Συνέδρια της Ελληνικής Επιστημονικής Ένωσης Τεχνολογιών Πληροφορίας & Επικοινωνιών στην Εκπαίδευση

Τόμ. 1 (2022)

7ο Πανελλήνιο Συνέδριο «Ένταξη και Χρήση των ΤΠΕ στην Εκπαιδευτική Διαδικασία»

Implementing Statistical Disclosure Control in Educational Data

Katerina Stathatou, Rozita Tsoni , Christos Panagiotakopoulos , Vassilios Verykios

Βιβλιογραφική αναφορά:

Stathatou, K., Tsoni , R., Panagiotakopoulos , C., & Verykios, V. (2023). Implementing Statistical Disclosure Control in Educational Data. *Συνέδρια της Ελληνικής Επιστημονικής Ένωσης Τεχνολογιών Πληροφορίας & Επικοινωνιών στην Εκπαίδευση*, *1*, 0079–0092. ανακτήθηκε από https://eproceedings.epublishing.ekt.gr/index.php/cetpe/article/view/5728

Implementing Statistical Disclosure Control in Educational Data

Stathatou Katerina[1], Tsoni Rozita[1], Panagiotakopoulos Christos[2] and Verykios Vassilios[1]

katerina_smath@yahoo.gr, rozita.tsoni@ac.eap.gr, cpanag@upatras.gr, verykios@eap.gr

[1] School of Science and Technology, Hellenic Open University

[2] School of Humanities and Social Sciences, University of Patras

Abstract

Before releasing statistical outputs, data suppliers have to assess whether or not the privacy of statistical units is endangered. In many countries, privacy laws require that agencies or data producers protect confidentiality. Statistical Disclosure Control (SDC) is thus an emerging field of research. This article gives a detailed view of basic SDC methods for data and administrative sources. It discusses the traditional approach of data anonymization by perturbation of data, the disclosure risk, and the data utility of anonymized data sets

Key Words: Disclosure risk, microaggregation, noise addition, local suppression, PRAM, general utility.

Introduction

Data anonymization refers to the process of de-identifying personal information from text which is a type of information sanitization to protect privacy. Businesses, governments, academic institutions, and citizens generate prodigious amounts of data every day and the race to collect and control them is intensifying. Particularly, as more of the services we rely on go online, and with the rise of personalization in marketing and product recommendations, start-ups seek to make their fortunes by exploiting and analyzing data. Governments around the world are harnessing data for effective policy-making while academic institutions have provisions for releasing microdata for research purposes.

Recent years have marked a shift in the way we think about personal information online. In the era of big data, personal privacy is a topic of increasing concern and high-profile data misuse has made us wary of to whom we choose to divulge our information. The 2018 Facebook-Cambridge Analytica scandal saw millions of Facebook users' personal data harvested by the consultancy to be used as fuel for political advertising. Users' confidence in Facebook's handling of privacy plunged and a growing number of users that deleted their accounts is reported. Also, many insist that companies delete their information from shadowy databases.

The fact is, trendy advertising firms are not alone in housing our personal data. Academic institutions, banks, medical services, and insurance companies all hold sensitive information about us for different reasons. So statistical organizations collect an increasing amount of data on persons and establishments and the demand for researchers to make statements about our society on an empirical basis is often only possible when investigating data with detailed information.

The problem starts when the data is sensitive, confidential, or simply private since this comes with a variety of legal, ethical, and technical challenges. In 2018, the European Commission matched the public mood and growing data privacy concerns by implementing

Panagiotakopoulos, C., Karatrantou, A., & Armakolas, S. (Eds.), *Proceedings of 7th Panhellenic Scientific Conference "Integration and Use of ICT in the Educational Process"*, pp. 79-92, Dept. of Educational Sciences and Social Work, University of Patras, Greece, 16-18 September 2022. ISBN 978-618-83186-7-0.

the Data Protection Privacy Regulation (GDPR), to avoid unconsented sharing of personal information. Companies and statistical producers are faced with the challenge of ensuring respondents' confidentiality when making microdata files accessible. In almost all countries, there are privacy laws that require the identity protection of respondents from surveys and censuses.

The goal is to choose an optimal method that manages disclosure risk while ensuring high-quality statistical data. This tension between complying with confidentiality requirements while at the same time requiring that microdata be released means that Statistical Disclosure Control (SDC) methods have to be applied, also known as microdata anonymization (Benschop et al., 2019). In this article, we discuss both the basic SDC methods on continuous and categorical variables and the effect of applying these methods on data utility.

Nowadays, academic institutions have provisions for releasing microdata for research purposes usually under special license and privacy laws. All these institutions must assess the disclosure risk in respect of microdata and if required choose appropriate SDC methods to apply to the data. But they do not share their knowledge and experience using SDC and the processes for creating safe data. To fill this gap, we evaluate SDC methods on microdata from a specific academic institution where they were previously treated to be safe to release. In any released microdata set, directly identifying key variables such as name or address are removed. Additional case studies are available in Templ et al. (2014).

The aim of SDC is to prevent sensitive information about individual respondents from being disclosed. The focus was on measuring the effects that various SDC methods would have on the risk-utility trade-off for microdata produced to measure common development indicators.

This paper is structured as follows: in the next Section the related work is presented followed by Section III where the methodology and the data for the implementation are presented. In Section IV techniques for implementation in continuous variables are discussed. The next Section involves techniques for categorical variables. The final Section contains a discussion of the implementation and the main conclusions are drawn.

Related work

As the field of privacy protection involves all data mining processes it is in the spotlight of researchers with several studies concerning methods and approaches that would allow de-identification and preserve civilian rights. In a highly influential paper (Verykios et al., 2014) classification of privacy-preserving methods is provided. According to this classification, five dimensions summarize the numerous approaches:

1. data distribution
2. data modification
3. data mining algorithm
4. data or rule hiding
5. privacy preservation

Within this context to enable data sharing organization will most likely try to hide some sensitive patterns before sharing its data with others. The algorithm Local Distortion Hiding (LDH) has been evaluated on the assumption of an opponent using the J48 (C4.5) classification algorithm. In the extension that is presented in (Feretzakis et al., 2020), the CART algorithm was used in a medical dataset hiding case study of a processed by LDH. Also, the article

(Shlomo & De Waal, 2008). demonstrates how placing controls in the perturbation processes preserve the logical consistency of the records by minimizing micro edit failures and focuses on minimizing information loss to preserve data utility.

A challenging task is to preserve privacy in record linkage. Since data is often distributed in different sources, linking data is used as a preprocessing step in many data mining and analytics projects to clean, enrich, and understand data for quality results. However, the linkage usually relies on quasi-identifiers that not only allow uniquely identifying individuals but also reveal private and sensitive information about them (Vatsalan et al., 2019). Karapiperis et al. (2017) proposed a record linkage framework that implements methods for anonymizing both string and numerical data values, which are typically present in data records. The framework relies on a strong theoretical foundation for rigorously specifying the dimensionality of the anonymization space, into which the original values are embedded, to provide accuracy and privacy guarantees under various models of privacy attacks. Additionally, an implementation of a framework for privacy-preserving large-scale linkage of electronic health records (Karapiperis et al., 2018) offers a robust and distributed solution in a very sensitive domain.

In the field of educational data mining, Tsoni et al. (2021) created a data pipeline to preserve privacy in an educational setting. Accordingly, the pipeline assesses the re-identification risk by comparing the original dataset with the anonymized data. The constant demand for public use of educational data in order to improve the teaching and learning process has led researchers to develop techniques for preserving the privacy of sensitive patterns when inducing decision trees and demonstrates the application of a heuristic to an educational data set (Feretzakis et al., 2021). The educational datasets are usually of large volume. Therefore, Krasadakis et al. (2020) extended their previous work by proposing an approach to improve large computational aspects of the hiding methodology, and in particular to accommodate bigger datasets by customizing the hiding scheme and allowing it to run in parallel while ensuring the hiding of the sensitive knowledge in its entirety. Several works proposed different techniques to ensure that there is compliance with rules and regulations of privacy in learning analytics in higher education (Kyritsi et al., 2018; 2019, Jones, 2019; Chicaiza, 2020). While an important number of research papers focuses on methods and techniques, the ethical aspect of data use and the related privacy-preserving policies have also triggered researchers' interest (Pardo & Siemens, 2014; Prinsloo & Slade, 2017; Kitto & Knight, 2019; Jones, 2019; Slade, & Tait, 2019) as the data from online teaching and learning activity are rapidly accumulated in databases. Thus, there is a need to balance students' privacy and the "tremendous potential" of open data (Daries et al., 2014). Slade and Prinsloo (2013) several years before the General Data Protection Regulation was established in 2016, had proposed six important principles for LA:

1. Learning analytics as moral practice
2. Students as agents (not only produces but most of all recipients)
3. Student identity and performance are temporal dynamic constructs
4. Student success is a complex and multidimensional phenomenon
5. Transparency (regarding the purposes and conditions)
6. Higher education cannot afford to not use data

Educational institutions are legally bound by data-protection laws to respect the "right to privacy", while the GDPR also introduced the "right to access" and the "right to be forgotten" (Voigt, & Von dem Bussche, 2017). Hoel and Chen (2016) highlighted the principles of openness, transparency, and the continuous negotiation between data subjects and data controllers as the most important implication of the GDPR in LA. Therefore, the incorporation of SDC methods and the assessment of the alteration of the produced dataset after the anonymization can improve educational research by providing wider access to data. In the next Section the methodology and the dataset that we used to implement it is described.

Methodology and sample data

Testing methods with real-life data

The methods discussed in this article originate from a large body of literature on SDC. Thus, the implementation is split into two main parts: Section IV that describes anonymization methods for continuous variables, risk and utility measurement with some elaborate examples, and Section V describes basic SDC methods, risk and utility measurements for categorical key variables.

For the examples in this article, we use the open-source and free package for SDC called sdcMicro as well as the statistical software R. sdcMicro is an add-on package to the statistical software R. The package was developed and is maintained by Matthias Templ et al. (2015).

The description of the data

A dataset of 315 records was used to evaluate the SDC methods. Each record contains information about a student or a tutor from two postgraduate courses of the School of Science and Technology at the Hellenic Open University.

The numerical key variables "w1" to "w6" stand for grades in the six written assignments that students had to submit during the academic year. The students of the postgraduate courses are divided into 8 classes represented by the categorical variable "Class".

The categorical variable "Type" describes whether the participant is a tutor or a student. Two different metrics capture participants' social interaction that is expressed by their participation in the discussion forum community.

The continuous variable "Views" shows the total number of forum views per participant in the academic year.

The categorical variable "Forum participant" indicates whether a student or a tutor has participated actively in the forum that is to have made at least one post in a discussion thread. Thus, it is a binary variable that takes the values "yes" and "no".

Techniques for continuous variables

Disclosure Risk

Risk measures for continuous variables are posterior measures as they are based on comparing the microdata before and after anonymization and are based on the proximity of observations between the original and perturbed data or record linkages. This approach assesses to what extent records in the perturbed data file can be correctly matched with those in the original data file.

A risk measure called *Interval disclosure* is the proportion of original values that fall into an interval, constructed around each masked value. Values that are within the interval around the initial value after anonymization are considered too close to the initial value and hence unsafe and need more perturbation.

An Interval disclosure is illustrated in Table 1. Another approach is the *outlier detection*. Continuous variables are often skewed and this means that there are a few outliers with high values relative to the other observations of the same variable. In practice, identifying the values of continuous variables that are larger than a predetermined p%-percentile might help identify outliers and thus units at greater risk of identification.

Table 1: Disclosure risk and information loss before applying any anonymization method

```
Listing 4.1
Numerical key variables: w1, w2, w3, w4, w5, w6
Disclosure risk is currently between [0.00%; 100.00%]
Current Information Loss:
 - IL1: 0.00
 - Difference of Eigenvalues: 0.000%
```

Since no anonymization has been applied to the continuous key variables, which represent the grades of students in each assignment, the disclosure risk can be high (up to 100%).

SDC methods for continuous variables

Microaggregation

Microaggregation is a perturbing method typically applied to continuous variables. Perturbative methods perturb values to limit disclosure risk by creating uncertainty around the true values.

Microaggregation is also a natural approach to achieving k-anonymity. Firstly, a small group of individuals is formed that is homogeneous concerning the values of selected variables, such as groups with similar forum views.

Subsequently, the values of selected variables of all group members are replaced with a common value, e.g., the mean. In sdcMicro multivariate microaggregation is implemented in the function microaggregation(). After applying microaggregation, in Table 2, we can observe that the disclosure risk decreased considerably beside the disclosure risk in the previous Table 1.

Table 2: Disclosure risk and information loss after applying microaggregation to continuous key variables

```
Listing 4.2
Numerical key variables: w1, w2, w3, w4, w5, w6
Disclosure risk is currently between [0.00%; 26.03%]
Current Information Loss:
 - IL1: 12213.19
 - Difference of Eigenvalues: 66.850%
```

Noise Addition

Adding noise is a perturbative method typically applied to continuous variables. This means that small values are added to the original values of a variable in order to protect data from exact matching with external files. There are several noise addition algorithms such as uncorrelated additive noise and correlated additive noise.

In sdcMicro noise addition is implemented in the function addNoise().

Table 3 shows the disclosure risk and information loss after correlated noise addition on the variables that represent students' grades.

Table 3: Disclosure risk and information loss after applying additive noise to continuous key variables

Listing 4.3 Numerical key variables: w1, w2, w3, w4, w5, w6 Disclosure risk is currently between [0.00%; 5.08%] Current Information Loss: - IL1: 5244.45 - Difference of Eigenvalues: 13.750%

We can see that addNoise is a more suitable method than microaggregation for these objects since the latter method includes lower risk and the data utility measure is comparable. For the variables w1, w2, w3, w4, w5, w6 the disclosure risk is about 26% after microaggregation while only 5% after noise addition.

Shuffling

Shuffling generates new values for selected sensitive variables based on the conditional density of sensitive variables given non-sensitive variables. The idea is to rank the individuals based on their original variables. Then fit a regression model with the variables to be protected as regressands and a set of variables that predict this variable well as regressors. As a rough illustration, assume we have two sensitive variables, income and savings, which contain confidential information. We first use education, age, gender, occupation variables as predictors in a regression model to simulate a new set of values for income and savings. We then apply shuffling to replace ranked new values with the ranked original values for income and savings. This regression model is used to generate n synthetic (predicted) values for each variable that has to be protected. These generated values are also ranked and each original value is replaced with another original value with the rank that corresponds to the rank of the generated value. This means that all original values will be in the data.

General utility measures for continuous variables

IL1 information loss measure

IL1 is a distance measure between the original dataset and the treated dataset for continuous variables. The measure is useful for comparing different methods of anonymization. The smaller the value of IL1, the closer the values are to the original values and the higher the utility.

As we can see in Table 2 and Table 3, the data utility is lower for microaggregation than for adding correlated noise.

Eigen

Another way to evaluate the information loss is to compare the robust eigenvalues of data. The output is the differences in eigenvalues before and after anonymization. Eigenvalues can be estimated from a robust or classical version of the covariance matrix. However, covariance-based measures are only suitable in the multivariate context without any missing values and zeros in the data. The greater the value, the larger the changes in the data and the information loss.

Since the smaller the value of the measure, the closer the values are to the original values

and the lower the information loss, we see that the data utility is lower for microaggregation than for adding correlated noise.

Assessing data utility with the help of data visualizations

Visualizations can be a useful tool to assess the impact on the data utility of anonymization methods and help to choose the appropriate anonymization technique for the data. We present the following visualizations:

- histograms plots
- boxplots

Histogram plots are useful for quick comparisons of variable distribution before and after anonymization. Histograms can be used for continuous variables and the advantage is that the results we can take are exact. In Fig. 1 and Fig. 2 there are examples to illustrate the changes in the variable "Views".

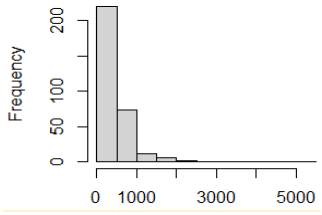

Figure 1: "Views" distribution (original data)

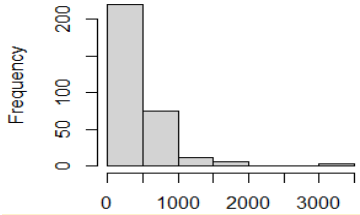

Figure 2: "Views" distribution (data after anonymization)

Box plots also give a quick overview of the changes in continuous variables before and after anonymization. The result in Fig. 3 shows an example for the variable "Views" after applying microaggregation. We can see clearly that the variability in the views of students decreased as a result of the anonymization method applied.

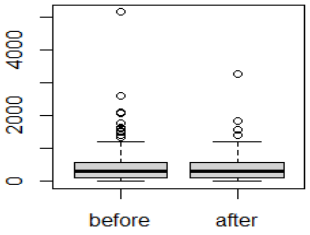

Figure 3: "Views" before and after microaggregation

Techniques for categorical variables

Disclosure risk for categorical key variables

Frequency counts

Disclosure risk is defined based on assumptions of disclosure scenarios, that is, how the intruder might exploit the released data to reveal information about a respondent. In general, the rarer a combination of values of the quasi-identifiers (key values) of observation in the sample, the higher the risk of identity disclosure. If the sample frequency equals 1, this individual has a unique combination of values of quasi-identifiers hence a high risk of re-

identification. So computing frequency counts serve as a basis for many disclosure risk estimation methods.

k-Anonymity and l-diversity

Assuming that sample uniques are more likely to be re-identified, one way to protect confidentiality is to ensure that each distinct pattern of key variables is possessed by at least k records in the sample. An individual violates k-anonymity if the sample frequency count for this key is smaller than the specified threshold k. For example, if an individual has the same combination of quasi-identifiers as three other individuals in the sample, these individuals satisfy 4-anonymity but violate 5-anonymity.

In some cases while k-anonymity is satisfied, sensitive information might still be disclosed. This might occur in cases where the data contains sensitive categorical variables that have the same value for all individuals that share the same key. The concept of l-diversity addresses this limitation of k-anonymity. l-diversity ensures that the sensitive variable has at least l distinct values for each group of observations with the same pattern of key variables.

Special Uniques Detection Algorithm (SUDA)

An alternative measure to determine disclosure risk is based on the concept of special uniqueness. An observation is defined as a special unique with respect to a variable set Q if it is a sample unique both on Q and on a subset of Q (Elliot eta al., 1998). To find special uniques, algorithms, called SUDA, have been developed. SUDA algorithms are based on the concept of special uniqueness, or on Minimal Sample Uniques (MSUs), which are unique variable sets without any unique subsets within a sample. SUDA identifies all the MSUs in the sample, which in turn are used to assign a SUDA score to each observation. This score indicates the risk using the size and distribution of MSUs within each record. The potential risk of the record is determined based on two issues:

- within an observation, the risk of the observation is higher as the number of variables needed to reach uniqueness (i.e., the smaller the size of MSU) gets smaller
- the risk of the observation is higher as the number of MSUs in an observation gets larger.

In order to estimate the observation-level disclosure risks, SUDA scores can be used in combination with the Data Intrusion Simulation (DIS) metric, which is a method for assessing a global disclosure risk for the entire data set. To receive the DIS score an iterative algorithm based on sampling of the data and matching of subsets of the sampled data with the original data is applied. This algorithm calculates the probabilities of correct matches given unique matches. SUDA and DIS-SUDA measures can be calculated in sdcMicro. It is important, after applying SDC methods, that one would recalculate the SUDA scores and compare them to the original values. Also, it may be useful to use histogram plots of these scores.

Table 4: Evaluating SUDA scores

Listing 5.1
Dis suda scores table:

	Interval	Number of records
1	== 0	309
2	(0.0, 0.1]	6
3	(0.1, 0.2]	0
4	(0.2, 0.3]	0
5	(0.3, 0.4]	0
6	(0.4, 0.5]	0
7	(0.5, 0.6]	0
8	(0.6, 0.7]	0
9	> 0.7	0

Attribute contribution:

	variable	contribution
1	Class	100.00000
2	Type	71.42857
3	Forum participant	28.57143

In Table 4, we can see that six observations have considerable high risk.

SDC methods for categorical variables

Recoding

Global recoding is a non-perturbative method that can be applied to both categorical and continuous key variables. It is a deterministic method used to decrease the number of distinct categories or values for a variable. For categorical variables, the idea of recoding is to combine several categories into fewer categories with higher frequency counts and less detailed information. This means that a global recoding achieves anonymity by mapping the values of the categorical key variables to generalized or altered categories. For continuous variables, global recoding constructs intervals, and the variable is changed into a categorical one. In both cases, the goal is to reduce the total number of possible values of a variable. For example, one could combine multiple levels of schooling (e.g., secondary, tertiary, postgraduate) into one (e.g., secondary and above) or a continuous income variable into a categorical variable of income levels.

Local suppression

If unique combinations of categorical key variables remain after recoding, local suppression could be applied to the data to achieve k-anonymity. Suppression of values means that values of a variable are replaced by a missing value (NA), thus reducing the record-level disclosure risks. The most common function included in sdcMicro, is localSuppression() and allows the use of suppression on specified quasi-identifiers to achieve a certain level of k-anonymity for these quasi-identifiers. This approach sets the parameter k and tries to achieve k-anonymity with minimum suppression of values. In Table 5, local suppression is applied to achieve the k-anonymity threshold of 3 on the quasi-identifiers “w1” to “w6”.

Table 5: Application of local suppression without importance vector

Listing 5.2
Local suppression:

KeyVar	Suppressions (#)	Suppressions (%)
w1	4	1.270
w2	12	3.810
w3	9	2.857
w4	7	2.222
w5	15	4.762
w6	26	8.254

In Table 6, we can see that 3-anonymity is ensured.

Table 6: Display of observations violating k-anonymity after local suppression

Listing 5.3
Infos on 2/3-Anonymity:
Number of observations violating
- 2-anonymity: 0 (0.000%) | in original data: 55 (17.460%)
- 3-anonymity: 0 (0.000%) | in original data: 73 (23.175%)
- 5-anonymity: 1 (0.317%) | in original data: 87 (27.619%)

Furthermore, it is possible to specify the desired ordering of key variables. The aim is that the higher the importance of a variable, the fewer suppressions are taken for this variable. Without ranking the importance of variables, the value of the variable "w6" is more likely to be suppressed, since this is the variable with most categories (Listing 5.2). The value in the importance vector can range from 1 to the number of quasi-identifiers. We can assume that the variable "w6" is very important, giving the importance 1. In Table 7, it can be seen that other variables are mainly used for suppression. In variable "w6", for example, only 1 instead of 26 local suppressions are made. The importance vector should be specified only in cases where the variables with many categories play an important role in data utility for the data users.

Table 7: Application of local suppression with importance vector

Listing 5.4
Local suppression:

KeyVar	Suppressions (#)	Suppressions (%)
w1	29	9.206
w2	15	4.762
w3	20	6.349
w4	6	1.905
w5	2	0.635
w6	1	0.317

Post-randomization Method (PRAM)

PRAM (Gouweleeuw et al., 1998) is a probabilistic, perturbative method for protecting categorical variables. This method swaps the categories for selected variables, such that

intruders that attempt to re-identify individuals in the data do so, but with positive probability, the re-identification made is with the wrong individual. This means that the intruder might be able to match several individuals between external files and the released data files, but cannot be sure whether these matches are to the correct individual. The method is based on a pre-defined transition matrix, which specifies the probabilities for each category to be swapped with other categories.

To illustrate, consider the variable "Type", with two categories: Type1=Student, Type2=Tutor. We define a 2 by 2 transition matrix, where p_{ij} is the probability of changing category i to j. For example, in the following matrix,

$$P = \begin{pmatrix} 0.9800250 & 0.0199750 \\ 0.3732828 & 0.6267172 \end{pmatrix}$$

the probability that the value "student" of the variable will stay the same after the perturbation is 0.9800250 and similarly the probability of "tutor" is 0.6267172. The value "student" will be changed to "tutor" with a probability of 0.0199750.

PRAM protects the records by perturbing the original data file, while at the same time, the characteristics of the original data can be estimated from the perturbed data file, since the probability mechanism used is known. PRAM is applied to each observation independently and randomly. This means that different solutions are obtained for every run of PRAM.

General utility measures for categorical variables

Number of missing values

Missing values (NA) might be accounted for as an informative utility measure. It counts the missing values in the original data and then in the anonymization data. Missing values are often introduced after suppression and more suppressions indicate a higher degree of information loss. Counting and comparing the number of missing values in the original and treated data can be useful to see the proportional increase in the number of missing values.

Table 8: Missing values in the original data and then the anonymization data

Listing 5.5

	NAw1	NAw2	NAw3	NAw4	NAw5	NAw6
before	16	16	16	16	16	140
after	20	28	25	27	35	172

The results agree with the number of missing values introduced by local suppression in Listing 5.2 since the variable w1 has 16 missing values in the original data and 4 suppressions, therefore it has 20 missing values after anonymization.

Comparing contingency tables

Another useful way to measure information loss in categorical variables is to compare contingency tables between pairs of variables. These tables should stay approximately the same, before and after anonymization in order to maintain the analytical validity of a dataset. Contingency tables can also be visualized using mosaic plots in order to compare the impact of anonymization methods.

Assessing data utility with the help of data visualizations

In this section, we use mosaic plots to assess at a glance how much the data has changed after anonymization. Mosaic plot is a useful visualization for showing changes in the tabulation of categorical variables. With mosaic plots, we can, for instance, quickly see the effect of different levels of k-anonymity that is achieved by choosing different parameters or differences in the importance vector in the local suppression algorithm.

We illustrate the changes for each category in the tabulation of the variable "*Class*" before and after applying PRAM. Looking at the mosaic plot in Fig. 4, we see the original sample frequencies and the sample frequencies from the perturbed data. It can be seen that PRAM has a slight influence on the distribution.

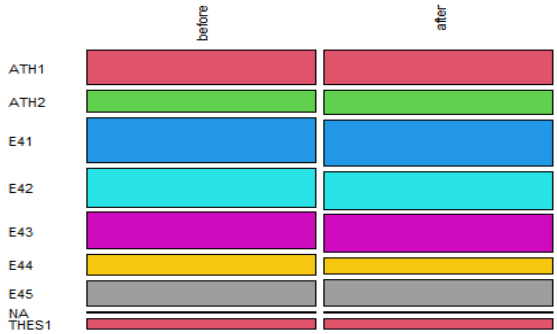

Figure 4: Sample frequencies for the "class" variable before and after anonymization

Discussion and conclusions

In this article, we have demonstrated SDC methods that are simple to implement and are commonly carried out at statistical agencies for sample microdata. More complex multivariate methods exist but the techniques that we presented can be applied for preserving edits and sufficient statistics. The important observation is that for any method, simple or complex, it is possible to increase the quality and utility of the perturbed microdata based on the proposed approaches. By combining SCD methods and designing innovative techniques for implementation, we can obtain consistent data, and release statistical outputs with higher degrees of utility at little cost to the risk of disclosure.

It is hard to determine the "best" SDC method to protect a dataset in general, since what is the "best" method depends on the intended uses of the data on the part of the users, the willingness of the statistical agency to disseminate this data set and the legal aspects of releasing these data and the structure of the data.

As regards the utility measures, the choice should be made in accordance with the variable types and anonymization method employed. The employed utility measures can be a combination of both general and user-specific measures. Additionally, it is important not only to focus on the characteristics of variables one by one but also on the interactions between variables. Hopefully this article would help to improve the understanding of several such SDC methods in an educational setting.

References

Benschop, T., Machingauta, C., & Welch, M. (2019). *Statistical disclosure control: A practice guide.*

Chicaiza, J., Cabrera-Loayza, M. C., Elizalde, R., & Piedra, N. (2020). Application of data anonymization in Learning Analytics. In Proceedings of the 3rd International Conference on Applications of

Intelligent Systems (pp. 1-6).
Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., ... & Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9), 56-63.
Elliot, M., Skinner, C., & Dale, A. (1998). Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, 2.
Feretzakis, G., Kalles, D., & Verykios, V. S. (2020). Local Distortion Hiding Algorithm in Medical Data: A Case Study Using CART. In *The Importance of Health Informatics in Public Health during a Pandemic* (pp. 99-102). IOS Press.
Feretzakis, G., Kalles, D., & Verykios, V. S. (2021). Knowledge Hiding in Decision Trees for Learning Analytics Applications. In *Advances in Core Computer Science-Based Technologies* (pp. 37-54). Springer, Cham.
Gouweleeuw, J. M., Kooiman, P., & De Wolf, P. P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of official Statistics*, 14(4), 463.
Hoel, T., & Chen, W. (2016). *Implications of the European data protection regulations for learning analytics design.* In Workshop paper presented at the international workshop on learning analytics and educational data mining (LAEDM 2016) in conjunction with the international conference on collaboration technologies (CollabTech 2016), Kanazawa, Japan-September (pp. 14-16).
Jones, K. M. (2019). " Just Because You Can Doesn't Mean You Should": Practitioner Perceptions of Learning Analytics Ethics. portal: *Libraries and the Academy*, 19(3), 407-428.
Jones, K. M. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, 16(1), 1-22.
Karapiperis, D., Gkoulalas-Divanis, A., & Verykios, V. S. (2017). FEDERAL: A framework for distance-aware privacy-preserving record linkage. *IEEE Transactions on Knowledge and Data Engineering*, 30(2), 292-304.
Karapiperis, D., Gkoulalas-Divanis, A., & Verykios, V. S. (2018). *FEMRL: A framework for large-scale privacy-preserving linkage of patients' electronic health records.* In 2018 IEEE International Smart Cities Conference (ISC2) (pp. 1-8). IEEE.
Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology*, 50(6), 2855-2870.
Krasadakis, P., Verykios, V. S., & Sakkopoulos, E. (2020). *Parallel based hiding of sensitive knowledge.* In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1249-1254). IEEE.
Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2018). Privacy Issues in Learning Analytics. *Blended and Online Learning*, 218.
Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2019). The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. Journal of Interactive Media in Education.
Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438-450.
Prinsloo, P., & Slade, S. (2017). Ethics and learning analytics: Charting the (un) charted. SoLAR.
Shlomo and De Waal (2008). Protection of Micro-data Subject to Edit Constraints. *Journal of Official Statistics*, Vol. 24, No. 2, pp. 229–253
Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510-1529.
Slade, S., & Tait, A. (2019). Global guidelines: Ethics in learning analytics. *International council for open and distance education*.
Templ, M., Kowarik, A., & Meindl, B. (2014). sdcMicro case studies (Vol. 1). *Research Report CS-2014*.
Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67, 1-36.
Tsoni, R., Zorkadis, V., & S. Verykios, V. (2021). *A Data Pipeline to Preserve Privacy in Educational Settings*. In 25th Pan-Hellenic Conference on Informatics (pp. 138-142).
Vatsalan, D., Karapiperis, D., & Verykios, V. S. (2019). Privacy-Preserving Record Linkage. *Encyclopedia of Big Data Technologies*.
Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). *State-of-the-*

art in privacy preserving data mining. ACM Sigmod Record, 33(1), 50-57.
Voigt, P., & Von dem Bussche, A. (2017). *The EU general data protection regulation (gdpr). A Practical Guide*, 1st Ed., Cham: Springer International Publishing, 10(3152676), 10-5555.