WP.34
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

DEALING WITH CONFIDENTIALITY IN DISSEMINATION: THE EXPERIENCE OF THE BASQUE STATISTICS OFFICE

Supporting Paper

Prepared by Marta Mas (Technical Assistance, Vitoria-Gasteiz) and Cristina Prado (Basque Statistics Office, Spain)

Dealing with Confidentiality in Dissemination: The experience of the Basque Statistics Office

Marta Mas[1] and Cristina Prado[2]

[1] Technical Assistance, Vitoria-Gasteiz, Basque Country (SPAIN)
[2] Basque Statistics Office (EUSTAT), Vitoria-Gasteiz, Basque Country (SPAIN)

Abstract. One of the main goals of a statistical agency is to maintain and provide statistical confidentiality for its respondents. Confidentiality should be preserved in all the stages of statistical production and especially in the dissemination phase. If we consider the wide range of formats in which statistical information is available (tables, microdata, metadata, etc.) and the detailed classifications and fine-scaled geographical levels released, the problem of data protection has become a far from trivial issue lately. This paper describes not only the experience of the Basque Statistics Office (EUSTAT) in providing protection for its published products, but also the development of a comprehensive policy that includes the establishment of standard protection criteria, the constitution of an expert group and a commitment to future tasks.

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

1 Legal framework and preliminary issues

One of the main goals of a statistical agency is to maintain and provide statistical confidentiality for its respondents. The sole use of information for statistical purposes should be also guaranteed. Privacy rights are preserved by our Constitution and considered in statistical laws. Specifically, Chapter IV of the Basque Statistics Law (23rd April, Law 4/1986) concerns the duty to keep statistical secret and the type of data protected:

> *"[…]the duty to keep statistical secret protects any identifiable data as belonging to an specific person […]"*

In addition, at national level, the Organic Law of Personal Data Protection (13th December, Law 15/1999) guarantees the protection of personal data, defining this concept as:

> *"[…] any information related to an identified or identifiable person"*

But, what does *identifiable* mean?. This question was partially answered and defined by European Directive 95/46/EC which considers identification by *direct* or *indirect* means. This is to say that not only direct identifiers (ID numbers, names, surnames, addresses, telephone numbers, etc.) must be protected against disclosure but also indirect identifiers (sex, age, marital status, relation to activity, etc) or combinations of them should be considered to avoid identification.

Since it was founded in 1986, EUSTAT has implemented physical and technological measures in order to protect published products against direct and indirect identification. As a result of the internal statistical project *"Research and Development in Statistical Data Protection Techniques"*, several actions have been taken during the last ten years:

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.

Table 1.1 Summary of actions in *"R&D in Statistical Data Protection Techniques"*

A lot of work has been done but data protection practices are still applied by statisticians as a part of a "non-written" code of practice, based more on "know-how" rather than on stated rules. Therefore, standards should be discussed and fixed. In fact, one of the recommendations on Statistical Confidentiality included in Principle 5 of the European Statistics Code of Practice[1], refers to this point:

> *"[…] Instructions and guidelines are provided on the protection of statistical confidentiality in the production and dissemination processes. These guidelines are spelled out in writing and made known to the public […]"*

Since 2006, important efforts have been devoted to fulfilling this principle. Finally, in April 2007, a first draft on standard safety criteria was agreed after months of discussion. This whole process, the decisions adopted and the actions carried out are described in the following sections.

[1] Adopted by Statistical Programme Committee on 24 February 2005

2 Current situation

2.1 Confidentiality Board.

One of the main points of this process is the need for an expert group to make decisions about data protection and to give advice on confidentiality matters. Such a group should integrate members from all areas of EUSTAT in order to cover both technical and legal aspects.

The Confidentiality Board has been constituted this year at EUSTAT with the highest representative of each area and the General Direction. These are some of its main duties:

- To establish rules and criteria about confidentiality issues.
- To establish and make decisions concerning sensitive topics and sensitive variables.
- To discuss and approve public-use microdata structure.
- To decide about on-site access conditions.
- To solve specific queries (research-use microdata, etc.)
- To advise other statistical agents from the Basque statistics system on confidentiality matters and data protection procedures.
- To keep a coherent and updated system.

2.2 Establishing confidentiality criteria in dissemination.

2.2.1 Research of sources and other experiences.

A small group of experts (mainly from the methodological area) was constituted in order to make a preliminary analysis of the situation. In the first stage, this analysis consisted of an external search of sources and experiences from other statistical offices, regarding their policies on reporting and implementing confidentiality. The results of this phase were as varied as the sources consulted, but general conclusions could be drawn from the study:

- Legal framework is available to all sources consulted and it is considered essential as a starting point. In addition, guidelines about confidentiality treatment were found in all of them.
- It is less common to find information about the sensitivity rules applied and the values for the parameters of such rules, which are, in most cases, confidential.
- Disclosure control methods are applied to tables and microdata in most cases.
- Geographical thresholds are applied in many cases with diverse values and mainly in microdata releases.
- Almost all the sources provide microdata products (for research use and/or public use)

In the second stage, a summary of the most common data protection practices used by EUSTAT in dissemination was included. On the one hand, only the economic statistics area applies a standard procedure to protect the released data. In spite of the fact that any counting (frequency table) of establishments and companies is allowed by the Basque Statistics Law, no economical magnitudes are published if a cell frequency is less than three (only two or less establishments or enterprises contribute to one cell). Recoding of categories and manual suppressions are applied in order to avoid disclosure. On the other hand, the socio-demographic statistics area often applies 'ad-hoc' protection for each particular case, if a problem of a breach of confidentiality arises.

From the general EUROSTAT guidelines, the experience of other statistical offices and our own practices, an initial proposal on confidentiality criteria for standard dissemination has been developed at EUSTAT.

2.2.2 Microdata protection rules

Although there is no standard release of microdata at this moment in EUSTAT, some rules have been developed to be taken into account for specific demands of information and future public-use files:

- Microdata files released should not include, in any case, either direct identifiers or personal data.
- In general, microdata files will not include geographical indicators referring to areas under a fixed threshold (10,000 inhabitants).
- Aggregation level for other variables included in the file will depend on geographical level released and sensitivity of the variable itself. Therefore, the more geographical detail, the less conceptual level and the greater the sensitivity of the variable the greater the aggregation of categories.
- As an additional protection, disclosure control techniques (perturbation methods, record swapping, noise addition, etc.) could be applied to microdata, always preserving the statistical properties of data.

2.2.3 Tabular data protection rules

Some new rules have been added to the current uses of table protection at EUSTAT. The existing ones have been specified in more detail or modified in some way:

- Low values should be avoided in frequency tables with multiple crossings, where at least one of the variables is sensitive and the geographical indicator refers to an area under a fixed threshold (10,000 inhabitants).
- Dominant contributions should be avoided in magnitude tables in order to prevent accurate estimation of sensitive data belonging to a contributor in a cell. Sensitivity rules (minimum frequency rule or concentration rules (n, K – rule, pq-rule, etc.)) will be applied to detect sensitive cells.

- Appropriate protection techniques for tabular data (recoding of variables, primary and secondary cell suppression, etc.) will be applied in order to protect sensitive cells from disclosure.
- As a general rule, specific demands for information should respect the same protection criteria as standard dissemination. However, certain cases could be studied and discussed by the Confidentiality Board and specific measures might be taken.

2.3 Checking confidentiality criteria.

Having made a proposal on safety criteria, the next step consists of checking these rules against the data published at this moment by EUSTAT. Nowadays, the main results of statistics and data products are released through our website (www.eustat.es) by means of statistical tables and the data bank. Both sources will be revised.

Throughout this process, we shall focus on two main aspects: the geographical scope and the sensitivity of the variables used in each revised table. According to the confidentiality rules recently approved, low frequencies should not be published if the geographical detail refers to areas under a fixed threshold and at least one sensitive variable is involved. Therefore, each table considered should fulfil both conditions. In addition, dominant contributions in magnitude tables (mainly in economical surveys) will be checked.

At the moment of writing this contribution, we are in full checking process of our published products. The results of the checking will be shown to the members of the Confidentiality Board, who will have to discuss and decide about the problems or weaknesses found.

3 Future tasks

3.1 Towards a safe-standard microdata structure

A step forward in EUSTAT dissemination policy is the general release of microdata as a statistical product. Apart from the Economical Activity Directory, which is public by law[2], only two anonymised microfiles containing social survey data have previously been ceded in response to specific demands for information. However, the objective is to develop a standard product which will be available for the general public but with all the guarantees of confidentiality protection.

[2] The releases of Directory files containing identifying information about economical establishments, enterprises and other entities are not covered by the duty to keep statistical secret. (Art.20.3 Law 4/1986 of 23 April - Basque Statistics Law)

This is not a merely trivial issue. In fact, the establishment of a "safe" microdata structure requires the consideration of multiple "intruder" scenarios and many other aspects enumerated below:

- Type of statistics (census or sampling survey)
- Hierarchical structure of the data (i.e.: families and individuals)
- Geographical indicators included
- Identifying variables and possible combinations (identifying keys)
- Level of detail (number of categories) of the variables included
- Sensitive variables included (if any)
- Risk indicator
- Disclosure control methods to be applied
- Information loss measure (Utility measure)

Nevertheless, EUSTAT is in a good position to face this challenge and it has already developed some confidentiality rules for future microdata releases which will be the starting point of this complex task.

3.2 On-site access facility

An alternative to microdata accessibility consists of providing users (mainly researchers) with an 'in-situ' workstation where microdata could be accessed under specific conditions. Recently, EUSTAT has been asked about the possibility of accessing health data in order to perform multivariate analysis and develop a mathematical model to prevent child leukaemia. It is being considered as a pilot experience for a future on-site facility.

4 Conclusions

EUSTAT has been working for a long time on the implementation of a complete data protection system which considers all the phases of statistical production. In fact, a more general report has also been produced this year which considers the treatment of confidentiality in the whole production process, from data collection to physical protection measures and computer security.

However, in this paper we have focused on the dissemination phase and the development of standard criteria to protect statistical products: microdata and tabular data. Reaching an agreement about these confidentiality rules has been a hard process, and the discussion about what should be considered as identifiable or sensitive is still ongoing. However, these criteria should be in continuously updated, reflecting changes in the legal framework, in the technological environment and in social reality.

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.