# 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]Senior Data Analyst, Bank of Portugal*

## Abstract

This article highlights the growing demand for robust Statistical Disclosure Control (SDC) methodologies in an era of large amounts of data and statistical information available. The balance between the data utility and the preservation of privacy is extremely important today, particularly given the evolution of legal frameworks such as the European Union's General Data Protection Regulation (GDPR). The article provides a comprehensive overview of the main concepts, different anonymization methodologies and emphasizes the importance of assessing the risk of identification and the loss of information. The main disclosure (identification) risk assessment measures for categorical and numerical variables are presented. Particular attention is paid to the risk of identification in longitudinal data, with proposed methodologies aimed at improving privacy protection. At the end, there is a discussion on the application of these concepts in real-world scenarios, namely in a financial database, highlighting ongoing research efforts to address privacy challenges in making individual microdata available on a panel.

**Keywords:** data utility, disclosure/identification risk, statistical disclosure control, data confidentiality

## 1. Introduction and Terminology

Nowadays, with the increase in demand for data and statistical information, it is essential to implement a set of confidentiality control methodologies. Data privacy concerns generally arise from legal reasons, related to the protection of individual confidentiality, with the aim of providing the best and greatest amount of information possible without compromising its quality and privacy (Mendes, 2011). Most of the data from certain statistical units corresponds to individual information, known as microdata. Statistical institutions have the great challenge of guaranteeing the confidentiality of statistical units in the dissemination of a detailed microdata set (Viana, 2014).

The Statistical Disclosure Control (SDC) techniques consist of a set of tools that can improve the level of confidentiality of any dataset, preserving to a greater or lesser extent its information detail, which allows institutions to publish their data in a safe and efficient way for the user (Benschop et al., 2021). The identification risk is the probability of an intruder identifying at least one respondent in the available microdata. Given the existence of different identification risk scenarios and SDC methods, a balance between data utility and privacy must be considered (Templ and Sariyar, 2022).

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. Approved in 2016 by the European Parliament and the European Council, the GDPR came into force on May 25, 2018, with direct application in the legal system of the different Member States of the European Union (EU) and replacing the 1995 Data Protection Directive, which determined minimum standards for processing data in the EU (POCH, 2019).

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

In this study, consider a data set ($W = \{X, S\}$) of size $N$, where $X$ is a set of $P$ sensitive or confidential variables and $S$ is a set of $Q$ non-confidential variables. Let $Y$ be the set of perturbed (or modified) variables $X$ and $Z$ the set of key variables that belongs to $S$.

This section introduces the need to assess identification risk and presents the main terminology definitions for this paper. Then, Section 2 describes the main concepts, while Section 3 defines the identification risk and main measures for its assessment. Section 4 presents the application results to a real microdata set and finally, Section 5 shows the main conclusions.

## 2. Main Concepts

This section introduces some relevant concepts for the application of Statistical Disclosure Control (SDC) methodologies.

### 2.1 Variables Classification

In Statistical Disclosure Control methods, it is important to identify variables according to the risk of identification. According to Benschop et al. (2021) we can classify the variables as:

- **Direct identifiers:** variables that provide direct information about the individuals; example: 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; example: combination of age, sex and residence.
- **Non-identifiers:** variables that do not provide direct and indirect information to identify individuals; example: socioeconomic, demographic or behavioral characteristics.

In SDC methods, the variables are also classified according to their level of sensitivity or confidentiality. Only indirect identifiers or non-identifiers can be classified as **sensitive variables** and **non-sensitive variables**. Sensitive variables require increased care in their

analysis and/or disclosure, as they 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. On the other hand, 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 (Benschop et al., 2021).

### 2.2 Risk of Identification and Loss of Information

When applying SDC methods, it is important to calculate the identification risk, as well as assess the possible loss of information. Even if the data is anonymized, there may still be a risk of identifying individuals, which leads to compromising the privacy and security of information (Templ et al., 2014).

**Identification risk** is defined as the probability of a user anticipating the values $X$ from the conditional density function $f(X|S)$ (Morais, 2022). Once the set of perturbed/modified confidential variables $Y$ is disclosed, users have additional information and the conditional density function $f(X|S,Y)$ is considered to obtain the values of $X$. However, if $f(X|S,Y) = f(X|S)$, access to the modified data does not offer additional information to users and in this case, the risk of identification is minimum. Therefore, it is important to evaluate the identification risk of the original data set $(W = \{X,S\})$, as well as the modified data set $(W^* = \{Y,S\})$ for choosing the most appropriate statistical disclosure control (SDC) method.

**Information loss** occurs when there is a decrease in the quantity or quality of original data during the dissemination process. Therefore, it is necessary to evaluate the information that was lost, comparing the published values with the values in the original microdata set. Prior to applying the methods to the original data, it is considered that these data have zero lost information (Templ et al., 2014).

There are several measures to evaluate the information loss in the SDC process, which seek to establish a balance between the risk of identification and data utility for users.

### 2.3 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). Anonymization must be an irreversible process, but considering technological developments, all costs, resources and knowledge necessary for possible re-identification must be considered (Sampaio, 2023). In this context, other terminologies become relevant in this study, as they are associated with data anonymization.

De-identification and pseudonymization are also techniques used to reduce the likelihood of identifying individuals in a personal database.

**De-identification** aims to remove or hide all personal information from a dataset to make it impossible to identify individuals. It is not necessarily an irreversible method, as it is possible to have a mapping table that is capable of reversing this method, linking the original values to the ones values that were de-identified. Typically, this method generates changes to the indirect identifier through generalization processes (such as modifying the scale of an attribute) or through the inclusion of uncertainty factors based on the original values (Sampaio, 2023).

**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. The data controller must de-identify the information to process and store the information separately and securely so that the two parties are not able to be brought back together. Therefore, it is not possible to identify the individual in the legally pseudonymized dataset (Sampaio, 2023).

Since the GDPR came into force, pseudonymization and anonymization have played an important role with regard to data processing, security and access. Pseudonymized data is still considered personal data while anonymized data is not (Kamińska, 2022).

### 2.4 Statistical Disclosure Control Methods (SDC)

SDC methods are typically known as anonymization methods, that is, they use strategies to ensure that the disclosed data does not reveal significant or recognizable information about individuals or entities. These methods can be classified as:

- **Perturbative methods:** involve the manipulated introduction of disturbances to the original data to preserve privacy, that is, adding noise or modifying the data in some way, in order to maintain the utility of the data and reduce the risk of identification (Templ, 2017).
- **Non-perturbative methods:** aim to protect privacy without directly introducing noise into the data, that is, changing the structure of the data in a way that ensures that the identity of individuals cannot be easily discovered (Templ, 2017).
- **Synthetic data generation:** refers to the creation of data sets that are artificially generated to resemble real data while maintaining relevant statistical and structural characteristics. The use of these methods presents a lower risk of identification, despite being a difficult process to apply (Viana, 2014).

In addition to perturbative and non-perturbative methods and the generation of synthetic data, it is also possible to divide the methods into probabilistic or deterministic (Morais, 2022). **Probabilistic methods** consist of probabilistic mechanisms or random number generating. **Deterministic methods** are based on a specific algorithm and generate the same results if applied consecutively to the same database.

## 3. Identification Risk

Some measures are presented to calculate the identification risk for categorical and numerical variables, and it is shown how to calculate the individual and the global risk.

### 3.1 Identification Risk Measures for Categorical Variables

***Frequency Counts.*** In the context of identification risk, frequency counts are usually performed for the group of key variables. Thus, consider $f_k$ and $F_k$ as the number of observations of the combination $k$ of key variables in the sample and population, respectively (Templ, 2017).

***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 $K$ (Benschop et al., 2021).

***L-diversity.*** This measure 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 (Templ, 2017).

### 3.2 Disclosure Risk Measures for Numerical Variables

***Record Linkage.*** It is a method that evaluates the correct number of links between published values and original values (Templ et al., 2014). Let $y_{ip}$ be the modified observation of the original $x_{ip}$. Consider $x_{1p}$ and $x_{2p}$ to be the closest observations to yip 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.*** Intervals are created around each published value, and it is checked whether the original value belongs to the established interval (Domingo-Ferrer, 2001).

***Outliers Count.*** It is carried out by identifying values that are higher or lower than a certain percentile (Templ, 2017).

### 3.3 Individual and Global Risk Identification Risk

Individual risk is the probability of identifying an individual observation, while global risk is the proportion of observations that can be identified by a user. The **individual risk** is calculated as $r_i = 1/F_k$, where $F_k$ corresponds to the population frequency of the combination $k$ of key variables, to which observation $i$ belongs. The **global risk** is often calculated by the arithmetic average of all individual risks (Morais, 2022): $R = (1/N)\sum_{i=1}^{N} r_i$.

As an alternative to the individual identification risk, there is the **Special Uniques Detection Algorithm (SUDA)**, which allows identifying observations with the highest risk. The application of the SUDA algorithm presupposes the classification of sets of key variables as Minimal Sample Unique (MSU - smallest set of unique combinations of key variables in the sample/population) in addition to assigning a **SUDA score** to each observation in the microdata set, which corresponds to the risk of an individual being identified. The identification risk is greater the smaller the size of an MSU set or the greater the number of MSU sets (Benschop et al., 2021; Morais, 2022).

### 3.4 Identification Risk for Panel Data

The identification risk in panel data is potentially much greater than in conventional measures that are adopted for cross-sectional data. Li et al. (2023) propose a new methodology for calculating $k$ -anonymity for panel data, as existing approaches for cross-sectional data cannot be directly applied to panel data. The authors bring three approaches: unicity, snowball unicity (sno-unicity) and a new proposed method, the graph-based minimum movement $k$-anonymization - $k$-MM.

This study shows the calculation of identification risk in a microdata set from the Central Credit Responsibility (CCR) for a given moment in time. However, this is a work in progress and therefore, we intend to explore and implement these new methodologies for panel data in the future.

## 4. Case Study

The database under study belongs to the Central Credit Responsibility (CCR) of Banco de Portugal (BdP). The main focus in this study is on the bases of individuals, namely on the set of key variables that can allow their identification. The database, which was made available by the Banco de Portugal Microdata Research Laboratory (BPLIM), is already de-identified and pseudonymized, as it does not have direct identifiers and the unique identifier was replaced by a fictitious code.

The database under study contains 6342255 observations relating to the credit records of Portuguese individuals in December 2022. Table 1 describes the key variables considered for this study. For more information about the categorical variables of the study, see **Appendix A**.

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 |

In this study, we compute individual and global risks using the R package sdcMicro. Considering the set of 6 categorical variables as the key variables of the database under study, the results for the *K*-anonymity measure are shown in Figure 1.

Figure 1: Initial $K$-anonymity results

```
--------------------------
134438 obs. violate 2-anonymity
249763 obs. violate 3-anonymity
--------------------------
```

We can see that there is a high number of observations that does not guarantee a minimum number of observations for the combination of key variables. Taking into account that the municipality of residence variable is very disaggregated, with 309 categories, we will consider the variable **nuts3,** which contains level 3 of the Nomenclature of Territorial Units for Statistics (NUTS III - see Table 7 in Appendix A). The results for the *K*-anonymity measure are presented in Figure 2.

Figure 2: $K$-anonymity results when using the variable nuts3

```
--------------------------
423 obs. violate 2-anonymity
891 obs. violate 3-anonymity
--------------------------
```

Replacing the **concelho** variable with the **nuts3 variable** strongly reduced the number of unique combinations, which went from 134438 to just 423. This number is very small considering the database's large size, but we can apply SDC methods such as suppression or recoding to ensure that it is not possible to identify individuals or that this risk is very low. The R code used to obtain these results can be found in Appendix A.

## 5. Conclusions

This article addresses the need to use robust Statistical Disclosure Control (SDC) methodologies and to establish a balance the data utility and the preservation of privacy, especially in an era of large amounts of data and legal restrictions, such as those established by the General Regulation Data Protection Regulation (GDPR) of the European Union. The main concepts discussed include variable classification, identification risk, information loss and anonymization techniques. Through a case study involving a financial database, specifically the microdata base of the Credit Responsibility Center (CCR), it demonstrates the practical application of these methodologies and highlights ongoing research efforts to address privacy challenges.

## References


Benschop, T., Machingauta, C., & Welch, M. (2021). Statistical disclosure control: A practice guide. *The World Bank.*

Domingo-Ferrer, J. (2001). Confidentiality, Disclosure and Data Access. *Elsevier Science.*

ISO, International Organization for Standardization (2011). ISO/IEC 29100:2011. https: //www.iso.org/standard/45123.html. Online; Access 13/03/2024.

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.

Mendes, E. (2011). Confidencialidade de Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística (Master's dissertation, University of Porto, Portugal).

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.

Kamińska, J. (2022). Pseudonymization vs anonymization: differences under the GDPR. https://www.statice.ai/post/pseudonymization-vs-anonymization. Online; Access 28/02/2024.

Viana, I. (2014). Métodos de Geração de dados sintéticos para a criação de microdados de uso público (Master's dissertation, University of Porto, Portugal).

## Appendix A

Table 2: Variable gender

| Code | Designation |
|---|---|
| **000** | Not available |
| **001** | Female |
| **002** | Male |

Table 3: Variable escEtario

| Code | Designation |
|---|---|
| **<=19** | Up to 19 years old |
| **[20-29]** | Between 20 and 29 years old |
| **[30-39]** | Between 30 and 39 years old |
| **[40-49]** | Between 40 and 49 years old |
| **[50-59]** | Between 50 and 59 years old |
| **60+** | 60 or more years old |

Table 4: Variable agregFam

| Code | Designation |
|---|---|
| **1** | 1 person |
| **2** | 2 people |
| **3** | 3 people |
| **4** | 4 people |
| **5** | 5 people |
| **6** | 6 people |
| **7+** | 7 or more people |

Table 5: Variable habLit

| Code | Designation |
|---|---|
| **000** | Not available |
| **001** | No education |
| **002** | Basic |
| **003** | Secondary |
| **004** | Higher |

Table 6: Variable sitProf

| Code | Designation |
|---|---|
| **000** | Unknown |
| **001** | Student |
| **002** | Retired |
| **003** | Employee |
| **004** | Self-employed |
| **005** | Unemployed |
| **006** | Out of the job market* |

*Out of the job market - professional situation of individuals who do not have a job and are not looking for one.

Table 7: Variable nuts3

| Code | Designation |
|---|---|
| **111** | Alto Minho |
| **112** | Cávado |
| **119** | Ave |
| **11A** | Área Metropolitana do Porto |
| **11B** | Alto Tâmega |
| **11C** | Tâmega e Sousa |
| **11D** | Douro |
| **11E** | Terras de Trás-os-Montes |
| **150** | Algarve |
| **16B** | Oeste |
| **16D** | Região de Aveiro |
| **16E** | Região de Coimbra |
| **16F** | Região de Leiria |
| **16G** | Viseu Dão Lafões |
| **16H** | Beira Baixa |
| **16I** | Médio Tejo |
| **16J** | Beiras e Serra da Estrela |
| **170** | Área Metropolitana de Lisboa |
| **181** | Alentejo Litoral |
| **184** | Baixo Alentejo |
| **185** | Lezíria do Tejo |
| **186** | Alto Alentejo |
| **187** | Alentejo Central |

Code 1: R code used for the results presented in this article

```
rm(list=ls())
library(sdcMicro)
library(simPop)

dataset <- read.csv("/bplimext/projects/p168_ValentinaLirio/work_area/dados.csv", encoding = "UTF-8")
head(dataset)

dataset$escEtario <- ifelse(dataset$idade<=19,"<=19",
                      ifelse(dataset$idade>19 & dataset$idade<=29,"[20-29]",
                        ifelse(dataset$idade>29 & dataset$idade<=39, "[30-39]",
                          ifelse(dataset$idade>39 & dataset$idade<=49, "[40-49]",
                            ifelse(dataset$idade>49 & dataset$idade<=59, "[50-59]",
                              ifelse(dataset$idade>=60, "60+",""))))))

dataset$genero<-factor(dataset$genero)
dataset$escEtario<-factor(dataset$escEtario)
dataset$agregFam<-factor(dataset$agregFam)
dataset$habLit<-factor(dataset$habLit)
dataset$sitProf<-factor(dataset$sitProf)
dataset$concelho<-factor(dataset$concelho)
dataset$nuts3<-factor(dataset$nuts3)
summary(dataset)

#alpha=0 ignores missing values
f1<-freqCalc(dataset, keyVars = c("genero", "escEtario", "agregFam","habLit", "sitProf", "concelho"),
             w=NULL, alpha = 0 )
f1

#alpha=1 - without ignoring missing values
f2<-freqCalc(dataset, keyVars = c("genero", "escEtario", "agregFam","habLit", "sitProf", "concelho"),
             w=NULL, alpha = 1 )
f2

#with the nuts3 variable instead of concelho
f3<-freqCalc(dataset, keyVars = c("genero", "escEtario", "agregFam", "habLit", "sitProf", "nuts3"),
             w=NULL, alpha = 1 )
f3
```