Statistical Confidentiality in European Business Statistics

Aleksandra Bujnowska*

* Eurostat - Statistical office of the European Union, aleksandra.bujnowska@ec.europa.eu

Abstract: aper to be submitted to UNECE/Estat worksession on statistical confidentiality 20-22 Sept 2017 Skopje

1 European statistics and the European Statistical System (ESS)

European statistics are statistics necessary for the performance of the activities of the European Union. They help measure the progress and efficiency of the Union's policies. The annual and multiannual statistical programmes define for which economic or social domains European statistics are necessary[1]. European statistics are in principle produced by Eurostat on the basis of the data transmitted by statistical offices in the EU countries. Eurostat publishes figures for the countries of the European Union (EU), European Economic Area (EEA: EU plus Iceland, Liechtenstein, Norway) and Switzerland, and compiles EU aggregates, including also for the Euro area. The statistical offices in the EU and the EEA countries transmit the data on the basis of specific subject-matter regulations[2]. These regulations define the variables, timeliness, quality and the necessary breakdowns of the data. In most domains the rules require that confidential data are to be sent to Eurostat alongside with non-confidential data[3]. Confidential figures may be used for production of EU aggregates.

[1] Statistical programmes are legal acts laying down priorities concerning the needs for information of the European Union. The current programme covers the period 2013-2017. It was established by the Regulation (EU) No 99/2013 of the European Parliament and of the Council of 15 January 2013 on the European statistical programme 2013-2017. The five-year programmes are backed up by annual programmes that set more detailed objectives for each year.

[2] Transmission of data from Switzerland is covered by a separate bilateral agreement between the EU and the Swiss Government.

[3] Definition of confidential data in the statistical law: 'confidential data' means data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit;

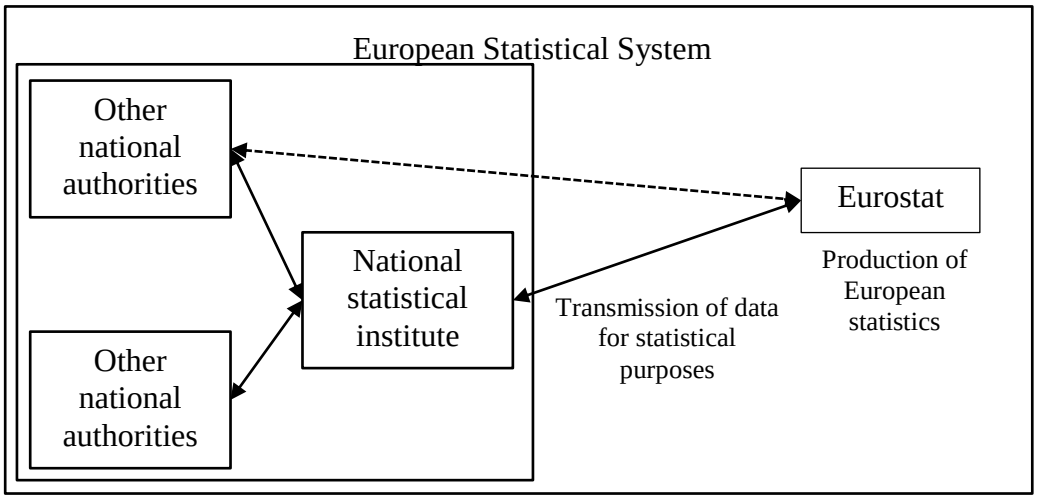

Fig. 1 Members of the European Statistical System.

The European Statistical System consists of all partners involved in the production of European statistics, namely: national statistical institutes (NSIs), other national authorities (ONAs, for example: regional statistical offices, ministries) and Eurostat. Statistical offices in the EU countries (NSIs and ONAs) are often referred to as national statistical authorities (NSAs). The members of the ESS cooperate to produce good quality European statistics on the basis of aligned methods, tools and processes.

2 Business data transmitted to Eurostat

The level of detail of the data transmitted by national statistical authorities (NSAs) to Eurostat is different for different data collections. Usually the data are sent as hypercubes according to dimensions specified in the regulations. A cell in a hypercube is flagged as confidential if there are less than "n" contributors, or one or two contributors dominate. The confidentiality parameters "n" and "k" ("k" is the dominance threshold) are defined by the individual national authorities.

In some domains the microdata are transmitted to Eurostat.

Annex 1 contains the list of standard SDMX flags used to mark the data as confidential or otherwise sensitive.

3 Recommended options for less confidential cells in European business statistics

In several business statistics domains the value of the data is severely undermined by the high share of confidential data cells. In Structural Business Statistics, for individual Member States up to 40% (average 15%) of their data cells are confidential. Out of the EU aggregates 30% would be missing if the aggregates were not rounded as protection measure. For Prodcom (statistics on the production of manufactured goods), up to 83% (average 42%) of the number of codes is confidential for individual EU countries, and 18% of EU aggregates would be confidential without rounding. In Foreign affiliates trade statistics (FATS) on average half of the non-zero data cells are confidential for a country, and one third of the EU aggregates are confidential.

To address this problem, Eurostat, together with ESS Expert Group on Statistical Disclosure Control (EG SDC)[4], in 2016 prepared **Recommendations for confidentiality management in business statistics in the ESS**. The Recommendations propose the following options to reduce missing data in business statistics due to confidentiality: A) optimal table design, B) use of waivers (disclosure permission given by responding business entities) and C) reviewing too strict confidentiality parameters.

These options are described in more detail below.

3.1 Optimal table design

The recommendations say that:

When designing the standard data requirements for European business statistics, careful consideration needs to be given to the level of detail in all the dimensions of the tables, in order to optimise the balance between the desired high level of detail and the resulting confidentiality suppressions especially in small Member States.

While this recommendation seems easy to implement, it may actually lead to more information loss than traditional suppression of confidential figures[5]. Reducing the level of detail of the data published (e.g. from NACE level 3 to 2) obviously increases the number of safe cells but at the same time it reduces the granularity of the table and

[4] The Expert Group on Statistical Disclosure Control brings together around 15 representatives of the national statistical institutes who advise and formulate recommendations to the ESS on statistical confidentiality issues.

[5] Statistics Finland conducted an experiment in Foreign Affiliates Statistics to find out which level of publication detail would be best to balance confidentiality and usability of data. See results here: Soininvaara K., Oinonen T., Nissinen A., "Balancing confidentiality and usability", Statistics Finland Working Papers 3/2014.

therefore its value for users. In the ESS context, the fact that the table design is defined in the relevant regulations adds rigidity and renders it more difficult to find an optimal solution. It is also difficult to define tables that are relevant both at EU and at national level because of differences in the industrial structure of countries. A natural solution would be to allow countries to release less detailed tables for most critical industries. However, this would be difficult to manage at European level, both with regard to the presentation of the results and the calculation of the European aggregates.

3.2 Use of waivers

Recommendations:

The EU Statistical Law[6] allows publishing data with the agreement of the company in question. This agreement could be given already in the questionnaire, or the companies can be asked for their agreement selectively when important data would have to be suppressed. These practices are encouraged.

When feasible in the relevant legislation, even changing the default from active to passive confidentiality could be considered.

In order to facilitate a more widespread deployment of waivers, Eurostat, together with the EG SDC, drafted a paper[7] discussing the key issues, namely:
- the legal basis,
- costs and benefits,
- different ways to apply waivers: systematic (e.g. an option to select in the standard questionnaire[8]) or targeted (only the biggest contributors are approached),
- the content of waivers (duration, variables covered etc.),
- methodological issues (which confidentiality rule to use for data covered by waivers; see annex 2).

In the various discussions[9] on waivers several use cases have been shared. It seems that the best results can be achieved by asking for waivers systematically from a larger number of critical companies. Survey-based statistical domains, where there are many missing values due to confidentiality, are ideal candidates for waivers. The implementation of waivers requires some organisational and technical investments.

[6] Regulation (EC) No 223/2009 on European statistics.
[7] Ala-Kihnia J., Bujnowska A., Kloek W.; "Use of waivers to reduce confidentiality suppressions in business statistics", Eurostat, NTTS2017 paper, session 8B.
[8] A tick-box is one way to do but is not deemed feasible in all cases on national level.
[9] The paper on waivers was discussed by the various stakeholders groups: ESS Working Group on Methodology and ESS Business Statistics Directors' Group bringing together high level representatives of the national statistical institutes.

The views on waivers differ across countries and across statistical domains. The discussions very often lead to passive confidentiality, which allows the publication of figures (even if it leads to disclosure of individual information) unless protection is explicitly requested by the statistical unit (see comparison in table 1). Passive confidentiality is relatively easy to manage and results in a smaller number of cells suppressed due to confidentiality. Its propagation is not straightforward though. In most countries passive confidentiality is not applied outside the domain of international trade statistics. Passive confidentiality in trade statistics is explicitly recognized by EU statistical law.

Table 1 Comparison of the different approaches to statistical confidentiality

Approaches to confidentiality	**Standard approach**	**Exceptions**	
	Active confidentiality	**Agreement of statistical unit (waiver)**	**Passive confidentiality**
By default data is:	Confidential	Confidential	Non-confidential
Do statistical offices need to ensure that published statistics do not lead to disclosure of information on an individual statistical unit?	Yes, always	Yes, but if the statistical unit agreed, its data can be disclosed	No, only if the statistical unit requested so, its data need to be protected

3.3 Confidentiality parameters and methods

The recommendations suggest revising confidentiality parameters "n" (minimum number of cell contributors) and "k" (maximum share of contribution to the cell by one or two biggest contributors) in order to maximise the potential of the data to be disseminated. The recommendation aims at limiting the range of the reference parameter values used in EU countries to protect tabular data in business statistics.

A revision of confidentiality parameters was also discussed intensively with NSAs. In some cases the change of threshold levels for frequency or dominance rules led to more figures being published. But many NSAs claimed that the parameter values are already set at the optimal level and cannot be modified any further.

4 Other options

In addition to the options in the Recommendations, some other proposals aiming at reducing the number of confidential cells in business statistics could be suggested. One is a time limitation for confidential data. Business data becomes less sensitive after some years. Hence, a period of validity of statistical confidentiality could be agreed in the ESS on the basis of current practices and a consultation with users in the EU countries. The end of the confidentiality status of a cell must be carefully decided in order not to disclose the protection method and parameters.

Another suggestion that could lead to fewer confidential cells is to enforce the use of confidentiality flags only in case of statistical confidentiality. Currently, data is flagged by NSAs as confidential for several reasons and any cell flagged as confidential must be hidden and its recalculation made impossible. Instead, data that is "not to be published due to insufficient quality" (or for any other reason) should be flagged as such rather than as confidential so that it may be used for recalculations and therefore secondary confidentiality does not apply.

A similar issue is flagging of data as confidential due to its sensitive nature. This should be avoided if the transmission of data to Eurostat is based on European Regulations. The sensitivity of the variable is no argument in establishing whether or not a cell is statistically confidential.

5 Confidentiality charters

The recommendations propose options for reducing the number of primary confidential cells. They are mainly addressed to NSAs. At Eurostat level SDC focuses on the release of safe EU aggregates. Normally the inside part of the table (countries' contributions) can not be changed. Only the data that are flagged as "confidential"/"not to be published" can be suppressed. Other data might have been published already nationally and therefore should remain public. This additional constraint leads to the suppression of many EU aggregates.

The Eurostat confidentiality charters lay down rules for the publication of EU aggregates composed of confidential national figures. The standard model can be adapted to the needs of the particular statistical domains.

The information needed to treat confidentiality at EU level is at least the confidential data itself. If, in addition, the reasons for confidentiality (e.g. number of statistical

units in the cell, shares of the first and the second largest contributors etc.) are available, Eurostat can determine the confidentiality of the EU aggregates more accurately. If the information about the number of individual contributors to a confidential national cell is unknown the national cell value is to be considered as based on one contributor.

The more information is received from NSAs about the nature of confidential cells, the easier it is to take a decision on the publication of EU aggregates.

To facilitate the publication of EU aggregates, the confidentiality charter allows including in the confidentiality cluster[10] data not to be published due to insufficient accuracy.

6 EBS Manual

National statistical authorities and Eurostat are developing the Framework Regulation Integrating Business Statistics (FRIBS). FRIBS aims to streamline and rationalise the reference framework for European business statistics, reducing unnecessary statistical burden on respondents.

In parallel to the legal drafting, a manual describing current practices and guidelines has been written. The main purpose of this European Business Statistics (EBS) manual is to serve as an umbrella for the metadata and methodologies of business statistics, primarily targeting data compilers, but also interested end-users. Chapter 17 deals with statistical disclosure control. The manual has been published as a dynamic online Statistics Explained publication and as a static pdf publication. The pdf version is available here.[11]

7 Conclusions

The European Statistical System observes growing difficulties with protecting business statistics. This is due to business concentration, internationalisation and the complexity of business structures. Eurostat recommends to the ESS members to review and to revise their approaches to confidentiality in view of publishing more cells, if possible, and without compromising the privacy and business interests of individual data providers. The recommendations were developed to define lines along which such a revision can be organised. The conclusion after several rounds of discussions on the recommendations is that there is no simple solution that could

[10] The confidential cluster is the group of countries contributing to an EU aggregate and whose data is confidential for a particular variable.
[11] http://ec.europa.eu/eurostat/documents/54610/7779382/EBS-manual-table-of-contents-and-introduction.pdf

remedy the problem of high numbers of confidential cells. A mix of options adapted to the national business structure should be tried and implemented in line with local rules and practices.

References

Soininvaara K., Oinonen T., Nissinen A., "Balancing confidentiality and usability", Statistics Finland Working Papers 3/2014.

Ala-Kihnia J., Bujnowska A., Kloek W. "Use of waivers to reduce confidentiality suppressions in business statistics", Eurostat, NTTS2017 paper, session 8B.

Annex 1

SDMX STATISTICAL GUIDELINES

SDMX Cross-Domain Code Lists

Name: Code list for Confidentiality Status (CONF_STATUS).

Description: This code list provides coded information about the sensitivity and confidentiality status of the data.

Established international standard used as input for the code list: None.

Version: 1.1.

Date: 26 June 2014.

Recommended code value	Recommended code description	Annotation
F	Free (free for publication)	Used for observations without any special sensitivity considerations and which can thus be freely shared. Usually, source organisations provide information and guidance on general requirements for re-dissemination (like mentioning the source) either on their websites or in their paper publications. In some institutional environments the term "unclassified" is used in a sense that still denotes implied restrictions in the circulation of information. If this is the case, the organisations concerned may probably consider that "free" (value F) is not the appropriate tag for this kind of "unclassified" category and that "Not for publication, restricted for internal use only" (value N) may be more appropriate.
N	Not for publication, restricted for internal use only	Used to denote observations that are restricted for internal use only within organisations.
C	Confidential statistical information	Confidential statistical information (primary confidentiality) due to identifiable respondents. Measures also should be taken to prevent not only direct access, but also indirect deduction or calculation by other users and parties, probably by considering and treating additional observations as "confidential" (secondary confidentiality management).
D	Secondary confidentiality set by the sender, not for publication	Used by the sender of the data to flag (beyond the confidential statistical information) additional observations in the dataset so that the receiver knows that he/she should suppress these observations in subsequent stages of processing (especially dissemination) in order to prevent third parties to indirectly deduct the observations that are genuinely flagged with "C".

S	Secondary confidentiality set and managed by the receiver, not for publication	If senders do not manage the secondary confidentiality in their data and/or there are also other countries' data involved (with the intention to eventually compile a regional-wide aggregate that is going to be published), the value "S" is used by the receiver to flag additional suppressed observations (within sender's data and/or within the datasets of other senders) in subsequent stages of processing (especially, dissemination) in order to prevent third parties to indirectly deduct the observations that were genuinely flagged with "C" by the sender.
A	Primary confidentiality due to small counts	A cell is flagged as confidential if less than m units ("too few units") contribute to the total of that cell. The limits of what constitutes "small counts" can vary across statistical domains, countries, etc.
O	Primary confidentiality due to dominance by one unit	Used when one unit accounts for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc.
T	Primary confidentiality due to dominance by two units	Used when two units account for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc.
G	Primary confidentiality due to dominance by one or two units	Used when one or two units account(s) for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc.
M	Primary confidentiality due to data declared confidential based on other measures of concentration	Cells declared confidential using mathematical definitions of sensitive cells, e.g. p-percent, p/q or (n,k) rules.

Annex 2

Why should a (2,k)-rule be replaced by a p%-rule, if waivers are used?

The p%-rule and the (2,k)-dominance rule are similar. Both rules take into account that the two biggest contributors A and B are best placed to estimate each other's value, knowing the subtotal T A (or T B) by deducting their own contribution. Using this subtotal for an estimate both contributors would overestimate the contribution of the other one by T A B. To determine which cell should be suppressed both rules take into account the precision of the estimate T A B.

The p%-rule compares the estimation error T A B to the value to be estimated, e.g. either B (if A is assumed to estimate B) or A (if B is assumed to estimate A). Notably, in a situation where A is much larger than B, the estimation error T A B can be relatively small compared to A and at the same time relatively big compared to B.

The (2,k) dominance rule, on the other hand, compares the estimation error to a percentage (100 k) of the total T. In this comparison, it does not matter, if B is assumed to estimate A, or A is assumed to estimate B. However, if B is relatively small, the estimation error T A B can be large relative to the contribution of B; A is not able to make an accurate estimate of B. To avoid over-suppression, a proper rule should not flag a cell, if the estimation error T A B is large compared to the objective of the estimation (e.g. the contribution B).For this reason it is recommended that countries using (2,k) dominance rule replace this rule by a p%-rule, if they use waivers. According to the SDC Handbook 4.2.1, p. 122, the natural choice for the parameter p of the rule is then setting p to (100-k)/k 100.