

Topic IV: Progress in the implementation of SDC methods and techniques in central and eastern Europe

PROTECTION OF CONFIDENTIAL DATA IN INSSE

Contributed paper

Submitted by INSSE, Romania¹

I. INTRODUCTION

1. In order to carry out its duties related to the dissemination of statistical data, the Romanian National Institute of Statistics and Economic Studies (INSSE) has to deal with confidentiality protection issues, resulting from the growing demand for individual and aggregated data from statistical and external sources, for a great variety of users: Governmental organizations, NGOs and Union Trades, academic and research institutions, banking and financial institutions, mass-media and public, as well as international organizations. Dissemination is done through publications, information from databases, electronic products and Internet.

2. The paper presents the current status and expectations in the field of statistical disclosure control and protection of confidentiality. The main practices used in INSSE for the limitation of statistical disclosure are mentioned, as well as the needs for future improvements of such practices.

II. LEGAL FRAME

3. The Romanian Statistical Law adopted in the Government Ordinance no. 9/1992, last updated in 2000, stipulates that:

- all activities of official statistics conform to the confidentiality principle: data obtained from statistical surveys or from administrative sources or otherwise is protected against disclosure, both for natural persons and legal units;
- statistical data is considered confidential when it allows statistical subjects to be identified, either directly or indirectly, disclosing individual information; as an exception, data referring to legal units, i.e. name, address, type of activity, social capital, turnover and number of employees is not considered confidential;
- confidential statistical information collected through official statistics can be used exclusively for statistical purposes; it cannot be used, neither as evidence in courts of law, nor to establish certain rights or obligations for the statistical subjects;
- statistical data resulting from data processing can only be published or disseminated if, directly or indirectly, the identification of the natural or legal persons is not possible;
- official statistical services must adopt administrative, organizational and technical measures to ensure the protection of individual statistical data; the personnel employed in the official statistical services with access to statistical data is bound to observe its confidentiality.

¹ Prepared by Alexandru Brodeala, Liana Marina and Gabriela Tihohod.

III. INTERNAL REGULATIONS

4. The importance of individual data protection has underlined the need to address data confidentiality topics in a more systematic way. Following the principles stated in the Statistical Law, it was necessary to adopt Internal Regulations, as a set of rules concerning data protection and statistical disclosure control. In order to do that, a Committee on Statistical Confidentiality was set up in October 2000 to:

- gather all internal practices from various domains: business statistics, statistics based on household surveys, demography (including the Population Census), statistics that use external sources;
- propose the Internal Regulation on Confidentiality.

Activities undertaken so far by this Committee are:

- study of methodological papers presented within meetings on Statistical Confidentiality and Disclosure Control;
- inventory of procedures used in-house;
- draft of the Regulations.

5. The Internal Regulations try to provide a policy to cover various aspects of confidentiality such as: organizational, methodological, technical, and staff training measures. Not only data confidentiality treatment, but also the broader topic of data security is approached in this document. The confidentiality policy contained in the Regulations will apply both to the 42 Districts Statistical Offices and the headquarters.

6. Different types of data are treated separately: distinctions are made between natural persons and legal units, between individual and tabular data, between statistical and administrative data. All the activities performed by the statistical office are considered: data collection, data storage and processing, data transfer/communication and data dissemination/release.

7. Some decisions were already taken, e.g.: there is no separate organizational unit with permanent and exclusive responsibilities on confidentiality issues, each production and dissemination unit being involved; updates of the internal Organization and Functioning Regulations follow; this means that some aspects of the Confidentiality rules will be decentralised. Other points are still under discussion, e.g.: whether or not should we consider different degrees of confidentiality (classes) and different categories of protection measures, accordingly.

IV. CURRENT PRACTICES

8. Data collected by INSSE through the statistical surveys is declared to be used only for statistical purposes; this statement is mentioned on all statistical questionnaires carrying confidential information. The respondents confidence in the statistical institute is a very important concern, in order to maintain public trust and future cooperation in surveys.

9. INSSE receives and uses also files with individual data from external sources, e.g. demographic data from the Ministry of Interior (Population division) and from the Ministry of Health, data on education from the Ministry of National Defense; in these cases, the confidentiality rules of the organization which produced the data are added.

10. Individual data, whatever the statistical domain, is not published; access to raw individual data is restricted to the corresponding staff inside INSSE. Microdata files, either about persons / households or enterprises, can be released outside the institute, to academic media and international organizations upon request approved by the INSSE President or based on previous agreements, only after anonymisation; this means several steps to be followed:

- removing the direct identification (such as name, address);
- changing the real identification code for persons, households, enterprises with a surrogate one;

- grouping key variables with many values into broader categories, performing thus variables recoding;
- limiting the number of variables in the file to a minimum (only those which are relevant for the request).

The Personal Identification Number for the natural persons is used only for validation purposes; after performing all the checks, the PIN is removed from the files.

11. As tabular data is concerned, distinctions can be made between methods applied to social and demographic data on one side and business data on the other side.

For aggregated data from social statistics domain, thresholds are used, e.g. in the published results of the Labour Force Survey, only cells with more than 2000 persons or households are shown.

Data from the Population Census (last census in January 1992) does not appear in a table on villages (corresponding to NUTS5 level) if the value of the cell is less than 10 people; the most sensitive variables are the nationality and the religious confession.

For aggregated data on enterprises, each cell value in the table must obey the following rules, otherwise it is not published:

- “cell size” rule: the minimum number of statistical units contributing to the aggregated value in the cell should not be less than 3; in certain circumstances depending on the statistical indicator presented and on the detail degree of the classification used, this limit can be greater than 3;
- “dominance” rule: the cell value is suppressed if one or two statistical units represent at least 80% of the aggregated value.

12. As general comments on the practices used for disseminating individual or tabular data, it should be said that:

- rather than keeping in such data sets as much information as possible, an exaggerated (and unnecessary) data removing is preferred;
- algorithms for protection of confidential values, through adding, altering or replacing records in a micro-data or in a tabular data file, are not used.

13. The “professional secrecy” was a very deeply rooted notion in both statisticians and users minds, for more than 40 years in Romania, but things had to change in the last decade, when INSSE became far more open to the public; nevertheless, a considerable effort is still needed for the improvement of public perception on the quality of statistical data.

14. The analysis and algorithms used for limitation of statistical disclosure are entirely based on personal professional experience and judgement of statisticians, gained mainly as a result of bilateral cooperation with EU countries; there is some computer assistance in implementing such algorithms, but no specialised software is used.

V. FUTURE WORK: NEEDS AND EXPECTATIONS

15. In the next months, INSSE will be focussing on improvement and finalizing of the Internal Regulations on data security, including confidentiality rules; exchanges of experience with other statistical offices that adopted such an internal document would be more than welcome.

16. An important issue in the treatment of confidentiality is to address the confidentiality methodology to a wider extent: methods already in use must be better understood, new techniques must be investigated; e.g., protection of confidential values in tabular data should be refined, paying more attention to problems linked to “secondary confidentiality”. Training on commonly used techniques for evaluation of disclosure risk and estimation models will therefore be needed and encouraged; there is in particular one aspect of the disclosure control strategy that needs to be improved, and that is minimizing the volume of information “lost” as a consequence of the sensitive values protection.

17. An additional mean for maintaining data confidentiality is knowledge about methods implemented by specialised software packages; procurement and training for such software is envisaged.

The increasing Eurostat role must be stressed, in facilitating contacts and discussions between national statistical offices representatives, through meetings, working groups and training measures. Within the frame of statistical pilot projects supervised methodologically by Eurostat, statistical disclosure control procedures should be provided, as an example for the participating countries with less experience in the field.

VI. CONCLUSION

18. With the objective of ensuring data protection and confidentiality, as well as avoiding overprotection of published aggregates, INSSE will put in place the confidentiality policy for data release and dissemination, as a framework for treating systematically the above rules. In parallel, significant progress is expected from an appropriate training programme concerning the methodological aspects of the topic, as well as the practical experiences.