Estudo de Estimativa de Sucesso de Re-Identificações em Divulgações Estatísticas Brasileiras

Lucas Starling de Paula Salles

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

lucasstarling1@gmail.com

Resumo. *Dados demográficos estatísticos são importantíssimos para direcionar a política de um estado democrático, porém essas informações também representam um risco à privacidade dos cidadãos. Para garantir o direito da população à privacidade foi criada a LGPD, em 2018, que estabelece que dados demográficos precisam ser anônimos para poderem ser publicados. A anonimização por desidentificação e amostragem é a metodologia mais frequentemente usada no Brasil, por instituições como o Instituto Brasileiro de Geografia e Estatística, para proteger esse tipo de dados. Esse trabalho busca verificar se essa estratégia de anonimização resulta, ou não, em dados que permanecem vulneráveis à re-identificação. São avaliados dados da pesquisa PeNSE, realizada pelo IBGE, e é observado que re-identificações bem sucedidas são extremamente prováveis, o que pode configurar uma violação da LGPD.*

1. Introdução

Privacidade, o direito à reserva de informações pessoais e da própria vida pessoal, é um conceito cada dia mais discutido. A privacidade foi intensamente afetada pela revolução digital, desrespeitada pela coleta e uso constantes de informações de usuários em todos os dispositivos conectados à internet. Contudo a coleta massiva de dados não serve apenas para fomentar uma indústria baseada em anúncios guiados. As estatísticas oficiais constituem um elemento indispensável no sistema de informação de uma sociedade democrática, oferecendo ao governo, à economia e ao público dados sobre a situação econômica, demográfica, social e ambiental de um país. Por isso é imperativo que seja possível conduzir pesquisas e divulgar estatísticas sem que o direito à privacidade dos cidadãos seja comprometido. Para esse fim, bancos de dados são anonimizados antes da publicação de suas informações.

Existem muitas técnicas para anonimizar um bancos de dados, frequentemente isso é feito através da combinação das técnicas de desidentificação e amostragem. Essa abordagem de anonimização é usada regularmente por dificultar a re-identificação de um usuário através da desidentificação, e fornecer negação plausível para o caso de se identificar algum indivíduo devido a amostragem.

O Instituto Brasileiro de Geografia e Estatística (IBGE) é o principal provedor de informações geográficas e estatísticas do país. Provendo mais de uma centena de estudos e pesquisas, o IBGE é responsável pela coleta e divulgação de dados agrícolas,

demográficos, econômicos, sociais, populacionais, entre outros. O Instituto é independente de organizações políticas e administrativas, não é suscetível a interferências externas. Governos, empresas, domicílios e o público devem, por lei, permitir o acesso do IBGE a dados destinados à elaboração de estatísticas oficiais. Por isso, conta com um volume imenso de dados sobre os mais diversos aspectos da população brasileira e de suas instituições.

Riqueza e abrangência de dados é essencial para a realização de estudos compreensivos e para a obtenção de resultados que representem de maneira adequada a população e o país. Contudo, os dados coletados pelo IBGE contém informações que, conforme a Lei Geral de Proteção de Dados Pessoais ([LGPD] – Lei n° 13.709, Art. 5, de 14 de agosto de 2018), se enquadram como dados pessoais sensíveis (dados sobre origem racial ou étnica, saúde, vida sexual, dentre outras) e representam um risco à privacidade dos respondentes. A Legislação prevê a confidencialidade desses dados e a garantia de que serão usados exclusivamente para fins estatísticos, conforme a Lei nº 5.534, de 14 de novembro de 1968, Art. 1, além de assegurar a anonimização dos dados pessoais ([LGPD], Art. 7). Para honrar esses direitos o IBGE se compromete com a rígida preservação do sigilo das informações individuais ou identificadas.

Para assegurar a confidencialidade exigida de seus dados, o Instituto implementa um conjunto de medidas. Seus protocolos e bancos de dados seguem diretrizes que buscam promover a segurança e integridade. Dados estimados obtidos através da agregação de indivíduos são associados a um erro amostral e publicados sem passarem por desidentificação, porque considera-se que a natureza de sua obtenção não permite a re-identificação dos informantes. Os microdados não desidentificados, informações mais desagregadas nas quais cada registro representa um indivíduo, são sujeitos a restrição de acesso. Usuários externos devem seguir protocolos de confidencialidade para poderem acessá-los e seu uso deve servir à finalidade de pesquisa e/ou à análise estatística.

A publicação de microdados é feita mediante a restrição de dados. O IBGE considera que a utilização de conceitos, classificações e métodos internacionais pelos órgãos de estatísticas de cada país promove a coerência e eficiência dos sistemas em todos os níveis oficiais. Por isso, a proteção dos informantes de pesquisas cujos microdados são divulgados é feita através da anonimização por amostragem e desidentificação. O IBGE considera essa estratégia suficiente, por se assemelhar a instituições internacionais de divulgação estatística.

A anonimização por desidentificação e amostragem tem sido o principal paradigma usado academicamente e institucionalmente para publicar dados sem comprometer a privacidade de seus titulares. Apesar disso, têm ocorrido casos de re-identificações em bancos de dados anonimizados dessa maneira. Um exemplo disso ocorreu na Alemanha em 2016: um grupo de pesquisadores conseguiram re-identificar figuras públicas em um banco de dados anonimizado contendo o histórico de navegação de mais de 3 milhões de cidadãos, descobrindo informações médicas e sobre suas vidas sexuais, [Hern 2017]. A vulnerabilidade dessa metodologia de anonimização foi também evidenciada por [Rocher, Hendrick e Montjoye 2019]. Os pesquisadores mostram que é possível estimar a probabilidade de ser realizada a re-identificação bem sucedida de um indivíduo em bancos de dados anonimizados. Foram obtidos resultados que indicam que 99.98% dos cidadãos americanos seriam re-identificados corretamente em qualquer banco de da-

dos anonimizado com essa técnica e que contenha pelo menos 15 atributos demográficos.

A reidentificação bem sucedida de indivíduos através de dados públicos provenientes de pesquisas demográficas resultaria na divulgação de informações sensíveis, configurando uma violação da LGPD. Por isso levanto o questionamento sobre a eficácia da anonimização feita nos microdados publicados pelo IBGE. Esse estudo se propõe, portanto, a mostrar se os cidadãos brasileiros estão, ou não, vulneráveis à re-identificação como consequência da exploração de dados incompletos e desitentificados, provenientes de divulgação estatística.

2. Referencial Teórico

Diversos tipos de ataques podem ser realizados em um banco de dados com dados demográficos, variando de acordo com a estratégia e objetivo do atacante, assim como pelo tipo de informações disponíveis. Contudo, a técnica proposta em [Rocher, Hendrick e Montjoye 2019], foi selecionada para avaliar a probabilidade de sucesso para re-identificações por ser agnóstica à estratégia de ataque. Essa abordagem possibilita um estudo compreensivo sobre dados incompletos e anonimizados, e sobre as informações que podem ser extraídas desses, sem restrições impostas por estratégias específicas de ataque. Além disso, também passa a ser possível gerar resultados conclusivos sem que seja necessário associar um registro específico contido nas amostras à identidade do respondente, evitando assim agregar à exposição das informações sensíveis dos indivíduos contemplados pela pesquisa. É necessário então compreender o objetivo do processo de anonimização, assim como as técnicas de amostragem, de desidentificação e de re-identificação.

Segundo à LGPD, Art. 5 **Anonimização** é definida como a utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo.

Desidentificação consiste na supressão de variáveis que propiciem a identificação do informante, [IBGE 2018]. Processo de remoção de identificadores pessoais (nome, número telefônico, endereço, qualquer informação ou número único para uma pessoa) e supressão ou generalização de informações sobre um indivíduo que por si só não o pode identificar, mas que pode se tornar identificadora quando combinada com outros dados. Deve ser considerado desidentificado o dado que, quando avaliado estatisticamente, apresenta baixo risco de ser usado para identificar indivíduos ou sobre o qual não haja conhecimento da existência de nenhuma informação identificadora, [OCR 2015].

Amostragem consiste em selecionar parte de uma população para observar, de modo que seja possível estimar alguma característica sobre toda a população, [IBGE 2018]. Duante Operações Estatísticas (entrevista de cidadãos para obtenção de dados) o Instituto questiona apenas uma fração da população resultando em dados amostrais. Ao fazer isso é garantido negação plausível à re-identificações realizadas com os dados obtidos, porque não é possível garantir que o indivíduo identificado foi de fato entrevistado em primeiro lugar.

Re-identificação consiste no esforço bem sucedido de identificação de um indivíduo integrante de uma base de dados desidentificada, resultando na associação de informações sensíveis a uma pessoa, [OCR 2015].

2.1. Cópulas Gaussianas

A Cópula Gaussiana é uma técnica de modelagem usada para o estudo de estruturas de dependência. A metodologia possibilita a estimativa da densidade de uma distribuição de probabilidade a partir das distribuições marginais para os atributos de uma variável aleatória. Esse modelo permite, portanto, que seja estimada a distribuição de probabilidade conjunta de todos os atributos de uma amostra de dados. Isso possibilita que seja calculada a probabilidade de um registro com atributos específicos estar presente em um espaço amostral específico.

Quando alimentadas com amostras que representam uma população, as Cópulas Gaussianas aproximam bem a densidade da mesma, até quando a amostra é composta por uma pequena fração de registros. Por isso, Cópulas podem ser usadas para calcular a probabilidade da existência de um indivíduo com características definidas em uma população sobre a qual se conhece apenas uma amostra.

3. Metodologia

A técnica de estimativa de sucesso em re-identificações em bancos de dados incompletos proposta por [Rocher, Hendrick e Montjoye 2019] funciona exclusivamente com base nos dados disponíveis. Isso possibilita a obtenção de resultados genéricos para o banco de dados como um todo, que representa uma amostra estatisticamente significativa da população pesquisada.

3.1. Estimativa de sucesso de re-identificação

Essa técnica para estimar o sucesso de re-identificação é bem fundamentada e apresentou resultados estatisticamente significantes para amostras de bancos de dados desidentificados sobre a população americana. Quando aplicada em uma amostra a metodologia resulta em um modelo que pode ser usado para obter três tipos de resultados. Podemos usar este método para avaliar a unicidade absoluta na amostra, percentual de registros contendo uma combinação de atributos única no banco de dados. Além disso, pode ser obtido o valor estimado de unicidade para a população original da amostra através do treinamento de um modelo que recebe uma amostra da população. Uma vez treinado esse modelo pode gerar uma população simulada de tamanho arbitrário, respeitando a distribuição de valores para os atributos da entrada, então pode ser calculada a proporção de indivíduos que são únicos nessa população estimada para um conjunto de atributos. O modelo também permite a avaliação de indivíduos, gerando uma probabilidade estimada do indivíduo escolhido existir, e ser único, na população com base em seus atributos.

De acordo com os autores, em [Rocher, Hendrick e Montjoye 2019], a abordagem para geração do modelo descrito consiste em cinco etapas:

1. **Inferência das distribuições marginais**. Nessa etapa são geradas distribuições marginais para os atributos. Isso é feito através da máxima verossimilhança logarítmica ajustada para contagem dos valores e distribuições binomial negativa e logarítmica de cada atributo.
2. **Inferência dos parâmetros da cópula latente**. Esse passo consiste essencialmente no cálculo da matriz de covariância entre os atributos. Para cada par de atributos é atribuída a distância euclidiana mínima entre a informação mútua das distribuições marginais e a informação mútua da distribuição conjunta inferida para o par.

3. **Modelagem da estrutura de associação de atributos**. Aqui a informação mútua entre dois atributos é calculada para medir a força da associação entre o par. Para minimizar os erros, o cálculo é ajustado através do cálculo da informação mútua entre a distribuição do primeiro atributo com uma série de permutações aleatórias do segundo atributo para capturar melhor a correlação entre esses.
4. **Estimativa da unicidade da população**. A cópula, agora ajustada pela amostra, é usada para gerar uma população estimada pelas distribuições. Essa população sintética então é consultada para ser calculada a proporção de indivíduos únicos, que é a unicidade estimada da população original.
5. **Cálculo da probabilidade de unicidade e corretude para indivíduos**. Para esse cálculo primeiro se integra sobre a cópula latente para gerar a distribuição de probabilidade de um indivíduo estar presente na população. A seguir, o indivíduo de interesse é alimentado nessa distribuição para gerar a probabilidade de sua presença na população. Esse valor é então subtraído de um, gerando a probabilidade de re-identificação correta que então é descontada caso existam outros indivíduos com os mesmo atributos no amostra.

Apesar de complexa, por usar ferramentas e conceitos estatísticos avançados, os autores da metodologia disponibilizam a implementação usada por eles para gerar resultados sobre a população americana. É possível então estudar o sucesso de re-identificação em outras bases de dados sem maiores preocupações com erros de codificação. O uso da metodologia é dependente apenas de características dos dados. A ferramenta disponibilizada aceita apenas dados numéricos e os autores do trabalho de 2019 concluem que a riqueza e abrangência dos atributos demográficos dita a precisão dos resultados obtidos. A recomendação é que sejam usados pelo menos 15 atributos demográficos mas é observado que dependendo do conjunto de dados e do tamanho da amostra proporcional à população é possível obter bons resultados com menos atributos demográficos.

4. Dados

Para determinar se a anonimização realizada pelo IBGE em seus dados garante a confidencialidade das informações pessoais, conforme definido pela Legislação, é necessário escolher dados para que seja feita a avaliação experimental. As pesquisas mais suscetíveis a ataques de reidentificação são aquelas que têm amostras de microdados publicadas, mesmo depois do processo de anonimização. A Pesquisa Nacional de Saúde do Escolar (PeNSE) é particularmente vulnerável por conter diversos atributos demográficos e muitas informações sensíveis sobre os respondentes. Por isso as amostras desta pesquisa foram selecionadas para avaliação.

A escolha dos dados é sujeita às restrições discutidas anteriormente, sendo necessário encontrar um conjunto de microdados amostrais numéricos, provenientes do IBGE, que contenha atributos demográficos para viabilizar o uso da ferramenta. A existência de dados, como os da pesquis PeNSE, já corrobora a viabilidade da estimativa de sucesso de re-identificações em dados estatísticos brasileiros.

Para aplicação da metodologia proposta, foram selecionadas as amostras da pesquisa PeNSE. Esses microdados são numéricos, contém atributos demográficos sobre os cidadãos entrevistados e foram desintentificados. Por isso, se encaixam nos requisitos para utilização da técnica escolhida. Essas amostras são boas candidatas para um estudo

como esse também porque contém dados pessoais sensíveis sobre seus informantes, tornando a re-identificação de indivíduos através dessas informações uma violação do Art. 7 da LGPD.

Foi realizada a revisão das metodologias implantadas pelo IBGE para anonimizar seus dados, com o intuito de determinar se existe possibilidade de re-identificação nos dados publicados. Conforme abordado anteriormente, o Instituto considera que a desidentificação de microdados amostrais é suficiente para a proteção dos respondentes, por isso amostras como as da pesquisa PeNSE são disponibilizadas publicamente. Já foi demonstrado que essa estratégia de anonimização tem sido alvo de ataques de re-identificação bem sucedidos, portanto podemos inferir que é possível realizar esse tipo de ataque nas amostras selecionadas, [Hern 2017].

4.1. PeNSE

A Pesquisa Nacional de Saúde do Escolar teve sua primeira edição em 2009 e tem como objetivo fornecer informações "sobre as características básicas da população estudantil, do sexto ano do ensino fundamental até o terceiro ano do ensino médio. Abrange aspectos socioeconômicos, como escolaridade dos pais, inserção no mercado de trabalho e posse de bens e serviços; contextos social e familiar; fatores de risco comportamentais relacionados a hábitos alimentares, sedentarismo, tabagismo, consumo de álcool e outras drogas; saúde sexual e reprodutiva; exposição a acidentes e violências; hábitos de higiene; saúde bucal; saúde mental; e percepção da imagem corporal, entre outros tópicos."[IBGE].

Ocorreram três edições da PeNSE: 2009, 2012 e 2015, sendo que a pesquisa mais recente foi a primeira que contou com a coleta de informações de alunos do 6^o ano do ensino fundamental até o 3^o ano do ensino médio. Até então eram questionados apenas os alunos do 9^o ano do ensino fundamental. Isso significa que os dados de 2015 são mais volumosos e completos, portanto são mais adequados para um estudo como esse. Os dados dessa edição da pesquisa são divididos em duas amostras, cada uma contendo três bases de dados. Em ambas amostras a base de dados "ALUNO_ESCOLA"contém mais informações, porém essas bases não podem ser agregadas porque contém atributos demográficos distintos. Assim sendo foi feita a opção pelo uso do banco de dados "ALUNO_ESCOLA"da amostra número um.

O conjunto de microdados escolhido contém mais de cem mil registros compostos por mais de cento e setenta atributos. Desses, foram selecionados 8 atributos demográficos, facilmente reconhecíveis sobre um indivíduo, para estimar a probabilidade de re-identificação:

Atributo	Significado
VB01001	Sexo do estudante
VB01002	Cor/Raça do estudante
VB01004	Mês de nascimento do estudante
VB01005	Ano de nascimento do estudante
UFCENSO	Unidade federativa
TIPO_MUNIC	Capital ou não do estado
escola	Código da escola do estudante
turma	Código da turma do estudante

5. Resultados

O primeiro experimento realizado tratou de calcular a unicidade da amostra para os atributos selecionados. Nesse contexto, os atributos são usados como quasi-identificadores (QIDs), porque agem como um identificador composto por diversas informações distintas. Para determinar como esses atributos afetam a unicidade dos registros da amostra, todas as combinações deles foram avaliadas. Para cada uma das 256 (2^8) combinações foi feito o cálculo da unicidade dos registros da amostra:

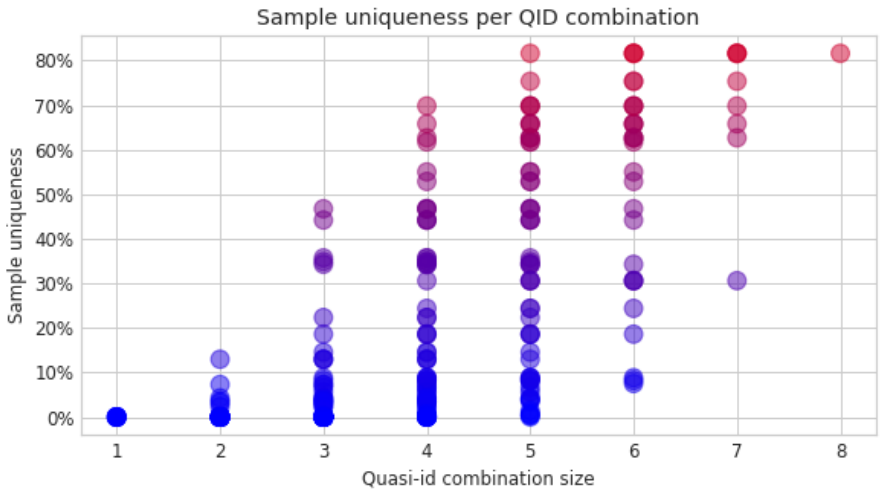

Figura 1. Unicidade na amostra por combinação de atributos

Não é viável testar todas as combinações de quasi-identificadores ao estimar a unicidade da população, porém é razoável imaginar que os conjuntos que geram maior unicidade na amostra também produziriam bons resultados na população, visto que o modelo de estimativas usa a amostra para modelar as distribuições de cada atributo. Por isso, com os valores de unicidade para as combinações de atributos calculados, foram selecionadas as melhores combinações para cada tamanho de conjunto, para serem usadas futuramente.

Tamanho	Conjunto de atributos	Unicidade na amostra
1	**escola**	0.001%
2	**VB01004, turma**	12.959%
3	**VB01004, VB01002, turma**	46.741%
4	**VB01004, VB01005, VB01002, turma**	69.823%
5	**VB01001, VB01004, VB01005, VB01002, turma**	81.630%
6	**VB01001, VB01004, VB01005, VB01002, TIPO_MUNIC, turma**	81.630%
7	**VB01001, VB01004, VB01005, VB01002, UFCENSO, escola, turma**	81.630%
8	**VB01001, VB01004, VB01005, VB01002, UFCENSO, TIPO_MUNIC, escola, turma**	81.630%

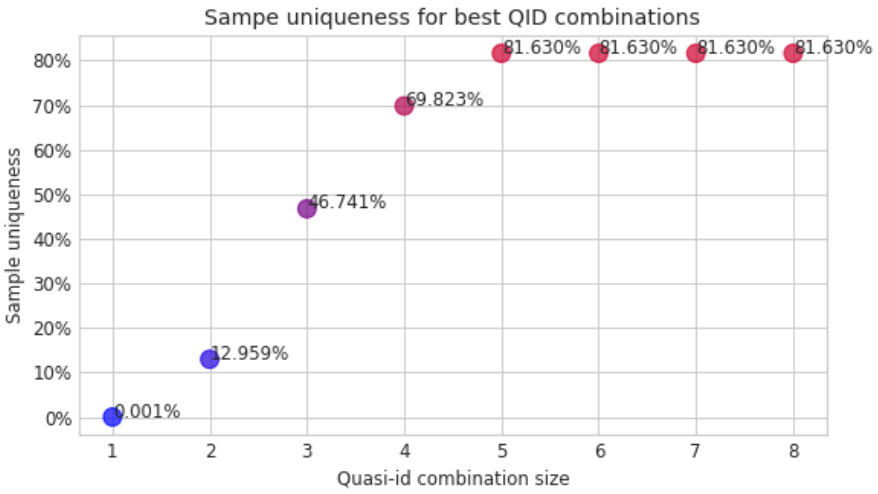

Figura 2. Unicidade na amostra para as melhores combinações de atributos

Podemos observar que, como é de se esperar, quanto mais atributos são usados, maior é a unicidade da amostra. Apesar dessa tendência, a unicidade para de subir depois de usados 5 quasi-identificadores. Isso provavelmente se deve ao tamanho reduzido da amostra e não interfere no prosseguimento dos experimentos.

Conhecendo as combinações de atributos mais suscetíveis à re-identificação pela amostra, resta calcular estimativa de sucesso na população escolar. Para cada conjunto de atributos o modelo de estimativa foi treinado e usado para gerar uma população estimada, do tamanho da população escolar de 2015, cerca de 27 milhões de alunos, [INEP]. Com essa população é feito o cálculo da unicidade. Nesse caso esse valor representa a chance de sucesso de uma re-identificação, visto que indica o número estimado de registros únicos na população para os atributos combinados.

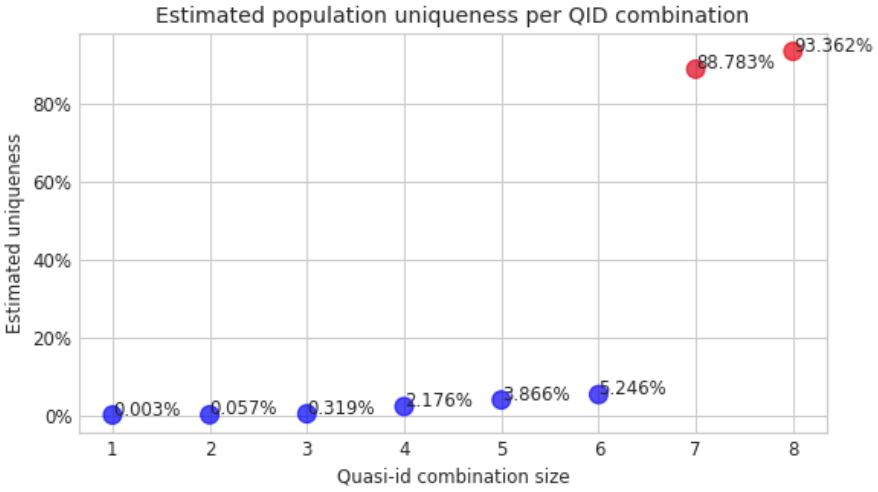

Figura 3. Unicidade estimada na população por combinação de atributos

Tamanho	Conjunto de atributos	Unicidade estimada
1	**escola**	0.003%
2	**VB01004, turma**	0.057%
3	**VB01004, VB01002, turma**	0.319%
4	**VB01004, VB01005, VB01002, turma**	2.176%
5	**VB01001, VB01004, VB01005, VB01002, turma**	3.866%
6	**VB01001, VB01004, VB01005, VB01002, TIPO_MUNIC, turma**	5.246%
7	**VB01001, VB01004, VB01005, VB01002, UFCENSO, escola, turma**	88.783%
8	**VB01001, VB01004, VB01005, VB01002, UFCENSO, TIPO_MUNIC, escola, turma**	93.362%

Os valores de unicidade estimados para a população seguem a tendência observada para a unicidade na amostra, quanto mais atributos são usados maior é a unicidade. É observado um salto no percentual de registros únicos quando o atributo "escola"é re-introduzido ao conjunto de QIDs. Isso poderia indicar overfit no modelo, contudo é coerente que especificamente esse atributo reduza significativamente a generalidade dos registros, apenas um pequeno número de alunos podem estudar na mesma escola.

6. Discussão

Esse trabalho propôs uma avaliação da qualidade da anonimização realizada pelo Instituto Brasileiro de Geografia e Estatística. Usando o modelo estatístico de estimativa de sucesso de identificações em bancos de dados incompletos foi possível analisar uma amostra da pesquisa PeNSE. Os resultados indicam que as informações demográficas publicadas, através de microdados, sobre os estudantes entrevistados são suficientemente específicas para que sejam realizadas re-identificações bem sucedidas em grande parte dos registros da amostra.

Um ataque simples pode ser modelado para usar os valores estimados de unicidade da população para re-identificar estudantes presentes na amostra. Com o conjunto de oito quasi-identificadores do banco de dados pode ser encontrado um aluno x que estava matriculado no ensino básico em 2015 e que corresponda a um registro da amostra. Contanto que o registro correspondente seja único na amostra podemos afirmar que a probabilidade da re-identificação de x estar correta é de pelo menos 93.362%.

Os dados analisados contém diversas informações privadas sensíveis sobre os respondentes. A alta probabilidade de sucesso de re-identificação, portanto, indica que os compromissos com a privacidade dos cidadãos Brasileiros não são atendidos pela anonimização realizada pelo IBGE.

7. Trabalhos futuros

A utilização nesse experimeto apenas de amostras da pesquisa PeNSE impede que a técnica de estimativa tenha sua precisão avaliada. Uma continuação importante desse estudo seria a utilização da metodologia em dados completos sobre a população brasileira.

Isso permitiria a medição de taxas de erro que fortaleceriam os resultados encontrados e garantiria maior confiança às conclusões deste trabalho.

Existem outras formas de anonimizar bancos de dados, como *Instituto Brasileiro de Geografia e Estatística* e *privacidade diferencial*. A metodologia usada nesse estudo permitiria a comparação da eficácia das diversas estratégias de anonimização. Conhecendo a estratégia que melhor protege as informações no contexto de pesquisas estatísticas no Brasil, poderiam ser estabelecidas novas regulamentações que garantam maior privacidade dos cidadãos brasileiros.

Referências

[Hern 2017]HERN, A. *Anonymous'browsing data can be easily exposed, researchers reveal*. [S.l.]: The Guardian, 2017.

[IBGE]IBGE. *PeNSE - Pesquisa Nacional de Saúde do Escolar*. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/educacao/9134-pesquisa-nacional-de-saude-do-escolar.html?=&t=o-que-e>.

[IBGE]IBGE. *Todas as Pesqusas e Estudos*. Disponível em: <https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas.html>.

[IBGE 2013]IBGE. *Princípios Fundamentais das Estatísticas Oficiais*. 2013. Disponível em: <https://www.ibge.gov.br/acesso-informacao/institucional/codigos-e-principios.html?option=com_content&view=article&id=16148>.

[IBGE 2014]IBGE. *Código de Boas Práticas das Estatísticas do IBGE*. 2014. Disponível em: <https://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Cartilha_Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE.pdf>.

[IBGE 2018]IBGE. *Confidencialidade no IBGE*. 2018. Disponível em: <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101636>.

[INEP]INEP. *Censo Escolar 2015*. Disponível em: <https://download.inep.gov.br/educacao_basica/censo_escolar/notas_estatistica/2017/notas_estatisticas_do_censo_escolar_2015_matriculas.pdf>.

[LGPD]LGPD. *Lei Geral de Proteção de Dados Pessoais*. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.html>.

[OCR 2015]OCR. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. U.S. Department of health Human Resources, 2015. Disponível em: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale>.

[Rocher, Hendrick e Montjoye 2019]ROCHER, L.; HENDRICK, J. M.; MONTJOYE, Y.-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Applied and Environmental Microbiology*, v. 10, p. 1–9, 2019.