UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Study and analysis for the elaboration and dissemination of microdata of sociodemographic information

Marta Mas Moreno, Ana María Miranda and Marina Ayestaran (Basque Statistics Office, Eustat)
m-mas@eustat.eus, ana_miranda@eustat.eus, m-ayestaranarregi@eustat.eus

Abstract

Microdata files are a particularly interesting product for the research community, since they offer greater flexibility when designing the different analyses and interpretation of results. At Eustat, we offer this information for some surveys and administrative record based statistics, generally in the area of population. All microdata files are protected prior to publication, that is, they do not include direct identifiers and have been treated to make the disclosure from indirect identifiers extremely difficult. This paper describes the process for obtaining a safe microdata file from survey data. The first step is to asses which records can be easily identified and the second step is to implement the necessary protection measures. The objective is to maintain a balance between the risk of identification and the usefulness of the information provided to the user. To illustrate this process, the analysis carried out for the Labour Force Survey in the Basque Country will be shown. The final product consists of a safe microdata file with its associated metadata.

KEY WORDS: Microdata, Confidentiality, Dissemination, Metadata

1 Introduction

Maintaining the privacy of the data providers, preserving the confidentiality of the information they provide and its use only for statistical purposes must be fully guaranteed within the statistical activity. This principle largely underpins the credibility of a statistical organization and must be present in all phases of statistical production.

The Basque Statistics Office (Eustat) is responsible for publishing and disseminating statistical data in accordance with the Basque Statistics Plan and annual statistical programs. Compliance with regulations regarding statistical secrecy and confidentiality of data from households, individuals, businesses, and administrations is essential.

The analysis of risk in the dissemination stage play a crucial role in maintaining confidentiality and protecting sensitive information. This article focuses on the risk analysis of microdata, prior to publication, mainly using the case study of the Population Survey in Relation to Activity (PRA) to illustrate their practical application.

2 Micro-Data Protection

2.1 Generation of ready-to-access microdata

The microdata are the individual data of respondents that are used to prepare tables of results. They are usually presented as tables in which each row ("record") stores the information of a unit and each column ("field") is a variable or a characteristic of the unit.

The microdata files that are available for public access in our website will be protected, that is, they will not include direct identification data and will be provided in such a way that the possible disclosure of data based on indirect identifiers is extremely difficult.

To protect a microdata file, the first phase consists of evaluating which records can be easily identified and the second phase consists of applying some protection measure. The assessment of the statistical disclosure risk (or statistical risk) of microdata sets is based on measuring in some way the occurrence of rare records.

A key combination is a selection of certain variable values that are considered identifiers for records because they are rare in some way. In short, a key combination allows to detect the rare records in the data set. Those records that have the values established in the key combination will be the records that can reveal confidential information because they are easily identifiable and therefore measures will have to be taken to protect the information.

There is no systematic procedure to establish the combinations of key variables; the person in charge of the statistical operation or the microdata file should set and test those combinations. However, for statistics belonging to the same area, the combinations of key variables analysed will be very similar, since these combinations often include characteristics common to all of them (for example: sex, age, geographical areas, in sociodemographic statistics or sector of activity or employment strata in economic statistics).

Once the combinations to be used have been determined, they are applied to the microdata and the frequencies of the combinations are observed in the file. If the frequencies are lower than a set limit, some protection measure must be applied.

Protection with information restriction methods is based on reducing the amount of information offered, either because it is directly suppressed, or because it is given with less detailed level.

The most common method is global recoding; this method consists of giving the information with a lower level of disaggregation: for example, at the province level instead of the municipal level, ages in five-year groups instead of year-by-year, economic activity at a digit instead of two etc. Global recoding applies to the entire file,

not just the records to be protected, and can be applied to both qualitative and quantitative variables. Disaggregation thresholds can be established for the variables common to different statistical operations within the same scope (e.g.: maximum geographical disaggregation, age intervals, economic classification, etc.).

We can also resort to protection with disturbance methods, this is based on altering the information offered, trying to maintain the global characteristics of the whole. There are several techniques that can be applied; the most used is the exchange of records (data swapping). The exchange of data between records consists of changing certain characteristics of the records to make them non-identifiable. Closeness criteria are normally established between the records to be exchanged so as not to alter the global characteristics of the microdata set, for example, exchanging records that are in the same municipality and have the same age, or have the same number of employees, or that they are in the same branch of activity etc.

If it is decided to apply disturbance methods, the user must be warned of the application of such methods for reasons of statistical secrecy, but no details will be given about the records affected neither about the parameters of the protection method.

In general, the microdata files that are disseminated will not present geographic identifiers that refer to areas with less than 10,000 inhabitants. This threshold is considered a suitable limit for the Basque geographical context. Therefore, geographic variables will be added to meet this criterion, this includes those referring to place of birth, place of residence, etc.

2.2 Example of application to PRA microdata

The growing demands from researchers, policy makers and others for more and more detailed statistical information leads to a conflict. The respondents are only willing to provide a statistical office with the required information if they can be certain that their data will be used with the greatest care, and in particular will not jeopardise their privacy. So statistical use of data by outside users should not lead to a compromise of confidentiality. However, making sure that microdata cannot be misused for disclosure purposes requires, generally speaking, that they should be less detailed, or modified in another way that hampers the disclosure risk.

This is in direct conflict with the wish of researchers to have as detailed data as possible. Much detail allows, not only more detailed statistical questions to be answered, but also more flexibility, that is, the user can lump together categories in a way that suits his purposes best. The field of statistical disclosure control in fact feeds on this trade-off: How should a microdata set be modified in such a way that another one is obtained with acceptable disclosure risk, and with minimum information loss? How exactly can one define disclosure risk? How should one quantify information loss? Once these problems have been solved - no matter how provisionary- the question is how all this wisdom can actually be applied in case of real microdata. If a certain degree of sophistication is reached, the conclusion is inescapable: specialised software is needed to cope with this problem and μ−Argus is such software.

Producing safe micro data is not a trivial issue. It should first be explained when microdata are considered safe or unsafe. It should also be explained how unsafe data can be modified to become safe.

All the microdata that we publish in the Eustat have been and continue to be subject to review in order to provide the maximum information with the minimum risk. The first survey we analysed was PRA - Population in Relation to Activity. (Labour Force Survey).

The Population Survey in Relation to Activity operation is a continuous source of information on the characteristics and dynamics of the workforce of the A.C. from Euskadi. It includes the relationship with the productive activity of the population residing in family households, as well as the changes produced in their employment situation; prepares indicators of quarterly variations on the evolution of the active population; it also estimates the degree of participation of the population in activities that are not economically productive. It offers information at the level of historical territories and capitals.

We start from the dataset with the PRA microdata of the last available quarter to determine the risk of the same we have used the µ-Argus program. The objective of Argus is to hinder the re-identification of individuals represented in the data to be published, that is, to prevent the disclosure of confidential data (disclosure). When a file is considered unsafe, anonymization techniques (SDC) will be applied, which will produce modifications in the data, so that an adequate level of security is reached, that is, adequate depending on the use that is going to be made of them: public or scientific.

What concepts must we handle to understand µ-Argus?

1. Key variable: variables that allow the informants to be identified. Important note: they must be defined as qualitative variables (Categorical).
2. Combination: crossing of variables that forms a table.
3. Dimension (dimension): number of variables that cross in a table.
4. Threshold: the limit at which a frequency or risk, for a combination, is considered safe or unsafe: values below the threshold will be unsafe, above safe.

First, we are going to carry out an individual risk analysis. The key variables that can generate potentially identifying combinations in the file are the following:

- TERH – Province of residence
- SEXO – Sex
- LNAC – Place of birth
- EDAD – Age
- NACI – Nationality

The weight variable that we use to calculate the risk is ELEV2.

Considering those variables directly without any recoding what we get in µ-Argus is:

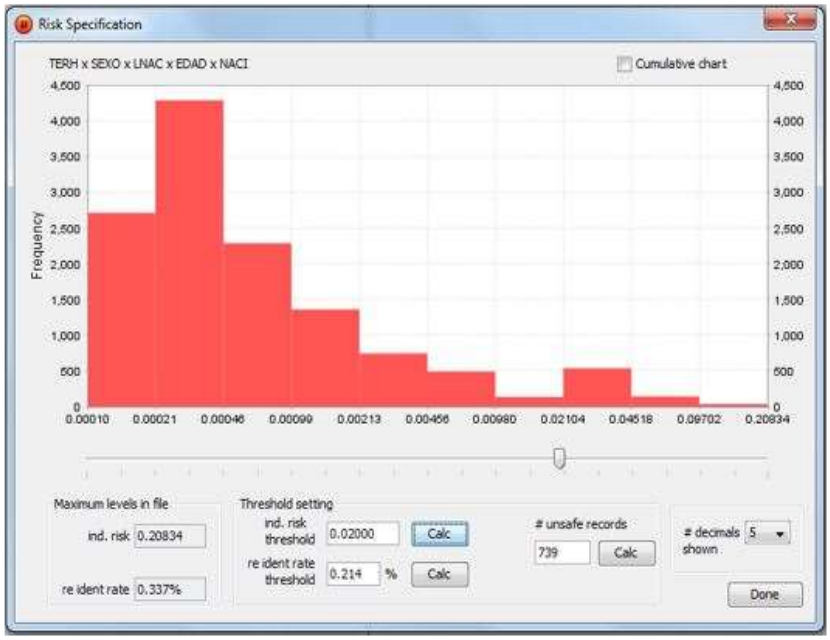


Fig. 1. µ-Argus risk chart without global recoding.

The dataset has 12,749 records, the re-identification ratio in the file is low (0.337%) but there is still a number of unsafe records (739) (considering as unsafe a risk greater than 0.02%). The individual risk threshold is set by the data protector and it depends on the type of data and its subsequent use. In general, it is desirable that it be small.

We review the frequencies and see that age is the variable with the greater number of unsafe records in all dimensions:

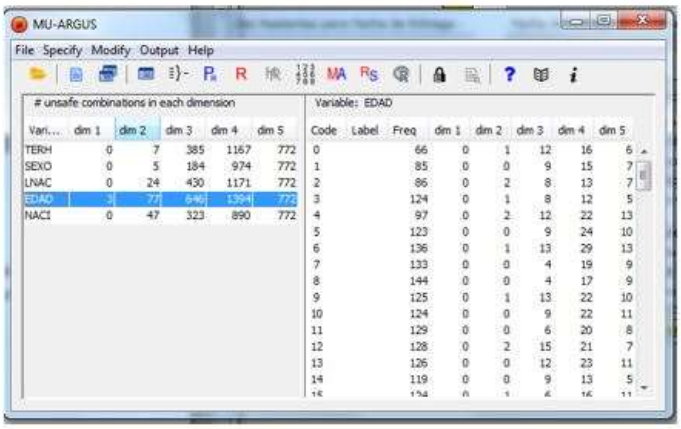


Fig. 2. Mu-Argus unsafe combinations (EDAD).

We decide to group the age into five-year groups. We recalculate the individual risk of each record after recoding and we have:

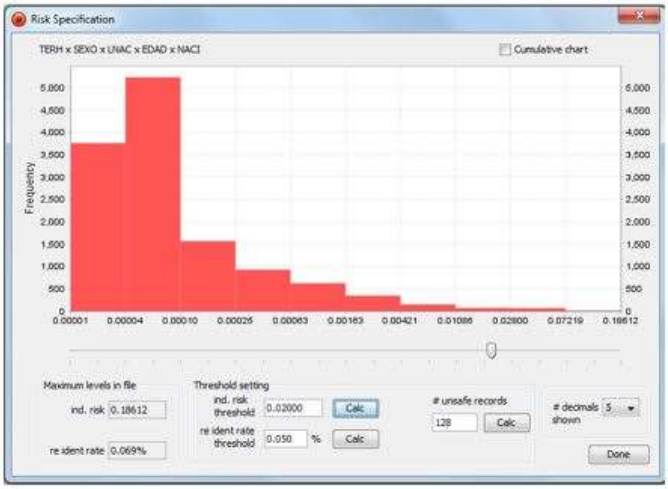


Fig. 3. Mu-Argus risk chart with global recoding (EDAD- five groups).

The risk of re-identification for all records has dropped considerably and the same happens with the re-identification ratio (0.069%) and the number of unsafe records (128). Given this scenario, we would assume that 128 of the records could be identifiable with reasonable effort.

The next variable with the most unsafe records is NACI (nationality), we are going to group this into two categories: Spanish and Foreign.

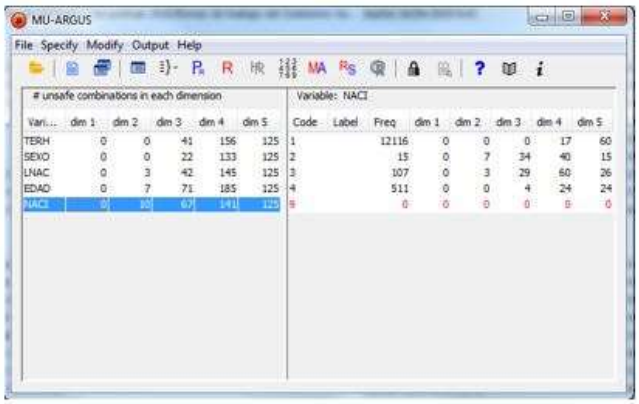


Fig. 4. µ-Argus unsafe combinations (NACI).

Once recoded, we calculate the individual risk of re-identification again and we have:

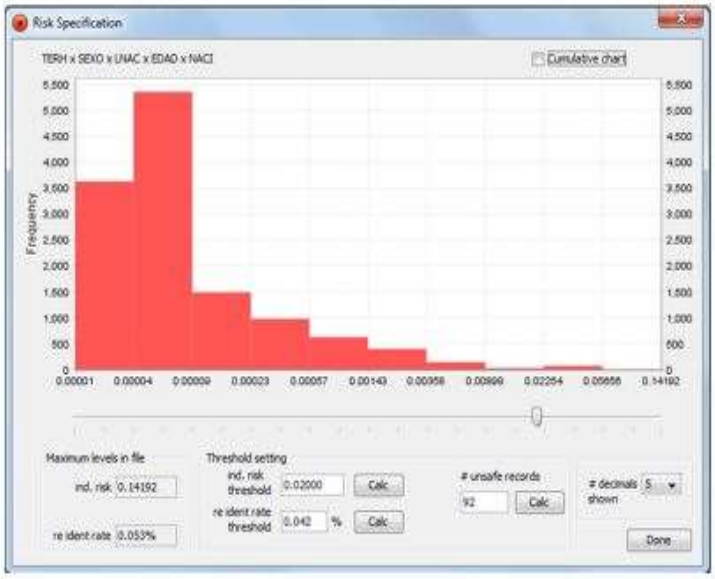


Fig. 5. µ-Argus risk chart with global recoding (EDAD and NACI)

We now have half as many unsafe records as in the previous step. The risk of re-identification has also decreased (0.053%); in such a scenario, we would assume that 92 of the records could be identified with reasonable effort. However, µ-Argus offers the possibility to suppress values in the unsafe combinations at the end of the protection process.

Summary of the process:

Step 1: Choose the key variables that we are going to study (you can think of adding some to the chosen ones)
Step 2: Calculate the combinations of all of them and the individual risk for each record according to those combinations.
Step 3: Choose the variables to be recoded and what this recoding will be based on the frequencies in each of the crossings.
Step 4: Recalculate the risk for new recoding for as many combinations as we consider until we find the recoding that fits what we want to give based on the information we offer and the protection of it.
Step 5: Generate the protected file, deleting information, or not, and review which records are problematic.

2.3 Microdata and metadata release

Once the confidentiality of the microdata file has been ensured up to a certain level of risk, the file is made available to our users.

Eustat offers microdata for public use and research purposes on its website. Public use microdata is available without additional requirements and free of charge, while research access requires on-site access at Eustat's facilities to ensure proper data handling.

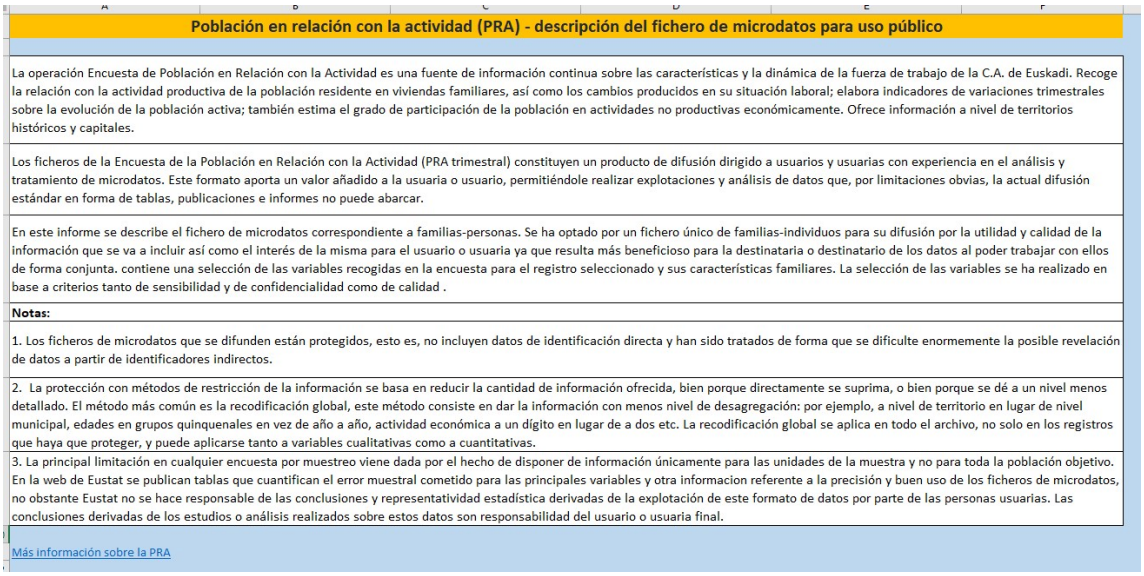
Metadata is also provided, describing the dataset's characteristics, variables, methodology, and limitations of use. Metadata provides additional and descriptive information about the collected and processed data. It is like a label that accompanies the data, providing details about its origin, the methodology used, and other relevant aspects.

Users can access this metadata to understand how the data was obtained, how it was processed, and what limitations may exist. This allows them to assess the quality and accuracy of the statistical results and use them appropriately in their own analyses.

In addition, by facilitating access to metadata, users can recreate the procedures used and verify the results obtained. This promotes transparency and knowledge sharing, improving the reliability and validity of the data and statistical analyses on our website.

At Eustat, we have generated a common metadata document for all the microdata files that are disseminated. This document consists of an Excel file with two sheets that provide essential information about the data, the conditions of use and the protection of the privacy of the respondents.

The first sheet, called "Metadata", provides specific information about the statistics. It includes details about the methodology used, the sample selected, the data collection period, and any relevant special considerations. In addition, the conditions of use are detailed, including the terms of reference and the restrictions applicable to the use of the data. It is important that users read and understand these conditions before using the statistical data.



Población en relación con la actividad (PRA) - descripción del fichero de microdatos para uso público

La operación Encuesta de Población en Relación con la Actividad es una fuente de información continua sobre las características y la dinámica de la fuerza de trabajo de la C.A. de Euskadi. Recoge la relación con la actividad productiva de la población residente en viviendas familiares, así como los cambios producidos en su situación laboral; elabora indicadores de variaciones trimestrales sobre la evolución de la población activa; también estima el grado de participación de la población en actividades no productivas económicamente. Ofrece información a nivel de territorios históricos y capitales.

Los ficheros de la Encuesta de la Población en Relación con la Actividad (PRA trimestral) constituyen un producto de difusión dirigido a usuarios y usuarias con experiencia en el análisis y tratamiento de microdatos. Este formato aporta un valor añadido a la usuaria o usuario, permitiéndole realizar explotaciones y análisis de datos que, por limitaciones obvias, la actual difusión estándar en forma de tablas, publicaciones e informes no puede abarcar.

En este informe se describe el fichero de microdatos correspondiente a familias-personas. Se ha optado por un fichero único de familias-individuos para su difusión por la utilidad y calidad de la información que se va a incluir así como el interés de la misma para el usuario o usuaria ya que resulta más beneficioso para la destinataria o destinatario de los datos al poder trabajar con ellos de forma conjunta. contiene una selección de las variables recogidas en la encuesta para el registro seleccionado y sus características familiares. La selección de las variables se ha realizado en base a criterios tanto de sensibilidad y de confidencialidad como de calidad .

Notas:

1. Los ficheros de microdatos que se difunden están protegidos, esto es, no incluyen datos de identificación directa y han sido tratados de forma que se dificulte enormemente la posible revelación de datos a partir de identificadores indirectos.

2. La protección con métodos de restricción de la información se basa en reducir la cantidad de información ofrecida, bien porque directamente se suprima, o bien porque se dé a un nivel menos detallado. El método más común es la recodificación global, este método consiste en dar la información con menos nivel de desagregación: por ejemplo, a nivel de territorio en lugar de nivel municipal, edades en grupos quinquenales en vez de año a año, actividad económica a un dígito en lugar de a dos etc. La recodificación global se aplica en todo el archivo, no solo en los registros que haya que proteger, y puede aplicarse tanto a variables cualitativas como a cuantitativas.

3. La principal limitación en cualquier encuesta por muestreo viene dada por el hecho de disponer de información únicamente para las unidades de la muestra y no para toda la población objetivo. En la web de Eustat se publican tablas que cuantifican el error muestral cometido para las principales variables y otra informacion referente a la precisión y buen uso de los ficheros de microdatos, no obstante Eustat no se hace responsable de las conclusiones y representatividad estadística derivadas de la explotación de este formato de datos por parte de las personas usuarias. Las conclusiones derivadas de los estudios o análisis realizados sobre estos datos son responsabilidad del usuario o usuaria final.

Más información sobre la PRA

The second sheet, titled "Variables", is a detailed guide to the variables present in the microdata file. Each variable is described in terms of its meaning, its category, and the associated level of protection. The ‘description’ column provides a clear and concise explanation of the variable, which makes it easier to understand its relevance in the context of the statistical operation. The ‘category’ column indicates the thematic classification of the variable, enabling more efficient navigation within the dataset. Finally, the level of protection indicates the degree of confidentiality and anonymization applied to each variable, thus ensuring the privacy of the informants.

Población en relación con la actividad (PRA) - diseño de registro del fichero de microdatos para uso público					
Número de orden	Nombre	Tipo	Descripción	Categorias	Tratamiento
1	NUMH	Num	Número de hogar		
2	AENC	Num	Año de encuestación		
3	TENC	Num	Trimestre de referencia		
4	TERH	Char	Territorio	**01** Alava **20** Gipuzkoa **48** Bizkaia	
5	MUNI	Char	Capital	**1** Bilbao **2** Vitoria-Gasteiz **3** Donostia / San Sebastián **9** Resto	Variable identificativa, se agrega por motivos de confidencialidad
6	SEXO	Char	Sexo	**1** Hombre **6** Mujer	
7	LNAC	Char	Lugar de nacimiento	**1** CAE **2** Resto de España **3** Resto del mundo	Variable identificativa, se agrega por motivos de confidencialidad
8	EDAD	Char	Edad	**01** 0-4 **02** 5-9 **03** 10-15 **04** 16-19 **05** 20-24 **06** 25-29 **07** 30-34 **08** 35-39 **09** 40-44 **10** 45-49 **11** 50-54 **12** 55-59 **13** 60-64 **14** 65-69 **15** 70-74 **16** 75-79 **17** 80-84 **18** >= 85	Variable identificativa, se agrega por motivos de confidencialidad
9	NACI	Char	Nacionalidad	**1** Española **2** Extranjera	Variable identificativa, se agrega por motivos de confidencialidad

This Excel workbook not only provides detailed statistical information but also ensures the proper use of the data. By offering a complete description of the variables and their levels of protection, users can leverage the data effectively while respecting the privacy of the respondents. It is a valuable tool for researchers, analysts, and anyone interested in reliable and secure statistical data analysis.

3 Conclusions

Risk analysis and the responsible dissemination of statistical microdata are crucial for maintaining confidentiality while providing valuable information to users. Eustat adheres to regulations, employs rigorous risk analysis methods, and offers a range of statistical products and access options. By ensuring data protection and providing comprehensive metadata, Eustat promotes the proper use of statistical information and contributes to knowledge advancement.

In conclusion, risk analysis and responsible dissemination of microdata are essential in the field of statistics. Eustat, as a statistical authority, follows regulations, applies rigorous protection measures, and offers diverse statistical products. Through these efforts, Eustat ensures confidentiality and promotes the proper utilization of statistical information for various purposes.