Qualidade de Dados em Bases de Dados Anonimizadas: Uma Abordagem de Avaliação Mista

Paulo Pombinho,[1] pmimatos@fc.ul.pt
Luís Cavique[1,2], luis.cavique@uab.pt
Luís Correia[1], luis.correia@ciencias.ulisboa.pt

[1] *LASIGE, DI - Faculdade de Ciências da Universidade de Lisboa, Portugal*
[2] *Universidade Aberta, Lisboa, Portugal*

1. Introdução

A qualidade dos dados é crucial, tanto ao nível empresarial como governamental, para uma análise adequada da situação que aqueles representam e para as consequentes tomadas de decisão (Batini, 2009). Em projetos de prospeção de dados é especialmente relevante evitar dados com qualidade inferior uma vez que se usam algoritmos que dependem de dados corretos para criar modelos e previsões precisos. Um conjunto de dados com problemas de qualidade pode implicar custos elevados tanto a nível económico como social, com a possibilidade de decisões erróneas serem tomadas quando se olha para dados incorretos (Batini e Scannapieco, 2016; Wang e Strong, 1996). Neste texto, propomos uma abordagem de avaliação de qualidade que considera métricas que lidam com atributos individuais e, adicionalmente, uma análise longitudinal de fluxo, que permite fazer uma avaliação de qualidade que tem em consideração informação contextual. São propostas métricas de Qualidade de Dados por Entrada e Qualidade de Dados por Atributo e, finalmente, é proposta uma medida de Qualidade Global de Dados baseada nessas métricas.
A avaliação da qualidade dos dados não é um tema de investigação novo e existem diversos trabalhos que analisam as diferentes métricas de avaliação de dados. No entanto, a maioria das propostas existentes focam-se na precisão de dados que podem ser comparados com as entidades do mundo real que representam. Esta comparação não é, por vezes, possível se, por diversas razões, a informação sobre a entidade do mundo real não estiver disponível a partir de outras fontes (Coleman-Sebastian, 2013). Este problema é especialmente relevante quando, devido à anonimização dos dados, surgem problemas nos próprios identificadores das entradas na base de dados.
Neste artigo, propomos uma abordagem de avaliação de qualidade que utiliza um conjunto diversificado de métricas de qualidade. Consideramos não só a informação sobre as diferentes dimensões individuais do atributo, mas também o seu contexto semântico. Propomos métricas de avaliação que permitam classificar os dados individuais e agregados e, através desta classificação, atribuir uma qualidade global.
Na secção seguinte apresentamos o trabalho relacionado que aborda os diferentes tipos de métricas de qualidade. Na secção 3, discutimos os problemas que surgem da anonimização dos dados e, na secção 4, os tipos de erros que podem existir. A secção 5 enumera os diferentes tipos de avaliações de qualidade utilizadas e, na secção 6, as métricas de avaliação são calculadas obtendo as medições de qualidade dos dados por entrada e por atributo. Na secção 7, a qualidade global de dados é definida e, finalmente, na secção 8 apresentamos uma análise preliminar da utilização das métricas de qualidade e, na secção 9 as conclusões e o trabalho futuro.

2. Trabalho Relacionado

Embora existam diversas propostas de métricas de qualidade de dados, não existe um acordo real sobre as dimensões que definem a qualidade dos dados nem sobre a sua nomenclatura (Batini et al., 2009).

Há uma grande diversidade de terminologias sobre métricas de qualidade de dados, muitas vezes com significados semelhantes, que são frequentemente demasiado focadas apenas em métricas simples de precisão. Com uma quantidade cada vez maior de dados disponíveis, as instituições estão cada vez mais a utilizar estes dados para apoiar a sua tomada de decisão, embora com a preocupação de não utilizarem dados incorretos, que possam causar más decisões, com grande impacto social e financeiro (Heinrich, 2018).

Adicionalmente, a crescente utilização de técnicas de inteligência artificial com conjuntos de dados muito grandes também pode ver a sua precisão muito afetada pela baixa qualidade dos conjuntos de dados originais (Ding e Li, 2018). As qualidades subjacentes ao *Big Data*, caracterizadas pelos três V's: Volume, Velocidade e Variedade, são um desafio significativo para garantir dados com qualidade e a adição de um quarto V: Veracidade, é sugerida para indicar a importância da qualidade dos dados (Lukoianova and Rubin, 2013).

Em seguida, descreveremos alguns dos trabalhos mais relevantes que se relacionam com as métricas de qualidade que usamos como base para a nossa proposta de avaliação da qualidade de dados.

Pipino, Lee e Wang (2002) enumeram diferentes métricas de avaliação. Listam uma dimensão livre de erro que representa a correção de dados; a completude que pode variar desde um nível de esquema de base de dados, para indicar quão completo é o próprio desenho da base de dados, até à completude por coluna, que se refere ao número de entradas na base de dados com um valor presente vs. em falta; e consistência, que avalia se o mesmo valor é consistente entre diferentes entradas de dados redundantes.

Cai e Zhu (2015) definem cinco dimensões diferentes, sendo as quatro primeiras: disponibilidade, avaliando se os dados são atualizados regularmente; usabilidade, relativamente à fiabilidade da fonte de dados e se o intervalo de valores é aceitável; relevância, avaliando se os dados são pertinentes para o tema em que são utilizados; qualidade de apresentação, relacionada com a compreensibilidade dos dados. A dimensão final é a fiabilidade, em que os autores incluem quatro subelementos: precisão, avaliando se os dados representam o estado real da informação; consistência, estimando se os dados são consistentes com outras fontes; integridade, verificando a sua consistência com as regras estruturais e de integridade dos conteúdos; e a completude para avaliar se problemas ou falta de informação num componente individual de dados compostos por múltiplos componentes, pode ter impacto na precisão e integridade globais.

Sidi et al. (2012) classificam os problemas de qualidade em problemas de fonte única ou de múltiplas fontes, respetivamente se estiverem contidos num conjunto de dados, ou se os problemas forem originários de fontes heterogéneas e das suas interações. Diferenciam entre problemas ao nível do esquema da base de dados e ao nível da instância se o problema de qualidade estiver relacionado com o desenho da base de dados ou se estiver na própria entrada, respetivamente. Os autores também descrevem um conjunto extenso de diferentes métricas a partir das quais destacam a consistência, que avalia se os dados são apresentados nos mesmos formatos ou se são compatíveis; precisão, para avaliar se os valores dos dados correspondem às suas congéneres do mundo real; completude, para aferir se os dados podem representar todos os estados possíveis e significativos; e a atualidade, identificando se os dados estão atualizados.

Jugulum (2014) define as quatro dimensões fundamentais da qualidade dos dados como completude, uma medida que avalia quais os elementos que precisam de estar presentes para atingir os objetivos; conformidade, para perceber se os dados são compatíveis com os formatos exigidos; validade, que aborda a correspondência dos valores dos dados com os seus tipos e intervalos corretos; e a precisão, que avalia se o valor dos dados reflete o mundo real.

Os trabalhos apresentados tratam principalmente da qualidade das entradas individuais na base de dados, não considerando que algumas entradas possam estar relacionadas com outras. No nosso trabalho vamos utilizar uma combinação das métricas descritas por Pipino, Lee e Wang (2002), Sidi et al (2013) e Cai e Zhu (2015) e adaptá-las para avaliar não só as entradas individuais, mas também realizar uma análise longitudinal contextual que considera a relação entre diferentes entradas, permitindo uma deteção mais robusta dos problemas de qualidade nos dados.

3. O Problema dos Dados Anonimizados

A investigação e a tomada de decisões políticas, por exemplo nas áreas da saúde pública ou da educação, estão cada vez mais a utilizar sistemas de informação com conjuntos de dados muito grandes para

apoiar as suas decisões (Cavique, 2020). A utilização de dados de baixa qualidade tem sido, no entanto, frequentemente reportada e leva a efeitos negativos (Chen, 2014).
A utilização de dados pessoais, de acordo com o Regulamento Geral sobre a Proteção de Dados (GDPR, 2016), implica que os conjuntos de dados são frequentemente anonimizados. Isto cria um desafio para a avaliação da qualidade dos dados, uma vez que impede a identificação e confirmação de erros, tornando impossível o uso de outro conjunto de dados que possa ser usado como uma comparação de quais os valores são corretos. Além disso, a anonimização eficaz geralmente está associada a outras proteções para evitar a reidentificação, de modo que uma entrada anonimizada não possa ser identificada através de uma análise cruzada de todos os dados disponíveis. Um exemplo é o conceito de *k*-anonimato (Samarati, 2001), onde os dados divulgados não permitem que uma associação seja feita a um número de indivíduos menor que *k*.
Isto significa que, para alguns domínios de aplicação, todos os dados extraídos devem estar na forma agregada sem entradas individuais. Como tal, os dados utilizados em algoritmos externos podem utilizar apenas dados de grupos de entradas com atributos semelhantes e um contador de quantos deles estão nessas condições específicas.
Estas limitações têm como consequência tornar as métricas tradicionais inadequadas para avaliar a qualidade dos dados anonimizados (Fletcher e Islam, 2014). Apesar destas questões, a investigação tem-se centrado mais nos próprios procedimentos de anonimização, do que na utilização de dados anonimizados na prospeção de dados e na sua avaliação de qualidade. Além disso, se for realizado um procedimento de anonimização impróprio, a utilidade potencial da prospeção de dados pode ser arruinada até mesmo por ganhos marginais de privacidade (Brickell e Shmatikov, 2008). Isto é especialmente importante se considerarmos que o fator mais importante para o sucesso dos projetos de ciência de dados é a utilização das dimensões corretas e o seu tratamento, sendo a parte mais morosa destes projetos (Domingos, 2012).
Algumas soluções para ultrapassar as questões da anonimização propõem a recolha de estatísticas, que são divulgadas com os dados anonimizados, para ajudar na avaliação da qualidade (Inan et al., 2009). Outros investigadores propõem que se invista na escolha dos procedimentos de anonimização certos para resolver os problemas que lhe estão associados ou até mesmo permitir a obtenção de melhores resultados (Buratovic et al., 2012; Silva et al., 2017). Tais abordagens, no entanto, não são viáveis para serem utilizadas por investigadores que utilizam conjuntos de dados grandes e já existentes, que foram previamente anonimizados a nível institucional.
A nossa abordagem visa ser usada em conjuntos de dados anonimizados, onde não há controlo sobre os procedimentos de anonimização. Na secção seguinte, descreveremos os tipos de erros em que nos focamos, neste contexto.

4. Tipos de Erros

Embora existam várias formas de introduzir erros nos conjuntos de dados, existem três problemas de qualidade principais que originam erros: dados em falta, erros de inserção e problemas com identificadores.

Dados em Falta dizem respeito a atributos que não são preenchidos em algumas das entradas. Estes podem ser originados por causas distintas. Podem faltar informações sobre um determinado atributo devido a uma omissão ao introduzir os dados na base de dados ou devido a problemas ou escolhas na própria estrutura da base de dados que possam, em determinadas circunstâncias, permitir a existência de valores nulos. Além disso, consideramos também falta de informação em situações em que, embora o atributo esteja preenchido, tem um conteúdo "não responde", o que é o caso, por exemplo, em bases de dados com informações que são propositadamente omitidas pelas entidades em causa.

Erros de Inserção podem resultar de erros diretos ao inserir informação na base de dados como, por exemplo, escrever um algarismo extra num atributo de idade ou um número de telefone. Estes erros também podem derivar de diferentes formatos utilizados em diferentes momentos, o que introduz inconsistências nos dados, especialmente se as variáveis utilizadas não forem suficientemente restritivas. Por exemplo, escrever um valor de data, num formato não específico, alternando entre o dia/mês/ano e o mês/dia/ano.

Problemas de Identificadores são aqueles que derivam dos procedimentos de anonimização e podem, por vezes, produzir problemas nos dados. Uma vez que não temos nenhuma base de verdade para comparar os resultados, estes problemas são mais difíceis de detetar e corrigir. Os problemas de identificação podem resultar em problemas graves se estivermos a analisar a relação entre diferentes entradas da mesma entidade do mundo real e o processo de anonimização introduziu ruído nas próprias chaves primárias ou estrangeiras. Isto é especialmente relevante em conjuntos de dados que consistem em séries temporais onde cada entidade tem várias entradas de dados recolhidas em determinados intervalos temporais.
As imagens que se seguem mostram exemplos de como identificadores inconsistentes podem criar uma perceção errada e potencialmente grave do que é a realidade.
A figura 1 mostra um exemplo de um conjunto de dados de evolução da saúde de pacientes, onde uma sobreposição de dois identificadores diferentes pode incorretamente mostrar um paciente doente a ser curado quando, na verdade, ambos os pacientes permaneceram na mesma situação em dois pontos de dados temporais diferentes.

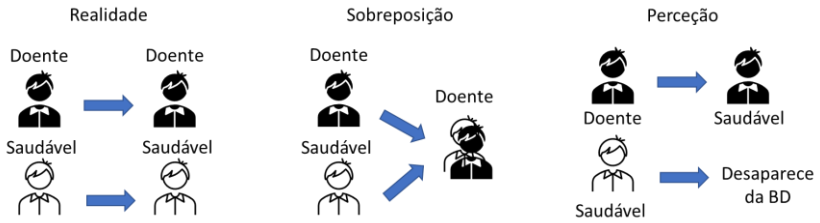

Figura 1. sobreposição de dois identificadores diferentes.

Da mesma forma, a figura 2 mostra como, num conjunto de dados de infratores criminais, se um novo identificador for incorretamente gerado para uma entidade existente, uma infração recorrente pode ser identificada como sendo de um novo infrator na sua primeira infração.

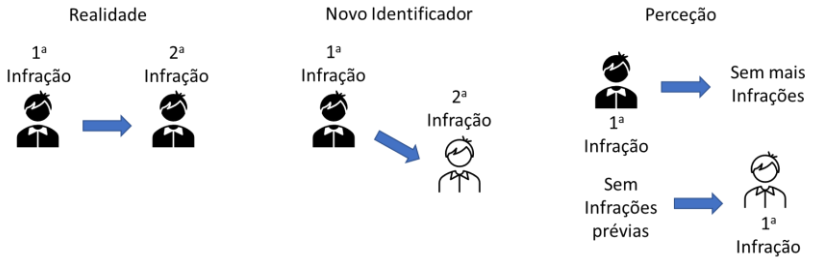

Figura 2. atribuição incorreta de um novo identificador a um existente.

5. Tipos de Avaliação da Qualidade

Embora alguns dos erros presentes na base de dados possam ser facilmente identificados, geralmente estes são apenas um pequeno subconjunto do total de erros presentes. Por exemplo, se as diferenças em dois atributos diferentes dos pacientes não forem extremas, os erros podem ser difíceis de detetar. Situações como as descritas nas figuras 1 e 2 são difíceis de sinalizar como erros, pois produzem evoluções de séries temporais que podem estar muito próximas dos casos reais.
Por conseguinte, os potenciais erros devem ser analisados utilizando o contexto envolvente. É provável que erros graves, como os mencionados, sejam associados a outras inconsistências entre os dados.

Neste trabalho, propomos a combinação de diferentes tipos de análise que permitem uma classificação mais robusta de qualidade e tendo em conta as diferentes dimensões de dados.
Consideramos diferentes tipos de análise:

- Análise Individual - considera cada entrada da base de dados isoladamente e visa identificar problemas nos valores dos atributos isoladamente.

- Análise longitudinal de fluxo – considera o fluxo temporal de cada entrada da entidade e avalia cada entrada tendo em consideração as entradas, da mesma entidade, que a precedem e sucedem, para permitir possíveis problemas serem identificados através da sua comparação. Esta análise só é possível em conjuntos de dados com séries temporais.

- Análise de atributos – Foca cada atributo para identificar aqueles que podem ter mais problemas, ao longo de todo o conjunto de dados. Em vez de realizar análises para cada entrada, todas as entradas são consideradas.

Para avaliar a qualidade de cada entrada na base de dados, utilizamos uma combinação das diferentes análises descritas.
Adaptamos a nomenclatura proposta por Pipino, Lee e Wang (2002), Sidi et al. (2013) e Cai e Zhu (2015) descritas na secção 2 e consideramos quatro tipos de métricas de qualidade. Estas métricas são fornecidas através de uma configuração adequada durante o processo ETL (*Extract, Transform, Load*).

Completude - avalia se todos os atributos de uma entrada são preenchidos ou se faltam valores em alguns atributos. É uma métrica binária.

Consistência - avalia se as informações contidas nos atributos da entrada tiveram de ser normalizadas durante a fase de processamento ETL. Ou se dados inseridos não estavam formatados de acordo com o resto do conjunto de dados. É uma métrica binária

Unicidade - identifica as entradas como duplicadas se, por exemplo, houver duas entradas no mesmo período de tempo que estão em conflito entre si. É uma métrica com valores reais contidos no intervalo entre 0 e 1.

Precisão - lida com a exatidão dos valores de atributo de uma determinada entrada. Permite a identificação de entradas com atributos preenchidos de forma incorreta. Como exemplo, se considerarmos a precisão de um atributo numérico, podemos identificar cenários em que o valor é inválido porque está fora do domínio de valores possíveis, tais como valores negativos quando apenas os positivos são válidos. É uma métrica de número real entre 0 e 1.

No que diz respeito à avaliação longitudinal do fluxo dos dados, consideramos uma avaliação de que analisa diferentes dimensões contextuais:

- Variações de atributos – avalia situações em que uma alteração do valor de um atributo, entre dois pontos de tempo consecutivos, de uma mesma entidade, é maior do que seria expectável. Isto só é possível nos conjuntos de dados de séries temporais.

- Variação do valor – executa um cálculo que avalia a probabilidade de um valor de um atributo considerando os outros atributos da mesma entrada.

- Incompatibilidade de Atributos – compara os diferentes atributos na mesma entrada, para os quais se sabe que existe uma correlação, em busca de valores incompatíveis.

Na secção seguinte descreveremos como calculamos a qualidade dos dados de cada entrada.

6. Qualidade Individual de Dados

Utilizando os tipos de avaliação definidos na secção anterior, podemos calcular quatro métricas diferentes de qualidade de dados. Todas estas métricas têm um intervalo entre 0 e 1, onde 0 representa qualidade nula e 1 uma qualidade perfeita.

6.1. Métricas de Qualidade Individual

Para calcular a métrica de **completude,** utilizamos informações relativas à contagem de atributos em falta, numa entrada, sobre o número total de atributos, M (equação 1).

$$Comp = 1 - \frac{1}{M} \sum_{i=1}^{M} [i \text{ está em } falta] \quad (1)$$

A **consistência** dos dados também é classificada, de forma semelhante à completude, utilizando a contagem de atributos com problemas de consistência sobre o número total de atributos, M (equação 2).

$$Cons = 1 - \frac{1}{M} \sum_{i=1}^{M} [i \text{ não é } consistente] \quad (2)$$

Finalmente, a **unicidade** é classificada de acordo com a identificação de entradas duplicadas, caso o identificador esteja em mais que uma entrada no mesmo ponto de dados temporais e, no caso de isto acontecer, avaliando quão diferentes são as entradas duplicadas (equação 3).

$$Unic = \begin{cases} 1, & não\ duplicado \\ 0.8, & duplicado\ com\ diferenças\ menores \\ 0, & duplicado\ com\ diferenças\ substanciais \end{cases} \quad (3)$$

A métrica de **precisão** avalia a exatidão e a probabilidade de determinados valores de atributos. Como indicado anteriormente, classificamos um atributo como inválido, se o valor ficar fora dos domínios desse atributo. No entanto, esta análise pode ser melhorada através da utilização de informação semântica dos valores de atributos esperados, ao considerar outras informações contextuais e permitir uma classificação com diferentes graus de probabilidade (equação 4). V é o valor do atributo, V_{max} e V_{min} são os valores máximos e mínimos válidos, e VP_{max} e VP_{min} são os respetivos valores máximos e mínimos prováveis para o contexto atual da entrada.

$$ProbabilidadeDeValor(V) = \begin{cases} 1, & VP_{min} < V < VP_{max} \\ \frac{V - V_{min}}{VP_{min} - V_{min}}, & V_{min} < V < VP_{min} \\ 1 - \frac{V - VP_{max}}{V_{max} - VP_{max}}, & VP_{max} < V < V_{max} \\ 0, & V < V_{min} \vee V > V_{max} \end{cases} \quad (4)$$

Como exemplo, podemos ter uma análise de um atributo com idade de utentes em idade reprodutiva numa maternidade, em que V é a idade atual, V_{max} e V_{min} são os valores absolutos máximo e mínimo que definem o intervalo de idades válidas, por exemplo valores entre 0 e 130, e VP_{min} e VP_{max} são os valores máximo e mínimos em que é expectável o valor do atributo, por exemplo entre 15 e 44, respetivamente.

Da mesma forma, um atributo pode ser classificado utilizando informações sobre a variação do valor atual V_a (obtida comparando com o ponto de dados anterior da mesma entidade) e as variações mínimas e máximas possíveis, respetivamente V_{min} e V_{max} (equação 5).

$$VariaçãoDeValor(V_a) = 1 - \frac{|V_a| - |V_{min}|}{|V_{max}| - |V_{min}|} \quad (5)$$

Como exemplo, podemos ter uma base de dados em que se registe o consumo mensal de água de clientes e em que se analisa a variação do consumo. Neste caso, V_a é a variação do consumo entre dois

pontos temporais, e V_{min} e V_{max}, as variações de consumo mínimas e máximas que são prováveis entre dois meses consecutivos.

Finalmente, a precisão de atributos em que se sabe existir una correlação pode ser classificada dependendo da comparação entre ambos os atributos e a existência de valores incompatíveis (equação 6).

$$Compatibilidade(V_1, V_2) = \begin{cases} 1, & V_1\ compatível\ com\ V_2 \\ 0, & V_1 incompatível\ com\ V_2 \end{cases} \quad (6)$$

Uma aplicação desta avaliação pode ser exemplificada num conjunto de dados que contenha dados de alunos e em que o atributo idade de um aluno não seja compatível com o ano curricular em que este esteja inscrito.

Tendo calculado os diferentes tipos de métricas de precisão, a precisão global da entrada é definida como a média dos seus diferentes componentes em todos os atributos M (equação 7).

$$Prec = \frac{\sum_{i=1}^{M} ProbabilidadeDeValor_i + \sum_{i=1}^{M} VariaçãoDeValor_i + \sum_{i=1}^{M} Compatibilidade_i}{3M} \quad (7)$$

6.2. Qualidade de Dados por Entrada

A métrica de avaliação de qualidade descrita na secção anterior permite-nos compreender, para cada entrada, a sua fiabilidade em cada dimensão.

Cada métrica é útil para ser capaz de identificar potenciais problemas com os dados. No entanto, uma vez que se espera que problemas de dados mais graves criem problemas de qualidade em várias dimensões em simultâneo (como é o caso, por exemplo, dos problemas que surgem de inconsistências com os identificadores), utilizamos a combinação das diferentes métricas para conseguir obter uma qualidade global para cada entrada na base de dados.

A Qualidade dos Dados por Entrada (QDE) é obtida como a soma ponderada de todas as métricas anteriormente descritas e é classificada como um valor entre 0 (sem qualidade) a 1 (melhor qualidade) (equação 8). Uma vez que, dependendo dos objetivos dos projetos, alguns dos problemas detetados com os dados podem ter uma influência mais severa na utilidade dos conjuntos de dados resultantes. Optámos por usar pesos em cada métrica de forma a poder dar mais importância às métricas mais importantes. Estes pesos são valores reais positivos, cuja soma é 1.

$$QDE_i = w_{Prec} \times Prec_i + w_{Comp} \times Comp_i + w_{Cons} \times Cons_i + w_{Unic} \times Unic_i \quad (8)$$

A qualidade dos dados de entrada permite-nos não só realizar a filtragem inicial das entradas de dados que têm a qualidade mais baixa, mas também, reconhecer os identificadores que estão associados a potenciais problemas de identificação e reportá-los para serem corrigidos.

7. Qualidade de Dados Global e por Atributo

Para utilizar os resultados em algoritmos de prospeção de dados e, quando apenas os dados agregados podem ser extraídos, é também necessário calcular um valor médio de qualidade de dados de entrada (QDE_{med}) de todas as entradas que compõem cada grupo de dados, com *N* entradas (equação 9).

$$QDE_{med} = \frac{1}{N} \sum_{i=1}^{N} QDE_i \quad (9)$$

Embora seja possível realizar a filtragem inicial utilizando a QDE, a utilização de uma qualidade de dados de grupo, no âmbito das técnicas de prospeção de dados, é relevante para ser capaz de filtrar os

dados, de forma dinâmica. Desta forma é possível utilizar diferentes limiares de qualidade, com os quais é possível afinar quais os que dão os melhores resultados e produzem as previsões mais corretas.
Anteriormente já foi descrita a avaliação de qualidade que pode ser alcançada para cada entrada do conjunto de dados. É também vital obter informações para a qualidade de cada atributo, utilizando uma visão global do conjunto de dados em vez de olhar para cada entrada separada.
A qualidade dos dados do atributo (QDA) pode ser obtida usando os diferentes cálculos para diferentes tipos de atributos. Para atributos para os quais só identificamos, numa classificação binária, se existe ou não um problema, utilizamos a proporção de entradas com problemas relativos ao número total de entradas *N* (equação10).

$$QDA\_bin = 1 - \frac{1}{N} \sum_{i=1}^{N} [i\ tem\ problemas] \tag{10}$$

Se, por outro lado, tivermos cálculos de precisão disponíveis, podemos utilizar a média de todos os valores de precisão das entradas nesse atributo, onde *N* é o número de entradas no conjunto de dados (equação 11). Uma alternativa que poderia ser explorada no futuro, é a utilização de um modelo não linear onde é possível, por exemplo, ter um crescimento exponencial que dependa do número de entradas que têm um erro nesse atributo.

$$QDA\_prec = \frac{1}{N} \sum_{i=1}^{N} \frac{ProbDeValor_i + VariaçãoDeValor_i + Compatibilidade_i}{3} \tag{11}$$

Por último, para obter uma visão geral da qualidade no conjunto de dados, um primeiro passo foi incluir pesos para cada tipo de métrica no cálculo da qualidade individual. Isto permite-nos definir uma maior importância para o tipo de problemas de qualidade que podem ser mais severos para os objetivos do projeto. Além disso, para obter uma avaliação verdadeiramente global do conjunto de dados (ou subconjuntos de todo o conjunto de dados), calculamos a qualidade global dos dados (QGD) adicionando as qualidades individuais médias (*N* é o número de entradas no conjunto de dados) com as qualidades médias dos atributos (*M* é o número de atributos) e usando pesos para ajustar a importância dos atributos versus as qualidades individuais (equação 12).

$$QGD = \frac{w_{QDE}}{N} \sum_{i=1}^{N} QDE_i \times + \frac{w_{QDA}}{M} \sum_{j=1}^{M} QDA_j \tag{12}$$

8. Análise Preliminar

As métricas propostas foram aplicadas numa base de dados, anonimizada, com cerca de 20 milhões de entradas, na qual era possível apontar problemas decorrentes de erros de identificadores. Estes podiam ser detetados através da comparação de atributos entre diferentes pontos temporais e da verificação da incompatibilidade dos valores existentes, nomeadamente em (i) entradas duplicadas (referentes a um mesmo ponto temporal, mas com atributos diferentes), (ii) inconsistências nos atributos que definiam uma idade, e (iii) em atributos com variação demasiado alta entre pontos temporais sequenciais.
Uma vez que, para o conjunto de dados utilizado, os erros com maior potencial de provocar efeitos negativos eram os decorrentes da **precisão** de alguns atributos, optou-se por utilizar um peso maior para esta métrica e também um peso ligeiramente maior para a **unicidade**, em detrimento das métricas de **consistência** e **completude**.
Após o cálculo das métricas, foram extraídos os dados com diferentes linhas de corte e feita uma verificação do número de entradas filtradas em cada caso (Tabela 1).

Tabela 1. Percentagem de entradas filtradas por linha de corte.

Linha de Corte	Diferença de entradas (%)
0.4	0.00%
0.5	-0.20%
0.6	-0.22%
0.7	-0.25%
0.8	-0.56%
0.9	-11.04%

Optou-se por filtrar todos os dados com uma QDE inferior a 0.7, uma vez que permitia retirar apenas 0.25% dos dados, correspondendo a menos de 40 mil entradas.
Na análise preliminar efetuada aos dados filtrados, apesar do reduzido número de entradas retiradas, foi possível verificar uma diminuição significativa de potenciais entradas com erros graves (Tabela 2). Os erros considerados mais importantes, como entradas com atributos com valores improváveis (que neste caso correspondia ao atributo idade), foram filtrados 98% dos problemas, e as entradas com duplicação de dados com inconsistências também reduzidas em 46%. As restantes métricas tinham pesos menores e, como tal, provocaram um menor número de dados filtrados. Como exemplo, a existência de um número muito alargado de entradas com atributos não preenchidos, em conjunto com o menor peso utilizado para a métrica de completude resultou apenas numa leve diminuição do número de entradas com este problema.

Tabela 2. Entradas com erro e, destas, qual a percentagem que é filtrada.

Tipo de Erro	Dados com Erro (%)	Entradas com Erro Retiradas (%)
Completude	2.25%	-0.38%
Unicidade	0.02%	-46.24%
Valor Improvável	0.23%	-97.67%
Variação de Valor	0.33%	-7.36%
Erros Combinados	**0.36%**	**-67.98%**

Apesar de a avaliação do impacto das métricas separadas ser útil para perceber que tipo de erros estão a ser detetados e resolvidos, conjuntos de dados anonimizados, com problemas de identificadores, provocam normalmente problemas em diversas métricas em simultâneo. Como tal, é importante verificar que o conjunto de entradas com diversos erros combinados, potencialmente provocados por problemas nos identificadores, foi reduzido em quase 68%.
A análise apresentada permite ter apenas uma noção de como este tipo de avaliação de qualidade pode ser utilizada para melhorar os dados, sem que seja necessário cortar um número demasiado grande de dados. É, no entanto, importante efetuar uma análise mais detalhada, comparando diferentes pesos e linhas de corte, bem como qual o ganho efetivo em cada caso. O número de entradas com erros poderia ter tido uma redução maior através da utilização de uma linha de corte mais alta, resultando, no entanto, num maior número de entradas cortadas. É, assim, necessário efetuar uma avaliação aprofundada, de forma a entender qual o maior ganho de qualidade para um menor número de entradas cortadas.
É importante destacar ainda que a utilização de pesos adequados é essencial para a obtenção de cálculos de qualidade que sejam relevantes para o tipo de utilização que se pretende para os dados, derivada da avaliação preliminar e do tipo de erros que é necessário resolver.

9. Conclusões

Neste artigo, foi apresentada uma abordagem para obter uma classificação de qualidade de dados. A nossa proposta utiliza uma agregação de diferentes tipos de avaliação de qualidade. Além disso,

permite o cálculo de uma qualidade de dados global e utiliza pesos para ajustar a importância de cada métrica usada, para refletir melhor os objetivos do caso em estudo.
Os problemas de identificação, derivados de procedimentos de anonimização, a menos que filtrados, podem ter grande impacto na criação de modelos válidos dificultando o processo, nomeadamente por poderem produzir leituras incorretas dos dados, misturando entradas de diferentes entidades.
A abordagem aqui proposta permite calcular uma métrica individual de qualidade de dados que avalia quão fiável é uma entrada específica no que diz respeito à sua precisão, consistência, completude e unicidade. Além disso, propõe-se igualmente uma métrica de qualidade de dados a utilizar com resultados agregados extraídos das bases de dados.
As avaliações propostas têm a vantagem de permitir que o conjunto de dados seja filtrado dinamicamente, para permitir remover as entradas com má qualidade de acordo com limiares de qualidade específicos.
No que diz respeito ao trabalho futuro, existe o objetivo de avaliar como a utilização das ferramentas de análise de qualidade propostas podem melhorar as previsões feitas pelos algoritmos de prospeção de dados, comparando conjuntos de dados filtrados, com diferentes limiares de qualidade, bem como com conjuntos não filtrados.
Pretendemos também comparar conjuntos de dados com qualidade heterogénea com conjuntos homogéneos, para compreender como a existência de entradas com qualidade muito diversa influência a qualidade global e explorar a adição de uma medida que avalie esta variação.
Por último, as métricas propostas utilizam uma avaliação interna, considerando apenas os dados existentes na base de dados. É relevante adicionar uma avaliação que também permita fazer uma comparação entre os valores globais do conjunto de dados e os disponibilizados por fontes externas, como é o caso das instituições que fornecem informação estatística.

Agradecimentos. Este trabalho foi parcialmente financiado pelos projetos FCT, na unidade de investigação BioISI, ref. UID/MULTI/04046/2103, unidade de investigação LASIGE, ref. UIDB, UIDP/00408/2020 e DSAIPA/DS/0039/2018.

Referências

Batini, C., Cappiello, C., Francalanci, C. e Maurino, A. (2009). In *ACM Computing Surveys*, 41(3), article 16.

Batini, C. e Scannapieco, M. (2016). Data and Information Quality: Dimensions, Principles and Techniques. *Data-Centric Systems and Applications,* Springer.

Brickell, J., e Shmatikov, V.. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 70-78.

Buratović, I., Miličević, M. e Žubrinić, K.. (2012). Effects of Data Anonymization on the Data Mining Results. *2012 Proceedings of the 35th International Convention MIPRO*, 1619-1623.

Cai, L e Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the *Big Data* Era. In *Data Science Journal*, 14(2), 1-10.

Cavique, L., Pombinho, P., Tallón-Ballesteros, A. e Correia, L.. (2020) Data Pre-Processing and Data Generation in the Student Flow Case Study. *IDEAL 2020: International Conference on Intelligent Data Engineering and Automated Learning*.

Chen, H., Hailey, D., Wang, N., e Yu, P. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health.* 2014, 11, 5170-5207.

Coleman-Sebastian, L. (2013). Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. *(1st edition). Morgan Kaufmann, Elsevier.*

Ding, J. e Li, X. (2018). An Approach for Validating Quality of Datasets for Machine Learning. *2018 IEEE International Conference on Big Data*, 2795-2803.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM, 55(10),* 78-87.

Fletcher, S., e Islam M.Z.. Quality Evaluation of an Anonymized Dataset. *2014 22nd International Conference on Pattern Recognition*, 3594-3599.

GDPR, General Data Protection Regulations (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, OJ L 119, 4.5.2016, 1–88.

Heinrich B., Hristova, D., Klier, M., Schiller, A., e Szubartowicz, M. (2018). Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, ISSN: 1936-1963.

Inan, A., Kantarcioglu, M., e Bertino, E.. (2009). Using Anonymized Data for Classification. *IEEE International Conference on Data Engineering,* 429-440.

Jugulum, R. (2014). Competing with high quality data: Concepts, tools, and techniques for building a successful approach to data quality. *(1st edition) Wiley*.

Lukoianova, T., e Rubin, V. (2013). Veracity Roadmap: Is *Big Data* Objective, Truthful and Credible?. *Advances in Classification Research Online* 24(1), 4–15.

Pipino, L., Lee, Y. and Wang, R. (2002). Data Quality Assessment. In *Communications of the ACM,* 45(4), 211-218.

Sidi, F., Panahy, P., Affendey, L., Jabar, M., Ibrahim, H. and Mustapha, A. (2012). Data Quality: A Survey of Data Quality Dimensions. P*roceedings of the 2012 International Conference on Information Retrieval & Knowledge Management*. 300-304.

Silva, H. O., Basso, T., e Moares, R.. (2017). Privacy and data mining: evaluating the impact of data anonymization on classification algorithms. *2017 13th European Dependable Computing Conference,* 111-116.

P. Samarati. (2001) Protecting Respondents' Identities in Microdata Release, *IEEE Transactions on Knowledge and Data*, 13, (6), 1010-1027.

Engineering, Vol 13 (6), 2001, pp. 1010–1027. International Journal on Uncertainty, Fuzziness and Knowledgebased

Systems, 10 (5), 2002; 557-570.

Wang, R. e Strong D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumer. *Journal of Management Information Systems,* 12(4), 5-34.