Errors of Identifiers in Anonymous Databases: Impact on Data Quality

Paulo Pombinho[1(✉)], Luís Cavique[2], and Luís Correia[1]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
pmimatos@fc.ul.pt, luis.correia@ciencias.ulisboa.pt
[2] Universidade Aberta, Lisboa, Portugal
luis.cavique@uab.pt

Abstract. Data quality is essential for a correct understanding of the concepts they represent. Data mining is especially relevant when data with inferior quality is used in algorithms that depend on correct data to create accurate models and predictions. In this work, we introduce the issue of errors of identifiers in an anonymous database. The work proposes a quality evaluation approach that considers individual attributes and a contextual analysis that allows additional quality evaluations. The proposed quality analysis model is a robust means of minimizing anonymization costs.

Keywords: Data pre-processing · Anonymized data · Data quality

1 Introduction

Data quality is crucial in business and governmental usage to allow for a correct analysis and subsequent decisions to be made appropriately. However, a dataset with quality problems can imply high costs both on economic and social levels, with the possibility of erroneous decisions to be made when looking at incorrect data [1, 2].

The use of personal data following the GDPR [3] implies that datasets are often anonymized. Anonymization creates a challenge for evaluating the quality of the data since it prevents the identification and confirmation of errors, making it impossible to use another dataset to identify which values are correct.

These limitations make traditional metrics inadequate for assessing the quality of anonymized data [4]. Current research has focused more on anonymization procedures than data prospecting and quality assessment. In addition, if an improper anonymization procedure is performed, the potential usefulness of data prospecting can be ruined even by marginal privacy gains [5]. The essential factor for the success of data science projects is the use of the correct dimensions and their treatment, being the most time-consuming part of these projects [6].

Data pre-processing is essential for dealing with complexity in Business Analytics. However, most scientific and corporate literature proposes solutions with many synonyms depending on the methodology or tool used to answer these challenges. Data Pre-processing also appears with Data Wrangling or Feature Engineering designations.

P. García Bringas et al. (Eds.): SOCO 2022, LNNS 531, pp. 547–556, 2023.
https://doi.org/10.1007/978-3-031-18050-7_53

Different designations are also usual for each pre-processing stage, and data collection is often used as a synonym for data selection. Moreover, data cleaning appears as data cleansing or data quality. Four stages were chosen, and the pre-processing data stages are sequenced in the following pipeline:

Data Collection → Data Quality → Data Transformation → Data Reduction.

In what follows, Sect. 2, Data Collection, presents the data enrichment, types of errors, and noisy identifiers. Section 3, Data Quality, enumerates the different quality assessments used. In Sect. 4, we discuss the computational results of applying the pre-processing. In Sect. 5, we conclude this study.

2 Data Collection

We instantiate the pipeline applied to a dataset with over twenty million student enrollment data entries to better analyze each step. This data has been anonymized following the General Data Protection Regulation [3], with all students identified only by a number. Since this dataset is used as input to data mining algorithms, we must ensure that the data we have is as accurate as possible to minimize the noise given as input to the algorithms and maximize the quality of the modeling and predictions made.

In this section, we will describe the data selection and data cleaning processes.

2.1 Data Enrichment

The database used consists of information on student enrollments. Each entry concerns a student's enrollment in a specific school year with information detailing the characteristics of the enrollment and its outcome. It has the following relevant dimensions:

- School year (e.g., 2008–2009)
- Grade (e.g., 12^{th} grade)
- Nature (e.g., public or private)
- Geographic Information
- Enrolment event
- Age and gender

Since we want to model the student flow, we need to correctly perceive how a student's path through the education system changes and what events are associated with these changes. We analyze how a student's enrollment changes in sequential years and add an attribute to characterize the student's state. We consider the path of a student as a sequence of states with three different types of state inputs and three types of outputs (Fig. 1):

- Dropout: there is no information about the student in the next school year,
- Retention: the student is in the same grade in the following school year,
- Pass: the student is enrolled in a higher grade in the following school year,
- Ingress: there is no information about the student in the previous school year.

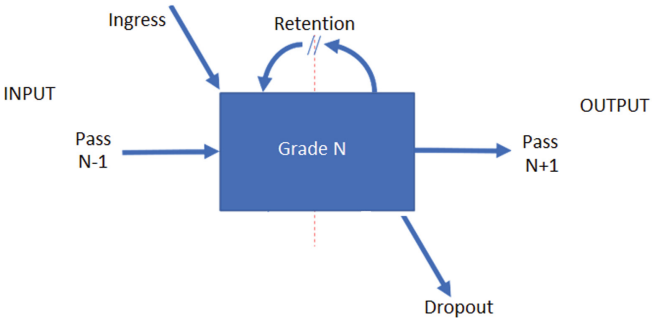

Fig. 1. Types of input/output enrolment events.

2.2 Errors of Identifiers

Identifier problems are derived from the anonymization procedures and can sometimes produce problems in the data. Since we have no ground truth to compare the results, these problems are more challenging to detect and correct than non-anonymized data. Identifier problems can result in severe problems if we analyze the relationship between different entries of the same real-world entity. Suppose the anonymization process has introduced noise in the primary/foreign keys of the data. In that case, we may find that a single entity has multiple data entries collected at certain time intervals in datasets consisting of time series.

Figure 2 shows how incorrect identifiers in two sequential entries can create a wrong and potentially severe deviation in the perception of reality. Figure 2 (a) shows how overlapping identifiers in two different students may produce an incorrect reading of

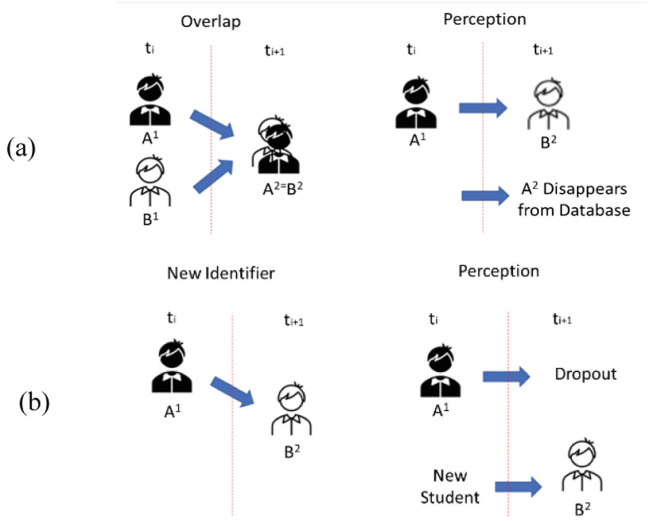

Fig. 2. (a) overlapping of two different students' identifiers (b) incorrect creation of a new identifier

reality. Instead of two different students passing grades, the overlap of the first student over the second will produce an incorrect result where student A will, in the next year, transition to the grade year of student B instead of its own. In contrast, student B will be shown as dropping out of school. Likewise, Fig. 2 (b) gives an example of how the attribution of a new identifier to an existing one can also cause issues. In this instance, student A will be perceived as dropping out in the next year, while a new student will be shown enrolling for the first time in the next year.

Since, in some instances, these errors create patterns that can be mistaken for normal data, they are not easily detected. Therefore, there is a need to use a set of different metrics that, in their combination, can allow us to detect these types of problems.

3 Data Quality

In this paper, we propose the combination of different types of analysis that allow for more robust classification of each data entry quality and the different data dimensions themselves. We adapt the nomenclature proposed in [7–9] and consider four types of individual quality assessment:

Completeness evaluates whether all attributes of an entry are populated, or values are missing in some attributes.

Consistency assesses whether the information contained in the entry's attributes must be normalized during the ETL (Extract, Transform, Load) processing phase. Alternatively, it also assesses if entered data was not formatted according to the rest of the dataset.

Uniqueness identifies the entries as duplicated if, for example, two entries in the same period are conflicting.

Accuracy measures the accuracy of the attribute values of a given entry and allows the identification of entries with incorrectly populated attributes. For example, let us consider the accuracy of a numeric attribute. We can identify scenarios where the value is invalid because it is outside the domain of possible values, such as negative values when only positive values are valid. Regarding the longitudinal assessment of the dataflow, we consider an accuracy assessment that analyses three different contextual dimensions:

I. Attribute variations – evaluates situations where a change in the value of an attribute in two consecutive time points of the same entity is greater than expected;
II. Value variation – performs a calculation that evaluates the probability of a value of an attribute considering the other attributes of the duplicate entry;
III. Attribute Incompatibility – compares, in the duplicate entry, the different attributes for which it is known that there is a correlation in search of incompatible values.

In the following subsections, we will describe the calculation of the individual quality metrics and the use in calculating the overall quality for each entry.

3.1 Quality Metrics

We calculate four different data quality metrics using the types defined in the previous section. These range between 0 and 1; '0' represents zero quality and '1' a perfect quality.

To calculate the **completeness** metric, we used information regarding the missing attribute count. The summation argument uses an Iverson bracket[1] to count the number of missing attributes, in entry a, about the total number of attributes, M. (Eq. 1).

$$Completeness = 1 - \frac{1}{M}\sum_{i=1}^{M}\left[a_i \; is \; missing\right] \tag{1}$$

Data **consistency** is also classified, similarly to the completeness, using the count of attributes with consistency issues over the total number of attributes, M (Eq. 2).

$$Consistency = 1 - \frac{1}{M}\sum_{i=1}^{M}[a_i \; is \; not \; consistent] \tag{2}$$

Uniqueness is classified according to the identification of duplicate entries if the identifier is in more than one entry at the same time data point and, in case this happens, evaluating how different the duplicate entries are (Eq. 3).

$$Uniqueness = \begin{cases} 1, & not duplicated \\ 0.8, & duplicate \; with \; minor \; differences \\ 0, & duplicate \; with \; major \; differences \end{cases} \tag{3}$$

The **accuracy metric** evaluates the accuracy and probability of specific attribute values. As indicated earlier, we classify an attribute as invalid if the value is outside the domains of that attribute. However, this analysis can be improved by using semantic information of expected attribute values, considering other contextual information, and allowing a classification with different degrees of probability (Eq. 4). V is the value of the attribute, V_{max} and V_{min} are then valid maximum and minimum values, and VP_{max} and VP_{min} are the respective maximum and minimum values likely for the current context of the entry.

$$Accuracy\_Type\_I(V) = \begin{cases} 1, & VP_{min} < V < VP_{max} \\ \frac{V - V_{min}}{VP_{min} - V_{min}}, & V_{min} < V < VP_{min} \\ 1 - \frac{V - VP_{max}}{V_{max} - VP_{max}}, & VP_{max} < V < V_{max} \\ 0, & V \langle V_{min} \vee V \rangle V_{max} \end{cases} \tag{4}$$

This metric may calculate how likely a student's age is, given his enrolled grade. The first component of the equation gives maximum quality to ages inside the valid, and expected, intervals, while, on the opposite, ages outside the valid values will have zero quality. Simmilarly, the second and third components of the equation evaluate the quality of valid age value, but below or above the likely values interval, respectively.

Similarly, an attribute can be classified using information about the variation of the current V value (obtained comparing with the previous data point of the same entity) and

[1] Iverson bracket https://en.wikipedia.org/wiki/Iverson_bracket.

its comparison with the minimum (most probable) and maximum variations possible, respectively, V_{min} and V_{max} (Eq. 5).

$$Accuracy\_Type\_II(V_a) = 1 - \frac{|V| - |V_{min}|}{|V_{max}| - |V_{min}|} \tag{5}$$

Like the previous metric, the value variation can be used to assess, for example, the change in grades of a specific student, allowing the detection of abnormal jumps in the student's path in the education system. This way expected variations of only one school year will have maximum quality while higher variations will decrease the quality value.

Finally, the accuracy of attributes heavily influenced by others can also be assessed by comparing pairs of attributes, searching for values that should not coexist, using a two-argument function (Eq. 6).

$$\text{Accuracy\_Type\_III}(V_1, V_2) = \begin{cases} 1, V_1 \textit{ compatible with } V_2 \\ 0, V_1 \textit{ incompatible with } V_2 \end{cases} \tag{6}$$

The compatibility metric can identify situations where two attributes are incompatible; for example, a student enrolled in early grades but a modality exclusive of higher education grades.

After the calculations of the different types of accuracy metrics, the overall accuracy is defined as the average of its different components in all M attributes (Eq. 7).

$$Accuracy = \frac{\sum_{i=1}^{M} Acc\_I_i + \sum_{i=1}^{M} Acc\_II_i + \sum_{i=1}^{M} Acc\_III_i}{3M} \tag{7}$$

3.2 Aggregated Data Quality

The quality assessment metric described in the previous section allows us to understand each entry and its reliability in each dimension. Each metric helps identify potential problems with the data. However, because more severe data problems are expected to simultaneously create quality problems in multiple dimensions (such as problems that arise from inconsistencies with identifiers), we use the combination of the different metrics to compute the overall quality for each entry in the database.

For each entry, the Aggregated data Quality (AQ) is obtained as the weighted sum of all metrics previously described and is classified as a value from 0 (no quality) to 1 (best quality) (Eq. 8). All weights are positive, and their sum is equal to 1.The use of weights on each metric allows us to emphasize which metrics are most important for a specific dataset or task. If all metrics are of equal importance then each weight will have equal value.

$$\begin{aligned} AQ_i = {} & w_{Comp}.Completness_i + w_{Cons}.Consistency_i \\ & + w_{Uniq}.Uniqueness_i + w_{Acc}.Accuracy_i \end{aligned} \tag{8}$$

The data quality allows us to perform initial filtering of data entries that have a quality below some threshold and recognize identifiers associated with potential identification problems and report them to be fixed. In the analysis performed, and described in the next section, we used a weight of 0.4 for the accuracy and 0.2 for the other three metrics.

4 Discussion

4.1 Data Cleaning

The proposed metrics were applied in the dataset described in Sect. 3, in which it was possible to point out problems arising from identifier errors. These could be found by comparing attributes between different time points and verifying the incompatibility of existing values, namely in (i) duplicate entries (referring to the same time point but with different attributes), (ii) inconsistencies in the attributes that defined an age, and (iii) in attributes with too high variation between successive time points.

After calculating the metrics, the data with different cutting lines were extracted, and the number of entries filtered in each case was checked (Table 1).

Table 1. Percentage of removed lines by criteria.

Criteria	Removed lines (%)
AQ $\geq$ 0.4	0.00%
AQ $\geq$ 0.5	0.20%
AQ $\geq$ 0.6	0.22%
AQ $\geq$ 0.7	0.25%
AQ $\geq$ 0.8	0.56%
AQ $\geq$ 0.9	11.04%

We chose to filter all data with an AQ below 0.7 since it allowed to remove only 0.25% of the data, corresponding to less than 40,000 entries.

4.2 Impact of the Anonymization on the Data Quality

In the preliminary analysis of the filtered data, despite the small number of entries withdrawn, it was possible to verify a significant decrease in entries with serious errors.

The errors considered to be more serious, such as entries with attributes with values with a low probability (Accuracy_type_I), which corresponded to improbable ages, considering the grade the students were enrolled in, allowed filtering out 98% of the existing problems (Table 2). Moreover, the entries with inconsistent data duplication were reduced by 46%. The remaining metrics had smaller weights and, as such, caused a smaller number of filtered data. For example, the existence of many entries with unfilled attributes and the lower weight used for the completeness metric resulted only in a slight decrease in the number of entries with this problem.

Evaluating the impact of each metric separately helps understand what type of errors are present in the database allowing the evaluation of potential solutions to solve them. For example, almost all the consistency errors present in the preliminary analysis could be solved by using several scripts that normalized some attributes in the database that were identified as having problems.

Table 2. Types of errors with occurrences and percentage removed.

Type of Error	Lines with Error (%)	Errors Removed (%)
Completeness	2.25%	0.38%
Uniqueness	0.02%	46.24%
Accuracy Type I	0.23%	97.67%
Accuracy Type II	0.33%	7.36%
Combined Errors	0.36%	67.98%

However, problems caused by identifier errors derived from the anonymization procedures are harder to identify as they affect only a smaller subset of the data. Moreover, as these types of errors deal with the whole entry and not only one of its attributes, they will trigger inconsistencies across several of the affected entries attributes, which will, in turn, affect multiple metrics simultaneously.

If, for example, two students' data are mixed due to an overlapping identifier, there will probably be inconsistencies in several attributes. In this type of scenario, the students' data could, for instance, show a jump in a curricular year, an age that would not match the curricular year she is enrolled in, a wrong teaching modality for the students' course or even a duplicate enrollments in the same school year. For this reason, it is precious to be able to detect these problems.

The fact that anonymization issues create more significant problems can be used to identify them as entries with several combined errors and, thus, a lower quality.

Using the aggregated data quality allowed us to identify most identifier problems. With aggregated data quality, it was possible to reduce, by almost 68%, the number of entries that had a combination of errors.

The analysis presented gives only a notion of how this type of quality assessment can be used to improve the data without cutting too many entries. However, performing a more detailed analysis is essential, comparing different weights and cutting lines and the effective gain in each case. The number of entries with errors could have been reduced further by using a higher cutting line, resulting in a more substantial number of filtered entries. Therefore, it is necessary to conduct an in-depth evaluation to find the best compromise between the most significant quality gain and the smaller number of cut entries.

It is also important to highlight that appropriate weight values are essential to obtain quality values relevant to the intended type of use. Consequently, the weight values need to consider the data derived from the preliminary assessment and the type of errors that need to be resolved.

5 Conclusions

This paper discusses the impact of GDPR and anonymization on data quality. Anonymization procedures introduce new errors that must be taken into account. Our proposal uses an aggregation of distinct types of quality assessment. In addition, it allows

the calculation of global data quality. It uses weights to adjust the importance of each metric used, better to reflect the objectives of the case under study.

Unless filtered, identifier problems derived from anonymization procedures can significantly impact the creation of valid models, making the process complicated because they can produce incorrect readings of the data, mixing entries from different entities.

The proposed approach calculates an individual data quality metric that assesses how reliable a specific entry is regarding its accuracy, consistency, completeness, and uniqueness.

Concerning future work, the objective is to evaluate how the proposed quality analysis tools can improve predictions made by data mining algorithms by comparing filtered data sets with different quality thresholds and unfiltered sets.

We will also compare datasets with heterogeneous quality with homogeneous quality datasets to understand how the existence of inputs with very diverse quality influences the overall quality and explore the addition of a measure that evaluates this variation.

The proposed metrics use an internal evaluation, considering only the existing data in the database. It is essential to add an assessment that also allows a comparison between the overall values of the data set and those made available from external sources, such as institutions that provide statistical information.

Finally, the application of GDPR has hidden costs that must be considered. As anonymization is unavoidable, the proposed quality analysis model is a robust means of minimizing these costs.

Acknowledgments. The authors would like to thank the FCT Projetct of Scientific Research and Technological Development in Data Science and Artificial Intelligence in Public Administration, 2018–2022 (DSAIPA/DS/0039/2018), for its support, and also acknowledge support by BioISI (UID/MULTI/04046/2103) and LASIGE Research Unit (UIDB/00408/2020, UIDP/00408/2020) center grants.

References

1. Batini, C., Scannapieco, M.: Data and information quality: dimensions, principles and techniques. Data-Centric Systems and Applications, Springer (2016). https://doi.org/10.1007/978-3-319-24106-7
2. Heinrich, B., Hristova, D., Klier, M., Schiller, A., Szubartowicz, M.: Requirements for data quality metrics. J. Data Info. Quality 1936–1963 9 (2), 1–32 (2018)
3. GDPR, General Data Protection Regulations: 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, pp. 1–88 (2016)
4. Fletcher, S., Islam M.Z.: Quality evaluation of an anonymized dataset. In: 22nd International Conference on Pattern Recognition, 3594–3599 (2014)
5. Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 70–78 (2008)
6. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
7. Pipino, L., Lee, Y., Wang, R.: Data quality assessment. Commun. ACM **45**(4), 211–218 (2002)

8. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14**(2), 1–10 (2015)
9. Sidi, F., Panahy, P., Affendey, L., Jabar, M., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management, pp. 300–304 (2012)