IMPACTO DA ADEQUAÇÃO À LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS NA METRIFICAÇÃO DA QUALIDADE DE DADOS

Impact of adequacy to the Brazilian General Data Protection Law on Data Quality Metrification

Leandro Furlam Turi, Giovanni Comarela
Universidade Federal do Espírito Santo
leandro.turi@edu.ufes.br, gc@inf.ufes.br

Resumo:
A necessidade de avaliar a qualidade de dados vêm ganhando importância, uma vez que a real contribuição (o valor de negócio) dos dados só pode ser estimada no seu contexto de uso. Tais obtenções de vantagens econômicas por meio da mineração de dados sofrem intervenção da Lei Geral de Proteção de Dados Pessoais (LGPD), que define regras sobre o processo de coleta, armazenamento e compartilhamento de informações. Considerando essa relação entre qualidade e legislação corrente, este trabalho avalia e compara o impacto de técnicas de adequação à legislação aos processos de mensuração de qualidade de dados. A partir dos resultados obtidos, notou-se que, em maior ou menor grau, todas as formas de corresponder-se à legislação mostraram-se passíveis a alterações da qualidade da base de dados em relação à base de dados original, demonstrando, assim, que a adequação à legislação deve ser conformada ao fim no qual o projeto se dará.
Palavras-chave: Qualidade de Dados; Lei Geral de Proteção de Dados Pessoais; Ciência de Dados.

Abstract:
The demand to assess data quality is gaining importance, since the real contribution (business value) of data can only be estimated in its context of use. Such economic advantages obtained through data mining are subject to the Brazilian General Personal Data Protection Law (LGPD), which defines rules on the process of collecting, storing and sharing information. Considering this relationship between quality and current legislation, this work evaluates and compares the impact of adaptation techniques to legislation on data quality measurement processes. From the results obtained, it was noted that, to a greater or lesser extent, all forms of complying with the legislation proved to be susceptible to changes in the quality of the database in relation to the original database, thus demonstrating that the adequacy to the legislation must be conformed to the purpose in which the project will take place.
Keywords: Data Quality; Brazilian General Personal Data Protection Law; Data Science.

1. 1 Introdução

O acelerado desenvolvimento tecnológico tem modificado a maneira com que se vêm desenvolvendo modelos de negócio no cenário global. Técnicas de processamento e análise de dados têm se mostrado ferramentas bastante usuais na perspectiva do mercado, de modo que os dados vão se constituindo no cerne da tomada de decisão e do planejamento estratégico de empresas (ARDAGNA et al, 2008).

Estratégias de mineração de dados estão surgindo como uma nova solução para os problemas encontrados ao se processarem grandes quantidades de dados. O problema é que nem todos os dados são padronizados: “uma das dificuldades fundamentais é que as informações extraídas podem ser tendenciosas, ruidosas, desatualizadas, incorretas, enganosas e, portanto, não confiáveis” (BERTI-EQUILLE, L.; BORGE-HOLTHOEFER, 2015).

As obtenções de vantagens econômicas por meio da mineração de dados também sofrem intervenção da Lei Geral de Proteção de Dados Pessoais (LGPD) (PINHEIRO, 2020). Essa Lei define regras sobre o processo de coleta, armazenamento e compartilhamento de informações, de forma a fazer valer observância à boa-fé e aos demais princípios elencados no Art. 6º, com especial destaque para os critérios de finalidade (tratamento com propósitos legítimos), prevenção de potenciais danos ao titular dos dados e responsabilização de danos patrimoniais ou morais causados decorrente da atividade de tratamento de dados pessoais, de modo a coibir abusos e favorecer um ambiente mais seguro aos usuários.

Considerando a relação entre qualidade e legislação corrente, este trabalho avalia e compara o impacto de técnicas de anonimização em acordo com a Lei Geral de Proteção de Dados Pessoais aos processos

de mensuração de qualidade de dados. Para isto, serão levantados conjuntos de dados pertencentes a diferentes contextos, aos quais serão aplicadas diferentes técnicas de anonimização acordadas à LGPD. Com cada conjunto anonimizado, é calculada a qualidade dos dados com métricas definidas pela ISO/IEC 25012:2008, cujos cômputos serão comparados e avaliados em relação aos resultados obtidos com a aplicação à base original. Com este processo, pretende-se responder a seguinte pergunta de pesquisa: Como a LGPD influencia na mensuração da qualidade de uma base de dados? Isto reflete-se em projetos de ciência de dados?

2. 2 Objetivos

Analisar a viabilidade e a adaptabilidade de processos de mensuração de qualidade de dados através de diferentes formas de adequação à LGPD: (1) categorizando as abordagens e os requisitos gerais atuais para cada métrica a ser analisada, demonstrando estruturas e desafios; (2) implementando e automatizando tais técnicas em conjunto com algoritmos de suporte para o processo de qualidade de dados; (3) comparando e avaliando os resultados obtidos com cada técnica de anonimização em relação à base original.

3. 3 Procedimentos Metodológicos

A metrificação da qualidade de dados constitui-se de um esquema cíclico, iniciando no mapeamento da documentação e do comportamento dos dados, através da observação de trechos das bases. Em seguida, são definidas as variáveis de teste. Após, ocorre a obtenção e avaliação dos resultados obtidos, recorrendo, e se necessário retificando, conclusões obtidas nos passos anteriores (MERINO, 2016). Todos os códigos utilizados, bem como a base de dados apresentada, estarão disponíveis de forma pública na publicação do trabalho.

Embora a literatura apresente muitas dimensões de qualidade de dados, trazer à tona padrões internacionais como ISO/IEC 25012:2008 e ISO/IEC 25024:2015 pode ser muito conveniente, e eles podem ser utilizados como guias de referência. A ISO/IEC 25012:2008 contém um modelo de qualidade de dados com um conjunto de características que os dados de qualquer sistema de informação devem cumprir para atingir níveis adequados de qualidade de dados externos. A ISO/IEC 25024:2015 fornece medidas gerais para quantificar a qualidade externa e a interna dos dados em conformidade com características da ISO/IEC 25012:2008. Nesse sentido, seis métricas inerentes a sistemas foram selecionadas: *Completude*, *Acurácia*, *Consistência*, *Credibilidade*, *Atualidade* e *Unicidade*. Uma descrição de cada métrica é apresentada a seguir, enquanto exemplos de aplicação poderão ser encontrados na discussão de resultados.

Completude caracteriza a taxa de preenchimento dos atributos. *Acurácia* visa a detectar se a informação registrada reflete o evento ou o objeto descrito, isto é, verificar se o dado cadastrado está em concordância com o evento observado. Tem três aspectos principais: (1) Acurácia sintática, definida como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado sintaticamente correto; (2) Acurácia de alcance, que define os intervalos nos quais os valores dos dados devem ser definidos em um domínio considerado semanticamente correto; (3) Acurácia Semântica, definida como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado semanticamente correto. *Consistência* avalia se dois ou mais atributos estão livres de contradição e são coerentes com outros dados de contextos coerentes. *Credibilidade* avalia se os atributos são considerados verdadeiros e críveis pelos usuários. *Atualidade* avalia se os dados representam o período factível e coerente. *Unicidade* mensura o grau de duplicidade nos dados.

O cômputo dos resultados numéricos ocorre de modo cascata, gerando-se a cada nível um domínio de aplicação. Por exemplo, em um fluxo completo contendo todas as métricas em uso, os registros submetidos ao teste de acurácia devem ser não nulos; os registros submetidos ao teste de credibilidade devem estar acurados; os registros submetidos aos testes de

consistência deverão ser credíveis; ao teste de atualidade devem ser consistentes; e os registros submetidos aos testes de unicidade devem ser atuais. Destaca-se que, quando não for possível metrificar para uma variável específica, a métrica antecedente à que seria utilizada, a nível de registro, deve ser trazida para definir o domínio de aplicação.

Acerca das bases de dados analisadas, procurou-se levantar exemplos de conjuntos de dados nacionais e internacionais em que determinadas informações pessoais foram publicadas como parte do registro público. A inspiração partiu de um tópico[1] da *Open Knowledge Foundation* e pelos tuítes de *@jwyg*, onde a discussão acerca de privacidade decorreu. A saber, as bases utilizadas foram: *athleteEvents*, *Canada*, *eleicoes*, *EuropeanSoccer*, *Poland* e *Sinasc*. A natureza e o tratamento realizado de cada base está disponibilizado no repositório aberto vinculado à pesquisa[2].

Destaca-se que não necessariamente as informações obtidas são objetos da LGPD. Entretanto, como claramente caracterizam-se como informações pessoais (mesmo que referindo-se a pessoas públicas, excluídas da legislação), servem como exemplo para a proposta deste trabalho.

Retornando à adequação à legislação vigente, a LGPD inclui na lista de dados sensíveis aqueles que dizem respeito à "origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural". Uma vez definidos os objetos da anonimização, basta definir os mecanismos. O inciso XI do Art. 5º define anonimização como sendo "utilização de meios técnicos razoáveis disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo". Buscando eliminar tais elementos identificadores, o trabalho foi feito sobre quatro grandes técnicas conhecidas em computação: *Supressão (SUP)*, *Generalização (GEN)*, *Randomização (RAND)* e *Pseudoanonimização (PSAN)*.

Supressão consiste na ação de suprimir (ou anular) determinada variável referindo-se a um registro. Por exemplo, excluir os dígitos de um número de telefone ou todos os nomes de uma base de dados. *Generalização* consiste em substituir os registros específicos por categorias mais amplas e genéricas. Por exemplo, idades exatas são convertidas em faixas etárias e um CEP é trocado apenas pela cidade ou região do país. *Randomização* é o processo de tornar o registro aleatório. Neste trabalho, os registros de cada variável submetida a esse processo são permutados aleatoriamente. *Pseudoanonimização* consiste no mecanismo do encobrimento da informação, substituindo-se um atributo por outro. Neste trabalho, o hash SHA224[3] foi utilizado.

4. 4 Resultados

4.4.1 Completude

Nessa dimensão são detectados valores faltantes mediante a busca por tipos ou valores representando a informação nula na base de dados. Destaca-se que para esta métrica faz-se necessário observar a dependência entre as variáveis definidas nos dicionários de dados adotados, como por exemplo a existência de uma anomalia em contrapartida ao código representativo desta (*Sinasc*). A Fig. 1 apresenta a completude de cada base de dados analisada para cada aplicação de técnica de anonimização, com o respectivo percentual numérico. Destaca-se a grande queda de completude dos registros para a técnica de supressão, em todas as bases analisadas, seguida pela queda para a técnica de generalização.

A queda para a supressão decorreu do fato de que, durante a ação de anular determinadas variáveis, essas passam a ser incompletas, causando as quedas observadas. Já para a generalização, a queda decorreu da aplicação da técnica de forma total para as variáveis.

[1] *https://lists-archive.okfn.org/pipermail/mydata-open-data*

[2] *https://github.com/leandrofturi/qualityLGPD*

[3] *https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf*

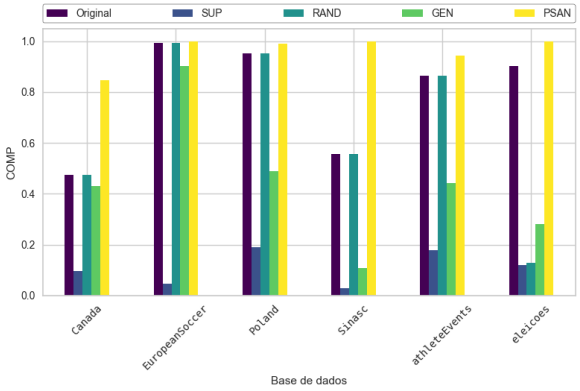

Figura 1 – Resultados para métrica de completude

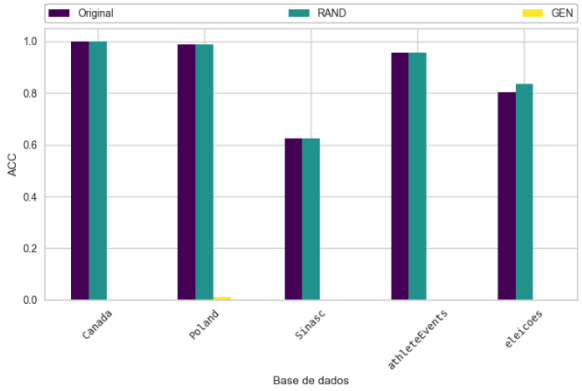

Figura 2 – Resultados para métrica de acurácia

4.4.2 Acurácia

Verificou-se se os dados apresentam os padrões descritos no dicionário de dados adotado como referência, tanto de quantidade de caracteres, quanto de valores esperados, em três aspectos principais: (1) sintático, definido como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado sintaticamente correto. Isso abrange todos os domínios finitos definidos pelo dicionário de dados, bem como formato dos valores (também conhecido como conformidade); (2) alcance, definido pelos intervalos aos quais os valores dos dados devem ser definidos, abrangendo principalmente datas; e (3) semântico, definido como a proximidade dos valores dos dados a um conjunto de valores definidos em um domínio considerado semanticamente correto, definido pelos dicionários de dados, abrangendo principalmente pesos, idades e medidas dentro do contexto no qual a base de dados se encontra. O resultado percentual por variável está descrito na Fig. 2.

Destaca-se a grande queda de acuracidade dos registros para a técnica de generalização, uma vez que várias verificações, que antes eram passíveis de serem aplicadas na base de dados original, pela aplicação não são mais necessárias. Isso inclui testes envolvendo conformidades ao dicionário de dados, faixas de valores para datas e verificações de idades e acuracidade nos códigos de cidades e municípios.

4.4.3 Credibilidade

Verificou-se o grau em que os dados têm atributos que são considerados verdadeiros e críveis, incluindo o conceito de autenticidade (a veracidade das origens, atribuições, compromissos) ISO/IEC 25012:2008. Nesse contexto, testes envolvendo a origem da informação (municípios de nascimento para a base de dados *Sinasc* e unidade eleitoral para a base de dados *eleicoes*) foram realizados, objetivando verificar se de fato a informação disponibilizada refere-se ao recorte realizado (estado do Espírito Santo). Poucos foram os problemas identificados, conforme apresentado na Fig. 3.

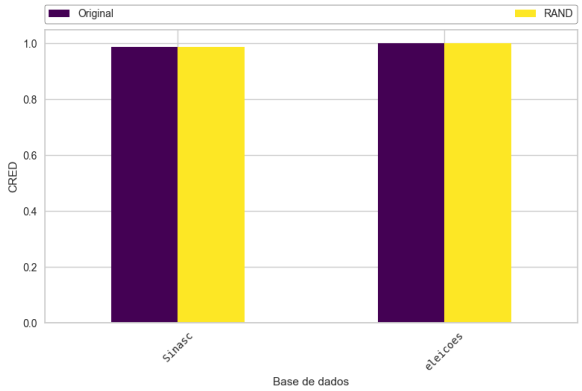

Figura 3 – Resultados para métrica de credibilidade

4.4.4 Consistência

Os resultados descritos na Fig. 4 são de testes aplicados a um mesmo registro, ou seja, mesma linha do conjunto de dados. Estes testes detectam principalmente problemas na entrada de dados envolvendo condições específicas de inconsistências e dependentes do contexto. Por exemplo, para a base de dados *Sinasc*, um teste

envolvendo a data de nascimento do recém-nascido e a data de nascimento da mãe varre os registros identificando se a data do nascimento do recém-nascido é maior que a data de nascimento da mãe. Todos os demais testes aplicados estão descritos no repositório aberto desta pesquisa.

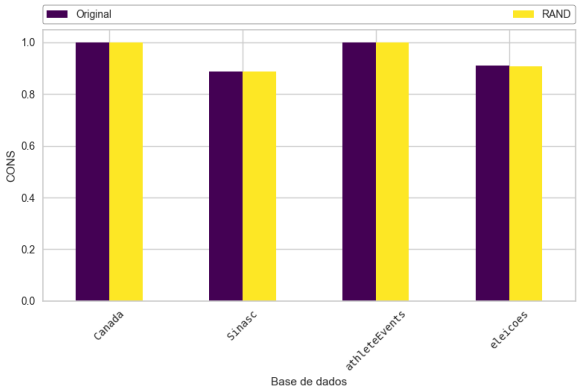

Figura 4 – Resultados para métrica de consistência

4.4.5 Atualidade

Nessa métrica foi verificado o grau em que os dados têm atributos que estão na temporalidade adequada para o uso. Poucos testes foram identificados: em *EuropeanSoccer*, verificar se cada registro possui informações contínuas (sem lacunas entre anos); e em *Sinasc*, se a informação do nascimento era submetida ao sistema em no máximo um ano. Os resultados são apresentados na Fig. 5.

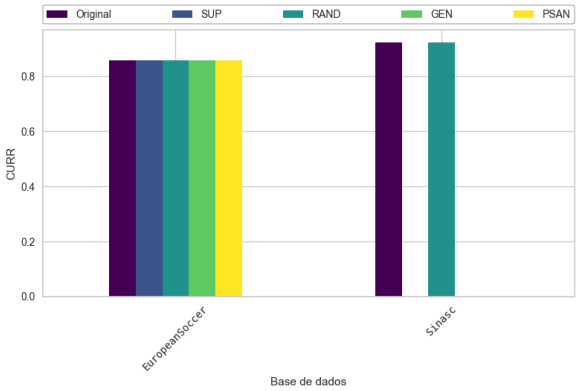

Figura 5 – Resultados para métrica de atualidade

4.4.6 Unicidade

Por fim, na métrica de unicidade foi analisado o grau de duplicidade dos registros, por meio da identificação do registro por um identificador único e pela verificação das constantes que definem os registros, como por exemplo nome, data e local de nascimento. Os resultados percentuais são apresentados pela Fig. 6.

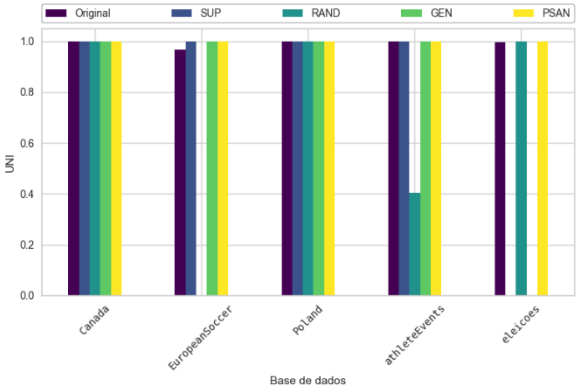

Figura 6 – Resultados para métrica de unicidade

De fato, a técnica de randomização é a mais favorável a apresentar erros quando existe mais de uma linha relativa a um mesmo registro, uma vez que essa técnica irá embaralhar a ordem das informações. Por outro lado, as técnicas de supressão, generalização e pseudoanonimização são favoráveis a manter a quantidade de falhas, uma vez que as variáveis se tornam mais amplas, e menos susceptíveis a erros. Tome por exemplo o campo referindo-se ao código de seis dígitos do município de nascimento do recém-nascido vivo de acordo com a tabela de códigos e municípios do IBGE[4], na base de dados *Sinasc*. Generalizando-se essa informação, ou seja, mantendo-se apenas os dois primeiros dígitos referentes apenas ao estado de nascimento, a possibilidade de haver erros nessa informação é reduzida. Isso vale para os demais testes.

5. 4 Considerações Finais

A avaliação realizada é especialmente oportuna, tendo em vista o cenário nacional e o atual empenho em fomentar o debate em torno da qualidade das informações e o respeito à privacidade do brasileiro. Nesse contexto, este trabalho explorou métricas de qualidade de dados presentes em normas técnicas, objetivando avaliar a qualidade dos dados de conjuntos de diferentes contextos e o impacto que a adequação à LGPD ocasiona nestes projetos. Para isso, foram

[4] *https://www.ibge.gov.br/explica/codigos-dos-municipios.php*

metrificadas as qualidades das bases de dados originais (*athleteEvents*, *Canada*, *eleicoes*, *EuropeanSoccer*, *Poland* e *Sinasc)*, e comparadas com o resultado obtido após conformação com a legislação.

Analisados os resultados obtidos pela métrica de completude, observou-se uma grande queda para a técnica de supressão, em todas as bases analisadas, seguida pela queda para a técnica de generalização. A queda para a supressão decorreu do fato de que, durante a ação de anular determinadas variáveis, estas passam a ser incompletas, causando os resultados observados. Já para a generalização, a queda decorreu da aplicação da técnica de forma completa para algumas variáveis.

Sobre os resultados dos testes de acurácia, destacou-se a grande queda de acuracidade dos registros para a técnica de generalização, uma vez que várias verificações que antes eram passíveis de serem aplicadas na base de dados original, pela generalização não são mais necessárias. Isso inclui testes envolvendo conformidades ao dicionário de dados, faixas de valores para datas e verificações de idades e acuracidade nos códigos de cidades.

Com relação aos resultados de credibilidade, consistência, atualidade e unicidade, observou-se pouca variação da metrificação, quando a técnica aplicada conseguiu manter a semântica da informação, mesmo que mascarada.

Baseado nestes resultados, nota-se uma técnica que se sobressai às demais, a saber, a pseudoanonimização. Uma vez que esta altera apenas a sintaxe dos registros (os valores visíveis, da forma como são escritos), não se deve esperar que haja diferenciações na aplicação desses algoritmos, desde que, claro, as informações originais não sejam de maioria numérica (perdendo-se, com a aplicação da técnica, a semântica intrínseca da informação). Tal fato foi tanto levantado na metrificação da qualidade e pode ser elencado em projetos posteriores de ciência de dados.

Visando observar a fundo tal comportamento evidenciado pela técnica de pseudoanonimização, realizar análises via algoritmos de aprendizado de máquina (de natureza geométrica como um *KMeans*[5]) em que a semântica intrínseca aos registros não seja requisito para o bom funcionamento podem ser redigidas e avaliadas.

6 Referências

ARDAGNA, D. et al. Context-aware data quality assessment for big data**. Future Generation Computer Systems**, v. 89, p. 548–562, 2018. ISSN 0167-739X.

BERTI-EQUILLE, L.; BORGE-HOLTHOEFER, J. Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics. **Synthesis Lectures on Data Management**, Morgan & Claypool Publishers, v. 7, n. 3, p. 1–155, 2015.

ISO/IEC JTC 1/SC 7 Software and systems engineering technical committee. **ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model**. 1. ed. Vernier, Geneva, Switzerland, 2012.

ISO/IEC JTC 1/SC 7 Software and systems engineering technical committee. **ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality**. 1. ed. Vernier, Geneva, Switzerland, 2015.

PINHEIRO, P. P. **Proteção de Dados Pessoais: Comentários à Lei n. 13.709/2018-LGPD**. 2. ed. São Paulo, São Paulo, Brasil: Saraiva, 2020. ISBN 9788553617487.

MERINO, J. et al. A data quality in use model for big data. **Future Generation Computer Systems**, v. 63, p. 123–130, 2016.

[5] *http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf*