

Anonymized Data Assessment via Analysis of Variance: An Application to Higher Education Evaluation

Maria Eugénia Ferrão[1,2], Paula Prata[1,3(✉)], and Paulo Fazendeiro[1,3]

[1] University of Beira Interior, Covilhã, Portugal
{meferrao,fazendeiro}@ubi.pt, pprata@di.ubi.pt
[2] Centre for Applied Mathematics and Economics (CEMAPRE), Lisbon, Portugal
[3] Instituto de Telecomunicações (IT), Covilhã, Portugal

Abstract. The assessment of the utility of an anonymized data set can be operationalized by the determination of the amount of information loss. To investigate the possible degradation of the relationship between variables after anonymization, hence measuring the loss, we perform an *a posteriori* analysis of variance. Several anonymized scenarios are compared with the original data. Differential privacy is applied as data anonymization process. We assess data utility based on the agreement between the original data structure and the anonymized structures. Data quality and utility are quantified by standard metrics, characteristics of the groups obtained. In addition, we use analysis of variance to show how estimates change. For illustration, we apply this approach to Brazilian Higher Education data with focus on the main effects of interaction terms involving gender differentiation. The findings indicate that blindly using anonymized data for scientific purposes could potentially undermine the validity of the conclusions.

Keywords: Data anonymization · Differential privacy · Data utility · Data quality · ENADE

1 Introduction

The challenge of monitoring sociodemographic gaps in educational outcomes shares some characteristics with other regulatory problems. Regarding the achievement of social justice, the cycle of public policies requires statistical evidence to inform policymakers. How can they promote a just society and what actions can they take to achieve this goal? The Brazilian education evaluation system is mentored to provide such evidence. Specifically, the Exame Nacional de Desempenho dos Estudantes [National Examination of Students Performance] (ENADE) [1] refers to higher education, and allows the computing of several indicators and measures. In addition, ENADE data have been used for scientific research in several scientific fields [2]. For instance, researchers [3] may ask "Are exam-based evaluative methods really adequate to evaluate institutions and courses?", "Does students' performance in standardized tests enable a reliable assessment about the quality of undergraduate courses?" These are just a few valuable research

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14105, pp. 130–141, 2023.
https://doi.org/10.1007/978-3-031-37108-0_9

questions on the topic. It is not our main concern to answer any of these questions in this study. Instead, we stand the need of conducting scientific research to support these or other related research questions. We are also certain on the benefits of leveraging evidence-based public policy and, thus, on the potential use of educational data assessment for improvement of education systems. The same is to say that making science for policy in social sciences implies the statistical modeling of administrative data. Thus, with such promising potential benefits in mind, our main concern is the data privacy issues which have emerged with the General Data Protection Regulation (GDPR) initiatives in effect across several regions around the world. In fact, restrictions to microdata access, even when the access intends scientific use, may compromise the contribution of scientific research for social justice.

At least since the 1970s, the subject of data privacy has been discussed in the specialized literature [4, 5]. Despite the great methodological developments, the adoption of data anonymization procedures still does not meet the needs for scientific research purposes in social and behavioral sciences [6]. Following previous work [7, 8], we intend to show obstacles of anonymized data by differential privacy for statistical modelling applied to higher education data.

The article is organized as follows: Sect. 2 presents methodological aspects related to ENADE database, it includes some details of differential privacy as the method applied here for the purpose of data anonymization and its implementation. It also includes a subsection on methods to assess the quality and utility of anonymized data. Section 3 presents the main results and interpretation. Finally, the conclusion is in Sect. 4.

2 Methodology

2.1 Dataset Description

Since 2004, the National Student Performance Exam (ENADE) has been held annually in Brazil to evaluate the performance of graduates from higher education institutions across various dimensions and skills [1, 9]. As a component of Brazilian higher education evaluation system (Sinaes), which includes institutional and program evaluations, the ENADE is instrumental in monitoring the quality indicators of higher education. Participation in the exam is mandatory for all students, and the assessment instruments cover multiple domains based on their field of study. The evaluation cycle for various academic programs is conducted over a three-year period, with different disciplinary areas being assessed each year, including baccalaureate and degree courses in health sciences, agrarian sciences, engineering, architecture, humanities, social sciences, and higher technology courses in various domains.

In 2019, 433930 students were involved (ENADE 2019 - version updated in May 13th, 2021). The exam generates a set of detailed information on the participating students, as well as the courses and Higher Education Institutions evaluated, that is, the microdata composed of 139 variables. This article considers both the students' specific score, such as their grade point average, *gpa1*, and their general score, *gpa2*, as well as their sociodemographic variables, including gender, age, self-declared race/skin color, mother's and father's education levels, working status, family income, and region. We introduced two new variables: (i) the "number of years needed to finish the graduate

studies", *years2finish*, computed by the difference between the current year (2019) and the "year of beginning graduation" plus one, and (ii) the number of years of school leave, *years_leave*, computed by the difference between the starting year of graduation, *year_in*, and the last year of secondary school, *year_hs*. Since some values recorded in the data set were not plausible according to the purposes of ENADE and the Brazilian Educational System, it was necessary to pre-process the dataset.

After removing missing and duplicated data, the original data set was reduced to 89.68% of the original size, that is, 389180 rows. Moreover, it was necessary to remove nonsensical data. Notice that in Brazil the academic year agrees with the civil year, the exam occurs at the end of the course (2019 in the present case) and the course takes at least 3 years to be concluded. We selected only the cases where the starting year of graduation, *year_in*, lies in the interval [2002, 2017]. To further remove outliers, the values of the last year of secondary school, *year_hs* were filtered to the interval [1992, 2017]. The cases where the last year of secondary school was bigger than the starting year of graduation were eliminated. Students aged under 16 were also removed from the dataset. This whole process of data cleansing resulted in the elimination of 72343 records from the downloaded data set. Thus, the resulting data set, with 361587 (83.32% of the initial size) records, will be hereafter referred to in this paper as the original (prior to anonymization) data set, DS0.

Table 1, shows for each of the selected variables, the label used in the experiments, its range in the original data set and its classification for the purpose of data anonymization. In the process of data anonymization the attributes are classified as direct identifiers, quasi-identifiers, sensitive, or insensitive, according to their characteristics [10]. Direct identifiers, that is, the ones that are associated with a high risk of re-identification were not present in the ENADE data. A quasi-identifier is an attribute that in combination with quasi-identifiers from other data sets can be useful for re-identification. Six quasi-identifiers were considered in this study: region, age, gender, race/skin color and the two computed variables *years2finish* and *years_leave*. The other attributes were considered as insensitive, that is, they are not associated with any privacy risk.

2.2 Differential Privacy

The concept of differential privacy (DP) was introduced in 2006 [11, 12] as a mathematical definition of privacy designed to statistical data analysis. According to [11], a randomized function R gives ε-Differential Privacy (ε-DP) if for all data sets D1 and D2 differing on at most one record, and for all potential output of R that could be predicted, S ⊆ Range(R),

$$\Pr[R(D1) \in S] \leq \exp(\varepsilon) * \Pr[R(D2) \in S] \quad (1)$$

Informally speaking, "differential privacy protects the privacy of individuals by requiring that the information of an individual does not significantly affect the output" of any computation that can be performed on the data [13]. The ε parameter, called privacy budget, is a measure of privacy loss. A variant of DP is "approximate differential privacy" introduced in [14] and called (ε, δ)-DP. It is a relaxation of pure DP where the guarantee of Eq. (1) needs to be satisfied only for events whose probability is at least ≈ 1–δ [15, 16].

Table 1. Selected variables, respective scales in DS0 and attribute classification.

Variable	*Label*	*Range*	*Attribute type*
University id	uni_id	Between 1 and 23410	Insensitive
Program id	prog_id	Between 3 and 5,001,378	Insensitive
Region	reg	1 = North (N) 2 = Northeast (NE) 3 = Southeast (SE) 4 = South (S) 5 = Central-West (C-W)	Quasi-identifier
Age	age	Between 19 and 86	Quasi-identifier
Gender	gen	M = Male F = Female	Quasi-identifier
Grade point average (specific score)	gpa1	Minimum = 0; Maximum = 93.0	Insensitive
Grade point average (general score)	gpa2	Minimum = 0; Maximum = 98.0	Insensitive
Race / skin color	race	A = White B = Black C = Yellow D = Pardo E = Indigenous F = Not declared	Quasi-identifier
Father's education	edu_f	1 = None 2 = 1st – 5th grade 3 = 6th – 9th grade	Insensitive
Mother's education	edu_m	4 = Secondary school 5 = Graduation 6 = Post-graduation	Insensitive
Family's income	inco	1 = Up to 1.5 minimum wages 2 = 1.5 to 3 minimum wages 3 = 3 to 4.5 minimum wages 4 = 4.5 to 6 minimum wages 5 = 6 to 10 minimum wages 6 = 10 to 30 minimum wages 7 = Above 30 minimum wages	Insensitive
Working status	work	A = Not working B = Work eventually C = Work up to 20 h a week D = Work 21 to 39 h a week E = Work 40 h a week or more	Insensitive

(*continued*)

Table 1. (*continued*)

Variable	*Label*	*Range*	*Attribute type*
Scholarship for non-local college attendance	schol	A = None B = Housing allowance C = Meal allowance D = Housing and food allowance E = Permanence allowance F = Other type of aid	Insensitive
Academic Scholarship	ac_schol	A = None B = Scientific initiation scholarship C = Extension scholarship D = Monitoring/tutoring sch E = PET scholarship F = Other type of academic sch	Insensitive
Number of years to finish the graduation	years2finish	Between 3 and 18	Quasi-identifier
Years between high school and college	years_leave	Between 0 and 25	Quasi-identifier

Nowadays, several implementations of DP techniques are ready to use. Some change the data by addition of random noise [17–20] and others by random sampling [16, 21]. In this work, we used the SafePub algorithm [16] that implements (ε, δ)-DP using the last approach. This algorithm, available in the open source ARX - Data Anonymization Tool, starts by randomly sampling the dataset. Usually, in a previous step users define generalization hierarchies for the quasi-identifiers for which they intend to decrease precision, for instance by changing the scale or order of magnitude. In the ENADE data, some of the attributes are already generalized as can be seen in Table 1 (see e.g., Father's and Mother' education or Family's income). In this work no hierarchy was defined for the studied attributes. From the parameters ε and δ the algorithm derives a value of k and proceeds performing k-anonymization [22]. This means that all the records appearing less than k times in the current dataset are suppressed. That process is repeated for each possible generalization hierarchy, searching for an optimal solution for a utility model chosen by the user. Finally, the output satisfies k-anonymity and has approximate differential privacy, (ε, δ)-DP, as shown in [16].

2.3 Utility and ANOVA Modelling

The assessment of the utility of an anonymized dataset should be based on data purpose, i.e. utility models typically evaluate data utility by quantifying the amount of information loss. For example, by measuring differences or similarities between the input

and the output dataset [23]. Other examples include the sum of squared errors, average distinguishability [24], non-uniform entropy [25] and discernibility [26].

In addition, the analysis of variance (ANOVA) is a useful statistical tool for analyzing differences between groups and can provide valuable insights into the relationships between different variables. In this work the analysis of variance model ($\alpha = 0.05$) is used to statistically compare the results between the original and the anonymized datasets. It is applied to identify student's sociodemographic characteristics (gender, race/skin color, mother's education, family's income, working status, region) which, *ceteris paribus*, may influence (specific) grade point average, *gpa1*, and (general) grade point average, *gpa2*. The statistical model includes the main effects and the following 1st order interactions: gender–working status, gender–race/skin color, gender–family's income, gender–region. Other significant interactions were explored with a Bonferroni correction. The substantive evidence will be presented in Table 5 with the purpose of showing the terms that become not statistically significant depending on the anonymization scenario. Details about the ANOVA model may be found in Scheffé [27].

3 Results

After preprocessing the ENADE 2019 data, the resulting data set, DS0, was anonymized using the ARX tool (version 3.9.1). The privacy model (ε, δ)-DP was applied varying the values of ε and δ for five cases illustrated in Table 2. In the data sets DS1, DS2 and DS3 the value of δ was fixed ($\delta = 0.01$) and the value of ε decreases from 3 to 1. In the data sets DS3, DS4 and DS5 the value of ε was fixed ($\varepsilon = 1$) and the value of δ decreases from 0.01 to 0.0001. For all data sets the process of anonymization was configured with the following parameters: suppression limit, 100%; utility measure, loss; aggregate function, rank; the search strategy was "best-effort, button up" with a limited number of 5 000 steps; all the remaining parameters took the ARX default values.

Table 2. Privacy model, number of records, value of k and estimated risk for each dataset.

Dataset	*Privacy model*	*Nº of records*	*K value*	*Estimated risk*
DS0	-	361587	-	100%
DS1	(3, 0.01)-DP	186245	68	1.5%
DS2	(2, 0.01)-DP	185636	38	2.6%
DS3	**(1, 0.01)-DP**	**144332**	**19**	**5.2%**
DS4	(1, 0.001)-DP	127032	33	3.0%
DS5	(1, 0.0001)-DP	117145	45	2.2%

From Table 2, it is possible to observe that, for the studied data set, the number of records remaining after anonymization decreases when the value of ε decreases keeping the value of δ, as well as the number of records decreases when the value of δ decreases keeping the value of ε. Despite DS1 having the largest number of records, it has the

highest value of k. This may seem counter-intuitive because a high value of k could lead to the assumption of a large deletion of records, but as rightfully noted in [28] "privacy is a subjective concept, it varies from person to person, from time to time, even for the same person".

Finally, the last column of Table 2 presents the re-identification risk estimated by ARX for each data set. Usually, three attacker's profiles are considered leading to journalist, prosecutor, and marketer risk. These classifications are related with the way of modelling the adversary's background knowledge about an individual [29]. The models are described in [30] and it is shown that prosecutor risk $\geq$ journalist risk $\geq$ marketer risk. We present the value obtained for the attacker's model that results in the highest risk of re-identification, that is, the estimated prosecutor risk. As can be observed, DS1 has the lowest risk, there just 1.5% of the records are at risk of re-identification. The data set with the lowest value of k, DS3, has the highest risk, that is, 5.2% of the records are at risk of re-identification.

Following the three main steps of data anonymization, apply a privacy model, quantify the risk of re-identification and assess data utility [23], next we present the results of measuring the quality of the studied data sets starting by quality models provided by ARX. Table 3 shows for each data set and for each quasi-identifier the value of the non-uniform entropy [31]. It compares the frequencies of quasi-identifier values in each anonymized dataset with the according frequencies in the original dataset. The ARX tool scales and normalizes the values returned by all the implemented quality models into the range [0, 1] where 0 represents a dataset from which all information has been removed and 1 represents the unmodified input dataset [23]. From Table 3, we can see that DS3 has the best values for the non-uniform entropy and DS5 the worst ones.

Table 3. Quality models: non-uniform entropy for the quasi-identifiers.

Dataset	*region*	*age*	*gender*	*race*	*years2finish*	*years_leave*
DS1	51.5%	48.4%	54.8%	45.4%	48.1%	43.5%
DS2	58.2%	55.0%	61.1%	52.2%	54.9%	50.5%
DS3	**63.9%**	**61.1%**	**67.8%**	**58.3%**	**60.8%**	**57.3%**
DS4	55.8%	52.6%	58.8%	49.6%	52.3%	48.0%
DS5	51.0%	47.9%	54.2%	44.9%	47.6%	42.9%

Table 4 presents for each data set and for each quasi-identifier the values of the sum of squared errors (SSE). The SSE [32] calculates the sum of squares of attribute distances between records in DS0 and their versions in the anonymized data set.

As can be seen from Table 4, the best results are obtained again for DS3. If we notice that DS3 is the data set with the highest re-identification risk, this is an expected result. More useful information means greater risk.

Beyond these standard quality metrics to further enhance the quality assessment we apply the ANOVA model to the studied data sets with *gpa2* as the dependent variable. Considering DS0 as the data reference, the estimates obtained suggest that main effects

Table 4. Quality models: sum of squared errors for the quasi-identifiers.

Dataset	*region*	*age*	*gender*	*race*	*years2finish*	*years_leave*
DS1	49.8%	60.2%	55.2%	59.1%	55.65	61.0%
DS2	56.4%	66.6%	61.5%	65.1%	61.8%	67.1%
DS3	**62.3%**	**72.4%**	**67.1%**	**68.0%**	**67.5%**	**72.1%**
DS4	54.1%	64.7%	59.2%	60.2%	60.0.%	64.8%
DS5	49.3%	60.1%	54.6%	55.9%	55.0%	60.5%

and interactions on *gpa2* are statistically different from zero. The proportion of variance explained by the model (R^2) is 38.3%. The Table 5 pinpoints the terms that do not present interactions statistically different from zero at a level of significance of 5%, i.e., the 'X' marks terms not statistically different from zero at 5%. Depending on the dataset the results change. For example, ANOVA model applied to DS4 and DS1 suggests that gender has no effect on *gpa2*, nor gender interactions with race/skin color or family income. Another example, by applying the model to DS2 the only interaction term that is statistically non-significant is gender-income.

Table 5. Original and anonymized data ANOVA results: gpa2 as dependent variable.

Dataset *Variable*	*DS0*	*DS1*	*DS2*	*DS3*	*DS4*	*DS5*
Region	X	-	-	-	-	-
Gpa1	X	-	-	-	-	-
Gender	X	X	-	-	X	-
Race/Skin color	X	-	-	-	-	-
Mother's education	X	-	-	-	-	-
Family's income	X	-	-	-	-	-
Working Status	X	-	-	-	-	-
Gender – Working status	X	-	-	-	-	-
Gender – Race/skin color	X	X	-	X	X	X
Gender – Family's income	X	X	X	X	X	X
Gender – Region	X	-	-	-	-	-
Adjusted R^2	38.3%	36.8%	36.8%	37.0%	36.9%	36.7%

As a matter of fact, it is possible to verify that several terms cease to present statistically significant effects according to the remaining data modelling of the anonymized datasets. For instance, DS3 and DS5 seem to indicate that the interactions of gender

with race/skin color or with family income potentially have no effect on gpa2 – contrary to what is verified in the original dataset. Please notice that the variance explained is approximately 37% for every model.

To emphasize the difficulty of identifying the loss of the effects of some of the interaction Gender-Income term after anonymization Fig. 1, 2 and 3 depict the results of some visualization tools traditionally used in exploratory data analysis. The estimated marginal means of *gpa2* for the interaction gender with family income is reported in Fig. 1 for DS0 and in Fig. 2 for DS3, whereas the correlation between all the variable pairs is visually conveyed via bivariate association in the heatmap of Fig. 3.

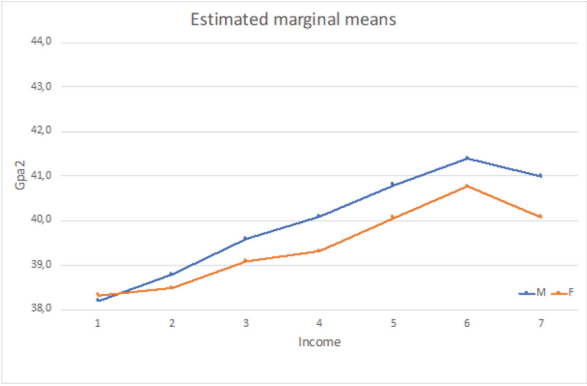

Fig. 1. Gender–Family's income estimated marginal means for DS0

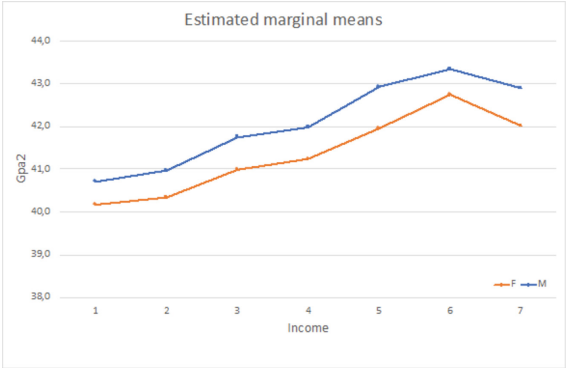

Fig. 2. Gender–Family's income estimated marginal means for DS3

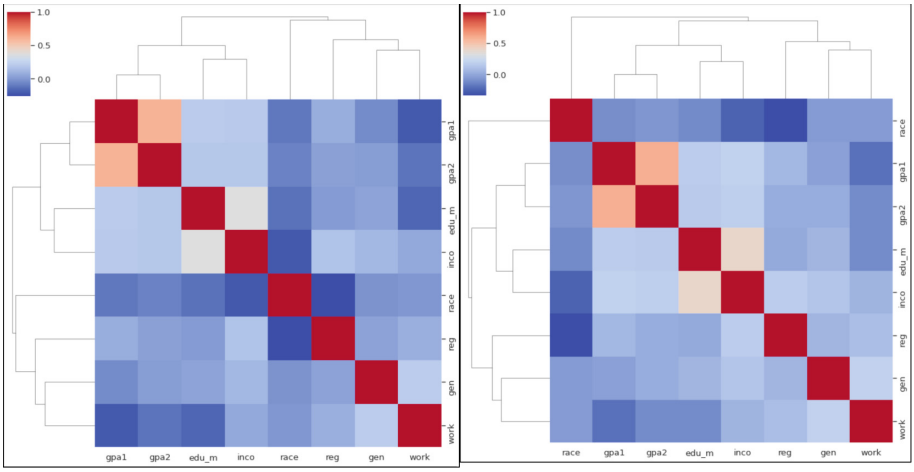

Fig. 3. Correlation matrix for DS0 and DS3

4 Conclusion

In this study, we used the research topic of equity in higher education to demonstrate that even the most suitable anonymization method for big data sets, i.e. differential privacy [20], can become ineffective for anonymizing data whose primary purpose is scientific research in social and behavioral sciences. The quantitative research on equity in higher education is based on the correlational structure between the variables involved. Therefore, we used as an example the Brazilian data from the National Student Performance Exam (ENADE) to investigate the relationship between various sociodemographic variables and the specific performance of students at the end of the course through an ANOVA model.

We particularly focused on the main effects of interaction terms involving gender differentiation. The risk of re-identification for anonymized datasets is around 5% or less. Considering the dataset that is closest to the original, based on two quality and/or utility metrics (entropy and sum of squared errors), the substantive conclusion that can be drawn from modeling anonymized data is that the two interactions "Gender – Race" and "Gender - Family's income" are not statistically significant (at a significance level of 5%). In fact, we found the opposite with the original data! This means that the object of research, that is equity in higher education, may be distorted based on anonymized data. The same is to say that anonymization may lead to a reduction in the amount of information available, and in turn, to a loss of statistical power to detect significant effects or interactions.

While the analysis and modeling of microdata from administrative records are essential for scientific research in various fields [6], the implementation of GDPR has brought additional challenges. In some countries, its implementation makes such research impossible. At the time of publication of this article, neither in Portugal nor in Brazil would it be possible to conduct research on equity in higher education based on recently collected

microdata from administrative records outside of the secure rooms located in Lisbon [8] and Brasilia, respectively. Researchers in these countries often work outside of the capital cities. Unfortunately, when the primary or sole method of accessing microdata is through anonymized data, it can negatively impact the quality of their research.

Acknowledgements. This work was partially funded by FCT- Fundação para a Ciência e a Tecnologia through project number CEMAPRE/REM - UIDB/05069/2020 and by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020.

References

1. An, P.E.: MANUAL DO ENADE Exame Nacional de Desempenho dos Estudantes. Dados (2004)
2. Fernandes, A. de O., Gomes, S. dos S.: Exame Nacional de Desempenho de Estudantes (Enade): Tendências da produção científica brasileira (2004–2018). Educ. Policy Anal. Arch. 30 (2022). https://doi.org/10.14507/epaa.30.6547
3. Bertolin, J.C.G., Marcon, T.: O (des)entendimento de qualidade na educação superior brasileira – Das quimeras do provão e do ENADE à realidade do capital cultural dos estudantes. Avaliação. **20**, 105–122 (2015). 10.590/S1414-40772015000100008
4. Dalenius, T.: Towards a methodology for statistical disclosure control. Stat. Tidskr. Stat. Rev. 15, 429–444 (1977)
5. Dalenius, T.: Finding a needle in a haystack. J. Off. Stat. **2**, 329–336 (1986)
6. Hand, D.J.: Statistical challenges of administrative and transaction data. J. R. Stat. Soc. Ser. A Stat. Soc. **181**, 555–605 (2018). https://doi.org/10.1111/rssa.12315
7. Santos, W., Sousa, G., Prata, P., Ferrao, M.E.: Data anonymization: K-anonymity sensitivity analysis. In: 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. IEEE, Sevilla (2020)
8. Ferrão, M.E., Prata, P., Fazendeiro, P.: Utility-driven assessment of anonymized data via clustering. Sci. Data. **9**, 1–11 (2022). https://doi.org/10.1038/s41597-022-01561-6
9. INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: ANRESC (Prova Brasil). https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados
10. Cox, L.H.: Suppression methodology and statistical disclosure control. J. Am. Stat. Assoc. **75**, 377–385 (1980). https://doi.org/10.1080/01621459.1980.10477481
11. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
12. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
13. Beimel, A., Nissim, K., Stemmer, U.: Private learning and sanitization: pure vs. approximate differential privacy. Theory Comput. **12**, 1–61 (2016). https://doi.org/10.4086/toc.2016.v012a001
14. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006). https://doi.org/10.1007/11761679_29

15. Kasiviswanathan, S.P., Smith, A.: On the "semantics" of differential privacy: a bayesian formulation. J. Priv. Confidentiality. **6** (2014). https://doi.org/10.29012/jpc.v6i1.634
16. Bild, R., Kuhn, K.A., Prasser, F.: SafePub: a truthful data anonymization algorithm with strong privacy guarantees. Proc. Priv. Enhancing Technol. **2018**, 67–87 (2018). https://doi.org/10.1515/popets-2018-0004
17. Avraam, D., Boyd, A., Goldstein, H., Burton, P.: A software package for the application of probabilistic anonymisation to sensitive individual-level data: a proof of principle with an example from the ALSPAC birth cohort study. Longit. Life Course Stud. **9**, 433–446 (2018). https://doi.org/10.14301/llcs.v9i4.478
18. Goldstein, H., Shlomo, N.: A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. J. Off. Stat. **36**, 89–115 (2020). https://doi.org/10.2478/jos-2020-0005
19. Jagannathan, G., Pillaipakkamnatt, K., Wright, R.N.: A practical differentially private random decision tree classifier. ICDM Work. In: 2009 - IEEE International Conference on Data Mining, pp. 114–121 (2009). https://doi.org/10.1109/ICDMW.2009.93
20. Jain, P., Gyanchandani, M., Khare, N.: Differential privacy: its technological prescriptive using big data. J. Big Data **5**(1), 1–24 (2018). https://doi.org/10.1186/s40537-018-0124-9
21. Li, N., Qardaji, W., Su, D.: On sampling, anonymization, and differential privacy or, k - anonymization meets differential privacy. In: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security - ASIACCS 2012, p. 32. ACM Press, New York (2012)
22. Sweeney, L.: A model for protecting privacy. Ieee S&P '02. 10, 1–14 (2002)
23. Prasser, F., Eicher, J., Spengler, H., Bild, R., Kuhn, K.A.: Flexible data anonymization using ARX—Current status and challenges ahead. Softw. Pract. Exp. **50**, 1277–1304 (2020). https://doi.org/10.1002/spe.2812
24. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional K-anonymity. In: 22nd International Conference on Data Engineering (ICDE 2006), p. 25. IEEE (2006)
25. Gionis, A., Tassa, T.: k-anonymization with minimal loss of information. In: Arge, L., Hoffmann, M., Welzl, E. (eds.) ESA 2007. LNCS, vol. 4698, pp. 439–450. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75520-3_40
26. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st International Conference on Data Engineering (ICDE 2005), pp. 217–228. IEEE (2005)
27. Scheffé, H.: The Analysis of Variance. Wiley, Hoboken (1999)
28. Yu, S.: Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access. **4**, 2751–2763 (2016). https://doi.org/10.1109/ACCESS.2016.2577036
29. El Emam, K.: Guide to the De-Identification of Personal Health Information. Auerbach Publications, Boca Raton (2013)
30. Kniola, L.: Plausible adversaries in re-identification risk assessment. In: PhUSE Annual Conference (2017)
31. Prasser, F., Bild, R., Kuhn, K.A.: A generic method for assessing the quality of De-identified health data. Stud. Health Technol. Inform. **228**, 312–316 (2016). https://doi.org/10.3233/978-1-61499-678-1-312
32. Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D., Martinez, S.: t-closeness through microaggregation: Strict privacy with enhanced utility preservation. IEEE Trans. Knowl. Data Eng. **27**, 3098–3110 (2015)