



**Balancing Privacy and Utility in the Reuse of
Administrative Educational Data: Challenges for Open
Science and Equity**

Paper 2646

ECER - EUROPEAN CONFERENCE ON EDUCATIONAL RESEARCH

Balancing Privacy and Utility in the Reuse of Administrative Educational Data: Challenges for Open Science and Equity

- Paula Prata (1,2), Maria Eugénia Ferrão (1,3) and João Marôco (4,5)
- (1) Universidade da Beira Interior
- (2) Instituto de Telecomunicações
- (3) ISEG Research
- (4) Universidade Lusófona
- (5) INTREPID Lab

Student Persistence and degree Completion in Higher Education: Theoretical and Empirical Models
(StdPerson) Project FCT: 2023.18332.ICDT



Open Research

- **Open Research** is increasingly central in educational research.
- It promotes transparency, reproducibility, collaboration, and accessibility throughout the research lifecycle.
- **Core principles:**
 - Transparency;
 - Reproducibility;
 - Societal relevance.

Open Research 2

- Societal relevance ensures that open research contributes not only to scientific knowledge, but also to informed decisions and real-world educational improvement.
- Open research practices **include**:
 - Open access publications;
 - Open methodologies ;
 - Data and code sharing;
 - Preregistration (declaring the research plan before analyzing the data).

EMIS data

- **Administrative educational data** are increasingly valuable.
- Educational Management Information Systems (EMIS) provide:
 - Population-level coverage;
 - Longitudinal information;
 - High temporal granularity;
 - Low-cost secondary research opportunities.

Open Data vs Data Protection

- Educational datasets often contain:
 - Demographic information;
 - Academic records;
 - Behavioral traces;
 - Institutional data.

Open Data vs Data Protection 2

- Challenges:
 - Presence of minors and vulnerable populations;
 - High-dimensional and linkable datasets;
 - Re-identification risks.
- **Key Question**
 - How can educational data be shared openly while preserving privacy and research utility?

Data Deidentification/Anonymization

- **Data Anonymization**

- Process of removing or modifying personally identifiable information from datasets, to assure individual privacy.
- Goal: irreversible prevention of identification.

- **Role in Responsible Open Science**

- Ethical requirement;
- Legal Safeguard (European GDPR – General Data Protection Regulation);
- Methodological necessity.

Main Data Deidentification Techniques

- **Perturbative techniques** vs Nonperturbative techniques:
 - In **perturbative** techniques, the original data are modified to protect privacy but preserving as much information as possible;
 - Includes: adding noise; data swapping, micro aggregation, etc., ...
- They prevent re-identification, but changing the values may hinder subsequent associations and analyses!

Main Data Deidentification Techniques 2

- Perturbative techniques vs **Nonperturbative techniques**:
 - **Nonperturbative** techniques focus on removing or masking data points without changing their values;
 - Includes: record suppression, dropping variables, recoding, aggregation.
- They reduce the risk of reidentification, but reducing the level of information impacts the research utility of the data!

Research Objectives

This study reflects on:

- Preparing educational microdata for reuse;
- Anonymization challenges in real-world datasets.
- Balancing:
 - Privacy protection;
 - Analytical utility;
 - Inclusiveness.

Datasets Used

- 2019 Brazilian National Student Performance Exam (ENADE) microdata.
- Register of students enrolled in and graduated from higher education (RAIDES) made available by the Portuguese Directorate-General of Education and Science Statistics (DGEEC).

Methodological Framework - data set audit

Identification of:

- **Direct** identifiers
 - attributes that directly makes possible the identification of an individual;
- **Quasi –identifiers**
 - in combination with other attributes (within the dataset or from other datasets) allow for individual re-identification;
- **Sensitive** attributes
 - variables whose content belongs to the private domain of individuals.

Methodological Framework - techniques applied

- **Dropping variables**
 - remove direct identifiers.
- **Generalization**
 - combining categories of a variable to create broader groups.

After defining possible **generalization hierarchies for the quasi-identifiers**, an anonymization algorithm should be used to find the minimal generalization that satisfies the desired privacy model.

- **Suppression**
 - remove records that have potential for reidentification.

Privacy Models

- **K-anonymity Principle:**

- Each record must be indistinguishable from at least $(k-1)$ other records with respect to the quasi-identifier attributes.

- **Differential Privacy (DP)**

- Provides formal mathematical privacy guarantees.

- Key idea**

- The inclusion or exclusion of one individual should not significantly affect the output.

Privacy Models 2

In this study

- **The approximate DP or (ϵ, δ) -DP model is applied using the SabPub algorithm** (algorithm designed for publishing anonymized datasets in tabular form).
- ϵ controls the primary level of privacy (the smaller it is, the stronger the privacy guarantee).
- δ represents a small probability that the guarantee may be violated. It is typically chosen to be extremely small.

Risk-Utility Assessment

Iterative Evaluation Process

- **Disclosure Risk**
 - Equivalence class sizes; outlier detection, plausibility of linkage attacks;
- **Data utility**
 - Comparison between original and anonymized data;
- **Objective**
 - Balance:
 - Privacy protection;
 - Analytical usefulness.

Results – ENADE 2019 dataset

DS0 - 361587 records after preprocessing

Insensitive attributes:

- University id
- Program id
- Gpa1 (specific score)
- Gpa2 (general score)
- Father's education
- Mother's education*
- Family's income*
- Academic Scholarship

Quasi-identifier attributes (group 1)

- Region
- Age
- Gender
- Skin color
- Number of years to finish the graduation
- Years between high school and college

Results – Anonymized Datasets

- Using ARX anonymization tool we applied approximate differential privacy (DP) and k-anonymity to DS0 considering three groups of quasi-identifiers, originating six anonymized datasets.

Quasi-identifiers:

- group 1
- group 2 = group 1 + Family's income
- group 3 = group 2 + Mother's education

Anonymized datasets:

- => DS1-K, DS1-DP
- => DS2-K, DS2-DP
- => DS3-K, DS3-DP

Anonymization parameters:

for (ϵ, δ) -DP we used $\epsilon = 1$ and $\delta = 0.01$; for k-anonymity we used $k = 20$

Results – Remaining Records

Records remaining after anonymization and risk of reidentification as evaluated by ARX:

Dataset	Remaining records	Highest Risk	K value
DS0	100%	100%	n.a.
DS1-K	73%	5.00%	20
DS2-K	50%	5.00%	20
DS3-K	29%	5.00%	20
DS1-DP	40%	5.26%	19
DS2-DP	26%	5.26%	19
DS3-DP	13%	5.26%	19

Results – Utility Analysis (gender)

Utility analysis - Multiple Linear Regression, with gpa2 as the dependent variable.

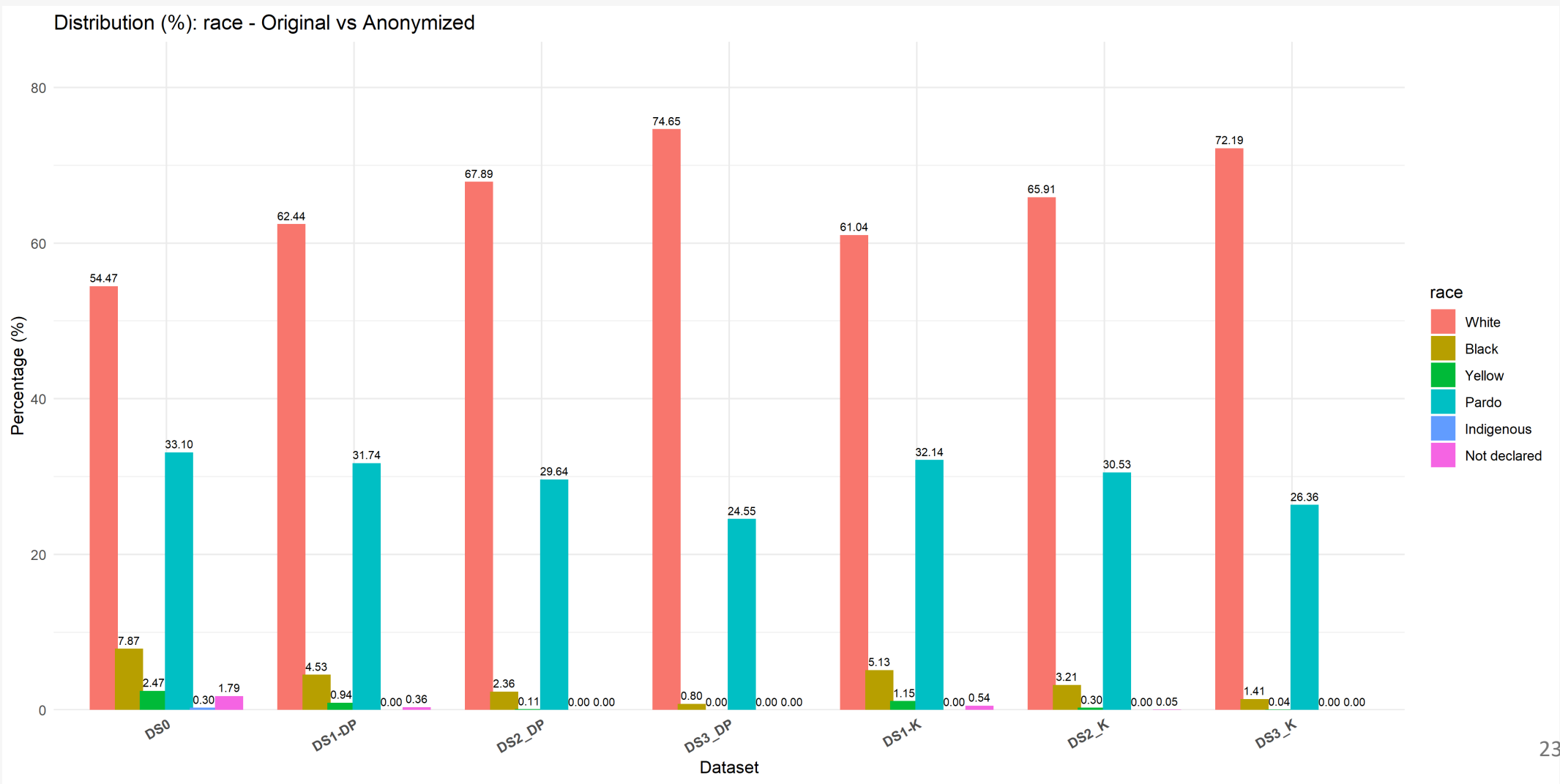
	DS0	DS1-K	DS2-K	DS3-K	DS1-DP	DS2-DP	DS3-DP
	Gender: Female vs Male						
Estimate	- 0.505	- 0.350	- 0.310	-0.125	- 0.262	- 0.371	- 0.391
Std error	(0.054)	(0.063)	(0.077)	(0.103)	(0.086)	(0.108)	(0.153)
P-value	< 0.001	< 0.001	< 0.001	< 0.223	< 0.002	< 0.001	< 0.011

- Compared with the estimate to DS0, with k-anonymity, the coefficient was reduced by approximately 31% in DS1-K and by about 75% in DS3-K. With DP the coefficient also became less negative in all cases, indicating an attenuation of the estimated effect.
- Generally, we can see that the anonymization produces a bias towards zero.

Results – Race/Skin Color Frequencies

- Observing the race frequencies in the next slide, we can see that racial minority groups in the original dataset virtually disappear following anonymization. This is particularly evident for the **Yellow**, **Indigenous**, and **Not Declared** categories.
- As a result, the corresponding coefficient estimates are either severely distorted or cannot be estimated at all.

Results – Graph with Race/Skin Color Frequencies



Results – Utility Analysis (skin color)

Utility analysis - Multiple Linear Regression, with gpa2 as dependent variable.

	DS0	DS1-K	DS2-K	DS3-K	DS1-DP	DS2-DP	DS3-DP
	Skin color: Yellow (Asian descents) vs White						
Estimate	- 1.267	- 0.310	- 2.259	1.609	- 0.380	- 2.294	N.A.
Std error	(0.162)	(0.288)	(0.668)	(2.283)	(0.431)	(1.528)	-
P-value	< 0.001	< 0.281	< 0.001	< 0.481	< 0.378	< 0.133	-

- The regression coefficient exhibited substantial instability across all the scenarios.
- For the smaller datasets (with more quasi-identifiers) it indicates an inversion in the direction of the relationship, or it cannot be calculated.

Results – RAIDES

Challenges in Accessing Portuguese Educational Data at the DGEEC Safe Center:

- Secure room may be hundreds of kilometers away.
- Access requires advance booking.
- Available computers may be outdated and slow.
- CSV files are suitable for small datasets but inefficient for large datasets and complex data relationships.
- Data may still not be released after anonymization because:
 - the criteria for classifying microdata quasi-identifiers are not publicly disclosed;
 - different anonymization tools use different methodologies to estimate re-identification risk, making results difficult to compare across tools.

Findings

Anonymization techniques may:

- Distort minority groups;
- Suppress rare categories;
- Introduce hidden bias.

Affected dimensions

- Socio-economic groups;
- Race/ethnicity;
- Institutional characteristics.

Consequence

Potentially misleading evidence for equity-oriented policies.

Conclusion

Anonymization is essential for Open Science in education

It is:

- Iterative;
- Context-dependent;
- Methodologically complex.

However

Even robust anonymization methods may:

- Reduce analytical validity;
- Bias representation;
- Limit educational inference.

Future Need

- Advanced privacy-preserving infrastructures;
- Better governance models;
- Balanced access policies for public-interest research.

Acknowledgments

Authors would like to acknowledge the valuable contribution of research fellows **Pedro Ferrão** and **Kelly Macedo** for their assistance in the collection of the results used in this study.

Thank You

Questions & Discussion