EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Assessment of disclosure risk on financial bases for individuals

Valentina Wolff Lirio[1], Rita de Sousa[2], Susana Faria[1]

[1]University of Minho, Portugal

[2]Bank of Portugal

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union

Introduction

- The **Statistical Disclosure Control (SDC)** techniques - set of tools that can improve the level of confidentiality of any dataset, which allows institutions to publish their data in a safe and efficient way for the user.
- The **identification risk** is the probability of an intruder identifying at least one respondent in the available microdata bases.
- The **General Data Protection Regulation (GDPR)** has the main objective of adapting data privacy laws in Europe by controlling the processing by individuals, companies or organizations of personal data.

Objective

- The main objective of this study is to explore **individual and global identification risk** assessment methodologies in **individual financial databases**, with **an application** to the microdata base of the Central Credit Register (**CCR**).

Variables Classification

- **Direct identifiers:** variables that **provide direct information** about the individuals; **examples:** name, tax identification number or address.
- **Indirect identifiers**: also known as **key variables** or **quasi-identifiers**, they do not provide direct identification information but, when combined with each other, enable the identification of individuals; **examples:** combination of age, sex and residence.
- **Non-identifiers:** variables that **do not provide direct and indirect information** to identify individuals; **examples:** socioeconomic, demographic or behavioral characteristics.

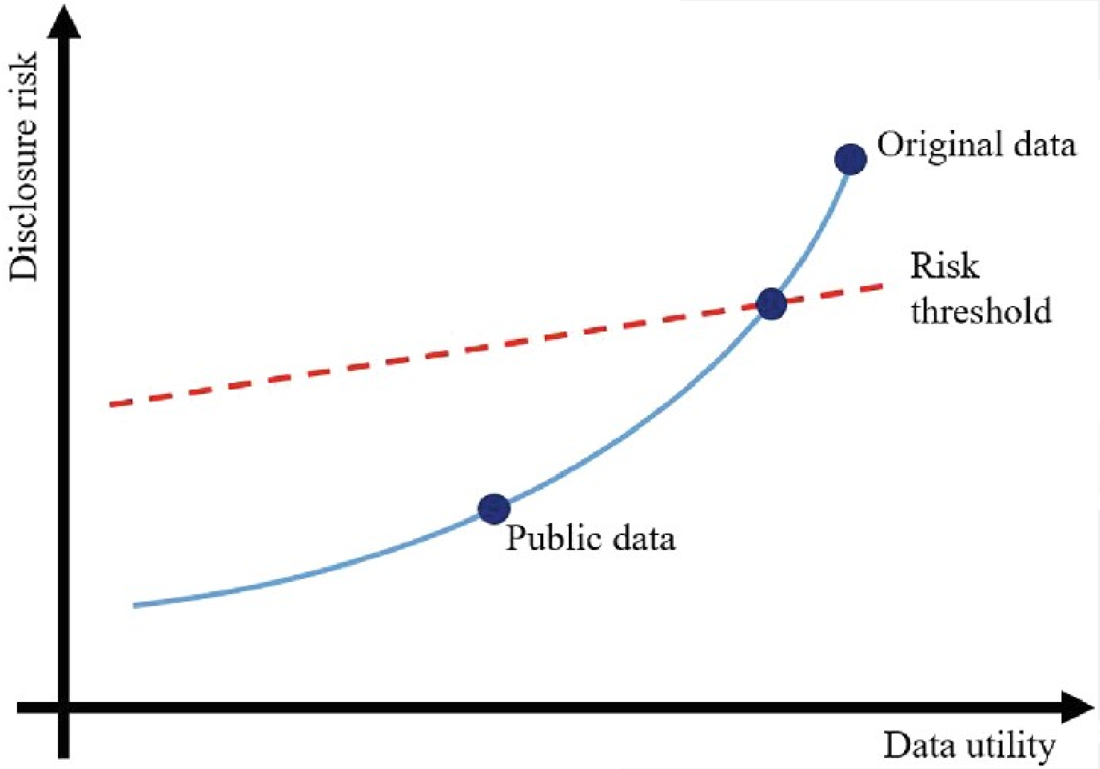
Variables Classification

- **Sensitive variables: may reveal sensitive personal information** of respondents. They normally depend on ethical and legalization issues to be linked. For example, data relating to health, religion, sexual orientation, socioeconomic status, income, criminal information, among others.

- **Non-sensitive variables: do not have confidential information** about individuals, but this does not mean that these variables are not relevant for research purposes and for the application of SDC methods.

Identification Risk
vs
Information Loss

Figure 1: Disclosure Risk *vs* Data Utility



Anonymization

According to the **ISO 29100:2011** standard, **anonymization** is a process in which **Personally Identifiable Information** (PII) is irreversibly modified, meaning that an entity cannot be identified either directly or indirectly (ISO, 2011).

Other **terminologies** become relevant in this study, as they are **associated with data anonymization**:

- **De-identification:** aims to **remove or hide all personal information** from a dataset to make it impossible to identify individuals.
- **Pseudonymization:** is a technique that aims to change all personal identifiers (for example, name, address and identification number) to pseudonyms: words or codes obtained artificially, which can act as masked representations of the original data.

Statistical Disclosure Control Methods (SDC)

- **Perturbative methods:** adding noise or modifying the data in order to maintain the utility of the data and reduce the risk of identification.
- **Non-perturbative methods:** aim to protect privacy without directly introducing noise into the data.
- **Synthetic data:** creation of data sets that are artificially generated to resemble real data while maintaining relevant statistical and structural characteristics.

Identification Risk Measures - Categorical Variables

- ***K*-anonymity:** the risk measure is based on the principle that the number of individuals in a sample/population sharing the same combination k of key variables should be higher than a specified threshold K.

- ***L*-diversity:** aims to ensure that each group of observations that share the same combination of key variables contains at least L distinct values for the sensitive variables.

Identification Risk Measures - Numerical Variables

- **Record Linkage:** evaluates the correct number of links between published values and original values. Let y_{ip} be the modified observation of the original x_{ip}. Consider x_{1p} and x_{2p} to be the closest observations to y_{ip} and calculate a distance between them. If either of them matches the original observation x_{ip}, then x_{ip} and y_{ip} are said to be linked.

- **Interval Measure:** created around each published value and it is checked whether the original value belongs to the established interval.

- **Outliers Count:** it is carried out by identifying values that are higher or lower than a certain percentile.

Individual and Global Identification Risk

- **Individual risk** - probability of identifying an individual observation: $r_i = 1/F_k$, where F_k is the population frequency of the combination k of key variables, to which observation i belongs.

- **Global risk** - proportion of observations that can be identified by a user. Often calculated by the arithmetic average of all individual risks:

$$R = (1/N)\sum_{i=1}^{N} r_i$$

- As an alternative to the individual identification risk, there is the **Special Uniques Detection Algorithm (SUDA)**, which allows identifying observations with the highest risk.

Case Study

- The database under study belongs to the **Central Credit Register** (CCR) of Banco de Portugal (BdP).
- The focus in this study is on the bases of individuals, mainly on the set of key variables that can allow their identification.
- The database under study contains **6342255 observations** relating to the credit records of **Portuguese individuals in December 2022**.

Case Study
Data Summary

```
    dtRef                  idEnt                 genero                 agregFam                                        sitProf
 Length:6342255     Min.   :      1               :1035114   1 pessoa  :2290646                                    :1368686
 Class :character   1st Qu.:1585564   Feminino :2831122   2 pessoas :1315452   Desempregado                     : 174765
 Mode  :character   Median :3171128   Masculino:2476019   3 pessoas : 634514   Empregado por conta de outrem:3129311
                    Mean   :3171128                        4 pessoas : 421425   Empregado por conta propria  : 292572
                    3rd Qu.:4756692                        5 pessoas : 106553   Estudante                    : 163506
                    Max.   :6342255                        6 pessoas :  16157   Fora do mercado de trabalho  : 275786
                                                           7+ pessoas:1557508   Reformado                    : 937629
                  concelho           nuts3           escEtario                habil
 1106 Lisboa           : 347490   170     :1799214   60+     :2171323                    :1322930
 1111 Sintra           : 242016   11A     :1089212   <=19    :  32386   Basico           :1334578
 1317 Vila Nova de Gaia: 195923           : 336549   [20-29]: 536332   Secundario       :2200297
 1312 Porto            : 147213   16E     : 270621   [30-39]: 942977   Sem escolaridade:  22718
 1105 Cascais          : 136722   150     : 263271   [40-49]:1371633   Superior         :1461732
 1107 Loures           : 121535   119     : 247929   [50-59]:1286586
 (Other)               :5151356   (Other):2335459   NA's   :   1018
```

Case Study Variables

Table 1: Study key variables

Variable	Type	Description
genero	Categorical	Individual's gender
escEtario	Categorical	Age group to which the individual belongs
sitProf	Categorical	The individual's professional status
agregFam	Categorical	Number of people in the household the individual belongs to
habLit	Categorical	Level of the individual's educational qualifications
concelho	Categorical	Individual's municipality of residence

Case Study

Figure 2: Initial K-anonymity results

```
Number of observations violating
  - 2-anonymity: 99265 (1.565%)
  - 3-anonymity: 189492 (2.988%)
  - 5-anonymity: 348482 (5.495%)
```

High number of observations that **do not guarantee a minimum of 2 or 3 observations** for each combination of key variables.

The municipality of residence variable is very disaggregated, with more than 300 categories, so we will consider the variable **nuts3**, which contains level 3 of the Nomenclature of Territorial Units for Statistics (NUTS III).

Figure 3: K-anonymity results when using the variable nuts3

```
Number of observations violating
  - 2-anonymity: 8394 (0.132%)
  - 3-anonymity: 17102 (0.270%)
  - 5-anonymity: 34472 (0.544%)
```

SDC Methods
Categorical Variables

Recoding

Reducing the number of categories of the number of people in the household **from 7 to 5**.

```
> dataset$agregFam2 <- ifelse(dataset$agregFam=="5 pessoas","5+ pessoas",
+                          ifelse(dataset$agregFam=="6 pessoas","5+ pessoas",
+                            ifelse(dataset$agregFam=="7+ pessoas","5+ pessoas",
+                               dataset$agregFam)))
```

```
Number of observations violating
  - 2-anonymity: 5713 (0.090%)
  - 3-anonymity: 12079 (0.190%)
  - 5-anonymity: 25374 (0.400%)
```

SDC Methods
Categorical Variables

Local Suppression

This method **replaces unique combinations** of key variables with **missing values**, such that the **identification risk does not exceed a threshold.**

```
> sdc <- localSupp(sdc, keyVar = "sitProf", threshold = 0.05)
> print(sdc)
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 525 (0.008%) | in original data: 5713 (0.090%)
  - 3-anonymity: 1325 (0.021%) | in original data: 12079 (0.190%)
  - 5-anonymity: 2723 (0.043%) | in original data: 25374 (0.400%)
```

```
> sdc <- localSupp(sdc, keyVar = "escEtario", threshold = 0.05)
> print(sdc)
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 15 (0.000%) | in original data: 5713 (0.090%)
  - 3-anonymity: 45 (0.001%) | in original data: 12079 (0.190%)
  - 5-anonymity: 129 (0.002%) | in original data: 25374 (0.400%)
```

Identification Risk

Unique Combinations

```
> sdcf@origData[sdcf@risk$individual[,2]==1,c("genero", "escEtario", "agregFam2", "habil", "sitProf", "nuts3")]
          genero escEtario  agregFam2            habil                      sitProf nuts3
111602  Feminino      <=19          4 Sem escolaridade                    Estudante   16H
236676             [50-59]          3 Sem escolaridade Empregado por conta de outrem   11E
493464             [50-59]          4 Sem escolaridade Empregado por conta de outrem   111
767541                 60+          3 Sem escolaridade   Empregado por conta propria   11D
905196                 60+          4 Sem escolaridade Empregado por conta de outrem   11E
963578                 60+          4 Sem escolaridade Empregado por conta de outrem   16B
1105916                60+          2 Sem escolaridade   Empregado por conta propria   11E
1708566            [30-39]          3 Sem escolaridade Empregado por conta de outrem   186
2175266            [50-59]          3 Sem escolaridade Empregado por conta de outrem   16I
2612229            [40-49]          4 Sem escolaridade                                 16G
2753516                60+          3 Sem escolaridade                    Reformado   16J
3913271            [50-59]          4 Sem escolaridade Empregado por conta de outrem   11D
4665663            [40-49]          4 Sem escolaridade                                 16F
5610168                60+ 5+ pessoas Sem escolaridade                                 11B
5643967                60+          3 Sem escolaridade                                 181
>
```

Individual risk

```
> summary(sdcf@risk$individual[,1])
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000199 0.0003035 0.0009709 0.0026719 0.0026954 1.0000000
```

Global risk

```
> sdcf@risk
$global
$global$risk
[1] 0.002671868
```

Conclusions

- Regarding the ***K*-anonymity** calculation we can see that, in general, for large databases there is a **high number of observations with a unique combination of key variables**.
- In this case, replacing the **municipality** variable with the **nuts3** variable **strongly reduced the number of unique combinations**, which went from 1.56% to just 0.13%.
- There are several **statistical disclosure control methodologies** that **reduce the risk of identification**, such as **recoding** and **local suppression** methods that generally apply to **key categorical variables**.

Challenge for Future

- **Identification Risk for Panel Data**.

References

- Benschop, T., Machingauta, C., & Welch, M. (2021). Statistical disclosure control: A practice guide. *The World Bank*.
- Templ, M., & Sariyar, M. (2022). A systematic overview on methods to protect sensitive data provided for various analyses. *International Journal of Information Security*, 21(6), pp. 1233-1246.
- Templ, M. (2017). Statistical disclosure control for microdata. *Springer*.
- Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014). Introduction to statistical disclosure control (sdc). *Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG*.
- **Li, S., Schneider, M. J., Yu, Y., & Gupta, S. (2023). Reidentification risk in panel data: Protecting for k-anonymity. Information Systems Research, 34(3), pp. 1066-1088.**
- Morais, J. (2022). Comparação de métodos perturbativos: utilidade e perda de informação em bases de microdados (Master’s dissertation, University of Minho, Portugal).
- POCH, Programa Operacional Capital Humano (2019). Regulamento Geral sobre a Proteção de Dados. https://www.poch.portugal2020.pt/ptpt/Candidaturas/Documents/POCH%20%20Gui%C3%A3o%20RGPD_Entidades%20Benefici%C3%A1rias_v8.0_rev.pdf. Online; Access 27/02/2024.
- Sampaio, S., Sousa, P. R., Martins, C., Ferreira, A., Antunes, L., & Cruz-Correia, R. (2023). Collecting, processing and secondary using personal and (pseudo) anonymized data in smart cities. *Applied Sciences*, 13(6), pp. 3830.
- Lee, Y.J., & Lee, K.H. (2018). What are the optimum quasi-identifiers to re-identify medical records? In: 20th International Conference on Advanced Communication Technology (ICACT), pp. 1025–1033. *IEEE*.

Thank you for your attention!