Análise empírica sobre a supressão dos microdados do Censo Escolar

Contexto

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) divulgou, no dia 18 de fevereiro, os dados detalhados do Censo Escolar 2021 e do Enem 2020. Sob a justificativa de atender à Lei Nº 13.709, de 14 de agosto de 2018, a Lei Geral de Proteção de Dados Pessoais (LGPD), o órgão realizou mudanças importantes nas bases de dados. No caso do Enem, removeu as variáveis que traziam informações sobre as escolas e os municípios dos participantes. Em relação ao Censo Escolar 2021, divulgou uma espécie de "sinopse estatística expandida", excluindo informações no nível do aluno e dos docentes (não pode ser chamada de microdados). Além disso, as bases de dados de edições anteriores do Censo Escolar foram removidas do site do Inep.

Há, portanto, duas situações distintas impostas pelo Inep: no caso do Enem 2020, trata-se de adequação à LGPD, que foi entendida e aplicada de forma bastante restritiva. No caso do Censo Escolar 2021, mais grave, trata-se da supressão de dados: não foram divulgados os microdados da edição 2021, como também foram retiradas do site as bases anteriores do Censo.

Neste documento, em específico, nos atemos ao Censo Escolar. Propomos um exercício com a base de dados 2020 mostrando que, com a realização de ajustes em algumas variáveis e a exclusão de (poucas) outras, é possível pseudoanonimizar os dados de todos os estudantes. Segundo a LGPD (art.13, § 4º) a pseudonimização é o "tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro".

Mas a LGPD obriga a pseudononimização dos dados de todos os estudantes para que possam ser utilizados por institutos de pesquisa?

Não, de forma alguma. Em seu 7º artigo, IV, a LGPD traz que, o tratamento de dados pessoais poderá ser realizado para: "a realização de estudos por órgão de pesquisa, garantida, **sempre que possível**, a anonimização dos dados pessoais". Aqui, uma conclusão lógica precisa ser ressaltada: a lei fala em "sempre que possível" e não em "obrigatoriedade". Portanto, caso a anonimização não seja viável, esta não é um requisito essencial à continuidade de estudos e pesquisas.

O que pretendemos com o exercício que é descrito a seguir?

A intenção é mostrar, empiricamente, que mesmo uma interpretação extrema — e equivocada — da LGPD não inviabilizaria a divulgação dos microdados. Há muitas opções, a partir do ajuste em variáveis, de reduzir drasticamente ou mesmo zerar a possibilidade de re-identificação dos titulares dos dados (estudantes, professores e diretores, no caso concreto).

Reitera-se: a proteção aos dados pessoais é a justificativa do Inep para a não divulgação de dados fundamentais à sociedade. O órgão cita, em nota de esclarecimento à sociedade, publicada em seu

portal no dia 22/3, estudo de pesquisadores da Universidade Federal de Minas Gerais (UFMG), que analisaram o Censo Escolar 2019. Segundo eles, "o uso de três identificadores (mês, ano de nascimento e código da escola em que estuda) permite a identificação com probabilidade de acertode até 29,64%. Se usados quatro identificadores, a chance de sucesso aumenta para 49,86% e, com o uso de todos os dez identificadores, o risco é elevado para 75,51%".

O exercício feito pelo Iede, e descrito a seguir, pseudoanonimizou os dados de 99,99% dos estudantes do Censo Escolar, e com ressalva de que outras modificações ainda são possíveis de serem feitas para que este número chegue a 100%.

Metodologia

Optou-se por trabalhar com a base de matrículas dos estudantes do Censo Escolar 2020. O primeiro passo foi analisar as variáveis da base e identificar aquelas que poderiam ajudar na identificação dos estudantes. Estas foram classificadas em três grupos: **variáveis que poderiam ser adaptadas, variáveis que poderiam ser no máximo mascaradas** (são fundamentais à pesquisa) e **variáveis que poderiam ser excluídas** (facilitam a identificação dos estudantes, ao passo que não são as imprescindíveis à maioria das pesquisas).

O primeiro ajuste foi em relação à data de nascimento do aluno. Existem quatro variáveis relacionadas ao aniversário do estudante: o mês de nascimento, o ano de nascimento, a idade de referência (idade do aluno no mês de referência em que ocorre o Censo Escolar, 31 de maio), e a idade calculada pelo ano de nascimento do aluno. Criou-se duas variáveis com diferentes datas de referência, além da já existente no Censo: **a idade em abril** (considerando que o ingresso dos alunos no Ensino Fundamental se baseia na idade em 31 de março), e **a idade em julho** (dado que a idade de referência das estimativas populacionais é 1 de julho). Isso permitiu a exclusão da variável mês de nascimento.

Outra adequação foi a recodificação das variáveis relacionadas à nacionalidade do aluno (aqui já algo mais questionável, pois prejudicará um grupo de pesquisadores). Atualmente, o Censo conta com uma variável que assume três valores: 1. caso o aluno seja brasileiro; 2. quando é brasileiro nascido no exterior ou naturalizado e 3. referente a alunos estrangeiros. Para garantir a não identificação do indivíduo pela nacionalidade, agregou-se o 2 (alunos nascidos no exterior) ao 1 (somente alunos brasileiros). A variável passou a ter somente os valores 1 e 3, sendo 1 alunos brasileiros e 3 alunos estrangeiros.

A variável que informa o local onde o aluno recebe escolarização também sofreu alteração. Essa variável possui o valor 1 quando o aluno estuda no hospital; valor 2, quando o aluno recebe escolarização em domicílio; e o valor 3, quando ele não recebe escolarização fora da escola. Neste caso, foram agregados os valores 1 e 2. Logo, o valor 1 passou a significar que o aluno recebe educação no hospital ou em domicílio, e o valor 3 significa que ele não recebe escolarização fora do ambiente escolar. Isto também poderá afetar o olhar de algumas pesquisas.

Além dessas alterações, foram **excluídas** algumas variáveis do banco de dados: o **mês de nascimento** do aluno e os **códigos do país e do município de nascimento do aluno**. Também foi adotada uma metodologia que mascara os dados originais de alunos, escolas e municípios, seguida da exclusão desses. Por exemplo: para a variável de Identificação do aluno (id_aluno) foi gerado um número aleatório

que substituiu o número original. Cabe destacar aqui que **seria importante, no caso de uma decisão do gênero, passar a utilizar o mesmo código para demais bases ajustadas à LGPD, para não inviabilizar estudos longitudinais**. Para os códigos de matrícula e de identidade da escola e do município da escola, foi realizado o mesmo procedimento: gerou-se um número aleatório e exclui-se o código original. Cabe destacar aqui que **só seria cogitável essa política de criar máscaras para os códigos se houvesse uma política para garantir fácil acesso às máscaras** por gestores, jornalistas e pesquisadores, entre outros grupos que possivelmente precisarão destes dados.

Resultados

Esse exercício foi feito para a base de matriculados do Centro-Oeste. Num universo de 3.659.818 alunos, apenas 1.197 mostraram-se com "informações únicas" em sua região, considerando as variáveis cor/raça, idade, ser brasileiro ou estrangeiro e a mesorregião de sua escola.

Certas condições podem facilitar a identificação, como ser um estudante de uma idade mais avançada. Por isso, um exercício posterior foi agrupar os estudantes com 20 anos ou mais. Além disso, não foram consideradas como pessoas identificáveis aquelas que não declararam sua cor/raça. Após todos esses cuidados, de um universo de 3,7 milhões de alunos, chegou-se a apenas 224 matriculados que possuem "informações únicas" em suas mesorregiões, o que representa 0,006% do total. A tabela 1, abaixo, indica que dos indivíduos que, mediante o cruzamento de diversas variáveis, poderiam eventualmente ser identificados, há 185 estrangeiros e 39 brasileiros. Desses, 19 se autodeclaram indígenas e 11 de cor/raça amarela.

Tabela 1: o perfil dos 224 matriculados identificáveis

O perfil dos 224 matriculados indentificáveis			
	Brasileiro	Estrangeiro	Total
Amarela	11	48	59
Branca	3	41	44
Indígena	19	12	31
Parda	4	38	42
Preta	2	46	48
Total	39	185	224

Entende-se que a LGPD é importante e que a reflexão sobre a segurança dos dados pessoais nas bases de órgãos como Inep é algo relevante e que deveria acontecer com mais constância. A partir de um relatório de riscos (algo que não foi feito), e que demonstre, por exemplo, a possibilidade de danos aos titulares caso a proficiência de um aluno ou a sua cor/raça venha a conhecimento público, pode-se pensar em ajustes na base de dados para mitigar esses riscos. Uma opção, como a demonstrada no exercício, é alterar ou excluir algumas variáveis, que trariam pouco prejuízo aos estudos quantitativos na área.

Há muitas possibilidades. O que não é admissível, sob hipótese alguma, é a supressão do acesso a dados essenciais ao debate público, à realização de pesquisas e à criação e ao acompanhamento de políticas públicas.