11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"

BIS Basel, 25-26 August 2022

Joint secondary anonymisation of categorial and numerical variables in sensitive time series microdata – novel approach for Statistical Disclosure Control of a sensitive microdata set published in BELab data laboratory[1]

Eugenia Koblents and Alberto Lorenzo Megía,
Bank of Spain

[1] This presentation was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

Joint secondary anonymisation of categorical and numerical variables in sensitive time series microdata

A novel approach for Statistical Disclosure Control of a microdata set published in BELab data laboratory

Eugenia Koblents Lapteva and Alberto Lorenzo Megía

Abstract

In this paper, a Statistical Disclosure Control (SDC) approach for secondary anonymisation of sensitive time series microdata is proposed that allows the joint analysis of numerical and categorical key variables. This method has been developed at the Banco de España's BELab data laboratory in order to protect confidentiality of the recently published CIR dataset. This dataset contains yearly microdata on loans to legal entities, including multiple variables describing the loans and debtors. The main challenge faced in this work is the fact that the set of key variables, i.e. those that may allow debtor re-identification, includes both categorical and numerical loan and debtor variables. Additionally, debtors may have multiple loans and loans may have multiple debtors, which makes the direct use of existing SDC software tools for microdata protection (mu-argus, sdcMicro) unfeasible. For these reasons, a novel SDC procedure has been designed and implemented in order to protect the debtors appearing in the CIR dataset against re-identification, while jointly analysing categorical and numerical variables and addressing time series data protection.

Keywords: Statistical Disclosure Control, numerical and categorical key variables, time series data protection.

JEL classification: C4 (Econometric and Statistical Methods: Special Topics)

Contents

1. Introduction

This paper addresses Statistical Disclosure Control (SDC) of sensitive microdata in the context of the Banco de España's BELab data laboratory [5]. The goal of SDC is to minimize the risk of re-identification of individual samples while minimising the information loss produced by the anonymisation in order to retain information utility [1]. The SDC process requires the identification of feasible disclosure scenarios and key variables that might allow the re-identification of individual samples in the microdata set under analysis.

This work entailed analysis of a dataset containing information on loans (in the following referred to as the CIR dataset), which has recently been published in a BELab safe data room [6]. The CIR (Central de Información de Riesgos) dataset contains yearly microdata on loans extended to legal entities, resident and non-resident in Spain, that were reported to the Central Credit Register (CCR) between 2016 and 2020. This dataset includes multiple variables describing loans and debtors, and does not contain information on the financial institutions involved. This dataset required both primary and secondary anonymisation to avoid debtor re-identification, and has some distinctive features that make the direct use of existing SDC software tools for microdata protection (mu-argus, sdcMicro) unfeasible. Additionally, existing SDC software tools have some other limitations that have also been addressed in this work.

As a result, a novel SDC procedure has been designed and implemented in order to protect debtors appearing in the CIR dataset against re-identification. The implemented SDC procedure makes use of the open source R package sdcMicro [2,3]. Both sdcMicro and mu-argus [1,4] are Eurostat-supported SDC software tools used by many public institutions, such as national statistical institutes and central banks. Both implement a broad variety of SDC methods for individual risk evaluation, microdata protection and information loss assessment. In this work, the sdcMicro package has been used because its creators claim that it is better optimised for large volumes of data [2,3] as compared with mu-argus. Pre- and post-processing of the CIR dataset has been implemented in Python. The proposed secondary anonymisation procedure has been designed in close collaboration with the Banco de España's CIR Department and has been validated by a team of internal researchers who consistently work with the CIR dataset, to guarantee that the utility of the anonymised dataset is preserved.

Numerous previous works have addressed the independent protection of categorical and numerical key variables [1,8,10,11,12]. In particular, in [1] the authors provide an extensive overview of the main SDC methods for addressing microdata protection problems involving both categorical and numerical key variables. However, both types of variables are generally processed independently. In [8] the authors focus on the protection of outlying samples of numerical variables, and compare, for different masking methods, the information loss and disclosure risk related to outliers. A recent review of the state of the art of SDC and a discussion on future challenges (big data, machine learning, etc.) can be found in [11].

Previous attempts have also been made to jointly analyse categorical and numerical variables. In particular, in [7] the authors propose to exploit the hierarchy of categorical variables to compute a numerical mapping that quantifies their underlying semantics. This approach is similar to the one proposed in this work, in the sense that it aims to combine categorical and numerical variables by computing distances between categories, while we propose to transform numerical variables into

categorical ones. More recently, in [9] the authors discuss the limitations of existing SDC software and the need to jointly assess disclosure risk and information loss when both categorical and numerical key variables are identified. However, to the best of our knowledge, a general solution to this problem has not yet been found.

The rest of the paper is organised as follows. In Section 2 we describe the secondary anonymisation method designed and implemented in this work, including a description of the CIR dataset and the steps of the anonymisation procedure. The numerical results obtained are also presented. Finally, Section 3 is devoted to the conclusions and future lines of research.

2. Secondary anonymisation of CIR dataset

The goal of SDC is to protect statistical data by producing safe datasets with low (individual) risks and high data utility, and that can be securely released without compromising data confidentiality. The SDC procedure for protecting sensitive microdata consists of the following steps [2]:

1. Deletion of direct identifiers, to guarantee primary confidentiality.
2. Identification of key variables, to address secondary confidentiality.
3. Measurement of individual risks based on sample frequency counts.
4. Application of SDC methods to protect high-risk observations.
5. Assessment of the resulting disclosure risk and the information loss produced by the anonymisation procedure.

Software tools such as sdcMicro and mu-argus provide implementations for a broad variety of SDC methods that can be used in steps 3-5. However, the CIR dataset has a number of peculiarities that hinder the direct application of standard microdata protection methods. Therefore, a specific procedure has been designed and implemented to address this problem.

The CIR dataset currently available at BELab contains data describing loans extended to legal entities between 2016 and 2020. Around 25 million records are available, containing 19 variables describing debtors and loans. A complete sample is available representing the whole population. The CIR dataset contains the following variables describing debtors:

1. Debtor ID (anonymised)
2. Residence
3. Institutional sector
4. Economic activity
5. Enterprise size
6. Legal form

On the other hand, the variables describing loans are the following:

7. Loan ID (anonymised)
8. Type of instrument
9. Residual maturity
10. Currency
11. Collateral type
12. Collateral coverage
13. Personal guarantee
14. Personal guarantee coverage
15. Investment region
16. Joint debtor
17. Number of joint debtors
18. Drawn amount
19. Undrawn amoun

Variables 18 and 19 (drawn and undrawn amounts) are numerical variables while the rest of them are categorical. Debtor ID and loan ID have been anonymised by the

data provider prior to being shared with BELab to guarantee primary confidentiality. Loans are only active for a specific period of time. A detailed description of the CIR dataset is available in the user manual published by BELab [6].

Even though the original dataset contains information on loans, the sensitive entity to be protected is the debtor, which makes the direct use of existing SDC software tools unfeasible. The standard SDC procedure for microdata protection makes the assumption that each row in the dataset represents an individual respondent, which is not satisfied in this case. The CIR dataset includes debtors with multiple loans and loans with multiple debtors (joint loans). Additionally, several variables describing loans, such as investment region, currency and amounts, can also allow debtor re-identification, and must therefore also be considered as key variables.

Existing SDC software tools have other limitations which are relevant in this case. On the one hand, they do not allow a joint analysis of categorical and numerical variables and do not support time series data protection. On the other hand, the implemented anonymisation methods for numerical variables do not yield a good trade-off between re-identification risk and information loss in this case. Ideally, we would like to protect only those samples that turn out to be sensitive when numerical and categorical variables are jointly analysed and leave the rest of the samples unaffected. The computational cost should also be affordable even for large datasets.

The top/bottom coding method is very simple and fast and only affects a small number of samples (information loss is limited). However, this method does not allow the protection of samples that might turn out to be sensitive when numerical and categorical variables are jointly analysed. Thus, disclosure risk might not be sufficiently reduced under this approach. This method processes each numerical variable independently, ignoring correlations among variables, and requires the definition of individual thresholds for each variable that directly affect the resulting data utility and disclosure risk.

Alternatively, more complex methods based on micro-aggregation are widely used and recommended in the literature, since they allow a significant reduction in disclosure risk. However, their main limitation is the fact that all samples are affected by the anonymisation process, significantly reducing data utility in some scenarios. This family of methods also requires high computation times, which can hinder its application when working with large volumes of data.

Finally, perturbative methods, such as noise addition and rank swapping, have been discarded in this work due to the feedback received from the team of internal researchers, who claimed that data utility would be seriously affected by these transformations.

For all these reasons, a specific approach has been designed and implemented at BELab to address secondary anonymisation of the CIR dataset. The proposed method consists of encoding loan and debtor information into a so-called debtor profile, to ensure that the resulting dataset contains one single row per individual respondent and that standard SDC methods and tools can be used. Numerical key variables are discretised and incorporated into this profile, allowing the assessment of individual re-identification risks based on complete debtor and loan information. The proposed procedure consists of the following steps:

1. Identification of debtor and loan key variables.
2. Global recoding of selected key variables, reducing the number of classes.

3. Creation of a full profile for each debtor, including information on all of its loans (active at some point throughout the full time series).
4. Local suppressions performed on debtor profiles to ensure k-anonymity with the selected value of k.
5. Transfer of local suppression patterns identified for each debtor to the original loans dataset.

When new yearly data is incorporated into the dataset, the full process needs to be repeated (including all yearly data available to date) and a new anonymised time series dataset needs to be generated. Under this approach, each researcher has access to no more than one version of the time series dataset simultaneously. Otherwise, an intruder would be able to cancel the performed suppressions by comparing different datasets, since the suppression pattern would be different in both versions.

2.1. Key variable identification

Key variables are those that may lead to the disclosure of individual samples in feasible re-identification scenarios. Key variable selection is a challenging problem that requires close collaboration between the SDC expert and the data provider. In this case, even though the original dataset mainly contains information on loans, the entity to be protected is the debtor. For this reason, all variables describing debtors, except for the anonymised debtor ID, have been considered as key variables.

Additionally, selected variables describing loans have also been considered as key variables, since they can allow debtor re-identification in the defined disclosure scenarios. Anonymised debtor and loan IDs have not been taken into account for the anonymisation procedure, since they are internal identifiers that are not published externally and thus cannot be used for re-identification by potential intruders. Table 1 shows all the debtor and loan key variables identified in this case study.

Table 1. Debtor and loan key variables.

Debtor key variables	Loan key variables
Residence	Currency
Institutional sector	Personal guarantee
Economic activity	Investment region
Enterprise size	Drawn amount
Legal form	Undrawn amount

2.2. Global recoding of categorical key variables

Once the set of key variables has been identified, global recoding is commonly performed on selected categorical variables by grouping existing classes [1]. This method significantly reduces disclosure risk while incurring an acceptable information loss, since the recoded classes and the grouped categories are selected based on expert knowledge in order to maximise data utility. Global recoding does not involve sample suppression but reduces the level of detail of all samples. In this work, this process has been agreed upon with the data provider and a number of internal researchers to guarantee high information utility for the resulting data. The process has been implemented in Python instead of using sdcMicro, because a large number of categories had to be grouped in this case. Table 2 shows the original and modified number of categories for the recoded variables. The reduction in disclosure risk

achieved through this procedure cannot be numerically assessed, since at this point the data still does not contain one single row per individual respondent and is thus unsuitable for risk disclosure assessment using sdcMicro. However, global recoding significantly reduces disclosure risk when a large number of categories are grouped.

Table 2. Original and modified number of categories for the debtor and loan key variables selected for global recoding.

Categorical variables	Original categories	Modified categories
Institutional sector (debtor)	16	3
Economic activity (debtor)	167	21
Currency (loan)	56	4
Personal guarantee (loan)	5	4
Investment region (loan)	55	18

2.3. Creation of debtor profiles

The key step in the designed anonymisation procedure is the creation of a detailed profile for each debtor containing information on all of its loans throughout the time series. In the original CIR dataset each row contains loan and debtor information, while the entity to be protected in this case is the debtor. Loans can affect multiple debtors and debtors can be involved in multiple loans. For this reason, the original dataset does not contain a single row per individual respondent, which makes the direct use of standard SDC procedures unfeasible. To overcome these difficulties, the original dataset has been transformed in such a way that each row, called the debtor profile, contains all the information of a given debtor. All debtor and loan key variables, both categorical and numerical, need to be represented in the debtor profile in order to assess individual risks based on complete debtor information. To achieve this goal, categorical loan key variables have been encoded using one-hot encoding, while numerical loan key variables have previously been discretised.

All debtor key variables (Table 1, left-hand column) have been directly included in the profile. These variables correspond to fixed debtor attributes that are usually constant over time (company size, economic activity, etc.). However, in some cases these attributes might also change during the time series. In those cases, the most frequent value has been considered. Where there are multiple modes, the most recent value has been selected.

Additionally, categorical and discretised numerical variables describing loan operations (Table 1, right-hand column) have also been incorporated into the debtor profile using one-hot encoding. For categorical loan variables, an auxiliary binary variable has been created for each category of the original key variables. In particular, 18 auxiliary binary columns have been created for the "Investment region" variable, 4 columns for the "Currency" variable and 4 columns for the "Personal guarantee" variable. The presence and absence of loan operations with specific "Currency", "Guarantee" and "Investment region" attributes has been encoded as 1 and 0 respectively in the debtor profile. For example, if a debtor has invested in Andalusia and Madrid, a 1 will appear on those two columns and the rest of the values for "Investment region" will be 0. Table 3 shows several examples of the resulting codification of categorical loan key variables.

Table 3. Synthetic examples of debtor profiles containing auxiliary binary variables corresponding to the original "Currency" and "Investment region" categorical loan variables.

Debtor ID	EUR	USD	GBP	Other currencies	Madrid	Catalonia	Andalusia
52364	1	0	1	0	1	1	0
76354	1	1	0	0	0	0	1
75345	1	1	0	0	0	1	1
34564	1	0	0	1	1	1	1
45634	0	1	1	0	1	0	1

On the other hand, the two numerical key variables describing loans (drawn and undrawn amount) have also been incorporated into the debtor profiles in the following way. First, the maximum of the two amounts for each operation has been computed. Then, this maximum value has been discretised according to the number of digits, thus taking only a small number of possible values (loans with less than 6, 7, 8, 9, 10 and 11 digits or more). This discretisation process assumes that an intruder might know the order of magnitude of a debtor's operations but not the exact amounts. This assumption has been considered reasonable by the data provider.

A new binary variable has then been added to the debtor profile for each of these new categories. The existence of operations with a given number of digits for a given debtor has been encoded as 1 in the debtor profile matrix. For example, if a debtor has operations with 8 and 9 digits, there will be 1s in those columns in its profile and 0s in the rest of the columns representing amounts. Table 4 shows synthetic examples of the section of the debtor profile corresponding to the discretised amount variables.

Table 4. Synthetic examples of debtor profiles containing auxiliary binary variables corresponding to the original loan amount variables.

Debtor ID	1-6 digits	7 digits	8 digits	9 digits	10 digits	11 digits or more
52364	1	1	1	1	0	0
76354	1	1	1	1	1	0
75345	1	1	0	0	0	0
34564	1	0	1	0	0	0
45634	1	1	1	1	1	1

As a result of this process, a full profile for each debtor has been created, including information on all its loans, which can then be processed using standard SDC software tools such as mu-argus or sdcMicro. In this particular case, a total of 1,430,503 debtors with active operations in any of the years across the time series has been obtained. A profile containing 37 variables has been created for each debtor, containing 5 debtor variables and 32 one-hot encoded loan key variables.

2.4. Local suppressions performed on debtor profiles

Once a detailed profile has been created for each debtor including the relevant debtor and loan information, the sdcMicro tool has been used to evaluate disclosure risk and anonymise the debtor information [2,3]. In particular, local suppressions to achieve k-anonymity with k=3 have been performed. A dataset satisfies k-anonymity when there are at least k samples with the same combination of key variables. sdcMicro allows the local suppression of specific values in the dataset in order to guarantee that k-anonymity is satisfied for all samples. Perturbative SDC methods for

protecting categorical key variables have been discarded in this work because they introduce randomness that can significantly reduce the value of the resulting data for researchers.

The disclosure risk assessment conducted in sdcMicro revealed that 68,450 debtor profiles (4.78%) do not satisfy k-anonymity with k=3. Individual risk for each debtor is computed as the inverse of its number of replicas in the dataset. Samples with an individual risk above a threshold of 1/k = 0.33 thus need to be protected. After performing 78,394 suppressions (corresponding to 0.15% of the debtor information), k-anonymity with k=3 is satisfied for all samples. The execution time for this process was 13 hours. Table 5 shows the results provided by the sdcMicro tool.

Table 5. Number and percentage of samples violating k-anonymity in the original and anonymised debtor profiles, computed by sdcMicro.

k-anonymity	Original debtor profiles	Anonymised debtor profiles
2-anonymity	46,936 (3.28%)	0 (0%)
3-anonymity	**68,450 (4.78%)**	**0 (0%)**
5-anonymity	95,733 (6.69%)	14,353 (1%)

2.5. Local suppressions performed on the original loans dataset

Finally, the local suppression pattern obtained for each debtor has been transferred to the original CIR dataset for each of the operations. For example, if a local suppression of the variable "Investment region=Madrid" has been performed for a given debtor, only those operations of that debtor with the attribute "Investment region=Madrid" will be affected by local suppressions, but not operations with any other investment region. On the other hand, if a discretised 10-digit amount has been suppressed for a certain debtor, only those operations with 10-digit amounts will be affected. This process produced a total of 4,300,076 local suppressions in the whole time series, which corresponds to 0.95% of the suppressed values, over a total of 1,430,503 rows and 37 columns. Table 6 shows a summary of the results.

Table 6. Summary of the number of local suppressions per variable in the full time series.

Debtor and loan key variables	Number of suppressions	Percentage of suppressions
Residence	70,720	0.27
Institutional sector	81,436	0.31
Legal form	528,404	1.98
Economic activity	3,006,248	11.29
Enterprise size	447,127	1.68
Currency	44,428	0.17
Guarantee	15,623	0.06
Drawn amount	2,993	0.01
Undrawn amount	2,993	0.01
Investment region	100,244	0.38
TOTAL	**4,300,076**	**0.95**

As can be seen in Table 6, the variables with the highest number of suppressions are the debtor key variables "Economic activity", "Legal form" and "Enterprise size", with a 11.29%, 1.98% and 1.68% of suppressions, respectively. Only 0.01% of numerical values of drawn and undrawn amounts have been suppressed.

Fig. 1 (left-hand chart) depicts the joint density estimation based on a 1% sample of the original data and the marginal densities for both numerical variables in logarithmic units. The plot shows that the joint density has two main modes. The marginal density of the variable "Undrawn amount" is highly concentrated at low values (the median is actually 0). Suppressed numerical values with added noise are also represented. It can be observed that multiple outlying samples have been suppressed. However, operations with lower amounts that turn out to be sensitive when analysed in combination with other key variables (e.g. the largest loans per sector or per region) have also been protected. By contrast, other outlying samples have been protected by suppressing some of the categorical key variables instead of the numerical ones. This suppression pattern has been obtained using sdcMicro, which identifies the optimal solution to achieve k-anonymity with the suppression of a minimum number of sample values. Obtaining this suppression pattern, which allows the protection of sensitive numerical samples within their whole range of values, is not possible without performing the described analysis.

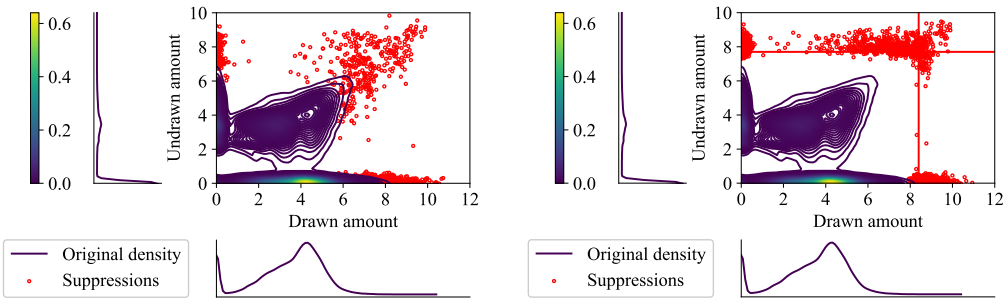


Fig. 1. Joint and marginal density estimation of the original data and suppressed values of the two numerical variables using the proposed method (left) and top/bottom coding (right).

For comparative purposes, Fig. 1 (right-hand chart) shows the suppression pattern obtained with top/bottom coding (with added noise), together with the threshold used for each variable. sdcMicro does not allow the joint protection of multiple numerical variables with top/bottom coding and each of them is protected independently. The suppression threshold for each variable has been adjusted so that the total number of suppressions is very similar for both methods (2,963 samples for top/bottom coding versus 2,993 samples for the proposed method). It can be observed that only samples with the highest values of one or both of the variables have been protected by top/bottom coding. Some of those samples are not really sensitive, as they satisfy k-anonymity and have thus been left unaffected by the proposed method. However, other samples that may indeed be sensitive when jointly analysing multiple key variables, have not been protected and disclosure risk may still be higher than desired.

Even though the number of suppressions is very similar in both cases, statistical distributions and summary statistics are affected in different ways, since top/bottom coding consistently suppresses the highest values in the population while the proposed method suppresses values in the whole range of values of each variable. To illustrate these differences, Fig. 2 shows the violin plot obtained from the original data, data anonymised using the proposed method and top/bottom coding, for the variables "Drawn amount" (left-hand chart) and "Undrawn amount" (right-hand chart) in logarithmic units. The plots show that only the tail of the distribution of both

variables is affected by anonymisation, while the rest of the distribution is preserved. It can also be seen that the proposed method affects the tails less than top/bottom coding.

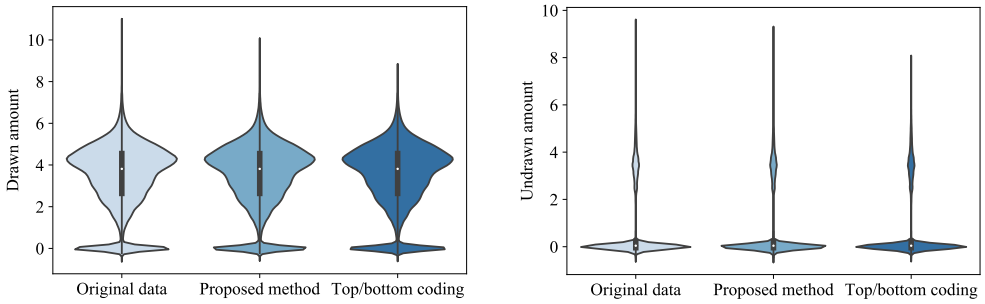


Fig. 2. Violin-plot of the "Drawn amount" (left) and "Undrawn amount" (right) variables. Original and anonymised data using the proposed method and top/bottom coding are shown.

Table 7 shows a comparison of the main summary statistics obtained using top/bottom coding and the proposed method. The deviation with respect to non-anonymised data has been computed for each variable. It can be observed that the maximum value of the data anonymised with top/bottom coding is reduced by 99% on average with respect to the maximum value of the original dataset, while only a 68% reduction is obtained with the proposed method. Similarly, the mean and standard deviation is far more affected when using top/bottom coding than the proposed method. The median is very slightly modified by both anonymisation procedures. These values illustrate how disclosure risk and information loss are significantly reduced when applying a suppression pattern that jointly analyses categorical and numerical variables.

Table 7. Comparison of summary statistics (deviation with respect to non-anonymised data) obtained with top/bottom coding and the proposed method for both numerical loan variables.

Summary statistics	Drawn amount		Undrawn amount	
	Top/bottom coding	Proposed method	Top/bottom coding	Proposed method
Maximum	-99.6%	-83.7%	-98.4%	-53.0%
Mean	-24.0%	-12.9%	-36.1%	-11.7%
Standard deviation	-88.7%	-68.3%	-79.7%	-34.6%
Median	-0.03%	-0.04%	-	-

3. Conclusions and future lines of research

In this paper, a strategy for performing secondary anonymisation of microdata combining numerical and categorical variables is proposed. The described method has been evaluated using a real and especially sensitive time series dataset, recently published by the BELab data laboratory. This dataset contains information on loans, while the sensitive entity to be protected is the debtor. Additionally, the set of key variables include both categorical and numerical debtor and loan variables. These particularities make the use of standard SDC methods unsuitable for this problem.

The key step in the proposed method is the creation of a debtor profile that incorporates complete information on all loans active at some point during the full

time series. The profiles include debtor key variables as well as one-hot encoded loan key variables. Once complete debtor profiles are created, standard SDC software tools such as scdMicro or mu-argus can be used to assess disclosure risk, protect sensitive samples and evaluate information loss. In this work, local suppressions to ensure k-anonymity with k=3 are performed to protect debtor information. The resulting suppression patterns are finally transferred to the original loan dataset.

The proposed method has a number of advantages over alternative procedures. It allows the joint protection of categorical and numerical variables, and generates suppression patterns that protect sensitive samples only and leave the rest unaffected. Disclosure risk is thus significantly reduced with very limited information loss. Additionally, statistical distributions and summary statistics are less affected in comparison with alternative methods such as top/bottom coding. This procedure makes use of existing SDC software tools, is simple to implement and cost efficient, which makes it suitable for large datasets. It allows the protection of time series data and only requires the selection of the k-anonymity k parameter.

Future lines of research include modelling the uncertainty regarding the information available to intruders, and exploring the relationship between SDC and anomaly detection to optimise results in both fields.

References

1. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P. P. (2012). Statistical disclosure control. John Wiley & Sons.
2. Templ, M., Meindl, B., and Kowarik, A. (2013). Introduction to statistical disclosure control (SDC). Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG.
3. Templ, M., Kowarik, A., and Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. Journal of Statistical Software, 67(1), 1-36.
4. Hundepool, A., De Wolf, P. P., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A. and Domingo-Ferrer, J. (2014). mu-Argus User's Manual version 5.1. Statistics Netherlands: The Hague, The Netherlands.
5. https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/
6. https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/prestamos-a-personas-juridicas--cir--54433f87573ad71.html
7. Domingo-Ferrer, J., Sánchez, D., and Rufian-Torrell, G. Anonymization of nominal data based on semantic marginality. Information Sciences 242 (2013): 35-48.
8. Mateo-Sanz, J. M., Sebé, F., and Domingo-Ferrer., J. Outlier protection in numerical microdata masking. International Workshop on Privacy in Statistical Databases. Springer, Berlin, Heidelberg, 2004.
9. Pietrzak, M. Statistical Disclosure Control Methods for Microdata from the Labour Force Survey. Acta Universitatis Lodziensis. Folia Oeconomica 3.348 (2020): 7-24.
10. Dandekar, R. A., Domingo-Ferrer, J. and Sebé, F. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. Inference Control in Statistical Databases. Springer, Berlin, Heidelberg, 2002. 153-162.
11. Elliot, M. and Domingo-Ferrer, J. The future of statistical disclosure control. arXiv preprint arXiv:1812.09204 (2018).
12. Fienberg, S. E., and Steele, R. J. Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics 14.4 (1998): 485.

BANCO DE ESPAÑA
Eurosistema

JOINT SECONDARY ANONYMISATION OF CATEGORICAL AND NUMERICAL VARIABLES IN SENSITIVE TIME SERIES MICRODATA

A novel approach for Statistical Disclosure Control of a microdata set published in BELab data laboratory

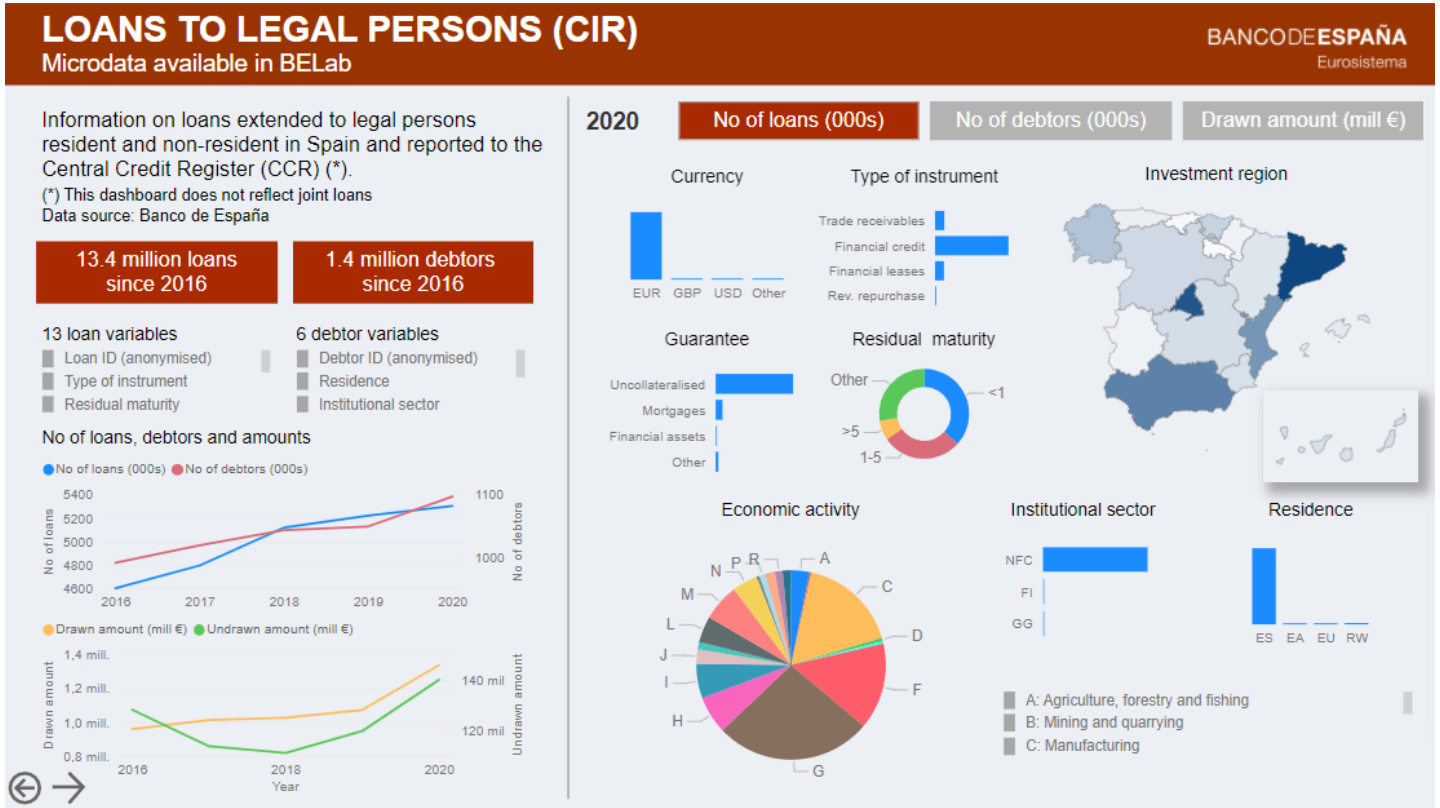
Eugenia Koblents
Alberto Lorenzo

ELEVENTH IFC CONFERENCE ON *"POST-PANDEMIC LANDSCAPE FOR CENTRAL BANK STATISTICS"*
BIS, BASEL, 25 AND 26 AUGUST 2022
INEXDA SESSION: MICRODATA DISCLOSURE CONTROL: A PRACTICAL PERSPECTIVE

25/08/2022

STATISTICS DEPARTMENT

- Banco de España launched **BELab** in July 2019 to provide access to the research community to high quality microdata via on-site and remote access. https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/
- In October 2021 CIR (*Central de Información de Riesgos*) Department of Banco de España provided a very **sensitive dataset** to BELab containing information on **loans to legal entities** extended between 2016 and 2020.
- Primary and **secondary anonymisation** was required to protect debtor confidentiality.
- Categorical and numerical debtor and loan **key variables** have been identified.
- CIR dataset contains **multiple rows per debtor and loan** (joint loans).
- A novel SDC approach for secondary anonymisation has been designed and implemented which allows to **jointly analyse categorical and numerical key variables**.
- The implemented approach makes use of the open-source R package **sdcMicro** for risk assessment and microdata protection and **Python** for data pre and post processing.



- Existing **SDC software tools** (sdcMicro, mu-argus) have some limitations which are relevant in this case:
  - They require data to contain **one single row per individual respondent**, which often is not the case.
  - They do not support a **joint analysis of categorical and numerical variables**.
  - Implemented anonymisation methods for **numerical key variables** (top/bottom coding, micro aggregation, noise addition, etc.) do not yield a good trade-off between disclosure risk and information loss for this problem.
  - They do not support **time-series data protection**.
- **A novel secondary anonymisation approach** has been designed and implemented which overcomes these difficulties:
  1. Identification of continuous and numerical debtor and loan **key variables** that can allow debtor re-id.
  2. **Global recoding** of selected key variables reducing the number of classes and disclosure risk.
  3. Creation of **full debtors' profiles** that incorporate information on all their loans throughout the full time series.
  4. **Debtor anonymisation:** local suppressions performed on debtor profiles with **sdcMicro** to guarantee k-anonymity.
  5. **Transfer of local suppression** patterns of debtors to the original **loans** dataset.
- When **new yearly data** is incorporated to the dataset the full process needs to be repeated, requiring that each researcher only has access to one version of the dataset to avoid the cancellation of local suppressions.

1. Identification of categorical and numerical debtor and loan **key variables** that can allow debtor re-id in feasible disclosure scenarios. This is a challenging problem that needs to be addressed in collaboration with the data provider:

Debtor key variables	Loan key variables
Residence	Currency
Institutional sector	Personal guarantee
Economic activity	Investment region
Enterprise size	Drawn amount
Legal form	Undrawn amount

2. **Global recoding** of selected categorical debtor and loan key variables by grouping existing classes to significantly reduce disclosure risk. This process is agreed with the data provider and data users to guarantee high **data utility**.

Categorical variables	Original categories	Modified categories
Institutional sector (debtor)	16	3
Economic activity (debtor)	167	21
Currency (loan)	56	4
Personal guarantee (loan)	5	4
Investment region (loan)	55	18

3. **Debtors' profiles** are created, which contain information on all their **loans** extended in the whole time-series.

- **Categorical key variables describing loans** have been incorporated into the profile using **one-hot encoding** (an auxiliary binary variable has been created for each category of the original key variables).

Debtor ID	EUR	USD	GBP	Other currencies	Madrid	Catalonia	Andalusia
52364	1	0	1	0	1	1	0
76354	1	1	0	0	0	0	1
75345	1	1	0	0	0	1	1

- **Continuous key variables describing loans** are discretized according to the number of digits (loans with 1-6, 7, 8, 9, 10 and 11 digits) and are incorporated into the profile in the same way as categorical variables.

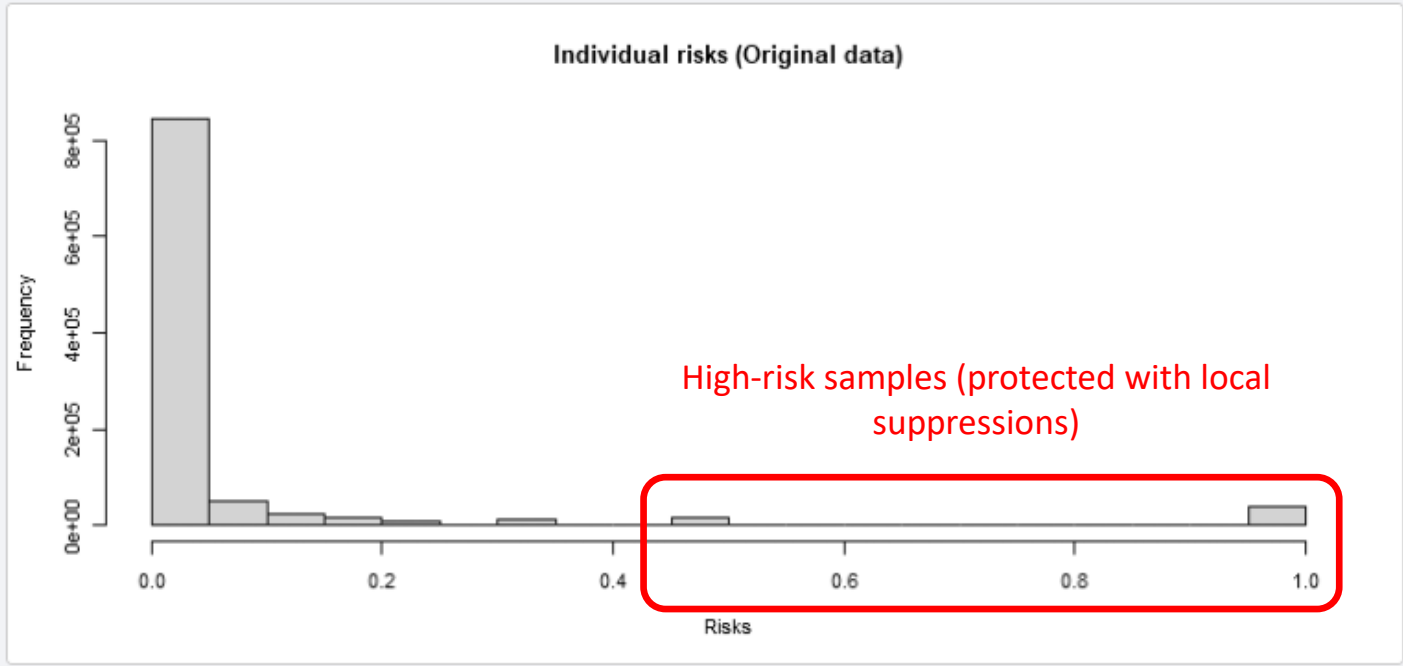
Debtor ID	1-6 digits	7 digits	8 digits	9 digits	10 digits	11 digits or more
52364	1	1	1	1	0	0
76354	1	1	1	1	1	0
75345	1	1	0	0	0	0

- As a result, **1.430.503 debtor profiles** (with one **single row per individual respondent**) have been created containing **37 variables**: 5 debtor key variables and 32 one-hot encoded loan key variables.

4. Debtor profiles contain **one single row per individual** and existing SDC software can be used to evaluate individual disclosure risks and to apply local suppressions to sensitive debtor profiles to guarantee **k-anonymity** (k=3).

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	46936 (3.281%)
3-anonymity	0 (0.000%)	68450 (4.785%)
5-anonymity	14353 (1.003%)	95733 (6.692%)

Disclosure risk evaluation performed by **sdcMicro** before and after applying local suppressions to debtors' profiles.

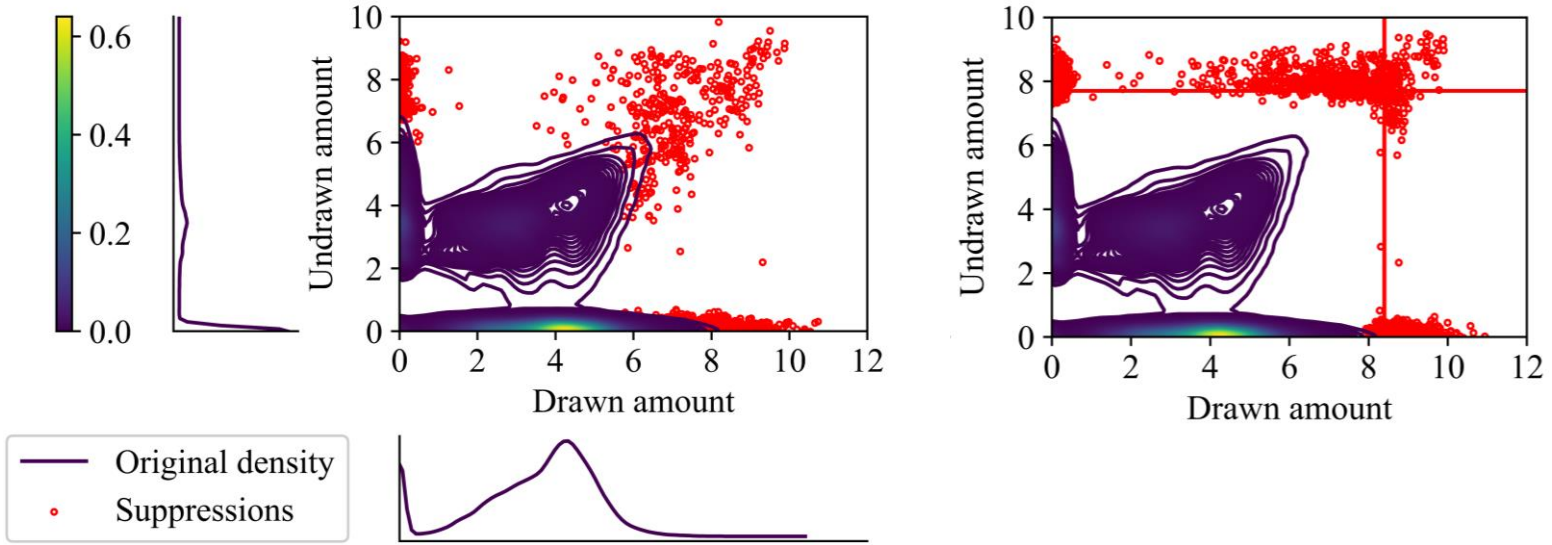


- **4.78%** of debtors are at risk of re-id in the original dataset, when combining categorical and numerical key variables.
- **0.15%** of debtors' information has been supressed (78.394 values of samples).
- As a result, all debtors in the anonymised dataset satisfy **k-anonymity** with k=3 (all individual risks are below 0.33).

5. Local suppressions obtained for debtors are **transferred** to the original loans dataset. **0.95%** of suppressions on average (mainly economical activity, legal form and company size). **0.01%** of numerical values suppressed.

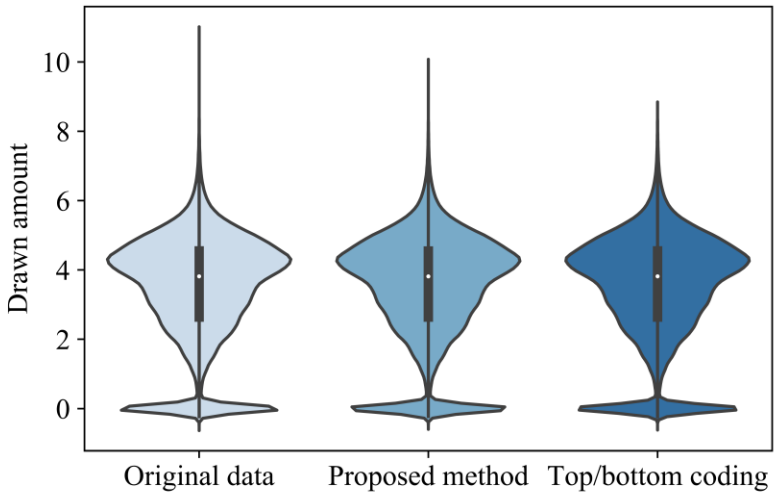
Debtor and loan key variables	Number of suppressions	Percentage of suppressions
Legal form	528,404	1.98
Economic activity	3,006,248	11.29
Enterprise size	447,127	1.68
Drawn amount	2,993	0.01
Undrawn amount	2,993	0.01
TOTAL	**4,300,076**	**0.95**

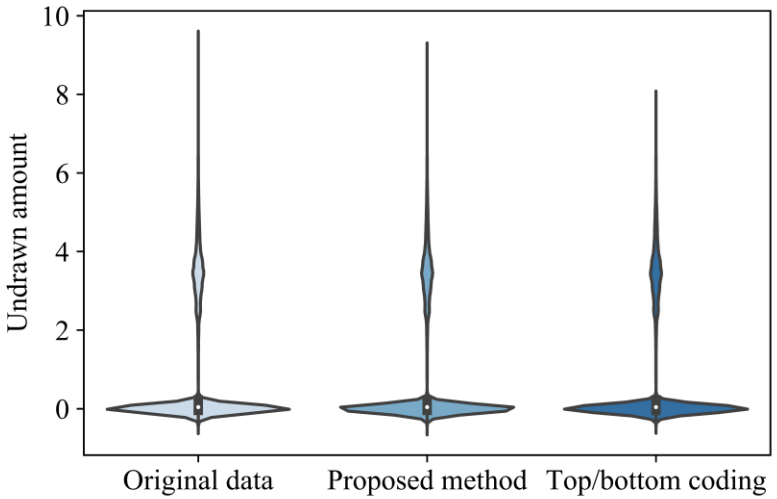
- Suppression pattern (red dots) obtained with the **proposed method** (left) and **top/bottom coding** (right).
- The proposed method protects samples that turn out to be sensitive when categorical and numerical variables are **jointly analysed** (largest loan per sector or region, etc.), while top/bottom coding consistently supresses **high-valued samples**.



- The proposed method affects the **data distribution and summary statistics** less than top/bottom coding. The obtained suppression pattern yields a **low information loss and disclosure risk**, since only sensitive samples are modified.

- Only the **tail of the distribution** is affected by anonymisation. The proposed method affects the tails less than top/bottom coding, because less outliers are suppressed.





- **Summary statistics** are significantly less affected using the proposed method than top/bottom coding.

Summary statistics	Drawn amount		Undrawn amount	
	Top/bottom coding	Proposed method	Top/bottom coding	Proposed method
Maximum	-99.6%	-83.7%	-98.4%	-53.0%
Mean	-24.0%	-12.9%	-36.1%	-11.7%
Standard deviation	-88.7%	-68.3%	-79.7%	-34.6%
Median	-0.03%	-0.04%	-	-

CONCLUSIONS

And future lines of research

- A novel **secondary anonymisation** approach has been designed and implemented to protect the CIR dataset as a result of a close **collaboration** between BELab and the data provider (CIR Department of Banco de España).
- This procedure has a number of **benefits** over alternative procedures:
  - It allows to jointly analyse and protect **categorical and continuous variables**. Information on loans is incorporated into the debtors data, yielding a very complete **profile** for each debtor and allowing to use existing **SDC software**.
  - The joint anonymisation of categorical and numerical variables **minimizes disclosure risk and information loss** since only sensitive samples are affected. Only **0.95%** of suppressions (**0.01%** of numerical values).
  - The full **time-series** dataset is protected as a whole.
  - Once the described procedure has been designed, its **implementation** and use is relatively simple and its **computational cost** is low, in comparison with alternative methods, such as micro aggregation.
- Note that the full process needs to be repeated every time **new yearly data** becomes available. Researches cannot have access to more than one version of the dataset simultaneously, to avoid the possibility of cancelling local suppressions.
- **Future research lines**:
  - Address the possibility of **modelling uncertainty** on the information available to intruders.
  - Analyse links between **microdata protection and anomaly detection**, since both topics are closely related.

BANCO DE ESPAÑA
Eurosistema

Thank you for your attention!

STATISTICS DEPARTMENT