risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 02/09/2020
Aceitação/Acceptance: 08/11/2020

Garantia de Privacidade Versus Utilidade dos Dados em Anonimização: um estudo no ensino superior

Paula Prata[1, 2], Maria Eugénia Ferrão[1, 3], Wilson Santos[1, 2], Gonçalo Sousa[1]

pprata@di.ubi.pt; meferrao@gmail.com

[1] Universidade da Beira Interior, Covilhã, Portugal

[2] Instituto de Telecomunicações (IT-UBI), Covilhã, Portugal

[3] REM - Research in Economics and Mathematics, CEMAPRE

DOI: 10.17013/risti.40.112–127

Resumo: No mundo digital, toda a atividade humana deixa um rasto de dados que constitui um recurso cada vez mais valioso, para avaliação e definição de estratégias nos mais variados domínios. A partilha desses dados, sendo socialmente importante, implica o respeito pela privacidade individual e portanto a sua anonimização. As atuais leis e regulamentos sobre privacidade oferecem orientações limitadas para lidar com um vasto leque de tipos de dados, ou com técnicas de reidentificação. Este trabalho pretende ilustrar um processo de anonimização, comparando para vários modelos de privacidade a perda de informação e a utilidade do conjunto de dados resultante. Encontrar o equilíbrio entre privacidade e utilidade é um desafio que pode ser mais facilmente alcançado por quem melhor conhece o significado dos dados e dos objetivos que se pretendem alcançar com eles.

Palavras-chave: Anonimização de dados, k-anonimato; ℓ-diversidade; t-proximidade; ENADE.

Privacy Preserving Versus Utility Preserving in Data Anonymization: a study in higher education

Abstract: In the digital world, all human activity leaves a trace of data that is growingly valued for the evaluation and definition of strategies in varied domains. The sharing of those data, being socially relevant, implies the respect for individual privacy and so, its anonymization. The current laws and regulations about privacy offer limited guidance to deal with the vast range of datatypes or with techniques of re-identification. This work aims at illustrating a process of anonymization, comparing to several models of privacy, the loss of information and the usefulness of that dataset resulting from the anonymization. Finding a balance between privacy and utility is a challenge that can be more easily found by those who know better the meaning of the data and objectives aimed at.

Keywords: Data anonymization; k-anonymity; ℓ-diversity; t-closeness; ENADE.

1. Introdução

Nos dias de hoje, a quantidade de dados sobre a atividade humana que é recolhida e armazenada digitalmente está em constante crescimento. Esses dados podem passar por todos os aspectos da nossa vida, como, por exemplo, a atividade nas redes sociais, rastos de localização recolhidos por telefones móveis, compras *online*, ou registos médicos. Transformar esses dados em conhecimento é uma mais-valia que tem tornado os dados num recurso cada vez mais valioso. O processamento e análise de dados possibilitam avanços socialmente importantes, em campos tão diversos como sistemas de suporte à decisão médica, criminologia computacional, protecção contra terrorismo informático ou marketing direccionado. Todos estes aspectos há muito idealizados (Chen et al., 2012; Adomavicius & Tuzhilin, 2005; Quiñonez et al., 2019) são, cada vez mais, possíveis devido à transversal digitalização da sociedade. O crescente interesse das mais variadas organizações em terem acesso aos nossos dados pode ser traduzido pela frase de Prasser et al. (2020) "The race for innovation has turned into a race for data" (p. 1277). No entanto, todo este potencial de análise de dados tem um custo associado. Os dados recolhidos, incluindo informação sensível, podem ser publicados e partilhados com entidades externas que as poderão usar para fins não previstos originalmente. Existe uma panóplia de riscos associados à partilha de dados pessoais, em especial se esses dados foram posteriormente associados com outras fontes, podendo a divulgação de dados pessoais sensíveis causar danos graves aos indivíduos em causa. Para evitar esses riscos, têm sido criados regulamentos de protecção de dados visando aumentar a garantia de protecção dos dados pessoais, (Directive 95, 1995) assim como existe inúmera investigação sobre os aspectos éticos, legais e sociais da partilha de dados (Kaye et al., 2012; Cambon-Thomsen, 2007). Em particular, com a entrada em vigor do Regulamento Geral sobre a Proteção de Dados, RGPD (GDPR, 2016), este tema está na ordem do dia e tem levado à consciencialização da sociedade para o problema da privacidade dos dados.

Vários exemplos de violação da privacidade têm sido descritos na literatura, como o conhecido caso do Governador do estado do Massachusetts, USA, William Weld que viu os seus dados médicos divulgados publicamente, quando uma base dados de um sistema de saúde foi tornada pública e os seus registos foram cruzados com dados de um caderno eleitoral que continha dados como "zip code", data de nascimento e género (Barth-Jones, 2012). Cada um destes atributos isolado não permite a identificação de um individuo, mas a sua combinação com outras fontes de dados pode levar a um conjunto mínimo de registos (Sweeney, 2002b). Geralmente, para a reidentificação ser possível, o adversário tem de conhecer *a priori* duas peças de informação: sabe que o registo da vítima está na base de dados e conhece algum atributo quase-identificador. No contexto de anonimização de dados, um adversário é alguém que tenta identificar indivíduos num conjunto de dados, supostamente anonimizado, e um atributo quase-identificador é definido como um atributo que não identifica um indivíduo, mas pode fazê-lo quando associado a outra informação. No caso anterior, o adversário sabia que a vítima tinha estado hospitalizada e os restantes dados foram fáceis de obter (Fung et al., 2010). Este caso teve grande impacto na procura por mecanismos de garantia de privacidade de dados pessoais. Foi demonstrado que 87% da população dos USA pode ser facilmente identificada com apenas três quase-identificadores: "zip code",

género e data de nascimento (Sweeney, 2000). Também o caso relatado em (Panduragan, 2014) mostra que dados supostamente anonimizados podem permitir a reidentificação. O número das licenças de cada táxi de Nova Iorque (composto por sete dígitos) foi anonimizado usando valores de dispersão. Os valores foram facilmente revertidos e informação sensível dos taxistas como percursos efectuados, o seu rendimento, e até a sua morada foram revelados. Mais recentemente, o estudo apresentado em (Sweeney et al., 2018) mostrou ser possível identificar univocamente estudantes de uma escola de Direito cujos dados tinham sido anonimizados de forma independente por 4 protocolos, correntemente usados. Muitos outros exemplos mostram quão importante e difícil é efectuar uma correta anonimização, assim como perceber os riscos associados à segurança dos nossos dados (Sweeney, 2015; Culnane et al., 2017; Koch, 2020).

Num processo de anonimização de dados pessoais, um aspeto, tão importante como garantir a privacidade de cada individuo, é garantir que os dados resultantes continuam a ter utilidade. Anonimizar significa retirar algumas características dos dados, e portanto, informação útil para os seus utilizadores pode ser perdida. Anonimizar deve ser um processo iterativo, em que a cada aplicação de um modelo de privacidade, e consequente avaliação do risco de reidentificação, se deve seguir a avaliação da utilidade dos dados obtidos. Todo o processo deve ser repetido, até se alcançar um equilíbrio razoável entre minimizar o risco de reidentificação e manter o máximo de utilidade dos dados (Prasser et al., 2020). Esta última pode ser avaliada pelo cálculo de uma simples proporção dos dados perdidos ou por métodos estatísticos, mais sofisticados, que indiquem em que medida as características dos dados anonimizados se distanciam dos dados originais. Todo o processo de anonimização depende do tipo de dados e do uso dos dados (Francis, 2018) ou propósito da análise de dados.

Neste trabalho, foram estudados, para um subconjunto dos dados públicos do ENADE - Exame Nacional de Desempenho do Estudantes de graduação do Brasil, vários processos de anonimização, comparando os resultados em termos de risco de reidentificação e de utilidade dos dados. Usando uma ferramenta de código aberto, foram aplicados dois modelos de privacidade, ℓ-diversidade e t-proximidade, considerando várias parametrizações, foi avaliado o risco de reidentificação associado e foi avaliada a utilidade dos dados resultantes, através de um modelo de análise de variância com múltiplos fatores principais e interacções de 2ª ordem. A partilha de dados pode trazer vários benefícios à sociedade, seja para avanços científicos, avaliação de políticas ou para melhoria de serviços. Este artigo contribui para a reflexão sobre o *trade-off* entre privacidade e utilidade dos dados. Quando os dados são provenientes de registos administrativos ou de órgãos governamentais, com grande potencial para fins de investigação científica, aspetos normativos e outros decorrentes da aplicação do RGPD podem inviabilizar ou até distorcer os fins da investigação científica. Adicionalmente, constitui uma abordagem exploratória de interesse para investigadores ou organizações que pretendam anonimizar os seus dados, tirando partido do elevado conhecimento do contexto e significado dos dados, e tornando o processo de anonimização tecnicamente explícito. Deste modo, o artigo contribui também para a adoção de práticas informadas e justificadas no processo de anonimização sem, contudo, por em causa os aspetos legais de privacidade impostos pelo RGPD.

Na secção 2 são descritos os modelos de privacidade utilizados, assim como o modelo que está na sua base, o modelo de k-anonimato. A secção 3 apresenta o modelo de utilidade escolhido para o propósito deste trabalho, isto é, o modelo de análise de variância com múltiplos fatores (ANOVA) e a secção 4 refere trabalho relacionado. A secção 5 contém o estudo experimental em três subsecções: descrição dos dados e do seu pré-processamento; a análise de privacidade e discussão dos resultados; a análise de utilidade e discussão dos resultados. Finalmente, a secção 6 apresenta as conclusões.

2. Modelos de Privacidade

As duas principais abordagens de anonimização são a aleatorização e a generalização. A aleatorização consiste em alterar os dados de forma a reduzir a possibilidade de associação entre os dados e o indivíduo. Uma técnica é, por exemplo, a adição de ruído aleatório a algumas variáveis, como proposto em Goldstein e Shlomo (2020). A generalização ou agregação consiste na junção de categorias ou classes de variáveis através de alteração da escala ou ordem de grandeza. Neste trabalho, vamos explorar dois modelos de privacidade baseados em generalização: ℓ-diversidade e t-proximidade. Estes dois modelos são evoluções de um modelo mais simples de privacidade que é o k-anonimato. Os três modelos vão ser descritos nas próximas subsecções. Ao aplicar um modelo de privacidade, pretende-se: reduzir o risco de identificação, isto é, evitar que um indivíduo seja associado a um registo específico; reduzir o risco de ligação, isto é, reduzir a possibilidade de associar dois registos do mesmo indivíduo quer estejam na mesma ou em diferentes bases de dados; reduzir o risco de inferência, isto é, não permitir que, após a anonimização, seja possível deduzir o valor de um atributo a partir dos valores de outros atributos de um dado individuo. Para avaliar o risco de reidentificação, são comuns três abordagens diferenciadas pelo que é suposto o possível adversário conhecer sobre os dados (Prasser & Kohlmayer, 2015; Kniola, 2017): modelo de promotor, em que se supõe que o adversário sabe que o indivíduo que procura está na base de dados; modelo de jornalista, em que o adversário desconhece se o indivíduo está na base de dados; modelo de marketing, em que o adversário quer identificar o maior número de indivíduos possível.

2.1. O Modelo k-anonimato

Um processo de anonimização começa por classificar os atributos do conjunto de dados. Atributos que permitam identificar directamente um indivíduo, como nome ou número de cartão de cidadão, são classificados como identificadores diretos. Atributos que não identificam um indivíduo diretamente, mas que permitam a associação com outros conjuntos de dados, são quase-identificadores. Os restantes atributos podem ainda ser classificados como sensíveis ou não sensíveis. Um atributo é sensível se o seu valor não deve ser descoberto por qualquer adversário, para nenhum indivíduo do conjunto de dados, caso contrário, o atributo será classificado como não sensível. Após a classificação dos atributos, é necessário suprimir ou modificar os atributos diretos. Como vimos nos exemplos apresentados na introdução, isso não é suficiente para evitar a reidentificação. Através de atributos quase-identificadores, é possível ligar os registos com outras bases

de dados e identificar indivíduos no conjunto de dados. Para evitar esse risco de ligação, foi proposto o modelo de privacidade k-anonimato (Sweeney, 2002a). Um conjunto de dados é k-anónimo, se cada registo é indistinguível de pelo menos k-1 outros registos, no que diz respeito aos atributos quase-identificadores. Formalmente, k-anonimato é definido da seguinte forma: "Seja a tabela RT(A1, ..., An), e QI_{RT} os quase-identificadores associados a essa tabela. RT satisfaz k-anonimização em relação a QI_{RT} se e só se cada sequência de valores em RT[QI_{RT}] tem no mínimo k ocorrências em RT[QI_{RT}]" (Sweeney, 2002a, p. 564). Para evitar que um indivíduo possa ser univocamente identificado através de ligação a outros conjuntos de dados, o modelo assegura que, para cada combinação dos seus atributos quase-identificadores, existem pelo menos k registos que partilham os mesmos valores. Registos que não verificam esta condição são eliminados.

Foram desenvolvidos inúmeros algoritmos que implementam o k-anonimato, como por exemplo, Datafly (Sweeney, 2002a), Incognito (LeFevre et al., 2005) e Mondrian (LeFevre et al., 2006). Segundo Ayala-Rivera et al. (2014) não existe um algoritmo melhor do que os outros. O melhor algoritmo em cada situação é influenciado por múltiplos fatores, como por exemplo o número de quase-identificadores, ou a distribuição dos dados na base de dados.

2.2. O Modelo ℓ-diversidade

O principal problema do modelo de k-anonimato é permitir a divulgação de informação, devido à falta de diversidade num ou vários atributos sensíveis. Se tivermos um conjunto de k registos, todos com os mesmos valores nos atributos quase-identificadores, e ocorrer que todos eles tenham um mesmo valor para um atributo sensível, então qualquer adversário que conheça um indivíduo que corresponda aos valores dos quase-identificadores irá poder inferir o valor do atributo sensível para esse indivíduo. Diz-se que esse conjunto de registos indistinguíveis constitui uma classe de equivalência. O modelo de privacidade ℓ-diversidade melhora o modelo de k-anonimato, reduzindo o risco de inferência de atributos, ao garantir que cada atributo sensível tem pelo menos ℓ valores distintos representados em cada classe de equivalência. Formalmente, considerando um bloco q que seja uma classe de equivalência relativa aos atributos quase-identificadores considerados, esse bloco q é ℓ-diverso se contém pelo menos ℓ valores distintos para os atributos sensíveis S. Uma tabela é ℓ-diversa se cada bloco q é ℓ-diverso (Machanavajjhala et al., 2007, p. 16). O modelo impõe assim que todos os registos que partilhem os mesmos quase-identificadores devem ter diversos valores para os atributos sensíveis. Existem diversas abordagens que tentam formalizar essa diversidade. A definição de (c, ℓ)-diversidade recursiva garante que o valor mais comum não apareça com demasiada frequência enquanto que os valores menos comuns não aparecem muito raramente. A definição formal é a seguinte: dado um bloco q, seja $r1$ o número de vezes que o valor do atributo sensível mais frequente aparece nesse bloco q; $r2$ será o número de vezes que o segundo valor mais frequente aparece e assim por diante até rm para um atributo sensível que tenha m valores possíveis. Dada uma constante c, o bloco q satisfaz (c, ℓ)–diversidade recursiva se $r1 < c(rℓ + rℓ+1, + ... + rm)$. A tabela T é ($c$, ℓ)–diversa recursiva se cada bloco q satisfaz (c, ℓ)-diversidade recursiva. Para ℓ = 1 a diversidade é sempre verificada (Machanavajjhala et al., 2007, p. 18).

2.3. O Modelo t-proximidade

O modelo de t-proximidade é um melhoramento da ℓ-diversidade, na medida em que tenta obter classes de equivalência com uma distribuição dos valores dos atributos sensíveis próxima da sua distribuição no conjunto original de dados. Segundo Li et al. (2007, p. 109), uma classe de equivalência é dita como tendo t-proximidade, se a distância entre a distribuição de um atributo sensível nessa classe e a distribuição do atributo em toda a tabela não é mais do que um valor limite t. A tabela é dita como tendo t-proximidade se todas as classes de equivalência têm t-proximidade. Para medir a distância entre as duas distribuições é proposto o uso da métrica *Earth Mover's Distance* (Rubner et al., 2000).

3. Modelo de Utilidade: Análise de variância

Para o propósito deste artigo usámos o modelo ANOVA com fatores principais e interação de 2ª ordem entre os fatores. Apresentamos a especificação do modelo com dois fatores e respetiva interação, podendo ser generalizado, através de termos aditivos, ao número de fatores e interações referentes à análise em causa. Considerando uma amostra de tamanho n $(i=1, .., n)$, a equação do modelo é a seguinte: onde denota a classificação final do i-ésimo estudante que pertence ao grupo p do fator e também pertence ao grupo k do fator δ. Ou seja, representa o primeiro fator, representa o segundo fator e refere-se ao efeito de interação entre os dois fatores, $p=1,..,P;\ k=1,...,K$. Decorre que o fator tem P grupos, o fator δ tem K grupos e há PK subgrupos de interação. O termo aleatório do modelo é representado por , com os seguintes pressupostos: distribuição normal com média nula, homocedasticidade ou homogeneidade das variâncias, elementos independentes entre si. Para mais detalhes sobre o modelo ver, por exemplo, Scheffé (1999).

4. Trabalho relacionado

A maioria dos trabalhos experimentais sobre anonimização de dados lida com dados médicos que, pela sua natureza, contêm informação sensível. Em Spengler e Prasser (2019) uma base de dados biomédicos é usada para avaliar o risco e a utilidade dos dados anonimizados usando os modelos de ℓ-diversidade, t-proximidade e β-semelhança. Também para dados médicos, em Lee et al. (2017) é apresentado um modelo de preservação da utilidade e da privacidade baseado em k-anonimização e "h-ceiling" um método que limita a generalização de dados. Na área da educação, Chicaiza et al. (2020) apresenta um estudo sobre análise de dados de aprendizagem usando k-anonimato e modelos de regressão linear para avaliar a utilidade dos dados. Em Santos et al. (2020) a utilidade de dados educacionais k-anonimizados é analisada calculando estatísticas descritivas para vários valores de k. Estudos recentes introduzem modelos de aprendizagem automática para garantir a privacidade dos dados e avaliar a sua utilidade (Eicher et al., 2020; Esquivel-Quirós et al., 2019).

5. Estudo Experimental

Na componente experimental, que descrevemos de seguida, foram usados os dados do Exame Nacional de Desempenho dos Estudantes de graduação no Brasil (ENADE)

disponíveis em *http://portal.inep.gov.br/enade*. Na anonimização dos dados foi usada a *framework* de código aberto, ARX (*https://arx.deidentifier.org/)* e para o estudo de utilidade foi usado o *software* estatístico SPSS.

5.1. Conjunto de Dados

Foram considerados para análise os dados do ENADE de 2018, no qual estiveram envolvidos 548 127 estudantes. O grande volume de registos, mais de meio milhão, pode dar uma falsa sensação de segurança, transmitindo a ideia de que registos únicos são raros, mas uma simples k-anonimização do subconjunto de dados apresentados na Tabela 1, para k=2 mostrou um número de registos únicos muito elevado. Apesar de os dados não conterem identificadores diretos, possuem quase-identificadores que poderão permitir a inferência de dados sensíveis ou ainda a associação a registos de outras bases de dados com possível reidentificação, o que justifica o estudo de anonimização realizado. Os atributos seleccionados foram o código da área do curso, região onde funcionou o curso, idade, género, raça/cor e média final do estudante, os níveis de educação da Mãe e do Pai e o rendimento do agregado familiar. Foi ainda calculado o número de anos entre terminar o ensino secundário e iniciar o curso superior, que designámos por "espera ingresso", e foi calculado o número de anos para concluir a graduação, "tempo diploma". A Tabela 1 mostra os nomes das variáveis usadas, a sua descrição e como foram classificadas para efeitos de anonimização.

Variável	Descrição	Classificação
Código Curso	Código da área de enquadramento do curso	Quase-identificador
Região	Código de região de funcionamento do curso	Quase-identificador
Idade	Generalizada nas categorias: [4,26[e [26,95[	Quase-identificador
Género	M ou F	Quase-identificador
Média Final	Média da classificação final obtida pelo estudante	Não sensível
Espera Ingresso	*Anos entre terminar secundário e início superior*	*Quase-identificador*
Tempo Diploma	*Tempo para obtenção do diploma*	*Quase-identificador*
Raça Cor	*Auto declaração*	*Quase-identificador*
Educação Pai	*Generalizada nas categorias: [A,B] [C,D] [E,F]*	*Quase-identificador*
Educação Mãe	*Generalizada nas categorias: [A,B] [C,D] [E,F]*	*Quase-identificador*
Rendimento Familiar	*Número de salários mínimos do agregado familiar*	*Sensível*

Tabela 1 – Variáveis seleccionadas e respectiva classificação.

Os dados resultantes foram pré-processados, tendo sido removidos registos com valores pouco plausíveis, como, por exemplo, registos em que o ano em que terminavam o ensino superior era inferior a 2018, ou ainda registos cujo valor calculado para o "tempo diploma" dava negativo. O conjunto resultante ficou com 536 466 registos. De seguida, foram generalizadas três variáveis: idade, educação da Mãe e educação do Pai. Os valores da idade foram recodificados em menor de 26 ou maior e igual que 26. Os níveis de educação do Pai e da Mãe foram generalizados em 3 categorias em vez das 6 originais.

O dicionário de dados completo pode ser consultado no *site* do ENADE. Finalmente, o atributo rendimento familiar foi classificado como sensível, a média final como não sensível e todos os restantes atributos foram classificados como quase-identificadores.

5.2. Análise de Privacidade

Os dados resultantes do pré-processamento foram anonimizados com (*c, ℓ*) - diversidade recursiva e com t-proximidade, fazendo variar os valores de *c, ℓ e t*. Para cada uma das parametrizações foi quantificada a percentagem de registos eliminados e foi calculado o risco máximo e o risco médio de reidentificação usando o modelo do prossecutor implementado no ARX.

5.2.1. Anonimização por ℓ-diversidade

A Tabela 2 apresenta os resultados da anonimização por (*c, ℓ)*-diversidade, fazendo variar o valor de ℓ de 2 a 5 para um valor de *c* = 3. Para cada conjunto anonimizado obtido, apresenta-se o número de registos (dimensão), a percentagem de registos eliminados, o risco médio e máximo de reidentificação. Como se pode observar, ao aumentar o valor de ℓ e portanto ao aumentar o número de registos de cada classe de equivalência a percentagem de registos eliminados aumenta drasticamente, subindo de 34,08% para ℓ = 2 até 82,85% para ℓ = 5. Por outro lado, o risco médio reduz gradualmente de 13,27% para 2,78%. Em relação ao risco máximo de reidentificação, ele será de 100/ℓ uma vez que os registos são agrupados em grupos de ℓ registos com valores iguais para os quase-identificadores. O atributo sensível que está a ser diversificado é o rendimento familiar.

(3, ℓ) - diversidade	**(3,2)**	**(3,3)**	**(3,4)**	**(3,5)**
N	353 637	264 634	171 107	91 991
Registos eliminados (%)	34,08%	50,67%	68,10%	82,85%
Risco médio (prossecutor)	13,27%	7,52%	4,63%	2,78%
Risco máximo	50%	33.3%	25%	20%

Tabela 2 – Dimensão (*N*) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após (*c, ℓ*)-diversidade, para ℓ a variar de 2 a 5, com *c*=3.

A Tabela 3, apresenta os mesmos valores mas agora para os dados resultantes de (*c, ℓ*)-diversidade fixando o valor de ℓ em 5, e fazendo variar o valor de *c* de 2 a 4. Aumentar o valor de *c*, significa aumentar o número de vezes que o valor do atributo sensível mais frequente pode ocorrer em cada classe de equivalência (ver Secção 2.2). Como se pode observar, a percentagem de registos eliminados diminui de 89,39% para 78,76% quando c aumenta de 2 para 4. Em relação ao risco, este aumenta ligeiramente quando *c* aumenta, no entanto esse resultado resulta apenas do aumento do número de registos. A avaliação do risco pelo modelo do prossecutor implementada no ARX apenas mede o risco de reidentificação e não o risco de inferência do atributo sensível. A avaliação do risco de inferência do valor do atributo sensível virá a ser tratada num próximo trabalho. Podemos no entanto afirmar que ao introduzirmos a diversidade, o risco de inferência diminui.

(*c*, *5*) - diversidade	(2,5)	(3,5)	(4,5)
N	56 935	91 991	113 966
Registos eliminados (%)	89,39%	82,85%	78,76%
Risco médio (prossecutor)	2,24%	2,78%	3,12%
Risco máximo	20%	20%	20%

Tabela 3 – Dimensão (*N*) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após (*c*, ℓ)-diversidade, para ℓ = 5 e *c* a variar de 2 a 4.

5.2.2. Anonimização por t-proximidade

Para estudar o modelo de t-proximidade, começamos por definir uma dimensão *k* para as classes de equivalência. O valor de *t* determina a distância entre a distribuição dos valores do atributo sensível nessas classes de equivalência e a distribuição no conjunto original. A Tabela 4 apresenta os resultados para os conjuntos de dados produzidos para k=2 e k=5 fazendo t=0,15 e t=0,3. v

t-proximidade	k=2, t=0,3	k=2, t=0,15	k=5, t=0,3	k=5, t=0,15
N	348 519	231 645	259 190	195 235
Registos eliminados (%)	35,03%	56,82%	51,69%	63,60%
Risco médio (prossecutor)	14,65%	10,77%	6,20%	5,78%
Risco máximo	50%	50%	20%	20%

Tabela 4 – Dimensão (*N*) do conjunto de dados, percentagem de registos eliminados, risco médio e máximo de reidentificação após t-proximidade (k = 2 e k =5 com t= 0,15 e t= 0,3).

Podemos observar que para um mesmo valor de *t*, a percentagem de registos eliminados aumenta quando *k* aumenta, como seria de esperar. Para o mesmo *k*, a percentagem de registos eliminados diminui quando *t* aumenta. Se exigimos maior proximidade na distribuição dos valores sensíveis, obtemos menos registos. Comparando os resultados de t-proximidade com os obtidos por diversidade, para conjuntos com a mesma dimensão das classes de equivalência, isto é, quando ℓ é igual ao k, podemos observar o seguinte: para k=2, (3, 2)-diversidade tem menos registos eliminados (34,08%) que qualquer dos conjuntos obtidos por proximidade 35,03% para t=0,3 e 56,82% para t=0,15; no entanto para k=5, a diversidade elimina entre 78 a 89% dos registos, enquanto a proximidade elimina no máximo 63,6% para t=0,15. Na próxima secção, iremos fazer a análise de utilidade para o conjunto obtido por (3, 5)-diversidade e para os casos de t-proximidade em que a dimensão das classes de equivalência é igual à do caso anterior, k=5, com t=0,15 e t=0,3. O conjunto obtido por diversidade tem um risco médio de reidentificação baixo (2.78%) e o atributo sensível tem bastante diversidade, no entanto, isso ocorre à custa da supressão de mais de 80% dos registos. Os conjuntos obtidos por proximidade perderem respectivamente cerca de 64% e 52% dos registos originais.

5.3. Análise de Utilidade

O modelo ANOVA foi aplicado aos dados ENADE descritos e ajustado considerando como variável dependente a média final e as restantes variáveis como fatores. A versão 24 do SPSS apresentou problemas de execução com elevado número de variáveis em particular quando cada uma delas tem diversas categorias tal como código do curso. O processador usado foi um Intel(R) Core(TM) i3-7100U CPU @ 2.40GHz com 8 GB de RAM. Esta limitação foi ultrapassada através da selecção de variáveis. Foram considerados 5 fatores: região, idade, género, raça/ cor, educação do Pai e educação da Mãe.

As Tabelas de 5 a 8 apresentam os resultados da estatística de teste F e valor de prova, respectivamente para os dados originais, os dados anonimizados através do modelo de privacidade ℓ-diversidade (com c=3 e ℓ =5) e para os dados anonimizados através do modelo de privacidade t-proximidade com k=5 e t=0,15 e t=0,3. Os testes de hipóteses consideram, sob H0, que cada um dos fatores e cada um dos termos de interação são iguais a zero.

Através da análise efetuada à Tabela 5, verificamos que, com excepção do termo principal associado ao fator região, todos os demais termos principais e termos de interacção são estatisticamente significativos ao nível de significância de 5% (valor p < 0,05). Ou seja, de acordo com tais resultados e em presença de todos os termos aditivos, só não é possível rejeitar a hipótese nula para o efeito principal de região. Apesar disso, os termos de interação entre região e idade, região e educação do pai e da mãe, região e sexo, região e raça/cor autodeclarada constituem-se como grupos diferenciadores na sua relação com a variável dependente média final obtida pelo/a estudante. Notamos, adicionalmente, que a maioria dos termos é estatisticamente significativa ao nível de 1%. No entanto, após a anonimização (3, 5)-diversidade, para o mesmo nível de significância, a maior parte das variáveis deixa de ter impacto direto na explicação da variável dependente (Tabela 6). Apenas os fatores raça/cor autodeclarada e educação da mãe continuam como fator estatisticamente diferente de zero, na associação à média final obtida pela/o estudante. Quanto aos termos aditivos de interacção, os resultados também se modificam com o processo de anonimização. Entre os 15 termos de interação, 5 deixam de ser estatisticamente significativos ao nível de significância de 5%.

De forma diferente acontece com as duas parametrizações do modelo de privacidade t-proximidade (Tabelas 7 e 8). Embora registando alterações relativamente à distribuição original, a explicação das variáveis do preditor linear sobre a variável resposta é em tudo mais idêntica aos dados originais. Ora, isto pode sugerir uma distorção menos drástica dos dados por parte deste procedimento de anonimização. Em detalhe, verificamos que, mesmo em tais cenários de anonimização, os resultados nem sempre confirmam os obtidos com os dados originais. Compare-se a título de exemplo o efeito principal de região, que nas Tabelas 7 e 8 se constitui como fator diferenciador da média final do estudante e o termo de interação entre idade e raça/cor autodeclarada que deixa de ser estatisticamente significativo.

Fonte de variação	F	Valor p
Região	0,930	0,445
Idade	5,400	0,000
Género	8,139	0,004
Raça Cor	10,825	0,000

Fonte de variação	F	Valor p
Educação Pai	9,760	0,000
Educação Mãe	8,819	0,000
*Idade * Educação Pai*	1,190	0,018
*Idade * Género*	3,079	0,000
*Idade * Educação Mãe*	1,344	0,000
*Idade * Raça Cor*	1,550	0,000
*Região * Idade*	1,559	0,000
*Género * Educação Pai*	17,223	0,000
*Educação Pai * Educação Mãe*	22,695	0,000
*Raça Cor * Educação Pai*	1,560	0,037
*Região * Educação Pai*	7,422	0,000
*Género * Educação Mãe*	28,581	0,000
*Género * Raça Cor*	23,235	0,000
*Região * Género*	10,456	0,000
*Raça Cor * Educação Mãe*	3,878	0,000
*Região * Educação Mãe*	3,885	0,000
*Região * Raça Cor*	12,860	0,000

Tabela 5 – ANOVA com termos principais e interação, aplicado aos dados originais.

Fonte de variação	F	Valor p
Região	2,135	0,074
Idade	0,384	0,535
Género	2,633	0,105
Raça Cor	9,621	0,000
Educação Pai	2,825	0,059
Educação Mãe	4,570	0,010
*Idade * Educação Pai*	0,483	0,617
*Idade * Género*	16,059	0,000
*Idade * Educação Mãe*	9,673	0,000
*Idade * Raça Cor*	2,296	0,076
*Região * Idade*	2,961	0,019
*Género * Educação Pai*	5,657	0,003
*Educação Pai * Educação Mãe*	14,162	0,000
*Raça Cor * Educação Pai*	1,129	0,341
*Região * Educação Pai*	1,998	0,043

Fonte de variação	F	Valor p
*Género * Educação Mãe*	5,140	0,006
*Género * Raça Cor*	4,080	0,003
*Região * Género*	2,786	0,025
*Raça Cor * Educação Mãe*	1,887	0,079
*Região * Educação Mãe*	1,980	0,045
*Região * Raça Cor*	0,876	0,597

Tabela 6 – ANOVA com termos principais e interação, (3,5)-diversidade.

Considerando os casos válidos, os pressupostos do modelo de utilidade foram verificados para todos os conjuntos de dados. Apresentamos na Tabela 9 a assimetria, curtose e desvio padrão referentes à distribuição dos dados originais e à distribuição dos dados anonimizados com ℓ -diversidade (3,5). Tais estatísticas são as necessárias para usar o teste Jarque-Bera (Bera & Jarque, 1981; Greene, 2003) segundo o qual a normalidade da distribuição é testada sob H0. Aplicando o teste, em ambos os conjuntos de dados a hipótese nula não é rejeitada ao nível de significância de 5%. A comparação das estatísticas de distribuição para a variável dependente permitem-nos verificar que com o processo de anonimização a distribuição se altera, e.g. a curtose acentua-se. A alteração da distribuição já era esperada uma vez que no processo os casos extremos/raros são suprimidos ou agregados. Para os restantes conjuntos de dados os resultados conduzem a interpretação semelhante.

Fonte de variação	F	Valor p
Região	15,937	0,000
Idade	22,903	0,000
Género	7,801	0,005
Raça Cor	29,824	0,000
Educação Pai	1,869	0,154
Educação Mãe	13,390	0,000
*Idade * Educação Pai*	12,456	0,000
*Idade * Género*	14,365	0,000
*Idade * Educação Mãe*	5,224	0,005
*Idade * Raça Cor*	0,432	0,786
*Região * Idade*	5,879	0,000
*Género * Educação Pai*	2,925	0,054
*Educação Pai * Educação Mãe*	7,959	0,000
*Raça Cor * Educação Pai*	1,545	0,136
*Região * Educação Pai*	2,433	0,013
*Género * Educação Mãe*	7,447	0,001
*Género * Raça Cor*	12,115	0,000

Fonte de variação	F	Valor p
*Região * Género*	4,396	0,001
*Raça Cor * Educação Mãe*	2,465	0,011
*Região * Educação Mãe*	2,905	0,003
*Região * Raça Cor*	10,381	0,000

Tabela 7 – ANOVA com termos principais e interação, t-proximidade (k=5, t=0,15)

6. Conclusões

Este trabalho analisou, para dados reais do sistema de ensino superior Brasileiro, estratégias para alcançar o equilíbrio entre privacidade e utilidade dos dados no processo de anonimização. Para estes dados verificou-se que, com classes de equivalência de dimensão 5, o que já garante um risco baixo de reidentificação, o modelo de t-proximidade pode levar a uma menor perda de registos do que o modelo de ℓ-diversidade recursiva, garantindo maior utilidade dos dados. Os nossos resultados também permitem verificar que os resultados do modelo de utilidade estão condicionados ao desenho do modelo de privacidade e podem tornar-se inúteis ou mesmo falaciosos. Neste caso, é necessário acautelar as possíveis interpretações substantivas e eventuais contribuições ou recomendações de política e prática, pois poderiam produzir efeito no sentido oposto ao que seria desejável. A comparação das estatísticas de distribuição referentes aos diferentes conjuntos de dados também nos permite afirmar que pressupostos teóricos estabelecidos para o modelo de utilidade podem deixar de se verificar após o processo de anonimização, podendo eventualmente comprometer a inferência estatística e a tomada de decisão subsequente.

Fonte de variação	F	Valor p
Região	21,292	0,000
Idade	23,688	0,000
Género	12,181	0,000
Raça Cor	83,183	0,000
Educação Pai	7,913	0,000
Educação Mãe	25,070	0,000
*Idade * Educação Pai*	12,076	0,000
*Idade * Género*	42,438	0,000
*Idade * Educação Mãe*	8,050	0,000
*Idade * Raça Cor*	0,769	0,545
*Região * Idade*	7,254	0,000
*Género * Educação Pai*	8,341	0,000
*Educação Pai * Educação Mãe*	37,334	0,000
*Raça Cor * Educação Pai*	3,309	0,001
*Região * Educação Pai*	3,637	0,000

Fonte de variação	F	Valor p
*Género * Educação Mãe*	28,428	0,000
*Género * Raça Cor*	15,721	0,000
*Região * Género*	6,938	0,000
*Raça Cor * Educação Mãe*	2,107	0,032
*Região * Educação Mãe*	5,365	0,000
*Região * Raça Cor*	13,121	0,000

Tabela 8 – ANOVA com termos principais e interação, t-proximidade (k=5, t=0,30).

Conjunto de dados	N válido	N omisso	Assimetria	Curtose	Desvio padrão
Original	452 578	83 888	0,217	-0,344	14,392
(3, 5)-diversidade	88 931	3 060	0,102	-0,459	14,739

Tabela 9 – Estatísticas de distribuição

Agradecimentos

Este trabalho é parcialmente financiado pela FCT/MCTES através de fundos nacionais e quando aplicável cofinanciado por fundos comunitários no âmbito dos projetos UIDB/50008/2020 e CEMAPRE/REM - UIDB/05069/2020.

Referências

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6) 734–749.

Ayala-Rivera, V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Privacy*. 7(3), 337–370.

Bera, A., & Jarque, C. (1981). Efficient tests for normality, heteroscedasticity, and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7, 313–318.

Barth-Jones, D. (2012). The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. https://ssrn.com/abstract=2076397.

Cambon-Thomsen, A., Rial-Sebbag, E. & Knoppers, B. M. (2007). Trends in ethical and legal frameworks for the use of human biobanks. *Eur. Respiratory Journal*, 30(2), 373–382. https://erj.ersjournals.com/content/30/2/373.

Chen, H, Chiang, R. H. L. & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 36(4) 1165–1188.

Chicaiza, J., Cabrera-Loayza, M. C., Elizalde, R., & Piedra, N. (2020). Application of Data Anonymization in Learning Analytics. *In 3rd Int. Conf. on Applications of Intelligent Systems*, ACM. https://doi.org/10.1145/3378184.3378229.

Culnane, C., Rubinstein, B. I. P., & Teague, V. (2017). *Health data in an open world*. arXiv:1712.05627v1 [cs.CY].

Directive 95. (1995). http: //data.europa.eu/eli/dir/1995/46/oj.

Eicher, J., Bild, R., Spengler, H. *et al.* (2020). A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. *BMC Med Inform Decis Mak,* 20(29). https://doi.org/10.1186/s12911-020-1041-3

Esquivel-Quirós, L. G., Barrantes, E. G., & Darlington, F., E. (2019). Marco de medición de la privacidad. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (31), 66-81. https://dx.doi.org/10.17013/risti.31.66-81.

Francis, P. (2018). *Can anonymized data still be useful? part deux*. https://aircloak.com/can-anonymized-data-still-be-useful-part-deux/.

Fung, B. C. M., Wang, K., Fu, A., & Yu, P. S. (2010). *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, Taylor & Francis Group.

GDPR (2016). Regulation (EU) 2016/679, L 119, pp. 1–88. https://gdpr-info.eu/recitals/no-26/.

Goldstein, H., & Shlomo, N. (2020). A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *Journal of Official Statistics*, 36, 89–115.

Greene, W.H. (2003). Econometric Analysis (5th edition). Prentice Hall.

Kaye, J., Meslin, E., Knoppers, B., Juengst, E., Deschênes, M., Cambon-Thomsen, A., Chalmers, D., Edwards, K., Hoppe, N., Kent, A., Adebamowo, C., Marshall, P., & Kato, K. (2012). Elsi 2.0 for genomics and society. *Science*, 336, 673–674.

Kniola, L. (2017). Plausible Adversaries. *In Re-Identification Risk Assessment. PhUSE Annual Conference*.

Koch, R. (2020). Political campaigns and your personal data. ProtonMail; https://protonmail.com/blog/political-campaigns-and-your-personal-data/.

Lee, H., Kim, H., Kim, J. W., & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *Medical Informatics and Decision Making*. 17(104). https://doi.org/10.1186/s12911-017-0499-0.

LeFevre, K., DeWitt, D., & Ramakrishnan, R., (2005). Incognito: Efficient full-domain k-anonymity. In ACM *SIGMOD Int. Conf. on Management of Data*, *(*pp.49–60).

LeFevre, K., DeWitt, D J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. *In 22nd Int. Conf. on Data Engineering (ICDE'06).*

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *In IEEE 23rd Int. Conf. on Data Eng*. (pp. 106–115).

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), article 3, 52 pages. https://doi.org/10.1145/1217299.1217302.

Panduragan, V. (2014) *On taxis and rainbows*. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

Prasser, F., Eicher, J., Spengler, H., Bild, R., & Kuhn, K. A. (2020). Flexible data anonymization using ARX—Current status and challenges ahead. *Softw Pract Exper*. 50, 1277–1304. https://doi.org/10.1002/spe.2812.

Prasser, F. & Kohlmayer, F. (2015) Putting statistical disclosure control into practice: the ARX data anonymization tool. In A. Gkoulalas-Divanis & G. Loukides (Eds.) Medical Data Privacy Handbook (p.111–48). Springer International Publishing.

Quiñonez, Y., Lizarraga, C., Peraza, J., & Zatarain, O. (2019). Sistema inteligente para el monitoreo automatizado del transporte público en tiempo real. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (31), 94-105. https://dx.doi.org/10.17013/risti.31.94-105.

Rubner, Y., Tomasi, C. & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2), 99–121.

Santos, W., Sousa, G., Prata, P., & Ferrão, M. E. (2020). Data Anonymization: K-anonymity Sensitivity Analysis. 15th Iberian Conf. on Information Systems and Technologies (CISTI) (pp. 1-6). https://doi.org/10.23919/CISTI49556.2020.9141044.

Scheffé, H. (1999). The analysis of variance. New York and London: Wiley.

Spengler, H., & Prasser, F. (2019). Protecting biomedical data against attribute disclosure. Studies in health technology and informatics, 267, 207–214.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Carnegie Mellon Univ. http://dataprivacylab.org/projects/identifiability/

Sweeney, L. (2002a). K-anonymity: A model for protecting privacy. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5) 557–570.

Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5) 571–588.

Sweeney, L. (2015). Only you, your doctor, and many others may know. *Technology* Science. https://techscience.org/a/2015092903/.

Sweeney, L., Loewenfeldt, M. von, & Perry, M. (2018). Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. Technology Science. https://techscience.org/a/2018111301/