

STATISTICAL DISCLOSURE CONTROL

Preface

This chapter introduces the term "statistical disclosure control" (SDC) and presents the various methods that may be applied by statistical offices to protect data released by statistical institutes in tables or as microdata.

The chapter is part of the [European Business Statistics Manual](#) and focusses as such on the protection of business data.

Contents

1. [European Statistical System and European statistics](#)
2. [Legal rules on SDC](#)
 - 2.1. [Definition of "confidential data"](#)
 - 2.2. [Legal obligation to protect confidential data](#)
 - 2.3. [Exceptions to the obligation to protect statistical unit' confidentiality](#)
 - 2.4. [Sending confidential data for statistical purposes](#)
 - 2.5. [Use of data from public sources](#)
 - 2.6. [European Statistics Code of Practice](#)
3. [Statistical units in business statistics](#)
4. [SDC in different types of output \(tables, microdata files, regression models, graphs, maps etc.\)](#)
5. [SDC rules and methods for tabular data](#)
 - 5.1. [Identifying confidential cells in tables](#)
 - 5.2. [Protecting of confidential cells in the tables](#)
 - 5.3. [Protecting EU aggregates](#)
6. [SDC for business microdata](#)
 - 6.1. [Disclosure risk in the microdata files](#)
 - 6.2. [SDC methods for microdata](#)
 - 6.3. [Microdata protection process](#)
 - 6.4. [Information loss due to microdata protection](#)

7. [SDC tools](#)
8. [Confidential data transmission between NSAs and Eurostat \(flags and meta-information\)](#)
9. [SDC aspects in confidential data exchange](#)
10. [Risk management \("five safes" model\)](#)
11. [See also](#)
12. [Further Eurostat information](#)
13. [External links](#)
14. [Contact](#)

1. European statistical system and European statistics

According to Regulation (EC) No 223/2009 on European statistics:

- The European statistical system is the partnership between Eurostat and the bodies responsible for developing, producing and distributing European statistics in each Member State.

These national statistical authorities (NSAs)¹ comprise national statistical institutes and other national authorities.

- European statistics are those statistics needed by the EU to perform its activities. They are determined in the [European statistical programme](#).

2. Legal rules on SDC

Statistical confidentiality is a fundamental principle of official statistics enshrined in the Treaty.

Article 338 of the Treaty on the functioning of the European Union

1. Without prejudice to Article 5 of the Protocol on the Statute of the European System of Central Banks and of the European Central Bank, the European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall adopt measures for the production of statistics where necessary for the performance of the activities of the Union.

2. The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and **statistical confidentiality**; it shall not entail excessive burdens on economic operators.

¹ [The European Statistical System](#)


Commission Regulation (EC) No 223/2009 on European statistics defines confidential data and its Chapter V covers statistical confidentiality. The parts relevant for statistical disclosure control in business statistics are described below.

2.1 Definition of "confidential data"

Article 3 "Definitions" of Regulation (EC) No 223/2009 defines "confidential data" in the following way:

'Confidential data' means data which allow a statistical unit (i.e. the person, company or organisation to which the data refers) 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.

 The risk of a statistical unit being identified is the *only* factor that qualifies data as confidential. It is not important *which* information is disclosed and if this information is sensitive or not.

2.2 Legal obligation to protect confidential data

The legal obligation of the members of the European Statistical System to protect confidential data is emphasised in Chapter V of Commission Regulation (EC) No 223/2009 on European statistics. Article 20 "Protection of confidential data" says that:

Within their respective spheres of competence, the NSIs and other national authorities and the Commission (Eurostat) shall take all necessary regulatory, administrative, technical and organisational measures **to ensure the physical and logical protection of confidential data (statistical disclosure control)**.

The NSIs and other national authorities and the Commission (Eurostat) shall take all necessary measures to ensure the **alignment** of principles and guidelines as regards the physical and logical protection of confidential data.

Physical protection of confidential data refers to different aspects of data security. National statistical authorities must make sure that only authorised users in the office have access to confidential data.

Logical protection means the authorities must check whether the published statistics allow the statistical unit to be identified. The domain of statistics which proposes methods for such checks is called "statistical disclosure control".

2.3 Exceptions to the obligation to protect statistical units' confidentiality

There are only two exceptions where identifiable data can be published (Article 20 of Regulation (EC) No 223/2009):

Statistical results which may make it possible to identify a statistical unit may be disseminated by the NSIs and other national authorities and the Commission (Eurostat) in the **following exceptional cases**:

- (a) where specific conditions and modalities are determined by an act of the European Parliament and of the Council acting in accordance with Article 251 of the Treaty and the statistical results are amended in such a way that their dissemination does not prejudice statistical confidentiality whenever the statistical unit has so requested; or
- (b) where the statistical unit has unambiguously agreed to the disclosure of data.

Exception (a)

This refers to "**passive confidentiality**". Where allowed in a separate legal act, the NSAs do not have to protect the data against identification of statistical unit unless explicitly requested by the importer or exporter.

This measure is used in trade statistics². In the other domains of business statistics NSA apply "active confidentiality". Active confidentiality means that statistical units do not have to explicitly ask the NSAs to have their data protected; NSAs have to protect the data of all statistical units.

Exception (b)

This refers to the situation opposite to "passive confidentiality". It says that, by default, all data is confidential, but if a given statistical unit explicitly agrees, the individual data referring to them may be disclosed. This approach is sometimes referred to as "waivers approach". Table 1 compares different approaches to statistical confidentiality.

Table 1 Comparison of the different approaches to statistical confidentiality

Approaches to confidentiality	Standard approach	Exceptions	
	Active confidentiality	Passive confidentiality	Agreement of statistical unit
By default data is:	Confidential	Non-confidential	Confidential
Statistical offices need to ensure that published statistics do not lead to disclosure of information on individual statistical unit?	Yes, always	No, only if the statistical unit requested NSAs to protect its data and only for the data provided by this statistical unit	Yes, but if the statistical unit agreed to disclosure, its data can be disclosed

² See for example Article 10 of Regulation (EC) No 471/2009 of the European Parliament and of the Council on Community statistics relating to external trade with non-member countries.

2.4 Sending confidential data for statistical purposes

Regulation 223/2009 allows confidential data to be sent between members of the ESS if *"this transmission is necessary for the efficient development, production and dissemination of European statistics or for increasing the quality of European statistics"* (Article 21 (1) of the Regulation)".

In practice, domain-specific legal acts define at which level of detail data is sent between ESS authorities, especially between NSAs and Eurostat:

- In some domains (Structure of Earnings Survey), unit level data (microdata) is transmitted.
- In many other domains (e.g. SBS), data is sent according to predefined breakdowns (aggregates). If the breakdowns are very detailed, this data may also be disclosive (may allow the individual contribution made by a statistical unit to be recalculated). NSAs transmit the data together with relevant flags allowing Eurostat to know how to treat it.

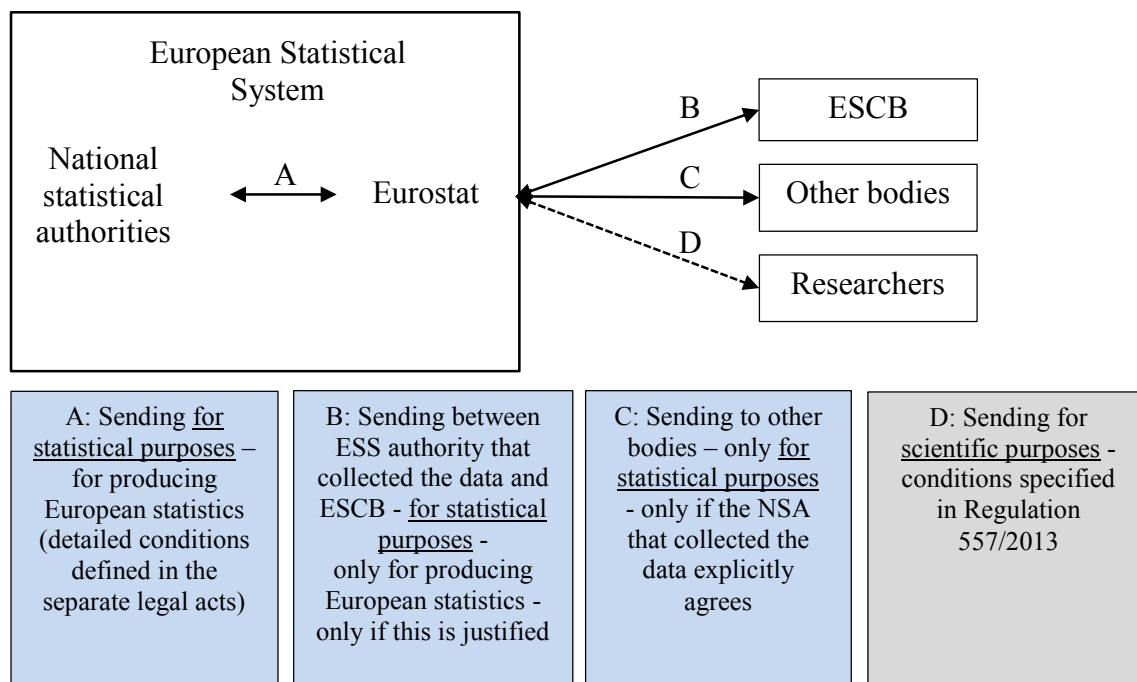
Article 21 allows confidential data to also be sent beyond ESS. The data may be transferred between ESS authority that collected the data and the member of European System of Central Banks (ESCB) if it is *"necessary for the efficient development, production and dissemination of European statistics or for increasing the quality of European statistics, within the respective spheres of competence of the ESS and the ESCB, and (if) this necessity has been justified."*

Any further sending of confidential data to other bodies and for statistical purposes requires explicit authorisation from the authority that collected the data.

Figure 1 represents the different conditions for transferring the data between the different bodies. The figure is a simplification. It does not show transfer of data between statistical authorities inside a country and it also does not show the exchange of data between national authorities from different countries for statistical purposes (see paragraph 9).

ESS data transmission to researchers is covered by separate legislation (see chapter on microdata service for researchers), which includes a strict authorisation procedure and requires NSAs to be consulted. Some organisations can be in different roles; a research department of an NSA or a Central Bank can also be recognised for access to microdata for research purposes.

Figure 1 Sending confidential data in the European Statistical System



2.5 Use of data from public sources

Article 25 of Regulation 223/2009 says that "*data obtained from sources lawfully available to the public and which remain available to the public according to national legislation shall not be considered confidential for the purpose of dissemination of statistics obtained from those data*".

This article applies when data has been collected from publicly available sources. It allows NSAs to publish the data even if publication would allow statistical units to be identified. This article has been applied in the case of business registers, for example. In some countries some basic information about companies is public and so does not need to be hidden in statistics distributed by NSAs.

2.6 European Statistics Code of Practice

The principle of statistical confidentiality is interpreted in the Code of Practice as the obligation on NSAs to guarantee the privacy of data providers (households, companies, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes.

Compliance with this obligation should be measured with reference to the following indicators:

- Statistical confidentiality is guaranteed in law.
- Staff sign legal confidentiality commitment on appointment.
- Penalties are prescribed for any wilful breaches of statistical confidentiality.

- Staff is given guidelines and instructions on the need to protect statistical confidentiality when producing and disseminating data. The confidentiality policy is made known to the public.
- Physical, technological and organisational provisions are in place to protect the security and integrity of statistical databases.
- Strict protocols apply to external users accessing statistical microdata for research purposes (in ESS, access to confidential data can be granted "*to researchers carrying out statistical analysis for scientific purposes*"³).

3. Statistical units in business statistics

Statistical unit means the basic **person, household or organisation to which the data refers** (officially: "*a basic observation unit, namely a natural person, a household, an economic operator and other undertakings, referred to by the data*"⁴).

In producing business statistics, the term statistical unit is used for the different units that statistical output refers to⁵; here the statistical unit is defined from an economic point of view. For the statistical description of the different economic processes (e.g. production, financial), separate statistical units are used – e.g. the **kind-of-activity unit**, the **enterprise** and the **enterprise group**.

All these units consist of one or more natural persons or legal persons; the legal entities are the building block for the system of statistical units used in business statistics. There is also a strict relationship between the unit concepts: an enterprise group consists of one or more enterprises and an enterprise consists of one or more kind-of-activity units.

For most companies, the 3 types of statistical unit coincide; for many others, the unit structure is simple. This has 2 implications for statistical disclosure control:

- The use of statistical units to produce economic statistics makes it a bit more difficult for an intruder to link information to concrete legal entities, as the structures of big companies are complex, dynamic and international (the enterprise group will be more easy to identify than the enterprise).

This case will not be dealt with in this text, which will assume that the link between the statistical unit and legal units is public and known.

³ [Article 23 of Regulation \(EC\) No 223/2009](#) on European statistics.

⁴ Definition in Regulation (EC) No 223/2009 on European statistics and [Regulation \(EEC\) No 696/93](#) .

⁵ See the separate chapter on [statistical units](#).

- The assessment of disclosure risks cannot be limited to one type of statistical unit, but should consider all unit types that present the same or similar variables in an integrated way. This creates more complex disclosure scenarios⁶.

The disclosure risk is not necessarily in the observed company; product information, for instance, may also lead to disclosure of big distributors or suppliers that did not report the information themselves.

4. SDC in different types of output (tables, microdata files, regression models, graphs, maps etc.)

*"Statistical disclosure control (SDC) means methods to reduce the risk of disclosing information on statistical units, usually based on restricting the amount of, or modifying, the data released"*⁷.

Traditionally SDC methods were associated with protecting tables. The tables are nowadays complemented by graphs, maps, models and the like. The approaches to SDC differ depending on the input data and how it is presented. In this chapter we focus on SDC methods for tables and for protecting microdata.

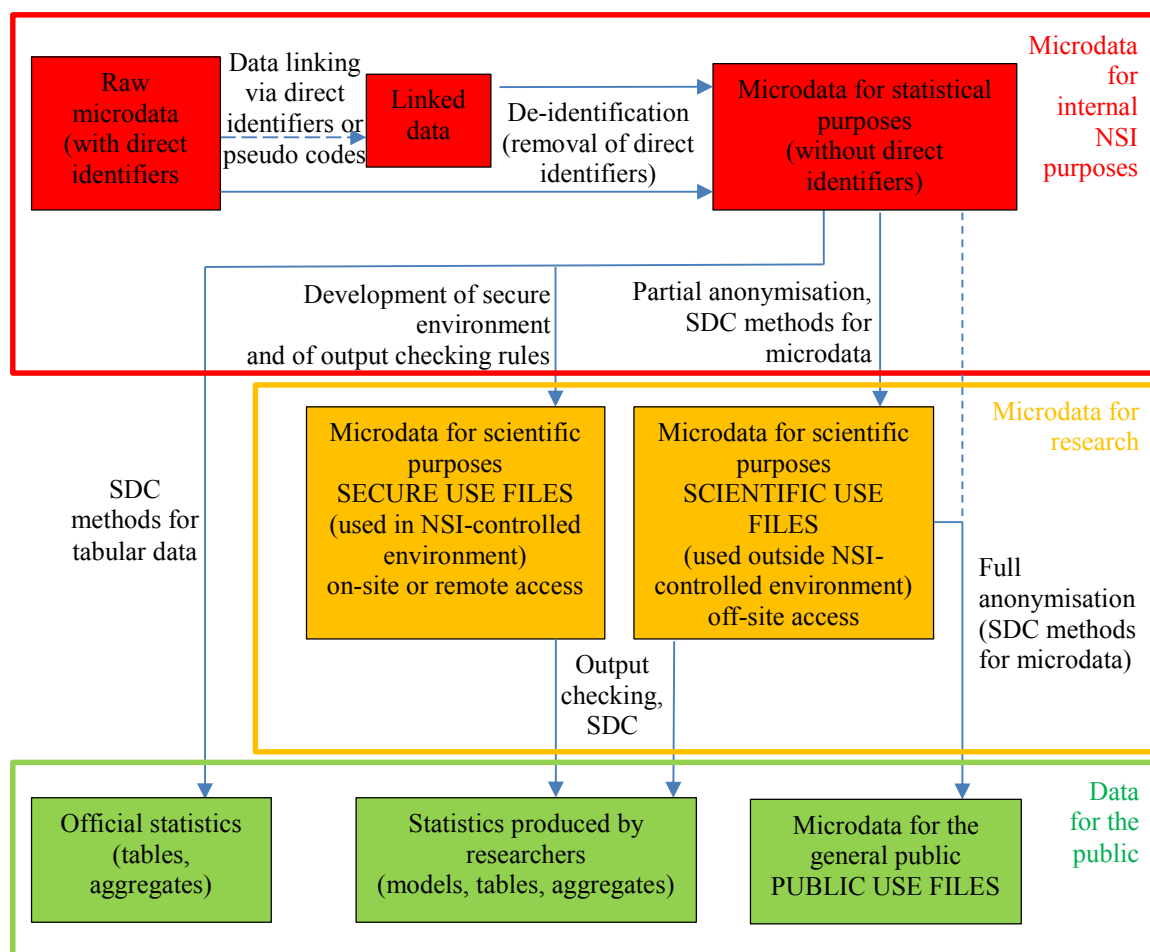
The SDC methods need to be consistently applied, taking into account the links between the data. Consistency needs to be ensured between linked tables and between different types of data presentation⁸.

⁶ Disclosure scenario: the means, motives and opportunities a potential intruder has to disclose confidential information (e.g. what information does the intruder have and how can this information be linked).

⁷ Definition in [Regulation \(EU\) No 557/2013](#) on access to confidential data for scientific purposes.

⁸ The obligation to apply the SDC methods before the data is published applies not only to statistical institutes but also to researchers and other users with lawful access to confidential data. NSIs typically apply SDC when they publish tables and/ or produce microdata. The output produced by researchers/other users need to be verified for SDC as well.

Figure 2 Place of statistical disclosure control in the process of producing statistics (tables), microdata files and other types of output



5. SDC rules and methods for tabular data

Statistical disclosure control for tabular data consists of 2 steps:

- 1) identifying the cells with disclosure risk
- 2) applying protection methods.

At the end of this section is a short presentation on the confidentiality of European aggregates calculated by Eurostat.

5.1 Identifying confidential cells in tables

Tabular data reports on categories of companies with similar characteristics (e.g. economic activity and size class). Publication of rare combinations should be avoided, as the statistical units can be easily identified.

National legislation and traditions differ slightly, but it is always required that the value in the table represents at least 3 observations. Values based on 1 or 2 observations are confidential.

In business statistics, the distribution of the target variable is often skewed: there are only a few big companies and also within a particular cell, 1 or 2 companies might be dominant. This would make it easy to disclose the information on the dominant company with a high level of accuracy. For this reason also, cells with dominant companies are confidential.

The rules used to determine the confidentiality of a cell are summarised in table 2.

Table 2 Confidentiality rules

Rule	Cell is confidential if
Threshold rule	The number of contributors is less than a pre-specified threshold (e.g. 5).
(n,k) rule / dominance rule	The n largest contributions to that cell make up for more than k% of the cell total (e.g. if the rule is (2,90), the cell is confidential if the two biggest contributors to the cell account for more than 90% of the turnover of the whole cell)
p % rule	A contributor to that cell is able to derive an estimate of some other contributor to the same cell within p% of its true value

Note: adapted from the Memobust handbook

Primary confidentiality

The term primary confidentiality is used for cells identified as confidential using these rules.⁹ Treating only the cells identified by these rules is usually not sufficient, as in a table with totals, the confidential cell can be re-calculated by taking the difference.

Secondary confidentiality

Secondary confidentiality is treating a non-confidential cell as confidential, to prevent disclosure of a confidential cell, by making it impossible for a user to recalculate the values of confidential cells¹⁰.

For example, you want to publish the following information on the mining industry (table 3).

⁹ **Primary confidentiality** – concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are: too few units in a cell and dominance of 1 or 2 units in a cell. Source: Handbook on Statistical Disclosure Control, version 1.2, Jan 2010


¹⁰ **Secondary confidentiality** – to reach the desired protection of confidential cells, it is necessary to treat additional non-confidential cells.

Table 3 Example: need for secondary confidentiality (mining for metal ores)

Industry	Production
Mining of iron ores	666
Mining of uranium and thorium ores	confidential
Mining of other non-ferrous metal ores	222
Mining of metal ores (= total)	999

The confidential cell will only remain confidential if one of the other (non-confidential) values is also treated as confidential. The usual solution is to keep the higher level aggregates, where possible by treating another cell as secondary confidential (e.g. *Mining of other non-ferrous metal ores*).

Totals are the easiest and most frequent example of relations between data that can lead to disclosure. Totals can occur at several levels in a classification (e.g. the NACE classification of economic activity) and in tailor-made groupings of categories (e.g. groupings of innovative or labour intensive activities). The need for secondary confidentiality also occurs with other mathematical relationships (e.g. ratios).

 In determining whether a cell should be confidential, bear in mind the effect of **publication at other times** (either earlier or possibly in future).

Example

Suppose that in table 3 above the total (mining of metal ores) in the bottom line was omitted. In this case, there would be no need for secondary confidentiality.

But publishing the data in this way blocks you from publishing the total mined amounts for metal ores in any future table.

So it is good to review the whole publication programme in an integrated way.

Consider also publications ***on the same subject for other reference periods***. If time series are important, the same approach should be kept as much as possible. This goes especially for confidentiality patterns where cells are suppressed, as a value for a confidential cell from some other reference period is usually a good basis for estimating and thus disclosing information.

In short-term statistics, the disclosure risk is reduced by publishing data in the form of growth rates.

Pre-tabular treatment (microdata)

Until now we have assumed that confidentiality will be treated table by table or as a set of tables. An attractive alternative is to treat the **microdata file** that is the basis for the tables.

This approach guarantees that all possible tables are treated the same way.

The methods available for this are partly the same as those described for protecting tabular data below: **redesign** (especially reducing in size), **suppression**, **rounding** and adding **noise**.

Some specific methods are described in section 6.3 *Microdata protection*: record swapping, micro aggregation and post-randomisation (PRAM).

However, it is not feasible to produce a microdata publication file that would only create safe tables, without significant loss of information. The current solution is to build some basic protection into the microdata file, and to add additional protection to tables where needed.

5.2 Protecting confidential cells in the tables

A well-known method for protecting confidential cells is to suppress them, so the values appear as missing. This rather crude method is safe, but it undermines the usability of the data. Missing values usually interrupt the transformation of complete tables.

Table 4 below shows an overview of protection methods. Many of them are built on the idea of adding some noise to the information, in order to still offer a table without missing values.

Table 4 Most frequently used SDC protection methods for tables

Method	Type of table	Description
Table redesign	Magnitude or frequency	Collapsing rows and/or columns
Cell suppression	Magnitude or frequency	Completely suppress the value of some cells (put a “cross’) ⚠️ <i>This does not suppress all information, as an interval can be established</i>
Rounding • Controlled • Deterministic • Random	Magnitude or frequency	Round each cell value to a pre-specified rounding base
Controlled tabular adjustment	Magnitude	Selectively adjust cell values: unsafe cells are replaced by either of their

		closest safe values. Other cell values are adjusted to restore additivity.
Perturbation	Magnitude	Add random noise to cell values.

Note: adapted from the Memobust handbook

Combining methods

If a table contains many confidential cells, you should consider redesign as a first step, possibly followed by rounding. But this might still be insufficient for some very large companies; suppression seems to be the only alternative.

Protecting a few very large companies can undermine the usefulness of the information. An effective solution is to come to an agreement with the companies concerned about the way they can be represented in statistics ('waivers'). As creating and maintaining waivers is expensive, there is a natural limit to this approach.

5.3 Protecting EU aggregates

Eurostat is responsible for calculating the European aggregates from the national data provided by the national statistical authorities.

This also makes Eurostat responsible for the confidentiality treatment of the newly derived information. Although the ideas behind the approach remain the same, the situation is somewhat different:

- the data published by the national statistical authorities are a given; Eurostat cannot suppress or add noise to data already published at the national level.
- Eurostat is dependent on the background information provided about the national confidential cells. This varies by domain; in some, the Regulation requires Member States to provide the values for the confidential cells as well as full information on why the cell is classified as confidential. At the other extreme, neither the values nor the justification are provided; the cells just appear as missing with a confidentiality flag.

If the confidential data is provided with full background information, the European aggregates can be calculated and their confidentiality determined more or less as in the national case. As the national rules and approaches are not aligned, it is important to agree on a transparent approach on how confidentiality rules will be applied at European level. This can be documented in a Confidentiality Charter.

Example

The total for the EU is built-up from 28 national values. Of these national values, 2 are confidential and therefore suppressed in the national publication:

- country A uses a threshold of 3 and the actual number of contributors is 2
- country B uses a threshold of 5 and the actual number of contributors is 4.

At European level, a threshold of 5 is agreed; the European total can be published as the missing value is based on 6 contributions. The example would be more complex with dominance rules, but the approach remains the same.

If the confidential data is not provided, the data can be treated as missing at European level. Eurostat could estimate data missing at national level as input to an estimated European aggregate (flagged as a Eurostat estimate).

6. SDC for business microdata

Business data is usually presented in tables. But some NSAs (and Eurostat) also offer access to business microdata for scientific purposes¹¹.

In the microdata file, each statistical unit is represented in a separate record. As some statistical units are very easily identifiable, microdata need to be protected, to make the identification (recognition) of statistical unit more difficult (or impossible).

There are different stages in the microdata preparation/protection process, which result in different types of microdata files (*see table 5 below for comparison*):

1. **De-identification or pseudonymisation** – the process of removing direct identifiers (like name, ID and address) from the confidential data (and replacing them with pseudo names). Microdata used by the NSAs to produce statistics and secure use files are prepared in this way¹².
2. **Partial anonymisation** – application of SDC methods on de-identified microdata, to reduce the risk of the statistical unit being identified. Scientific use files are the result of partial anonymisation.
3. **Complete anonymisation** – application of SDC methods to completely eliminate the risk of identification for the statistical unit (directly or indirectly)¹³. Public use files contain completely anonymised records.

¹¹ In some countries, release of business microdata is forbidden by law.

¹² See table 5 below and the microdata access chapter for the definition of the 3 types of files: **secure use** files, **scientific use** files and **public use** files.

¹³ In some countries, anonymisation may mean something different than making the data completely anonymous (non-confidential). For instance, in Germany anonymisation usually means de-identification.

Table 5 Microdata protection and resulting types of microdata files

Microdata protection	Level of disclosure risk in the resulting files	Resulting files	Users
De-identification	High risk	Microdata for statistical purposes	NSAs staff
		Secure use files Files to be used in the physical (safe centre access) or virtual (remote access) environment controlled by NSAs	Eligible researchers
Partial anonymisation	Medium/Low risk (risk reduced), the actual level depends e.g. on the sensitivity of response variables	Scientific use files Files that can be used outside NSA's secure environments; the security of the data is the responsibility of the data receiver (researcher)	Researchers
Complete anonymisation	Risk eliminated under predefined attacker scenarios	Public use files Files prepared in such a way that the statistical units cannot be identified. <i>This type is practically non-existent for business microdata, it would result in too much information loss)</i>	All

In this section, we present the major factors that have to be considered when business microdata is released for scientific purposes. The reference material provides further details on the microdata protection methods.

6.1 Disclosure risk in the microdata files

Disclosure risk is the risk that a particular statistical unit is identified and some new information about it is disclosed.

This goes beyond *identity risk* – the risk of identifying the statistical unit without disclosing new information.

There might be some additional ambiguity in cases where the statistical unit consists of more than one legal unit (see paragraph 3). But such cases will not be addressed in this text, which will assume the link between the statistical unit and legal units is public and known.

Analysis of disclosure

Disclosure risk is higher in business microdata than in social (population) microdata. This is because business data is usually much more skewed than personal data – big companies are normally very visible in a microdata set. A lot of information on companies is public and this can be used to identify them in a data set.

Disclosure of business data can be driven by economic incentives: gaining crucial market information, notably on competitors, suppliers or costumers.

This makes it very hard to make a safe business microdata set¹⁴. Another important factor is that business data is not always sampled. Sampling provides additional ambiguity about recognition. If there is no sampling, a user can be relatively sure that the company they are interested in is contained in the data.

When preparing the microdata file, you also have to take into account the corresponding **target audience**. If the files are prepared for scientific purposes, the permissible disclosure risk may be higher, because researchers are considered to be trusted users:

- they sign the necessary commitments to get access to confidential data.
- any failure to respect confidentiality rules would have a negative impact on their reputation and that of the organisation they belong to.
- they are not normally interested in individual information.
- there is no evidence of cases where authorised researchers deliberately seek to re-identify observations.

However, disclosure risk still exists, because researchers may disclose individual information **accidentally** – for example when they publish un-safe tables without checking confidentiality rules. Scientific use files are protected, to limit these risks.

¹⁴ [Handbook on statistical disclosure control](#)

6.2 SDC methods for microdata

Various SDC methods are available for microdata protection. In general, they reduce information about the statistical unit or introduce data perturbation.

Some methods try to prevent identity disclosure, whereas others try to create uncertainty about the attribute. Table 6 below compares the most frequent methods.

Table 6 Most frequent methods for business microdata protection

SDC method	Definition	Example
Global recoding (information reduction method)	Re-categorisation applied to the whole dataset	Employment size-classes 250-499, 500- 999 1000+ are merged into a single class, 250+
Top/bottom coding (information reduction method)	All values above or below a specified value are set to that value, or to a code indicating the class	All turnovers over €500 000 are set to equal €500 000 or to 500+.
Micro-aggregation (data perturbation)	Records are grouped, based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates for those variables. What is released is the aggregates (e.g. the mean of the aggregated values), not the individual record values.	Records are ordered by turnover, in ascending order – for each group of e.g. 3 records the real turnover is replaced by the average of the group
PRAM (Post randomisation method) (data perturbation)	The scores of a categorical variable are changed, with certain probabilities, into other scores. The method can be defined as <i>intentional misclassification with known misclassification probabilities</i> .	
Suppression (information reduction method)	<i>Whole-variable suppression</i> - a variable is no longer released for the whole file. <i>Whole record suppression</i> - a whole record is suppressed.	

	<i>Local suppression</i> - one or more records have a variable value suppressed	
Record swapping (data perturbation)	Swapping pairs of records that are partially matched on a set of key variables but are e.g. in different geographical locations.	
Rounding (information reduction method)	Replaces original values of response variables with rounded values	

6.3 Microdata protection process

Protecting microdata is a process divided in several steps.

It is important to distinguish different types of variables: direct identifiers, indirect identifiers and response variables:

- **direct identifiers** – such as name and unique national identification number.
- **indirect identifiers** – such as NACE category, size class or region – which divide the total population into subpopulations; rare combinations may lead to the statistical unit being identified.
- **response variables** – represent the information about the statistical unit which would be disclosed if the unit is identified. If they have extreme values, response variables can also lead to the statistical unit being identified.

Direct identifiers are always removed from the microdata files for researchers, as they are clearly disclosive. They are usually separated from the data and replaced by a statistical identifier that has no administrative function (pseudonymisation), early in the statistical process, to reduce risks during statistical processing.

When preparing microdata, the combinations of variables that may lead to identification of statistical units are analysed. **The variables most likely to lead to identification are called quasi-identifiers.** For business microdata these would typically be variables such as NACE, NUTS and size class.

The steps to follow to protect business microdata files for researchers, based on already pseudonymised microdata, are¹⁵:

1. Define quasi-identifiers;

¹⁵ Based on the methodology for developing scientific use files for Community Innovation Survey.

2. Decide on the allowed share of combinations of quasi-identifiers leading to small frequencies (e.g. at most 5% of combinations should lead to low frequencies, e.g. 1 or 2 units – this is an acceptable level of identification risk).

The actual level of identification risk will vary, depending on the targeted type of microdata file and data sensitivity.

3. Calculate frequencies for the combinations of quasi-identifiers;
4. Perform a global recode on the quasi-identifiers;
5. Apply SDC methods like micro-aggregation on the variables and records requiring protection (*for business microdata, a typical variable to micro-aggregate is turnover*).
6. Perform local suppressions on the identifying variables, if there are still records requiring protection;
7. Repeat steps 3-6 until you achieve an acceptable level of identification risk and utility in the file.

6.4 Information loss due to microdata protection

The more microdata is protected, the more information is lost in the data.

If possible, carry out the microdata preparation ***in collaboration with potential data users***. This helps to preserve variables that are extremely important for data analysis and identify those that could be "sacrificed" for protection purposes.

For measures of information loss, see the chapter "*Information loss in microdata protection*" in the [Handbook on Statistical Disclosure Control](#). The measures are usually based on:

- comparison of records in the original and protected dataset;
- comparison of some statistics computed on the basis of the original and protected dataset.

7. SDC tools

Applying SDC is a complex process – it has to guarantee the anonymity of statistical units yet not lead to unnecessary loss of information through excessive suppressions/modifications.

SDC tools have been developed to cope with these problems. There are 2 families:

- tools protecting confidential data presented in **tables**;
- tools protecting confidential data in microdata **files**.

Tools for tables

Standard tools for data presented in tables include **tau Argus** and R-based **sdcTable**.

According to the questionnaire conducted in the ESS (in 2016), most NSAs were familiar with tau Argus. Standard tools are often complemented by other tools (SAS, STATA, Excel) and manual procedures.

Specialised SDC tools identify primary confidential cells according to rules defined by the user. The more information is provided to the tool, the better the data is protected.

Ideally input data is microdata. The tools identify secondary confidential cells depending on the chosen treatment method; both primary and secondary confidential cells are protected with the selected method. The data may be protected by suppression, rounding etc.

Tools for microdata protection

These include **mu Argus** and R-based **sdcMicro**. These tools are accompanied by standard statistical tools such SAS, STATA, SPSS. The specialised microdata protection tools apply SDC methods on the microdata listed above.

Other types of output

There are no tools to protect other types of output: especially models, graphs etc. The complexity of such output usually makes it very difficult to disclose information about individual statistical units.

For some useful guidelines on how to deal with these specific forms of output, see [Guidelines for the checking](#).

Eurostat supports migration of SDC tools towards open source solutions. Since 2015 Argus tools have been open source and since 2016 the Argus codes are available on GitHub¹⁶. User support is also offered on GitHub¹⁷.

8. Confidential data transmission between NSAs and Eurostat (flags and meta-information)

Data transmitted to Eurostat or between NSAs needs to be appropriately flagged. For European statistics this is usually defined in the relevant subject-matter regulations.

The minimum information Eurostat needs to treat confidentiality at EU level is the confidential data itself.

¹⁶ GitHub is a web-based repository hosting open-source software projects. It offers version control and source code management.

¹⁷ All information is available here: <https://github.com/sdcTools>

If the NSAs can also provide the *reasons* for confidentiality (e.g. number of statistical units in the cell, shares of the first and the second largest contributors etc.), Eurostat can set the confidentiality of the EU aggregates more accurately.

Having the reasons for confidentiality enables us to apply a set of criteria to check whether or not data on individual statistical units is safe when we publish EU-level aggregates and, if not, to take appropriate measures to rectify that.

There are **4 typical cases** of confidential data transmission between NSAs and Eurostat:

1. No confidential figures sent to Eurostat, no information on the reasons for confidentiality;
2. Confidential data sent to Eurostat (hypercubes, tables), but no reasons for confidentiality;
3. Confidential data sent to Eurostat together with information about the reason (sensitivity rule) for confidentiality and confidentiality parameters (e.g. number of contributors to the cell, share of largest contributor);
4. Data provided at micro level.

In general: the more detailed the data received, the more efficient the treatment of statistical confidentiality, in terms of balancing data protection and the provision of aggregates.

Information at micro level allows Eurostat to protect statistical units by taking into account unique or dominating statistical units across different Member States.

If the data is transmitted in semi-aggregated format (e.g. as hypercubes, or in SBS format) and if some additional information is provided on the reasons for confidentiality, the decision to publish EU aggregates can still be taken without unnecessary suppressions.

In case (2) above (no explanatory information provided on the reasons for confidentiality), the national contribution is treated as the contribution of a single statistical unit (of course, the worst case scenario).

To ensure a consistent system of flags in the statistical domains that use SDMX format to send the data, the code list on "confidentiality status" was defined by the SDMX Statistical Working Group¹⁸.

It is a mixed list, indicating whether observations are free for publication, for internal use only (for reasons not related to statistical confidentiality) or confidential. It also provides reasons for confidentiality (e.g. dominance rule).

¹⁸ http://sdmx.org/wp-content/uploads/CL_CONF_STATUS_v1_1_26-6-2014.doc

Table 7 Confidentiality-related flags used in SDMX

Recommended code value	Recommended code description	Annotation
F	Free (free for publication)	<p>Used for observations without any special sensitivity considerations and which can thus be freely shared.</p> <p>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. <i>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.</i></p>
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 <i>sender</i> , 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 <i>receiver</i> , 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 <i>small counts</i>	A cell is flagged as confidential if less than <i>m</i> 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.

9. SDC aspects in confidential data exchange

Members of the ESS are permitted by Regulation 223/2009 to exchange confidential data, if this is necessary for developing, producing and distributing European statistics or improving their quality.

Such data exchanges are particularly useful for statistics measuring cross-border flows of goods, capital, or people.

In some statistical domains (e.g. tourism statistics, Euro-Group register), ESS members exchange confidential data through Eurostat as an intermediary. They often use this data to perform comparisons, to check the quality of their own statistics (e.g. outward tourism from country A to country B (as reported by country A) should mirror inward tourism from country A to country B, as reported by country B).


Direct exchanges of confidential data between ESS members for purposes of *producing* statistics will be a new practice under FRIBS (see [Data sources](#), in particular section 4.3 on the microdata exchange for intra-EU trade).

There are 2 basic approaches to SDC when confidential data is exchanged:

1. The sender protects confidential data and the receiver receives data that are already safe, to be processed further or published (e.g. tourism statistics)
2. The sender sends confidential data together with instructions on how to protect it. If both sender and receiver publish the data, they must agree on which SDC rules for protecting it they will apply, and must both apply them equally.

Situation (2) applies for example when NSAs send data to Eurostat to produce European statistics. The instructions on how to protect the data are sent together with

confidential data, and are agreed with Eurostat and the NSAs in confidentiality charters.

 *It is important to keep the receiver of the data well informed about any **change in the status of the particular cell** (confidential or not) and the underlying reasons for this.*

Any inconsistency may lead not only to disclosure of the data – both for the country concerned and possibly also other countries (e.g. if the data concerned was part of a confidential cluster in an EU aggregate¹⁹).

In SIMSTAT, the SDC rules must be agreed by all exchanging parties, based on core principles for identifiable micro-data exchange that they establish among themselves.

10. Risk management ("5 safes" model)

The "5 safes" model is useful for considering various elements of data protection and security. It refers to:

1. safe **projects**
2. safe **settings**
3. safe **data**
4. safe **outputs**
5. safe **people**.

It is often used for microdata release²⁰, but can also be helpful for reviewing some aspects of confidential data protection in NSAs' standard distribution process. The table below shows how the model is applied to business statistics.

¹⁹ Confidential cluster: the group of countries contributing to an EU aggregate and whose data is confidential for a particular variable.

²⁰ See more: [Self-study material for the users of Eurostat microdata sets](#)

Table 8 – The "5 safes" model, applied to business statistics

The "5 safes"	What makes it "safe"?
Safe projects	Are the confidential data used for lawful (statistical or scientific) purposes?
Safe settings	Are confidential data used in appropriately safe environment? Are they securely stored? Is access limited to authorised staff/researchers?
Safe data	Are the data safe to be published? For microdata, are they prepared adequately, regarding access settings and microdata type?
Safe output	For microdata, are the results of the research safe to be made public?
Safe people	Are the users of confidential data aware and respectful of confidential data handling conditions?

11. See also

Overview of methodologies of European business statistics: [EBS manual](#)

Legal provisions related to Statistical Disclosure Control can be found in the following [overview](#)

12. Further Eurostat information

Key regulations:

- [Regulation \(EC\) No 223/2009](#) of the European Parliament and of the Council on European statistics
- [Commission Regulation \(EU\) No 557/2013](#) on access to confidential data for scientific purposes

Methodological guidelines

- [Guidelines for output checking](#) (*Data without boundaries* project).
- [Handbook on statistical disclosure control](#), version 1.2; January 2010; ESSnet on Statistical Disclosure Control; Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Lenz R., Naylor J., Schulte Nordholt E., Seri G., de Wolf P-P.
- [Self study material for users of Eurostat microdata sets](#)
- [sdcTools manuals](#)

- [Memobust handbook](#), Statistical Disclosure Control, 26 March 2014, ESSnet on Methodology for modern business statistics
- [Model for Confidentiality Charters](#), Eurostat April 2016 (in the annex)
- Recommendations for confidentiality management in business statistics in the ESS, September 2016 (document available for the ESS on request)

13. External links

14. Contacts

Aleksandra Bujnowska (aleksandra.bujnowska@ec.europa.eu)

Wim Kloek (wilhelmus.kloek@ec.europa.eu)

Estat-microdata-access@ec.europa.eu