Statistical Regulation, Planning, Standardization and Normalization Division
(DIRPEN)

PROTOCOL FOR MICRODATA ANONYMIZATION

February 2012

NATIONAL ADMINISTRATIVE DEPARTMENT OF STATISTICS

JORGE BUSTAMANTE R.
Director

CHRISTIAN JARAMILLO HERRERA
Deputy Director

MARIO CHAMIE MAZILLO
Secretary General

TECHNICAL DIRECTORS

NELCY ARAQUE GARCÍA
Statistical Regulation, Planning, Standardization and Normalization

EDUARDO EFRAÍN FREIRE DELGADO
Statistical Methodology and Production

LILIANA ACEVEDO ARENAS
Census and Demography

MIGUEL ÁNGEL CÁRDENAS CONTRERAS
Geostatistics

ANA VICTORIA VEGA ACEVEDO
Synthesis and National Accounts

CAROLINA GUTIÉRREZ HERNÁNDEZ
Dissemination, Marketing and Statistical Culture

Bogotá, D. C., February 2012

Authors: Fredy Rodríguez Galvis, Mauricio Ricaurte.

Proofreading in Spanish: Sonia Marcela Naranjo Morales.

English translation: Juliana Mosquera Dueñas.

Proofreading in English: Ximena Díaz.

With the support of: DANE IT Department

CONTENTS

PRESENTATION

The National Administrative Department of Statistics, DANE, as the coordinator entity of The National Statistical System (NSS), promotes the improvement of the quality and credibility of the statistics that are required by the general public, the government and the international community, and that correspond to the economic, demographic, social and environmental domains. These statistics also enable the design, formulation, monitoring and assessment of public plans, programs and policies.

Therefore, the entity designed and developed Protocols for the Statistical Activity in the framework of the *National Code of Good Practice for Official Statistics*. The protocols aim to facilitate the implementation of the practices outlined in the Code, and thus enhance the quality of processes and production of official statistics generated by the entities belonging to NSS. This material helps to strengthen the technical capacity of the National Statistical System through guidelines and proven practices that are easy to interpret and implement in the statistics-producing entities.

The protocols consider different international benchmarks as a reference, such as the UN Fundamental Principles, quality dimensions of the Organization for Economic Cooperation and Development (OECD), the UK Statistics Authority and Statistics New Zealand, as well as the major international benchmarks with high standards for the production of statistics, and that show specialized advances in the statistics field.

INTRODUCTION

The information available to the public requires the inclusion of a value added, thus enabling its use in the development of social and economic plans, or for research of phenomena in different fields. Thus, both individuals and companies require disaggregated data to facilitate the generation of new information for particular uses. However, the provisions of Law 79 of 1993 have to be observed. The act states that "…data provided to the National Administrative Department of Statistics (DANE), in the development of censuses and surveys, may not be disclosed to the public, nor to official entities, agencies, or public authorities, except in numerical summaries"[1].

Pursuant to the above, the delivery of disaggregated data to users is possible. However, there are problems in defining mechanisms to prevent the identification of those who provide information. This is mainly due to the number of variables established in the statistical operations that when crossed can facilitate the identification of individuals or businesses. Such information can be used for inappropriate purposes, causing confidence of those who provide information to DANE to drop and significantly affecting the institution's credibility. The question then arises: ¡What should DANE do to deliver data and complete information, and at the same time, protect data and the information itself against the aforementioned problems?

The answer lies in the rigorous application of the existing legislation on protection of confidentiality, and that is set out in the Political Constitution and laws, such as the one corresponding to statistical confidentiality and the habeas data. At this point, it should be remembered that it is necessary to work on the creation a statistical law that analyzes these issues in depth and that ensures to a full extent, protection of confidentiality and exemplary sanctions for violating the fundamental right to privacy and the habeas data.

The law does not establish mechanisms or instruments for the protection of information; hence DANE as the governing body should promote the protection of individual privacy of persons and companies that report information to DANE. In view of this situation, the entity is required to generate a protocol for microdata anonymization, as well as to develop and apply techniques based on international standards and good practices.

This document develops the protocol that establishes general guidelines for the implementation of microdata anonymization processes in statistical operations produced by the entity. Reference is made to some techniques that enable reducing the risk of identification of data sources.

[1] Congress of Colombia. Law 79 of 1993, which regulates the development of Housing and Population Censuses in all the national territory. October, 1993.

The document initially presents the basic definitions, so as to understand the microdata anonymization process. It then presents the definition of the process and the participants involved and, finally, explains some anonymization techniques.

1. DEFINITIONS

Microdata anonymization: This refers to the process that prevents the identification of the units under study that are sources for individual records of the set of microdata[2].

Archive: Set of records with statistical data, where one or more variables are involved (categorical or numerical) and that presents information on individuals, businesses considered as observation units[3].

Sensitivity criteria: Rules applied for the detection of sensitive (risky) cells in frequency and / or magnitude tables. These rules can be based on the number of contributions to the cell (threshold value rule) or the value of the dominant contributions of the cell (dominance rules). The former detects cells with small frequencies and the latter detects cells with dominant contributions to the value of the cell. In both cases these cells are potentially "dangerous" in relation to their dissemination as they may contain, or result in the disclosure of, sensitive information[4].

Data of a personal nature: This refers to any numerical, alphabetical, graphic, photographic, acoustic, or any other type of information that can be collected, registered, processed and transmitted. This information pertains to identified or identifiable physical persons (such as name, last name, marital status, sex, age, address, social security number, employee registration number, personal identification, phone number, etc.).[5]

Statistical data: Refers to any numerical, alphabetical, graphic, photographic, acoustic, or any other type of information, pertaining to statistical units (natural or legal persons, public entities or bodies, etc.) that is collected for statistical purposes and, therefore, subject to the rules governing statistical confidentiality, such as names, addresses and identity numbers[6].

[2] DANE. Decree 1503 of 2011. The statistical confidentiality committee is formed. 2011

[3] This definition applies only to this document, so it should not be applied to other publications of the entity.

[4] Eustat. *Tratamiento de la confidencialidad en las operaciones estadísticas de Eustat* (Treatment of confidentiality in Eustat's statistical operations). 2007.

[5] Ibid

[6] Ibid

Indirect identifications: This refers to the characteristics that can be shared by various respondents where the combination of these could lead to the re-identification of one of the respondents. For example, the combination of variables such as place of residence, age, sex and profession could be identified if only one person of that particular sex, age and profession lived in that particular place[7].

Sensitive information: This refers to information regarded as strictly confidential. The information and characteristics related to age, origin, health, race, religion, ideology, membership, finances, etc., are considered sensitive in nature and require special protection[8].

Microdata: Data related to the characteristics of population units (individuals, households, establishments, among others.) which constitute a unit of information in a database, and are collected by means of a statistical operation[9].

Statistical operation: Set of processes and activities starting from the systematic collection of data to the production of aggregate results[10].

Microdata Pre-anonymization: The process by which criteria for including certain identification variables in the design of data collection instruments are established, considering the needs of the users.

Statistical confidentiality: Data that have been provided to DANE, in the development of censuses and surveys, may not be disclosed to the public or official bodies, but in numerical summaries, wherefrom no type of individual information that could be used for other purposes can be deduced[11].

Pseudo-anonymization: The process that is used to hide identities. The purpose of using pseudonyms is that of collecting more data on the same person without the need to know his/her identity. Its use is particularly relevant in the statistical and research environments[12].

---

[7] Ibid

[8] Ibid

[9] National Administrative Department of Statistics. Decree 1503 of 2011. The statistical confidentiality committee is formed. 2011

[10] Ibid

[11] Congress of the Republic of Colombia. Law 79 of 1993, by means of which the development of the Housing and Population Censuses in all the national territory. October, 1993 was regulated.

[12] Opinion 4/2007 on the concept of personal data, Article 29 Working party, European Council.

Sensitive Variable: A numerical or categorical variable that contains sensitive information[13].

[13] Eustat. Tratamiento de la confidencialidad en las operaciones estadísticas de Eustat (Treatment of confidentiality in Eustat statistics operations). 2007.

2. RELATED PRINCIPLES

Microdata anonymization is an international practice to ensure access to information and the manner in which different agencies establish instruments with general guidelines to manage microdata. Concerning the above, The United Nations Statistics Division established principle number 6 which states: “Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes”[14].

Similarly, the National Code of Good Practice for Official Statistics refers to the anonymization of microdata in the following principle and two of its corresponding good practices:

Principle 5: *Confidentiality*: This principle states the importance of confidentiality as an essential factor for improving the credibility mentioned in the good practices.

5.3*. It must be ensured that the publication of official statistics does not allow the individual identification of respondents.*

5.6. *The access to anonymized microdata on behalf of external users must be subject to protocols that guarantee confidentiality.*

[14] United Nations. Fundamental Principles of Official Statistics. 1994.

3. MICRODATA ANONYMIZATION

The anonymization process is aimed at controlling the risk of identification to which natural or legal persons who provide information that is used for statistical purposes may be subject to. It is understood that the possibility of extracting data from an aggregation enables identification of the source, so that the implementation of a suitable anonymization process avoids the probable misusage of disaggregated information (microdata).

When the microdata anonymization process is being carried out, it should always be borne in mind that the ultimate goal of the information is its usefulness to users. Therefore, the use of data must be preserved, trying to add the least possible noise in the results and protecting the information sources' privacy.

Archives, macrodata or statistical information, graphs, documents and publications that are derived from any statistical operation and disseminated through the communications channels and means established by the entity, constitute the dissemination of any statistical operation. Under no circumstances should these enable direct individual identification, as provided in Law 79, Article 5 on the Statistical Confidentiality.

The measures and criteria for data protection in the phase of statistical data dissemination are designed to avoid direct or indirect identification of individuals or entities. This can occur when publishing very-detailed analyses or disaggregations that may lead to the disclosure of sensitive or confidential information on them. Guidelines depend largely on the format in which information is disseminated and on the general or specific character thereof. Considering the above, the steps in the anonymization process as well as the specific activities in each case are presented below.

4. ANONYMIZATION PROCESS

This process has three stages that develop the procedure to prevent the identification of data sources. These stages are:

STAGE 1. Pre-anonymization

STAGE 2. Anonymization of microdata for internal use.

STAGE 3. Anonymization.

Each stage of the process shall be documented by means of a confidentiality record of the statistical operation. Below are the activities that should be followed in each stage.

STAGE 1. Pre-anonymization

Pre-anonymization should be carried out when the design of the statistical operation is being conceived. Pre-anonymization is understood as a preliminary step for the clear determination of variables, and of direct identifiers and other confidential data to be obtained in the development of the statistical operation.

The following should be considered when implementing this stage:

1. The thematic team responsible for the operation must classify each of the variables in the archive or database to perform the anonymization process considering these categories:

   i. Geographic identification variables.

   ii. Variables of direct identification of individuals or businesses.

   iii. Variables with magnitudes or numeric values.

   iv. Variables of a sensitive or confidential nature.

   v. Unrestricted variables for public access.

   vi. Categorical variables.

2. Once the categorization of the variables is defined, the mechanisms for the protection of the privacy of information sources should be determined, defining the amount of personal data necessary for the development of the statistical operation and determining the indirect variables that may lead to the identification of the source. The premise, in these cases, should be to minimize the amount of personal

information that should be used. This minimization has a direct relationship with the method used for information collection and the manner in which it should be recorded.

Data containing sensitive information requires special attention. This will require generating a procedure to define the need and the amount of sensitive data to be obtained by the different statistical operations. Likewise, the procedure should also define the way to proceed in case sensitive information has to be omitted from the database once it is collected. If possible, the conditions that the statistical units and sources should have for an effective control of their personal information (which is considered sensitive) should be specified. Statistical units and sources will be the ones that should give their free and informed consent to the statistical entity so that the latter can access their information.

STAGE 2. Anonymization of microdata for internal use

This stage should be completed during the Statistical Production and is characterized by the implementation of activities to protect the privacy of data once they become part of the entity's databases and have been collected by the officials in charge. The aim is to transfer the necessary variables to perform the corresponding analysis of the operation to the subject-matter experts who are directly involved in the implementation and analysis of the statistical operation. This is done without involving risks to the statistical unit and following the assurance of data confidentiality.

In this regard the anonymization of microdata for internal use consists in the deletion, suppression or concealment, under pseudonyms, of direct identifiers that are associated with an individual or company, and is an intermediate step in the process of anonymization that eliminates the possibility of a re-identification by using mathematical algorithm techniques.

In the phase of anonymization of microdata for internal use once the direct identifiers that have been eliminated are restored by crossing the pseudonymized database with another one containing the original variables, the risk of re-identification increases. Therefore, this is considered as an intermediate step. A fundamental difference between this step and that of anonymization itself, is that the former assumes that personal identifiers have been extracted from a database, where is not possible to identify the source unless the original archive is obtained. The source can be fully identified by crosschecking the information. On the other hand, the anonymization procedure assumes that all connections between the registers of an individual and the source have been irreversibly broken and, therefore, the identification of the source is less likely due to the algorithmic techniques.

The activities to be carried out are the following:

1. Criteria to prevent leakage of collected information should be established as well as procedures for the concealment, suppression and pseudonymization of personal data by the IT systems department in the statistical production process, so that the subject-matter experts who are involved in the statistical operation cannot easily infer sensitive and confidential data from the sources, unless these are strictly necessary for the analysis of the operation. This procedure is useful in cases where information is lost due to theft or mishandling by the statistical institute employees. If the thematic team or the custodian of the information routinely work with information in which direct identification variables have been suppressed, hidden or replaced by pseudonyms by the IT systems department, there would be less risk of data violation[15].

   OECD's international standards such as the Guidelines on the Protection of privacy and transborder Flows of Personal Data contain principles such as that of use limitation that states that personal information should not be disclosed for unidentified purposes, except for the cases in which there is consent of the data subject or by the authority of law. In such conditions, an effective procedure for the anonymization of microdata for internal use will help to comply with this principle even though the action generating the risk is not malicious.

The variables of direct identification of a source in the list below are likely to be eliminated or masked. In any case the thematic team is responsible for defining which are the variables to be subject to this process:

1. Names.
2. All elements pertaining to date of birth, date of incorporation in the chamber of commerce, except for the year.
3. Telephone and fax numbers.

[15] An example of the latter case is described in the text published by the Office of the Information and Privacy Commissioner of Ontario, Canada. "Dispelling the Myths Surrounding De-identification". The document notes that in cases where confidential information resting on a USB flash drive or any kind of hardware equipment and which has previously undergone processing for the elimination of direct identification variables, is stolen or lost, it is unlikely that the individual (external to the entity) in possession of information has the technical capacity and the reasons to make it identifiable. However, if the information has not been processed and the direct identifiers are recorded in the database, it is feasible for the information to be used for malicious purposes. Many of these accidents regarding information leaks occur within statistical offices. This can be avoided if most of the personal data included in the databases are kept in a non-identifiable manner.

4. Identification Numbers: Colombian citizenship card, passport, Colombian identity card for those under 18 years old, number of affiliates in the social security system, driver's licenses, NIT (Tax Identification Number), RUT (Single Tax Registry), RUP (Single Proponents Registry), RUE (Single Business Registry), etc.
5. Email addresses.
6. Bank account numbers.
7. Vehicle identifiers, license plate, etc.
8. Mobile device identifiers and serial numbers.
9. IP addresses.
10. Biometric identifiers.
11. Photographs and similar images.
12. Any other single identification number.

Another activity aimed at decreasing the risk of identification consists in assigning pseudonyms to variables of direct identification of individuals: Names, ID cards, etc.; by establishing a blind identifier called pseudonym, thus obtaining a pseudonymized database. The identification of the individual is known only by the IT systems department or the subject-matter experts of the statistical operation, while the other members of the organization cannot establish the relationship between the blind and the original identifier, under certain conditions of cryptography. This activity, as indicated by Galindo[16], does not resolve the risks associated with the database indirect re-identification, it simply reduces the level of risk of direct identification within the statistical institute.

The following points can be considered to perform the pseudoanonymization technique:

a. To assign a single pseudonym to each object of the personal identifiable information.
b. The pseudonym should be used in place of formal identification numbers, such as citizenship identity cards, driving licenses, etc. It is recommended for pseudonyms to have the same length and format, to increase readability.
c. To consider the impact of information systems in the allocation of pseudonyms in relation to internal uses.

[16] Galindo's text is cited in the bibliography, and was written in the setting of a meeting at the European Union on issues related to anonymization.

d. If pseudonyms are used for external use, these must be different from the ones generated for internal use, and cannot have a relationship with each other.

e. The IT systems team should establish the cryptographic techniques to accomplish the incorporation of pseudonyms to replace direct identification variables.

As noted by Galindo, in his document *Microdata sharing via pseudonymization*, the pseudonymisation technique enables sharing databases with other working groups assigned to the organization responsible for managing data, as well as with other State entities requiring information for statistical purposes, who will have access to them without being able to derive direct identification of the source. The latter does not imply that this step is sufficient to elude the duty of statistical confidentiality and of statistical reserve of DANE's and other state entities' officials, since as previously defined, the anonymization of microdata for internal use is not the definitive step in the anonymization process.

The step of anonymization of microdata for internal use enables collected information to have a security and confidentiality criterion when handled exclusively by the subject-matter experts of the operation, who have access to the database without direct identifiers or these being replaced by pseudonyms. Therefore, the subject-matter experts are responsible for guarding the information.

In summary, the step of anonymization of microdata for internal use seeks to reduce the risks of identification. Among the activities described for this purpose the following are highlighted:

a. To eliminate the identifiers associated with a person or company.

b. To use ranges for identifiers, for example, range values in exchange for the specific age of the individual.

c. Using pseudoidentifiers (pseudonymisation).

STAGE 3. Microdata anonymization

This is the final step in the anonymization process, which aims to provide disaggregated data for the general public, and as such, involves the use of the techniques described below. This stage involves the elimination of the maximum risk of identification of the source, causing the least damage to data utility, and consists of the following activities:

1. The thematic team should establish the most appropriate anonymization technique that has to be applied to each of the variables subject to the anonymization process; In adddition the team should also include cases related to the application of the technique for each variable.

2. Subsequently, the anonymization proposal is sent to the team of experts of the Statistical Confidentiality Assurance Committee.

3. The IT Systems technical team that is responsible for the statistical operation implements anonymization algorithms proposed by the thematic team in charge, runs the anonymization process, performs tests and delivers the result to the thematic team for its final approval.

The following are some additional aspects that must be considered during the anonymization process:

- Disseminated Microdata archives will not include in any case direct and indirect identifiers of records, or data of personal nature.

- The publication of cells with numerical magnitudes or values wherefrom the contribution or contributions from any of the statistical units (individuals or companies) that add value to the cell can be easily derived, should be avoided. This occurs when there are few contributors, or when there are contributions which are dominant or above the average of the cell (sensitivity criteria technique).

- To prevent the occurrence of such cells, sensitivity criteria and variable recoding techniques can be applied and/or +suppression of cells methods that can properly protect the archive and preserve as much information as possible.

- The detail of other variables included in the microdata archive will depend on the geographical level provided and on the sensitivity of the variable itself; A wider disseminated geographical scope and a lower sensitivity degree of the variable enable a greater conceptual disaggregation.

- In order to provide greater protection, microdata reduction or disturbance techniques can be applied (the latter is recommended), modifying quantitative variables in small random quantities and / or exchanging attributes in a controlled manner between records of proximate geographic areas, respecting in all cases, distributions (means, totals, etc.) by historical territory.

- Anonymized archives are validated and approved by the experts that make up the Statistical Confidentiality Assurance Committee before their publication.

5. PARTICIPANTS OF THE MICRODATA ANONYMIZATION PROCESS

Statistical confidentiality involves all of DANE's staff and the natural or legal persons who have knowledge of individualized statistical information who are obliged not to disclose, directly or indirectly, individual or individualized data from information sources. This implies the prohibition of using the data obtained directly from the informants for non-statistical purposes.

All the personnel performing data collection, as well as any other work in DANE that is related to the statistical process, shall sign a confidentiality clause.

Resolution 1503 of 2011 regulates statistical dissemination and the formation of the assurance committee. According to this resolution the following participants should be considered for the carrying out of the process:

Statistical Confidentiality Assurance Committee

This expert committee is responsible for the validation and approval of the proposed techniques on anonymization by the thematic team responsible for the statistical operation in its planning stage, this in compliance with Resolution 1503 Article 1 numeral 3 and 6[17].

Similarly in pursuance of numeral 2 at the end of the anonymization process, the committee is responsible for validating and approving anonymized archives before their publication.

Thematic team

It is responsible for the anonymization proposal considering the activities of the process described above. It is worth noting that the thematic team should consider the process of anonymization of microdata from the statistical planning.

The thematic team is responsible for the statistical operation and for the anonymization techniques that will be applied to each of the variables that need to be anonymized.

IT systems team

This team supports the statistical operation and is in charge of developing and implementing the necessary algorithms to anonymize the microdata archives, according to

[17] National Administrative Department of Statistics. Decree 1503 of 2011. By which the Statistical Confidentiality committee is formed. 2011

the classification of variables and the definition of anonymization techniques for each variable. The IT Systems team returns anonymized microdata archives to the thematic team, for their corresponding validation.

6. MICRODATA ANONYMIZATION TECHNIQUES

There are several techniques that enable the anonymization of microdata. The main factor when selecting a technique is the international acceptance that this has; this refers to the relevance of the international benchmark regarding the themes and proven successful experiences when implementing the techniques in other national statistical institutes.

The methods refer to the limitation in statistical disclosure and can be classified into two categories: Data perturbation methods and data reduction methods.

Methods based on data perturbation

These refer to processes involving systematic data modification (sometimes in small random quantities), so that the figures are not precise enough to reveal information about individual cases. Including new data, deleting and / or modifying the existing ones can benefit statistical confidentiality.

The main data perturbation techniques are:

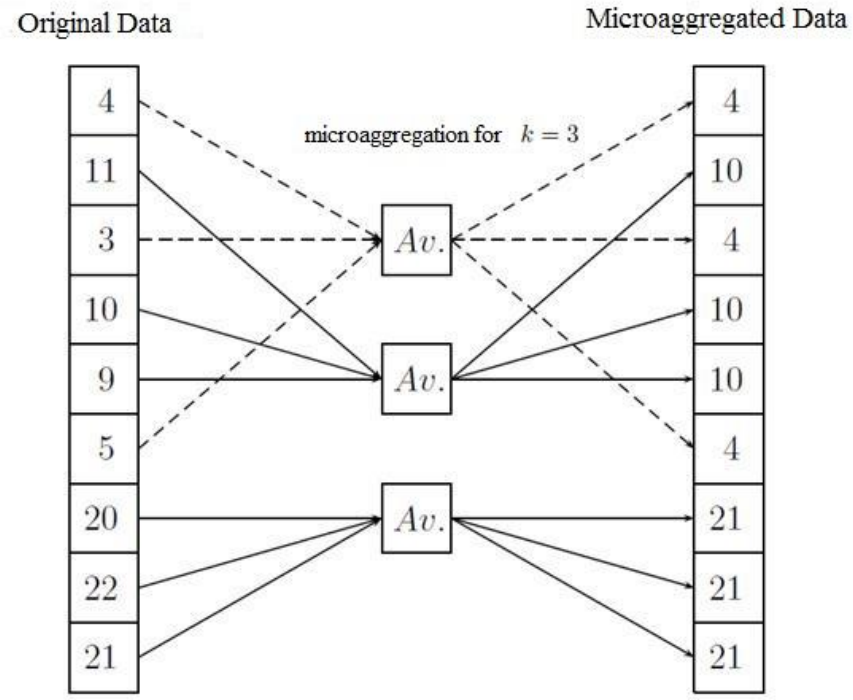
- Microaggregation[18]: A perturbation technique proposed by Eurostat as a form to disclose statistics for numeric variables. The idea is to replace an observed value with the average calculated over a small group of units (small aggregate or micro-aggregate) including the one under study. It consists in grouping individual records in small groups before their publication, maintaining the results when applying statistical operations. By setting a parameter k, the microaggregation of a numerical set of data would be:

- Grouping: The records contained in the original set are grouped into subsets of cardinality at least k by some criterion of similarity (e.g. Euclidean distance). The result of this process is a k-partition (a k-partition is a partition in which each one of the parts has at least k elements).

    - Substitution: each record of the original set is replaced by the middle record of the subset to which it was allocated in the previous stage[19].

[18] Agusti Solanas, Antoni Martínez-Ballesté, Josep Domingo-Ferrer, Josep M. Susana Bujalance and Mateo-Sanz, Microaggregation methods for k-anonymity: privacy in databases. Dept. Computer Science and Mathematics, University of Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain.

[19] U. Gonzalez-Nicolas and A. Solanas. Privacy Protection through Multivariate microaggregation based on Genetic Algorithms: Roulette selection vs. Uniform selection.



Source: U. Gonzalez-Nicolas and A. Solanas Privacy Protection by Multivariate Microaggregation based on Genetic Algorithms: Roulette selection vs. Uniform selection.

Units of the same group will be represented by the same value in the published archive. Groups contain a predefined minimum number k of units. K minimum accepted value is 3. For a given k, the problem consists in determining the partition of the set of units in groups of at least k units (k-partition), thus minimizing the loss of information, which is usually expressed as a loss of variability. Therefore, the groups are constructed according to a criterion of maximum similarity between units. Data protection is achieved through the micro-aggregation mechanism, ensuring that data are not in units of at least k with the same value in the data archive.

There are several methods that consist in modifying values (from vectors of continuous variables) according to different criteria such as: The Maximum Distance method (MD), the Maximum Distance to Average Vector method (MDAV) and the maximum distance to the variable size average vector or Variable size-MDAV (V-MDAV).

The MDAV method is the most common tool for microdata anonymization, but it has been demonstrated that in terms of efficiency the V-MDAV method has produced better results.

Other microdata perturbation[20] techniques are presented below:

- Post Randomization Method - PRAM[21]: A method for controlling statistical dissemination that can be applied to categorical data. It is a perturbation and probabilistic method for protecting microdata archives[22].

  Compared to other methods such as global recoding, local suppression and top and bottom coding that can lead to a high loss of information for safe data archives, the PRAM method is a better alternative as it maintains the level of detail while the level of dissemination control is carried out through the introduction of uncertainty in the results of the identification variables.

  The PRAM method can be used to produce microdata archives with the same structure as the original microdata archive, but with some kind of synthetic data. It can also produce safe data archives while leaving some of the characteristics of the archive more or less unchanged.

  PRAM is a method that is defined in terms of transition probabilities; it is summarized in a PRAM matrix. It produces microdata archives in which the values of some categorical variables are changed for certain records in relation with the values of the original microdata archive. It is usually applied to identification variables, i.e. variables that can be used to identify the respondent. The result is the obtention of microdata archives with incorrect values in the identification variables, which reduces the risk of identification to the minimum.

  The PRAM method can be considered as a form of classification error.

- Synthetic data. Data are randomly generated, preserving some statistics or internal relations of the original dataset.

  - Data Distortion by probability distribution: This method can be used both in categorical and continuous variables. It aims to obtain a randomly protected data set from the original data set.

[20] Hybrid Microaggregation, historical vision on statistical disclosure control. An overview of public statistics. Institute of Statistics of Andalusia, Spain.

[21] Elsa Cristina Pinto Mendes. Confidentiality of Data: Application and Comparison of Techniques of Statistical Disclosure Control. in 2010

[22] Ibid

The distortion procedure is carried out in three stages:

- To identify the underlying density function for each confidential variable in the data set and estimate the parameters associated with the density function.

- To generate a series which was randomly obtained from the density function for each confidential variable.

- Mapping: This refers to the classification of the altered and the original series in the same order and substituting each element of the original series with the corresponding element of the altered series. The mapping and the substitution processes are necessary only if the altered variables were to be used jointly with other unaltered variables.

- Hybrid Microdata approach: Consists in the calculation of masked data as a combination of original data and synthetic data. This combination enables a better control of totally synthetic data over the individual characteristics of masked data. Hybrid masking involves combining original data with synthetic data.

- Record exchange or permutation[23]: This is a method of disclosure control applied to microdata. which consists in exchanging the values of some variables contained in records paired by means of a representative key variable. This method is sometimes called "multidimensional transformation." It is a transformation technique that guarantees (under certain conditions) the maintenance of a set of statistics, such as averages, variances and univariate distributions.

- Rounding: This can be based on deterministic or random techniques (depending on whether rounding is applied to only one, or to several variables). It consists of substituting the value of the original variables by rounded values.

- Adjusting weights: If the type of sampling applied to the original set of data is known such sampling could be done backwards, so that reidentification can be carried out based on the weights. These methods modify such weights so that sampling cannot be easily taken backwards.

- Adding Noise: Consists in the addition of random noise, which has the same correlation structure of the original data. This technique involves the generation of random values that can be added to those reported by the respondent. This can be done in several ways, depending on whether it is applicable to individual or multiple variables, or noise is added without altering means, variances and co-variances. Furthermore, linear programming techniques can minimize the differences between the actual values and the altered ones.

[23] Taken from: http://www.eustat.es/documentos/datos/Documento_web-confidencialidad_c.pdf It refers to Rank Swapping.

- Re-sampling: This method was originally proposed to protect tabular data, but can also be used to protect microdata.

  Let V be an original variable in a data set with n records t independent samples X1, ..., Xt. All samples are sorted using the same classification criteria, then a masked variable is created as x1, ..., xn, where: n is the record number, xj is the mean of the j-th value classified in X1, ..., Xt.

  Assuming that microdata z1, ..., zn are aggregated to create macro data in a contingency table X, with I rows and J columns, and with certain specifications, xij is the original frequency of the i-th row and the j-th column. To create an anonymized table X ', the sample z'1, ..., z'n is obtained from the original data z1, ..., zn n times and with substitution. Thus, table X 'is an estimate of the original table X, not allowing the obtention of any precise information of X[24].

Methods based on data reduction[25]

There are methods based on data reduction where data are not altered when applying these techniques, but that instead produce partial suppressions or reductions in the level of detail of the original set. These procedures tend to avoid the presence of single or atypical recognizable individuals.

The main reduction techniques are:

- Variable elimination: The first application of this method is the elimination of direct identifiers from the data archive. A variable must be eliminated when it is highly identifiable and another method of protection cannot be applied. A variable can also be eliminated when it is considered too sensitive for public use, or irrelevant for analytical purposes. For example, information on race, religion, HIV, etc. not be entered in a public use archive, while it could be delivered in a license archive.

- Record Elimination: It may be taken as an extreme measure of data protection when the unit is identifiable despite applying other protection techniques. For example, in a data set of manufacturing and business surveys, a company may belong alone to a specific sector. In this case, it is preferable to eliminate this particular record, instead

[24] Elsa Cristina Pinto Mendes. Data Confidentiality Re-Sampling: Application and Comparison of Techniques of Statistical Disclosure Control. Page 38. in 2010

[25] Methods taken from: Alpa K. Shah. State-of-art in Statistical Anonymization Techniques for Privacy Preserving Data Mining. Vol. 3 No. 7 July 2012. International Journal of Computer Science & Engineering Technology (IJCSET). Retrieved on February 2, 2013 from: http://www.ijcset.com/docs/IJCSET12-03-07-020.pdf

of eliminating the variable "industry" of all records. Since it greatly affects the statistical properties of released data, the elimination of records should be avoided as much as possible. It is an appropriate procedure for categorical variables. In continuous variables, the risk of disclosure increases. When the records subject to elimination are selected according to a statistical planning by sampling this is denominated sampling.

- Global Recoding: Combines categories to form new less specific categories. In continuous variables, values are made discrete (from infinite to finite). The technique is applied to numerical variables, whether continuous or discrete and it affects all records in the data archive. Consider, for example, the variable “marital status” that is often observed in the following categories: single, married, separated, divorced, widow. The sampling frequency of the “separated” category could be low, especially when crossed with other variables. The two adjacent categories, separated and divorced can be joined into one “separated or divorced” category. The observed frequencies of the combinations of the participation in this new category would be higher than those for “Separated” and “Divorced” separately. The categories to be joined will be chosen considering the usefulness of data and the statistical control of the frequencies. The method can also be applied to key variables (such as geographic codes) in order to reduce their identification effects.

- Top and bottom Coding: This technique can refer to a special case of global recoding that can be applied to numeric or ordinal categorical variables. Variables such as “Salary” and “Age” are two typical examples. The highest values of these variables are generally very atypical and therefore identifiable. Top coding introduces new categories such as “monthly salary above 10 million pesos” or “above 75 years old”, leaving the observed values unchanged. The same reasoning applies to the lowest values observed and defines bottom coding. When working with ordinal categorical variables, an upper (or lower) category is defined by adding the “highest” (or “lowest”) categories.

- Cell Suppression: A method applied to tabular data, which comprises primary and complementary (secondary) suppression. Primary suppression consists in withholding all disclosive cells from publication which means not showing their values in the table but replacing them with a symbol (e.g., “missing” or “suppressed”, ...) to indicate the suppression. According to the definition of sensitivity criterion, disclosive cells would be those with a low value in frequency tables and cells with a low value in, or representing a case of dominance in quantitative variable tables should be primary suppressed. To achieve the desired degree of protection of disclosive cells, it is sometimes necessary to suppress additional cells that make it necessary to recalculate the value of the primary suppression; these are referred to as complementary (secondary) suppression.

  The criteria for selecting additional cells for suppression should be carefully chosen in order to ensure the desired level of protection and at the same time suppressing the least amount of information. For example, suppose that the combination “marital

status = widowed, Age = 17” is a single population. If the information on age is suppressed, the combination “marital status = widowed, Age = missing or deleted” will not be identifiable.

Implementing combinations of anonymization techniques by reduction is possible. For example, following a strict order, the global recoding technique can be applied and then the suppression of cells.

BIBLIOGRAPHY[26]

Alpa K. Shah.: State-of-art in Statistical Anonymization Techniques for Privacy Preserving Data Mining. Vol. 3, No. 7 July 2012. International Journal of Computer Science & Engineering Technology (IJCSET). Retrieved on February 2, 2013 from: http://www.ijcset.com/docs/IJCSET12-03-07-020.pdf

Basque Statistics Institute. Tratamiento de la confidencialidad en las operaciones estadísticas de EUSTAT. *(Treatment of confidentiality in EUSTAT's statistical operations).* Taken from:http://www.eustat.es/documentos/datos/Documento_web-confidencialidad_c.pdf

Congress of Colombia. Law 79 of 1993, which regulates the development of Housing and Population Censuses in all the national territory. October, 1993.

Duncan, George. Exploring the Tension Between Privacy and the Social Benefits of Governmental Databases. Paper presented at Security, Technology, and Privacy: Shaping a 21st Century Public Information Policy, 2003 April 24-25.

Elsa Cristina Pinto Mendes. Confidencialidad de Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística *(Confidentiality of Data: Application and Comparison of Techniques of Statistical Disclosure Control).* 2010.

Galindo David, Verheul Eric R. Microdata sharing via pseudonymization in: http://epp.eurostat.ec.europa.eu/portal/page/portal/conferences/documents/unece_es_work_session_statistical_data_conf/TOPIC%201-WP.03%20IP%20GALINDO.PDF

INE-National Statistics Institute of Chile. "Dimensiones de la calidad según OCDE y EUROSTAT" (Quality dimensions according to the OECD and Eurostat). November 2007.

National Administrative Department of Statistics. Resolution 1503 of 2011. by which Resolution No. 173 of April 2 is repealed, the Statistical Confidentiality Assurance Committee is formed and other regulations are established.

National Administrative Department of Statistics. National Code of Good Practice for Official Statistics.

OECD. "Quality Framework And Guidelines For OECD Statistical Activities", Version 2003.

The Office of the Information and Privacy Commissioner of Ontario, Canada. Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy. 2011.

[26] Translation of bibliographic titles and names is for reference purposes only.

United Nations. Fundamental Principles of Official Statistics.1994. http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp.

Ursula Gonzalez-Nicolas y Agusti Solanas. Protección de la Privacidad mediante Microagregación Multivariante basada en Algoritmos Genéticos: Selección por Ruleta vs. Selección Uniforme. *(Privacy Protection through Multivariate Microaggregation based on Genetic Algorithms: Roulette selection vs. Uniform Selection).* 2009.