ISSN: 2237-826X
ARTIGO | PAPER

Estimação do risco de revelação em pesquisas amostrais domiciliares

Disclosure risk estimation in household sample surveys

Bruno Freitas Cortez[1], Maysa Sacramento de Magalhães[2]

[1] *Escola Nacional de Ciências Estatísticas (ENCE), Rio de Janeiro – RJ, Brasil. ORCID: https://orcid.org/0000-0002-8667-6137*

[2] *Escola Nacional de Ciências Estatísticas (ENCE), Rio de Janeiro – RJ, Brasil.*

Autor para correspondência/*Mail to:* Bruno Freitas Cortez, bruno2977@gmail.com

Recebido/*Submitted*: 06 de fevereiro de 2023; **Aceito/*Approved*:** 30 de julho de 2023

Resumo

Introdução: Este artigo tem por objetivo estimar o risco de revelação – a probabilidade de se descobrir a identidade da unidade respondente em um banco de dados disseminado – de um arquivo de uso público da Pesquisa Nacional por Amostra de Domicílios Contínua. **Método:** A estimação foi realizada por meio de um modelo probabilístico, mais especificamente, o modelo Benedetti e Franconi, coloquialmente chamado de "abordagem italiana". **Resultados:** Observou-se que, embora a maioria dos registros possua um risco de revelação muito baixo, existem alguns que requerem maior atenção com relação à sua disseminação, devido ao alto risco. Isto ocorre mesmo para recortes geográficos de divulgação mais agregados. **Conclusão:** As estimativas do risco de revelação apresentadas apontam que técnicas de Controle Estatístico de Confidencialidade são ferramentas fundamentais para auxiliar os produtores de informação, em sua missão de garantir a confidencialidade das informações.

Palavras-chave: Risco de revelação; Controle estatístico de confidencialidade; Microdados; Pesquisas amostrais domiciliares.

Abstract

Introduction: *This article aims to estimate the disclosure risk – the probability of discovering the identity of the respondent unit in a disseminated database – from the Continuous National Household Sample Survey public use file.* ***Method:*** *The estimation was carried out using a probabilistic model, more specifically, the Benedetti and Franconi model, colloquially called the "Italian approach".* ***Results:*** *It was observed that although most records have a very low disclosure risk, there are some that require greater attention regarding their dissemination, due to the high risk. This occurs even for more aggregated geographical dissemination areas.* ***Conclusions:*** *The estimates of disclosure risk presented indicate that Statistical Disclosure Control techniques are fundamental tools to help information producers in their mission to guarantee the confidentiality of information.*

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

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 publicados na Internet (Santos & Kowata, 2018). Essa 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 intangíveis (Machado & Famá, 2011), e sua qualidade é fator de relevante vantagem competitiva (Rocha, 2019).

Entretanto, quando se trata de estatísticas oficiais, 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 com relação 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 (United Nations, 2015). Dessa 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 esses procedimentos podem não ser suficientes e cada vez mais vem sendo empregado técnicas de Controle Estatístico de Confidencialidade (CEC).

Benschop, Machingauta, e Welch (2019) definem o CEC como um conjunto de métodos que procura tratar o dado para que 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 mascaramento do dado. Elliot e Domingo-Ferrer (2018) argumentam que o CEC consiste fundamentalmente de duas etapas: análise do risco de revelação e controle de revelação. Risco de revelação pode ser definido como a probabilidade de um intruso – alguém que tentará, a partir do banco de dados disseminado, descobrir informação que ele previamente

desconhecia de um ou mais respondentes – ter êxito em seu intento. Particularmente a chamada revelação de identidade, que ocorre quando um intruso consegue determinar uma relação de um para um entre um registro nos microdados e uma entidade-alvo com um determinado grau suficiente de confiança (Waal & Willenborg, 1996), deve ser evitada. Em outras palavras, a partir do banco de dados divulgado, o intruso não deve descobrir quem é a unidade respondente, seja pessoa física ou jurídica.

Dessa forma, a primeira etapa visa mensurar o risco de revelação que, idealmente, deve ser mantido abaixo de um limite máximo aceitável pré-estabelecido. Identificados os registros com risco acima desse limite, a segunda etapa visa alterar o banco de dados, por meio da aplicação de métodos de mascaramento, para contornar esse problema. Este artigo vai se concentrar na primeira etapa, ou seja, calcular o risco de revelação dos registros. As estratégias de mascaramento, por sua vez, são diversas e decididas pelo produtor dos dados levando em conta as especificidades de cada pesquisa, a partir de algum arcabouço institucional definido para estas questões.

Elliot e Domingo-Ferrer (2018) acrescentam ainda que o campo de estudo do CEC foi uma área desenvolvida de forma lenta e gradual em resposta a desafios que os profissionais dos INEs 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 proliferação do número de organizações detentoras de dados sobre unidades individuais (microdados).

Essa crescente demanda de microdados por pesquisadores é justificada, uma vez que, para proferir 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). Dessa forma, os INEs sofrem pressão para divulgar microdados que possam promover novas pesquisas e descobertas, assim como subsidiar políticas públicas. Nesse sentido, muitos órgãos governamentais veem como parte de sua missão compartilhar dados detalhados no nível do registro individual (Taylor, Zhou, & Rise, 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 esses produtores não são apenas obrigados a proteger a confidencialidade das informações recebidas, mas ela também é crucial para manter a confiança dos entrevistados pelos INEs e garantir a honestidade e validade de suas respostas (Templ, 2017).

Sob o ponto de vista legal, o Instituto Brasileiro de Geografia e Estatística (IBGE) – o INE brasileiro – tem sua atividade regida pela Lei nº 5.534/1968. Em seu primeiro artigo, a contrapartida da obrigatoriedade da prestação de informações, é justamente que estas sejam usadas em caráter sigiloso, somente para fins estatísticos. A Lei nº 5.878/1973, que dispõe sobre o IBGE e dá outras providências, apresenta em seu sexto artigo similar redação (Instituto Brasileiro de Geografia e Estatística [IBGE], 2018). Adicionalmente, foi promulgada, em 2018, a Lei nº 13.709, conhecida como Lei Geral de Proteção de Dados (LGPD), que estabelece uma estrutura legal de direitos dos titulares de dados pessoais. Embora essa lei não impeça a prestação de informações ao IBGE, que é amparada pelas duas leis anteriores citadas, traz novos desafios para a Instituição, como, por exemplo, no aprimoramento de procedimentos para o uso de registros administrativos, bases de dados e dados não estruturados (IBGE, 2023).

No que diz respeito aos microdados disseminados de suas pesquisas, o IBGE adota uma série de medidas para garantir a confidencialidade das informações. Especificamente nas pesquisas amostrais domiciliares, objeto de estudo deste artigo, os dados são desindentificados, 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 (IBGE, 2018). Entretanto não são calculadas estimativas para o risco de revelação nos microdados de uso público, ou seja, aqueles disponibilizados em sua página oficial.

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, Elliot, e Salazar-González (2011) afirmam que a informação do domicílio aumenta o risco de revelação dos registros contidos nele, e que esse incremento pode ser particularmente grande para domicílios com muitos moradores.

Dessa forma, este artigo tem por objetivo estimar o risco de revelação de uma pesquisa amostral domiciliar, mais especificamente a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua). Ressalta-se que, para avaliar o risco, é preciso assumir que a tentativa de tal revelação ocorre de acordo com um cenário a ser especificado, que será visto com detalhes na seção de método. A motivação da escolha desta pesquisa foi por esta ser a principal do gênero, abrangendo todo o território nacional, tendo seus resultados divulgados para diversos recortes geográficos, variando desde o total do país, até para os municípios das capitais em seu nível mais desagregado.

A PNAD Contínua faz parte do Sistema Integrado de Pesquisas Domiciliares (SIPD) que o IBGE começou a implantar no ano de 2006. No que concerne às motivações temáticas, o SIPD foi pensando 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.

O objetivo PNAD Contínua é 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, 2021, p. 5). A pesquisa foi implantada, inicialmente de forma experimental, em outubro de 2011, ainda sem abrangência nacional, com o intuito de realizar os ajustes necessários em seus processos. 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 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, 2021).

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 desses 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. 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.

No momento do início de elaboração do estudo apresentado neste artigo, existiam microdados da PNAD Contínua de disseminação pública na página da internet do IBGE até o ano de referência de 2020. Todavia, 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. Considerou-se também o fato do ano de 2020 ser atípico devido ao início da pandemia do SARS-COV-2.

Para o ano de 2019, existem seis arquivos de microdados disponíveis para o acesso público na página do IBGE: os quatro arquivos de divulgação trimestral e os dois de divulgação anual previamente mencionados. 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. Os arquivos de divulgação trimestral possuem em torno de 550 mil registros, contra aproximadamente 440 mil dos de divulgação anual.

Dessa forma, determinado que seriam usados microdados de divulgação trimestral de 2019, a etapa seguinte seria definir qual trimestre utilizar. Para um intruso, geralmente quanto mais informação no banco de dados melhor, pela possibilidade de mais variáveis para facilitar a revelação e, em caso de sucesso na identificação da entidade, mais informações coletar. Na divulgação do segundo trimestre a PNAD Contínua apresentou um módulo suplementar sobre o tema de educação. Perguntas sobre esse tema, que vem sendo recorrente desde 2016 na pesquisa, possuem um potencial para conter variáveis que possam ser consideradas identificadoras do respondente. Sendo assim, definiu-se por utilizar neste artigo os microdados referentes à divulgação do segundo trimestre do ano escolhido.

MÉTODO

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 tente, 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 & 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. Este artigo se concentrará nos dados de uso público, justamente por sua maior vulnerabilidade ao ataque de intrusos.

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 a procura de respondentes em comum; ou o conhecimento próprio sobre características das unidades respondentes que ele conheça para tentar fazer a revelação (Benschop et al., 2019; Templ, 2017).

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 presentes no conjunto de dados-alvo, permitindo que as unidades respondentes possam ser identificadas. Benschop et al. (2019) acrescentam que o risco de revelação dependerá da inclusão ou exclusão das variáveis-chave escolhidas. Dessa forma, o processo de seleção do conjunto dessas variáveis deve, portanto, ser abordado com grande atenção e cuidado. Os autores sugerem, inclusive, que o primeiro passo seria o de realizar um inventário de todos os conjuntos de dados disponíveis no país. De posse dessa lista, deveriam ser analisadas quais variáveis estariam contidas nesses bancos de dados e que poderiam ser utilizadas por potenciais intrusos.

Dessa forma, foi realizada uma pesquisa em relação às bases de dados que possam servir de informações externas para um intruso vincular suas informações com a PNAD Contínua. Com relação às bases públicas, levou-se em consideração três critérios: escopo, atualização e abrangência geográfica dos dados disponíveis. O primeiro critério se refere à população-alvo das pesquisas, ou seja, a possibilidade de um entrevistado da PNAD Contínua esteja também nessa base externa. O segundo é relativo ao período de coleta das informações, pois, idealmente, elas devem ser o mais próximas possíveis para o intruso fazer uma revelação correta. Por fim, a abrangência geográfica é essencial, pois pelo menos alguns registros de ambas as bases devem pertencer a uma mesma área para ser possível a revelação. É importante frisar que as variáveis das bases externas devem ser idênticas ou, pelo menos, compatibilizáveis com as da PNAD Contínua.

Assim, levando-se em conta todas as questões descritas, a escolha das variáveis-chave teve como base os seguintes bancos de dados externos: o Cadastro Geral de Empregados e Desempregados (CAGED), a Relação Anual de Informações Sociais (RAIS), o Cadastro Único (CadÚnico) e o Censo da Educação Superior (Censo Superior). Adicionalmente com relação aos dados privados, observa-se que, com o grande avanço tecnológico presenciado 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 nessas 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, mau uso ou vazamentos de informações não podem ser descartados. É preciso, então, considerar que existam intrusos com, ao menos, algumas informações básicas das pessoas (por exemplo: nome, sexo, idade etc.) ao seu dispor. Assim, os dois cenários de revelação e as variáveis-chave consideradas são:

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, sejam públicos, privados ou uma mistura de ambos. A Tabela 1 apresenta as variáveis-chave e suas respectivas categorias selecionadas para este cenário.

Deve ser ressaltado que a categoria "não aplicável" para a variável anos de estudo, presente na Tabela 1, refere-se às pessoas com menos de 5 anos de idade na PNAD Contínua. É possível assumir que este grupo possui zero anos de estudo, mas, na prática, dificilmente um intruso teria interesse em fazer a revelação dessas 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).

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

Tabela 1. Variáveis-chave e suas categorias selecionadas para o cenário de revelação 1

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. Nesse 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. A Tabela 2 apresenta as variáveis-chave e suas respectivas categorias selecionadas para este cenário.

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

Tabela 2. Variáveis-chave e suas categorias selecionadas para o cenário de revelação 2

Como pode ser observado na Tabela 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 dessa informação, mas não necessariamente com total exatidão.

As estimativas do risco de revelação baseadas nas variáveis-chaves para ambos os cenários serão calculadas para todos os recortes geográficos divulgados pela PNAD Contínua: Brasil, Grandes Regiões, Unidades da Federação (UF), Regiões Metropolitanas (RM) que contém municípios das capitais e todos os municípios de capitais. Entretanto um intruso consegue ter acesso a dois recortes geográficos adicionais com as informações fornecidas no banco de dados. Quando é incluída uma variável para designar que um domicílio pertence a um município de capital ou a uma RM dentro de uma UF, o intruso sabe, por exclusão, que o registro que não contém esse valor está fora da capital ou da RM daquela respectiva UF. Logo, com a incorporação desses dois novos recortes geográficos, os resultados deste artigo contemplam sete domínios de análise distintos, quais sejam: Brasil; Grandes Regiões; Unidades da Federação; Regiões Metropolitanas que contêm municípios das capitais; Municípios fora destas RMs; Municípios das capitais; Municípios fora das capitais.

Ressalta-se ainda que, seguindo a recomendação de Hundepool et al. (2012), qualquer suposição deve ser conservadora, implicando em 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.

Como já mencionado anteriormente, é muito importante levar a estrutura hierárquica, presente na PNAD Contínua, 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.

A Abordagem italiana

Como previamente descrito na introdução, existem duas abordagens para se calcular o risco de revelação: a individual para cada registro, ou a global para o arquivo como um todo. Neste artigo, 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. De qualquer modo, o risco global do arquivo pode ser estimado por métodos que utilizem os riscos individuais.

A abordagem utilizada será por meio de modelo probabilístico. A ideia geral, sintetizada por Duncan et al. (2011), é que registros que apresentam valores de chave (o conjunto de respostas dadas pelo respondente nas variáveis-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 essa 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 essa 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 Benedetti e Franconi (1998) e posteriormente desenvolvida por Polettini e Stander (2004).

Essa última, coloquialmente chamada de "abordagem italiana", destaca-se sob o ponto de vista operacional, uma vez que esta se encontra incorporado tanto no *software* μ-*Argus* quanto no pacote "sdcMicro" do *software* R. De fato, Templ (2017) aponta que essa abordagem é usual na literatura. Dessa forma, tal método foi escolhido para a estimação do risco de revelação neste artigo.

Para a definição do risco de revelação individual, Benedetti, Capobianchi, e Franconi (1998) consideram uma amostra S de tamanho n selecionada de uma população finita 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:

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

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 registros na amostra que possuam o mesmo k são idênticos, em termos de risco de revelação. Dessa 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/f_k (distribuição condicional de F_k dado f_k), tal que:

$$r_k = E(\frac{1}{F_k}|f_k) = \sum_{h>=f_k} \frac{1}{h} P(F_k = h|f_k) \tag{1}$$

Para determinar a função de densidade de F_k/f_k, uma abordagem de superpopulação é introduzida (Bethlehem, Keller, & Pannekoek, 1990; Polettini, 2003; Rinott, 2003), em que é pressuposto:

$$F_k|\pi_k \sim Poisson(N\pi_k), F_k = 0, 1, ..., independentemente \tag{2}$$

tal que πk 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 esse 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. Benedetti et al. (1998) demonstram que, sob a hipótese de distribuição a posteriori de F_k/f_k ser binomial negativa, a equação (1) pode ser expressa como:

$$r_k = E(F_k^{-1}/f_k) = \int_0^\infty \left\{ \frac{p_k exp(-t)}{1 - q_k exp(-t)} \right\}^{f_k} dt \tag{3}$$

tal que p_k é a probabilidade de uma unidade populacional com o valor de chave k ser selecionado na amostra e $q_k = 1 - p_k$.

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. Benedetti et al. (1998), com base no estimador de superpopulação de máxima-verossimilhança de p_k e introduzindo a informação dada pelo desenho amostral D, resumida pelo estimador de Horvitz-Thompson de F_k, obtêm o estimador de p_k com base no desenho amostral:

$$\hat{p}_k^D = \frac{f_k}{\sum_{i:k(i)=k} {}^{w}\, i} \tag{4}$$

tal que w_i é o peso amostral do registro i com valor de chave k. Com vistas à criação de um algoritmo para a implementação computacional do modelo de Capobianchi, Polettini, e Lucarelli (2001), utilizando o estimador da equação (4), daqui para frente chamado apenas de k por questão de simplicidade, estimam o risco de revelação a partir da seguinte aproximação:

$$r_k = \frac{\hat{p}_k}{f_k - (1 - \hat{p}_k)} \tag{5}$$

RESULTADOS

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 desses registros, o modelo atribui uma estimativa do referido risco. 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 do arquivo de dados a ser disseminado.

Assim, não faz parte do escopo deste artigo definir qual deveria ser o limite máximo aceitável de risco de revelação que se deve 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 3 e 4 apresentam esses resultados de acordo com os dois cenários de revelação propostos.

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

Tabela 3. Registros de domicílios não seguros, dado um limiar de risco de revelação, por recorte geográfico de divulgação, segundo o cenário de revelação 1

De acordo com as Tabelas 3 e 4, observa-se que a maioria dos registros possui risco de revelação abaixo de 1% em todos os recortes geográficos considerados. Entretanto, quanto mais desagregado esse recorte, a proporção de registros de domicílios acima dos limiares propostos tende a aumentar. É possível também verificar que o cenário 2 possui mais registros acima dos limiares propostos. 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 desse recorte possuem risco estimado superior a 10%, e 1% possui um risco maior que 20%, como mostra a Tabela 4. Entretanto, ainda que nenhum recorte geográfico fosse divulgado, ou seja, que os dados se refiram apenas para o Brasil, ainda existiriam registros pontuais de domicílios com risco acima do limiar de 20% para ambos os cenários considerados, como pode ser observado em ambas as tabelas.

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

Tabela 4. Registros de domicílios não seguros, dado um limiar de risco de revelação, por recorte geográfico de divulgação, segundo o cenário de revelação 2

Outra análise importante 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 e 6, que as distribuições das estimativas em todos os recortes geográficos e cenários são assimétricas positivas. Corroborando as análises anteriores, verifica-se também que os valores das 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.

Medidas de posição	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

Tabela 5. Medidas de posição do risco de revelação dos domicílios, por recorte geográfico de divulgação, segundo o cenário de revelação 1

Medidas de posição	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

Tabela 6. Medidas de posição do risco de revelação dos domicílios, por recorte geográfico de divulgação, segundo o cenário de revelação 2

As Tabelas 5 e 6 reforçam a ideia de que a grande maioria dos registros apresenta estimativas de risco de revelação muito baixa. No entanto uma medida importante apresentada nas referidas tabelas é o 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 que podem ser suscetíveis à revelação por um intruso. Ao se considerar o cenário 2, nessa medida específica de posição, constata-se que, mesmo no recorte mais agregado possível, existe um registro com risco bem elevado. Este risco é maior ainda nos níveis mais desagregados, como esperado.

Se as informações das Tabelas 3 e 4 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 e 6 mostram que, dentre esses registros, existem alguns especialmente vulneráveis, muito acima dos limiares propostos. Entretanto há ainda uma questão a ser respondida, que é o motivo pelo qual o risco permanece alto em alguns registros, mesmo para os recortes geográficos mais agregados. Isto pode ser explicado, em grande parte, pela estrutura hierárquica da pesquisa. Tomando-se como exemplo o cenário 2, que apresentou maior quantidade de registros acima dos limiares propostos, a Tabela 7 apresenta a distribuição dos registros com risco acima de 20%, segundo o tamanho do domicílio, em número de moradores, nos recortes mais agregados.

É possível observar, com base na Tabela 7, que, embora os domicílios com oito ou mais moradores representassem menos de 1% da amostra, em todos os recortes estavam ali contidos o maior número absoluto dos casos não seguros dado o limiar de 20%. Em números relativos, esse montante é maior quanto mais agregado é o recorte geográfico. Para o segundo cenário, considerando uma disseminação apenas para o Brasil, 78 dentre os 95 casos de registros não seguros se encontram nesse estrato. 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. Essa questão já é, inclusive, tratada por outros produtores de dados oficiais, como o Office for National Statistics (ONS), o INE do Reino Unido. Especificamente em relação aos dados de uso público, pode ocorrer a supressão do registro inteiro no banco de dados, em casos de domicílios que ultrapassem um determinado número máximo de moradores (Government Statistical Service [GSS], 2014).

Sumarizando, os resultados, nesta seção, mostram que a estimação do risco de revelação pode fornecer importante insumo para auxiliar os produtores de informação, na tomada de decisão quanto à disseminação de informações em microdados.

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

Tabela 7. Registros de domicílios não seguros, dado um limiar de risco de revelação de 20%, por tamanho do domicílio e recorte geográfico, segundo o cenário de revelação 2

CONSIDERAÇÕES FINAIS

De acordo com os resultados apresentados, verificou-se que a grande maioria de registros na PNAD Contínua analisada possui um risco de revelação muito baixo. Entretanto existem alguns registros pontuais que podem requerer maior atenção com relação à sua disseminação. Embora quanto mais agregado seja o recorte geográfico divulgado essa situação tenda a diminuir, ainda persistem registros de domicílios com riscos altos, principalmente os com grande número de moradores, mesmo ao se considerar resultados para o Brasil.

Adicionalmente, observou-se que a divulgação de determinados recortes geográficos (como municípios em Regiões Metropolitanas ou municípios de capitais), permite ao intruso compor outros recortes de divulgação não previstos (domicílios naquela UF fora da Região Metropolitana ou municípios de capitais). Em outras palavras, ao se estimar o risco de revelação, deve-se ter em mente o uso dos dados sob a perspectiva de um intruso.

De modo geral, as estimativas do risco de revelação apresentadas apontam que técnicas de CEC são ferramentas fundamentais para auxiliar os produtores de informação, em sua missão de garantir a confidencialidade das informações.

REFERÊNCIAS

Benedetti, R., Capobianchi, A., & Franconi, L. (1998). Individual risk of disclosure using sampling design information. *Contributi Istat*, *14*. 1412003. Recuperado de https://www.researchgate.net/publication/243784265_Individual_risk_of_disclosure_using_sampling_design_information

Benedetti, R., & Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of new techniques and technologies for statistics* (p. 225–232).

Benschop, T., Machingauta, C., & Welch, M. (2019). *Statistical disclosure control: a practice guide.* Recuperado de https://sdcpractice.readthedocs.io/en/latest/

Bethlehem, J., Keller, W., & Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, *85*(409), 38–45. doi: 10.2307/2289523

Capobianchi, A., Polettini, S., & Lucarelli, M. (2001). *Strategy for the implementation of individual risk methodology into μ-ARGUS* (Relatório Técnico n. 1.2-D1). Roma: CASC project.

Castanha, R. C. G. (2021). A ciência de dados e a cientista de dados. *AtoZ: novas práticas em informação e conhecimento*, *10*(2), 1–4. doi: 10.5380/atoz.v10i2.79882

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). *Statistical confidentiality: Principles and practice.* New York: Springer. doi: 10.1007/978-1-4419-7802-8

Elamir, E. A. H., & Skinner, C. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, *22*(3), 525–539. Recuperado de https://www.researchgate.net/publication/247361113_Record_Level_Measures_of_Disclosure_Risk_for_Survey_Microdata

Elliot, M. J., & Domingo-Ferrer, J. (2018). The future of statistical disclosure control.. Recuperado de https://www.researchgate.net/publication/329884395_The_future_of_statistical_disclosure_control

Franconi, L., & Polettini, S. (2004). Individual risk estimation in μ-argus: a review. In J. Domingo-Ferrer (Ed.), *Privacy in statistical databases* (v. 2736, p. 262–272). Springer. Recuperado de https://link.springer.com/chapter/10.1007/978-3-540-25955-8_20 doi: 10.1007/978-3-540-25955-8_20

Government Statistical Service. (2014). *Gss/gsr disclosure control guidance for microdata produced from social surveys.* Recuperado de https://analysisfunction.civilservice.gov.uk/wp-content/uploads/2018/03/Guidance-for-microdata-produced-from-social-surveys-4.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P.-P. d. (2012). *Statistical disclosure control.* Wiley. doi: 10.1002/9781118285064

Instituto Brasileiro de Geografia e Estatística. (2018). *Confidencialidade no ibge: procedimentos adotados na preservação do sigilo das informações individuais nas divulgações de resultados das operações estatísticas.* IBGE.

Instituto Brasileiro de Geografia e Estatística. (2021). *Pesquisa nacional por amostra de domicílios contínua. notas técnicas versão 1.8.* IBGE.

Instituto Brasileiro de Geografia e Estatística. (2023). *Estratégia geral de tecnologia da informação e comunicação do ibge. egti 2023-2024.* IBGE.

Machado, J. H., & Famá, R. (2011). Ativos intangíveis e governança corporativa no mercado de capitais brasileiro. *Revista Contemporânea de Contabilidade*, *8*(16), 89–110. Recuperado de https://periodicos.ufsc.br/index.php/contabilidade/article/view/2175-8069.2011v8n16p89/20046 doi: 10.1590/S1519-68952011000100007

Polettini, S. (2003). Some remarks on the individual risk methodology. In *Joint ece/eurostat work session on statistical data confidentiality.* Recuperado de https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2003/04/confidentiality/wp.18.e.pdf

Polettini, S., & Stander, J. (2004). A bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases* (p. 247–261). Springer. doi: 10.1007/978-3-540-25955-8_19

Rinott, Y. (2003). On models for statistical disclosure risk estimation.. Recuperado de https://api.semanticscholar.org/CorpusID:16509807

Rocha, D. F. (2019). Concorrência em mercados digitais e desafios ao controle de atos de concentração. *Revista de Defesa da Concorrência*, *7*(2). Recuperado de https://revista.cade.gov.br/index.php/revistadedefesadaconcorrencia/article/view/413/236

Santos, Y. T., & Kowata, E. T. (2018). A importância do big data nas organizações. In *V congresso de ensino, pesquisa e extensão da ueg.* Recuperado de https://www.anais.ueg.br/index.php/cepe/article/view/13307

Skinner, C. J., & Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, *14*(4), 361–372. Recuperado de https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/estimating-the-re-identification-risk-per-record-in-microdata.pdf

Taylor, L., Zhou, X. H., & Rise, P. (2018). A tutorial in assessing disclosure risk in microdata. *Statistics in Medicine*, *37*(25), 3693–3706.

Templ, M. (2017). *Statistical disclosure control for microdata: methods and applications in r.* Springer.

United Nations. (2015). *United nations fundamental principles of official statistics: implementation guidelines* (Relatório Técnico). Recuperado de https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf

Waal, T., & Willenborg, L. C. (1996). A view on statistical disclosure control for microdata. *Survey Methodology*, *22*(1), 95–103. Recuperado de https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14381-eng.pdf?st=KKaA-W9y

Willenborg, L., & de Waal, T. (2001). *Elements of statistical disclosure control.* Springer.

Como citar este artigo (APA):
Cortez, B. F. & Magalhães, M. S. de (2023). Estimação do risco de revelação em pesquisas amostrais domiciliares. *AtoZ: novas práticas em informação e conhecimento, 12*, 1 – 11. Recuperado de: http://dx.doi.org/10.5380/atoz.v12.89682

NOTAS DA OBRA E CONFORMIDADE COM A CIÊNCIA ABERTA

CONTRIBUIÇÃO DE AUTORIA

Papéis e contribuições	Bruno Freitas Cortez	Maysa Sacramento De Magalhães
Concepção do manuscrito	X	X
Escrita do manuscrito	X	
Metodologia	X	
Curadoria dos dados	X	
Discussão dos resultados	X	X
Análise dos dados	X	X

EQUIPE EDITORIAL