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

Marta Mas Moreno

Ana Maria Miranda Ligüerzana

UNECE

DESTATIS
Statistisches Bundesamt

UNECE EXPERT MEETING ON STATISTICAL DATA CONFIDENTIALITY

26 to 28 September 2023

Hochschule RheinMain, Building A ("Audimax" room), Kurt-Schumacher-Ring, 18 65197 Wiesbaden

Eustat

Outline

1. Introduction
   - Regulations
   - Statistical products

2. Microdata Protection
   - Generation of ready-to-access microdata
   - Case study: Example of application to PRA microdata

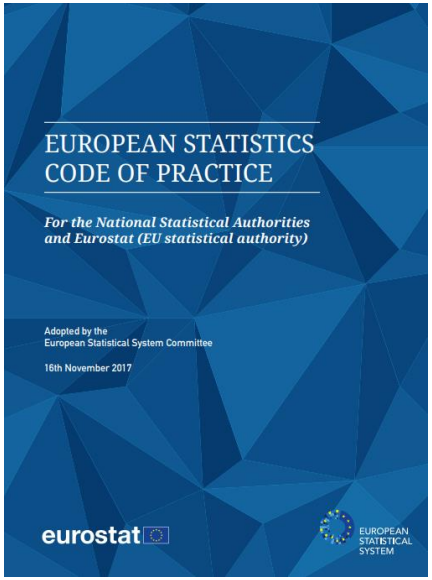
3. Microdata and metadata release

4. Conclusions

Regulations

- One of the main goals of a statistical office is to maintain and provide statistical confidentiality for its respondents.

- The information they provide, its use only for statistical purposes and the security of the data should be preserved in all the stages of statistical production.



PRINCIPLE 5

Statistical Confidentiality and Data Protection

The privacy of data providers, the confidentiality of the information they provide, its use only for statistical purposes and the security of the data are absolutely guaranteed.

Regulations

Law 1986/4 on Statistics of the Autonomous Community of the Basque Country.

CHAPTER IV: STATISTICAL CONFIDENTIALITY

Maintaining the **privacy** of the data providers (households, individuals, businesses and administrations), preserving the **confidentiality** of the information they provide and its use only for **statistical purposes** must be fully guaranteed within the statistical activity.

Regulations

Law 1986/4 on Statistics of the Autonomous Community of the Basque Country.

Article 29: COMPETENCIES OF THE BASQUE STATISTICS OFFICE

The **publication and dissemination** of the statistical results included in the Basque Statistical Plan and in the annual statistical programs correspond to Basque Statistics Office.

Dilemma

Guidelines

Manual and Guides for quality in statistical production

- Register of technical projects
- Treatment of confidentiality in Eustat statistics operations
- Manual of good practices in sample design and extraction
- Manual of outcomes for the calculation of response rates

In all the stages of statistical production

PUBLICATION AND DISSEMINATION

Protection measures for statistical products

- Frequency tables:

Population by locality in Alava, according to sex

		Population by sex		
		Total	Men	Women
0010	Jokano	34	18	16
0011	Luna	8	5	3
0013	Santa Eulalia	9	5	4
0014	Sendadiano	19	12	7
0016	Uribarri-Kuartango	10	9	1
0017	Urbina de Basabe	4	2	2

Note: The information of some localities has been grouped by geographical proximity criteria to preserve statistical confidentiality.

Disclosure risk: Low frequencies

Protection measures: recoding, grouping

- Magnitude tables:

Establishments and employed individual in the Basque Country

	Araba/Álava		Bizkaia		Gipuzkoa	
	Establishments	Individuals employed	Establishments	Individuals employed	Establishments	Individuals employed
Total	**24.208**	**152.557**	**87.379**	**455.454**	**56.896**	**303.645**
01. Agriculture, livestock, forestry and fishing	**1.943**	**2.843**	**1.568**	**4.215**	**1.542**	**2.971**
01. Agriculture, livestock, hunting and related service	1.901	2.746	1.208	1.692	1.340	1.790
02. Silviculture and forest exploitation	41	x	245	946	113	›
03. Fishing and fisheries	1	x	115	1.577	89	›

(x) Cell protected for confidentiality reasons

Disclosure risk: few contributions, dominant contributions

Protection measures: recoding, cell supression,...

Protection measures for statistical products

- **Geographical information systems.**
  - High-detailed information is available
  - Multiple area selection allowed

Disclosure risk: Low frequencies

Protection measures:

- Download is bounded
- Information on areas with less than 3 statistical units is not provided

Protection measures for statistical products

- Microdata: Statistical unit level data

Microdata file

Individual records

SEX	AGE	PLACE OF RESIDNCE	OCCUPATION	VAR1	VAR2
Male	38	Vitoria-Gasteiz	Sociologist	30.000	3
Female	26	Bilbao	Statistician	50.000	5
Female	50	Lanestosa	Astronaut	400.000	8
.	35	Donostia-S.Sebastian	Influencer	80.000	7
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Data

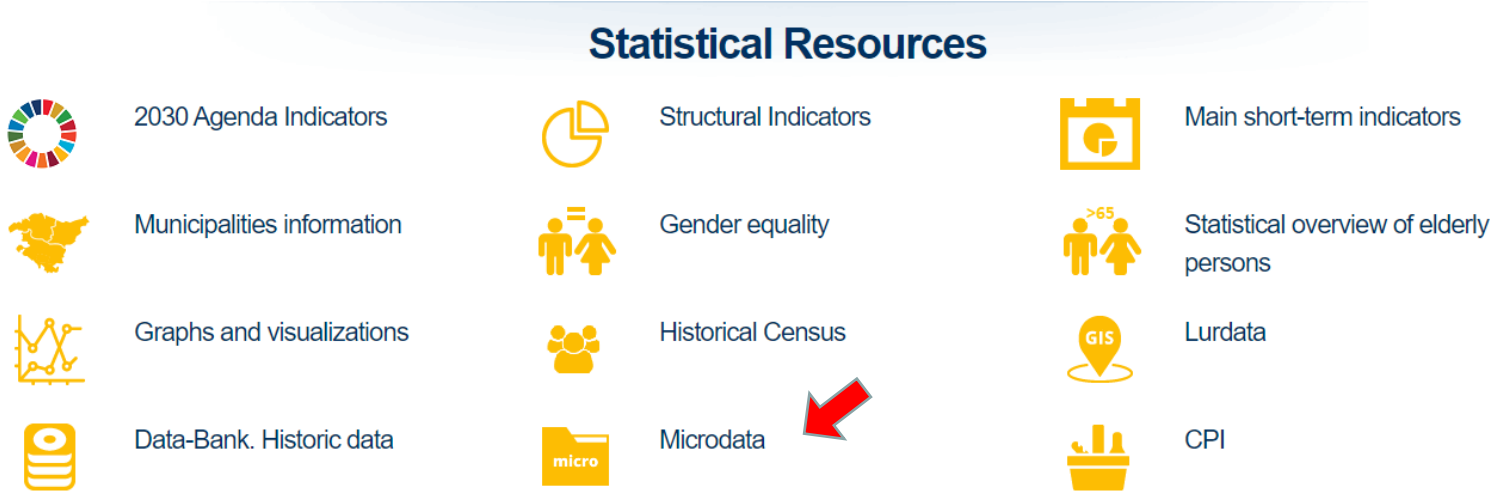
Disclosure risk: To infer the identity of a statistical unit with a high degree of certainty.

When “rare” combinations occurs in the microdata file

Microdata release

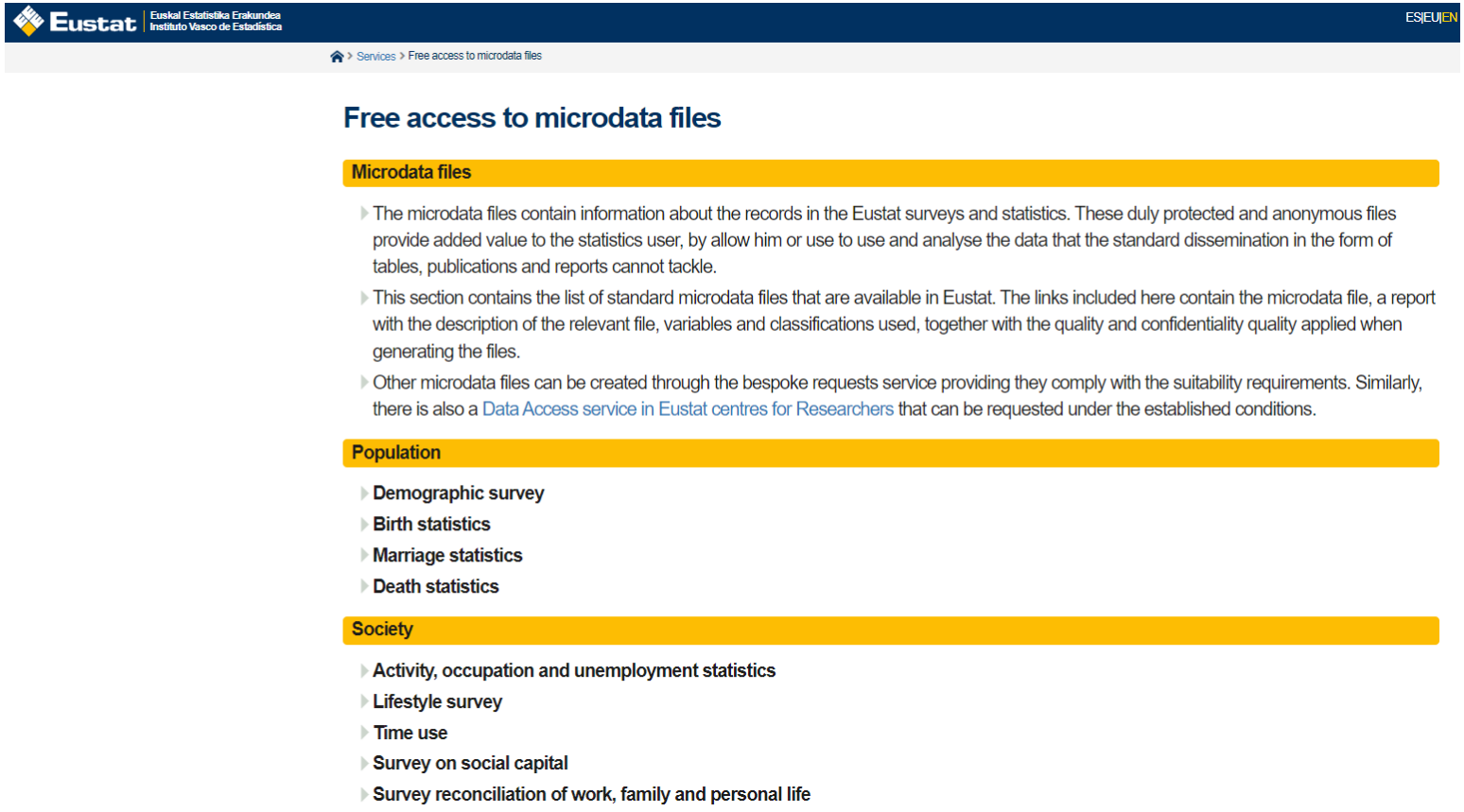
- Microdata for public use in Eustat

Protected and available to users without any type of requirement



https://en.eustat.eus/productosservicios/fich_microdatos_i.aspx

Microdata release



https://en.eustat.eus/indice.html

Microdata Eustat

Microdata protection

Generation of ready-to-access microdata

- Microdata files that are available for public access wil be protected.
- Phase 1: evaluating which records can be easily identified
- Phase 2: applying some protection measure.

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.

General protection criteria

- Geographic criteria – Geographic variables from areas below a given size will not be included.

- The identifying variables included (sex, age, marital status, profession, educational level,...) will be categorized according to the **risk analysis** of file identification.

- Special care with “fusion” variables with other public files (age, date of birth, etc.)

Example of application to PRA microdata

Survey on the population in relation to activity

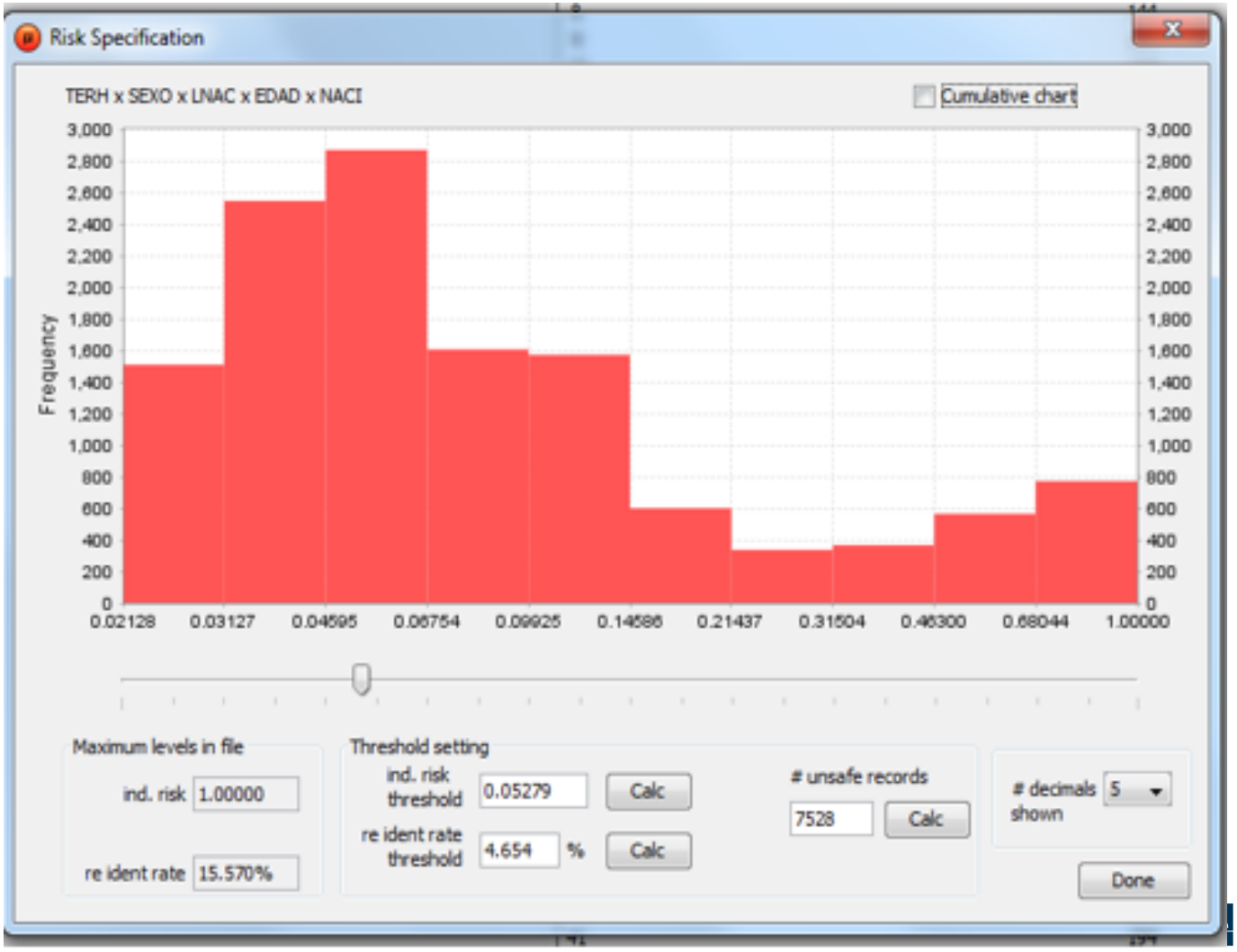
Products by operation > Survey on the population in relation to activity

The Survey on the Population in Relation to Activity operation is a continuous source of information on the characteristics and dynamics of the labour force of the Basque Country. It records the relation to productive activity of the population resident in family households, as well as the changes produced in labour situations; it produces indicators of conjunctural variations in the evolution of the active population; it also estimates the degree of participation of the population in economically non-productive activities. It offers information on the province and capital level.

Statistical tables	
Activity, employment and unemployment rates	
Activity and unemployment rate of the population aged 16 and over in the Basque Country by quarter according to province (%) (1). 2023/II	07/21/2023
Activity rate of the population of 16 and over in the Basque Country by quarter according to age and sex (%) (1). 2023/II	07/21/2023
Activity rate of the population of 16 and over in the Basque Country by quarter according to number of dependent children and sex (%) (1). 2023/II	07/21/2023
Employment rate of the population aged 16 to 64 in the Basque Country by quarter according to province, sex and age (%) (1). 2023/II	07/21/2023
Employment rate of the population of 16 and over in the Basque Country by quarter according to age and sex (%) (1). 2023/II	07/21/2023

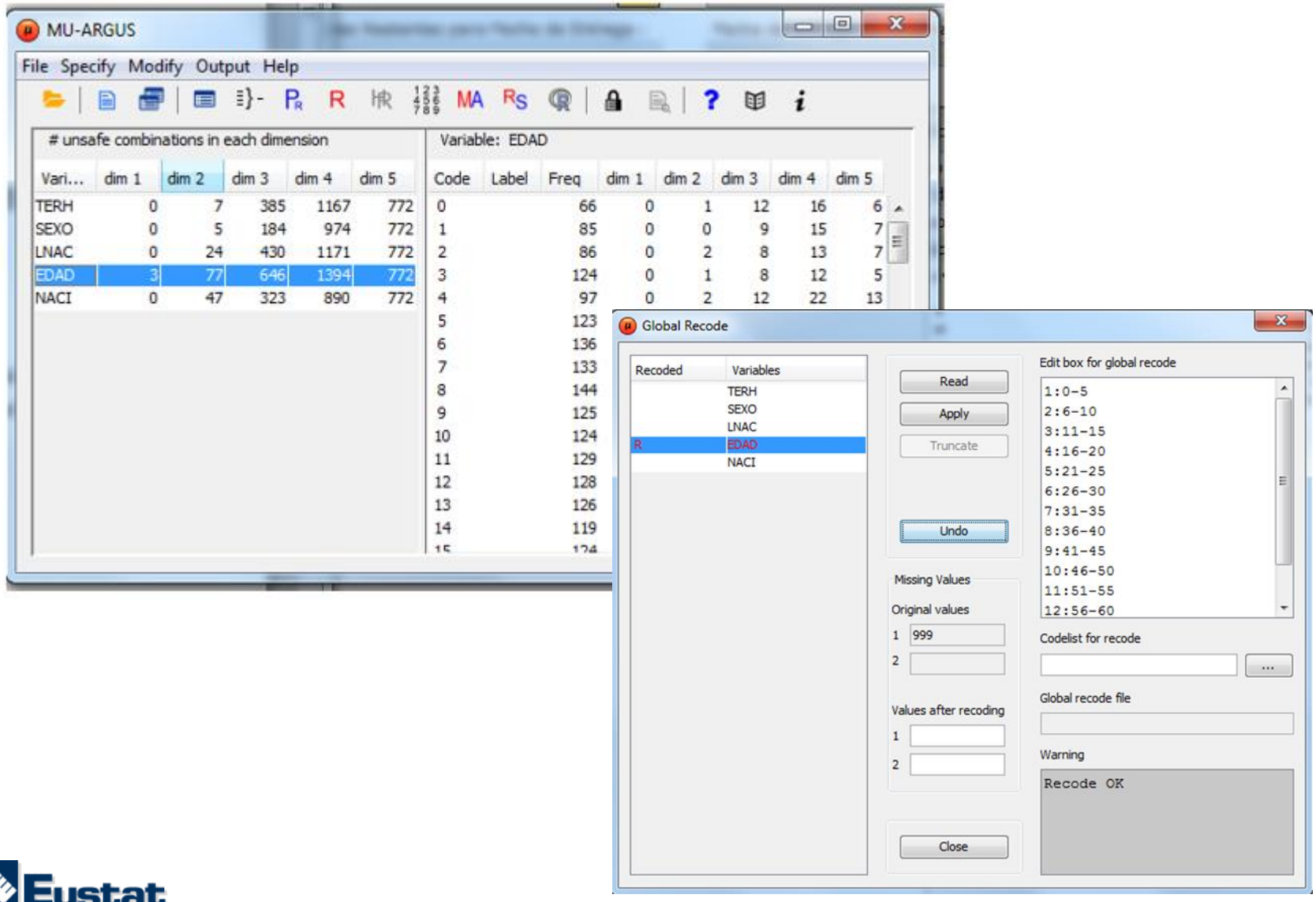
Example of application to PRA microdata

μ-ARGUS



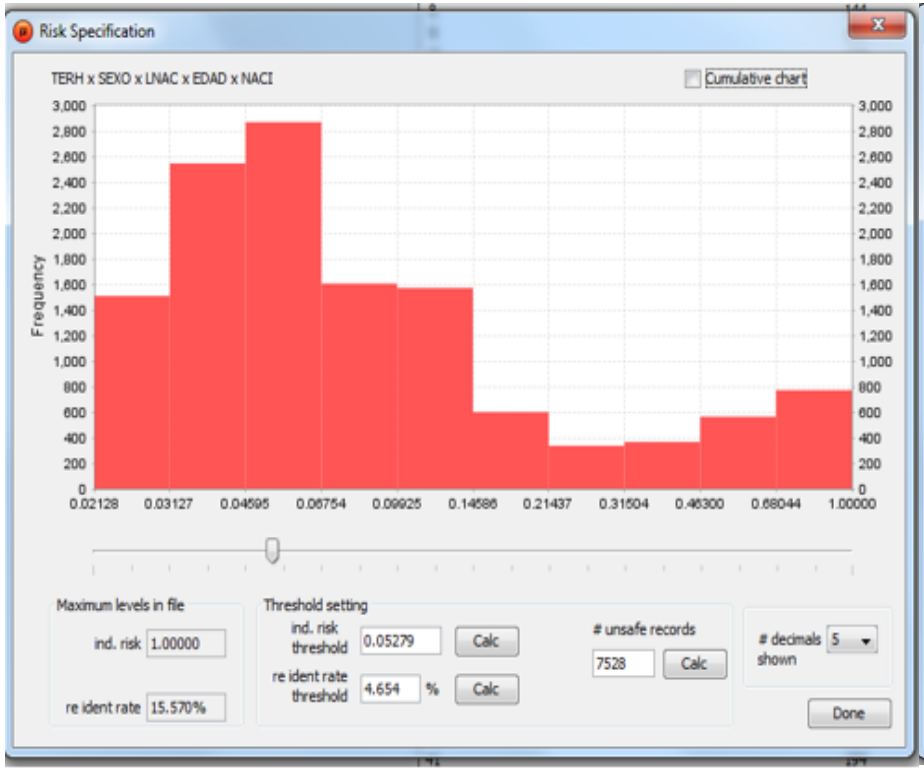
Eustat

Example of application to PRA microdata

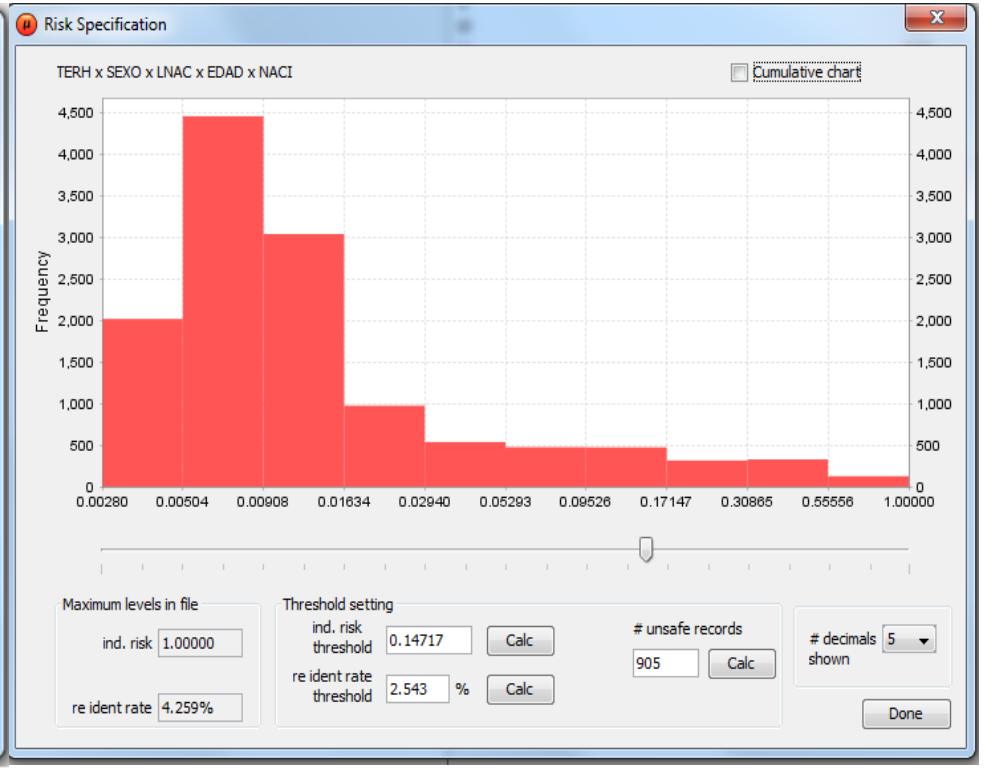


Example of application to PRA microdata

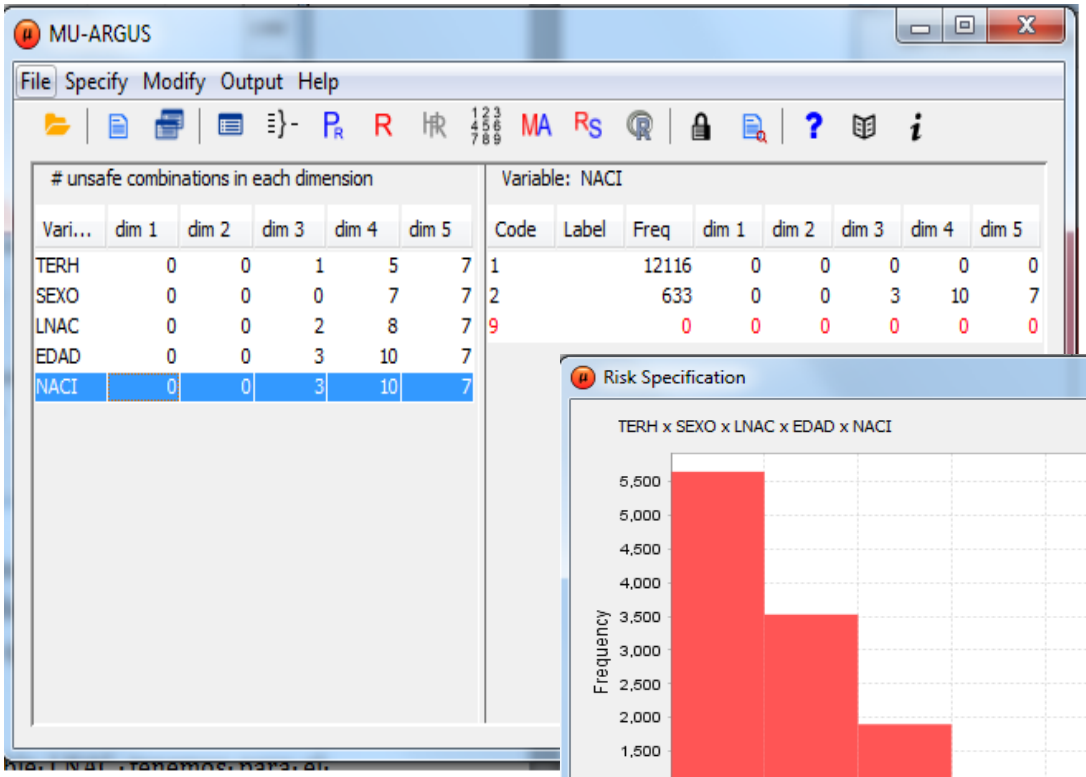
Unaggregated data

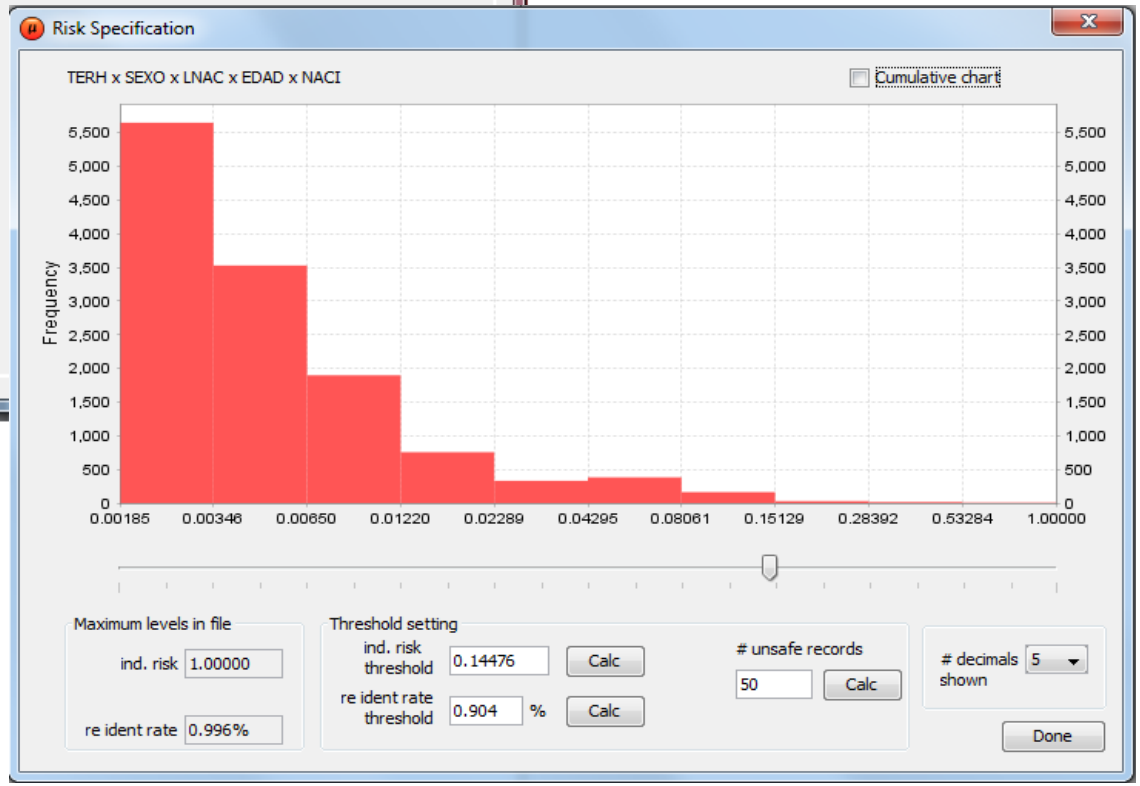


Added age



Example of application to PRA microdata





- Added age
- Birthplace added
- Nationality
- Territories and capitals

Example of application to PRA microdata

- ✓ The household identifier will be published, this identifier is NOT maintained from one quarter to the next so that a household cannot be identified from one quarter to the next. It will be a correlative number assigned to the homes that will be ordered randomly.

- ✓ The family variables generated from the information in the file have been removed, such as the number of assets in the home and other similar ones. Giving this variable made sense when we were not providing the dwelling identifier, now these variables can be obtained from the data in the file.

- ✓ We have added a variable to the file to identify the capitals (MUNI) and to be able to differentiate them from the rest of the municipalities. The maximum disaggregation will be territory in all the others. It has been generated in such a way that we could differentiate municipalities with more than a certain number of inhabitants in the event that we decided to give a greater geographical breakdown.

- ✓ Some variables have been excluded from the proposal, but not definitively. In some cases it is because they have not yet been published and are in the review phase, we will wait for their publication to add them to the microdata file. Some others have been excluded because it is considered that the data does not have sufficient quality, and when this changes we will reassess their inclusion in the file.

Microdata and metadata release

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
10	LEST	Char	Mayor nivel de estudios terminados	Actualmente información no disponible	
11	ENRE	Char	Sistema enseñanza reglada	S/N	

Metadata | Variables

Microdata and metadata release

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

Microdata and metadata release

Access to Researchers in Eustat centres

It is a service provided by this institute for access to data for scientific purposes.

The locations where the data are accessed from are under maximum security so as to preserve the confidentiality of these data, and are supervised by Eustat technicians.

The access centres are located in the Eustat offices in Vitoria-Gasteiz, Donostia-San Sebastián and Bilbao

Access Conditions:

1. - Request

A request must be made to the General Directorate of Eustat, with a summary of the research to be conducted attached to this request. The following must be included:

- Information on the Institution that is making the request (University, research centre, etc).
- Data of the person in charge of the research or project.
- Details of the person who will carry out the "in situ" analysis in the Eustat facilities.
- Purpose of the research or project and the need for access to the data.
- Detailed description of the research:
  - The total data for which access is required (relation of variables)
  - Methods of analysis to carry out
  - Computer requirements necessary for correct execution (software, hardware, etc).

Gracias

Eskerrik asko

Thank you