o risco estimado ($r_i$) tenha as seguintes condições:

$$r_i \geq 0{,}1 \text{ e } r_i \geq 2.(\tilde{r} + 2.MAD(r))$$

tal que $\tilde{r}$ é a mediana dos riscos individuais estimados e $MAD(r)$ é o desvio absoluto mediano destes riscos que, por sua vez, é calculado por:

$$MAD(r) = mediana(|r_i - \tilde{r}|)$$

Esta medida de risco global, ou seja, a contagem de registros com risco acima deste *benchmark* relativo, está implementada no já citado pacote "sdcMicro" e os resultados estão dispostos na Tabela 5.16.

Templ (2017), aponta que a terceira abordagem utiliza modelos log-lineares para estimar o número de unicidades amostrais que também são unicidades populacionais. Parte-se do pressuposto que o número de unidades da população com determinado valor de chave observado na amostra tenha uma distribuição de Poisson. Desta forma, o risco global poderia ser estimado por um modelo log-linear, utilizando efeitos principais e interações, a partir das variáveis-chave. Esta ideia foi introduzida por Skinner e Holmes (1998).

Uma vez que o foco principal desta tese é em relação aos riscos individuais, ou seja, evitar que um registro com risco considerado acima de um valor considerado aceitável seja divulgado, optou-se por calcular as medidas de risco global apenas pelas duas primeiras abordagens. Em termos operacionais tais abordagens praticamente não impactariam no fluxo de produção da pesquisa, por serem derivadas diretamente das estimativas dos riscos individuais previamente estimados.

Com base na Tabela 5.14 verifica-se, em linha com as análises anteriores, que a identificação esperada tende a aumentar para o cenário 2 e quanto mais desagregadas forem as áreas de divulgação. Benschop *et al*. (2019), no entanto, argumentam que as medidas de risco global devem ser avaliadas com cuidado. É possível que alguns registros de alto risco de revelação, sejam compensados por muitos registros de baixo risco, obtendo-se uma medida relativamente pequena de risco global. Os autores citam, por exemplo, que uma medida de risco médio com valor baixo, não excluiria a possibilidade da existência de alguns registros com alta probabilidade de identificação.

Este é o caso da coluna de percentuais da Tabela 5.14 que pode ser interpretada como, já mencionado nesta seção, a média dos riscos. Embora o valor médio mais alto seja pouco mais de 3% para o recorte de municípios das capitais para o cenário 2, a análise dos riscos individuais mostrou alguns registros com riscos bem elevados próximos aos 90%. Por outro lado, o valor absoluto da identificação esperada é dado pela soma dos riscos individuais dos registros. Assim, mesmo que todos os valores estimados do risco fossem muito baixos, a soma indicaria algum montante de

identificação esperada, mesmo que em nenhum registro individual fosse detectada vulnerabilidade relevante.

Tabela 5.14: Risco global de revelação, medido por identificação esperada dos domicílios da PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação, segundo os cenários de revelação.

| Recorte Geográfico | Registros de domicílios | Identificação esperada | | | |
|---|---|---|---|---|---|
| | | Cenário 1 | | Cenário 2 | |
| | | Total | % | Total | % |
| Brasil | 187.061 | 80,7 | 0,04 | 275,7 | 0,15 |
| Grandes Regiões | 187.061 | 348,1 | 0,19 | 928,9 | 0,50 |
| Unidades da Federação | 187.061 | 1.621,6 | 0,87 | 3.090,7 | 1,65 |
| RM ou RIDE | 59.691 | 776,6 | 1,30 | 1.331,0 | 2,23 |
| Municípios fora de RM ou RIDE | 108.783 | 1.039,4 | 0,96 | 1.870,2 | 1,72 |
| Municípios de Capitais | 43.994 | 864,4 | 1,96 | 1.469,2 | 3,34 |
| Municípios fora de capitais | 143.067 | 1.440,6 | 1,01 | 2.556,8 | 1,79 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

A mesma medida de identificação esperada pode ser calculada para os registros de pessoas. Deve ser ressaltado que o número de pessoas em um domicílio é variável, e observou-se na seção anterior que domicílios de maior tamanho tendem a ter maior risco de revelação. Assim, é esperado que a proporção de pessoas com identificação esperada tenda a ser maior, comparada à proporção de domicílios. Seu cálculo é dado pela soma do risco hierárquico de todos os registros de pessoa. Uma vez que pessoas no mesmo domicílio possuem o mesmo risco, sua fórmula é equivalente à identificação esperada do domicílio ponderada pelo seu tamanho. Os resultados estão dispostos na Tabela 5.15.

É possível observar, com base na Tabela 5.15, que os valores percentuais para todos os recortes geográficos e cenários de revelação são maiores se comparados aos da Tabela 5.14. Este resultado era previsto, pois advém da influência da hierarquia no

risco de revelação, em função do tamanho do domicílio. Dado que se almeja garantir a privacidade do respondente – a pessoa – a partir da confidencialidade de suas informações, este risco baseado nos registros de pessoas pode ser preferível. De fato, a medida de identificação esperada que está disponível no pacote "sdcMicro" é feita com base nestes registros.

Outra medida do risco global também contida no pacote "sdcMicro" é a contagem de registros acima do *benchmark*, já descrita nesta seção. Templ (2017) argumenta que enquanto as medidas apresentadas nas Tabelas 5.14 e 5.15 dão uma ideia do risco geral de identificação para o arquivo, esta abordagem é uma medida relativa que indica se a distribuição de riscos dos registros possui valores extremos. Os resultados estão dispostos na Tabela 5.16.

Tabela 5.15: Risco global de revelação, medido por identificação esperada de pessoas da PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação, segundo os cenários de revelação.

| Recorte Geográfico | Registros de pessoas | Identificação esperada | | | |
|---|---|---|---|---|---|
| | | Cenário 1 | | Cenário 2 | |
| | | Total | % | Total | % |
| Brasil | 551.348 | 311,9 | 0,06 | 1.640,1 | 0,30 |
| Grandes Regiões | 551.348 | 1.259,3 | 0,23 | 4.441,6 | 0,81 |
| Unidades da Federação | 551.348 | 5.828,5 | 1,06 | 12.833,8 | 2,33 |
| RM ou RIDE | 170.401 | 2.729,2 | 1,60 | 5.124,8 | 3,01 |
| Municípios fora de RM ou RIDE | 323.797 | 3.710,4 | 1,15 | 7.705,2 | 2,38 |
| Municípios de Capitais | 126.336 | 3.088,8 | 2,44 | 5.603,9 | 4,44 |
| Municípios fora de capitais | 425.012 | 5.193,5 | 1,22 | 10.581,5 | 2,49 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Com base na Tabela 5.16, verifica-se, de modo análogo às análises anteriores, que o número de registros acima do *benchmark* tende a aumentar para o cenário 2 e quanto mais desagregadas forem as áreas de divulgação. Uma exceção a este

comportamento é observada para o recorte de UF, que possui números relativos superiores aos recortes de municípios dentro ou fora de RM ou RIDE. Deve-se ter em conta que o registro possuir risco de revelação estimado maior que 0,1 é uma condição para ele ser enquadrado nesta categoria. As Tabelas A15 e A16 do Anexo A, por sua vez, mostram que as UF Acre e Roraima são as que possuem a maior proporção de domicílios com esta característica. Uma vez que não há RM ou RIDE nas mesmas, estes registros não são contabilizados nestes dois recortes mais desagregados. Esta também é uma indicação que estas UF podem concentrar o maior número de valores extremos do risco de revelação.

Tabela 5.16: Risco global de revelação, medido por registros de pessoas acima do *benchmark*, na PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação, segundo os cenários de revelação.

| Recorte Geográfico | Registros de pessoas | Acima do *benchmark* | | | |
|---|---|---|---|---|---|
| | | Cenário 1 | | Cenário 2 | |
| | | Total | % | Total | % |
| Brasil | 551.348 | 40 | 0,01 | 104 | 0,02 |
| Grandes Regiões | 551.348 | 94 | 0,02 | 248 | 0,04 |
| Unidades da Federação | 551.348 | 392 | 0,07 | 693 | 0,13 |
| RM ou RIDE | 170.401 | 93 | 0,05 | 153 | 0,09 |
| Municípios fora de RM ou RIDE | 323.797 | 133 | 0,04 | 257 | 0,08 |
| Municípios de Capitais | 126.336 | 67 | 0,05 | 101 | 0,08 |
| Municípios fora de capitais | 425.012 | 499 | 0,12 | 809 | 0,19 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Por outro lado, a soma dos recortes dos municípios dentro e fora de capitais, somados, perfazem o total de registros da PNAD Contínua. Podem ser considerados uma subdivisão em dois grupos disjuntos do recorte de UF. Desta forma, é possível observar que os maiores números absoluto e relativo de registros com valores extremos de risco de revelação estão localizados nos municípios fora da capital, recorte não divulgado,

mas que pode ser construído pelo intruso. Deve ser ressaltado que a fórmula apresentada para o cálculo desta medida leva em conta a mediana dos riscos individuais estimados deste conjunto específico de registros e o desvio absoluto mediano destes riscos. Logo, esta não é uma medida indicada para comparações entre diferentes arquivos de dados.

Por fim, é preciso destacar que os riscos globais podem ser utilizados em conjunto com os riscos individuais para nortear a decisão de divulgação de um arquivo. O produtor dos dados, além de determinar um limite máximo aceitável para o risco individual de todos os registros pode, por exemplo, tal como aponta Hundepool *et al*. (2012), igualmente estabelecer um parâmetro máximo para o risco global do arquivo.

### 5.7.4. Considerações gerais sobre os riscos de revelação estimados

Como já mencionado na Seção 5.6, a estimação dos riscos de revelação pelo modelo de Benedetti-Franconi está disponível no *software* R. Utilizando o computador pessoal especificado na Seção 5.5, os tempos de processamento, em segundos, necessários para a obtenção dos resultados nos diferentes recortes geográficos foram baixos e estão dispostos na Tabela 5.17.

Tabela 5.17: Tempo de execução da estimação dos riscos de revelação pelo modelo de Benedetti-Franconi benchmark, em segundos, na PNAD Contínua 2019 – 2º trimestre, por recorte geográfico de divulgação e cenários de revelação.

| Recorte Geográfico | Cenário 1 | Cenário 2 |
|---|---|---|
| Brasil | 6,8 | 7,4 |
| Grandes Regiões | 7,3 | 7,4 |
| Unidades da Federação | 7,2 | 7,3 |
| RM ou RIDE | 7,5 | 7,0 |
| Municípios de Capitais | 7,2 | 7,2 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Observa-se com base na Tabela 5.17 que o tempo de processamento foi de aproximadamente 7 segundos, praticamente não variando para os diferentes cenários e recortes geográficos. As variáveis-chave correspondentes aos cenários 1 e 2 foram

apresentadas nos Quadros 5.1 e 5.2. Caso não seja acrescentada nenhuma outra variável de recorte geográfico, são obtidas as estimativas para o Brasil, caso contrário obtém-se as estimativas para o recorte correspondente. Deve ser ressaltado que o acréscimo da variável indicadora de RM ou RIDE, possibilita a estimação de dois recortes geográficos: municípios dentro ou fora de RM ou RIDE. O mesmo é válido para o acréscimo da variável indicadora de capital para os recortes de municípios dentro e fora destas capitais. Desta forma, dada a disponibilidade e conhecimento do *software* na Instituição e o reduzido tempo de processamento[9], a implementação no processo de produção das pesquisas da estimação do risco de revelação seria relativamente simples.

Com relação às estimativas obtidas, de um modo geral, verificou-se que existem registros de pessoas e, por conseguinte, de domicílios, que apresentaram alto risco de revelação, ultrapassando 90% nos casos mais extremos. Embora o quantitativo de registros vulneráveis seja maior nos recortes geográficos mais desagregados, o problema persistiria mesmo na divulgação apenas para o Brasil. Parte desta questão é explicada pela estrutura hierárquica dos dados, tal que domicílios com muitos moradores possuem características únicas que podem facilitar sua identificação. Por fim, há recortes geográficos não divulgados, mas de fácil recomposição por parte do intruso, com expressivos riscos de revelação das informações para registros em determinadas unidades territoriais como, por exemplo, nos municípios fora da capital da UF Roraima.

[9] Os arquivos também foram processados tanto apagando-se metade de seus registros, quanto duplicando-os, para testar se alterações em seu tamanho poderia aumentar exponencialmente o tempo de processamento. Não houve, em princípio, evidências que corroborem esta hipotése. Tomando-se o exemplo do recorte geográfico para o total do Brasil no Cenário 1, os tempos de execução para arquivos com metade e o dobro do número de registros foram de 4,3 e 13,4 segundos, respectivamente.

# CAPÍTULO 6: MASCARAMENTO DOS DADOS E PERDA DE INFORMAÇÃO

Este capítulo aborda a questão do tratamento a ser feito nos registros com risco de revelação acima do limite tolerável, a mensuração da perda de informação decorrente deste tratamento e a incorporação destas questões no processo da pesquisa. No Capítulo 5 foi estimada uma série de riscos de revelação, para cada registro de pessoa e de domicílio, correspondente às diferentes combinações de cenários de revelação e recortes geográficos. Desta forma, é preciso definir primeiramente qual o conjunto de riscos estimados será levado em conta. Optou-se, então, por trabalhar com o cenário 2 no recorte geográfico de divulgação referente aos domicílios localizados em capitais. Esta escolha, consequentemente, implica a se trabalhar também com o recorte complementar dos domicílios em municípios fora da capital, o qual pode ser obtido pelo intruso.

A opção pelo cenário 2 foi motivada pelo fato deste conter uma maior incidência de registros considerados como unicidades amostrais e, principalmente, apresentar, em média, maiores valores das estimativas do risco de revelação, como mostrado no capítulo anterior. A escolha dos recortes geográficos se deu por razão análoga, uma vez que quanto mais desagregado é o recorte, maior tenderá a ser a presença de unicidades amostrais e o risco de revelação dos registros. Caso se garanta que o limite tolerado de risco não seja violado nestes recortes – que são subdivisões da UF – garante-se que em suas agregações (UF, Regiões e Brasil), o mesmo ocorrerá. A escolha deste recorte também foi motivada pela possibilidade de se trabalhar com o arquivo inteiro da PNAD Contínua. Caso a opção fosse trabalhar com a outra subdivisão possível da UF – recortes de domicílios em municípios dentro e fora de RM ou RIDE – isto não seria possível pois, como comentado no Capítulo 5, nem todas as UF possuem esta subdivisão.

É preciso, por fim, definir qual o limite tolerável do risco de revelação. Devido à estrutura hierárquica da PNAD Contínua, é o risco domiciliar (ou hierárquico) que deve ser levado em consideração. Nesta aplicação, optou-se pelo valor de 20%, o maior limiar proposto no Capítulo 5. Esta opção se deu em virtude do exposto no parágrafo anterior com relação ao cenário de revelação e recorte geográfico. Isto é, os valores aqui

analisados são referentes ao cenário com, em média maiores estimativas de risco e no recorte mais desagregado. Para os recortes mais agregados, garante-se que o risco será menor e há uma tendência do mesmo para o outro Cenário 1.

O objetivo deste capítulo é, então, abordar as questões de mascaramento de dados e perda de informações, a partir de um conjunto estimado de riscos de revelação. Entretanto, é possível que em uma pesquisa mais de um conjunto destes riscos tenha que ser levado em conta. Neste caso, a mesma abordagem poderia ser replicada para os distintos conjuntos de riscos.

## 6.1. Mascaramento dos dados

Se as estimativas dos riscos de revelação apontam quais registros no arquivo de dados necessitam o uso de métodos de mascaramento, a questão seguinte é: que métodos serão utilizados e em quais variáveis? Essa decisão deve levar em conta as especificidades de cada pesquisa, a partir de algum arcabouço institucional definido para estas questões. Não é o escopo desta tese definir tal arcabouço para o IBGE, tampouco determinar que variáveis e métodos devem ser utilizados no caso da PNAD Contínua, tarefa que deve levar em conta a equipe temática da pesquisa. Desta forma, o intuito desta seção é propor uma abordagem relativa às questões que devem ser consideradas na escolha destas variáveis e métodos, tendo a PNAD Contínua como exemplo.

### 6.1.1. Considerações para a escolha das variáveis

Como mencionado na seção anterior, o conhecimento da equipe temática da pesquisa é de essencial valia na decisão das variáveis a serem alvo de mascaramento. É importante, por exemplo, conhecer as principais formas que os usuários utilizam os dados e indicadores demandados. Benschop *et al.* (2019) argumentam que se por um lado o CEC deve priorizar a proteção aos respondentes, por outro não pode esquecer os usuários dos dados, limitando a perda de utilidade das informações. Desta forma, deve-se evitar aplicar algum método de mascaramento em variáveis consideradas muito importantes para o usuário, ou, pelo menos, tentar preservar o máximo possível sua

utilidade. A PNAD Contínua, por exemplo, tem como um dos principais objetivos mensurar a inserção da população no mercado de trabalho, sendo a taxa de desemprego um indicador de grande relevância. Assim, as variáveis utilizadas no cálculo deste indicador devem ter atenção especial em termos de sua utilidade.

Outro fator a ser levado em conta é a relação entre as variáveis no questionário da pesquisa. Há casos em que a resposta de um quesito é dependente ou tem relação com a resposta de outro. Um exemplo, no caso da PNAD Contínua, é a variável idade que vai definir se o entrevistado deve responder os quesitos de educação e trabalho, requeridos para pessoas com 5 e 14 anos ou mais de idade, respectivamente. Assim, quaisquer alterações em uma variável, por uso de algum método de mascaramento, devem levar este aspecto em consideração para não causar inconsistência no registro. Por este motivo, é importante se atentar às regras de crítica da pesquisa nesta etapa. A associação entre os processos de mascaramento e os de crítica e imputação das variáveis serão vistos com mais detalhes na Seção 6.5.

Pesquisas que possuem hierarquia entre registros, que é o caso da PNAD Contínua, também devem receber um cuidado adicional nesta fase. Benschop *et al.* (2019), argumentam que algumas variáveis que se referem ao nível hierárquico mais alto, tem seus valores replicados para todos os registros no nível mais baixo. Um exemplo seria a variável número de moradores do domicílio, ou seja, todos os registros de pessoas no domicílio possuem o mesmo valor para ela. Neste caso, é necessário ter atenção para que as variáveis do nível hierárquico mais alto se mantenham consistentes em todos os registros do nível mais baixo. No exemplo anterior significaria não haver um registro de pessoa com valor diferente para a variável número de moradores, dentro do domicílio. Caso contrário, a estrutura hierárquica permitiria uma posterior recomposição de um valor que foi mascarado.

Há também a possibilidade de se evitar que determinadas variáveis sejam alvo de mascaramento por questões de política interna do produtor dos dados. Neste caso, haveria uma definição pré-estabelecida que uma variável ou um conjunto delas não poderia ser alvo de mascaramento. Esta decisão poderia ser fruto de diversos motivos, um exemplo é supor uma pesquisa amostral que faça uso de ajuste nos fatores de expansão utilizando calibração, no intuito de melhorar a qualidade de suas estimativas.

Esta calibração faria com que determinadas estimativas de totais da pesquisa coincidissem com as de outras fontes externas. Assim, poderia ser interessante deixar as variáveis envolvidas neste processo inalteradas, para garantir que estes totais não se modificassem. Na PNAD Contínua, até a divulgação do trimestre móvel de junho/agosto de 2021, este recurso era utilizado para que o total populacional da pesquisa, em alguns recortes geográficos, coincidisse com o das estimativas populacionais produzidas pelo IBGE. A partir de então, os fatores de expansão da pesquisa também foram ajustados para coincidir com estimativas de sexo e classes de idade para o Brasil (IBGE, 2021c).

De um modo geral as variáveis-chave são as candidatas naturais para o uso de métodos de mascaramento, afinal é com base nestas que é feita a hipótese de tentativa de revelação pelo intruso. Não obstante, variáveis sensíveis podem ser alvo de mascaramento, principalmente na presença de *outliers* que, em conjunto com alguma outra variável, possa deixar o registro vulnerável. Um exemplo é um valor de renda muito alta, associada a uma unidade geográfica relativamente pequena. Duncan *et al.* (2011) resumem esta ideia quando apontam que mascarar as variáveis-chave torna mais difícil a identificação do registro, enquanto mascarar as variáveis sensíveis as fazem menos úteis para o intruso.

Por fim, há que se levar em conta que dentre as variáveis-chave, geralmente existe uma ou um subconjunto delas que é responsável pela maior parte das unicidades amostrais. Então, estas variáveis seriam as principais candidatas ao uso de métodos de mascaramento, pois minimizariam a alteração do banco de dados original (maximizariam a utilidade dos dados), dado um limite aceitável do risco de revelação.

### 6.1.2. Considerações para a escolha dos métodos de mascaramento

Como mencionado na Seção 2.10, os métodos de mascaramento podem ser divididos em duas classes: não perturbativos e perturbativos. Os primeiros consistem em supressões parciais ou reduções de detalhes dos dados originais, enquanto os segundos substituem os valores originais por meio de algum tipo de perturbação.

Com relação à escolha dos métodos a serem utilizados na base de dados, algumas considerações devem ser feitas. Em primeiro lugar, como mencionado na Seção

2.10, o tipo da variável a ser mascarada (categórica, ordinal, discreta ou contínua) deve ser levado em conta, pois, geralmente, os métodos são adequados para certos tipos específicos. Além disso, como argumenta Templ (2017), cada conjunto de dados tem suas características peculiares, assim como cada país possui regulações legais próprias, não sendo possível definir uma regra geral.

Entretanto, a literatura aponta o que é mais usualmente utilizado nas experiências internacionais. Benschop *et al.* (2019), por exemplo, argumentam que métodos como recodificação global e supressão local geralmente são aplicados nas variáveis-chave para prevenir a identificação do registro, enquanto recodificação de topo ou métodos perturbativos em outras variáveis identificadoras ou sensíveis são comumente usados para prevenir a revelação de atributo.

Com relação às variáveis-chave, Benschop *et al.* (2019) detalham que a recodificação global é geralmente o primeiro passo em um processo de anonimização. Ele é usado para reduzir o número de combinações únicas de valores das variáveis-chave. Entretanto, os autores ponderam que se os usuários necessitarem da informação mais detalhada, este método pode não ser apropriado. Após esta etapa é que se aplicaria o método da supressão local, uma vez que a recodificação anterior poderia diminuir muito as supressões necessárias, trazendo ganhos com relação à utilidade dos dados. Menos supressões também trariam ganhos em termos de tempo computacional necessário para a aplicação do método. Os autores lembram que, dependendo da complexidade do problema, o tempo de execução do algoritmo pode aumentar muito ou, no limite, não encontrar solução. Hundepool *et al.* (2012) acrescentam que modos de combinar recodificação global e supressão já são discutidos na literatura há décadas, citando como exemplo o trabalho de Waal e Willenborg (1995).

Especificamente no que concerne à etapa de supressão local, Benschop *et al.* (2019) apontam que a escolha das variáveis a serem suprimidas geralmente é guiada em função de minimizar o número de supressões e, consequentemente, a perda de informação. Adicionalmente, se uma variável é muito importante para o usuário, é possível escolher que esta não seja alvo deste método, a menos que seja estritamente necessário. Os autores ainda argumentam que caso a abordagem com recodificações e

supressões ainda não seja suficiente para alcançar o resultado desejado, métodos perturbativos poderiam ser considerados.

Templ (2017), por sua vez, traz estudos de casos nos quais são exemplificados tratamentos para diferentes tipos de pesquisa. No caso de uma pesquisa amostral domiciliar, o autor também faz uso dos métodos de recodificação global, seguido de supressão local para o tratamento de variáveis-chave selecionadas. No entanto, é feito um alerta que se existir muitas variáveis-chave categóricas (mais do que cinco, por exemplo) é possível que a recodificação não seja suficiente para reduzir o risco de revelação ao nível desejado, ou a supressão local possa acarretar grande perda de utilidade dos dados. Neste caso seria possível, por exemplo, adicionar outra camada de incerteza aos dados alterando algumas variáveis categóricas pelo método da pós-aleatorização (PRAM), descrito na Seção 2.10.2.

Usualmente não se utiliza o PRAM nas variáveis-chave, entretanto Benschop *et al.* (2019) argumentam que num cenário de revelação com muitas variáveis-chave tal que o algoritmo de supressão local não encontre solução, este método pode ser uma alternativa. Neste caso, o algoritmo consideraria um conjunto menor de variáveis-chave, enquanto as restantes seriam contempladas pelo PRAM, por exemplo. Uma vez que este método é perturbativo, o intruso até poderia fazer relações de um para um entre a informação que ele tem em sua posse e um registro do banco de dados divulgado, porém não teria mais a certeza de que o pareamento estaria correto. Entretanto, Hundepool *et al.* (2012) argumentam que para fazer inferências adequadas em um arquivo de microdados ao qual o PRAM foi aplicado – e o mesmo ocorre em outros métodos perturbativos – o usuário precisa incluir alterações sofisticadas em seus métodos usuais. Isso exige um bom conhecimento tanto do PRAM quanto da análise estatística a ser aplicada.

Como já mencionado, os métodos de mascaramento se concentram nas variáveis-chave. Isto ocorre porque elas não podem ser suprimidas na divulgação dos dados por seu alto valor analítico, ao passo que são as supostamente usadas para o intruso fazer a identificação dos registros (ELLIOT e DOMINGO-FERRER, 2018). Entretanto, as demais variáveis do banco de dados também podem ser alvo destes métodos. Nestes casos, porém, é muito difícil traçar uma estratégia geral de abordagem,

pois o motivo específico do mascaramento, assim como o tipo da variável em questão são determinantes na escolha do método. Se, por exemplo, a preocupação é com relação a um *outlier* em uma variável contínua, a recodificação de topo (ou fundo) poderia ser a solução indicada. Se o intuito é adicionar um nível extra de incerteza para o intruso, um possível recurso seria aplicar o PRAM em algumas variáveis categóricas adicionais, e assim por diante. Certamente o aspecto da utilidade dos dados deve ser também levado em conta neste processo de escolha.

## 6.2. Aplicação do mascaramento dos dados da PNAD Contínua 2019 – 2º trimestre

Esta seção apresenta um exemplo de aplicação das etapas envolvidas no processo de mascaramento de dados. Como já mencionado, optou-se por trabalhar com o cenário de revelação 2 e com o recorte geográfico mais desagregado, que subdivide a UF em domicílios localizados dentro ou fora de sua capital. O conjunto de variáveis-chave trabalhadas está disposto no Quadro 6.1, que são as mesmas especificadas do Quadro 5.2, acrescidas das concernentes à localização geográfica: UF e Capital. Estas duas últimas têm, por sua vez, suas categorias apresentadas no Quadro 5.3.

Desta forma, é necessário fazer uma observação com relação à variável Capital. O Quadro 5.3, seguindo a descrição do dicionário da pesquisa, mostra que a variável possui categorias que correspondem aos 27 municípios capitais de UF, além da possibilidade de um valor em branco caso registro não se localize nestes municípios. Entretanto, sob o ponto de vista do intruso, essa variável pode, na prática, ser reduzida para apenas duas categorias: uma delas indicando que o registro pertence ao município da capital da UF e outra – obtida pelo *missing* – indicando que o registro está em município fora desta capital. É esta configuração, representada no Quadro 6.1, que será usada nesta aplicação.

Quadro 6.1: Variáveis-chave e número de suas categorias na configuração escolhida para a aplicação dos métodos de mascaramento

| Código da variável | Descrição | Número de categorias |
|---|---|---|
| UF | Unidade da Federação | 27 |
| Capital | Indicadora de município de Capital | 2 |
| V2007 | Sexo | 2 |
| V2009 | Idade | 131 |
| V2010 | Cor/raça | 6 |
| VD3004 | Nível de instrução | 7 |
| V2001 | Tamanho do domicílio | 30 |

Fonte: PNAD Contínua 2019 – 2º trimestre

A ideia é aplicar o tratamento mais usual apontado na literatura, como já descrito, que é a combinação de recodificação global e posteriores supressões locais nas variáveis-chave. Estas variáveis são as utilizadas no cálculo das medidas do risco de revelação. Assim, após o seu mascaramento é possível recalcular tais medidas e aferir a redução de risco obtida. Este resultado deve ser analisado conjuntamente com a perda de utilidade proveniente deste processo, questão que será abordada na Seção 6.3.

É preciso destacar que, antes de quaisquer cálculos, tratamentos adicionais poderiam ser sugeridos como, por exemplo, mascaramento de *outliers* em determinadas variáveis. Entretanto, como já apontado na seção anterior, não é o escopo desta tese entrar em questões mais específicas da pesquisa, as quais devem ser alvo de ponderação em conjunto com sua equipe temática. Pretende-se aqui ilustrar, a partir deste exemplo da PNAD Contínua, o processo de mascaramento de dados.

### 6.2.1. Análise exploratória das variáveis-chave

Ao se definir pela utilização da técnica de recodificação global nos dados, é preciso então determinar quais variáveis serão mascaradas. De uma forma geral,

aquelas que apresentem categorias com pequena frequência amostral são possíveis candidatas, uma vez que estas categorias tendem a gerar unicidades ou valores de chave raros. Assim, examinar a distribuição de frequência das variáveis-chave é um importante passo inicial.

Entretanto, especificamente nesta aplicação, não será feita esta análise para as variáveis UF e Capital, afinal, o objetivo aqui é justamente testar os dados no recorte geográfico mais desagregado. Dentre as demais cinco variáveis presentes no Quadro 6.1, uma delas pertence ao nível hierárquico do domicílio ("tamanho do domicílio") e as outras quatro ao de pessoa ("sexo", "idade", "cor/raça" e "grau de instrução"). Benschop *et al.* (2019) argumentam que pelo fato das variáveis do nível mais alto serem replicadas para todos os registros do nível mais baixo, elas devem ser mascaradas primeiro.

É preciso, então, analisar primeiro a variável de tamanho do domicílio. A Tabela A.29 do Anexo A apresenta a frequência amostral desta variável para os registros dos dois níveis hierárquicos (domicílios e pessoas). É possível observar, como já previamente apontado no capítulo anterior, que registros de domicílios com muito moradores – principalmente a partir de oito – são escassos na amostra. Adicionalmente, quanto mais pessoas no domicílio, maior a probabilidade que sua estrutura (combinação de idade e sexo dos moradores, por exemplo) seja única. Assim, o risco de revelação para domicílios de grande tamanho tende a ser muito alto.

Entretanto, como também mencionado no capítulo anterior, a ideia não é adotar uma medida mais radical de suprimir estes registros. Nesta aplicação optou-se pela solução de manter os registros de pessoas dos domicílios com oito ou mais moradores, porém impossibilitando que o usuário possa saber a qual domicílio elas pertencem. Esta decisão, no entanto, permite que um intruso facilmente possa deduzir que pessoas sem vínculo a nenhum domicílio estejam residindo em um que contivesse oito ou mais pessoas. Na prática, esta medida equivale a uma recodificação global de topo e o número de categorias dessa variável cairia para oito ("1", "2", "3", "4", "5", "6", "7", "8 ou mais").

As Tabelas A30 até A33 do Anexo A, por sua vez, apresentam as frequências das variáveis-chave relativas ao nível hierárquico do indivíduo. Dentre elas, a que chama a

maior atenção é a variável idade apresentada na Tabela A33, uma vez que são observados 114 valores distintos respondidos (idades entre 0 e 114 anos, com exceção do valor 113). Isto faz com que cada categoria tenha, em média, uma menor frequência de registros na amostra e isso se agrava muito nas idades mais avançadas. Uma recodificação possível seria propor alguma configuração de grupos etários com a finalidade de formar categorias com maior volume de resposta.

De qualquer modo, mesmo em um cenário em que se tente preservar o máximo de informação possível, ainda assim parece razoável agrupar, pelo menos, os valores mais altos em uma categoria como, por exemplo, 85 anos ou mais. Gomes e Turra (2008) apontam que a contagem de indivíduos em idade avançada está sujeita a erros, muitas vezes por erro de declaração dos próprios entrevistados. Acrescentam ainda que essa é uma das principais razões pelas quais as tábuas de vida são geralmente finalizadas em 80 ou 85 anos e mais de idade. Desta forma, um agrupamento nestas categorias mais rarefeitas, traria pouca perda de informação, pois os valores individuais são questionáveis. Adicionalmente, mesmo que se pudesse ter a certeza de que as respostas são fidedignas, isso indicaria valores *outliers* que trariam grande risco de revelação para o registro.

A variável cor ou raça apresentada na Tabela A31, por sua vez, apresenta duas categorias com pouca frequência: “amarela” e “indígena”. Uma possível recodificação seria a junção das duas em uma categoria denominada, por exemplo, de “outros”. Sendo mais rigoroso, dado que essa nova categoria ainda poderia ser considerada pouco frequente, uma solução seria a recodificação conjunta de “negro”, “amarelo” e “indígena”. Entretanto, devido ao grande valor de análise para os usuários desta variável, sua recodificação – principalmente uma que seja mais radical, como esta do último exemplo – deve ser feita se absolutamente indispensável. Percebe-se ainda que existe um pequeno número de registros em que aparece a categoria “ignorado”. Dado que isto representa uma falta de informação, não é problema sob o ponto de vista do risco de revelação. Na verdade, isto é equivalente a uma supressão local na variável destes registros, omitindo informação de um intruso.

As outras duas variáveis-chave restantes não se apresentam como boas candidatas para uma recodificação. A variável sexo, de vital importância analítica para

os usuários, só possui duas categorias e ambas com grande frequência, como mostra a Tabela A30. É preciso destacar ainda que, na prática, não faz sentido recodificar uma variável de apenas duas categorias, pois a junção delas significa acabar com toda a informação. Neste caso, a solução seria remover a variável do banco de dados.

A variável do nível de instrução, por sua vez, é derivada, ou seja, ela é obtida com base em outras presentes na pesquisa. Desta forma, caso alguma categoria apresente frequência muito baixa, é preciso fazer uma análise conjunta com as variáveis das quais ela é obtida. Quaisquer modificações devem manter a consistência dos dados, assim como assegurar que um intruso não consiga recuperar a informação omitida, a partir dos valores remanescentes no arquivo. De qualquer modo, como mostra a Tabela A32, não existem categorias nesta variável com frequências raras na amostra. Deve ser ressaltado, ainda, que os registros em que consta a categoria "não aplicável" correspondem aos de pessoas com menos de 5 anos de idade, as quais não respondem o módulo de características de educação da pesquisa.

É preciso destacar que, pensando em termos de processo de pesquisa, esta análise das variáveis-chave pode ganhar um componente automatizado e/ou parametrizável. Por exemplo, definir que as variáveis com categorias abaixo de uma determinada frequência relativa possam ser candidatas à recodificação.

Por fim, deve ser salientado que a PNAD Contínua é uma pesquisa repetida ao longo do tempo. Embora nesta aplicação a análise se concentre em apenas uma divulgação trimestral da pesquisa, o ideal é que em caso de implantação do CEC esta análise leve em conta vários trimestres. É muito importante que as variáveis possam ser comparáveis ao longo das pesquisas, ou seja, não se recomenda propor configurações de recodificações de forma independente para cada divulgação.

### 6.2.2. Definindo as configurações de recodificação global e supressão local

A partir do estudo inicial das variáveis-chave, o processo de definição da configuração de mascaramento pode ser feito de forma interativa pela interface gráfica "sdcAPP" do pacote "sdcMicro" do *software* R. A cada nova junção de categorias propostas, a interface gera algumas estatísticas, que serão detalhadas nesta seção, e

recalcula automaticamente o risco de revelação para os registros de pessoa. Embora este não seja o risco que se deva considerar para definição de um registro seguro, dada a estrutura hierárquica da PNAD Contínua, ele serve como boa ferramenta de análise nesta etapa.

O processo interativo da interface gráfica dá liberdade para o analista testar inúmeras configurações diferentes em pouco tempo. A produção das estatísticas e recálculo dos riscos de revelação de cada nova junção de categorias, levou menos de 5 segundos no computador pessoal especificado na Seção 5.5. Desta forma, esta seção conterá um resumo da análise feita pelo autor, porém tendo o cuidado de deixar ilustrado todas as etapas que envolveram este processo.

A seção anterior apontou possíveis recodificações nas variáveis tamanho do domicílio, idade e cor/raça. Na primeira configuração testada optou-se por preservar o máximo de informação da variável idade, apenas criando uma categoria final para pessoas de 85 anos ou mais de idade. Já para tamanho do domicílio considerou-se o limite superior de 8 ou mais pessoas e, por fim, para cor/raça a junção de “indígena” com “amarelo”. As recodificações foram testadas separadamente e em conjunto, e a Tabela 6.1 apresenta uma síntese dos resultados obtidos.

Posto que a prioridade era recodificar as variáveis tamanho de domicílio e/ou idade, preservando a variável cor/raça, essas duas primeiras foram testadas em configurações de recodificação de forma separada e conjunta. Assim, a chamada “configuração 1” (C1) foi a que juntou apenas os valores mais altos da variável idade em uma categoria de 85 anos ou mais. A “configuração 2” (C2), por sua vez, foi a que agregou apenas os domicílios com 8 ou mais moradores. Já a “configuração 3” (C3) considerou ambas as recodificações propostas em C1 e C2. Por fim, a “configuração 4” (C4), englobou C3 adicionando a recodificação proposta para a variável cor/raça. Para estas configurações, são apresentadas na Tabela 6.1 o número de registros de pessoas acima de determinados riscos de revelação e o número de unicidades amostrais.

Tabela 6.1: Número de registros de pessoas acima de determinados limites de riscos de revelação e de unicidades amostrais, para os dados originais da PNAD Contínua e para as configurações de recodificação inicialmente propostas.

| Estatísticas | Número de registros | | | | |
|---|---|---|---|---|---|
| | Dados originais | Configurações de recodificação | | | |
| | | C1 | C2 | C3 | C4 |
| Risco de revelação | | | | | |
| >10,0% | 910 | 896 | 846 | 832 | 831 |
| > 9,0% | 1.345 | 1.327 | 1.256 | 1.238 | 1.235 |
| > 8,0% | 2.473 | 2.444 | 2.319 | 2.290 | 2.283 |
| > 7,5% | 3.207 | 3.169 | 2.999 | 2.961 | 2.951 |
| Unicidades amostrais | 135.780 | 133.171 | 132.255 | 129.643 | 129.484 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

É possível observar, de acordo com a Tabela 6.1, que os dados originais apresentavam 910 registros de pessoas com risco de revelação acima de 10%. Conforme o parâmetro do risco diminui, evidentemente o número de registros acima deste limite aumenta até atingir um total de 3.207, para um risco de 7,5%. A interface gráfica permite ao usuário escolher livremente os parâmetros de risco.

Os dados originais apresentavam também mais de 135 mil unicidades amostrais. Neste caso, o pacote "sdcMicro" disponibiliza automaticamente o número de registros violando a *k*-anonimidade, definida na Seção 2.11, para valores de *k* iguais a 2, 3 e 5. Como a PNAD Contínua possui uma fração amostral pequena, tal que mesmo uma unicidade amostral pode não ser problemática em termos de risco de revelação, optou-se por utilizar somente como estatística de análise o número de unicidades, ou seja, o equivalente a *k* igual 2.

Ao se comparar as estatísticas dos dados originais, com os dados modificados para cada uma das quatro configurações de recodificação observa-se, como esperado, que quanto maior a carga de recodificação, menor o número de registros acima dos parâmetros máximos estabelecidos de risco e de unicidades amostrais. A análise entre C1 e C2 mostra que a recodificação da variável de tamanho do domicílio tem um maior poder de redução de risco, em relação à configuração proposta para a variável idade. A

comparação entre C3 e C4, por sua vez, revela que a recodificação da variável cor/raça altera muito pouco as estatísticas apresentadas.

As análises prévias mostraram que variáveis de idade e tamanho de domicílio devem ser recodificadas, pelo menos, em seus valores extremos. Assim, descartou-se trabalhar como C1 e C2, uma vez que ambas as configurações contemplam apenas uma das variáveis. Este fator, adicionado aos poucos ganhos observados para C4, optou-se por preservar a variável cor/raça e trabalhar com C3 para a etapa de supressões locais.

Templ (2017) aponta que existem duas abordagens para a implementação da supressão local. A primeira é definir um parâmetro *k* e tentar alcançar a respectiva *k*-anonimidade, com o menor número possível de valores suprimidos, o que vai demandar o uso de algum algoritmo heurístico. Na segunda abordagem é escolhido um valor máximo de risco de revelação para o registro individual (sem hierarquia) e todos os registros com risco acima desse valor são suprimidos para uma determinada variável escolhida. Após esta primeira supressão o risco pode ser recalculado e se persistirem registros ainda acima desse limite, nova variável pode ser escolhida e assim por diante.

Como já foi comentado, a PNAD Contínua possui uma fração amostral pequena e mesmo uma unicidade amostral pode não ser problema. Desta forma, a primeira abordagem não faz sentido neste caso, pois acarretaria uma perda de informação muito acima do necessário. Mesmo que fosse escolhido o menor valor de *k* possível, ou seja, *k*=2 para eliminar as unicidades amostrais, isto significaria alterar informação em aproximadamente 130 mil registros, como mostra a Tabela 6.1. Por isso, a segunda abordagem será, então, utilizada.

Entretanto, já foi mencionado acima que uma característica desta abordagem – e que acaba sendo uma limitação dela – é que o parâmetro de risco a ser fixado é referente ao registro individual. Para o caso da PNAD Contínua o que se deve buscar, em termos de limite máximo de risco aceitável, é relativo ao risco hierárquico. Nesta aplicação, por exemplo, deseja-se que nenhum registro apresente risco de revelação acima de 20%. Assim, é preciso arbitrar um valor abaixo destes 20% para o risco individual, tendo em vista a fórmula de cálculo do risco hierárquico apresentado na Seção 5.3. Um valor relativamente “alto” (mais próximo dos 20%) implica menores alterações no banco de dados, porém com maior probabilidade de registros que ao final

do processo apresentem risco hierárquico acima do limite desejável. De forma análoga, quanto menor for o valor arbitrado, menor será a probabilidade de registros ultrapassarem o limite máximo, mas ao custo de mais supressões em registros que não necessitariam delas.

Para ilustrar essa afirmação, ao se supor um valor máximo de 1% para o risco individual, dificilmente o limite de 20% para o risco hierárquico seria violado por algum registro. Afinal, como apresentado na Seção 5.3, o risco hierárquico é igual a 1 menos a probabilidade de nenhuma pessoa ter sido identificada no domicílio. Entretanto, vários registros de pessoas acima deste limite – principalmente quanto mais perto deste valor de 1% – poderiam apresentar risco hierárquico abaixo de 20% mesmo sem qualquer tipo de supressão local. A Tabela 6.1 mostrou que ao se diminuir o parâmetro de risco o número de registros que necessitam de supressões locais aumenta quase que exponencialmente.

Não faz parte do escopo desta tese desenvolver um método que determine um possível valor "ótimo" para a determinação deste parâmetro, embora sem dúvidas seja um tema relevante para pesquisas futuras. Arbitrou-se aqui um valor que visa alterar uma pequena quantidade relativa de registros, ao passo que elimine quase todos os casos de risco hierárquico acima de 20%, ao fim do processo. Nos testes realizados pelo autor, que utilizaram os parâmetros da Tabela 6.1, optou-se por trabalhar com o limite individual de 7,5%, que implica na supressão de pouco mais de 0,5% dos registros totais (2.961) para C3.

A escolha que deve ser feita a seguir é o ordenamento das variáveis para a supressão. Conforme observado anteriormente, nesta abordagem é feita a supressão de uma variável para os registros com risco acima do parâmetro arbitrado e, restando casos ainda acima desse limite, nova variável é selecionada e assim por diante. Aqui é importante outra vez o conhecimento da equipe temática da pesquisa para a definição desta ordem, em função da importância das variáveis para os usuários ou outros critérios.

Nesta tese, analisou-se as variáveis do Quadro 6.1 e excluindo as variáveis de idade e tamanho do domicílio, já recodificadas, assim como a variável sexo por sua importância analítica para o usuário, restaram as variáveis referentes a cor/raça,

instrução e indicadora de município da capital[10]. Esta última é a que se pretendeu preservar a informação o máximo possível, uma vez que é justamente este recorte geográfico que está sendo posto à prova nesta aplicação. Como a variável de instrução é derivada, ou seja, uma alteração nela implica alterações em mais variáveis no banco de dados, esta ficou como segunda prioridade. Sendo assim, a primeira variável a ser suprimida será cor/raça. Os resultados desta etapa estão dispostos no Quadro 6.2.

Quadro 6.2: Resultados da etapa de aplicação da supressão local, para C3 em registros com r > 0,075

| Passo | Variável suprimida | Valores suprimidos | Unicidades amostrais | Registros não seguros |
|---|---|---|---|---|
| 1 | Cor/raça (V2010) | 2.961 | 127.593 | 1.660 |
| 2 | Instrução (VD3004) | 1.660 | 124.194 | 331 |
| 3 | Capital | 331 | 123.677 | 104 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Como é possível observar no Quadro 6.2, após a supressão dos valores da variável cor/raça para os 2.961 registros com risco acima do limite estabelecido, 1.660 deles continuaram acima deste patamar. Desta forma, o segundo passo foi aplicar nestes registros a supressão dos valores da variável de educação, mas, ainda assim, 331 deles continuaram com risco individual acima de 0,075.

É preciso ressaltar, antes de comentar o terceiro passo, que há uma peculiaridade em relação a uma possível supressão na variável indicadora de capital. Ela é, na prática, uma variável do nível hierárquico superior. O domicílio está localizado em determinado local e todas as pessoas que habitam nele terão o mesmo valor atribuído em seu registro. Assim, cuidados adicionais deverão ser tomados neste caso, os quais serão descritos ainda nesta seção. Por enquanto, o resultado importante, como mostrado no Quadro 6.2, é que mesmo se fosse ignorada esta peculiaridade, a supressão dos valores desta terceira variável ainda manteria 104 registros com risco acima de 7,5%. O relatório apresentado pela interface gráfica mostrou ainda que seis

[10] A variável UF não foi considerada porque, embora componha conjuntamente com a variável indicadora de capital este recorte geográfico, ela isoladamente é a indicadora de outro recorte geográfico.

registros dentre estes 104, possuíam risco individual acima de 20%, que seria o limite pretendido para o risco hierárquico. A solução neste caso é rever a configuração de recodificação global, pois ela se mostrou insuficiente.

Neste ponto, as estatísticas fornecidas automaticamente pela interface gráfica ajudam a diagnosticar o problema. A comparação entre a Tabela 6.1 e o Quadro 6.2 mostra que o número de unicidades amostrais diminuiu em uma pequena proporção, em relação aos dados originais. Isto se deve, em grande medida, ao grande número de categorias da variável idade. Mesmo ao se agregar as pessoas com 85 anos ou mais, a variável conta ainda com 86 possíveis categorias (0 a 85 ou mais).

Desta forma, uma primeira solução testada foi categorizar esta variável em grupos quinquenais. Ainda assim, os grupos finais seriam rarefeitos e persistiriam problemas de redução do risco de revelação mesmo com posteriores supressões locais. Foi preciso então, após testes realizados pelo autor, recodificar o último grupo etário de modo que ele passasse a englobar as pessoas com 65 anos ou mais de idade.

Deve ser ressaltado que quaisquer recodificações devem levar em conta possibilidades do intruso contorná-las. No caso da PNAD Contínua, a parte 4 do questionário, apresentada no Capítulo 5, refere-se às características de trabalho das pessoas de 14 anos ou mais de idade. Assim, considerando grupos quinquenais de idade, o intruso poderia facilmente desmembrar indivíduos com 14 anos de dentro do grupo “10 a 14 anos”, pelo fato do registro ter respostas nesta parte 4. Neste caso, então, essa foi a única categoria etária que permaneceu com a idade individual. Assim, a nova configuração de recodificação global, chamada de C3’, engloba as mesmas duas variáveis da configuração anterior: tamanho do domicílio e idade. A mudança ocorre na segunda variável, tal que são considerados os seguintes quinze grupos etários: “0 a 4”, “5 a 9”, “10 a 13”, “14”, “15 a 19”, “20 a 24”, “25 a 29”, “30 a 34”, “35 a 39”, “40 a 44”, “45 a 49”, “50 a 54”, “55 a 59”, “60 a 64” e “65 ou mais”.

Adicionalmente será considerada ainda uma outra configuração de recodificação, chamada C3’’, que possui a mesma recodificação global de C3 e C3’ para a variável de tamanho do domicílio, mas utiliza os grupos etários da PNAD Contínua atualmente disponíveis no SIDRA (Sistema IBGE de Recuperação Automática). O SIDRA é uma ferramenta digital que pode ser acessada por navegador de internet e permite ao

usuário consultar dados de várias pesquisas realizadas pelo IBGE, de forma simples e rápida. Neste caso, esta ferramenta permite que o usuário faça tabulações com a variável idade, mas não com os valores individuais. Os resultados são retornados em seis classe de grupos etários, que são: “0 a 13”, “14 a 17”, “18 a 24”, “25 a 39”, “40 a 59” e “60 ou mais”.

A vantagem de C3’’ reside no fato dos grupos manterem um padrão já conhecido pelos usuários, assim como estes grupos etários terem sido formados desta maneira passando pelo crivo da área temática. A desvantagem, em relação à C3’, é o fato do menor número de categorias trazer uma maior perda de informação. Nesta tese julgou-se interessante trabalhar com duas configurações em paralelo, uma vez que a comparação entre as duas ilustrará com mais detalhes os quesitos analisados neste capítulo (redução do risco e perda de informação), após o mascaramento dos dados.

Observa-se, com base na Tabela 6.2, que para C3’ e C3’’ há uma expressiva redução nos registros de pessoas acima dos parâmetros máximos estabelecidos de risco e de unicidades amostrais, em relação aos dados originais e para C3. Como esperado, essa queda é ainda maior para C3’’ que apresenta uma recodificação da variável idade em menor número de categorias. Estes resultados mostram a relevância desta variável no risco de revelação dos registros, sendo preciso medir seus efeitos antes da disseminação dos dados, principalmente se a intenção for disponibilizar as idades individuais dos respondentes.

Os resultados da etapa de supressão local para C3’ e C3’’ estão dispostos, respectivamente, nos Quadros 6.3 e 6.4. No caso de C3’ observa-se que após a supressão dos valores da variável cor/raça para os 987 registros com risco acima do limite estabelecido, 306 deles continuaram acima deste patamar. Ao aplicar nestes registros restantes a supressão dos valores da variável de educação, ainda 2 deles continuaram com risco individual acima de 0,075. Por fim, ao se suprimir o valor da variável capital destes dois registros, não restou mais nenhum com risco acima do limite proposto.

Tabela 6.2: Número de registros de pessoas acima de determinados riscos de revelação e de unicidades amostrais, para os dados originais da PNAD Contínua, e as configurações C3, C3’ e C3’’.

| Estatísticas | Número de registros | | | |
|---|---|---|---|---|
| | Dados originais | Configurações de recodificação | | |
| | | C3 | C3’ | C3’’ |
| Risco de revelação | | | | |
| >10,0% | 910 | 832 | 318 | 170 |
| > 9,0% | 1.345 | 1.238 | 458 | 251 |
| > 8,0% | 2.473 | 2.290 | 787 | 422 |
| > 7,5% | 3.207 | 2.961 | 987 | 506 |
| Unicidades amostrais | 135.780 | 129.643 | 36.829 | 17.185 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Quadro 6.3: Resultados da etapa de aplicação da supressão local, para C3’ em registros com r > 0,075.

| Passo | Variável suprimida | Valores suprimidos | Unicidades amostrais | Registros não seguros |
|---|---|---|---|---|
| 1 | Cor/raça (V2010) | 987 | 35.784 | 306 |
| 2 | Instrução (VD3004) | 306 | 34.836 | 2 |
| 3 | Capital | 2 | 34.831 | 0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Como já destacado, a variável indicadora de capital refere-se ao nível hierárquico domicílio, tal que todas as pessoas que habitam nele terão o mesmo valor atribuído em seu registro. Desta forma, a simples supressão em um registro de pessoa pode não ter valia, uma vez que o intruso consegue recompor este valor pela resposta contida nas demais pessoas daquele domicílio. Existem duas exceções para esta situação: um domicílio com apenas um morador, ou um domicílio com oito ou mais moradores, já que a recodificação proposta para este último caso retira a possibilidade de identificação das pessoas em relação ao domicílio. Nesta aplicação específica, dentre os dois registros

com a variável de capital suprimida, um correspondia a uma pessoa que morava sozinha, e o outro era de pessoa em domicílio de tamanho oito ou mais[11], logo nenhum outro procedimento para uma correta supressão foi necessário. Caso contrário medidas adicionais poderiam ser tomadas como, por exemplo, suprimir a informação para todos os outros moradores daquele domicílio.

Quadro 6.4: Resultados da etapa de aplicação da supressão local, para C3'' em registros com r > 0,075.

| Passo | Variável suprimida | Valores suprimidos | Unicidades amostrais | Registros não seguros |
|---|---|---|---|---|
| 1 | Cor/raça (V2010) | 506 | 16.584 | 98 |
| 2 | Instrução (VD3004) | 98 | 16.258 | 1 |
| 3 | Capital | 1 | 16.257 | 0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Com relação à C3'' verifica-se que após a supressão dos valores da variável cor/raça para os 506 registros com risco acima do limite estabelecido, 98 deles persistiram acima deste patamar. Ao aplicar nestes registros restantes a supressão dos valores da variável de educação, apenas 1 continuou com risco individual acima de 0,075. Este registro era relativo a um morador em domicílio com 8 ou mais pessoas, então não foi necessária outra medida além da supressão simples deste valor. Após esta ação, não restou mais nenhum registro com risco acima do limite proposto.

Deve ser destacado que a ordem das variáveis a serem alvo da supressão local, influenciam o total de valores suprimidos. Tomando como exemplo C3'', se o passo 1 consistisse na supressão da variável de educação, 88 dos 506 registos permaneceriam com risco acima do patamar de 0,075, o que corresponde a um valor pouco menor do que os 98 observados quando se optou pela variável cor/raça no passo inicial. Desta forma, este pode ser também um critério levado em conta pelo usuário na escolha do ordenamento das variáveis, ou seja, escolher a permutação que dará a menor perda de informação.

[11] A interface gráfica apresenta os registros com risco de revelação acima de quaisquer limites que o usuário desejar. Assim, é possível saber imediatamente suas características sem precisar de programações adicionais.

As combinações de configurações de recodificação e supressões locais garantem, como já destacado anteriormente, que os registros individuais tenham risco de revelação abaixo de 0,075. Assim, após esta etapa é preciso verificar se há registros ainda acima do risco de revelação hierárquico proposto de 0,2. A Tabela 6.3 apresenta número de domicílios, e o total de seus respectivos residentes, acima deste limite para os dados originais e após as duas propostas de mascaramento.

Tabela 6.3: Total de registros de domicílios e pessoas com risco de revelação acima de 0,2 para os dados originais e após as duas propostas de mascaramento.

| Tipo de registro | Registros com r > 0,2 | | |
|---|---|---|---|
| | Dados originais | Dados mascarados | |
| | | C3'+supressões | C3''+supressões |
| Domicílios | 1.333 | 51 | 20 |
| Pessoas | 9.150 | 329 | 129 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

É possível observar, com base na Tabela 6.3, que embora o total de registros de domicílios, e, consequentemente, o total de pessoas que neles residem, com risco acima de 0,2 tenha diminuído muito após o mascaramento dos dados ainda restam registros não seguros. Este número, como esperado, é menor para C3'' por ser a configuração de recodificação mais restritiva para a variável idade. A Tabela 6.4 apresenta os registros não seguros segundo o seu número de moradores.

A Tabela 6.4 mostra que os domicílios com risco de revelação acima de 0,2 possuem entre 4 e 7 moradores, sendo a maior parte deles concentrados neste limite superior. Este comportamento era esperado, uma vez que ao se assegurar somente o limite máximo para o risco individual, quanto mais registros de pessoas dentro de um domicílio, maior é a probabilidade que o risco hierárquico seja violado. Em outras palavras, embora nenhum deles possua risco individual acima de 0,075, o produto deste risco conjunto pode ultrapassar o valor de 0,2.

Tabela 6.4: Total de registros de domicílios com risco de revelação acima de 0,2 para as duas propostas de mascaramento, por tamanho do domicílio.

| Tamanho do domicílio | Registros com r > 0,2 | |
|---|---|---|
| | C3'+supressões | C3''+supressões |
| 4 | 1 | 1 |
| 5 | 5 | 1 |
| 6 | 15 | 6 |
| 7 | 30 | 12 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

É preciso destacar que a decisão de remover o identificador de domicílio para registros com tamanho igual ou maior que 8, faz com que não haja mais hierarquia nesse grupo. Assim, é o risco individual que deve ser considerado para os registros de pessoas nele contido. Desta forma, nesse subconjunto garante-se que não há registros com risco acima de 0,075, por isso o tamanho máximo de domicílio observado na Tabela 6.4 foi igual a 7.

A Tabela A.34, no Anexo A, apresenta a distribuição dos registros de domicílio não seguros, para as duas propostas de mascaramento, por Unidade da Federação. É possível observar que a maioria destes registros está localizada na Região Norte, principalmente no Acre, seguido de Roraima e Amapá. Estas três UF eram as que, de uma forma geral, já apresentavam os maiores números de registros não seguros para os dados originais, conforme analisado no Capítulo 5.

Identificados estes domicílios considerados não seguros após a recodificação global e supressão local, é necessário tomar alguma medida adicional para sanar este problema. Nesta tese optou-se por dar-lhes o mesmo tratamento que os domicílios com oito ou mais moradores, ou seja, remover a possibilidade de identificação do domicílio. Esta abordagem tem uma vantagem adicional: tira a certeza do intruso com relação ao tamanho mínimo domiciliar – antes igual a oito – dos registros de pessoas não identificados a um domicílio. Entretanto, independente da escolha de tratamento feito nesta fase, é preciso mensurar também esta perda de informação adicional, além das produzidas pelas recodificações e supressões iniciais.

## 6.3. Utilidade dos dados e perda de informação

Após a aplicação dos métodos de mascaramento nos dados originais, é necessário medir o seu efeito, ou seja, a perda de informação resultante. Duncan *et al.* (2011) argumentam que existem duas abordagens, que podem ser complementares, para a estimação desta perda: uma que compara diretamente os dados originais com os mascarados; e outra que compara certos indicadores da pesquisa calculados por estes dois tipos de dados. Templ (2017) faz essa mesma subdivisão apontando que enquanto a primeira seria mais usual, embora com certa limitação por ser genérica, a segunda poderia ter uma utilidade maior para o usuário, ao medir diferenças para os indicadores mais importantes.

Hundepool *et al.* (2012), no entanto, argumentam que geralmente os usos potenciais de dados são muito diversos e pode ser difícil identificar todos. Adicionalmente, ponderam que mesmo que todos os usos pudessem ser identificados, a emissão de várias versões do mesmo conjunto de dados original, para que a i-ésima versão tenha uma perda de informação otimizada para i-ésimo uso de dados, pode resultar em alguma revelação inesperada. Desta forma, os autores afirmam que dificilmente o mascaramento dos dados é feito visando um uso específico dos dados, logo as medidas genéricas de perda de informação são desejáveis.

De forma análoga ao que foi observado para os métodos de mascaramento, não há uma regra geral para a escolha das medidas de perda de informação a serem utilizadas. A peculiaridade do conjunto de dados mascarado deve ser levada em conta. Por exemplo, Hundepool *et al.* (2012) e Benschop *et al.* (2019) apresentam um conjunto de medidas que são adequadas para variáveis contínuas e outro para variáveis categóricas. Enquanto no primeiro caso é possível fazer comparações de médias, covariâncias e correlações das variáveis nos arquivos original e mascarado, no segundo caso a comparação pode ser feita, por exemplo, por meio de tabelas de contingência, contagem de valores faltantes, dentre outros.

Além do tipo de variável, o método de mascaramento utilizado também deve ser levado em conta na escolha das medidas de perda de informação. A contagem de valores faltantes, por exemplo, é relevante caso o banco de dados tenha sido tratado

com supressões locais. Por outro lado, comparações por meio de tabelas de contingência são adequadas caso os dados tenham sido alvo de métodos perturbativos, tal que os valores das categorias de determinadas variáveis tenham sido alterados. A Seção 6.3.1 apresenta algumas medidas descritas na literatura, indicando as situações em que eles podem ser utilizados.

### 6.3.1. Medidas de perda de informação

Como previamente observado na seção anterior, os diferentes autores não necessariamente classificam as possíveis abordagens para mensurar a perda de informação da mesma maneira. De modo análogo, o rol de métodos elencados nem sempre é o mesmo entre os autores. Desta forma, é aqui apresentado um conjunto, que não pretende ser exaustivo, de possibilidades de mensuração da perda de informação presentes na literatura.

a) Medidas de utilidade com base nas necessidades do usuário final: essa é a "segunda abordagem" categorizada por Duncan *et al.* (2011) e Templ (2017) mencionada na seção anterior. Para estes autores, todas as demais possíveis medidas a serem apresentadas nesta seção estariam dentro do escopo da "primeira abordagem".

Esta abordagem parte do pressuposto que embora nem todas as necessidades e usos dos dados de uma pesquisa possam ser inventariados, muitas vezes é possível identificar usos mais frequentes ou características importantes, que podem ser avaliadas antes e depois da anonimização. Um exemplo seria indicadores de pobreza numa pesquisa de rendimento (BENSCHOP *et al.,* 2019).

Assim, o produtor dos dados poderia selecionar estas estimativas mais importantes, chamados de "indicadores de *benchmarking*" ou "indicadores de qualidade". Após a escolha dos indicadores, o produtor dos dados definiria um critério para compará-los (antes e após o mascaramento). Esta comparação poderia ser com base em suas propriedades estatísticas, tais como estimativas pontuais, sobreposição dos intervalos de confiança, dentre outros. A partir destes resultados seria feita a avaliação sobre se a utilidade de dados protegidos é boa o suficiente para a divulgação (TEMPL, 2017).

b) Medidas genéricas de utilidade (ou de perda de informação): são as que comparam os valores ou algumas estatísticas dos dados originais com os mascarados. Todas as medidas são aplicadas a posteriori, pois medem a utilidade após o processo de mascaramento e requerem os dados antes e depois da sua execução. (BENSCHOP *et al.,* 2019). Templ (2017), por sua vez, denomina esta abordagem de "comparações relativas ao elemento" (tradução livre de "*element-wise comparisons*").

Estas medidas abrangem as já citadas contagem de valores faltantes e comparação por meio de tabelas de contingência. Outra medida também utilizada para variáveis categóricas é a contagem de registros modificados (geralmente calculada por variável). Para variáveis contínuas existe a possibilidade da comparação das estatísticas como a média, variância, covariância e estrutura de correlação, das variáveis mais importantes no conjunto de dados. Alternativas mais sofisticadas para este tipo de variáveis abrangeriam o cálculo de medidas de distância ou autovalores entre os dados antes e após o mascaramento, por exemplo (BENSCHOP *et al.,* 2019).

c) Modelagem estatística: é possível a utilização de modelos estatísticos de diferentes maneiras. Benschop *et al.* (2019) destacam uma abordagem mais direta que consiste em estimar modelos para ambos os conjuntos de dados e comparar os parâmetros da regressão. Isto possibilitaria comparar também relações com variáveis não contínuas, por exemplo, ao introduzir variáveis *dummy* na regressão. Devido ao caráter singular de um modelo regressão específico, este tipo de método é mais utilizado quando se sabe a finalidade para qual os dados serão utilizados.

Uma abordagem mais elaborada é o uso de escores de propensão, como descrito por Templ (2017). A ideia consiste em juntar os conjuntos de dados originais e mascarados e, em seguida, criar uma variável indicadora com valor "1" para os registros do primeiro conjunto e "0" para os do segundo. A seguir, um modelo de regressão logística é ajustado usando esta indicadora como variável resposta. As previsões deste modelo são então comparadas com a proporção (geralmente 1/2) de "0" e "1" observadas. Quanto mais próxima forem as proporções estimada e observada, maior será a utilidade dos dados.

d) Entropia: Ao desenvolver um modelo matemático para a área de comunicação, Shannon (1948) utiliza a palavra "entropia" para caracterizar uma medida

de informação associada à ideia de incerteza (MAGOSSI e PAVIOTTI, 2019). Na teoria da informação, a entropia de Shannon quantifica o conteúdo médio de informação que um receptor perde ao não conhecer o valor da variável aleatória (AGAFITEI e DEFAYS, 2011).

Hundepool *et al.* (2012) citam trabalhos De Waal e Willenborg (1999), Kooiman *et al*. (1998) e Willenborg De Waal (2001), em que se utiliza a entropia de Shannon para mensurar a perda de informação decorrente do uso de alguns métodos de mascaramento. Utiliza-se o conceito de entropia em CEC, partindo-se do pressuposto que o processo de mascaramento é um ruído que seria adicionado aos dados originais, no caso destes serem transmitidos por um meio com interferências. Os autores argumentam que embora o uso da entropia seja interessante sob o ponto de vista teórico, sua interpretação em termos de perda da informação de dados é menos óbvia do que para as medidas já apresentadas.

### 6.3.2. Cálculo das medidas de perda de informação para a PNAD Contínua 2019 – 2º trimestre

A adequação das medidas de perda de informação depende das variáveis mascaradas e dos métodos utilizados. Assim como exposto na Seção 6.2, foram utilizados métodos de mascaramento não perturbativos (recodificação global e supressão local) em cinco variáveis-chave. Desta forma, as abordagens apresentadas na seção anterior serão analisadas à luz da presente aplicação.

A PNAD Contínua, como descrito no Capítulo 4, tem como principal objetivo fornecer informações sobre a inserção da população no mercado de trabalho. O questionário da pesquisa é dividido em cinco partes mais uma seção de variáveis derivadas, tal como apresentado na Tabela 4.2. Nenhuma variável relacionada às partes 4 e 5 do questionário – características de trabalho e rendimentos de outras fontes, respectivamente – foi alvo de mascaramento, ou seja, todos os indicadores de trabalho e renda permaneceram inalterados[12]. Sendo assim, as principais estimativas da pesquisa, sob o ponto de vista dos usuários, não sofreram quaisquer alterações. Em

[12] Caso algum indicador fosse calculado para domínios referentes a uma variável mascarada (por exemplo, idade) poderia haver restrições (neste caso, limitada às classes ou soma de classes recodificadas).

outras palavras, não é preciso fazer uma análise a partir de “indicadores de qualidade”, uma vez que não houve perda de informação em relação a este aspecto.

A abordagem por modelagem estatística é utilizada principalmente quando são aplicados métodos de mascaramento perturbativos, o que não é o caso deste conjunto de dados. Desta forma, a análise da perda de informação será feita por meio de medidas genéricas – que variam conforme o tipo de variável e método de mascaramento aplicado – e pelo cálculo de uma função de entropia.

No processo de mascaramento dos dados, a análise foi iniciada pelas variáveis do nível hierárquico superior, tal que se optou por remover a identificação das pessoas aos domicílios com tamanho maior ou igual a 8. Uma vez que o intruso poderia recompor essa falta de informação, tratou-se como uma recodificação global de topo (categoria 8 ou mais moradores). Entretanto, após a última etapa de mascaramento, os registros de pessoas com risco hierárquico que persistiam acima de 0,2 também foram desidentificados de seus respectivos domicílios. Isso traz uma nova fonte de incerteza ao intruso que não pode mais garantir um tamanho mínimo do domicílio. Assim, para fins de perda de informação, a análise desta variável se assemelha a uma supressão local (falta de informação) e não mais a uma recodificação de topo. Desta forma, a análise da variável de tamanho do domicílio será feita em conjunto com as demais variáveis que se submeteram ao processo de supressão local.

A etapa seguinte do processo de mascaramento foi referente à recodificação global da variável idade. A interface gráfica “sdcAPP” disponibiliza automaticamente três medidas de perda de informação que comparam a variável antes e após o processo de recodificação, são elas: o número total e os tamanhos médios e mínimos das categorias. Estes resultados, para ambas as configurações propostas, estão dispostos na Tabela 6.5.

Na etapa de mascaramento foram propostas duas configurações, denominadas C3’ e C3’’, tal que esta última possui um menor número de categorias e, consequentemente, traz consigo uma maior perda de informação. Os resultados da Tabela 6.5 mostram que nos dados originais cada categoria da variável possuía, em média, 4.836 registros de pessoas. Este valor aumentou para 36.757 e 91.981 registros para C3’ e C3’’, respectivamente. É possível observar também que nos dados originais

existiam categorias – como já mostrado na Tabela A33 (no Anexo A) – com apenas um registro de pessoa. Após a recodificação a categoria mais rarefeita de C3’ tem 8.632 registros, enquanto à de C3’’ 34.874 registros. Esta última medida também pode ser utilizada para o produtor dos dados avaliar questões relativas à confidencialidade dos mesmos. Por exemplo, é possível definir um critério de tamanho mínimo de registros em uma categoria para cada variável destinada à disseminação pública.

Tabela 6.5: Medidas genéricas de perda de informação relativas à recodificação global da variável idade, para as configurações C3’ e C3’’.

| Estatísticas sobre as categorias da variável idade | Dados originais | Dados mascarados | |
|---|---|---|---|
| | | C3’ | C3’’ |
| Número total | 114 | 15 | 6 |
| Tamanho médio | 4.836 | 36.757 | 91.891 |
| Tamanho mínimo | 1 | 8.632 | 34.874 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Na etapa subsequente do processo de mascaramento, aplicou-se a supressão local em algumas variáveis-chave. A medida genérica mais intuitiva e usual de perda de informação para este caso é a comparação dos valores faltantes antes e após o mascaramento. A Tabela 6.6 apresenta estes resultados para ambas as configurações propostas.

A referida tabela mostra que, para todas as variáveis, em C3’’ há menos registros com supressões locais em relação à C3’, ou seja, menor perda de informação por este método de mascaramento. Isto é decorrência da recodificação global mais restritiva desta configuração. Como já mencionado anteriormente, a opção por remover o identificador do domicílio dos registros de pessoas, implica em valores faltantes para a variável de tamanho do domicílio. Os números da Tabela 6.6 correspondem aos 15.650 registros de pessoas em domicílios com tamanho maior ou igual a 8, acrescidos dos 329 e 129 registros para C3’ e C3’’, respectivamente, conforme mostrado na Tabela 6.3, que persistiram com risco hierárquico acima de 0,2 após a supressão local.

Tabela 6.6: Contagem de valores faltantes para as variáveis com valores suprimidos, para as configurações C3’ e C3’’.

| Variáveis | Registros de pessoas | Registros de pessoas com valores faltantes | | |
|---|---|---|---|---|
| | | Dados originais | Dados mascarados | |
| | | | C3’+supressões | C3’’+supressões |
| Tamanho Dom. | 551.348 | 0 | 15.979 | 15.779 |
| Cor/raça | 551.348 | 39 | 1.026 | 545 |
| Instrução | 551.348 | 0 | 306 | 98 |
| Capital | 551.348 | 0 | 2 | 1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Deve ser ressaltado que a opção por remover a identificação do domicílio de alguns registros, pode gerar outras supressões que devem ser contabilizadas nesta etapa também. Em outras palavras, não bastaria suprimir os valores da variável de tamanho do domicílio se o intruso pudesse recompor a identificação por meio de outras variáveis. No caso da PNAD Contínua, as poucas variáveis que poderiam servir para tal fim se encontram nas partes 1 (Identificação e Controle) e 2 (Características Gerais dos Moradores) de seu questionário.

A parte 1 contempla as variáveis de localização geográfica que serão mantidas, e as de controle que identificam de fato o domicílio devendo, então, ser suprimidas. Estas últimas (número de seleção do domicílio, painel, etc.) não trazem informação relevante ao usuário, além da impossibilidade de identificação do domicílio, já contabilizada pela variável de tamanho do domicílio. O mesmo caso se aplica para a V2003 (número de ordem da pessoa), na parte de características gerais dos moradores.

Na parte 2 há também a V2005 (condição do domicílio) que poderia ser suprimida de parte deste subconjunto de registros. Um exemplo é manter a informação no registro quando se trata pessoa de referência, que é muito útil para estudos socioeconômicos, mas não possibilitaria uma possível recomposição do domicílio. De toda maneira, se alguma medida de supressão fosse tomada com relação a V2005, seria preciso a inclusão de sua medida de perda de informação. Isso vale igualmente para as variáveis derivadas destas citadas. Por exemplo, quaisquer modificações em V2005

devem ser análogas em sua derivada VD2002, que possui o mesmo nome de “condição no domicílio”, mas com menos categorias.

O produtor dos dados pode usar a análise do número de valores faltantes para determinar um valor máximo percentual aceitável, ou seja, acima deste limite a utilidade da variável seria pouca para uma possível divulgação. Uma vez que os registros com valores faltantes da variável de tamanho do domicílio ficaram abaixo dos 3%, nesta aplicação não será contabilizada a perda de informação decorrente de uma possível supressão da variável de condição do domicílio. Um primeiro motivo é que, pelo exposto no parágrafo anterior, o montante de registros suprimidos seria um subconjunto dos verificados em V2001, não inviabilizando a utilidade da variável. O segundo é que em termos da comparação entre C3’ e C3’’, essa questão não faz diferença, visto que a variável de condição do domicílio teria seus valores suprimidos para os mesmos registros em ambas as configurações.

Com relação às demais variáveis da Tabela 6.6, é preciso destacar que a variável cor/raça já possuía 39 registros com a categoria “ignorado” nos dados originais. Assim, os valores de 1.026 e 545 registros faltantes, para as configurações C3’ e C3’, respectivamente, incluem estes 39 registros. Para a variável de instrução, embora o número de registros suprimidos da seja muito pequeno, em relação ao total de registros, cabe lembrar novamente que esta é uma variável derivada. Neste caso, é necessário um procedimento análogo nas variáveis originais.

Uma forma alternativa de mensurar a perda de informação, que permite a comparação de todas as variáveis conjuntamente, independentemente de terem sido alvo de métodos de mascaramento distintos, é por meio de uma medida de entropia. Templ (2017) sugere uma função para este cálculo para variáveis categóricas. Considerando $c_1, c_2, ..., c_k$ categorias da variável $X_j$, a entropia $E_{cj}$ é definida por:

$$E_{c_j} = -\frac{1}{n}\sum_{c_j \in X_j} f_{c_j}\, log\left(\frac{f_{c_j}}{n}\right) \qquad (6.1)$$

tal que $f_{cj}$ é a frequência da categoria $c_j$ da variável $X_j$ e $n$ o número total de observações. Quanto menor o valor retornado desta função, maior quantidade de informação a variável conterá. Por outro lado, supondo uma máxima perda de informação da variável,

equivalente a uma recodificação global que agregasse todas as categorias da variável em apenas uma, o valor de $E_{cj}$ seria igual a zero.

Desta forma, o autor argumenta que esta medida pode também servir de insumo para a decisão da estratégia de mascaramento. Por exemplo, a variável com o menor valor de entropia poderia ser escolhida para ser alvo de recodificação global. A Tabela 6.7 apresenta esta medida para todas as variáveis que foram mascaradas.

Tabela 6.7: Medida de entropia para as variáveis alvo de algum método de mascaramento, para as configurações C3’ e C3’’.

| Variáveis | Dados originais | Dados mascarados | |
|---|---|---|---|
| | | C3’+supressões | C3’’+supressões |
| Tamanho Dom. | -1,8487 | -1,8067 | -1,8069 |
| Idade | -4,4036 | -2,6614 | -1,7065 |
| Cor/raça | -0,9714 | -0,9703 | -0,9708 |
| Instrução | -1,7988 | -1,7981 | -1,7986 |
| Capital | -1,2639 | -1,2639 | -1,2639 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Retomando o argumento de Hundepool *et al.* (2012), os valores da função de entropia apresentados na Tabela 6.7 podem não ter uma interpretação óbvia. Entretanto, se analisados conjuntamente, eles podem ser considerados como interessante medida resumo da perda de informação. De forma isolada, os valores podem ser encarados como uma proxy da “quantidade de informação” na variável. Assim, ao se comparar os resultados dos dados originais em relação a C3’ e C3’’, para uma mesma variável, é possível interpretar este aumento do valor da entropia como a “quantidade de informação perdida” daquela variável.

Desta forma, observa-se, por exemplo, que o maior aumento do valor dessa função nos dados mascarados é relativo à variável de idade. De fato, esta variável foi alvo de recodificação global e esta foi mais restritiva ainda para C3’’. A Tabela 6.7 mostra que o valor da medida de entropia cai para menos da metade neste último caso (de -4,4036 para -1,7065). Nas demais variáveis, alvos de supressão local que alterou apenas pequena parte do total de registros, esta diferença é bem menor. Adicionalmente,

verifica-se que quanto menos valores suprimidos, menor é esta diferença, chegando ao limite de na variável indicadora de capital, não haver diferença considerando-se 4 casas decimais (esta diferença aparece apenas na sexta casa).

As análises feitas nesta seção mostraram que C3' apresentou, de modo geral, uma menor perda de informação. Por este motivo, tal configuração poderia ser preferível, uma vez que ambas garantem que não existem registros acima do limite máximo de risco estabelecido. Entretanto, outras medidas de risco podem ser consideradas para uma análise conjunta com as de utilidade, para ambas as configurações. Esta questão será abordada na próxima seção.

## 6.4. Considerando conjuntamente o risco de revelação e a utilidade da informação.

Duncan *et al.* (2011) argumentam que alguns métodos de mascaramento são dependentes de certos parâmetros. A seleção dos "melhores" valores para estes parâmetros é realizada, geralmente, por duas abordagens:

a) Maximizar a utilidade dentre aqueles possíveis valores que resultam em risco de revelação abaixo de um limiar $\rho$;
b) Maximizar um escore que seja uma média ponderada das medidas de risco ($R$) e utilidade ($U$), ou seja, a escolha de parâmetros para obter $max\{\lambda R+(1-\lambda)U\}$ dado determinado $\lambda$.

Os autores argumentam que a primeira abordagem é mais utilizada, assim como é mais familiar às pessoas com conhecimento estatístico, por seguir o princípio de um teste de hipótese em que se maximiza o poder do teste fixando-se a probabilidade de erro do "Tipo 1". Neste caso, a ênfase está na confidencialidade pois os dados mascarados só são divulgados quando o risco de divulgação está abaixo do limiar fixado. Na aplicação para a PNAD Contínua realizada nesta tese, a opção por C3' (e suas respectivas supressões) seria preferível em relação a C3''[13], uma vez que esta

[13] As comparações feitas nesta seção levam em conta apenas os fatores de risco de revelação e utilidade da informação. Em outras palavras, supõe-se que as configurações de possíveis grupos etários foram escolhidas pela área temática, sendo, a priori, igualmente satisfatórias analiticamente.

configuração apresentou menor perda de informação, garantindo-se a inexistência de registros com risco de revelação acima de 20%.

No entanto, os autores apontam que o critério desta primeira abordagem pode alguma vezes ser considerado extremo. Ao não se medir um possível *trade-off* entre risco e utilidade não seriam levadas em conta, por exemplo, situações em que uma adição muito pequena no risco poderia ser mais do que compensada por um grande aumento da utilidade. Neste caso, a segunda abordagem seria mais apropriada.

O trabalho de Domingo-Ferrer *et al.* (2001) é um exemplo do uso de escore que leva em conta medidas de risco e utilidade para comparar a performance de várias configurações de métodos de mascaramento e seus respectivos parâmetros, aplicados em um banco de dados. Os autores separam a análise para dados categóricos e dados contínuos, e os escores podem levar em conta mais de uma medida de risco ou de utilidade. Assim, uma função genérica de escore (s), previamente definida nessa seção, pode ser dada por:

$$s = \sum_{l=1}^{n} \alpha_l r_l + \sum_{m=1}^{n} \beta_m u_m \qquad (6.2)$$

tal que $r_l$ é a l-ésima medida de risco considerada ponderada por $\alpha_l$; $u_m$ é a m-ésima medida de utilidade considerada ponderada por $\beta_m$.

### 6.4.1. Escore aplicado para a PNAD Contínua 2019 – 2º trimestre

Na Seção 6.2 foram propostas duas configurações de mascaramento para garantir que não restasse quaisquer registros da PNAD Contínua trabalhada com risco de revelação maior que 20%. Posteriormente, verificou-se que a primeira configuração (C3') apresentava uma menor perda de informação para o banco de dados e, por isso, poderia ser escolhida como preferível em relação à segunda (C3''). Entretanto, pode ser que esta última proveja uma maior proteção à confidencialidade das informações que, na visão do produtor dos dados, compense a menor utilidade observada. Desta forma, essa seção tem o objetivo de propor um escore que leve em conta o *trade-off* entre risco e utilidade para auxiliar na decisão entre distintas configurações de mascaramento.

É preciso, então, definir quais serão as componentes ($\alpha$, $r$, $\beta$, $u$) da fórmula geral apresentada na equação (6.2). Com relação à medida de utilidade, optou-se por utilizar

a função de entropia apresentada na equação (6.1). Como já observado na seção anterior, os valores da entropia podem ser encarados como interessante medida resumo da perda de informação, assim como podem ser calculados independentemente do método de mascaramento utilizado.

Por outro lado, a medida de risco utilizada na Seção 6.2 era relativa à contagem de registros com risco de revelação acima de 20%. Uma vez que o processo de mascaramento visou reduzir esta medida a zero, para ambas as configurações propostas, é necessário escolher outras medidas para mensurar o risco que persiste nos registros após este processo. Desta forma, foram escolhidas duas métricas já apresentadas no Capítulo 5, que podem ser consideradas complementares: a contagem de registros com risco acima de 10% (chamada daqui em diante de $r_1$) e a medida de risco global referente ao número esperado de identificações, e que consiste na soma dos riscos individuais estimados (que será chamada de $r_2$).

A vantagem de combinar $r_1$ e $r_2$ é que enquanto a primeira medida aponta os valores mais extremos da distribuição que se desejaria minimizar, a segunda dá uma ideia geral do risco médio do arquivo, uma vez que a proporção esperada de registros identificados é, matematicamente, a média dos riscos estimados. Ambas medidas foram calculadas para os dados originais e para os após mascaramento, pelas duas configurações propostas e estão dispostas na Tabela 6.8. Os cálculos foram feitos somente com base nos registros de pessoas, uma vez que a identificação de domicílios com oito ou moradores e dos que persistiam com risco maior que 20% após a etapa de mascaramento, deve ser removida do arquivo de dados.

De acordo com a Tabela 6.8, é possível observar que $r_1$ e $r_2$ diminuem consideravelmente após a aplicação dos métodos de mascaramento, sendo esta queda ainda mais acentuada para a C3''. Nos dados originais 6,53% dos registros de pessoas possuíam risco de revelação acima de 10%, enquanto o risco médio por registro era de 2,94%. Após a aplicação dos métodos de mascaramento a proporção de $r_1$ cai para 0,74% e 0,33% para C3' e C3'' respectivamente, ao passo que o risco médio por registro cai para 0,78% e 0,42% nas mesmas duas configurações. Desta forma, observa-se que C3'', embora tenha apresentado maior perda de informação, mostrou maior redução do risco

de revelação, em relação à C3'. Assim, uma escolha que leve em conta o *trade-off* entre risco e utilidade é concebível neste caso.

Tabela 6.8: Medidas de risco dos registros de pessoas para os dados originais e para os dados após mascaramento pelas duas configurações consideradas.

| Arquivo de dados | Registros de pessoas | Medidas de risco | | | |
|---|---|---|---|---|---|
| | | $r_1$ | | $r_2$ | |
| | | Total | % | Total | % |
| Dados originais | 551.348 | 35.999 | 6,53 | 16.185,4 | 2,94 |
| C3'+supressões | 551.348 | 4.105 | 0,74 | 4.309,3 | 0,78 |
| C3''+supressões | 551.348 | 1.818 | 0,33 | 2.331,9 | 0,42 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

O próximo passo para a construção do escore é definir o seu sinal, ou seja, se os maiores ou os menores valores significam um melhor *trade-off*. A escolha de uma ou outra opção não apresenta vantagem em termos teóricos, mas vai orientar a forma que as variáveis de risco e utilidade irão compor o escore. Nesta aplicação, optou-se por um escore tal que um maior valor represente melhor *trade-off*. Assim, as informações utilizadas serão a proporção de utilidade preservada das variáveis, e a proporção de redução das medidas de risco. Em todos os casos, quanto maior o valor da variável, melhor seria a configuração de mascaramento proposta. Adicionalmente, o uso exclusivo de proporções garante a mesma ordem de grandeza das informações a constituir o escore. Em caso de ordem de grandezas distintas, é necessário verificar a necessidade de algum tipo de normalização das variáveis.

Com relação à proporção de utilidade preservada das variáveis (*u*), ela pode ser extraída diretamente dos valores da medida de entropia apresentada na equação (6.1). Na Seção 6.3 foi comentado que o zero desta medida é absoluto, ou seja, uma perda total de informação acarretaria um valor igual a zero para a entropia. Desta forma, a proporção desejada pode ser estimada pelo valor da entropia dos dados pós-mascaramento, dividido pela mesma medida em relação aos dados originais ($E_{cj\text{-}mascados} / E_{cj\text{-}originais}$). Estes resultados estão dispostos na Tabela 6.9

Tabela 6.9: Utilidade preservada para as variáveis alvo de algum método de mascaramento, para as configurações C3’ e C3’’.

| Variáveis | Utilidade preservada (*u*) | |
|---|---|---|
| | C3’+supressões | C3’’+supressões |
| Tamanho Dom. | 0,9773 | 0,9774 |
| Idade | 0,6044 | 0,3875 |
| Cor/raça | 0,9989 | 0,9994 |
| Instrução | 0,9996 | 0,9999 |
| Capital | 1,0000 | 1,0000 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Os valores da Tabela 6.9, por serem funções das medidas de entropia, mostram o mesmo panorama apresentado na Tabela 6.7. A variável de idade, por sofrer recodificação global, apresentou maior perda de utilidade, especialmente para C3’’ (0,3875 = -1,7065/-4,4036 extraídos da Tabela 6.7). Nas demais variáveis, alvos de supressão local, essa perda foi bem menor e C3’’, por ser a configuração que necessitou menor número de supressões, preservou um pouco mais a utilidade das variáveis se comparado a C3’. Entretanto, a tabela apresenta o valor da entropia por variável e o que se busca é um valor resumo da medida de utilidade para o arquivo de dados pós mascaramento. Assim, é preciso arbitrar algum tipo de ponderação entre os valores de entropia apresentados na Tabela 6.9 para chegar a este valor desejado. Deve-se ter em vista que esta decisão afetará diretamente a medida resumo de utilidade e, por conseguinte, o resultado do escore.

Qual ponderação utilizar é quase totalmente inexplorada na literatura da área, de modo que não foi possível identificar uma regra para esta escolha. Geralmente a comparação da medida de entropia se dá entre métodos que mascaram apenas uma variável. Exceção encontrada foi o trabalho de Agafitei e Defays (2011), no qual os autores partiram do pressuposto que as variáveis teriam a mesma importância e, assim, utilizaram a perda de informação média. Esta suposição, por sua vez, implica que o

produtor dos dados pode arbitrar pesos diferentes para as variáveis, caso julgue necessário.

No procedimento com a PNAD Contínua desta tese, embora cinco variáveis tenham sofrido algum tipo de mascaramento, a maior parte da perda da utilidade concentrou-se na variável de idade. Para melhor captar essa questão foi proposta uma ponderação em que o peso de cada variável que fosse proporcional à média, para as duas configurações propostas, desta perda observada. A Tabela 6.10 apresenta ilustra este procedimento.

Tabela 6.10: Obtenção do peso de ponderação para a construção da medida resumo de utilidade.

| Variáveis | Perda de utilidade (1-*u*) | | Perda média de utilidade | Peso |
|---|---|---|---|---|
| | C3'+supressões | C3''+supressões | | |
| Tamanho Dom. | 0,0227 | 0,0226 | 0,0227 | 0,0429 |
| Idade | 0,3956 | 0,6125 | 0,5041 | 0,9549 |
| Cor/raça | 0,0011 | 0,0006 | 0,0009 | 0,0017 |
| Instrução | 0,0004 | 0,0001 | 0,0003 | 0,0005 |
| Capital | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| Total | -- | -- | 0,5278 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

As duas primeiras colunas da Tabela 6.10 apresentam a perda de utilidade estimada, a partir das medidas de entropia, que é igual ao complementar dos valores dispostos na Tabela 6.9. A terceira coluna contém a média aritmética das duas colunas anteriores. O peso da ponderação vai ser, então, proporcional a esta medida, ou seja, o valor desta perda média dividido pelo valor total da terceira coluna. Estes pesos, aplicados aos valores da Tabela 6.9 resultarão no valor resumo da perda de utilidade a ser utilizada no escore, que será denominada de $u_1$', que está disposta na Tabela 6.11.

Cabe ressaltar, como já comentado na Seção 6.3, que a supressão local em uma variável, pode acarretar supressão de outras variáveis no mesmo registro que impossibilitem o intruso recompor essa informação. Entretanto, como a perda de

informação por este método foi pouca, acarretando um peso relativamente pequeno pela ponderação definida, optou-se por fazer o cálculo utilizando apenas as variáveis diretamente trabalhadas. A inclusão de outras variáveis teria impacto muito reduzido no resultado do escore e, como esta aplicação tem objetivo ilustrativo, optou-se por esta simplificação neste caso particular.

Na Tabela 6.11 também se encontram as medidas de redução de risco – denominadas de $r_1'$ e $r_2'$ – extraídas diretamente dos valores de $r_1$ e $r_2$, contidos na Tabela 6.8. Como mencionado anteriormente, deseja-se utilizar proporção de redução de $r_1$ e $r_2$, entre os dados originais e mascarados. Assim $r_{lu}' = (r_{lo} - r_{lu})/r_{lo}$, tal que $r_{lo}$ é a l-ésima medida de risco calculada para os dados originais (o), enquanto $r_{lu}$ é a l-ésima medida de risco calculada para os dados mascarados pela u-ésima configuração proposta. Estes valores também são apresentados na Tabela 6.11.

Tabela 6.11: Variáveis e seus respectivos valores utilizados para a construção do escore, para as configurações C3' e C3''.

| Variáveis | C3'+supressões | C3''+supressões |
|---|---|---|
| $u_1'$ | 0,6212 | 0,4142 |
| $r_1'$ | 0,8867 | 0,9495 |
| $r_2'$ | 0,7347 | 0,8571 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

O valor de $u_1'$, contido da Tabela 6.11, capta o que se pretendia com a ponderação escolhida: levar em conta prioritariamente a informação da perda de utilidade decorrente da etapa recodificação global, com uma correção resultante da etapa de supressão local. Em outras palavras, o valor de $u_1'$ é próximo ao de $u_{V2009}$, mas contendo um pequeno acréscimo que é maior para C3'', captando, assim, o ganho marginal de utilidade em virtude do menor número de supressões locais.

Definidas as variáveis de utilidade e risco, assim como seus valores, para entrar no escore (denominado de *s*), o próximo passo é definir como se dará a ponderação destas componentes. Com base na equação geral (6.2) o escore proposto será da forma:

$$s = \alpha_1 r_1' + \alpha_2 r_2' + \beta_1 u_1' \qquad (6.3)$$

tal que: $\lambda = \alpha_1 + \underline{\alpha_2}$ e $(1-\lambda) = \beta_1$.

O parâmetro $\lambda$ representa o peso que se dará às medidas de risco. Este é mais um caso que a literatura não indica um valor padrão, ficando a cargo do produtor dos dados a escolha, devido a especificidade do problema em questão. Domingo-Ferrer *et al.* (2001), por exemplo, optaram por arbitrar o valor de 0,5 para dar pesos idênticos entre as duas componentes básicas do escore. Entretanto, como a ideia prioritária do CEC é proteger a confidencialidade das informações, é razoável supor que o valor de $\lambda$ não pode ser menor que 0,5. Desta forma, em vez de escolher um valor específico para este parâmetro, o escore será calculado para cinco valores de $\lambda$ diferentes (0,5 até 0,9 em intervalos de 0,1) para melhor visualização de seu comportamento com relação à mudança de ponderação.

A partir da escolha do valor de $\lambda$, é preciso definir como ele será distribuído entre $\alpha_1$ e $\alpha_2$. Em outras palavras, dentre as medidas de risco utilizadas, se o produtor dos dados dará maior importância a alguma delas e, em caso afirmativo, qual seria esta magnitude. Nesta tese, por se tratar de uma aplicação de caráter mais ilustrativo, serão propostas duas abordagens para comparação dos resultados. Na primeira não será privilegiada nenhuma medida, ou seja, $\alpha_1$ e $\alpha_2$ serão iguais a $\lambda/2$. Já a segunda, parte do pressuposto que se deve evitar ao máximo registros de mais alto risco, mesmo que todos já estejam garantidos abaixo do limite máximo de 20%. Assim, a ideia seria priorizar $r_1'$ dando a ele um peso de $2\lambda/3$, enquanto a $r_2'$ será atribuído peso de $\lambda/3$. Os resultados dos escores estão dispostos na Tabela 6.12.

Com base nos resultados da Tabela 6.12, é possível observar que ao se atribuir a mesma importância para as duas componentes básicas do escore ($\lambda = 0,5$), C3' apresenta um melhor *trade-off* entre risco e utilidade. Em outras palavras, o valor do escore é maior para C3' para $\lambda = 0,5$, sejam $\alpha_i$ iguais ou diferenciado. No entanto, quanto maior peso é dado para a componente de risco $\lambda$, o escore para C3'' tende a superar o relativo a C3'. Este comportamento é esperado visto que C3'' apresentou as menores medidas de risco. Supondo $\alpha_i$ iguais, C3'' já era preferível, dadas as cinco opções de ponderação definidas, para $\lambda$ a partir de 0,7. No entanto, para pesos diferenciados de $\alpha_i$, C3'' seria escolhido somente a partir de $\lambda$ igual a 0,8.

Tabela 6.12: Resultado do escore proposto, segundo os parâmetros adotados, para as configurações C3’ e C3’’.

| Parâmetros | | | C3’+supressões | C3’’+supressões |
|---|---|---|---|---|
| $\lambda$ | $\alpha_1$ | $\alpha_2$ | | |
| $\alpha_l$ iguais | | | | |
| 0,50 | 0,25 | 0,25 | 0,716 | 0,659 |
| 0,60 | 0,30 | 0,30 | 0,735 | 0,708 |
| 0,70 | 0,35 | 0,35 | 0,754 | 0,757 |
| 0,80 | 0,40 | 0,40 | 0,773 | 0,805 |
| 0,90 | 0,45 | 0,45 | 0,792 | 0,854 |
| $\alpha_l$ diferenciados | | | | |
| 0,50 | 0,33 | 0,17 | 0,729 | 0,666 |
| 0,60 | 0,40 | 0,20 | 0,750 | 0,717 |
| 0,70 | 0,47 | 0,23 | 0,772 | 0,767 |
| 0,80 | 0,53 | 0,27 | 0,793 | 0,818 |
| 0,90 | 0,60 | 0,30 | 0,815 | 0,868 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Em resumo, caso a abordagem escolhida fosse maximizar a utilidade dentre as configurações que apresentassem risco revelação abaixo do limiar de 20% proposto, C3’ seria preferível. Caso a opção fosse pela abordagem de maximizar um escore que ponderasse medidas de risco e utilidade, C3’’ poderia passar a ser uma melhor opção caso o peso da componente de risco fosse elevado.

## 6.5. Relação entre as etapas de mascaramento e de crítica e imputação

Após decididos todos os métodos que serão utilizados no processo de CEC, o próximo passo é referente à sua implementação. Isto requer uma análise tantos dos recursos existentes para o produtor dos dados (capacidade operacional, *softwares* disponíveis, etc), quanto dos procedimentos de produção empregados na pesquisa (HUNDEPOOL *et al.*, 2012). Desta forma, nesta seção é proposta para uma abordagem

para a implementação do CEC, em pesquisas amostrais domiciliares, explorando as similaridades e sinergias com os processos de crítica e imputação já existentes na pesquisa.

Muitos métodos de mascaramento podem ser encarados como ações de imputação. Por exemplo, a supressão local e a recodificação são, em essência, imputações determinísticas tais que o valor original é substituído por outro – seja por *missing* no primeiro caso, ou pela nova classe no segundo – já previamente fixado. A adição de ruídos e o PRAM, por sua vez, são tipos de imputação probabilística, que seguirão a distribuição definida para o ruído ou para a matriz de transição, respectivamente. Templ (2017), inclusive, pondera que um método alternativo ao PRAM seria criar alguns valores faltantes de forma aleatória, para posteriormente imputá-los probabilisticamente na etapa concernente.

Como mencionado na Seção 2.10.3, o produtor das informações pode fazer a opção por disponibilizar dados sintéticos ao invés da utilização de métodos de mascaramento. Ainda assim, esta alternativa pode ser realizada por um processo de imputação múltipla, como aponta Hundepool *et al.* (2012). A ideia, neste caso, é tratar todas as unidades da população que não foram selecionadas na amostra como dados faltantes, imputá-las e extrair amostras desses registros imputados para divulgação pública. Drechsler (2011), em seu livro que trata especificamente sobre o tema de dados sintéticos, aprofunda bastante esta abordagem. Deve ser ressaltado que é possível combinar métodos de mascaramento e dados sintéticos. Duncan *et al.* (2011), por exemplo, consideram a ideia já descrita por Templ (2017) de suprimir valores para depois imputá-los como "dados sintéticos locais", e que poderia ser adotada em adição a métodos de mascaramento prévios.

Entretanto, apesar das similaridades, as etapas de mascaramento e crítica e imputação são distintas devendo ser definida a ordem destes processos no fluxograma de produção da pesquisa. Os novos valores mascarados, por exemplo, devem obedecer às regras de crítica. Assim, uma etapa de mascaramento posterior à de crítica deve garantir que não sejam criadas quaisquer inconsistências nos dados. Por outro lado, supondo uma etapa de imputação após o CEC, pode ser necessário adotar regras

adicionais como definir se dados já mascarados podem ser alvo de novas imputações, entre outras.

Independente da ordem entre estas duas etapas, é preciso levar em conta também se alterações no plano de crítica da pesquisa serão necessárias em virtude da estratégia adotada para o mascaramento dos dados. Tomando como ilustração a aplicação desta tese nos dados da PNAD Contínua deve ser observado, por exemplo, se há alguma regra de crítica que impeça a presença de valores faltantes para as variáveis alvo de supressão local. Adicionalmente, a recodificação global em grupos etários também vai demandar a revisão, e possível atualização, de regras de crítica que envolvam a variável idade.

Outra questão que deve ser levada em conta é em relação à presença de variáveis de marca de imputação. Esta variável indicadora informa ao usuário se o valor de determinada variável contida no registro é uma resposta do informante ou é fruto de imputação. Desta forma, o usuário tem a opção de, no limite, apagar estes valores e utilizar um outro método de imputação que prefira. Entretanto, sob o ponto de vista do intruso, uma variável imputada seria o equivalente a um valor faltante, uma vez que o desconhecimento sobre o valor verdadeiro torna essa informação sem utilidade para uma possível revelação. Logo, a presença da marca de imputação mostra ao intruso exatamente quais variáveis não são informações originais do usuário. Caso esta variável não fosse disponibilizada traria um grau de incerteza adicional, pois não se poderia garantir quais valores são imputados e, consequentemente, se uma identificação é correta. Hundepool *et al.* (2012) recomendam, de uma forma geral, a sua remoção. De qualquer maneira, cabe ao produtor dos dados avaliar a pertinência da presença da marca de imputação em cada caso específico.

### 6.5.1. Aplicação para os dados da PNAD Contínua 2019 – 2º trimestre

Esta seção pretende ilustrar a implementação do processo de CEC frente aos recursos e processos já existentes no IBGE, dando ênfase à etapa de crítica e imputação. A ideia é aplicar os conceitos da Seção 6.5 na prática, embora saiba-se que uma

implementação de fato do CEC deva contar com a colaboração indispensável da equipe metodológica da pesquisa.

Retomando a ideia de Hundepool *et al.* (2012) explicitada na seção anterior, com relação à análise dos recursos utilizados, deve ser destacado que todos os procedimentos de CEC foram realizados por meio do pacote "sdcMicro" do *software* R, no computador pessoal especificado na Seção 5.5. Este *software*, além de disponível gratuitamente, já é utilizado no IBGE. Ademais, foram observados reduzidos tempos de execução das etapas de CEC realizadas no computador pessoal, aquém em termos de processamento a *hardwares* mais potentes da Instituição. Assim, a implementação de todo o procedimento poderia ser realizada neste ambiente de modo relativamente simples.

A segunda análise proposta por Hundepool *et al.* (2012) é em relação aos processos já existentes. A PNAD Contínua utiliza o sistema *Canadian Census Edit and Imputation System* (CANCEIS) na etapa de crítica e imputação da pesquisa. Adicionalmente, os valores considerados discrepantes nos valores de rendimentos são substituídos por meio de imputação determinística. O valor imputado será o mais alto que seja considerado válido, isto é, não discrepante, dentro dos limites estabelecidos pelo método de identificação de *outliers* adotado. Deve ser ressaltado que esta imputação não tem motivação ligada à confidencialidade dos dados. Ela foi adotada após a verificação, em 2016, de um valor muito destoante de rendimento que gerou impactos artificiais no aumento do rendimento e da desigualdade daquele ano e, posteriormente, na queda da desigualdade, em 2017 (IBGE, 2021b). O arquivo de uso público disponível na página do IBGE e utilizado nesta tese, não apresenta variáveis de marca de imputação.

O processo de CEC, especialmente a estratégia de mascaramento dos dados, deve levar em conta como é feita a etapa de crítica e imputação da pesquisa. As variáveis mascaradas e o tipo de tratamento utilizado devem ser confrontados com o atual plano de crítica da pesquisa, para possíveis alterações no mesmo. Como já adiantado na seção anterior, deve ser observado se existem regras que impeçam a presença de valores faltantes para as variáveis alvo de supressão local, além da revisão nas regras que envolvam a variável idade, devido à recodificação global introduzida.

Outra questão é concernente aos tipos de métodos utilizados e a ordem de processamento deles. Na aplicação desta tese optou-se por métodos não perturbativos que, dado a sua natureza explicitada na seção anterior, podem ser introduzidos, por exemplo, em paralelo à etapa de imputação determinística da pesquisa. Caso haja o interesse ou a necessidade de uma etapa de proteção adicional, é possível aproveitar também a etapa de imputação probabilística do CANCEIS já implantada na pesquisa. Uma alternativa é a estratégia já comentada de Templ (2017) que consiste em criar valores faltantes de forma aleatória para depois imputá-los. Esta abordagem seria beneficiada, em termos de adicionar um grau maior de incerteza aos possíveis intrusos, pela ausência de marcas de imputação no arquivo de dados de uso público.

Deve ser ressaltado que, mesmo no exemplo anterior em que se admite que os processos de mascaramento e imputação determinística possam ocorrer dentro de uma mesma etapa, há que se definir uma ordem de execução. Em outras palavras, é preciso estabelecer se o mascaramento ocorre antes e, neste caso, se os valores mascarados podem ser ou não alvo de imputações posteriores; ou a imputação deve ocorrer antes e, em seguida, decidir sobre a possibilidade ou não de mascaramento dos valores imputados.

Decisões sobre a sequência de execução de determinados passos já são tomadas dentro do próprio processo de imputação. Há que se definir não somente a ordem entre as etapas de imputação determinística e probabilística, como esta última pode conter subetapas que devem ser encadeadas de alguma forma. Por exemplo, é possível supor que variáveis demográficas e de trabalho já tenham sido alvo de imputação probabilística anterior e estejam fixadas para servir de insumo, na execução da imputação probabilística das variáveis de rendimento. Desta forma, a etapa de mascaramento poderia ser encarada como mais uma subetapa e ser incluída do modo relativamente simples na sequência já existente de processos da pesquisa.

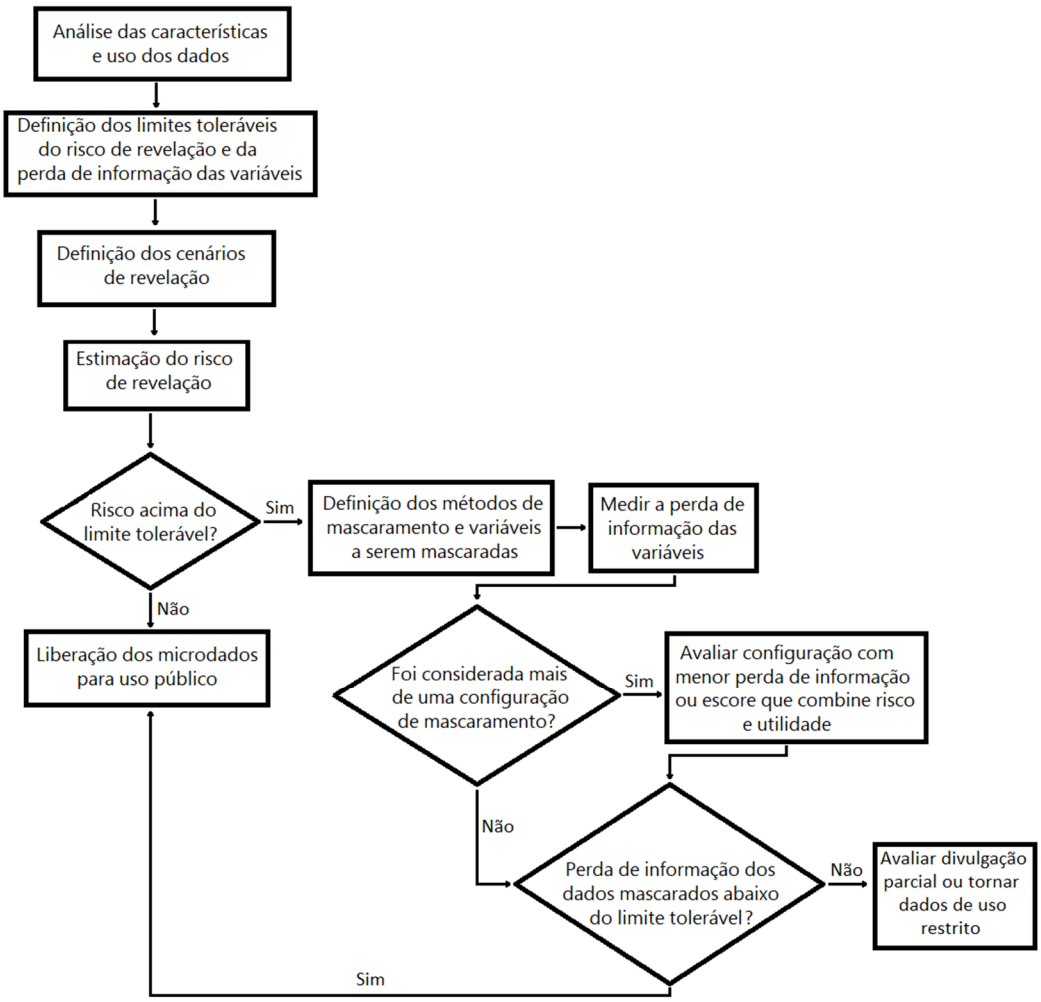
Para a aplicação desta tese, uma sugestão seria colocar a etapa de mascaramento por recodificação global e supressão local antes da etapa de imputação. Uma vez calculados os riscos de revelação pós-mascaramento, possíveis imputações posteriores só iriam adicionar um grau de incerteza maior ao intruso, ou seja, estaria garantida uma cota máxima para o risco. Desta forma, não haveria restrições, pelo

menos *a priori*, para que os dados mascarados pudessem ser alvo de imputações subsequentes.

## 6.6. Considerações gerais sobre todas as etapas de CEC apresentadas

Todas as etapas para a implementação do CEC tratadas nesta tese, desde a análise inicial das características dos dados da pesquisa até a decisão para a liberação final destes dados para uso público, estão presentes no fluxograma contido na Ilustração 6.1. No entanto, é importante frisar que cada microdado é único, possuindo características e usos por parte dos usuários muito peculiares. Cada etapa apresentada no fluxograma, pode ser desmembrada em subetapas que, por sua vez, podem diferir conforme a pesquisa analisada. Por exemplo, ao se considerar a "definição dos métodos de mascaramento e variáveis a serem mascaradas", assunto tratado nas Seções 6.1 e 6.2, foi possível observar o grande número de perguntas a serem consideradas e, consequentemente, decisões que devem ser tomadas com base no microdado alvo. Desta forma, o fluxograma contido nesta seção pretende passar uma ideia da ordem e da interação entre as fases de forma resumida.

Ilustração 6.1: Fluxograma com as etapas da abordagem proposta de CEC, nesta tese, para pesquisas amostrais domiciliares.




Fonte: Elaborado pelo autor

Por fim, é importante frisar que as etapas contidas neste fluxograma, ou em qualquer abordagem de CEC já estabelecida, não são imutáveis. Conforme novas pesquisas, com suas especificidades, são analisadas, ou conforme novos riscos e desafios à confidencialidade das informações surjam devido aos constantes avanços tecnológicos observados na sociedade, há a necessidade de uma contínua revisão do que está sendo feito, objetivando sempre o aprimoramento dos processos existentes.

# CAPÍTULO 7: CONSIDERAÇÕES FINAIS

Assegurar que os respondentes não possam ser identificados a partir dos dados publicados é mais do que uma questão estritamente legal. A confiança da sociedade em relação ao INE é essencial para a obtenção de dados fidedignos, os quais, por sua vez, dependem da boa vontade e cooperação da população.

O campo de estudo do CEC foi desenvolvido justamente em resposta a desafios em relação à confidencialidade das informações que os profissionais dos INE enfrentavam na prática, resultado principalmente de inovações tecnológicas e novas formas de disseminação de dados. Estes dois últimos elementos continuam fazendo parte da realidade atual e não há motivos para duvidar que será diferente nas próximas décadas. Desta forma, a confidencialidade das informações estará constantemente exposta a novas ameaças, ressaltando cada vez mais a relevância do CEC tanto no âmbito dos INE, quanto para a sociedade como um todo.

Uma das etapas mais importantes do CEC é a estimação do risco de revelação. Entretanto, isto envolve etapas anteriores como, por exemplo, a determinação de um cenário de revelação e das variáveis-chave em que serão baseados tais cálculos, o mapeamento de fontes externas de dados que possam auxiliar possíveis intrusos, todas tratadas nesta tese. Com relação à busca de bases de dados públicas disponíveis que possam ser utilizadas por um intruso, foi proposto nesta tese que se leve em consideração aspectos de escopo, atualização e abrangência geográficas dos dados. A análise das bases selecionadas, em conjunto com PNAD Contínua, levou à escolha das variáveis-chave. No que concerne aos cenários de revelação, a literatura aponta que é prudente estabelecer mais de um quando se trata de dados de uso público, e aponta dois caminhos usualmente utilizados. No primeiro, pressupõe-se que o intruso vai usar de alguma informação externa para fazer a revelação e, no segundo, que ele utilizará o conhecimento próprio sobre características das unidades respondentes. Foram então estabelecidos dois cenários com base nesses pressupostos supracitados.

Nesta tese os riscos de revelação foram estimados por modelagem estatística, mais especificamente pela "abordagem italiana". Este método é popular na literatura e se encontra incorporado no pacote "sdcMicro" do *software* R sendo, então, de implementação relativamente simples no processo de produção das pesquisas do IBGE. Entretanto, para pesquisas amostrais com fração amostral maiores (acima de 1%), há a possibilidade de se utilizar métodos heurísticos para a estimação do risco, como SUDA que se encontra disponível também no pacote "sdcMicro". Assim, uma opção para trabalhos futuros é a análise dos casos em que é mais vantajoso utilizar estas ou ainda outras diferentes abordagens.

De acordo com os resultados apresentados, observou-se que existem registros na PNAD Contínua analisada com alto risco de revelação. Quanto mais desagregado é o recorte geográfico divulgado, este problema tende a aumentar, porém, mesmo para o nível mais agregado, persistem registros de domicílios com risco de revelação elevados. Em outras palavras, existem registros com padrões de respostas muito peculiares. O efeito da estrutura hierárquica tem papel relevante neste contexto, uma vez que domicílios com muitos moradores possuem características únicas, que os deixariam vulneráveis à revelação. Os resultados mostraram que a maior concentração dos registros de domicílios com risco de revelação acima de 20%, era referente ao estrato com 8 ou mais moradores. Desta forma, é importante a avaliação da forma de divulgação dos domicílios a partir de um determinado número de moradores.

É possível ainda observar uma grande heterogeneidade para os valores do risco de revelação dentro dos recortes geográficos. Há um percentual muito maior de registros de domicílios acima dos limiares de risco de 10% e 20% na UF Roraima, do que em UF mais populosas como São Paulo e Minas Gerais, por exemplo. Adicionalmente, é preciso destacar que ao se divulgar determinados recortes geográficos (como municípios em Regiões Metropolitanas ou municípios de capitais), existe a possibilidade de o intruso recompor outros recortes de divulgação não previstos (domicílios naquela UF fora da Região Metropolitana ou fora de capitais) que podem deixar registros

vulneráveis à revelação. Assim, é necessário considerar que informações, a princípio sem intenção de divulgação, podem ser obtidas, com base nos dados divulgados.

Foram também apresentadas medidas de risco global que podem ser consideradas medidas resumo do risco de revelação. Elas possuem a vantagem de não impactar o fluxo de produção da pesquisa, por serem obtidas a partir das estimativas dos riscos individuais. O produtor dos dados pode utilizar medidas de risco global para auxiliar na decisão de divulgação de um conjunto de dados. Desta forma, uma possível linha de pesquisa futura é a pertinência da incorporação destas medidas, no sentido de o Instituto estabelecer limites máximos aceitáveis para elas.

Após a estimação dos riscos de revelação, caso seja necessária a adoção de métodos de mascaramento, é preciso definir quais variáveis serão alvo desta etapa e que métodos serão utilizados. Contudo, há a possibilidade de o produtor dos dados estabelecer políticas internas para orientar, por exemplo, que determinadas variáveis permaneçam com sua utilidade máxima preservada. Neste caso, estudos futuros com as áreas de temáticas de cada pesquisa, levando em conta suas especificidades, são de fundamental valia.

No que diz respeito aos métodos de mascaramento, há que se ter em conta que para muitos métodos perturbativos, o usuário precisa incluir alterações em seus procedimentos de análise usuais que exigem um bom conhecimento do método aplicado. Assim, num primeiro momento, pode ser recomendável priorizar adoção de métodos não perturbativos exclusivamente. De qualquer modo, a diversidade de métodos existentes e a possibilidade de combinação deles, abre um leque amplo de possibilidade para trabalhos futuros

Ressalta-se ainda, que independente das variáveis ou métodos escolhidos, é necessária atenção adicional para pesquisas repetidas ao longo do tempo, como é o caso da PNAD Contínua. É de suma importância que as variáveis possam ser comparáveis ao longo das pesquisas, desta forma há que se ter muito cuidado ao se propor alterações em uma configuração de mascaramento que já está em uso.

Nesta tese julgou-se interessante trabalhar com duas configurações de mascaramento de variáveis, para que a comparação entre as duas ilustrasse com detalhes os quesitos de redução do risco e da perda de informação, após o mascaramento dos dados. De forma análoga, estas configurações foram analisadas pelas duas abordagens existentes na literatura: a que maximiza a utilidade das informações, e a que maximiza uma função que leva em conta conjuntamente as medidas de risco e de utilidade das informações. Para este último caso, foi desenvolvido um escore que tem esta finalidade.

Após a decisão sobre todos os procedimentos e métodos a serem utilizados no processo de CEC, o próximo passo é referente à sua implementação. Foi proposta uma abordagem com foco nas similaridades com os processos de crítica e imputação já existentes na pesquisa. Muitos métodos de mascaramento podem ser encarados como ações de imputação. Desta forma, ao se adotar estes métodos, é muito importante avaliá-los conjuntamente com o processo de crítica e imputação já existente na pesquisa, que poderá sofrer algumas alterações. Não obstante, possíveis sinergias entre estas duas etapas têm potencial de tornar mais simples a incorporação do CEC na pesquisa.

Observou-se, nesta tese, que durante todo o processo de CEC determinadas escolhas, que não possuem uma regra geral sendo estabelecidas pelo produtor dos dados, devem ser tomadas. Definição dos cenários de revelação e das variáveis-chave, limite máximo de risco de revelação tolerado, escolha das variáveis a serem mascaradas e do método de mascaramento, assim como decisão sobre a utilidade da informação a ser preservada nas variáveis são alguns exemplos. Foram, então, discutidas formas de se abordar estas questões e se arbitraram decisões para elas com relação aos dados da PNAD Contínua utilizada. Cada uma destas questões pode, então, ser alvo de estudos futuros. Não apenas para aplicação no processo de pesquisa em termos práticos, mas também para elaboração de normas internas sobre o tema para o IBGE.

Em suma, a partir do que foi desenvolvido neste trabalho, os resultados obtidos indicam que a adoção de técnicas de CEC disponibiliza ferramentas fundamentais para garantir a confidencialidade das informações dos respondentes.

# REFERÊNCIAS


ABOWD, J.M.; NISSIM, K.; SKINNER, C.J. First issue editorial. **Journal of Privacy and Confidentiality**, v. 1, n. 1, 2009.

ABS. 1160.0 - **ABS Confidentiality Series**, Aug 2017. Australian Bureau of Statistics. 2017.

AGAFITEI, M.; DEFAYS, D. Analysis of information loss in European data due to confidentiality. **Joint UNECE/Eurostat work session on statistical data confidentiality.** Tarragona, Spain, 26-28 October 2011.

ARANTES, S.B.; DE MAGALHÃES, M.S; Primary analysis of disclosure risk in tabular data from a Brazilian economic survey. Conference Of European Statisticians**. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.** Netherlands. 2019.

ARANTES, S.B.; DE MAGALHÃES, M.S; BRITO, J.A.M. Disclosure risk assessment for frequency tables of a Brazilian economic survey. **63rd World Statistics Congress 2021.** International Statistical Institute (ISI). 2021.

BENEDETTI, R.; FRANCONI, L. Statistical and technological solutions for controlled data dissemination. In **Pre-proceedings of New Techniques and Technologies for Statistics** p. 225–232. 1998.

BENEDETTI, R.; A. CAPOBIANCHI, A.; FRANCONI, L. Individual risk of disclosure using sampling design information. **Contributi Istat**, v. 1412003. 1998.

BENSCHOP, T.; MACHINGAUTA, C.; WELCH, M. **Statistical disclosure control: A practice guide**. 2019.

BERGEAT, M. La gestion de la confidentialité pour les données individuelles. **Documents de travail de l'Insee M**, v. 2016, 2016.

BETHLEHEM, J.; KELLER, W.; PANNEKOEK, J. Disclosure control of microdata. **Journal of the American Statistical Association**, v. 85, n. 409, p. 38–45, 1990.

BIANCHINI, Z.M. **O tratamento da questão do sigilo das informações**. Diretoria de Pesquisas - DPE, Divisão de Metodologia – DME, IBGE, 1994.

BIANCHINI, Z.M. et al. **Grupo de Sigilo: Atividades Realizadas, Desafios e Perspectivas para o Futuro**. IBGE, 1999.

CAPOBIANCHI, A.; POLETTINI, S.; LUCARELLI M. **Strategy for the implementation of individual risk methodology into µ-ARGUS.** Technical report, Report for the CASC project. No: 1.2-D1. 2001.

CARLSON, M. Assessing microdata disclosure risk using the poisson-inverse gaussian distribution. **Statistics in Transition**, v. 5, n. 6, p. 901–925. 2002.

CASTANHA, R.C.G. A ciência de dados e a cientista de dados. **Atoz: novas práticas em informação e conhecimento**. v. 10, n. 2, p. 1-4, 2021.

CHARTRAND, R.L. The Federal Data Center: Proposals and reactions. In: **Washington, D. C.: Library of Congress Legislative Reference Service**, p. 354-362. December 1967.

CONNORS, E; KRUPNIKOV, Y; RYAN, J B. How Transparency Affects Survey Responses. **Public Opinion Quarterly**, v. 83, n. S1, p.185–209, 2019.

COUPER, M.P., SINGER, E., CONRAD F.G., GROVES R.M. Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. **Journal of Official Statistics**, v. 24, n. 2, p. 255–275. 2008

COUPER, M.P., SINGER, E., CONRAD F.G., GROVES R.M. Experimental Studies of Disclosure Risk, Disclosure Harm, Topic Sensitivity, and Survey Participation**. Journal of Official Statistics**, v. 26, n. 2, p. 287–300. 2010

DANDEKAR, R.A.; DOMINGO-FERRER, J.; SEBÉ, F. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In: **Inference Control in Statistical Databases**. Springer, Berlin, Heidelberg, p. 153-162. 2002.

DANE; CEPAL. **Código Regional de Buenas Prácticas en Estadísticas para América Latina y el Caribe**. Departamento Administrativo Nacional de Estadística de Colombia y Comisión Económica para América Latina y el Caribe. 2011.

DANE. **Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional**. Dirección de Regulación, Planeación, Estandarización Y Normalización (DIRPEN) – DANE. Agosto de 2018.

DALENIUS, T. Towards a Methodology for Statistical Disclosure Control. **Statistisk Tidskrift**, v. 5, p. 429-444. 1977.

DALENIUS, T.; REIS, S.P. Data-swapping: a technique for disclosure control (extended abstract). **Proceedings of the ASA Section on Survey Research Methods**, pp. 191–194. American Statistical Association, Washington DC. 1978.

DEFAYS D.; NANOPOULOS P. Panels of enterprises and confidentiality: the small aggregates method. **Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys**, pp. 195–204. Statistics Canada, Ottawa. 1993.

DESTATIS. **Methods – Approaches – Developments**. Information of the German Federal Statistical Office. Edition 2/2018. DESTATIS. 2018.

DE WAAL, A.G.; WILLENBORG, L.C.R.J. Global recodings and local suppressions in microdata sets. **Proceedings of Statistics Canada Symposium'95**, p. 121–132. Statistics Canada, Ottawa. 1995.

DE WAAL, A.G.; WILLENBORG, L.C.R.J. Information loss through global recoding and local suppression. **Netherlands Official Statistics**, v. 14, p. 17–20. 1999.

DE WAAL, T.; WILLENBORG, L.C.R.J. A view on statistical disclosure control for microdata. **Survey Methodology**, v. 22, n. 1, p. 95-103. 1996.

DIEESE. Sistema PED - **Pesquisa de Emprego e Desemprego**. Departamento Intersindical de Estatística e Estudos Socioeconômicos. Disponível em: <www.dieese.org.br>. Acesso em: outubro de 2021.

DOMINGO-FERRER, J.; MATEO-SANZ, J.M.; TORRA. V. Comparing SDC methods for microdata on the basis of information loss and disclosure risk**. Pre-proceedings of ETK-NTTS'2001**, vol. 2, p. 807–826. Eurostat, Luxemburg. 2001.

DRECHSLER, J. **Synthetic datasets for statistical disclosure control: theory and implementation**. Springer Science & Business Media, 2011.

DUNCAN G. T.; KELLER-MCNULTY, S.; STOKES, S. **Disclosure risk vs. data utility: the r-u confidentiality map.** Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico, 2001.

DUNCAN, G.; LAMBERT, D. Disclosure-limited data dissemination. **Journal of the American Statistical Association**, v. 81, n. 393, p. 10–18. 1986.

DUNCAN, G.T.; ELLIOT, M.; SALAZAR-GONZÁLEZ, J.J. **Statistical Confidentiality: Principles and Practice**. New York: Springer, 2011.

DUNN, E.S. The idea of a national data center and the issue of personal privacy. **The American Statistician**, v. 21, n. 1, p. 21-27. 1967.

ELAMIR, E. A. H.; SKINNER, C. Record Level Measures of Disclosure Risk for Survey Microdata. **Journal of Official Statistics**. v. 22, n. 3, p. 525–539, 2006.

ELLIOT M.J. DIS: a new approach to the measurement of statistical disclosure risk. **International Journal of Risk Management**, v. 2, n. 4, p. 39–48, 2000.

ELLIOT, M.J.; DOMINGO FERRER, J. The future of statistical disclosure control. **The National Statistician's Quality Review**. London, December 2018.

ELLIOT, M.; MANNING, A. M. Using dis to modify the classification of special uniques. In **Joint UNECE/Eurostat work session on statistical data confidentiality**. Luxembourg. 2003.

ELLIOT, M.J.; MANNING, A.M.; FORD, R.W.: A computational algorithm for handling the special uniques problem. **International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems**. V. 10, n. 5, p. 493–509. 2002.

EUROSTAT. **European Statistics Code Of Practice**. For the National Statistical Authorities and Eurostat (EU statistical authority). Eurostat. 2017.

ENAP. **Elaboração de Plano de Dados Abertos**. Brasília: Escola Nacional de Administração Pública. 2017.

FADEL, A.C. Aplicação de métodos de controle estatístico de sigilo nos microdados da PeNSE 2019. Draft**. Encontro COMEQ 2021**. Rio de Janeiro. 2021

FADEL, A.C.; OCHI, L.C.; BRITO, J.A.M.; SEMAAND, G.S. Microaggregation heuristic applied to statistical disclosure control. **Information Sciences**. v. 548, p. 37-55. 2021.

FELLEGI, I.P. On the Question of Statistical Confidentiality. **Journal of the American Statistical Association**, v. 67, n. 337, p. 7-18. 1972.

FRANCONI, L.; POLETTINI, S. Individual risk estimation in μ-Argus: A review. In: **International workshop on privacy in statistical databases**. Springer, Berlin, Heidelberg, p. 262-272. 2004

GOMES, M.M.F.; TURRA, C.M. **Quantos são os centenários no brasil? Uma estimativa indireta da população com 100 anos e mais com base no número de óbitos**. Cedeplar, Universidade Federal de Minas Gerais, 2008.

GOUWELEEUW, J.M.; KOOIMAN P.; WILLENBORG L.C.R.J.; DE WOLF P.P. Post randomisation for statistical disclosure control: theory and implementation. **Journal of Official Statistics**, v. 14, n. 4, p. 463–478. 1998.

GRIFFITHS, E.; et al. **Handbook on Statistical Disclosure Control for Outputs**. Safe Data Access Professionals Working Group. 2019.

GSS. GSS/GSR **Disclosure Control Guidance for Microdata Produced from Social Surveys**. Government Statistical Service. Reino Unido. 2014.

GT ART.29. **Dictamen 05/2014 sobre técnicas de anonimización**. Grupo De Trabajo Sobre Protección De Datos Del Artículo 29. 2014.

HOLVAST, J. History of Privacy. International Federation for Information Processing. V. Matyáš et al. (Eds.): **The Future of Identity**, IFIP AICT 298, p. 13–42, 2009.

HOUAISS, A. **Dicionário sinônimos e antônimos**. 2. ed. São Paulo: Publifolha, 2008.

HOSHINO, N.; TAKEMURA, A. On the relation between logarithmic series model and Other superpopulation models useful for microdata disclosure risk assessment. **CIRJE F-Series 98-F-7**, CIRJE, Faculty of Economics, University of Tokyo. 1998.

HUNDEPOOL, A.; DE WOLF, P. **Statistical Disclosure Control**. Method Series. Statistics Netherlands, 2012.

HUNDEPOOL, A. *et al*. **Statistical Disclosure Control**. Wiley Series in Survey Methodology: Wiley. ISBN 9781118348222, 2012.

IBGE. Sistema Integrado de Pesquisas Domiciliares – SIPD. **Textos para discussão Diretoria de Pesquisas**, n. 24. IBGE. Rio de Janeiro. 2007

IBGE. **Código de Boas Práticas das Estatísticas do IBGE**. IBGE. Rio de Janeiro. 2013.

IBGE. **Pesquisa Nacional por Amostra de Domicílios Contínua. Notas Metodológicas**. IBGE. Rio de Janeiro. 2014.

IBGE. **Confidencialidade no IBGE. Procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas**. IBGE. Rio de Janeiro. 2018a.

IBGE. **Censo Demográfico 2010. Nota técnica 01/2018. Releitura dos dados de pessoas com deficiência no Censo Demográfico 2010 à luz das recomendações do Grupo de Washington**. IBGE. Rio de Janeiro. 2018b.

IBGE. **PNAD CONTÍNUA: Nota Técnica – Informações referentes à divulgação dos dados do segundo trimestre de 2020**. IBGE. Rio de Janeiro. 2020.

IBGE. **Código de Boas Práticas das Estatísticas do IBGE**. 2ª edição. IBGE. Rio de Janeiro. 2021a.

IBGE. **Pesquisa Nacional por Amostra de Domicílios Contínua. Notas técnicas Versão 1.8.** IBGE. Rio de Janeiro. 2021b.

IBGE. **PNAD CONTÍNUA: Nota técnica 04/2021 - Sobre a alteração do método de calibração dos fatores de expansão da PNAD Contínua**. IBGE. Rio de Janeiro. 2021c.

INDEC. **Secreto estadístico**. Disposición INDEC nº 176/99. INDEC. Buenos Aires, 1999.

INEGI. **Aspectos relevantes sobre la gestión de confidencialidad y protección de datos personales en el INEGI y el SNIEG**. Instituto Nacional de Estadística y Geografía. 2019.

INEI. **Código de Buenas Prácticas Estadísticas**. Instituto Nacional de Estadística e Informática. Peru. 2012.

INE,IP. **Política de confidencialidade estatística**. Instituto Nacional de Estatística. Portugal. 2019.

INEP. **Resumo técnico do Censo da Educação Superior 2019**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Brasília. 2021.

ISTAT. **Methods and IT tools for statistical production**. Instituto Nazionale di Statistica. Itália. 2018.

KOOIMAN P.; WILLENBORG L.C.R.J.; GOUWELEEUW J. **PRAM: a method for disclosure limitation of microdata**. Technical report, Statistics Netherlands (Voorburg, NL). 1998.

MACHADO, J.H.; FAMÁ, R. Ativos intangíveis e governança corporativa no mercado de capitais brasileiro. **Revista Contemporânea de Contabilidade**, v. 8, n. 16, p. 89-109. 2011.

MAGOSSI, J.C.; PAVIOTTI, J.R. Incerteza em entropia. **Revista Brasileira de História da Ciência**, Rio de Janeiro, v. 12, n. 1, p. 84-96, 2019.

McKENNA, L. **A History of the Current Population Survey and Disclosure Avoidance**. Research and Methodology Directorate. U.S. Census Bureau. 2019a.

McKENNA, L. **A History of the Economic Census and Disclosure Avoidance**. Research and Methodology Directorate. U.S. Census Bureau. 2019b.

McKENNA, L. **A History of the Survey of Income and Program Participation and Disclosure Avoidance**. Research and Methodology Directorate. U.S. Census Bureau. 2019c.

McKENNA, L. **Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples**. Research and Methodology Directorate. U.S. Census Bureau. 2019d.

MINISTÉRIO DA ECONOMIA. **Manual de Orientação. CAGED - Cadastro Geral de Empregados e Desempregados.** LEI Nº. 4923/65. Brasília. 2021.

MINISTÉRIO DO TRABALHO E EMPREGO. **Manual de Orientação. RAIS - Relação Anual de Informações Sociais**. Ano-Base 2020. Brasília. 2021.

MILLER, A. R. The national data center and personal privacy. **The Atlantic**, v.11, p. 53-57. 1967.

MURALIDHAR, K.; SARATHY, R. Data shuffling - A new masking approach for numerical data. **Management Science**, v. 52, n. 5, p. 658-670, 2006.

O'KEEFE, C.M.; SHLOMO, N. Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data. **Transactions on Data Privacy**, v. 5, n. 2, p. 403-432, 2012.

ONS. **Disclosure control guidance for birth and death statistics**. Briefing note (revised Jan 2014) on the publication of tabular data. Office for National Statistics. Reino Unido. 2014.

ONS. **Review of the Dissemination of Health Statistics: Confidentiality Guidance**. Office for National Statistics. Reino Unido. 2006.

ONU. **United Nations Fundamental Principles of Official Statistics**. Organização das Nações Unidas, 2015.

POLETTINI, S. Some remarks on the individual risk methodology. **In Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality**, Luxembourg, 2003.

POLETTINI, S., STANDER, J.: A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In: **International Workshop on Privacy in Statistical Databases**. Springer, Berlin, Heidelberg, p. 247-261. 2004.

REISS, S.P.; POST, M.J.; DALENIUS, T. Non-reversible privacy transformations. In: **Proceedings of the 1st ACM SIGACT-SIGMOD Symposium on Principles of Database Systems**. 1982. p. 139-146.

RINOTT, Y. On models for statistical disclosure risk estimation. In: **Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality**. 2003.

ROCHA, D.F. Concorrência em Mercados Digitais e Desafios ao Controle de Atos de Concentração. **Revista de Defesa da Concorrência**, v. 7, n. 2, p. 99-121, 2019.

RODRIGUEZ, R.A. Disclosure avoidance and the american community survey. In: **2021 ACS Data Users Conference**. 2021.

SAMBIASE, A.G.F.; BASTOS, B.B.; DE ANDRADE, K.R. Estratégias para o aprimoramento contínuo do cadastro único: Plano de Monitoramento da Gestão Municipal. **Revista Brasileira de Avaliação**, v. 6, p. 92-107, 2020.

SANSANA, A.G. **Privacidade, Consentimento, Legítimo Interesse e a Nova Lei Geral de Proteção de Dados Pessoais**. 2018. Trabalho de conclusão de curso (Pós-graduação Lato Sensu em Direito Empresarial – LLC) - Insper, São Paulo, 2018.

SANTOS, Y.T.; KOWATA, E.T. A importância do Big Data nas organizações. **V Congresso de Ensino, Pesquisa e Extensão da UEG.** Universidade Estadual de Goiás. Goiás. 2018.

SAWYER, J.; SCHECHTER, H. Computers, privacy, and the national data center: The responsibility of social scientists. **American Psychologist**, v. 23, n. 11, p. 810, 1968.

SEBÉ, F. et al. Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In: **Inference Control in Statistical Databases**. Springer, Berlin, Heidelberg, p. 163-171. 2002.

SHANNON, C.E. A mathematical theory of communication. **The Bell system technical journal**, v. 27, n. 3, p. 379-423, 1948.

SHLOMO, N. Statistical disclosure limitation: New directions and challenges. **Journal of Privacy and Confidentiality**, v. 8, n. 1, 2018.

SINGER, E.; VON THURN, D.R.; MILLER, E.R. Confidentiality assurances and response: A quantitative review of the experimental literature. **Public Opinion Quarterly**, v. 59, n. 1, p. 66-77, 1995.

SILVA, K.M.L. **Avaliação do Risco de Revelação em Microdados: Análise dos Microdados da Amostra do Censo Demográfico 2010**. Dissertação de Mestrado, ENCE, Rio de Janeiro, 2020.

SILVA, P.L.N. O sigilo das informações estatísticas: ideias para reflexão. **Textos para discussão**, IBGE, v. 1, n. 4, 1988.

SKINNER, C.J. Statistical disclosure control for survey data. In: **Handbook of statistics**. Elsevier, 2009. p. 381-396.

SKINNER, C.J.; HOLMES, D.J. Estimating the re-identification risk per record in microdata. **Journal of Official Statistics**, v. 14, n. 4, p. 361, 1998.

SNORRASON, H.; POURBAIX, I.; MATĚJČEK, J. Peer review report. **On compliance with the code of practice and the coordination role of the National Statistical Institute. Spain.** Eurostat. 2015.

STATCAN. **Statistics Canada Quality Guidelines**. Sixth Edition. Statistics Canada / Statistique Canada. 2019.

SULLIVAN, G.R. **The use of added error to avoid disclosure in microdata releases**. Phd Thesis. Iowa State University, 1989.

TAYLOR, L.; ZHOU, X.; RISE, P. A tutorial in assessing disclosure risk in microdata. **Statistics in Medicine**, v. 37, n. 25, p. 3693-3706, 2018.

TEMPL, M. Statistical disclosure control for microdata using the R-package sdcMicro. **Transactions on Data Privacy**, v. 1, n. 2, p. 67-85, 2008.

TEMPL, M. **Statistical disclosure control for microdata**. Cham: Springer, 2017.

US CENSUS BUREAU. **American Community Survey Information Guide**. U.S. Census Bureau. Issued October 2017.

WAAL, A.G.; WILLENBORG, L.C.R.J. Global recodings and local suppressions in microdata sets. **Proceedings of Statistics Canada Symposium'95**, p. 121–132. Statistics Canada, Ottawa, 1995.

WESTIN, A.F. **Privacy & Freedom.** The Bodley Head, London. 1967.

WILLENBORG, L.; DE WAAL, T. **Statistical disclosure control in practice**. Springer Science & Business Media, 1996.

WILLENBORG, L.; DE WAAL, T.. **Elements of statistical disclosure control**. Springer Science & Business Media, 2001.

# APÊNDICE – PROGRAMAS EM R

```
library("PNADcIBGE")
library("sdcMicro")

#Ler o arquivo salvo da PNAD
pnadc=read.table("C:\\Docs\\Doutorado\\Tese\\PNADC\\pnadc_ita2.txt",sep=";",head
er = T)

#cálculos para recorte “capital” e no cenário 2. Repetem-se as mesmas funções para
#outros recortes e cenário 1.

#Conta o número de chaves – usado para cálculo da unicidade amostral
conta2cap=freqCalc(pnadc,keyVars=c("V2001","V2007","V2009","V2010",
"VD3004","UF","Capital"))$fk #freq cen2 Cap

#especifica variáveis-chave, peso e hierarquia para cenários e recortes geográficos
sdc_cap2 <- createSdcObj(dat=pnadc,keyVars=c("V2001","V2007","V2009", "V2010",
"VD3004","UF","Capital"), numVars=NULL,weightVar="V1028",hhId="domic",
strataVar=NULL,pramVars=NULL,excludeVars=NULL,seed=0,randomizeRecords=FALSE,
alpha=c(1))

#Calcula o risco de revelação individual pela abordagem italiana
risco_cap2 <- get.sdcMicroObj(sdc_cap2, type="risk")$individual

#Cálculo do risco global
print(sdc_cap2, "risk")

#Análise exploratória - faz frequência das variáveis para ver o q pode ser recodificado
#globalmente
freqV2001 <- table(pnadc$V2001)
freqV2007 <- table(pnadc$V2007)
freqV2009 <- table(pnadc$V2009)
freqV2010 <- table(pnadc$V2010)
freqVD3004 <- table(pnadc$VD3004)

#Mascaramento das variáveis idade e tamanho do domicílio por uma configuração de
#recodificação – gerado automaticamente pelo SdcApp
obj <- NULL
if (!exists("pnadc")) { top('object "pnadc" is missing; make sure it exists.`', call.  FALSE)}
obj$inputdata <- readMicrodata(path="pnadc", type="rdf", convertCharToFac=FALSE,
drop_all_missings=FALSE)
inputdataB <- obj$inputdata
inputdata <- varToFactor(obj=inputdata, var="UF")
```

```
inputdata <- varToFactor(obj=inputdata, var="Capital")
inputdata <- varToFactor(obj=inputdata, var="V2001")
inputdata <- varToFactor(obj=inputdata, var="V2007")
inputdata <- varToFactor(obj=inputdata, var="V2009")
inputdata <- varToFactor(obj=inputdata, var="V2010")
inputdata <- varToFactor(obj=inputdata, var="VD3004")

sdcObj <- createSdcObj(dat=inputdata,
      keyVars=c("UF","Capital","V2001","V2007","V2009","V2010","VD3004"),
      numVars=NULL, weightVar=c("V1028"), hhId=c("domic"),  strataVar=NULL,
      pramVars=NULL, excludeVars=NULL, seed=0, randomizeRecords=FALSE,
      alpha=c(1))

## Store name of uploaded file
opts <- get.sdcMicroObj(sdcObj, type="options")
opts$filename <- "pnadc"
sdcObj <- set.sdcMicroObj(sdcObj, type="options", input=list(opts))

## Recode Variable - idade
sdcObj <- groupAndRename(obj=sdcObj, var="V2009",
before=c("0","1","2","3","4","5","6","7","8","9","10","11","12","13"), after=c("0a13"),
addNA=FALSE)
sdcObj <- groupAndRename(obj=sdcObj, var="V2009", before=c("14","15","16","17"),
after=c("14a17"), addNA=FALSE)
sdcObj <- groupAndRename(obj=sdcObj, var="V2009",
before=c("18","19","20","21","22","23","24"), after=c("18a24"), addNA=FALSE)
sdcObj <- groupAndRename(obj=sdcObj, var="V2009",
before=c("25","26","27","28","29","30","31","32","33","34","35","36","37","38","39"),
after=c("25a39"), addNA=FALSE)
sdcObj <- groupAndRename(obj=sdcObj, var="V2009",
before=c("40","41","42","43","44","45","46","47","48","49","50","51","52","53","54","
55","56","57","58","59"), after=c("40a59"), addNA=FALSE)
sdcObj <- groupAndRename(obj=sdcObj, var="V2009",
before=c("60","61","62","63","64","65","66","67","68","69","70","71","72","73","74","
75","76","77","78","79","80","81","82","83","84","85","86","87","88","89","90","91","
92","93","94","95","96","97","98","99","100","101","102","103","104","105","106","1
07","108","109","110","111","112","114"), after=c("60_+"), addNA=FALSE)

## Recode variable - tamanho dom
sdcObj <- groupAndRename(obj=sdcObj, var="V2001",
before=c("8","9","10","11","12","13","14","15","16","17","19","20","22"),
after=c("8_+"), addNA=FALSE)

## Suppression of risky observations above threshold in specified variable
sdcObj <- localSupp(sdcObj, threshold=0.075, keyVar="V2010")
```

```
## Suppression of risky observations above threshold in specified variable
sdcObj <- localSupp(sdcObj, threshold=0.075, keyVar="VD3004")

## Suppression of risky observations above threshold in specified variable
sdcObj <- localSupp(sdcObj, threshold=0.075, keyVar="Capital")

#Após o sdcApp, feito o mascaramento, abrir os arquivos resultantes
anom_3l <- dat
sdc_anom3l <- createSdcObj(dat=anom_3l,keyVars=c("V2001","V2007","V2009",
"V2010", "VD3004","UF","Capital"), numVars=NULL,weightVar="V1028",
hhId="domic", strataVar=NULL, pramVars=NULL,excludeVars=NULL,seed=0,
 randomizeRecords=FALSE,alpha=c(1))
risco_anom3l <- get.sdcMicroObj(sdc_anom3l, type="risk")$individual

#Calcula a medida de entropia
entropy <- function(fk, n){ 1/n * sum(fk*log(fk/n)) }
n <- nrow(pnadc)

#para os dados originais
entropy(as.numeric(table(pnadc$V2001)), n)
entropy(as.numeric(table(pnadc$V2009)), n)
entropy(as.numeric(table(pnadc$V2010)), n)
entropy(as.numeric(table(pnadc$VD3004)), n)
entropy(as.numeric(table(pnadc$Capital)), n)

#para os dados mascarados
entropy(as.numeric(table(anom_3l$V2009)), n)
entropy(as.numeric(table(anom_3l$V2010)), n)
entropy(as.numeric(table(anom_3l$VD3004)), n)
entropy(as.numeric(table(anom_3l$Capital)), n)
```

# ANEXO A

Quadro A1: Categorias da variável referente à variável indicadora da Unidade da Federação da PNAD Contínua 2019 e seus respectivos rótulos.

| Categoria | Descrição |
|---|---|
| 11 | Rondônia |
| 12 | Acre |
| 13 | Amazonas |
| 14 | Roraima |
| 15 | Pará |
| 16 | Amapá |
| 17 | Tocantins |
| 21 | Maranhão |
| 22 | Piauí |
| 23 | Ceará |
| 24 | Rio Grande do Norte |
| 25 | Paraíba |
| 26 | Pernambuco |
| 27 | Alagoas |
| 28 | Sergipe |
| 29 | Bahia |
| 31 | Minas Gerais |
| 32 | Espírito Santo |
| 33 | Rio de Janeiro |
| 35 | São Paulo |
| 41 | Paraná |
| 42 | Santa Catarina |
| 43 | Rio Grande do Sul |
| 50 | Mato Grosso do Sul |
| 51 | Mato Grosso |
| 52 | Goiás |
| 53 | Distrito Federal |

Fonte: PNAD Contínua 2019 – 2º trimestre

Quadro A2: Categorias da variável referente à variável indicadora de município da capital da PNAD Contínua 2019 e seus respectivos rótulos.

| Categoria | Descrição |
|---|---|
| 11 | Município de Porto Velho (RO) |
| 12 | Município de Rio Branco (AC) |
| 13 | Município de Manaus (AM) |
| 14 | Município de Boa Vista (RR) |
| 15 | Município de Belém (PA) |
| 16 | Município de Macapá (AP) |
| 17 | Município de Palmas (TO) |
| 21 | Município de São Luís (MA) |
| 22 | Município de Teresina (PI) |
| 23 | Município de Fortaleza (CE) |
| 24 | Município de Natal (RN) |
| 25 | Município de João Pessoa (PB) |
| 26 | Município de Recife (PE) |
| 27 | Município de Maceió (AL) |
| 28 | Município de Aracaju (SE) |
| 29 | Município de Salvador (BA) |
| 31 | Município de Belo Horizonte (MG) |
| 32 | Município de Vitória (ES) |
| 33 | Município de Rio de Janeiro (RJ) |
| 35 | Município de São Paulo (SP) |
| 41 | Município de Curitiba (PR) |
| 42 | Município de Florianópolis (SC) |
| 43 | Município de Porto Alegre (RS) |
| 50 | Município de Campo Grande (MS) |
| 51 | Município de Cuiabá (MT) |
| 52 | Município de Goiânia (GO) |
| 53 | Município de Brasília (DF) |

Fonte: PNAD Contínua 2019 – 2º trimestre

Quadro A3: Categorias da variável referente à variável indicadora de Região Metropolitana ou Região Administrativa Integrada de Desenvolvimento da PNAD Contínua 2019 e seus respectivos rótulos.

| Categoria | Descrição |
|---|---|
| 13 | Região Metropolitana de Manaus (AM) |
| 15 | Região Metropolitana de Belém (PA) |
| 16 | Região Metropolitana de Macapá (AP) |
| 21 | Região Metropolitana de Grande São Luís (MA) |
| 22 | Região Administrativa Integrada de Desenvolvimento da Grande Teresina (PI) |
| 23 | Região Metropolitana de Fortaleza (CE) |
| 24 | Região Metropolitana de Natal (RN) |
| 25 | Região Metropolitana de João Pessoa (PB) |
| 26 | Região Metropolitana de Recife (PE) |
| 27 | Região Metropolitana de Maceió (AL) |
| 28 | Região Metropolitana de Aracaju (SE) |
| 29 | Região Metropolitana de Salvador (BA) |
| 31 | Região Metropolitana de Belo Horizonte (MG) |
| 32 | Região Metropolitana de Grande Vitória (ES) |
| 33 | Região Metropolitana de Rio de Janeiro (RJ) |
| 35 | Região Metropolitana de São Paulo (SP) |
| 41 | Região Metropolitana de Curitiba (PR) |
| 42 | Região Metropolitana de Florianópolis (SC) |
| 43 | Região Metropolitana de Porto Alegre (RS) |
| 51 | Região Metropolitana de Vale do Rio Cuiabá (MT) |
| 52 | Região Metropolitana de Goiânia (GO) |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A1: Unicidades amostrais nos registros de pessoas da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Grandes Regiões.

| Grandes Regiões | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **551.348** | **8.516** | **1,5** | **27.438** | **5,0** |
| 1.Norte | 77.416 | 1.983 | 2,6 | 6.470 | 8,4 |
| 2.Nordeste | 186.029 | 1.604 | 0,9 | 6.116 | 3,3 |
| 3.Sudeste | 142.691 | 1.507 | 1,1 | 5.350 | 3,7 |
| 4.Sul | 90.806 | 1.689 | 1,9 | 4.601 | 5,1 |
| 5.Centro-Oeste | 54.406 | 1.733 | 3,2 | 4.901 | 9,0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A2: Unicidades amostrais nos registros de domicílios da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Grandes Regiões.

| Grandes Regiões | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **187.061** | **7.578** | **4,1** | **19.550** | **10,5** |
| 1.Norte | 22.911 | 1.738 | 7,6 | 4.345 | 19,0 |
| 2.Nordeste | 60.585 | 1.471 | 2,4 | 4.447 | 7,3 |
| 3.Sudeste | 51.446 | 1.377 | 2,7 | 3.923 | 7,6 |
| 4.Sul | 33.318 | 1.440 | 4,3 | 3.300 | 9,9 |
| 5.Centro-Oeste | 18.801 | 1.552 | 8,3 | 3.535 | 18,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A3: Unicidades amostrais nos registros de pessoas da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidades da Federação.

| Unidades da Federação | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **551.348** | **43.796** | **7,9** | **96.902** | **17,6** |
| 11.Rondônia | 9.314 | 1.507 | 16,2 | 2.763 | 29,7 |
| 12.Acre | 10.997 | 1.461 | 13,3 | 3.039 | 27,6 |
| 13.Amazonas | 16.609 | 1.582 | 9,5 | 3.980 | 24,0 |
| 14.Roraima | 6.594 | 1.448 | 22,0 | 3.020 | 45,8 |
| 15.Pará | 20.705 | 1.757 | 8,5 | 4.256 | 20,6 |
| 16.Amapá | 5.027 | 1.077 | 21,4 | 2.445 | 48,6 |
| 17.Tocantins | 8.170 | 1.460 | 17,9 | 2.788 | 34,1 |
| 21.Maranhão | 37.753 | 1.783 | 4,7 | 4.944 | 13,1 |
| 22.Piauí | 13.082 | 1.545 | 11,8 | 3.231 | 24,7 |
| 23.Ceará | 28.316 | 1.645 | 5,8 | 4.096 | 14,5 |
| 24.Rio Grande do Norte | 11.492 | 1.706 | 14,8 | 3.348 | 29,1 |
| 25.Paraíba | 14.534 | 1.685 | 11,6 | 3.422 | 23,5 |
| 26.Pernambuco | 22.377 | 1.713 | 7,7 | 3.859 | 17,2 |
| 27.Alagoas | 21.101 | 1.645 | 7,8 | 3.896 | 18,5 |
| 28.Sergipe | 10.600 | 1.571 | 14,8 | 3.015 | 28,4 |
| 29.Bahia | 26.774 | 1.842 | 6,9 | 4.343 | 16,2 |
| 31.Minas Gerais | 40.558 | 1.502 | 3,7 | 4.060 | 10,0 |
| 32.Espírito Santo | 20.227 | 1.812 | 9,0 | 3.582 | 17,7 |
| 33.Rio de Janeiro | 39.129 | 1.685 | 4,3 | 4.393 | 11,2 |
| 35.São Paulo | 42.777 | 1.838 | 4,3 | 4.535 | 10,6 |
| 41.Paraná | 29.155 | 1.654 | 5,7 | 3.696 | 12,7 |
| 42.Santa Catarina | 33.103 | 1.489 | 4,5 | 3.343 | 10,1 |
| 43.Rio Grande do Sul | 28.548 | 1.620 | 5,7 | 3.347 | 11,7 |
| 50.Mato Grosso do Sul | 12.082 | 1.753 | 14,5 | 3.211 | 26,6 |
| 51.Mato Grosso | 13.845 | 1.671 | 12,1 | 3.506 | 25,3 |
| 52.Goiás | 18.486 | 1.747 | 9,5 | 3.511 | 19,0 |
| 53.Distrito Federal | 9.993 | 1.598 | 16,0 | 3.273 | 32,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A4: Unicidades amostrais nos registros de domicílios da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidades da Federação.

| Unidades da Federação | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **187.061** | **36.252** | **19,4** | **64.231** | **34,3** |
| 11.Rondônia | 3.115 | 1.198 | 38,5 | 1.823 | 58,5 |
| 12.Acre | 3.343 | 1.145 | 34,3 | 1.900 | 56,8 |
| 13.Amazonas | 4.414 | 1.170 | 26,5 | 2.296 | 52,0 |
| 14.Roraima | 1.870 | 1.000 | 53,5 | 1.521 | 81,3 |
| 15.Pará | 6.070 | 1.440 | 23,7 | 2.642 | 43,5 |
| 16.Amapá | 1.342 | 746 | 55,6 | 1.134 | 84,5 |
| 17.Tocantins | 2.757 | 1.129 | 41,0 | 1.741 | 63,1 |
| 21.Maranhão | 11.083 | 1.580 | 14,3 | 3.378 | 30,5 |
| 22.Piauí | 4.115 | 1.255 | 30,5 | 2.077 | 50,5 |
| 23.Ceará | 9.272 | 1.436 | 15,5 | 2.822 | 30,4 |
| 24.Rio Grande do Norte | 3.703 | 1.382 | 37,3 | 2.106 | 56,9 |
| 25.Paraíba | 4.796 | 1.393 | 29,0 | 2.300 | 48,0 |
| 26.Pernambuco | 7.642 | 1.431 | 18,7 | 2.627 | 34,4 |
| 27.Alagoas | 7.086 | 1.386 | 19,6 | 2.604 | 36,7 |
| 28.Sergipe | 3.608 | 1.272 | 35,3 | 1.927 | 53,4 |
| 29.Bahia | 9.280 | 1.595 | 17,2 | 3.005 | 32,4 |
| 31.Minas Gerais | 14.279 | 1.405 | 9,8 | 2.952 | 20,7 |
| 32.Espírito Santo | 7.198 | 1.565 | 21,7 | 2.595 | 36,1 |
| 33.Rio de Janeiro | 14.731 | 1.540 | 10,5 | 3.270 | 22,2 |
| 35.São Paulo | 15.238 | 1.605 | 10,5 | 3.246 | 21,3 |
| 41.Paraná | 10.324 | 1.391 | 13,5 | 2.588 | 25,1 |
| 42.Santa Catarina | 11.984 | 1.248 | 10,4 | 2.331 | 19,5 |
| 43.Rio Grande do Sul | 11.010 | 1.361 | 12,4 | 2.342 | 21,3 |
| 50.Mato Grosso do Sul | 4.135 | 1.424 | 34,4 | 2.075 | 50,2 |
| 51.Mato Grosso | 4.724 | 1.386 | 29,3 | 2.339 | 49,5 |
| 52.Goiás | 6575 | 1.506 | 22,9 | 2.550 | 38,8 |
| 53.Distrito Federal | 3367 | 1.263 | 37,5 | 2.040 | 60,6 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A5: Unicidades amostrais nos registros de pessoas da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e RM / RIDE.

| RM ou RIDE | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **170.401** | **29.629** | **17,4** | **56.219** | **33,0** |
| 13.RM Manaus | 8.846 | 1.258 | 14,2 | 2.842 | 32,1 |
| 15.RM Belém | 6.023 | 1.350 | 22,4 | 2.644 | 43,9 |
| 16.RM Macapá | 3.942 | 978 | 24,8 | 2.125 | 53,9 |
| 21.RM Grande São Luís | 6.772 | 1.324 | 19,6 | 2.738 | 40,4 |
| 22.RIDE Grande Teresina | 5.885 | 1.825 | 31,0 | 2.891 | 49,1 |
| 23.RM Fortaleza | 9.432 | 1.542 | 16,3 | 3.093 | 32,8 |
| 24.RM Natal | 4.732 | 1.414 | 29,9 | 2.408 | 50,9 |
| 25.RM João Pessoa | 5.299 | 1.393 | 26,3 | 2.459 | 46,4 |
| 26.RM Recife | 8.032 | 1.433 | 17,8 | 2.822 | 35,1 |
| 27.RM Maceió | 6.188 | 1.265 | 20,4 | 2.400 | 38,8 |
| 28.RM Aracaju | 3.856 | 1.104 | 28,6 | 1.948 | 50,5 |
| 29.RM Salvador | 5.686 | 1.460 | 25,7 | 2.580 | 45,4 |
| 31.RM Belo Horizonte | 9.896 | 1.504 | 15,2 | 3.127 | 31,6 |
| 32.RM Grande Vitória | 9.477 | 1.687 | 17,8 | 2.949 | 31,1 |
| 33.RM Rio de Janeiro | 25.895 | 1.843 | 7,1 | 4.122 | 15,9 |
| 35.RM São Paulo | 17.852 | 1.755 | 9,8 | 3.768 | 21,1 |
| 41.RM Curitiba | 8.417 | 1.392 | 16,5 | 2.461 | 29,2 |
| 42.RM Florianópolis | 4.221 | 1.009 | 23,9 | 1.633 | 38,7 |
| 43.RM Porto Alegre | 9.604 | 1.404 | 14,6 | 2.553 | 26,6 |
| 51.RM Vale do Rio Cuiabá | 3.808 | 1.200 | 31,5 | 2.135 | 56,1 |
| 52.RM Goiânia | 6.538 | 1.489 | 22,8 | 2.521 | 38,6 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A6: Unicidades amostrais nos registros de domicílios da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e RM / RIDE.

| RM ou RIDE | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **59.691** | **22.274** | **37,3** | **33.965** | **56,9** |
| 13.RM Manaus | 2.599 | 951 | 36,6 | 1.620 | 62,3 |
| 15.RM Belém | 1.818 | 963 | 53,0 | 1.400 | 77,0 |
| 16.RM Macapá | 1.045 | 637 | 61,0 | 926 | 88,6 |
| 21.RM Grande São Luís | 2.043 | 947 | 46,4 | 1.505 | 73,7 |
| 22.RIDE Grande Teresina | 1.817 | 1.139 | 62,7 | 1.463 | 80,5 |
| 23.RM Fortaleza | 3.139 | 1.224 | 39,0 | 1.971 | 62,8 |
| 24.RM Natal | 1.544 | 956 | 61,9 | 1.299 | 84,1 |
| 25.RM João Pessoa | 1.737 | 990 | 57,0 | 1.399 | 80,5 |
| 26.RM Recife | 2.813 | 1.095 | 38,9 | 1.765 | 62,7 |
| 27.RM Maceió | 2.483 | 987 | 39,8 | 1.488 | 59,9 |
| 28.RM Aracaju | 1.306 | 748 | 57,3 | 1.058 | 81,0 |
| 29.RM Salvador | 2.069 | 1.098 | 53,1 | 1.535 | 74,2 |
| 31.RM Belo Horizonte | 3.519 | 1.243 | 35,3 | 2.040 | 58,0 |
| 32.RM Grande Vitória | 3.412 | 1.350 | 39,6 | 1.980 | 58,0 |
| 33.RM Rio de Janeiro | 9.851 | 1.637 | 16,6 | 3.006 | 30,5 |
| 35.RM São Paulo | 6.416 | 1.468 | 22,9 | 2.558 | 39,9 |
| 41.RM Curitiba | 3.006 | 1.074 | 35,7 | 1.565 | 52,1 |
| 42.RM Florianópolis | 1.683 | 715 | 42,5 | 982 | 58,3 |
| 43.RM Porto Alegre | 3.780 | 1.096 | 29,0 | 1.671 | 44,3 |
| 51.RM Vale do Rio Cuiabá | 1.248 | 817 | 65,5 | 1.101 | 88,2 |
| 52.RM Goiânia | 2.363 | 1.139 | 48,2 | 1.633 | 69,2 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A7: Unicidades amostrais nos registros de pessoas, em municípios fora da RM/RIDE, da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidade da Federação.

| Municípios fora da RM/RIDE na UF | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **323.797** | **31.405** | **9,7** | **63.467** | **19,6** |
| 13.Amazonas | 7.763 | 1.198 | 15,4 | 2.595 | 33,4 |
| 15.Pará | 14.682 | 1.619 | 11,0 | 3.451 | 23,5 |
| 16.Amapá | 1.085 | 516 | 47,6 | 753 | 69,4 |
| 21.Maranhão | 30.007 | 1.679 | 5,6 | 4.350 | 14,5 |
| 22.Piauí | 8.171 | 1.404 | 17,2 | 2.539 | 31,1 |
| 23.Ceará | 18.884 | 1.504 | 8,0 | 3.405 | 18,0 |
| 24.Rio Grande do Norte | 6.760 | 1.458 | 21,6 | 2.388 | 35,3 |
| 25.Paraíba | 9.235 | 1.445 | 15,6 | 2.612 | 28,3 |
| 26.Pernambuco | 14.345 | 1.533 | 10,7 | 3.112 | 21,7 |
| 27.Alagoas | 14.913 | 1.585 | 10,6 | 3.343 | 22,4 |
| 28.Sergipe | 6.744 | 1.368 | 20,3 | 2.247 | 33,3 |
| 29.Bahia | 21.088 | 1.769 | 8,4 | 3.885 | 18,4 |
| 31.Minas Gerais | 30.662 | 1.604 | 5,2 | 3.777 | 12,3 |
| 32.Espírito Santo | 10.750 | 1.550 | 14,4 | 2.838 | 26,4 |
| 33.Rio de Janeiro | 13.234 | 1.741 | 13,2 | 3.288 | 24,8 |
| 35.São Paulo | 24.925 | 1.782 | 7,1 | 3.845 | 15,4 |
| 41.Paraná | 20.738 | 1.628 | 7,9 | 3.254 | 15,7 |
| 42.Santa Catarina | 28.882 | 1.405 | 4,9 | 3.099 | 10,7 |
| 43.Rio Grande do Sul | 18.944 | 1.448 | 7,6 | 2.795 | 14,8 |
| 51.Mato Grosso | 10.037 | 1.567 | 15,6 | 2.930 | 29,2 |
| 52.Goiás | 11.948 | 1.602 | 13,4 | 2.961 | 24,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A8: Unicidades amostrais nos registros de domicílios, em municípios fora da RM/RIDE, da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidade da Federação.

| Municípios fora da RM/RIDE na UF | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **108.783** | **25.530** | **23,5** | **41.854** | **38,5** |
| 13.Amazonas | 1.815 | 803 | 44,2 | 1.304 | 71,8 |
| 15.Pará | 4.252 | 1.255 | 29,5 | 2.085 | 49,0 |
| 16.Amapá | 297 | 252 | 84,8 | 289 | 97,3 |
| 21.Maranhão | 8.753 | 1.471 | 16,8 | 2.950 | 33,7 |
| 22.Piauí | 2.585 | 1.058 | 40,9 | 1.550 | 60,0 |
| 23.Ceará | 6.133 | 1.307 | 21,3 | 2.280 | 37,2 |
| 24.Rio Grande do Norte | 2.159 | 1.101 | 51,0 | 1.452 | 67,3 |
| 25.Paraíba | 3.059 | 1.144 | 37,4 | 1.711 | 55,9 |
| 26.Pernambuco | 4.829 | 1.256 | 26,0 | 2.049 | 42,4 |
| 27.Alagoas | 4.603 | 1.269 | 27,6 | 2.104 | 45,7 |
| 28.Sergipe | 2.302 | 1.052 | 45,7 | 1.390 | 60,4 |
| 29.Bahia | 7.211 | 1.515 | 21,0 | 2.653 | 36,8 |
| 31.Minas Gerais | 10.760 | 1.471 | 13,7 | 2.787 | 25,9 |
| 32.Espírito Santo | 3.786 | 1.256 | 33,2 | 1.918 | 50,7 |
| 33.Rio de Janeiro | 4.880 | 1.454 | 29,8 | 2.279 | 46,7 |
| 35.São Paulo | 8.822 | 1.519 | 17,2 | 2.684 | 30,4 |
| 41.Paraná | 7.318 | 1.358 | 18,6 | 2.284 | 31,2 |
| 42.Santa Catarina | 10.301 | 1.183 | 11,5 | 2.157 | 20,9 |
| 43.Rio Grande do Sul | 7.230 | 1.184 | 16,4 | 1.943 | 26,9 |
| 51.Mato Grosso | 3.476 | 1.277 | 36,7 | 1.952 | 56,2 |
| 52.Goiás | 4.212 | 1.345 | 31,9 | 2.033 | 48,3 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A9: Unicidades amostrais nos registros de pessoas da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e municípios das capitais.

| Capitais | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **126.336** | **30.045** | **23,8** | **55.540** | **44,0** |
| 11.Porto Velho (RO) | 3.020 | 942 | 31,2 | 1.659 | 54,9 |
| 12.Rio Branco (AC) | 5.346 | 1.177 | 22,0 | 2.231 | 41,7 |
| 13.Manaus (AM) | 7.137 | 1.184 | 16,6 | 2.562 | 35,9 |
| 14.Boa Vista (RR) | 4.167 | 1.107 | 26,6 | 2.214 | 53,1 |
| 15.Belém (PA) | 4.072 | 1.172 | 28,7 | 2.140 | 52,6 |
| 16.Macapá (AP) | 2.773 | 822 | 29,6 | 1.702 | 61,4 |
| 17.Palmas (TO) | 1.752 | 620 | 35,4 | 1.108 | 63,2 |
| 21.São Luís (MA) | 4.929 | 1.132 | 23,0 | 2.385 | 48,4 |
| 22.Teresina (PI) | 3.388 | 1.083 | 32,0 | 1.761 | 52,0 |
| 23.Fortaleza (CE) | 6.265 | 1.375 | 21,9 | 2.495 | 39,8 |
| 24.Natal (RN) | 2.881 | 1.098 | 38,1 | 1.838 | 63,8 |
| 25.João Pessoa (PB) | 3.432 | 1.072 | 31,2 | 1.911 | 55,7 |
| 26.Recife (PE) | 3.266 | 1.053 | 32,2 | 1.823 | 55,8 |
| 27.Maceió (AL) | 5.008 | 1.156 | 23,1 | 2.161 | 43,2 |
| 28.Aracaju (SE) | 2.501 | 888 | 35,5 | 1.550 | 62,0 |
| 29.Salvador (BA) | 3.654 | 1.182 | 32,3 | 2.046 | 56,0 |
| 31.Belo Horizonte (MG) | 5.346 | 1.176 | 22,0 | 2.371 | 44,4 |
| 32.Vitória (ES) | 2.512 | 834 | 33,2 | 1.486 | 59,2 |
| 33.Rio de Janeiro (RJ) | 13.034 | 1.674 | 12,8 | 3.415 | 26,2 |
| 35.São Paulo (SP) | 9.879 | 1.553 | 15,7 | 3.071 | 31,1 |
| 41.Curitiba (PR) | 4.805 | 1.077 | 22,4 | 1.795 | 37,4 |
| 42.Florianópolis (SC) | 2.328 | 682 | 29,3 | 1.148 | 49,3 |
| 43.Porto Alegre (RS) | 4.182 | 1.007 | 24,1 | 1.756 | 42,0 |
| 50.Campo Grande (MS) | 3.872 | 1.183 | 30,6 | 1.989 | 51,4 |
| 51.Cuiabá (MT) | 2.407 | 958 | 39,8 | 1.609 | 66,8 |
| 52.Goiânia (GO) | 4.387 | 1.242 | 28,3 | 2.041 | 46,5 |
| 53.Brasília (DF) | 9.993 | 1.598 | 16,0 | 3.273 | 32,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A10: Unicidades amostrais nos registros de domicílios da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e municípios das capitais.

| Capitais | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **43.994** | **21.270** | **48,3** | **31.180** | **70,9** |
| 11.Porto Velho (RO) | 951 | 610 | 64,1 | 832 | 87,5 |
| 12.Rio Branco (AC) | 1.720 | 879 | 51,1 | 1.255 | 73,0 |
| 13.Manaus (AM) | 2.147 | 865 | 40,3 | 1.434 | 66,8 |
| 14.Boa Vista (RR) | 1.173 | 715 | 61,0 | 1.038 | 88,5 |
| 15.Belém (PA) | 1.237 | 777 | 62,8 | 1.053 | 85,1 |
| 16.Macapá (AP) | 743 | 502 | 67,6 | 693 | 93,3 |
| 17.Palmas (TO) | 586 | 400 | 68,3 | 541 | 92,3 |
| 21.São Luís (MA) | 1.482 | 779 | 52,6 | 1.229 | 82,9 |
| 22.Teresina (PI) | 1.058 | 712 | 67,3 | 920 | 87,0 |
| 23.Fortaleza (CE) | 2.108 | 1.028 | 48,8 | 1.505 | 71,4 |
| 24.Natal (RN) | 966 | 675 | 69,9 | 879 | 91,0 |
| 25.João Pessoa (PB) | 1.146 | 715 | 62,4 | 1.020 | 89,0 |
| 26.Recife (PE) | 1.170 | 708 | 60,5 | 999 | 85,4 |
| 27.Maceió (AL) | 2.035 | 876 | 43,0 | 1.326 | 65,2 |
| 28.Aracaju (SE) | 856 | 565 | 66,0 | 766 | 89,5 |
| 29.Salvador (BA) | 1.367 | 840 | 61,4 | 1.159 | 84,8 |
| 31.Belo Horizonte (MG) | 1.961 | 905 | 46,1 | 1.412 | 72,0 |
| 32.Vitória (ES) | 949 | 572 | 60,3 | 807 | 85,0 |
| 33.Rio de Janeiro (RJ) | 5.129 | 1.392 | 27,1 | 2.328 | 45,4 |
| 35.São Paulo (SP) | 3.631 | 1.217 | 33,5 | 1.990 | 54,8 |
| 41.Curitiba (PR) | 1.758 | 766 | 43,6 | 1.082 | 61,5 |
| 42.Florianópolis (SC) | 968 | 459 | 47,4 | 678 | 70,0 |
| 43.Porto Alegre (RS) | 1.742 | 732 | 42,0 | 1.064 | 61,1 |
| 50.Campo Grande (MS) | 1.322 | 793 | 60,0 | 1.106 | 83,7 |
| 51.Cuiabá (MT) | 814 | 617 | 75,8 | 761 | 93,5 |
| 52.Goiânia (GO) | 1.608 | 908 | 56,5 | 1.263 | 78,5 |
| 53.Brasília (DF) | 3.367 | 1.263 | 37,5 | 2.040 | 60,6 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A11: Unicidades amostrais nos registros de pessoas, em municípios fora da capital, da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidade da Federação.

| Municípios fora da Capital na UF | Registro de pessoas | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **425.012** | **39.165** | **9,2** | **80.240** | **18,9** |
| 11.Rondônia | 6.294 | 1.388 | 22,1 | 2.115 | 33,6 |
| 12.Acre | 5.651 | 1.125 | 19,9 | 2.098 | 37,1 |
| 13.Amazonas | 9.472 | 1.260 | 13,3 | 2.870 | 30,3 |
| 14.Roraima | 2.427 | 974 | 40,1 | 1.566 | 64,5 |
| 15.Pará | 16.633 | 1.658 | 10,0 | 3.713 | 22,3 |
| 16.Amapá | 2.254 | 762 | 33,8 | 1.371 | 60,8 |
| 17.Tocantins | 6.418 | 1.409 | 22,0 | 2.432 | 37,9 |
| 21.Maranhão | 32.824 | 1.708 | 5,2 | 4.541 | 13,8 |
| 22.Piauí | 9.694 | 1.451 | 15,0 | 2.676 | 27,6 |
| 23.Ceará | 2.2051 | 1.578 | 7,2 | 3.685 | 16,7 |
| 24.Rio Grande do Norte | 8.611 | 1.566 | 18,2 | 2.755 | 32,0 |
| 25.Paraíba | 11.102 | 1.518 | 13,7 | 2.883 | 26,0 |
| 26.Pernambuco | 19.111 | 1.666 | 8,7 | 3.540 | 18,5 |
| 27.Alagoas | 16.093 | 1.605 | 10,0 | 3.436 | 21,4 |
| 28.Sergipe | 8.099 | 1.450 | 17,9 | 2.507 | 31,0 |
| 29.Bahia | 23.120 | 1.791 | 7,7 | 3.978 | 17,2 |
| 31.Minas Gerais | 35.212 | 1.530 | 4,3 | 3.881 | 11,0 |
| 32.Espírito Santo | 17.715 | 1.756 | 9,9 | 3.395 | 19,2 |
| 33.Rio de Janeiro | 26.095 | 1.700 | 6,5 | 3.957 | 15,2 |
| 35.São Paulo | 32.898 | 1.787 | 5,4 | 4.106 | 12,5 |
| 41.Paraná | 24.350 | 1.621 | 6,7 | 3.406 | 14,0 |
| 42.Santa Catarina | 30.775 | 1.454 | 4,7 | 3.223 | 10,5 |
| 43.Rio Grande do Sul | 24.366 | 1.539 | 6,3 | 3.044 | 12,5 |
| 50.Mato Grosso do Sul | 8.210 | 1.584 | 19,3 | 2.696 | 32,8 |
| 51.Mato Grosso | 11.438 | 1.608 | 14,1 | 3.168 | 27,7 |
| 52.Goiás | 14.099 | 1.677 | 11,9 | 3.198 | 22,7 |
| 53.Distrito Federal | 0 | 0 | -- | 0 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A12: Unicidades amostrais nos registros de domicílios, em municípios fora da capital, da PNAD Contínua 2019 – 2º trimestre, por cenário de revelação e Unidade da Federação.

| Municípios fora da Capital na UF | Registro de domicílios | Unicidade - Cenário 1 | | Unicidade - Cenário 2 | |
|---|---|---|---|---|---|
| | | Total | % | Total | % |
| **Total** | **143.067** | **31.691** | **22,2** | **52.681** | **36,8** |
| 11.Rondônia | 2.164 | 1.060 | 49,0 | 1.394 | 64,4 |
| 12.Acre | 1.623 | 810 | 49,9 | 1.175 | 72,4 |
| 13.Amazonas | 2.267 | 873 | 38,5 | 1.491 | 65,8 |
| 14.Roraima | 697 | 549 | 78,8 | 646 | 92,7 |
| 15.Pará | 4.833 | 1.316 | 27,2 | 2.263 | 46,8 |
| 16.Amapá | 599 | 450 | 75,1 | 566 | 94,5 |
| 17.Tocantins | 2.171 | 1.049 | 48,3 | 1.454 | 67,0 |
| 21.Maranhão | 9.601 | 1.509 | 15,8 | 3.098 | 32,3 |
| 22.Piauí | 3.057 | 1.121 | 36,7 | 1.676 | 54,8 |
| 23.Ceará | 7.164 | 1.368 | 19,1 | 2.488 | 34,7 |
| 24.Rio Grande do Norte | 2.737 | 1.232 | 45,0 | 1.714 | 62,6 |
| 25.Paraíba | 3.650 | 1.222 | 33,5 | 1.892 | 51,8 |
| 26.Pernambuco | 6.472 | 1.378 | 21,3 | 2.382 | 36,8 |
| 27.Alagoas | 5.051 | 1.301 | 25,8 | 2.210 | 43,8 |
| 28.Sergipe | 2.752 | 1.136 | 41,3 | 1.595 | 58,0 |
| 29.Bahia | 7.913 | 1.542 | 19,5 | 2.734 | 34,6 |
| 31.Minas Gerais | 12.318 | 1.417 | 11,5 | 2.835 | 23,0 |
| 32.Espírito Santo | 6.249 | 1.490 | 23,9 | 2.433 | 38,9 |
| 33.Rio de Janeiro | 9.602 | 1.527 | 15,9 | 2.903 | 30,2 |
| 35.São Paulo | 11.607 | 1.550 | 13,4 | 2.932 | 25,3 |
| 41.Paraná | 8.566 | 1.349 | 15,7 | 2.390 | 27,9 |
| 42.Santa Catarina | 11016 | 1.217 | 11,1 | 2.254 | 20,5 |
| 43.Rio Grande do Sul | 9268 | 1.287 | 13,9 | 2.151 | 23,2 |
| 50.Mato Grosso do Sul | 2813 | 1.227 | 43,6 | 1.675 | 59,5 |
| 51.Mato Grosso | 3910 | 1.304 | 33,4 | 2.104 | 53,8 |
| 52.Goiás | 4967 | 1.407 | 28,3 | 2.226 | 44,8 |
| 53.Distrito Federal | 0 | 0 | -- | 0 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A13: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por Grandes Regiões divulgadas, segundo o cenário de revelação 1.

| Grandes Regiões | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **187.061** | **7.467** | **4,0** | **189** | **0,5** | **31** | **0,1** |
| 1.Norte | 22.911 | 2.068 | 9,0 | 134 | 2,4 | 25 | 0,1 |
| 2.Nordeste | 60.585 | 1.378 | 2,3 | 14 | 0,2 | 1 | <0,1 |
| 3.Sudeste | 51.446 | 1.009 | 2,0 | 6 | 0,1 | 3 | <0,1 |
| 4.Sul | 33.318 | 1.415 | 4,2 | 17 | 0,3 | 0 | 0.0 |
| 5.Centro-Oeste | 18.801 | 1.568 | 8,3 | 18 | 0,8 | 2 | 0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A14: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por Grandes Regiões divulgadas, segundo o cenário de revelação 2.

| Grandes Regiões | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **187.061** | **19.271** | **10,3** | **1.079** | **0,6** | **255** | **0,1** |
| 1.Norte | 22.911 | 4.966 | 21,7 | 561 | 2,4 | 163 | 0,7 |
| 2.Nordeste | 60.585 | 4.461 | 7,4 | 217 | 0,4 | 40 | 0,1 |
| 3.Sudeste | 51.446 | 3.025 | 5,9 | 76 | 0,1 | 19 | <0,1 |
| 4.Sul | 33.318 | 3.240 | 9,7 | 99 | 0,3 | 10 | <0,1 |
| 5.Centro-Oeste | 18.801 | 3.579 | 19,0 | 126 | 0,7 | 23 | 0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A15: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por Unidades da Federação divulgadas, segundo o cenário de revelação 1.

| Unidades da Federação | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **187.061** | **38.095** | **20,4** | **1.383** | **0,7** | **220** | **0,1** |
| 11.Rondônia | 3.115 | 1.432 | 46,0 | 47 | 1,5 | 1 | <0,1 |
| 12.Acre | 3.343 | 2.033 | 60,8 | 306 | 9,2 | 48 | 1,4 |
| 13.Amazonas | 4.414 | 1.309 | 29,7 | 90 | 2,0 | 28 | 0,6 |
| 14.Roraima | 1.870 | 1.467 | 78,4 | 385 | 20,6 | 85 | 4,5 |
| 15.Pará | 6.070 | 1.384 | 22,8 | 12 | 0,2 | 0 | 0,0 |
| 16.Amapá | 1.342 | 896 | 66,8 | 107 | 8,0 | 24 | 1,8 |
| 17.Tocantins | 2.757 | 1.344 | 48,7 | 54 | 2,0 | 8 | 0,3 |
| 21.Maranhão | 11.083 | 1.909 | 17,2 | 33 | 0,3 | 2 | <0,1 |
| 22.Piauí | 4.115 | 1.369 | 33,3 | 25 | 0,6 | 1 | <0,1 |
| 23.Ceará | 9.272 | 1.424 | 15,4 | 4 | 0,0 | 0 | 0,0 |
| 24.Rio Grande do Norte | 3.703 | 1.413 | 38,2 | 18 | 0,5 | 0 | 0,0 |
| 25.Paraíba | 4.796 | 1.485 | 31,0 | 21 | 0,4 | 0 | 0,0 |
| 26.Pernambuco | 7.642 | 1.349 | 17,7 | 7 | 0,1 | 0 | 0,0 |
| 27.Alagoas | 7.086 | 1.851 | 26,1 | 67 | 0,9 | 6 | 0,1 |
| 28.Sergipe | 3.608 | 1.437 | 39,8 | 21 | 0,6 | 0 | 0,0 |
| 29.Bahia | 9.280 | 1.175 | 12,7 | 9 | 0,1 | 0 | 0,0 |
| 31.Minas Gerais | 14.279 | 1.070 | 7,5 | 7 | 0,0 | 0 | 0,0 |
| 32.Espírito Santo | 7.198 | 1.791 | 24,9 | 25 | 0,3 | 0 | 0,0 |
| 33.Rio de Janeiro | 14.731 | 1.469 | 10,0 | 8 | 0,1 | 0 | 0,0 |
| 35.São Paulo | 15.238 | 636 | 4,2 | 14 | 0,1 | 7 | <0,1 |
| 41.Paraná | 10.324 | 1.333 | 12,9 | 10 | 0,1 | 0 | 0,0 |
| 42.Santa Catarina | 11.984 | 1.389 | 11,6 | 29 | 0,2 | 1 | <0,1 |
| 43.Rio Grande do Sul | 11.010 | 1.305 | 11,9 | 7 | 0,1 | 0 | 0,0 |
| 50.Mato Grosso do Sul | 4.135 | 1.592 | 38,5 | 36 | 0,9 | 4 | 0,1 |
| 51.Mato Grosso | 4.724 | 1.486 | 31,5 | 18 | 0,4 | 4 | 0,1 |
| 52.Goiás | 6.575 | 1.471 | 22,4 | 13 | 0,2 | 0 | 0,0 |
| 53.Distrito Federal | 3.367 | 1.276 | 37,9 | 10 | 0,3 | 1 | <0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A16: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por Unidades da Federação divulgadas, segundo o cenário de revelação 1.

| Unidades da Federação | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **187.061** | **65.004** | **34,8** | **4.703** | **2,5** | **992** | **0,5** |
| 11.Rondônia | 3.115 | 2.061 | 66,2 | 163 | 5,2 | 17 | 0,5 |
| 12.Acre | 3.343 | 2.881 | 86,2 | 740 | 22,1 | 179 | 5,4 |
| 13.Amazonas | 4.414 | 2.459 | 55,7 | 284 | 6,4 | 88 | 2,0 |
| 14.Roraima | 1.870 | 1.823 | 97,5 | 787 | 42,1 | 239 | 12,8 |
| 15.Pará | 6.070 | 2.548 | 42,0 | 74 | 1,2 | 12 | 0,2 |
| 16.Amapá | 1.342 | 1.242 | 92,5 | 304 | 22,7 | 92 | 6,9 |
| 17.Tocantins | 2.757 | 1.929 | 70,0 | 169 | 6,1 | 27 | 1,0 |
| 21.Maranhão | 11.083 | 3.992 | 36,0 | 279 | 2,5 | 58 | 0,5 |
| 22.Piauí | 4.115 | 2.199 | 53,4 | 125 | 3,0 | 19 | 0,5 |
| 23.Ceará | 9.272 | 2.828 | 30,5 | 79 | 0,9 | 11 | 0,1 |
| 24.Rio Grande do Norte | 3.703 | 2.142 | 57,8 | 95 | 2,6 | 12 | 0,3 |
| 25.Paraíba | 4.796 | 2.408 | 50,2 | 99 | 2,1 | 19 | 0,4 |
| 26.Pernambuco | 7.642 | 2.514 | 32,9 | 62 | 0,8 | 8 | 0,1 |
| 27.Alagoas | 7.086 | 3.141 | 44,3 | 325 | 4,6 | 56 | 0,8 |
| 28.Sergipe | 3.608 | 2.082 | 57,7 | 124 | 3,4 | 21 | 0,6 |
| 29.Bahia | 9.280 | 2.356 | 25,4 | 70 | 0,8 | 6 | 0,1 |
| 31.Minas Gerais | 14.279 | 2.361 | 16,5 | 59 | 0,4 | 8 | 0,1 |
| 32.Espírito Santo | 7.198 | 2.946 | 40,9 | 148 | 2,1 | 20 | 0,3 |
| 33.Rio de Janeiro | 14.731 | 3.110 | 21,1 | 54 | 0,4 | 13 | 0,1 |
| 35.São Paulo | 15.238 | 1.445 | 9,5 | 39 | 0,3 | 14 | 0,1 |
| 41.Paraná | 10.324 | 2.485 | 24,1 | 68 | 0,7 | 5 | <0,1 |
| 42.Santa Catarina | 11.984 | 2.536 | 21,2 | 141 | 1,2 | 17 | 0,1 |
| 43.Rio Grande do Sul | 11.010 | 2.254 | 20,5 | 33 | 0,3 | 4 | 0,0 |
| 50.Mato Grosso do Sul | 4.135 | 2.262 | 54,7 | 118 | 2,9 | 15 | 0,4 |
| 51.Mato Grosso | 4.724 | 2.483 | 52,6 | 122 | 2,6 | 17 | 0,4 |
| 52.Goiás | 6.575 | 2.457 | 37,4 | 65 | 1,0 | 4 | 0,1 |
| 53.Distrito Federal | 3.367 | 2.060 | 61,2 | 77 | 2,3 | 11 | 0,3 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A17: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por RM ou RIDE divulgadas, segundo o cenário de revelação 1.

| RM ou RIDE | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **59.691** | **20.963** | **35,1** | **545** | **0,9** | **70** | **0,1** |
| 13.RM Manaus | 2.599 | 992 | 38,2 | 23 | 0,9 | 1 | <0,1 |
| 15.RM Belém | 1.818 | 953 | 52,4 | 4 | 0,2 | 0 | 0,0 |
| 16.RM Macapá | 1.045 | 757 | 72,4 | 124 | 11,9 | 32 | 3,1 |
| 21.RM Grande São Luís | 2.043 | 1.015 | 49,7 | 23 | 1,1 | 1 | <0,1 |
| 22.RIDE Grande Teresina | 1.817 | 1.245 | 68,5 | 130 | 7,2 | 13 | 0,7 |
| 23.RM Fortaleza | 3.139 | 1.178 | 37,5 | 6 | 0,2 | 0 | 0,0 |
| 24.RM Natal | 1.544 | 946 | 61,3 | 18 | 1,2 | 1 | 0,1 |
| 25.RM João Pessoa | 1.737 | 1.048 | 60,3 | 28 | 1,6 | 0 | 0,0 |
| 26.RM Recife | 2.813 | 1.029 | 36,6 | 8 | 0,3 | 2 | 0,1 |
| 27.RM Maceió | 2.483 | 1.078 | 43,4 | 15 | 0,6 | 0 | 0,0 |
| 28.RM Aracaju | 1.306 | 780 | 59,7 | 16 | 1,2 | 0 | 0,0 |
| 29.RM Salvador | 2.069 | 786 | 38,0 | 14 | 0,7 | 0 | 0,0 |
| 31.RM Belo Horizonte | 3.519 | 988 | 28,1 | 19 | 0,5 | 1 | <0,1 |
| 32.RM Grande Vitória | 3.412 | 1.458 | 42,7 | 26 | 0,8 | 3 | 0,1 |
| 33.RM Rio de Janeiro | 9.851 | 1.506 | 15,3 | 3 | 0,0 | 0 | 0,0 |
| 35.RM São Paulo | 6.416 | 425 | 6,6 | 17 | 0,3 | 12 | 0,2 |
| 41.RM Curitiba | 3.006 | 1.037 | 34,5 | 4 | 0,1 | 0 | 0,0 |
| 42.RM Florianópolis | 1.683 | 732 | 43,5 | 22 | 1,3 | 3 | 0,2 |
| 43.RM Porto Alegre | 3.780 | 1.057 | 28,0 | 2 | 0,1 | 0 | 0,0 |
| 51.RM Vale do Rio Cuiabá | 1.248 | 860 | 68,9 | 37 | 3,0 | 1 | 0,1 |
| 52.RM Goiânia | 2.363 | 1.093 | 46,3 | 6 | 0,3 | 0 | 0,0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A18: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por RM ou RIDE divulgadas, segundo o cenário de revelação 2.

| RM ou RIDE | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **59.691** | **31.936** | **53,5** | **1.699** | **2,8** | **285** | **0,5** |
| 13.RM Manaus | 2.599 | 1.662 | 63,9 | 105 | 4,0 | 16 | 0,6 |
| 15.RM Belém | 1.818 | 1.398 | 76,9 | 41 | 2,3 | 0 | 0,0 |
| 16.RM Macapá | 1.045 | 1.007 | 96,4 | 300 | 28,7 | 94 | 9,0 |
| 21.RM Grande São Luís | 2.043 | 1.605 | 78,6 | 117 | 5,7 | 20 | 1,0 |
| 22.RIDE Grande Teresina | 1.817 | 1.586 | 87,3 | 247 | 13,6 | 40 | 2,2 |
| 23.RM Fortaleza | 3.139 | 1.896 | 60,4 | 33 | 1,1 | 3 | 0,1 |
| 24.RM Natal | 1.544 | 1.296 | 83,9 | 73 | 4,7 | 4 | 0,3 |
| 25.RM João Pessoa | 1.737 | 1.472 | 84,7 | 96 | 5,5 | 7 | 0,4 |
| 26.RM Recife | 2.813 | 1.665 | 59,2 | 31 | 1,1 | 8 | 0,3 |
| 27.RM Maceió | 2.483 | 1.611 | 64,9 | 100 | 4,0 | 13 | 0,5 |
| 28.RM Aracaju | 1.306 | 1.098 | 84,1 | 75 | 5,7 | 9 | 0,7 |
| 29.RM Salvador | 2.069 | 1.160 | 56,1 | 37 | 1,8 | 5 | 0,2 |
| 31.RM Belo Horizonte | 3.519 | 1.708 | 48,5 | 53 | 1,5 | 10 | 0,3 |
| 32.RM Grande Vitória | 3.412 | 2.155 | 63,2 | 107 | 3,1 | 12 | 0,4 |
| 33.RM Rio de Janeiro | 9.851 | 2.814 | 28,6 | 17 | 0,2 | 2 | 0,0 |
| 35.RM São Paulo | 6.416 | 955 | 14,9 | 34 | 0,5 | 17 | 0,3 |
| 41.RM Curitiba | 3.006 | 1.504 | 50,0 | 30 | 1,0 | 0 | 0,0 |
| 42.RM Florianópolis | 1.683 | 1.030 | 61,2 | 58 | 3,4 | 9 | 0,5 |
| 43.RM Porto Alegre | 3.780 | 1.622 | 42,9 | 16 | 0,4 | 0 | 0,0 |
| 51.RM Vale do Rio Cuiabá | 1.248 | 1.129 | 90,5 | 104 | 8,3 | 14 | 1,1 |
| 52.RM Goiânia | 2.363 | 1.563 | 66,1 | 25 | 1,1 | 2 | 0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A19: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, fora de RM ou RIDE que podem ser deduzidas pelo intruso, por sua UF, segundo o cenário de revelação 1.

| Municípios fora da RM/RIDE na UF | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **108.783** | **25.906** | **23,8** | **632** | **0,6** | **61** | **0,1** |
| 13.Amazonas | 1.815 | 1.026 | 56,5 | 124 | 6,8 | 32 | 1,8 |
| 15.Pará | 4.252 | 1.186 | 27,9 | 14 | 0,3 | 1 | <0,1 |
| 16.Amapá | 297 | 264 | 88,9 | 48 | 16,2 | 6 | 2,0 |
| 21.Maranhão | 8.753 | 1.823 | 20,8 | 37 | 0,4 | 4 | <0,1 |
| 22.Piauí | 2.585 | 1.127 | 43,6 | 17 | 0,7 | 0 | 0,0 |
| 23.Ceará | 6.133 | 1.329 | 21,7 | 5 | 0,1 | 0 | 0,0 |
| 24.Rio Grande do Norte | 2.159 | 1.149 | 53,2 | 23 | 1,1 | 0 | 0,0 |
| 25.Paraíba | 3.059 | 1.233 | 40,3 | 24 | 0,8 | 0 | 0,0 |
| 26.Pernambuco | 4.829 | 1.219 | 25,2 | 8 | 0,2 | 0 | 0,0 |
| 27.Alagoas | 4.603 | 1.818 | 39,5 | 99 | 2,2 | 8 | 0,2 |
| 28.Sergipe | 2.302 | 1.220 | 53,0 | 30 | 1,3 | 0 | 0,0 |
| 29.Bahia | 7.211 | 1.159 | 16,1 | 8 | 0,1 | 0 | 0,0 |
| 31.Minas Gerais | 10.760 | 1.144 | 10,6 | 9 | 0,1 | 0 | 0,0 |
| 32.Espírito Santo | 3.786 | 1.487 | 39,3 | 38 | 1,0 | 0 | 0,0 |
| 33.Rio de Janeiro | 4.880 | 1.494 | 30,6 | 42 | 0,9 | 3 | 0,1 |
| 35.São Paulo | 8.822 | 739 | 8,4 | 9 | 0,1 | 0 | 0,0 |
| 41.Paraná | 7.318 | 1.310 | 17,9 | 12 | 0,2 | 1 | <0,1 |
| 42.Santa Catarina | 10.301 | 1.313 | 12,7 | 28 | 0,3 | 2 | <0,1 |
| 43.Rio Grande do Sul | 7.230 | 1.139 | 15,8 | 15 | 0,2 | 0 | 0,0 |
| 51.Mato Grosso | 3.476 | 1.364 | 39,2 | 15 | 0,4 | 4 | 0,1 |
| 52.Goiás | 4.212 | 1.363 | 32,4 | 27 | 0,6 | 0 | 0,0 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A20: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, fora de RM ou RIDE que podem ser deduzidas pelo intruso, por sua UF, segundo o cenário de revelação 2.

| Municípios fora da RM/RIDE na UF | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **108.783** | **41.621** | **38,3** | **2.430** | **2,2** | **422** | **0,4** |
| 13.Amazonas | 1.815 | 1.476 | 81,3 | 304 | 16,7 | 95 | 5,2 |
| 15.Pará | 4.252 | 1.987 | 46,7 | 65 | 1,5 | 13 | 0,3 |
| 16.Amapá | 297 | 295 | 99,3 | 74 | 24,9 | 14 | 4,7 |
| 21.Maranhão | 8.753 | 3.535 | 40,4 | 267 | 3,1 | 52 | 0,6 |
| 22.Piauí | 2.585 | 1.605 | 62,1 | 82 | 3,2 | 11 | 0,4 |
| 23.Ceará | 6.133 | 2.373 | 38,7 | 84 | 1,4 | 12 | 0,2 |
| 24.Rio Grande do Norte | 2.159 | 1.522 | 70,5 | 71 | 3,3 | 9 | 0,4 |
| 25.Paraíba | 3.059 | 1.771 | 57,9 | 80 | 2,6 | 18 | 0,6 |
| 26.Pernambuco | 4.829 | 2.007 | 41,6 | 66 | 1,4 | 9 | 0,2 |
| 27.Alagoas | 4.603 | 2.736 | 59,4 | 360 | 7,8 | 63 | 1,4 |
| 28.Sergipe | 2.302 | 1.529 | 66,4 | 119 | 5,2 | 16 | 0,7 |
| 29.Bahia | 7.211 | 2.161 | 30,0 | 61 | 0,8 | 7 | 0,1 |
| 31.Minas Gerais | 10.760 | 2.239 | 20,8 | 57 | 0,5 | 6 | 0,1 |
| 32.Espírito Santo | 3.786 | 2.158 | 57,0 | 163 | 4,3 | 22 | 0,6 |
| 33.Rio de Janeiro | 4.880 | 2.304 | 47,2 | 105 | 2,2 | 23 | 0,5 |
| 35.São Paulo | 8.822 | 1.436 | 16,3 | 48 | 0,5 | 4 | 0,0 |
| 41.Paraná | 7.318 | 2.192 | 30,0 | 76 | 1,0 | 10 | 0,1 |
| 42.Santa Catarina | 10.301 | 2.359 | 22,9 | 131 | 1,3 | 16 | 0,2 |
| 43.Rio Grande do Sul | 7.230 | 1.882 | 26,0 | 42 | 0,6 | 5 | 0,1 |
| 51.Mato Grosso | 3.476 | 2.058 | 59,2 | 91 | 2,6 | 12 | 0,3 |
| 52.Goiás | 4.212 | 1.996 | 47,4 | 84 | 2,0 | 5 | 0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A21: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por municípios da capital divulgados, segundo o cenário de revelação 1.

| Capitais | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **43.994** | **21.149** | **48,1** | **920** | **2,1** | **95** | **0,2** |
| 11.Porto Velho (RO) | 951 | 696 | 73,2 | 60 | 6,3 | 0 | 0,0 |
| 12.Rio Branco (AC) | 1.720 | 1.316 | 76,5 | 255 | 14,8 | 29 | 1,7 |
| 13.Manaus (AM) | 2.147 | 880 | 41,0 | 7 | 0,3 | 0 | 0,0 |
| 14.Boa Vista (RR) | 1.173 | 967 | 82,4 | 242 | 20,6 | 35 | 3,0 |
| 15.Belém (PA) | 1.237 | 781 | 63,1 | 2 | 0,2 | 0 | 0,0 |
| 16.Macapá (AP) | 743 | 564 | 75,9 | 70 | 9,4 | 8 | 1,1 |
| 17.Palmas (TO) | 586 | 445 | 75,9 | 35 | 6,0 | 0 | 0,0 |
| 21.São Luís (MA) | 1.482 | 822 | 55,5 | 13 | 0,9 | 0 | 0,0 |
| 22.Teresina (PI) | 1.058 | 737 | 69,7 | 18 | 1,7 | 0 | 0,0 |
| 23.Fortaleza (CE) | 2.108 | 1.001 | 47,5 | 2 | 0,1 | 0 | 0,0 |
| 24.Natal (RN) | 966 | 682 | 70,6 | 7 | 0,7 | 0 | 0,0 |
| 25.João Pessoa (PB) | 1.146 | 743 | 64,8 | 15 | 1,3 | 0 | 0,0 |
| 26.Recife (PE) | 1.170 | 677 | 57,9 | 1 | 0,1 | 0 | 0,0 |
| 27.Maceió (AL) | 2.035 | 964 | 47,4 | 14 | 0,7 | 0 | 0,0 |
| 28.Aracaju (SE) | 856 | 582 | 68,0 | 8 | 0,9 | 1 | 0,1 |
| 29.Salvador (BA) | 1.367 | 575 | 42,1 | 3 | 0,2 | 0 | 0,0 |
| 31.Belo Horizonte (MG) | 1.961 | 825 | 42,1 | 6 | 0,3 | 0 | 0,0 |
| 32.Vitória (ES) | 949 | 676 | 71,2 | 46 | 4,8 | 1 | 0,1 |
| 33.Rio de Janeiro (RJ) | 5.129 | 1.242 | 24,2 | 0 | 0,0 | 0 | 0,0 |
| 35.São Paulo (SP) | 3.631 | 382 | 10,5 | 21 | 0,6 | 14 | 0,4 |
| 41.Curitiba (PR) | 1.758 | 742 | 42,2 | 3 | 0,2 | 0 | 0,0 |
| 42.Florianópolis (SC) | 968 | 484 | 50,0 | 25 | 2,6 | 4 | 0,4 |
| 43.Porto Alegre (RS) | 1.742 | 720 | 41,3 | 2 | 0,1 | 0 | 0,0 |
| 50.Campo Grande (MS) | 1.322 | 828 | 62,6 | 22 | 1,7 | 0 | 0,0 |
| 51.Cuiabá (MT) | 814 | 639 | 78,5 | 25 | 3,1 | 1 | 0,1 |
| 52.Goiânia (GO) | 1.608 | 903 | 56,2 | 8 | 0,5 | 1 | 0,1 |
| 53.Brasília (DF) | 3.367 | 1.276 | 37,9 | 10 | 0,3 | 1 | <0,1 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A22: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, por municípios da capital divulgados, segundo o cenário de revelação 2.

| Capitais | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **43.994** | **30.367** | **69,0** | **2.669** | **6,1** | **452** | **1,0** |
| 11.Porto Velho (RO) | 951 | 896 | 94,2 | 165 | 17,4 | 21 | 2,2 |
| 12.Rio Branco (AC) | 1.720 | 1.672 | 97,2 | 574 | 33,4 | 131 | 7,6 |
| 13.Manaus (AM) | 2.147 | 1.455 | 67,8 | 70 | 3,3 | 7 | 0,3 |
| 14.Boa Vista (RR) | 1.173 | 1.160 | 98,9 | 520 | 44,3 | 137 | 11,7 |
| 15.Belém (PA) | 1.237 | 1.057 | 85,4 | 35 | 2,8 | 0 | 0,0 |
| 16.Macapá (AP) | 743 | 725 | 97,6 | 192 | 25,8 | 39 | 5,2 |
| 17.Palmas (TO) | 586 | 563 | 96,1 | 97 | 16,6 | 11 | 1,9 |
| 21.São Luís (MA) | 1.482 | 1.286 | 86,8 | 96 | 6,5 | 10 | 0,7 |
| 22.Teresina (PI) | 1.058 | 948 | 89,6 | 69 | 6,5 | 5 | 0,5 |
| 23.Fortaleza (CE) | 2.108 | 1.467 | 69,6 | 22 | 1,0 | 1 | <0,1 |
| 24.Natal (RN) | 966 | 874 | 90,5 | 57 | 5,9 | 2 | 0,2 |
| 25.João Pessoa (PB) | 1.146 | 1.049 | 91,5 | 66 | 5,8 | 1 | 0,1 |
| 26.Recife (PE) | 1.170 | 957 | 81,8 | 11 | 0,9 | 0 | 0,0 |
| 27.Maceió (AL) | 2.035 | 1.424 | 70,0 | 88 | 4,3 | 11 | 0,5 |
| 28.Aracaju (SE) | 856 | 783 | 91,5 | 57 | 6,7 | 7 | 0,8 |
| 29.Salvador (BA) | 1.367 | 847 | 62,0 | 8 | 0,6 | 0 | 0,0 |
| 31.Belo Horizonte (MG) | 1.961 | 1.307 | 66,6 | 31 | 1,6 | 6 | 0,3 |
| 32.Vitória (ES) | 949 | 883 | 93,0 | 135 | 14,2 | 11 | 1,2 |
| 33.Rio de Janeiro (RJ) | 5.129 | 2.147 | 41,9 | 10 | 0,2 | 0 | 0,0 |
| 35.São Paulo (SP) | 3.631 | 793 | 21,8 | 32 | 0,9 | 20 | 0,6 |
| 41.Curitiba (PR) | 1.758 | 1.056 | 60,1 | 22 | 1,3 | 0 | 0,0 |
| 42.Florianópolis (SC) | 968 | 726 | 75,0 | 47 | 4,9 | 9 | 0,9 |
| 43.Porto Alegre (RS) | 1.742 | 1.050 | 60,3 | 17 | 1,0 | 0 | 0,0 |
| 50.Campo Grande (MS) | 1.322 | 1.147 | 86,8 | 68 | 5,1 | 2 | 0,2 |
| 51.Cuiabá (MT) | 814 | 766 | 94,1 | 78 | 9,6 | 8 | 1,0 |
| 52.Goiânia (GO) | 1.608 | 1.269 | 78,9 | 25 | 1,6 | 2 | 0,1 |
| 53.Brasília (DF) | 3.367 | 2.060 | 61,2 | 77 | 2,3 | 11 | 0,3 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A23: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, fora de municípios da capital que podem ser deduzidos pelo intruso, por sua UF, segundo o cenário de revelação 1.

| Municípios fora da Capital na UF | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **143.067** | **32.937** | **23,0** | **1.431** | **1,0** | **274** | **0,2** |
| 11.Rondônia | 2.164 | 1.245 | 57,5 | 49 | 2,3 | 1 | <0,1 |
| 12.Acre | 1.623 | 1.297 | 79,9 | 272 | 16,8 | 57 | 3,5 |
| 13.Amazonas | 2.267 | 1.075 | 47,4 | 110 | 4,9 | 31 | 1,4 |
| 14.Roraima | 697 | 674 | 96,7 | 350 | 50,2 | 115 | 16,5 |
| 15.Pará | 4.833 | 1.247 | 25,8 | 15 | 0,3 | 1 | <0,1 |
| 16.Amapá | 599 | 507 | 84,6 | 124 | 20,7 | 36 | 6,0 |
| 17.Tocantins | 2.171 | 1.215 | 56,0 | 56 | 2,6 | 9 | 0,4 |
| 21.Maranhão | 9.601 | 1.876 | 19,5 | 37 | 0,4 | 3 | <0,1 |
| 22.Piauí | 3.057 | 1.216 | 39,8 | 28 | 0,9 | 1 | <0,1 |
| 23.Ceará | 7.164 | 1.378 | 19,2 | 6 | 0,1 | 0 | 0,0 |
| 24.Rio Grande do Norte | 2.737 | 1.268 | 46,3 | 25 | 0,9 | 2 | 0,1 |
| 25.Paraíba | 3.650 | 1.323 | 36,2 | 26 | 0,7 | 0 | 0,0 |
| 26.Pernambuco | 6.472 | 1.307 | 20,2 | 8 | 0,1 | 0 | 0,0 |
| 27.Alagoas | 5.051 | 1.849 | 36,6 | 90 | 1,8 | 7 | 0,1 |
| 28.Sergipe | 2.752 | 1.308 | 47,5 | 31 | 1,1 | 0 | 0,0 |
| 29.Bahia | 7.913 | 1.187 | 15,0 | 9 | 0,1 | 0 | 0,0 |
| 31.Minas Gerais | 12.318 | 1.060 | 8,6 | 9 | 0,1 | 0 | 0,0 |
| 32.Espírito Santo | 6.249 | 1.700 | 27,2 | 28 | 0,4 | 0 | 0,0 |
| 33.Rio de Janeiro | 9.602 | 1.503 | 15,7 | 15 | 0,2 | 0 | 0,0 |
| 35.São Paulo | 11.607 | 679 | 5,8 | 7 | 0,1 | 0 | 0,0 |
| 41.Paraná | 8.566 | 1.291 | 15,1 | 11 | 0,1 | 0 | 0,0 |
| 42.Santa Catarina | 11.016 | 1.350 | 12,3 | 28 | 0,3 | 1 | <0,1 |
| 43.Rio Grande do Sul | 9.268 | 1.219 | 13,2 | 11 | 0,1 | 0 | 0,0 |
| 50.Mato Grosso do Sul | 2.813 | 1.372 | 48,8 | 46 | 1,6 | 6 | 0,2 |
| 51.Mato Grosso | 3.910 | 1.406 | 36,0 | 21 | 0,5 | 4 | 0,1 |
| 52.Goiás | 4.967 | 1.385 | 27,9 | 19 | 0,4 | 0 | 0,0 |
| 53.Distrito Federal | 0 | 0 | -- | 0 | -- | 0 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A24: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, dado um limiar de risco de revelação, fora de municípios da capital que podem ser deduzidos pelo intruso, por sua UF, segundo o cenário de revelação 2.

| Municípios fora da Capital na UF | Registros de domicílio | Registros de domicílios não seguros por limiar | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1% | | 10% | | 20% | |
| | | Total | % | Total | % | Total | % |
| **Total** | **143.067** | **52.953** | **37,0** | **3.988** | **2,8** | **881** | **0,6** |
| 11.Rondônia | 2.164 | 1.571 | 72,6 | 120 | 5,5 | 8 | 0,4 |
| 12.Acre | 1.623 | 1.570 | 96,7 | 544 | 33,5 | 145 | 8,9 |
| 13.Amazonas | 2.267 | 1.714 | 75,6 | 305 | 13,5 | 92 | 4,1 |
| 14.Roraima | 697 | 692 | 99,3 | 486 | 69,7 | 202 | 29,0 |
| 15.Pará | 4.833 | 2.170 | 44,9 | 65 | 1,3 | 13 | 0,3 |
| 16.Amapá | 599 | 590 | 98,5 | 222 | 37,1 | 86 | 14,4 |
| 17.Tocantins | 2.171 | 1.608 | 74,1 | 160 | 7,4 | 26 | 1,2 |
| 21.Maranhão | 9.601 | 3.718 | 38,7 | 271 | 2,8 | 57 | 0,6 |
| 22.Piauí | 3.057 | 1.784 | 58,4 | 114 | 3,7 | 16 | 0,5 |
| 23.Ceará | 7.164 | 2.537 | 35,4 | 88 | 1,2 | 11 | 0,2 |
| 24.Rio Grande do Norte | 2.737 | 1.776 | 64,9 | 74 | 2,7 | 11 | 0,4 |
| 25.Paraíba | 3.650 | 1.981 | 54,3 | 92 | 2,5 | 19 | 0,5 |
| 26.Pernambuco | 6.472 | 2.287 | 35,3 | 63 | 1,0 | 9 | 0,1 |
| 27.Alagoas | 5.051 | 2.812 | 55,7 | 342 | 6,8 | 60 | 1,2 |
| 28.Sergipe | 2.752 | 1.738 | 63,2 | 118 | 4,3 | 19 | 0,7 |
| 29.Bahia | 7.913 | 2.235 | 28,2 | 69 | 0,9 | 7 | 0,1 |
| 31.Minas Gerais | 12.318 | 2.223 | 18,0 | 57 | 0,5 | 5 | 0,0 |
| 32.Espírito Santo | 6.249 | 2.708 | 43,3 | 143 | 2,3 | 16 | 0,3 |
| 33.Rio de Janeiro | 9.602 | 2.827 | 29,4 | 68 | 0,7 | 17 | 0,2 |
| 35.São Paulo | 11.607 | 1.428 | 12,3 | 42 | 0,4 | 3 | <0,1 |
| 41.Paraná | 8.566 | 2.287 | 26,7 | 70 | 0,8 | 6 | 0,1 |
| 42.Santa Catarina | 11.016 | 2.456 | 22,3 | 134 | 1,2 | 14 | 0,1 |
| 43.Rio Grande do Sul | 9.268 | 2.056 | 22,2 | 30 | 0,3 | 4 | 0,0 |
| 50.Mato Grosso do Sul | 2.813 | 1.820 | 64,7 | 130 | 4,6 | 15 | 0,5 |
| 51.Mato Grosso | 3.910 | 2.232 | 57,1 | 109 | 2,8 | 16 | 0,4 |
| 52.Goiás | 4.967 | 2.133 | 42,9 | 72 | 1,4 | 4 | 0,1 |
| 53.Distrito Federal | 0 | 0 | -- | 0 | -- | 0 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A25: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, em RM ou RIDE divulgada, dado um limiar de risco de revelação, segundo os cenários de revelação.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | |
|---|---|---|---|---|---|---|
| | | | Cenário 1 | | Cenário 2 | |
| | Total | % | Total | % | Total | % |
| **Total** | **59.691** | **100,0** | **70** | **100,0** | **285** | **100,0** |
| 1 | 10.556 | 17,7 | 4 | 5,7 | 6 | 2,1 |
| 2 | 16.598 | 27,8 | 10 | 14,3 | 15 | 5,3 |
| 3 | 14.925 | 25,0 | 9 | 12,9 | 9 | 3,2 |
| 4 | 10.814 | 18,1 | 10 | 14,3 | 22 | 7,7 |
| 5 | 4.215 | 7,1 | 9 | 12,9 | 26 | 9,1 |
| 6 | 1.497 | 2,5 | 13 | 18,6 | 45 | 15,8 |
| 7 | 613 | 1,0 | 4 | 5,7 | 43 | 15,1 |
| 8 ou mais | 473 | 0,8 | 11 | 15,7 | 119 | 41,8 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A26: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, fora de RM ou RIDE que pode ser deduzido pelo intruso, dado um limiar de risco de revelação, segundo os cenários de revelação.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | |
|---|---|---|---|---|---|---|
| | | | Cenário 1 | | Cenário 2 | |
| | Total | % | Total | % | Total | % |
| **Total** | **108.783** | **100,0** | **61** | **100,0** | **422** | **100,0** |
| 1 | 16.224 | 14,9 | 0 | 0,0 | 1 | 0,2 |
| 2 | 29.871 | 27,5 | 5 | 8,2 | 6 | 1,4 |
| 3 | 27.589 | 25,4 | 4 | 6,6 | 6 | 1,4 |
| 4 | 20.547 | 18,9 | 9 | 14,8 | 16 | 3,8 |
| 5 | 8.815 | 8,1 | 4 | 6,6 | 39 | 9,2 |
| 6 | 3.333 | 3,1 | 10 | 16,4 | 54 | 12,8 |
| 7 | 1.361 | 1,3 | 5 | 8,2 | 67 | 15,9 |
| 8 ou mais | 1.043 | 1,0 | 24 | 39,3 | 233 | 55,2 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A27: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, em municípios de capital divulgados, dado um limiar de risco de revelação, segundo os cenários de revelação.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | |
|---|---|---|---|---|---|---|
| | | | Cenário 1 | | Cenário 2 | |
| | Total | % | Total | % | Total | % |
| **Total** | **43.994** | **100,0** | **95** | **100,0** | **452** | **100,0** |
| 1 | 7.891 | 17,9 | 2 | 2,1 | 2 | 0,4 |
| 2 | 11.980 | 27,2 | 7 | 7,4 | 9 | 2,0 |
| 3 | 10.860 | 24,7 | 8 | 8,4 | 19 | 4,2 |
| 4 | 8.030 | 18,3 | 20 | 21,1 | 46 | 10,2 |
| 5 | 3.239 | 7,4 | 12 | 12,6 | 66 | 14,6 |
| 6 | 1.140 | 2,6 | 16 | 16,8 | 91 | 20,1 |
| 7 | 466 | 1,1 | 10 | 10,5 | 69 | 15,3 |
| 8 ou mais | 388 | 0,9 | 20 | 21,1 | 150 | 33,2 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A28: Registros de domicílios não seguros da PNAD Contínua 2019 – 2º trimestre, fora de municípios de capital que pode ser deduzido pelo intruso, dado um limiar de risco de revelação, segundo os cenários de revelação.

| Tamanho do domicílio | Registros de domicílios | | Registros de domicílios não seguros (limiar de 20%) | | | |
|---|---|---|---|---|---|---|
| | | | Cenário 1 | | Cenário 2 | |
| | Total | % | Total | % | Total | % |
| **Total** | **143.067** | **100,0** | **274** | **100,0** | **881** | **100,0** |
| 1 | 21.616 | 15,1 | 6 | 2,2 | 6 | 0,7 |
| 2 | 39.254 | 27,4 | 33 | 12,0 | 34 | 3,9 |
| 3 | 36.183 | 25,3 | 30 | 10,9 | 50 | 5,7 |
| 4 | 26.993 | 18,9 | 51 | 18,6 | 89 | 10,1 |
| 5 | 11.527 | 8,1 | 41 | 15,0 | 103 | 11,7 |
| 6 | 4.333 | 3,0 | 33 | 12,0 | 118 | 13,4 |
| 7 | 1.796 | 1,3 | 31 | 11,3 | 129 | 14,6 |
| 8 ou mais | 1.365 | 1,0 | 49 | 17,9 | 352 | 39,9 |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre

Tabela A29: Frequência da variável V2001 na amostra da PNAD Contínua 2019 – 2º trimestre.

| V2001 – Número de pessoas no domicílio | Frequência na amostra | |
|---|---|---|
| | Domicílios | Pessoas |
| Total | 187.061 | 551.348 |
| 1 | 29.507 | 29.507 |
| 2 | 51.234 | 102.468 |
| 3 | 47.043 | 141.129 |
| 4 | 35.023 | 140.092 |
| 5 | 14.766 | 73.830 |
| 6 | 5.473 | 32.838 |
| 7 | 2.262 | 15.834 |
| 8 | 927 | 7.416 |
| 9 | 442 | 3.978 |
| 10 | 188 | 1.880 |
| 11 | 90 | 990 |
| 12 | 57 | 684 |
| 13 | 25 | 325 |
| 14 | 8 | 112 |
| 15 | 7 | 105 |
| 16 | 3 | 48 |
| 17 | 3 | 51 |
| 19 | 1 | 19 |
| 20 | 1 | 20 |
| 22 | 1 | 22 |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A30: Frequência da variável V2007 na amostra da PNAD Contínua 2019 – 2º trimestre.

| V2007 - Sexo | Frequência na amostra | |
|---|---|---|
| | Absoluta | Percentual |
| Total | 551.348 | 100,0 |
| 1 - Masculino | 268.931 | 48,8 |
| 2 - Feminino | 282.417 | 51,2 |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A31: Frequência da variável V2010 na amostra da PNAD Contínua 2019 – 2º trimestre.

| V2010 – Cor ou raça | Frequência na amostra | |
|---|---|---|
| | Absoluta | Percentual |
| Total | 551.348 | 100,0 |
| 1 - Branca | 212.957 | 38,6 |
| 2 - Preta | 46.805 | 8,5 |
| 3 - Amarela | 2.796 | 0,5 |
| 4 - Parda | 285.892 | 51,9 |
| 5 - Indígena | 2.859 | 0,5 |
| 9 - Ignorado | 39 | <0,1 |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A32: Frequência da variável VD3004 na amostra da PNAD Contínua 2019 – 2º trimestre.

| VD3004 – Nível de instrução (variável derivada) | Frequência na amostra | |
|---|---|---|
| | Absoluta | Percentual |
| Total | 551.348 | 100,0 |
| 1 - Sem instrução e menos de 1 ano de estudo | 45.753 | 8,3 |
| 2 - Fundamental incompleto ou equivalente | 204.153 | 37,0 |
| 3 - Fundamental completo ou equivalente | 40.672 | 7,4 |
| 4 - Médio incompleto ou equivalente | 35.512 | 6,4 |
| 5 - Médio completo ou equivalente | 114.825 | 20,8 |
| 6 - Superior incompleto ou equivalente | 21.504 | 3,9 |
| 7 - Superior completo | 54.223 | 9,8 |
| 99 - Não aplicável | 34.706 | 6,3 |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A33: Frequência da variável V2009 na amostra da PNAD Contínua 2019 – 2º trimestre.

| V2009 – Idade | Freq. na amostra | V2009 – Idade | Freq. na amostra | V2009 – Idade | Freq. na amostra | V2009 – Idade | Freq. na amostra |
|---|---|---|---|---|---|---|---|
| Total | 551.348 | 28 | 7.604 | 57 | 6.091 | 86 | 799 |
| 0 | 6.137 | 29 | 7.551 | 58 | 6.205 | 87 | 710 |
| 1 | 7.108 | 30 | 8.272 | 59 | 6.022 | 88 | 626 |
| 2 | 6.716 | 31 | 7.702 | 60 | 6.130 | 89 | 550 |
| 3 | 7.334 | 32 | 8.151 | 61 | 5.574 | 90 | 472 |
| 4 | 7.411 | 33 | 8.385 | 62 | 5.299 | 91 | 367 |
| 5 | 7.293 | 34 | 8.225 | 63 | 5.138 | 92 | 287 |
| 6 | 7.265 | 35 | 8.486 | 64 | 4.988 | 93 | 247 |
| 7 | 7.405 | 36 | 8.607 | 65 | 4.999 | 94 | 179 |
| 8 | 7.833 | 37 | 8.736 | 66 | 4.440 | 95 | 133 |
| 9 | 7.813 | 38 | 8.453 | 67 | 4.161 | 96 | 113 |
| 10 | 8.149 | 39 | 8.292 | 68 | 4.135 | 97 | 72 |
| 11 | 7.983 | 40 | 8.845 | 69 | 3.708 | 98 | 62 |
| 12 | 8.259 | 41 | 7.745 | 70 | 3.705 | 99 | 55 |
| 13 | 8.544 | 42 | 7.877 | 71 | 3.300 | 100 | 34 |
| 14 | 8.632 | 43 | 7.614 | 72 | 3.087 | 101 | 22 |
| 15 | 8.664 | 44 | 7.376 | 73 | 2.822 | 102 | 20 |
| 16 | 8.755 | 45 | 7.394 | 74 | 2.685 | 103 | 7 |
| 17 | 8.823 | 46 | 7.172 | 75 | 2.543 | 104 | 11 |
| 18 | 9.130 | 47 | 6.812 | 76 | 2.258 | 105 | 5 |
| 19 | 8.944 | 48 | 6.985 | 77 | 1.986 | 106 | 2 |
| 20 | 8.485 | 49 | 7.082 | 78 | 2.048 | 107 | 3 |
| 21 | 8.187 | 50 | 7.605 | 79 | 1.830 | 108 | 1 |
| 22 | 8.362 | 51 | 6.656 | 80 | 1.705 | 109 | 1 |
| 23 | 8.433 | 52 | 7.030 | 81 | 1.451 | 110 | 2 |
| 24 | 8.121 | 53 | 6.985 | 82 | 1.343 | 111 | 1 |
| 25 | 7.909 | 54 | 7.063 | 83 | 1.343 | 112 | 1 |
| 26 | 7.703 | 55 | 7.195 | 84 | 1.039 | 114 | 1 |
| 27 | 7.596 | 56 | 6.705 | 85 | 931 | | |

Fonte: PNAD Contínua 2019 – 2º trimestre

Tabela A34: Total de registros de domicílios e pessoas com risco de revelação acima de 0,2 para as duas propostas de mascaramento, por Unidade da Federação.

| Unidade da Federação | Registros com r > 0,2 | |
|---|---|---|
| | C3'+supressões | C3''+supressões |
| 11.Rondônia | 1 | 1 |
| 12.Acre | 18 | 7 |
| 14.Roraima | 7 | 3 |
| 16.Amapá | 6 | 2 |
| 17.Tocantins | 1 | -- |
| 21.Maranhão | 2 | 1 |
| 24.Rio Grande do Norte | 2 | -- |
| 25.Paraíba | 2 | -- |
| 28.Sergipe | 1 | -- |
| 31.Minas Gerais | 1 | -- |
| 32.Espírito Santo | 1 | 1 |
| 42.Santa Catarina | 4 | 3 |
| 51.Mato Grosso | 4 | 2 |
| 52.Goiás | 1 | -- |

Fonte: Elaborado pelo autor a partir dos dados da PNAD Contínua 2019 – 2º trimestre