WP. 38
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (vii): Trans-border access to microdata

Basque Statistics Office Confidentiality project: final stages

Prepared by Anjeles Iztueta, Virginia Luengo, Marta Mas, Cristina Prado, EUSTAT, Spain

Basque Statistics Office Confidentiality Project: final stages

Anjeles Iztueta[2], Virginia Luengo[1], Marta Mas[3]Cristina Prado[2]
[1]Technical Assistance, Vitoria-Gasteiz, Basque Country (SPAIN)
[2]Basque Statistics Office (EUSTAT), Vitoria-Gasteiz, Basque Country (SPAIN)
[3] Basque Government Department of Education, Universities and Research. (SPAIN)

Abstract. The objective of this contribution is to present the main works in relation to confidentiality in statistical distribution, carried out by Eustat over the last few years. These works have been fundamentally centred on the elaboration of public microdata for its distribution on the Website, and making data accessible to in-situ researchers.

As well as establishing the history regarding microdata, the types of files created will be described, and the protection methods applied and the software used.

It will describe the established protocol and the new challenges set in relation to access for researchers. In the same vain, the motivating and coordinating role carried out from within the EUSTAT Confidentiality Council will be highlighted.

Keywords. Confidentiality, Re-identification, Statistical disclosure control, Microdata protection, Tabular data protection, On-site access

1 Introduction

In the Work Session on Statistical Data Confidentiality 2009 in Manchester, the different stages carried out by EUSTAT relating to distribution confidentiality were presented from a methodological and organisational point of view, and appear in the table below

Period	Action	Output
1988-1999	Research fellowship on data protection techniques and statistical confidentiality	Technical notebook on "*Statistical Data Protection Techniques*" edited by EUSTAT.
April 2000	International Seminar on "Confidentiality and statistical data protection techniques" organized by EUSTAT. Lecturer: L.H. Cox	Publication: *"Confidentiality and statistical data protection techniques"* L.H. Cox edited by EUSTAT.

September 2000	Security Analysis of Census Tables	Internal report about sensitive crosses and dissemination proposal
October 2000	Participation in OFISTAT (Official Statistics List of distribution) Seminar on Statistical Confidentiality	Discussion about the proposed document: *“Statistical Secret protection: basic elements of a data protection system”* by A.Garín, J. Urrutia
2001	Participation in The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Skopje, Macedonia, 14-16 March)	Article: *“A comparative test for several threshold values in frequency tables: A Tau-Argus performance example.”*
2002	Tabular Data protection of preliminary results of the Census 2001, using Tau-Argus (optimal method).	Publication of suppression patterns for frequency tables with fine geographical levels.
2003-2004	CASC project pursuit.	Testing of Argus software.
June 2004	Attendance of PSD (Privacy in Statistical Databases) Conference. (Barcelona, Spain, 6-9 June)	
2005	Staff training on disclosure control and protection software.	Internal Workshop on SDC techniques and ARGUS.
2006	Work on standard safety criteria	Internal report about analysis of sources and internal situation.
December 2006	Attendance of PSD Conference. (Rome, December)	Feedback and contacts.
2007	Creation of “Rules of Confidentiality in statistical distribution”	Internal distribution
2007	rules for Website tables	in the Confidentiality Council

2007	Constitution of Confidentiality Board	Group of experts to assess and deal with issues of confidentiality in terms of distribution.

Table 1.1 Summary of the methodological and organisatory actions carried out for the treatment of confidentiality in statistical distribution.

As described in Table 1.1, during the first phases of the project the actions regarding confidentiality were fundamentally centred on the understanding and application of data protection, through the knowledge and development of ARGUS software. This phase was developed in its entirety in the institute's area of methodology

Subsequently, efforts were centred on the creation of written rules and criteria for its use in the organisation, for which the participation of all of the areas of EUSTAT was fundamental: production, distribution, methodology and administration, both in the taking of decisions regarding which criteria to adopt, and in its fulfilment.

The constitution of the Confidentiality Council was crucial for this last point, where the rules to be applied were discussed, as well as the questions raised with specific information requests.

From 2008, the works of the project were centred on the creation of microdata files that were public or available on request, as well as on establishing in-situ data access for researchers.

2 Microdata for standard distribution

2.1 Background

Until 2008, the politics surrounding microdata in EUSTAT was completely restrictive. Previously, applications by researchers were attended to internally in the institute and socio-demographic microdata files were only created exceptionally, on request. It should also be noted that the consideration at the time was that there was no real demand as such, given that the requests were very occasional.

From 2008 a period of reflection was initiated via the Confidentiality Council with the purpose of creating standard microdata and offering them on the Website, due to the fact that it was a product that was starting to become commonplace in European statistics offices, and that one of the desires was to provide a better services to users, especially in the research field.

At the same time the first public microdata files were being created, with the aim of contrasting the applied criteria within the different areas.

One question that was put forward during the reflection period was that of creating standard microdata files or only creating them on request, given that the different needs of users is wide and one type of file for all users might not yield a return on the effort and resources used.

The creation of some microdata files and their being made available on the website was a factor that energised user requests from the university world, where requests were made for files that were offered as well as others not offered.

The selection of surveys for the creation of microdata files was established within the Confidentiality Council, based on the possibility that users could find it useful to be able to exploit them.

In the first phase files of surveys directed towards individuals and families were dealt with, followed by a case of files of surveys directed towards businesses.

The data related to businesses presented a specific problem since the protection of indirect identification was complicated, due to the commercial structure of the Basque Country, concentrated in certain sectors in a small number of businesses.

In cases of researchers who request microdata on businesses, this has relied on ad-hoc requests or in-situ access to data.

Standard files on offer are announced in a specific section of the Website http://en.eustat.es/ci_ci/productosServicios/fich_microdatos_i.html#axzz1YE0N3TT , where a form is provided to be completed, http://en.eustat.es/documentos/Mod_solicitud_microdatos_i.pdf, by petitioners where they must specify their details and the reason for the request.

2.2 Methodology for the creation of microdata

The methodology for the creation of standard microdata that Eustat has employed since 2008 is the following:

- Study the structure of the files separated, in hierarchies...
- Select the identification variables (sex, age, place of residence, civil status, profession,...).
- Identify sensitive variables (data with information related to ideology, union membership, religion, beliefs and health) taking into account the Law on the Protection of Personal Data.

A risk analysis is carried out on the group of identification variables. Microdata protection software (Mu-Argus) has been used, developed by the Dutch Statistics Institute (CBS) within the programs of the European framework. This software applies two methods for microdata protection, a traditional risk method and a method based on association probability.

2.2.1 Unique record identification

The traditional risk method is based on the identification of unique or “rare” records, with regards to a specific combination of identifying or “key” variables. A geographic variable is normally included amongst these variables. Keys that generate a large number of unique records are analysed and variables with the greatest identification power are detected.

2.2.2 Ratio of re-identification or average global risk.

The detection of a unique record in a sample does not guarantee that this is also unique in the population. For these cases Argus employs another method based on probabilities of association. This method models the strategy of a hypothetic intruder that tries to establish a link between the unit in the statistical file and a unit of an available register or population file.

The re-indentification is produced if this link is correct. It defines a probability of re-identification by record and imposes a threshold value for that probability.

The probability of re-identification for each record and with respect to the K key (combination of identification variables) is given by the expression:

$$r_i=1/F_k$$

where F_k= Frequency in the population of the combination of k values of the key.

For all the records that have the same combination of key values, the probability of re-identification will be the same.

A record will be “unsafe" if its probability of re-identification is greater than or equal to the fixed threshold.

When the F_k is unknown, it is modelled:

$F_k/f_k \approx$ BinNeg (p_k, f_k) p_k being $= \dfrac{f_k}{\sum_{i:k(i)=k} w_i}$ where w_i = sample weight of record i

X = Expected number of re-identifications in the file.

In a k cell:

$X_k \approx \text{Bin}(f_k, r_k) \Rightarrow E[X_k] = f_k{*}\, r_k$

For the whole file:

$$X = \sum_{k=1}^{K} E[X_k] = \sum_{k=1}^{K} f_k r_k$$

we define a re-identification ratio

$$Ratio.\mathrm{Re}-ident. = \frac{1}{n}\sum_{k=1}^{K} f_k r_k$$

This re-identification ratio is a more intuitive method and gives us a better risk interpretation of a file with regards to a k key (combination of identifying variables).

2.3 Surveys analysed

During the last few years various statistical surveys have been analysed with the aim of offering standard distribution microdata files:

- Survey on living conditions (ECV 2008)
- Survey on demographics and validation (EDV 2009)
- Survey on social capital (ECS 2010)
- Survey on environment - families (EMAF 2010)
- Survey on the information society - families (ESIF 2011)

Given that the process is similar, below we are going to describe the process followed for three survey types:

2.3.1 Survey on living conditions

The survey on living conditions is carried out by sampling. The size of the sample is 4909 families for the family questionnaire, and 4909 individuals for the individual questionnaire (one person per kish table is selected in each family). The risk has been analysed separately for each file (individuals and families).

Treatment for identifying variables for the individuals file and the families file

For the analysis, the following group of **identifying** variables is considered:

DESCRIPTION Individuals file	DESCRIPTION Families file
Province (3)	Province (3)
Municipality (104)	Municipality (104)
Zone (9)	Zone (9)
Age (100)	Age (100)
Profession (9)	Profession (9)
Sex (2)	Number of spaces (24)
Civil status (5)	Family size (9)
Place of birth (5)	Place of birth (5)
Level of education (4)	Place of birth (5)
Relation to activity (3)	Professional situation (7)

Unique records

In both files the recurring appearance of the variables municipality and age in the combinations with a greater number of unique records is observed. Attention should be drawn to the identifying power of the family size and number of household spaces variables, in combination with the two previous variables.

Re-identification ratios

A probabilistic method has been applied to the keys with a greater number of unique records and, following the observation of the results, necessary recoding has been carried out, achieving a reduction in the risk of re-identification of the individual and family files down to an acceptable level.

For both for the individual and families files the keys where the re-identification ratio is greater by biggest to smallest are:

<u>Individual file:</u>

Municipality (104) x Age (100) x Profession (9)
Municipality (104) x Age (100) x Level of education (4)
Municipality (104) x Age (100) x Place of birth (5)
Municipality (104) x Age (100) x Profession (5)
Municipality (104) x Age (100) x Sex (2)
Municipality (104) x Age (100) x Relation to activity (3)

<u>Family file:</u>

Municipality (104) x Age (100) x Profession (9)
Municipality (104) x Age (100) x Number of spaces (24)
Municipality (104) x Age (100) x Family size (9)
Municipality (104) x Age (100) x Professional situation (7)
Municipality (104) x Age (100) x Level of education (4)
Municipality (104) x Age (100) x Profession (5)

Recoding for the standard distribution families and individuals microdata file:

- ZONE: geographic variable coded in 12 modes in the families and individuals files (9 modes of districts and 3 modes of provincial capitals)
- AGE of the person of reference coded in five-yearly groups in the individuals and families files given its high power of identification in combination with the other demographic variables (18 modes).
- FAMILY SIZE coded in 5 modes in the family file.
- NUMBER OF SPACES coded in 5 modes in the family file.

Sensitive variables

The variables that have been identified as sensitive are related to household security devices, social environment and citizen security (directly related to physical aggression, crime, drug addiction or prostitution), as well as direct references to economic restrictions.

2.3.2 Survey on the information society - families (ESIF)

The ESIF is a sampling survey and its final structure has been formed by joining the families and the individuals files. Each record of this file is an individual with their individual and family characteristics. The sampling base consists of 4902 individuals.

Treatment for identifying variables

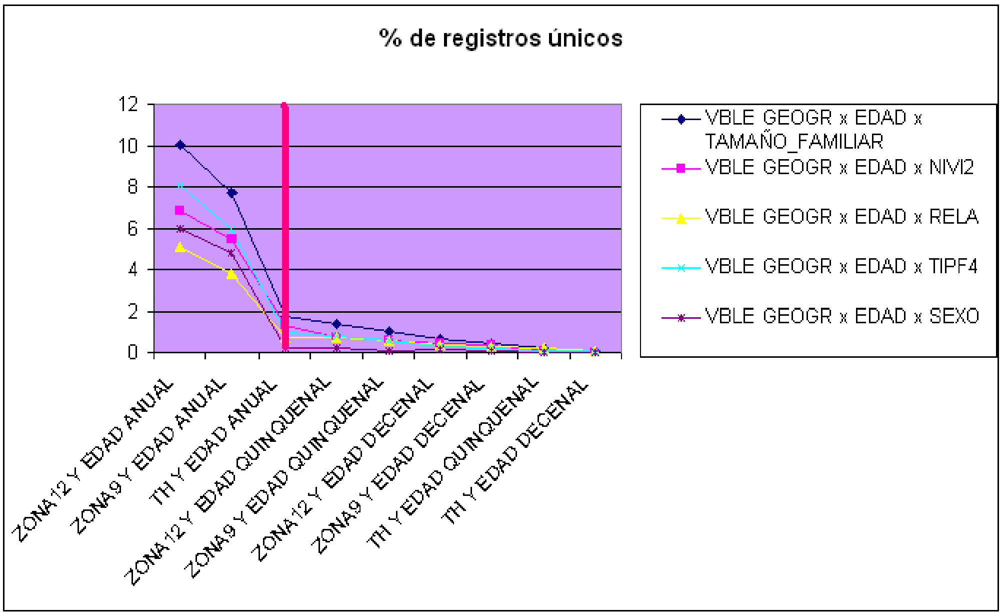
The identifying variables that have been used in the Argus analysis are the following:

DESCRIPTION **Individuals-families file**
Province (3)
Sex (2)

Age (100)
Level of education (3)
Relation to activity (3)
Family size (8)
Type of family (3)

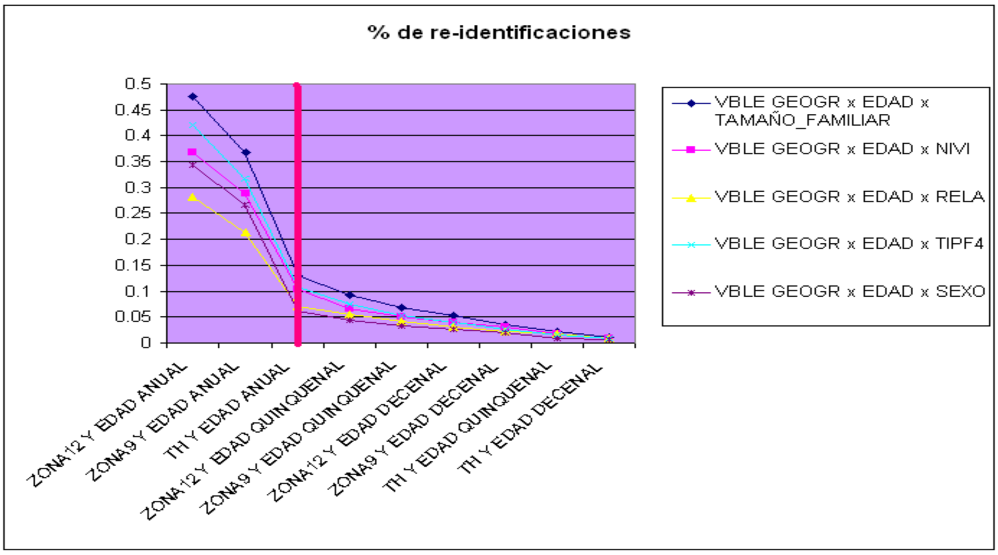
Unique records

In the following graph the reiterating appearance of the province and age variables in the combinations with a greater number of unique records can be observed. Also noteworthy is the identifying power of the family size, level of education, relation to activity and family type variables, in combination with the previous variables.



Re-identification ratios

The probabilistic method has only been applied to those keys that generate the greatest number of unique records and recoding has been carried out on the variables to reduce the risk of re-identification down to an optimum level.
The keys where the re-identification ratio is greater by order of higher to lower are:
Geographic variable x Age (100) x Family size (8)
Geographic variable x Age (100) x Level of education (3)
Geographic variable x Age (100) x Relation to activity (3)
Geographic variable x Age (100) x Family type (3)
Geographic variable x Age (100) x Sex(2)
All of this can be observed in the following graph:



Recoding for the standard distribution microdata file.

- Geographic variable (3 categories – province level).
- FAMILY SIZE has been coded in 4 categories.

Sensitive variables

No variables of a sensitive nature are included.

2.3.3 Future task

Continue during the 2011 and 2012 with the creation of new microdata files (Labour Force Survey, Natural Population Movement, Survey on Family Conciliation...), in a way that necessary routines are created amongst the producers, so that their creation is integrated into the productive process of statistics.

3 In situ access to researchers

This service has appeared in EUSTAT at the same time as the start of the creation of the microdata files and with the aim of giving a better service to researchers, above all in those cases where use of individualised data is required, where it could be indirectly identifiable, as in the case of economic data.

In these cases access to EUSTAT facilities was permitted from special computer hardware that had to comply with a rigorous protocol before, during and after the carrying out the operations requested by the researchers. The established protocol is set out below.

3.1 Request

As a prior step to in situ access, a request must be made to the General Directorate of the Institute that must contain:

- Information on the Institution that is making the request (University, research centre, etc).
- Details of the researcher in charge of the research or project.
- Details of the person who will carry out the "in situ" analysis in the facilities of the institute.
- 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
    - Description of results expected (exploitations, tables, reports...)
    - Computer requirements necessary for correct execution (software, hardware, etc).
- Work plan (estimated time to be spent in EUSTAT facilities)

3.2 Access to confidential data for scientific purposes

The request for access will be studied by the Eustat Confidentiality Council.
To this effect, authorisation will be given on the basis of two requirements:

- **Subjective requirement:** The request must originate from:
    - Universities or other higher education institutions.
    - Scientific research organisations or institutions
    - Other organisations, agencies or institutions, after obtaining the report from the Eustat Confidentiality Council.
- **Objective requirement:** Access to anonymous confidential data will only be permitted for scientific purposes, for which there must be an analysis of the scientific research that is to be carried out via access to protected data.

  In this sense, the European Data Protection Agency, in a report from 2002, has defined the term "scientific" in the following way:

  The term "scientific" from a semantic point of view implies belonging to a science. Such an expression, understood literally, has an all-encompassing range that would imply the possibility of connecting practically any treatment of personal data to a scientific speciality, both social and natural. Thus, even a study on markets, publicity, or commercial or advertising techniques would have or could establish a connection with a speciality or branch of knowledge (economic sciences, information sciences, etc.)

In the case of rejection of the request, the interested party will be informed of the reasons for the rejection, with the possibility of rectification on the part of the petitioner on the points of conflict.

3.3 Signing of a contract with conditions of access.

Once the request is accepted, the requesting party must sign a contract in which the following aspects are specified:

- Conditions of access to data.
- Obligations of researchers.
- Obligation of confidentiality.
- Measures for respecting confidentiality with regards to statistical data.
- Sanctions in the case of contract violation.

3.4 Access to data in Eustat centres

Access to anonymous individual statistical data will be carried out in Eustat facilities.

To this end a computer will be made available with the following physical and technical restrictions:

- All data entrance and exit ports (floppy disk drive, CD drive, USB ports, printer, etc.) will be blocked.
- The most common statistical tools will be installed with the possibility to install other pieces of software required by the researcher, as long as the relevant licence is provided.
- In no case will the equipment be connected to the production environment or the internal EUSTAT network and only local access will be provided to the data required by the researcher.
- No Internet or email access will be available.

It is not recommended that the facilities prepared for this purpose are equipped with telephones, faxes or any other method of telematic communication, because supervision is limited.

Access to facilities will be within the hours and days planned by the requesting party and agreed to by both parties in the access to microdata contract. A Eustat official will be designated who will be responsible for supervising the work of the researcher.

3.5 Results of the research

The results of the research will not be allowed to leave the Eustat facilities before they have been checked by the official charged with supervising the work, who must verify that they do not contain data which contains a risk of disclosure of statistical secrets.

3.6 Future tasks

At the moment the described in situ data access service is only offered in the main EUSTAT office, but solutions are being studied to offer it from the provincial offices in Bilbao and Donostia-San Sebastián This service is based on infrastructure of virtualitation of servers (VMWARE)

Regarding the possibility of Remote Access, which is considered the most comfortable solution for researchers, it would be the next phase of study, notwithstanding the evaluation of development costs for this type of solution and the number of requests that we could receive from researchers.

4 Conclusions

- The creation of microdata for distribution and access to data for research purposes is fundamental in terms of developing and promoting the production of statistics, especially in these times.
- The driving role of these types of transversal projects is indispensible, in our case this role is carried out by the Confidentiality Council integrated by the different departments of EUSTAT.
- The creation of microdata requires teamwork between methodologists, experts on protection techniques, and producers of statistics, the data experts.
- It is necessary for all of this know-how to extend to all producers of statistics so that the creation of microdata is one output more in statistical operations. This will be achieved through training.

References

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on *The Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of such data.*

Basque Statistics Office - EUSTAT (1999) *Statistical Data Protection Techniques.* Technical notebook.

Basque Statistics Office - EUSTAT (2007) *Treatment of Confidentiality in EUSTAT statistical operations*. Confidentiality protocol.

Garín, A., Urrutia, J., (2000). *Statistical Secret protection: basic elements of a data protection system*. OFISTAT Seminar.

National Institute of Statistics - INE (1994) . *Population and Households Census 1991: Methodology*. ISBN: 84-260-2889-6. Madrid.

Law 4/1986 of 23 April - *Basque Statistical Law.*

Law 15/1999 of 13 December - *Organic Law on Personal Data Protection.*

Statistical Programme Committee (2005) *European Statistics Code of Practice and Commission Recommendations.* Brussels.