HELLENIC REPUBLIC

MINISTRY OF ECONOMY AND FINANCE

GENERAL SECRETARIAT OF
THE NATIONAL STATISTICAL SERVICE
OF GREECE

Methods applied for protecting the confidentiality of microdata in the National Statistical Service of Greece[1]

Ioannis Nikolaidis*

1. Introduction

The purpose of data protection is to ensure that no unnecessary data are collected and that confidential data are not revealed to outsiders at any stage of the data processing. The aim of the data protection principle in the case of official statistics is to ensure the availability of exhaustive and reliable data through the maintenance of confidential relations with data suppliers. Data protection requirements define the way that data can be collected, how the data should be processed and in what form they should be published. Thus, data protection influences the procedures that are employed in the compilation of statistics and the content of the data that are published. Data protection represents a central principle of official statistics and a constituent of the external image of statistical authorities

2. Protection of statistical confidentiality

Data collected by the N.S.S.G may only be used for statistical purposes and may not be disclosed only in aggregated form so that no individual references could be extracted.With the view of guaranteeing the principle of independence and the transparency of statistical information, a Committee for the protection of statistical information, the so called Committee for the protection of Statistical Confidentiality (article 8 of the Law 2392/96), has been established within the N.S.S.G with the aim of :

- Ensuring the impartiality and the autonomy of statistical information
- Guaranteeing the compliance of such information with the regulations governing the confidentiality protection of the information, which is supplied to N.S.S.G.
- Ascertaining the quality of the statistical methods and data process techniques being used in the collection, storage and dissemination of data.
- Ascertaining the compliance of the surveys with directives and recommendations of international and Community organizations.

---

[1] Paper for the OECD Conference: "Assessing the feasibility of microdata", Luxembourg, 26-27 October 2006

*Head of the Methodology, Analysis and Research Section, e-mail: giannikol@statistics.gr

3. Access to statistical data

The data gathered by surveys conducted by the N.S.S.G are accessible to the public and are available for study or research purposes to those, who request them in accordance with the provisions of the articles 5 and 6 of the Law 2392/96.

Sample collections of individual data, made anonymous and purged of any reference linking them to individuals (physical or legal persons),may also be provided by the N.S.S.G to public bodies or agencies, legal persons, companies, associations and individual citizens on the basis of whether the request is justified and consent is granted by the Committee of Statistical Confidentiality.

The state registers and archives of the public services and the legal entities of the wider public sector are accessible by the N.S.S.G in order to promote the use of administrative archives and to reduce the of the respondents. In particular, with respect to the use of fiscal data of the enterprises derived from the Value Added Taxes archives, a Joint Ministerial Decision n.12833/C-528/1996 has been signed between the Ministry of National Economy (N.S.S.G administratively belongs to the Ministry of National Economy) and the Ministry of Finance)in which among other things the anonymity and the fiscal secrecy for the enterprises is fully respected.

For the implementation of the Community Directive 95/46/EC, an important national Law (n.2472/97) on the protection of individuals concerning the process of personal data entered into force. It incorporates and applies the principles set up by the above mentioned Community Directive.

According to the article 11 of the Law, the purposes and the modalities of treatment of the personal data gathered must be furnished to the respondent in order to :

- Clarify the aims and the modalities of treatment for which data will be collected
- Specify the compulsory or facultative nature of data release.

Furthermore, according to the article 5 of the Law, the treatment of personal data is allowed only with the explicit consent of the respondent, given in written form. Some provisions of the Law provide for the adoption of preventive and proper safety measures in order to minimize the risks, even accidental of data destruction and loss, of unauthorized access, of not allowed treatment or of non-conformity with purposes of the data collection.

Internal rules for data confidentiality have been adopted by the N.S.S.G. These provide for procedures and methodologies that prevent the disclosure of micro-data concerning households and enterprises. As far as enterprises are concerned these rules imply that it is not possible to provide individual data for less than three units in statistical tables. Anyone, who does not adopt the necessary measures to ensure the safety of personal data, is subject to punishment by imprisonment up to three years and to a fine up to 8.804 Euros, depending on the gravity of the effects of the event.

4. Microdata

Microdata are released for scientific purposes, in a format from which individuals cannot be identified either directly or indirectly. Data identifying statistical units other than individuals similarly are not released. The microdata are released by the Division

of Statistical Information and Editions and only after the opinion of the Committee of Statistical Confidentiality.

In NSSG the main tools for protecting microdata are excluding obvious identifiers, limiting geographic detail and limiting the number o variables on the file as follows:

- Sampling
- Issuing multiple files, one with more detailed geography and less detailed characteristics and other with less detailed geography and more detailed characteristics
- Grouping, by splitting continuous variables into rages to reduce detail
- Grouping and recoding into broad categories
- Eliminating any variables that can be used to link to external sources that contain individual identifiers

4.1 Sampling

Microdata come either from census or a sample. Data from a sample are much more safer since the sample uniqueness is not always population uniqueness and thus the chances of an intruder identifying an individual are low. Using sample data, an important topic is the determination of sample size so that the respondents cannot be identified.

Let N be the population and n the sample size. The sampling fraction $f = \frac{n}{N}$ should be determined so that the identification risk to be the lowest possible. Suppose that the identification key divides the population into K cells.

For one cell, say cell h ($h = 1,2,...,K$), it is necessary to introduce the following notation:

N_h : The population size

n_h : The sample size

w_{hi} : The sampling weight of the i order unit ($i = 1,2,..., n_h$)

$\widehat{N}_h$: The estimated population size from the sampling units. That is:

$$\widehat{N}_h = \sum_{i=1}^{n_h} w_{hi}$$

In the case that, the frequency N_h is not provided to the user or generally it is not known, a cell is considered to be sensitive, when $\widehat{N}_h$<3. Thus, each cell is considered to be safe if and only if

$$\sum_{i=1}^{n_h} w_{hi} \geq 3 \quad (1).$$

From the relation (1), if $n_h = 1$, the cell is considered to be safe when $w_{h1} \geq 3$, and if $n_h = 2$, the cell is considered to be safe when $w_{h1} + w_{h2} \geq 3$

After the application of the relation (1), if the cell is unsafe then suppression or recoding is carried out or in extreme cases the records belonging to the sensitive records are removed, so that the new created cells to be safe.

4.1.1 Sample of population census microdata

In the case that a sample of census microdata is provided to users then:

a. The sample is selected with equal probabilities and $\frac{n_h}{N_h} = \frac{n}{N} = f$

b. The sampling weights are equal to a constant quantity for all sampling units. That is: $w_{hi} = \frac{1}{f} = \frac{N}{n}, \ \forall \ h,i$

c. A cell is considered to be safe when $\widehat{N}_h \geq 3 \Rightarrow \sum_{i=1}^{n_h} w_{hi} \geq 3 \Rightarrow \sum_{i=1}^{n_h} \frac{1}{f} \geq 3 \ \Rightarrow$

$$\frac{n_h}{f} \geq 3 \quad (2)$$

From the relation (2), the determination of f is carried out so that:

$$f \leq \frac{n_h}{3}, \ \forall \ h \ (3)$$

The relation (3) is a condition that ensures the safety of the cell according to the threshold rule ($N_h \geq 3, \forall h$), as following:

$$f \leq \frac{n_h}{3} \Leftrightarrow f \leq \frac{N_h \cdot f}{3} \Leftrightarrow 3 \leq N_h \ (4)$$

As the minimum value of n_h is equal to $n_h^{\min} = 1$, from the relation (3) the maximum value of f is calculated as follows:

$$f^{\max} = \frac{1}{3} \ (5)$$

From the relation (5), in the case that a sample of census micro data is provided to users, the maximum value of the sampling fraction is equal to three.

5. Conclusions

i. The disclosure control method used by the NSSG for the release of microdata is the data reduction (sampling, grouping, coding, suppression). If sampling is applied the sum of sampling weights in each cell is less than three.

ii. Applying the data reduction method, it is necessary to be examined the amount of damage done to microdata, when sensitive cells (defined by identification keys) are suppressed, and thus the data are modified into safe microdata.

iii. In the case that a sample of micro census data is provided to users, the maximum sampling faction is equal to one third and it is avoided providing the frequencies of all cells, which are created from the division of the population from the identification keys.

References

Cochran W. G. (1977). Sampling Techniques. 3rd ed. New York: John Wiley.

Tzougas J. (1999). The national legislation on statistics in Greece. Proceedings of the Joint Eurostat/UN-ECE Work session on Statistical Data Confidentiality. European Communities.

Willenborg L., De Waal T, (1996). Statistical Disclosure Control in Practice. Springer-Verlang , New York, Inc.