Food and Agriculture Organization
of the United Nations

STATISTICAL DISCLOSURE CONTROL PROTOCOL

Contents

Figures

Tables

Acknowledgements

This document was originally prepared by Daniela Ichim (Statistician Italian National Statistical Institute), Michael Rahija (Statistician in the Office of Chief Statistician), and Aliou Mballo (Statistician in the Office of Chief Statistician). Anidi Oluwakayode (Statistician in the Office of Chief Statistician) also provided support in the finalization. It was develop in order to define the processes and procedures for populating the Food and Agriculture Microdata (FAM) catalogue. Financial support for the work of Daniela Ichim and Aliou Mballo was provided by the Bill and Melinda Gates Foundation, and United States Agency for International Development.

The World Bank microdata library team especially, Matthew Welch, provided a lot of help by sharing similar documents which govern the use and management of the World Bank microdata library.

I. Background and introduction

Purpose and definition of Statistical Disclosure Control

The compilation and dissemination of official statistics is one of the core responsibilities of a nation's government. Statistics provide the foundation for evidence based decision making, and effective public administration cannot be undertaken without timely and accurate statistics. Traditionally, in order to compile official statistics, national statistical systems collect microdata which refers to unit-level data usually collected through surveys, experiments, censuses and administrative systems. These data provide information on individual people or entities (otherwise known as "data subject(s)") such as individuals, households, business enterprises, facilities, farms or even geographic areas. More often than not, microdata contain personally identifying and/or confidential information on the data subjects.

Statisticians have long acknowledged the importance of securing this information in order to maintain the trust of the populations they serve. In this regard, the 6th principle of the Fundamental Principles of Official Statistics states "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes." Furthermore, FAO's Statistical Quality Assurance defines Principle 10 as "All data subject to national confidentiality policies (e.g. concerning people and legal entities, or small aggregates) are kept strictly confidential, and are used exclusively for statistical purposes, or for purposes mandated by legislation."

However, while acknowledging the importance of securing individual data, the United Nations also advocates for the free dissemination of microdata. Disseminating microdata allows users to engage in research, increases the transparency and accountability of national statistical institutions, and generate quality improvements through feedback from users (UNSD 2014).

The competing principles of data security and microdata dissemination are arbitrated through a domain of statistics called Statistical Disclosure Control (SDC). SDC methods allow for protecting a dataset through the application of statistical tools, allowing the institution to safely disseminate the micro dataset.

It is important to note here that this protocol incorporates the FAO Personal Data Protection Principles (AC No. 2021/01) and it is in accordance with the principles.

Micro datasets in FAO

FAO engages directly in the collection of microdata through household and farm surveys, and procures micro datasets from external sources. Regarding the former, the statistics division (ESS) regularly implements surveys to rural households and agricultural holdings for methodological research related to agricultural statistics. The social protection (ESP) division does the same for research related to decent work, migration, and social protection schemes. The livestock division (AGAL) also engages directly in survey data collection through projects such as the Pastoralist Knowledge Hub. On the other hand, the nutrition division (ESN), and one team in ESS procures their datasets from external sources. In some cases, these datasets are already publically available, in others they are not. When datasets are purchased from a survey company, the sales agreement often includes a clause giving FAO the right to disseminate the microdata.

Surveys implemented for methodological research are frequently used to compare different measurement methods in terms of bias and cost-effectiveness. For example, under the Global Strategy to Improve Agricultural and Rural Statistics, a sample survey of 495 households was implemented in 3 regions in Kenya during 2018. The purpose of the exercise was not to generate representative statistics on any particular segment of the population. Instead, it was used to evaluate the biases of different methodologies for measuring underemployment (Hillesland & Mwaniki 2018). This type of exercise is also common in ESP for testing different ways measuring social economic variables on social protection, migration, gender, etc.

Other surveys which are either collected, supported, or purchased by FAO are processed in order to generate national statistics. For example, ESS purchases household survey data from Gallup Polling Company which contain the *Food Insecurity Experience Scale* module. ESS processes these surveys in order to compute the proportion of the population experiencing food insecurity.[13]

The Office of Chief Statistician which is charged with developing corporate standards and quality assurance for all statistical activities advocates for the dissemination of all micro datasets collected by FAO, or in any way supported by the organization. In most if not all cases, the micro datasets contain personally identifying and/or confidential and sensitive information which require removal or protection prior to dissemination.

Table 1 shows a few micro datasets collected by FAO with examples of personally identifying information, and confidential/sensitive information that they contain.

Table 1: Examples of microdata containing personal and confidential/sensitive information

Division	Survey description	Sampling Unit	Examples of personally identifying variables	Examples of personally identifying variables Confidential/sensitive
Collected by FAO				
ESS	Methodological research	Agricultural holdings	• Holder name • Holder phone number • Household member names	• Financial details of holding • Personal details of hh members (religion, incomes, etc.)
			• Geo-references of main structure and plots	
ESP	Social protection, migration, etc.	Households	• Household head name • Household member names • Household address or ID number	• Migratory status of household members • Participation in government transfer programs
AGAL	Pastoralist Knowledge Hub	Pastoralists (census)	• Pastoralist name • Geo-references of household • Telephone number	• Ethnicity
*Collection **supported** by FAO*				
ESS	AGRISurvey	Agricultural holdings	• Holder name • Holder phone number • Holder ID # • Holding registration # • Geo-references of main structure and plots	• Financial details of holding • Salaries of workers
ESS	World Census of Agriculture	Agricultural holdings	• Holder name • Holder phone number • Holder ID # Holding registration #	• Financial details of holding • Ethnic group of holder

Note: AH refers to household and non-household sector holdings as defined by the WCA 2020.

This list of datasets is not exhaustive, but illustrates the presence of micro datasets in FAO which require the application of SDC methods prior to release. The release of personally identifying and/or sensitive or confidential information contained in datasets represents an institutional risk to the organization. It follows that a protocol on how FAO uses SDC to protect this information is needed.

Existing corporate and UN-level policies relevant to microdata dissemination

The FAO corporate microdata dissemination policies mandates that microdata dissemination follow UN Personal Data Protection and Privacy Principles[14], Principle 6 of the Fundamental Principles of Official Statistics, and the Corporate Data Protection and Privacy Policy.

Furthermore, this protocol is written under the assumption of full compliance with the following stipulations of the microdata dissemination protocol:

- The micro dataset will only be used for statistical and/or research purposes;
- Any results derived from the micro dataset will be used solely for reporting aggregated information, and not for any specific individual entities or data subjects;
- The users shall not take any action with the purpose of identifying any individual entity (i.e. person, household, enterprise, etc.) in the micro dataset(s). If such a disclosure is made inadvertently, no use will be made of the information, and it will be reported immediately to FAO;
- The micro dataset cannot be re-disseminated by users or shared with anyone other than the individuals that are granted access tothe micro dataset by FAO.

Objectives and how to use this protocol

This document is complementary to the Metadata and Microdata Curation and Dissemination Protocol. Accordingly, this document includes a wide range of methods from the academic literature which are divided into the following sections:

- Section 2 covers the topic of disclosure risk and provides key concepts, an overview of methods, and provides guidance on how to identify and measure disclosure risk.
- Section 3 addresses statistical disclosure limitation methods and provides instructions for how to use the main techniques totreat different variable types.
- Section 4 provides tools for how to evaluate a protected dataset including how to measure information loss, and data use specific utility measures.
- Section 5 describes how data curator in OCS will document SDC process and what information will be made available to whom.
- Section 6 defines the SDC workflow describing the typical steps undertaken for protecting a micro dataset prior to dissemination.

Rigorous development, formulas, and other technical details on the tools and methods are not included in this protocol; however, relevant literature and citations are provided.

As topics arise, text is included to indicate what team takes-on the major related responsibilities. The **data provider** refers to the technical unit, or officer who is considered the custodian of the dataset who submits it for dissemination. The **data curator** is the officer in OCS who will process and prepare the data provided for dissemination. The **Office of Chief Statistician (OCS)** is the technical unit within FAO which holds the mandate over standard setting, and quality assurance for statistical data and the author of this document.

II. Disclosure risk

Disclosure risk refers to the likelihood that an individual (i.e. data subject) whose information is registered in a dataset will be identified, or more information than is intended will be revealed. This section will provide the reader with the key concepts and tools, for describing different types of risk, disclosure limitation methods, the risk vs. utility trade-off, and identifying and measuring disclosure risk.

Key concepts

A **data subject** is an individual, group of individuals, business, farm, etc. whose information is being collected. In a micro dataset, each record registers information on a data subject. In FAO, data subjects are typically individual people, however, in some agricultural survey programs, data subjects may be agricultural holdings.

Disclosure control refers to the measures taken to protect data subjects in accordance with confidentiality requirements. The goal is to ensure that the confidentiality protection provisions are met while preserving as much as possible the analytical usefulness of the microdata file.

When dealing with detailed information, **breaches of confidentiality** might occur when a user intentionally or otherwise finds out more information than authorized about a data subject. An **intruder** is a user who wants to probe the statistical data and reveal sensitive information about a specific data subject, or group of data subjects. It is impossible to completely alleviate the chance that an intrusion could occur in any dataset without removing all of the useful information. As a result, FAO releases datasets only to individuals that have registered, confirmed their identity, and signed an agreement with a specific terms of use including that microdata can only be used for research purposes. Additionally, experiences from other national and international organizations which disseminate microdata indicate that researchers do not deliberately spend disproportionate efforts trying to identify statistical units.

Disclosure scenarios form the basis of possible means and opportunities of disclosure. From a conceptual point of view, the disclosure scenarios are the answers to the question “How could one reveal confidential information about the data subjects?”. While there is not a universal answer to this question, it is clear that a disclosure scenario is a series of hypotheses about the ways that confidentiality of the data subjects might be breached. As disclosure scenarios depend on the survey/microdata content, different disclosure scenarios might apply for different surveys. As different users might have access to the same microdata file, different disclosure scenarios might apply for the same microdata file.

There are two main types of disclosure defined by the type of information they reveal about data subjects (Willenborg and De Waal 2001):

a) **Identity disclosure** occurs when the intruder links a known legal person (individual or business or household) to a record in the disseminated microdata file. Such disclosure might happen in presence of unique or very rare characteristics. Examples are fiscal code (of individuals or enterprises), monopolistic enterprises, extremely large land surfaces, the only proprietary of some type of livestock, rich individuals resident in extremely poor regions, etc.

b) **Individual attribute disclosure** may be achieved even without identity disclosure. Attribute disclosure occurs when an intruder is able to discover sensitive attributes about data subjects without needing to uncover their identity. The sensitive attributes may be of interest to an intruder and, if disclosed, could cause harm to the data subjects. The following are two particular types of attribute disclosure:

- **Inferential attribute disclosure** occurs when an intruder discovers the value of some survey variables more accurately with the released microdata file than otherwise would have been possible. For example, suppose an intruder builds a regression model which can accurately predict the religion of an individual based on the data subjects age, location, and ethnic group. The intruder can make a reasonable inference about a sensitive variable (religion) without needing to match the data subject to a specific person.
- **Group attribute disclosure** occurs when an intruder can learn about a group of individuals but not about a single subject. Such disclosure might happen when all individuals in a group suffer from the same disease, all the individuals report the same nutrition problems or the same bad/good decent work characteristics or all the enterprises have negative value-added or all crops were destroyed during the survey reference period, etc.

Except for the group attribute disclosure, individual attribute disclosure often follows identity disclosure, meaning that identity disclosure is a necessary condition for attribute disclosure.

As part of the microdata curation processes, data providers will be required to classify certain variables which are important for the SDC process (Willenborg and De Waal 2001). Below are the classifications which will be used.

Direct identifiers are variables allowing a direct and unambiguous identification of the legal units without any additional effort/knowledge. Examples are names, phone numbers, web-sites, IP address, ID card number, etc.

Indirect identifiers are variables whose combination can be used to reveal the identity of a data subject. This can be achieved by trivial observation, linking or using administrative data (for example) or other information already known by the user. Consequently, the indirect identifiers are typically *visible* variables (e.g. roofing materials, external plumbing, etc.), *sensitive* variables (e.g. health information, income, etc.) or *registered* which means they are publicly available in public records, administrative systems, or datasets.

Indirect identifiers may be both categorical and continuous variables. Examples included in the first category are gender, place of birth, marital status, industrial sector, administrative territorial units of residence/activity/location, migratory status, participation in government transfer programs, etc. Examples of continuous indirect identifiers are age, income, number of cars/tractors, agricultural surface, number of heads (animals), turnover, number of employees, financial details of holdings, etc.

Registered indirect identifiers might be visible or not, e.g. sex or place of birth. Income is an example of indirect identifier which might be registered or not, but generally is not easily available to users in absence of a targeted survey.

Sensitive and/or extremely sensitive variables represent the confidential content of the microdata file, i.e. the information unavailable to the intruder before accessing the disseminated microdata file. Sensitive variables may be indirect identifiers. Examples of sensitive variables are income, economic accounts of firms, number of hours worked, or information related to time use, decent work, nutrition aspects, pastoral movements, food security, land use, etc. **Extremely sensitive variables** are those variables with a high sensitive content and if a disclosure is made, will cause physical, legal, or emotional harm to the data subject. Insome cases, they are indicated by laws or other legal provisions. Examples of extremely sensitive variables are those related to health, justice, political opinions, religious beliefs, sexual life, drug/alcohol use, etc.

Free text – while it is difficult to a-priori classify such variable, it might anyway contain some direct, indirect or sensitive information. Using some knowledge extraction methods and tools, user might find this information useful for their statistical analyses. An example of "problematic" free text is an answer including both name (direct identifier) and sensitive information (e.g. earnings), e.g. "My name is Mario Rossi. Last year I earned 10 dollars while this year I earned 1,000,000,000 dollars. I didn't know which is the right answer to the item 5 in the questionnaire".

Variables not related to legal units do not generally pose any confidentiality problems. Anyway, any possible link to legal units whose confidentiality must be protected should be carefully assessed before microdata dissemination. Examples of variables not related to legal units include food or product nutritional characteristics, bio-physical features of animals, soil or crop composition, etc. Anyway, if there is a unique producer of a given product, or a unique owner of a certain type of animals, the dissemination of too detailed information might reveal confidential information concerning legal units.

Both types of disclosure, i.e. identity and attribute disclosure, may be achieved by means of both direct and indirect identifiers.

Disclosure risk (Hundepool et al 2012) is defined as the probability of disclosure with respect to specified set of disclosure scenario(s). Accordingly, each type of disclosure and each disclosure scenario is associated with a level of disclosure risk which should be assessed. For each disclosure scenario, the confidentiality of the answers provided by a respondent is considered protected if the disclosure risk for this respondent and the respondent's answers is sufficiently low, i.e. below acceptable thresholds.

Individual risk is defined as the disclosure risk associated to each unit in the microdata file. A common approach is to consider some form of the re-identification risk (identity disclosure). The latter is concerned with the possibility that the intruder would be able to determine a correct link between a microdata record and a known unit. This definition of risk is appropriate only if the records in the microdata file are associated with units in the population.

The declared advantage of an individual risk measure is that only those records appearing unsafe for a given risk threshold need to be locally protected.

Global risk measures are global or file-level disclosure risk measures obtained by synthetizing/ aggregating record-level risk measures. Global risk measures are useful for evaluating the overall quality of the microdata file to be disseminated.

Microdata protection from disclosure implies controlling the possible re-identifications of the data subjects using **statistical disclosure limitation methods** by limiting the information content of the data, particularly of the indirect identifiers.

Statistical disclosure limitation methods[15] can be classified upon their impact on the data in:

- data (or information) reduction
- data perturbation

Data reduction methods, such as global recoding or local suppression, aim at increasing the number of individuals in the sample/population sharing the same or similar identifying characteristics shown by the units considered at risk. Such procedures tend to avoid/minimize the existence of unique or rare recognizable individuals, households or enterprises.

Perturbation methods like random rounding, adding noise or swapping achieve data protection from a two-fold perspective. First, if the data are modified, re-identification by means of matching algorithms, e.g. record linkage, is harder and more uncertain. Secondly, even when an intruder is able to re-identify a unit, he/she could not be confident that the disclosed data are consistent with the original data.

Data protection methods necessarily lead to a modification of the original data. Quantification of these changes is the goal of **information loss measures**. While there is no widely accepted information loss measure, there are three main approaches in evaluating the impact of the protection methods on statistical outputs:

- Measuring information content of the microdata file (mathematics-based)
- Comparison with outputs already disseminated (i.e. reports, official statistics, etc.) (office-based)
- Considering the impact on other uses of the dataset (user-based)

These approaches may be applied to both categorical and continuous variables.

The first one is generally based on the mathematical concept of entropy (and related) which can be applied to quantify the difference in the information content between the original and protected dataset. While entropy-based is a mathematically sound approach, it is rarely applied in practical microdata dissemination projects since it is not directly related to the survey objectives, to the finite populations estimation framework or to the analytical potential of the datasets.

The second approach is based on the derivation of coherence measures. The idea is to use the protected microdata file to reproduce the statistical outputs already disseminated. For example, the data curator will compare the estimated totals or means obtained using the microdata before and after the application of the statistical disclosure control methods.

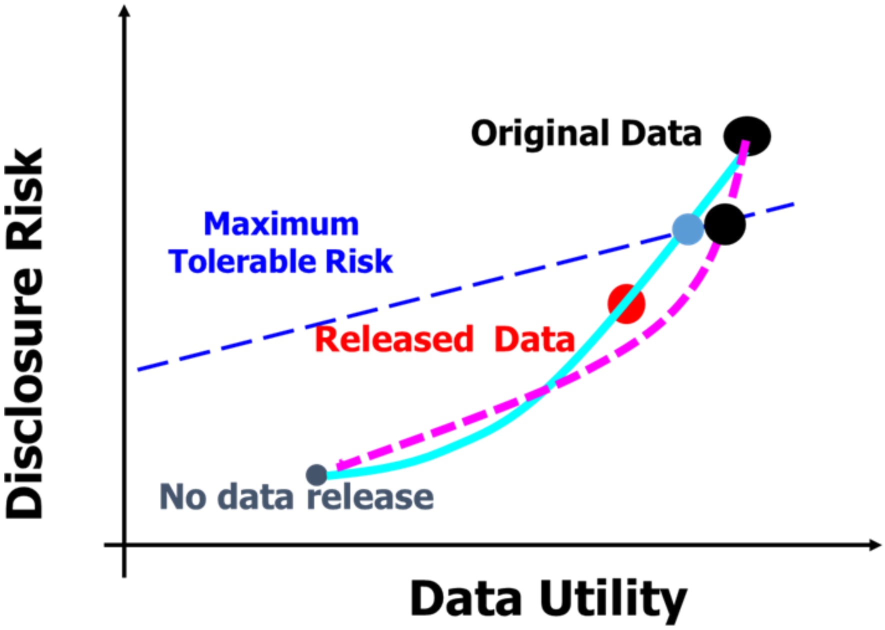
For the implementation of the third approach it is useful to understand the data, understand the users and their main research needs. In this framework, the information loss is measured by means of its counterpart, i.e. data utility which should be considered from the perspective of users of the microdata file.

Risk-utility trade-off

Reducing risk by using methods which result in information loss is known as the **risk-utility trade-off** (Duncan et al 2001). This may be formulated as an optimisation problem. In practice, different SDC methods can be evaluated and compared by considering the values of alternative measures of risk and utility. Sometimes it can be useful to construct a Risk-Utility map, such as Figure 1. As both disclosure risk and data utility assessments might involve some subjective choices, the Risk-Utility framework can assist in the selection of an optimal anonymization strategy.

In Figure 1, the risk of disclosure is represented on the vertical axis while the data utility is represented on the horizontal axis. The dots represent different versions of the anonymized microdata file (for example the same protection method applied using different parameters). Obviously, the lowest disclosure risk is associated with the “no data release” point. In the context of microdata dissemination, the same "no data release" point is associated with no data utility, and no risk. The opposite situation is represented by the "original data release", as if no statistical disclosure limitation method were applied. Such point simultaneously shows the maximum value of data utility and the maximum amount of disclosure risk. "Intermediate" data release points (red points) would show disclosure risk and data utility values between these extremes. The blue dashed line represents the maximum tolerable risk line. Depending on the problem setting, it might be parallel to the horizontal axis, i.e. when OCS a-priori defines the maximum disclosure risk threshold. The blue dashed line might also have a positive slope, when OCS is willing to accept more risk in change of a much greater data utility. As there is no mathematical/deterministic relationship between disclosure risk and data utility, the definition of thresholds for both disclosure risk and data utility is generally an extremely difficult task. In practice, SDC experts generally apply different disclosure limitation methods, evaluate both the remaining disclosure risk and data utility for a given set of measures or parameters. This choice will be made as an agreement between OCS and the data provider.

Figure 1: Disclosure risk and data utility map



Note: Author's own elaboration.

Identifying Disclosure Risk

The assessment of the disclosure risk is based on the analysis of the possible means (Hundepool et at. 2010) an intruder might use to identify data subjects. In few words, an intruder may:

- link data subjects in the microdata file with an external data source
- reveal the identity of a data subject through unique knowledge of the intruder

For the first case, direct and indirect identifiers both present clear disclosure risk. **Direct identifiers** are never released and should be removed by the data provider prior to submitting a microdata set to OCS. Indirect identifiers contain **key variables** which are unique combination of indirect identifiers used to breach confidentiality. To pose a risk for disclosure by linkage, key variables must be available for the population represented in the micro dataset in an external data source. Data providers will provide a list of key variables when submitting a dataset in order of importance.

An intruder in possession of external datasets could use key variables to match against the disseminated microdata file. This would enable identification of an individual or other data subjects if the register includes direct identifiers. Examples of such external data sources are (or related to) electoral lists, business registers, land registers or different property lists, cars (or other transportation/work means like tractors) registers, etc. The external registers may also be commercial ones. In case of sample surveys, special attention should be paid to the sampling lists/frames if they are publicly accessible.

Familiarity with the data should enable a list of key variables for each microdata file to be disseminated. A non-prescriptive shortlist of common key variables (likely to be found in external data sources) is:

- age
- sex
- education
- marital status
- geographical information (place of birth, place of residence, place of work, etc.)
- size or composition of household
- size or composition of the livestock
- land/crop characteristics (composition, surface, administration, etc.)
- dwelling characteristics (type, number of bathrooms/toilets, surface, number of rooms, etc.)
- income/wages/taxes (household or individual)
- occupation or industry
- number of employees
- information related to international trade
- ethnic group
- religion
- area planted

The availability and accessibility of external data sources containing the key variables should be assessed and it should be determined whether or not likely intruders have the expertise to apply the record matching techniques. It is important that the existence of external data sources for each micro dataset is identified, and considered by the data provider and data curator in OCS.

Record linkage or matching to external files is a standard method of intrusion. In register-based disclosure scenarios, it is assumed that an intruder would use the key variables in order to implement several deterministic or probabilistic record matching techniques. Several aspects related to the means an intruder might use should be discussed when evaluating such disclosure scenarios. Firstly, the availability and accessibility of such registers together with the level and quality of information registered in these data sources should be carefully assessed. Only common variables could be used for matching purposes.

Additionally, it should be assumed that an intruder might not have access to all the external registers, but different intruders might have access to different external data sources. Secondly, it should be assumed that the intruder has the technical means and knowledge to apply any record matching technique. Thirdly, it should be assumed that an intruder would link published information with the released microdata using only a selection of key variables. Indeed, it is known that the increase in the number of matching variables might reduce the number of correct matches. In these disclosure scenarios it is generally assumed that the intruder would use up to a maximum of 4 combinations of key variables. Finally, the quality of the information registered in both microdata files should have a significant impact on the matching results.

The second type of disclosure relies on unique knowledge an intruder has related to a data subject or group of data subjects. For example, **spontaneous recognition** occurs when an intruder uses his/her knowledge of specific data subject(s) to disclose their identity. Such identification may not necessarily be malicious and could be realized by the intruder unintentionally. He may still be referred to as an intruder, even in absence of a malicious intent.

For individuals, the personal knowledge an intruder might use are likely key variables about a friend, family member or work colleague e.g. name, age, sex, marital status, income, occupation, address, housing variables, such as accommodation type and tenure, etc. These variables may be known within a certain degree of accuracy. For example, an intruder might know (without any special effort) the approximate age of his colleague.

For businesses the intruder might know the industry sector, location, number of employees, lists of products, etc. In case of extremely rare values (for example monopolistic situations), such variables might lead to identification of units.

The **nosy neighbour scenario** is a sub-case of the spontaneous recognition scenario. This scenario relies exclusively on detailed private knowledge that only a reduced number of persons such as family members or neighbours/colleagues or friends would know. Accordingly, it is more likely that one particular individual will be targeted. Besides the key variables mentioned in the spontaneous recognition scenario, other common key variables used in the nosy neighbour scenario include number of cars/tractors/etc., household composition, well-being information (nutrition, decent work, financial situation). What distinguishes the nosy neighbour scenario is the extremely detailed information an intruder might have about a single person.

The risk of disclosure due to personal knowledge of the intruder can be quantified only on the basis of rare values of some variables in the microdata file. Methods for measuring this risk are described later in this section.

Finally, when identifying disclosure scenarios, it is important to consider the **disproportionate effort rule**. As discussed in the Risk-Utility framework, completely eliminating disclosure risk without eliminating data utility is impossible. The Terms of Use for micro datasets allow the use of microdata only for statistical, scientific and/or research purposes. Other usages are strictly prohibited. This fact minimizes the incentives for users to make huge efforts for identifying data subjects (i.e collecting and preparing external datasets, applying record matching techniques, etc.). OCS will take into consideration the amount of the effort required for intrusion vs. the potential reward when determining the appropriate level of risk and information loss.

Measuring record-level risk

The primary technique for measuring record-level risk is to analyze potential re-identifications using a set of key variables, their sample weights, and estimating records' uniqueness in the population. Over-simplified, the rarer some characteristic, or combination of characteristics are in a representative sample, the rarer they are likely to be in the population. The following provides an overview of the main methods for assessing record-level risk based on key variables which are either categorical or continuous under the assumption that no additional information will be used by an intruder to attempt a re-identification.

As part of OCS's responsibility to ensure the confidentiality published micro datasets, prior to publication, the data curator in OCS will analyse the set of key and sensitive (when applicable) variables using the techniques list below. See Section VI for further details.

Categorical variables

The simplest rule to measure the re-identification risk of a unit is to consider at risk those records that are sample uniques with respect to the given combination of key variables. The **threshold rule (**Willenborg and De Waal 2001**)**, and its extension **k-anonymity** are applied based on sample frequencies of a key variable or combination of key variables. The threshold rule implies that a record containing a unique value, or combination of key variables, is likely to be unique in the population, and therefore easier for an intruder to disclose. Obviously, in case of sample surveys characterized by small or extremely small rates, the validity of such assumption should be questionable.

K-anonymity (Sweeny 2002) is an extension of the threshold rule and states that records whose frequencies with respect to a set of key variables are smaller than a given threshold should be considered at risk. A commonly used threshold is equal to 3. The idea is that if there are at least three units in the sample sharing the same characteristics, there should be at least three units in the population sharing the same characteristics. Consequently, such records should be considered as safe records.

Special uniques (SUDA) (Elliot et al 2005) are sample uniques which also contain unique combinations of subsets of key variables. Suppose, for example, that there is one data subject for which the combination of their gender, dwelling type, highest education reached and employment status renders him/her unique. This data subject would also be a special unique if a combination (e.g. highest education reached and employment status) of a subset of these variables were also a sample unique. When this subset of key variables is obtained by collapsing geographical information, these special uniques seem to be far more likely population uniques.

ℓ-diversity (Machanavajjhala et al 2007) pertains to sensitive variables. This disclosure risk definition can be used to protect data against attribute disclosure by ensuring that each sensitive variable has at least ℓ values. In case the rule is not satisfied, all records should be considered at risk.

The threshold rule, k-anonymity, SUDA and **ℓ**-diversity are binary disclosure risk measures: a record may be at risk or not. Consequently, they do not incorporate the uncertainty associated to the disclosure scenarios, key variables selection, intruder technical knowledge and abilities, interests and efforts, etc.

Individual risk based on unknown population frequencies

The risk measures exclusively based on sample frequencies might lead to overprotection, especially in presence of small sampling rates. Several approaches may be applied to deal with population uniqueness too, for example by defining at risk those sample uniques which are population uniques, too. For census data, or when an administrative register covering the whole population is available, the population frequencies are known for each key variable and the risk measure can then be computed. For sample surveys, when population frequencies are unknown, risk measures based on inference procedures, i.e. model-based procedures, might be applied (Elamir and Skinner 2006). In practical situations, the most frequently used model-based measures of individual disclosure risk for microdata are:

- the probability that a sample unique is a population unique;
- the probability of a correct re-identification.

The probability that a sample unique is a population unique is generally estimated by means of inferential procedures based on log-linear models or on assumptions on the population frequencies distributions, e.g. negative binomial. As these model-based measures use calibration weights, they may be applied only in case of sampling surveys, paying attention to the mismatch between key and stratification variables.

In absence of calibration weights (representativeness property), the matching between the microdata file and an external data source is usually performed by means of probabilistic record matching techniques, thus obtaining the number of correct linkages. Of course, such technique might be applied when an external data source is available during the disclosure risk assessment stage (Shlomo and Skinner 2010).

Continuous variables

For continuous variables, techniques for measuring record-level risk based on rareness or uniqueness in a sample or population are useless as each value of a continuous variable is likely to be unique. Accordingly, the methods for measuring record level disclosure risk using continuous variables are limited.

One option is to consider all the records at risk of disclosure and to apply a protection method to the key variables. Finally, the records at risk of re-identification are those records correctly matched when linking the original and the protected versions of the dataset (Domingo-Ferrer and Torra 2003). Of course, this risk of re-identification depends on the selected protection method. Moreover, such an approach might lead to overprotection.

A second category of risk assessment procedures for continuous key variables is represented by several methods borrowed from the outlier detection framework. For example, one may define at risk all the units for which the univariate continuous key variable takes a value greater than the 95% quantile of its observed values. This outlier approach generates a fixed percentage of units at risk, depending on the subjective choice of the quantile. Moreover, the units at risk would be, by definition, on the right tail of the distribution. Boxplots and other outlier detection criteria may be used as well. These approaches obviously inherit the problems related to outlier detection, including the definition of an outlier.

The few clustering algorithms applied in the statistical disclosure control framework are distance-based algorithms. Anyway, the standard algorithms generally tend to find clusters with equal variance. Such algorithms would tend to define at risk only some units on the tails of the distribution of the continuous key variable. For example, when hierarchical ascending algorithms are applied, a cut-off value on the aggregation distance defines the units at risk; the last aggregated units are considered at risk. The clustering algorithms may not be suitable for continuous key variables with very skewed distributions.

It should be observed that when the outlier-based or cluster-based strategies are used, no individual risk measure is estimated; the risk is rather considered as a binary property.

Hierarchical data

For many micro datasets, the data has a hierarchical structure where an individual data subject belongs to a higher-level group of data subjects. Typical examples are individuals which belong to households, workers that belong to a holding, etc. The disclosure of a data subject in a lower level of the hierarchy can imply disclosures at a higher level and vice versa. Accordingly, these hierarchical data structures should be taken into account when measuring the disclosure risk. The information on the hierarchical structure, e.g. household type or household size, should be included among the key variables. Here we refer to households as they represent the most frequent situation.

Only the individual risk methodology allows for a hierarchical structure in the data. The household risk (Hundepool et al. 2014) is defined as the probability that at least one individual in the household is re-identified. For the individual level dataset, all the individuals in a given household share the same household risk. In practice, a risk threshold is set on the household risk. Then, the individuals belonging to a household are considered unsafe (at risk) when their individual risks are greater than the threshold divided by the household size. This is equivalent to considering that the household members share the same individual risk, i.e. the maximum value. Consequently, this is clearly a strongly prudential approach.

Agricultural surveys can contain more than two levels. For example, a holding may have parcels which contain plots which contain crops. Also, if the holding is a household sector holding, then the holding belongs to a household. Accordingly, the disclosure of a plot could result in disclosure of parcels, holding, and even household. Another example of disclosure could occur if there is only one holding growing a specific crop. Consider a household sector which is the only one that grows apples in a study area. The presence of apples in a crop roster could be used to disclose the household.

When external registers are available and when the levels do not vary over time, a hierarchical risk approach should be followed for each level in the dataset by defining first the risk threshold R_h at the higher hierarchical levels. Then, for the other levels, the risk threshold may be found by dividing R_h by the number of units belonging to that level. This is similar to the individual-household approach. For example, for the plot level dataset, a plot belonging to a household can be considered at risk if the risk associated to the plot is greater than the threshold set for the household risk divided by the number of plots the household has.

Structural information, on plots may be registered in some external databases. Moreover, the number of plots belonging to each household/holding is not expected to vary too much over time. Hence a hierarchical risk approach may be suitable for plots. Examples of structural information that may be used to assess the disclosure risk of plots are area, territorial position, management type, etc.

However, it is less likely there are register which contain crop and/or livestock level information. Furthermore, these characteristics are likely to exhibit large variations even over short periods of time. Consequently, the hierarchical risk approach based on the individual risk methodology is not the most appropriate. In this context, the most identifying information on crops or livestock should be directly included among the key variables in disclosure scenarios on households/holdings. In other words, besides scenarios based on socio-demographic variables like household size, region of residence, dwelling characteristics, etc., several additional disclosure scenarios based on agricultural information should be considered. Examples of crop or livestock based key variables may be type of crops, principal type of farming, number of herds, species, etc.

Measuring global risks

The objective of computing a global risk for a micro dataset is to compute the overall level of risk of a dataset. Measuring global risk is a matter of aggregating individual risk measures. It follows that global risk measures inherit the properties of the individual risk measure.

The total number of records with record-level disclosure risks greater than a threshold is a first example of global risk measure. The threshold may be set in absolute (e.g. all records having a risk greater than 5%) or relative values (all records with risks greater than a certain decile of individual risk, for example). In both cases, this global risk measure is positively correlated to the number of records: datasets with more records would appear to have a greater global risk.

Another global risk measure is defined as the sum of individual risks, i.e. the expected proportion of all data subjects that could be identified. The expected number of re-identifications may be then easily derived by multiplying it with the number of records in the files. These two measures are generally implemented in SDC software[16].

When only sample uniques may be considered to have some disclosure risk, two global risk measures based on log-linear models may be estimated: the number of sample uniques that are also population uniques and the expected number of correct re-identifications for sample uniques (Skinner and Shlomo 2008).

The disclosure risk estimated using statistical modelling provide a consistent disclosure risk measures at both the individual record-level and the overall global file-level.

In general, global risk measures suffer from the drawback that some individual records may show very high individual risks even in presence of an acceptable global risk. Indeed, few high individual risk records may be compensated by many low individual risk records. Consequently, both global and individual measures of risk should be used in order to decide whether a microdata file may be released.

III. Disclosure limitation methods

This section introduces commonly used disclosure limitation methods in official statistics. Disclosure limitation methods generally aim at reducing the disclosure risk, increasing the uncertainty in possible re-identifications or at decreasing the "benefits" of possible disclosures. In other words, the aim is to achieve some tolerable level of risk for release.

These methods can be classified into non-perturbative and perturbative. The former include methods which do not distort the data, instead they reduce the detail in the original data set through removing specific variables or values for specific records. The latter is a set of methods which changes the underlying data while controlling the differences in estimations that can be derived from the perturbed and original dataset (Hundepool et al 2012).

The choice of an optimal SDC strategy is an **iterative process,** wherein the statistician is applying methods, and immediately evaluating their impact on disclosure risk and information loss. There is not a one-size-fits all solution, so methods are applied and evaluated until the dataset reaches an acceptable point on the Risk-Utility map (Figure 1).

As a first step, the data provider will remove direct identifiers, and extremely sensitive variables. The data provider will also provide OCS with a list of key variables in order of importance. Data curator in OCS will then, evaluate the dataset, and recommend a series of appropriate SDC methods resulting in an overall strategy. The data provider and OCS will then reach an agreement on at minimum the following:

- the protection type: data reduction, data perturbation or a mixture of both
- the relative importance of each key variable, i.e. which ones should be preserved most
- which records to protect, i.e. only records at risk or all the records
- method for evaluation of information loss (e.g. mathematics based, comparison with outputs, impact on statistical analyses, etc.)

This section will describe the set of disclosure limitation methods summarized in Table 2 below.

Table 2: Disclosure limitation methods by variable type, and responsible unit

Method type	Method	Variable classification	Common Variable type(s)	Responsible
Non-perturbative	Removal	Direct Identifiers	● Text ● Geo-references	Data Provider
Non-perturbative	Removal	Extremely Sensitive	● Text ● Categorical ● Continuous	Data Provider
Non-perturbative	Global and local recoding	Key and Sensitive	● Categorical ● Continuous	Agreement by OCS and Data provider
Non-perturbative	Top/Bottom coding	Key and Sensitive	● Categorical ● Continuous	Agreement by OCS and Data provider

Non-perturbative	Sampling	All variables in census datasets	All	Agreement by OCS and Data provider
Perturbative	Record swapping	Key and Sensitive	• Categorical • Continuous	Agreement by OCS and Data provider
Perturbative	Rank swapping	Key and Sensitive	• Categorical • Continuous	Agreement by OCS and Data provider
Perturbative	Post Randomization (PRAM)	Key and Sensitive	• Categorical	Agreement by OCS and Data provider
Perturbative	Rounding	Key and Sensitive	• Continuous	OCS with agreement from Data P.
Perturbative	Microaggregation	Key and Sensitive	• Continuous	Agreement by OCS and Data provider
Perturbative	Noise addition	Key and Sensitive	• Continuous	Agreement by OCS and Data provider
Perturbative	Model-based	Key and Sensitive	• Continuous	Agreement by OCS and Data provider

Note: Agreement means agreement between OCS and data provider.

Variables which are removed

Direct identifiers

Direct identifies must be suppressed from the microdata. Direct identifiers are variables such as names, addresses, phone numbers, IP address, web-sites, e-mail addresses and any identification number, e.g. passport ID, VAT number, etc. Extremely detailed geographical information, e.g. latitude and longitude of dwelling, plot, farm headquarters addresses, etc., are direct identifiers.

Extremely sensitive variables

These variables are generally related to health, justice, religion, politics, drugs, sexual life/identity, etc. and they shall be suppressed because the harm they could cause is considered unaffordable.

Geographic variables

In some cases, geographical variables, e.g. place of birth, though are not direct identifiers, they provide enough information to allow for easy disclosure. In this case, geographical information shall be released using higher hierarchical levels of standard territorial classifications. Combination of different or non-nested territorial classifications generally increases the disclosure risk, and will be not used.

Some variables, e.g. postal codes, phone numbers prefixes, urbanisation-based indicators, local taxes, particular types of agricultural products or herds, implicitly include detailed geographical information, so their inclusion in the microdata may reveal a lower level of geography. Such variables should be released only if they are referred to high levels of geography, and represent an acceptable level of disclosure risk. Additionally, some categories of variables like country of birth or nationality may have low frequencies, thus representing a higher disclosure risk.

If removal of geographic variables present an unacceptable level of information loss , perturbative and non-perturbative methods described later in this section may be applied to geographical and other variables. For example, family structure, number of children, etc. could be suppressed and the enumeration areas (geo information) could be swapped for the records at risk of disclosure.

Non-perturbative disclosure limitation methods for key and sensitive variables

The disclosure limitation methods may be applied to both key and sensitive variables. Non-perturbative methods increase protection by reducing information. It follows that the selection and application of non-perturbative methods must closely consider the extent of information loss.

In the presence of a large number of indirect identifiers (e.g. 10 – 15), removal of the least important variables is often appropriate especially when they are correlated with other variables to be disseminated. The following provide methods and recommendations for other types of variables.

Local suppression

Local suppression is a technique that replaces actual values in a dataset to missing values. It is a generally applied to categorical variables, but may also be applied to continuous key variables to achieve the goal of k-anonymity or any other threshold-based rule. It is applied only to those records considered at risk of disclosure. The idea is to eliminate extremely rare combinations of key variables by suppressing one or more values of key variables that render them records unique. For example, consider a dataset in which there are 3 key variables,e.g. age, gender, and marital status. If there is only one data subject that is 18, married, and female, then the data subject's marital status could be removed, making the data subject indistinguishable based on age and gender.

When deciding which key variable to suppress, it is important to consider their importance so that less important variables receive more local suppressions[17]. In the previous example, if marital status were the most important variable, local values of age or gender would be suppressed. It is worth noting that local suppression violates the assumption of missing at random when analyzing data. Local suppression is recommended when the amount of unique combinations of the key variables is low.

Global and local recoding

Recoding is the most common SDC method used when releasing microdata, and involves combining several highly detailed categories into less categories with a lower level of specificity. It is also the first method which is applied in the anonymization process and can be applied to both categorical and continuous variables.

For a categorical key variable, several categories may be combined to form new (less specific) categories. For example, a marital status which may take the values married, divorced, widower, or single maybe combined to "married" or "never married". Other examples include variables which contain specific categories of animals or crops, may be grouped by some higher level classification into species. When standard classifications exist, they generally guide the global recoding. In other

---

[17] https://sdcpractice.readthedocs.io/en/latest

cases, low frequencies categories (i.e. the most disclosive ones) are combined with similar/contiguous categories.

For a key numeric variable, global recoding means replacing it by another variable which is a discretized version of the original. For example, years of age[18] may be grouped into age groups; taxes/income/salaries/number of animals heads/agricultural surfaces/etc. may be discretized. The new categories may be defined using a-priori defined or data-driven (e.g. quantiles/deciles) classes. When global recoding is applied, it is generally considered that the information loss for count variables (number of years, employees, number of heads, number of crops, number of tractors, etc.) may be acceptable. On the contrary, for continuous variables (taxes, turnover, value-added, budget, financial details, etc.), a large information loss due to recoding should be assumed.

Extremely detailed breakdown of categories which results in more than 20 categories should be avoided. In these cases, global recoding may be applied without any information loss since researchers are not interested in detail, but rather in some aggregated categories.

Global recoding is commonly applied to variables like age, marital status, dates, education, occupation (e.g. main job, previous job, etc.) and economic activity. Age is usually released in 5 or 10 year bands. Education, occupation and economic activity are often released up to two digits of their international standard classifications. Regarding dates, instead of disseminating the date of interview/birth/start or end of economic activities/etc., such information may be grouped into months/years, even measured from a reference date (i.e. time from an event). No huge information loss is generally introduced when global recoding is applied to dates.

Local recoding is similar to its global version, except that is applied only to risky records. While it preserves more information than global recoding, it might lead to unusual classifications which may be useless for statistical analyses.

Global recoding is a suitable method for different disclosure scenarios, i.e. nosy neighbour or register-based linkage scenarios. Global recoding also reduces the disclosure risk measured by means of threshold rules, e.g. k-anonymity, or the l-diversity. Anyway, when the “new” categories are too broad, the information loss might be too large.

Top/bottom coding

Top/bottom coding is a special case of global recoding which can be applied on continuous or categorical ordinal variables. The idea is that top values, i.e. those above a certain threshold, receive the value of the threshold itself. For example, any individual in survey microdata earning more than 10.000$ will have the income value replaced by 10.000$; any individual with age greater than 100 years will have the age replaced by 100 years. The same reasoning applies for bottom values (those values below a certain threshold).

Top/bottom coding could be also applied to count variables like number of heads, number of cars, etc. Attention should be paid those variables which could be derived from the data structure. For example, the variable household size might be top coded by creating a single maximum category, such as 10+. When the complete list of household members is also included in the microdata, such top coding would be useless.

---

[18] Age refers to time-related variables. Age may include age since full-education, age of the husband/wife, children’s age, time since starting/stopping an activity, etc.

Top coding is a suitable disclosure control method for financial and wealth variables, (e.g. property, pensions, etc.), i.e. in those cases where there is a high probability to deal with few “dominating” records. Incomes, salaries and related variables shall be top-coded. The threshold might be equal to ten times the average value in the microdata.

Sampling

Since surveys typically have very small sampling fractions, data curator could assume that a potential intruder would not know whether an individual is included in the survey microdata or not. Sampling therefore provides an inherent level of protection. The same principle could strengthen the protection level. Indeed, a random sample of the original set of records could be disseminated. Sampling increases the level of uncertainty.

Different sampling schemes may be applied, even taking into account inclusion probabilities inversely proportional with the disclosure risk. Sampling methods are suitable for categorical microdata, but they are less adequate for continuous variables because they leave completely unperturbed all the records in the microdata file. Sampling is particularly useful for disseminating micro datasets from censuses.

Perturbative disclosure limitation methods for key and sensitive variables

Unlike non-perturbative methods, perturbative methods change the values in a dataset. These methods introduce uncertainty to the dataset making it impossible for an intruder to be sure if he/she has made a disclosure. Since perturbative methods alter the microdata, it is important to check that the perturbed values are consistent with related values elsewhere in the dataset. For example, if a marital status is perturbed for a boy of age 6, then the intruder can deduce the perturbation. Another example may be if grown crops grown are perturbed, but are inconsistent with a geographical variable such as region; the perturbation can easily be reversed engineered undermining its purpose.

Data swapping

In data swapping (Dalenius and Reiss 1978) the basic idea is to transform a microdata file by exchanging values of key (or sensitive) variables among individual records. Two records having similar control variables are paired and the values of some variables are swapped, typically their geographical variables. For example, two individuals having the same sex, age, and marital status (control variables) would be paired and their place of work/ residence interchanged. Consequently, confidential information like income would be protected in register-based disclosure scenarios.

Records are generally exchanged in such a way that low-order frequency counts or marginals are maintained. Targeted swapping involves records at risk of disclosure and only few other records.

It is advisable to swap a small number of records and only after a number of variables have already been recoded or suppressed, and just a few records at risk remain. Indeed, the swapping rate might be lower if global recoding is applied first.

Rank swapping

Rank swapping (Greenberg 1987) is similar to data swapping, but it allows using a continuous variable to define pairs of records to be swapped. The pairs are those records that are close, i.e. within a certain range, to each other based on sorting defined by a continuous variable. Those records with close ranks on the sorted variable form the potential pairs for swapping and their sensitive/key values may be exchanged. A value can be swapped only with similar values.

For example, suppose the record #1 in Tables 3 and 4 is considered at risk of disclosure and that the variable Income is used as a sorting variable. As the closest record is record #2, by applying rank swapping, the sufficient nutrition status, i.e. the sensitive variable, of the first two records would be exchanged.

Table 3: Original dataset

Record ID	Age	Gender	Daily income	Nutrition status sufficient
1	15-24	Male	10	Y
3	15-24	Male	1,000	Y
2	>80	Female	12	N

Table 4: Perturbed dataset after rank swapping

Record ID	Age	Gender	Daily income	Nutrition status sufficient
1	15-24	Male	10	N
2	>80	Female	12	Y
3	15-24	Male	1,000	Y

Like data swapping, rank swapping preserves the univariate distributions. Moreover, preservation of relationships between variables may be controlled by the range parameter and by applying the rank swapping to several variables simultaneously. Rank swapping may be adapted to preserve other statistics[19].

In general, among data perturbation techniques, rank swapping has been identified as a particularly well-performing method in terms of the trade-off between disclosure risk and information loss.

Post-Randomisation Method (PRAM)

PRAM (Gouweleeuw et al 1998) is a kind of intended misclassification. For each record in a microdata file, the value of a categorical key variable maybe changed, following a predetermined probability mechanism, into another category of the same key variable. The transition matrix which specifies for each category of a variable a probability for each transition to another category should be a-priori defined. Then, for each record, a random number is generated from a uniform (0,1) distribution which determine whether or not the value is perturbed. PRAM is applied to each observation independently and the procedure is random.

For example, consider a dataset where there is a key variable that can take the values “urban”, “peri-urban”, or “rural”, and random number is drawn from a uniform distribution for each record as shown below:

Table 5 Original dataset prior to application of PRAM

Record ID	Area	Random number
1	Urban	.2

---

[19] R package sdcMicro.

2	Urban	.6
3	Rural	.4
4	Peri-urban	.8

Accordingly, the following transition matrix is defined where the first column shows the original value, and the values of the next three columns show the probability of transition to value of the column. In other words, approximately half of records with “Urban” will be unchanged, and the other half will be changed to “peri-urban”. “Urban” cannot ever be changed to “Rural”.

Table 6: Transition table defining probabilities of changing category

	Urban	Peri-urban	Rural
Urban	.5	.5	0
Peri-urban	.5	.5	0
Rural	0	0	1

The probabilities assigned in the transition matrix will have an effect on the final frequency counts. Accordingly, it is important that care is taken when considering the definition of these probabilities.

Based on the value of the random number, the value is perturbed as shown in the following table.

Table 7: Original dataset after the PRAM application

Record ID	Area	Random number	Perturbed Area
1	Urban	.2	Urban
2	Urban	.6	Peri-urban
3	Rural	.4	Rural
4	Peri-urban	.8	Urban

PRAM affects the edit rules. Some microdata consistencies could be enforced by appointing a transition probability of 0 to the illogical scores. The amount of inconsistencies may be controlled by the transition matrix; unfortunately, the edit rules only partially defined in terms of PRAM-related variables might fail and need to be corrected through edit and imputation operations.

Since PRAM uses a probability mechanism, a possible intruder could never be sure about the disclosed values: with a certain probability these were perturbed. PRAM is a suitable protection method when the number of key variables is large and the number of unique combinations is high.

The transition matrix should be disseminated; in order to make adequate inferences on a microdata file to which PRAM has been applied, the researchers need to implement complex changes to standard statistical methods.

Rounding

For a given continuous key variable, original values are replaced by rounded values. For example, agricultural areas like 1,003 km^2 should be released as 1,000 km^2. It is always a good practice to apply a “light” rounding, i.e. to less significant digits, to continuous variables; otherwise a respondent might recognize at least himself almost with certainty.

In multivariate keys, rounding is usually applied independently on each variable. The level of perturbation, hence the level of protection increases with the rounding base. Rounding may be also applied only to records at risk of disclosure and to the sensitive variables to reduce the risk of attribute disclosure.

Microaggregation

Microaggregation (Domingo-Ferrer and Torra 2003) is a perturbative method for continuous key variables. Application of this method leads to replacing individual values with values computed on small aggregates (microaggregates). The basic objective of microaggregation is to satisfy the k-anonymity rule. To obtain microaggregates, the records are combined to form groups of size at least k, for example by means of clustering algorithms which maximize the similarity within clusters. The usual value for k is 3. Ranking with respect to a numeric variable is another method than can be used to group records. The group dimension may be fixed or random. Then, for each variable, the average value over each group is computed and used to replace the original values. Instead of the average, other statistics may be used. Alternatively, the values of variables may be exchanged within each group.

Microaggregation is generally applied to the entire dataset, thus it implicitly considers that each record is at risk of disclosure. It may be adapted to be applied only on records at risk.

Microaggregation is applied simultaneously to all the continuous variables included in a multivariate key. The bias generated by microaggregation increases with the number of variables included in the key and with the dimension of groups. When microaggregation is applied to univariate key variables or independently on each variable in a multivariate key, it is called “individual ranking”. Microaggregation reduces more the disclosure risk, while individual ranking preserves more information in the continuous variables. When applied to variables having skewed distributions, e.g. earnings, microaggregation preserves the general shape of those distributions, especially when the grouping variables are correlated with the key. In general, less information loss occurs when the key variables are highly correlated[20].

Noise addition

Noise addition (Brand 2002) consists in adding (sometimes multiplying by) randomly generated numbers to the continuous key (or sensitive) variables values like taxes or income. For example, a random noise is generated from a Normal distribution with mean zero and a small variance for each record and this random noise is added to the individual’s value of income:

$$z_j = x_j + \epsilon_j \text{ where } \epsilon_j \sim \mathrm{N}(0, \sigma^2)$$

In practice, both correlated and uncorrelated noise addition may be applied. Both methods generally preserve the means of the key variables, but variances and covariances are generally not maintained by adding uncorrelated noise. On the contrary, correlated random noise can be added to continuous variables ensuring that not only means are preserved but also the variances and covariances (Fuller 1993).

The noise being added usually has mean zero and small variance, which suits well with continuous original data. No exact matching is possible with external files, but approximate (interval matching may still be feasible). When using a constant small variance, small values are strongly perturbed, while the large values remain less perturbed. Consequently, in its pure form, the noise addition method is not suitable for continuous variables with skewed distributions like expenditures, earnings, turnover, value-added, financial, etc. It is more suitable for normalized variables like labour productivity (value-added per employee), daily food consumption, agricultural production by km^2, mean oil consumption, etc.

The noise addition approach can also be modified to suit the following contexts as well:

- Additive noise could be generated within small homogenous sub-groups (Shlomo and De Waal 2008) in order to use different perturbation variance for each sub-group. Generating noise in sub-groups also causes less inconsistencies.
- Adding noise could be carried out on a selection of records, e.g. only on those records at risk of disclosure.
- Adding noise may be applied to the sensitive variables to reduce the risk of attribute disclosure.

In general, when dealing with skewed distributions, microaggregation or individual ranking offer a higher disclosure risk limitation than noise addition with small variances.

Model-based perturbation methods

In few words, the idea is to perform some regression analysis and to release the fitted values instead of the original ones. The main advantage is that relationships between variables are preserved. The fitted values may be released only for the records at risk of disclosure. When the regressions perform extremely well, model-based perturbation methods might not guarantee sufficient disclosure risk reduction. Shuffling is a kind of swapping using a regression model to determine which variables are swapped (Muralidhar and Sarathy 2006). Many regression models may be implemented in the SDC framework, e.g. linear, multivariate, multilevel, quantile-based. etc. While allowing a high degree of flexibility, model-based perturbation methods are rarely implemented in SDC software.

Special issues and additional considerations

Geo-references

When data providers need to preserve geo-references (Burgert et al 2013) in a dataset, different perturbative methods may be attempted. These methods are generally grounded on the previously described ideas, but they also consider the geographical distance between records. For example, risky records may have swapped their geo-references. The swapping may involve only the closest records or only records in contiguous areas or even entire contiguous areas, i.e. exchanging all the records in given areas.

Other methods include rounding, off-setting and microaggregation. Rounding consists in truncating the coordinates of the risky records to a specified number of significant digits; thus the points would appear at the vertices of a grid. Geographical off-setting is similar to adding noise. The geo-references of each risky record may be replaced by the geo-references of a randomly selected point within a circular buffer around the original point. Geographical microaggregation would consists in releasing the mean geo-reference for small and close groups of records.

Large households or large groups of units

No method is really adequate to protect large households (e.g. >10 members) while preserving the hierarchical data structure. The common approach is the suppression of such households from the microdata file. Alternatively, only some household members could be suppressed (partial suppression); re-computation of the household structure/size and of sampling weights is subsequently mandatory. Large households could also be split, but this approach is rarely used in official statistics.

The threshold and approach used for defining large households will be based on the dataset. The same reasoning will be applied to large groups of statistical units, when the group structure/size may be inferred from the data.

Free text

This kind of information may be registered as responses to non specific questions in surveys. This information may contain personal details or opinions. As free text could hardly be the aim of any scientific research, it should be removed from the microdata file. In exceptional cases, text mining techniques should be applied to recover some structured information to disseminate.

Weights

Sampling weights can provide additional identifying information to intruders, especially when the weights are very close to 1 and the reference population is a large population (this is the case of large samples from census data, for example). Identification of sampling weights approximatively equal to 1 should always be performed. If needed, a sub-sampling and subsequent weight calibration should be performed before microdata dissemination.

Age of the data

Elapsed time since the survey reference period should also be considered when releasing microdata files as large temporal differences represent additional protection. Indeed, outdated information is less “accessible” and it might be also considered as less “interesting”.

Longitudinal surveys

Longitudinal microdata sets are more at risk because the same individuals are surveyed several times and some of their characteristics naturally change, e.g. age. Special attention should be paid when publicly releasing them. In longitudinal surveys, the overall anonymization strategy must be developed before the entire survey ends, ideally in occasion of the first microdata release. Since one of this strategy's objectives is to define the variables to be released, assumptions must be made about how those variables evolve over time, particularly if some of them might become key variables, e.g. age. Recoding categories, classifications and thresholds should be maintained over the survey waves.

Edit rules and data structures

Perturbative methods might destroy the edit rules, e.g. crops not available in some regions, marital status of children under 5 years old, etc. Before microdata dissemination, a screening for illogical records should be performed; such records should be corrected by imputation methods and subsequent disclosure risk evaluation.

When applying SDC methods, the preservation of different types of data structures is not always guaranteed. The impact of SDC methods on relationships between units and variables must always be checked before microdata dissemination.

The hierarchical grouping of units, e.g. households, holdings, crops, etc. might be affected by SDC methods. For example, if individual unit is removed, the structure related variables, e.g. size, classifications, totals, etc., are no longer valid and they should be updated. The relationships between variables are generally affected by perturbative methods. Indeed, any modification of a single variable impacts on the relationship validity. Examples of such relationships include:

- group information related to individual information: household income as sum of individual's income, financial details a groups of enterprises as sums of financial details of each business, group classifications as derived from individual information, etc. As SDC methods are commonly applied to individual records, the group information has to be updated and checked for disclosure risk
- relationships between variables: total number of heads as sum of heads by species, total agricultural area as sum of different areas, annual income as sum of monthly incomes, ratio of variables, highly correlated variables, etc. After the application of the SDC methods, the known relationships between variables should be checked.

IV. How to evaluate protected dataset

The evaluation of SDC protected datasets is an integral part of any disclosure control process. A microdata file should always be evaluated considering the disclosure risk and data utility trade-off illustrated in Figure 1. While methods for assessing and reducing the disclosure risk were detailed in previous sections, this section illustrates the main approaches to data utility evaluation.

The quality evaluation is generally performed by a) assessing the information loss resulting from the application of statistical disclosure limitation procedures and b) simulating the analyses that will be performed by the end-users (data utility). In practice, both information loss and data utility measures are used by data providers and users. Data providers use these methods to make informed decisions about the optimal positioning on the Disclosure Risk and Data Utility Map (Figure 1), while users are helped to understand whether the applied SDC methods are ignorable or not in their statistical analyses. Accordingly, data curator in OCS will use data utility measures to evaluate a dataset prior to release while users will use them to implement possible adjustments when carrying out statistical analyses on protected data. The same data utility assessments may be used by both data curator and end-users.

The variables considered in the data utility evaluation step do not necessarily coincide with the key variables defined and used in the disclosure risk assessment and protection stage. Even if a variable was not at all modified, due to its relationships with other protected variables, it might be included in the data utility evaluation. For each microdata file to be released, the OCS and data providers will agree on the list of variables used to assess the data utility.

There are three main types of utility measures: 1. Coherence with published statistics 2: generic and 3: data use specific. Each type of measure will be described in this section. Notably, all of these data utility measures are relative not absolute; i.e., they measure the utility of the data compared with the original data.

Coherence with published statistics

A first category of utility measures concerns the coherence with already published statistics, e.g. reports, tabular data, official statistics, etc. Data providers will provide OCS with a list of the most important indicators that have been computed and published using a dataset. Consequently, OCS can make a direct comparison between the outputs derived from original and protected microdata files. The results generated from the protected dataset should be as close as possible and within an a-priori defined margin of error of the published statistics.

Generic measures

Generic measures evaluate how a collection of basic statistics (first, second moments, etc.) is affected by the application of disclosure limitation methods. The most general methods for data utility evaluation assess the distance between the original and protected data using some discrepancy function.

If the structure of the protected data set is very similar to the structure of the original data set, a high data utility may be assumed. In fact, the main purpose for preserving the structure of the data set is to ensure that the masked data set will be analytically valid and interesting. Additionally, except for direct identifiers, removal of any variable from the protected microdata file is generally considered a large information loss.

In the case of suppressed data, the suppression rate indicating the number of values suppressed compared to those released is often used as a data utility indicator. Data curator in OCS will produce these rates at different detail levels and for various respondent groups; this would avoid the complete information suppression for entire subpopulations. At least the total suppression rate should be provided to users. Only suppressions derived from the SDC process should be reported; other types of missing data should be reported elsewhere.

Generic data utility measures may be classified according to their level of comparison. A first set of methodologies is derived from the comparison of raw records in the original and the protected dataset. The more similar the protection method to the identity function, the lesser the SDC impact (and higher disclosure risk). This requires pairing records in the original dataset and records in the protected dataset. A second set is represented by those methodologies based on comparing individual distributions of each modified variable. Finally, the third class of data utility measures represents those measures incorporating information about more variables. In practice, different measures from each class are computed and analysed, even for several subdomains.

Generic data utility measures are generally defined independently for continuous and discrete variables. Only few measures may be defined for both categorical and numeric variables.

Analytical validity

Considering the entire dataset, a protected microdata set is analytically valid (Winkler 1998) if it approximately preserves the following with respect to the original data:

a. means and covariances on a small set of subdomains

b. marginal values for a few tabulations of the data

c. at least one distributional characteristic.

The subdomains (a) or variables (b and c) to be included in the data utility analysis should be selected according to their importance. This selection should be subject of an agreement between OCS and data provider.

Generic utility measures for continuous variables

For each perturbed continuous variable, descriptive statistics, e.g. minimum, maximum, mean and median, on the change derived from the perturbation process, should be computed and analysed in order to possibly identify outlying perturbations (Trent et al 2010). The sum of the absolute distances (Yancey et al 2002) between the corresponding observations in the raw and anonymized datasets represents another data utility measure included in standard SDC software:

$$IL1 = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - z_{ij}|}{\sqrt{2}S_j}$$

where p is the number of perturbed continuous variables; n is the number of records in the dataset; x_{ij} and z_{ij} are the values before and after anonymization for variable j and unit i, respectively; S_j is the standard deviation of variable j in the original data. $IL1$ should also be computed on subdomains.

For each perturbed continuous variable, the mean absolute difference between ranks of each observation indicates significant differences between the distributions of the original and protected data. A comparison of the rank-based scores computed for each perturbed variable should identify possible changes in the relationships between variables.

For each perturbed continuous variable, besides the computation of the perturbation rate, quantiles and means could be compared in order to assess the preservation of their distributional characteristics. Graphical tools such as boxplots and quantile-quantile plots, could be efficiently used to assess the degree of equivalence between the original and perturbed distributions. Alternatively, the Kolmogorov-Smirnov test may be used. Such goodness of fit tests (Gibbons and Chakraborti, 1992) provide only a qualitative indication about the agreement between two (empirical) distributions. They can be used to test whether two distributions can be considered equivalent, but no quantitative measure can be extrapolated.

In case of many continuous variables, discrepancy measures between covariance matrices, correlation matrices, principal component matrices, etc. could be computed. In case of small differences, a large data utility may be assumed. Matrix discrepancy can be measured in many ways, for example (Domingo-Ferrer et al 2001):

- Mean square error: sum of squared differences between pairs of matrices, divided by the number of cells in either matrix;
- Mean absolute error: sum of absolute differences between pairs of matrices, divided by the number of cells in either matrix;
- Mean variation: sum of absolute percentage variation of differences between pairs of matrices, divided by the number of cells in either matrix.

For example, if COV_O and COV_P are the covariances matrices estimated using the original and protected subset of continuous variables, the three discrepancy measures between the two matrices become:

- Mean square error $$= \frac{\sum_{j=1}^{N} \sum_{1 \le i \le j} (COV_O^{ij} - COV_P^{ij})^2}{N(N+1)/2}$$
- Mean absolute error $$= \frac{\sum_{j=1}^{N} \sum_{1 \le i \le j} |COV_O^{ij} - COV_P^{ij}|}{N(N+1)/2}$$
- Mean variation $$= \frac{\sum_{j=1}^{N} \sum_{1 \le i \le j} |(\frac{COV_O^{ij} - COV_P^{ij}}{COV_O^{ij}})|}{N(N+1)/2}$$

Obviously these types of data utility measures can only be used if the same continuous variables exist in the original and protected dataset (i.e. continuous variables are not suppressed or recoded). Additionally, a global data utility score (Domingo-Ferrer and Torra 2001) may be computed by taking the mean of the mean variations computed for different statistics, i.e. correlation, covariance, variance, etc. Such score would vary between 0% and 100%. The interpretation of values between both scales is as follows: small information loss from 0% to 10%, medium information loss from 11% to 20%, serious information loss from 21% to 30% and no data utility for 31% and over.

Generic utility measures for categorical variables

For categorical data three kinds of data utility measures may been considered: direct comparison of categorical values, comparison of contingency tables and entropy-based measures.

Direct comparison of original and protected categorical variables requires the definition of a distance for categorical variables. Definitions consider only the distances between pairs of categories that can appear when comparing a record and its protected version. When the range of a variable is an ordinal scale, the distance between category *a* and *b* is proportional to the number of categories between *a* and *b*. When the range of a variable is not ordinal, the distance is 1 if the values are different and 0 if they are not. Subsequently, the distance between original and protected units may be computed as the sum of distances of their corresponding attributes. In essence, a larger distance indicates a large information loss.

For categorical data, besides the previous distance based measures and goodness of fit tests, several measures of association (Agresti 2002) could also be used. Such association measures are based on the definition of concordant pairs (observations belonging to the same category in both the original and protected variables). Gamma and Goodman statistics are the most used in practice (Agresti 2002). Distance-based or concordance-based measures may be computed when the same categories appear in both original and protected data. Thus these measures are more suitable for protection methods like data swapping or PRAM and not suitable for disclosure limitation methods like global recoding.

An alternative to directly comparing the values of original and protected categorical variables is to compare their contingency tables. For a given subset of categorical variables, their corresponding contingency tables may be computed for a microdata file before and after applying the protection methods. The number of differences between both contingency tables is denoted by CTBIL (Contingency Table Based Information Loss (Domingo-Ferrer et al 2001). As the number of cells in a contingency table depends on the number of categories in the variable, a normalised expression of CTBIL is usually considered. Contingency tables may be analysed by means of the Pearson Chi-square test, too. Only pairing of categories is necessary for the application of this measure; thus a contingency table approach may be applied for almost all disclosure limitation methods detailed in the previous section.

The third approach for measuring data utility for categorical data is based on the mathematical concept of entropy. This approach may also be applied for continuous variables. To measure the difference in the information content between the original and protected dataset, the Shannon's entropy change might be considered. Formulas of information loss measures based on conditional entropy may be derived for different data protection techniques[21] (Willenborg and deWaal 2001). However, the entropy based measures have also issues when dealing with data sparsity. Theoretically, being formal measures, entropy based measures can be computed for any protection method. In practical applications, their computational cost is high. Moreover, from a data utility point of view, it is not clear how such measures are related to the surveyed phenomena or to the user requirements.

Data use-specific utility measures

The utility of microdata that has undergone disclosure limitation methods is based on whether the same statistical analyses and inferences may be drawn on perturbed data and original data. The general idea is to compare statistical analyses using both original and protected data. If the differences are not too large, the data utility may be considered good. To assess data utility in microdata files, proxy measures include measuring distortions to distributions and the impact on bias, variance and other statistics. The comparison of statistics may be performed by means of percent relative differences, for example.

Ideas for selecting the statistical analyses to be used in data utility assessment step generally come from the literature review or from reports researchers publish based on similar previously released microdata. OCS and the data provider should discuss and agree on which statistical analyses are likely to be performed by the data users. In general, both the outputs and model assessment statistics will be compared.

Some protection methods have a known impact on statistical analyses (Abowd and Schmutte 2015):

- adding random noise generated from a Normal distribution with a mean of zero and a small variance, has no impact on the point estimate of a total or mean, but increases the variance and cause a wider confidence interval
- microaggregation decreases the variance and cause a narrower confidence interval
- swapping and PRAM preserve few marginal distributions
- swapping variables (geography) on a sub-sample of records introduces biases in joint distributions
- top and bottom coding change the quantiles, thus introducing some distributional bias

This known impact should be quantified and reported to the users, at least in generic formula.

When evaluating data utility, statistics different from the ones preserved by the protection methods should be used. The following are three methods for evaluating differences, but many more may be developed (Rinott and Shlomo 2007; Gomatam and Karr 2003; Hundepool et al 2012) depending on the context.

Impact on measures of association

The test of independence for a two-way table having R rows and C columns is based on the Pearson Chi-square statistic $X^2 = \sum_i \sum_j \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ where o_{ij} is the observed count while $e_{ij} = (n_{i.} n_{.j})/n$ is the expected count. Typically, the Cramer's V $CV = \sqrt{\frac{X^2/n}{min(R-1,C-1)}}$ is used as a measure of association between categorical variables. The information loss measure that is derived is the percent relative difference between original and perturbed data:

$$RCV = 100\,\frac{CV_{pert} - CV_{orig}}{CV_{orig}}$$

For multiple dimensions, log-linear modelling is often used to examine associations. A similar measure to RCV can be calculated by taking the relative difference in the deviance obtained from the estimated model based on the original and perturbed microdata, respectively.

Impact on a regression analyses

For continuous variables, it is useful to assess the impact on the correlation and in particular the R^2 of a regression or ANOVA analyses. For example, in an ANOVA analysis, the test involves checking whether a continuous dependent variable has the same mean across groups defined by a categorical explanatory variable. The goodness of fit criterion R^2 is surely a statistic whose relative difference should be assessed. Another utility measure is based on assessing differences in the means of a response variable across categories of an explanatory variable having K categories. Let $\bar{y}_k$ be the mean in category k and define the "between" variance of this mean by: $B(\bar{y}) = \frac{1}{K-1}\sum_k (\bar{y}_k - \bar{y})^2$. Then, the information loss may be measured by $RB = 100\,\frac{B_{pert}-B_{orig}}{B_{orig}}$.

Additionally, a comparison of estimates of coefficients when applying a regression model on both the original and perturbed microdata is usually performed by means of the overlapping degrees of their confidence intervals. In general, for other types of regression models, e.g. logistic, Poisson, multivariate linear or nonlinear models, hierarchical, etc., the same approach should be followed: compare both the estimates and some model assessment statistics like R^2 or between variances. If necessary, regressions should also be performed by subdomains.

Impact on clustering methods

Depending on the clustering method and on whether the disclosure limitation method increases or decreases the variances, the number of clusters might increase or decrease. Consequently, a simple comparison on the number of clusters obtain using original and protected data should be performed. Moreover, when the composition of clusters is very much changed, a large information loss may be assumed. Other clustering statistics that may be compared include variances and shapes of individual clusters or minimum distance to other clusters.

Documenting the SDC Process

For archiving purposes, it is important that detailed records are kept regarding the SDC strategy applied to datasets for a variety of reasons. For example, authorized internal users may wish to better understand the SDC process that was applied, and internal auditing may help improve SDC processes. Finally, it is important that all users of a dataset are aware of at least general terms of how the dataset has been altered during the SDC process. As mentioned in the previous sections, some methods such as local suppression can result in non-random missing values. Also, due to the application of SDC, tabulations using a protected dataset may not match exactly the results of some already published report. An overview of the methods applied to the dataset during the SDC process can help the users understand these discrepancies. In some cases it is even possible for users to adjust their analyses in order to take into account the “distortions” introduced by the SDC methods.

This section will provide an overview of the steps data curator in OCS will take to document the SDC process, and who will be allowed access.

Documenting SDC process in internal archive

As referenced previously, data providers will remove direct identifiers and extremely sensitive variables prior to submitting the dataset to OCS for dissemination. Upon receipt, data curator in OCS will create a corresponding file in the internal archive and carry out the steps described in the Micro and Meta data curation and dissemination protocol.

For quality assurance, and risk assessment, the data curator will follow a code template referred to as the “Microdata processing report” (MPR) developed in RMarkdown which will ensure that all processes are fully documented, integrated directly with the code, and reproducible. The MPR will include the following sections:

- Configuration: Loading packages, datasets, and setting working directory
- Data Quality Checks: Defined in Section 3.2.1 of the Micro and Meta data curation and dissemination protocol
- Risk Assessment: Description of disclosure scenarios, set key and sensitive variables, and measure record and global risk
- Application of SDC methods, evaluation of dataset: Iterative process of applying SDC methods, and evaluating disk risk and utility

- Summary of changes: Table listing variables which have changed, number of records changed, the SDC methods applied, and any parameters; information loss/data utility measures to be reported to the end-users.

After the process is complete, the MPR will be generated as a .pdf file and saved in the directory of the internal archive.

SDC process information in published metadata

Providing limited information on the SDC process to external users is important for transparency, and analysis. This information will be provided through the addition of a DDI element to the published metadata. A metadata element called, "Anonymization Process" will be included as an element in the metadata on data processing.

This metadata element will include a summary of the methods applied, and potential impacts on the analytical potential of the data identified during the evaluation. If important variables are changed which could result in tabulations that differ from official results, this will be noted.

Finally, a list of variables that were altered/removed during the SDC process will be included with the following information for each:

- Methods applied
- Parameters (if applicable)
- Approximate number of changed records

This should provide the user with enough relevant information to consider in their analyses, but not enough to allow any reverse engineering to commit an intrusion. If users request information, OCS will consult with the data provider prior to releasing any additional information.

IV. Overview of SDC workflow

The first previous sections of this protocol describe the main methods and techniques that data curator in OCS will use for preparing micro datasets for dissemination, and documenting the process. This section defines the main steps in the SDC workflow referring to those concepts and tools. Throughout the steps, identity disclosure will be the type of disclosure considered. The workflow will be subject to updates to improve efficiency and methods.

The workflow is divided in 5 main steps:

6. Removal of direct identifiers and extremely sensitive variables – Data provider
7. Definition of key variables, disclosure scenarios, preferred terms of access[22], and published statistics – Data provider
8. Measure risk and apply disclosure limitation methods – Data curator
9. Evaluate protected dataset and document – Data curator and Data Provider
10. Approval by the Chief Statistician and release of the anonymized microdata file - OCS

There is some overlap between steps 3 and 4 because measuring risk, applying methods, and evaluate occur in cycles.

Step 1: Removal of direct identifiers and extremely sensitive variables

As a first step for preparing a dataset, the data provider shall remove all direct identifiers and extremely sensitive variables. In most cases, direct identifiers are not needed for performing statistical analyses, and the retention of this information exposes OCS to unjustified liability. The data provider should also remove extremely sensitive variables for the same reason. If the data provider has any doubts or difficulties identifying these variables, the data provider can contact OCS for support in this initial step. However any support provided will take place at the data provider workstation so that no data is transferred to OCS with direct identifiers or extremely sensitive variables.

Step 2: Definition of key variables, disclosure scenarios, preferred terms of access, and published statistics

Once the data provider is certain that step 1 is complete, he/she will submit the dataset through a data deposit system. The data deposit system consists of a form containing fields to be completed by the data provider, as well as a space to upload datasets, and other resources. The fields will include the following information: key variables, disclosure scenarios, sensitive variables and the preferred terms of access. Additionally, each field will contain relevant definitions, instructions, and explanations to assist the data provider. Finally, reports and other tables can be uploaded for comparison with published statistics. If code is available used to generate the published statistics, this can also be provided.

The field for the key variables will require that the variables are entered in order of importance so that data curator in OCS can make informed decisions about applying methods such as suppression. There will also be a field to describe potential disclosure scenarios including any known registers or datasets containing key variables for the study population and may therefore be used for record linkage or matching. The data provider will also provide a list of sensitive variables, and indicate the preferred terms of access.

Step 3: Measure risk and apply disclosure limitation methods

When data curator in OCS receives a dataset, the first SDC related action will be to compare the list of key variables provided to a list of common key variables such as those listed in Section 2. Data curator will also check the potential sensitive variables. If additional key and sensitive variables are found, OCS will request the data provider to review their submission. Once the lists of key and sensitive variables are finalized, data curator in OCS will enter the information into MPR Rmarkdown template described in Section 5.

Using the key and sensitive variables, data curator will measure the risk using individual risk methodology, k-anonymity, and l-diversity approaches, for example. The threshold of 3-k will serve as a general guide for sample surveys, but is subject to the judgement of the data curator based on the sample size, and the availability of external registers. Following the risk assessment, data curator will apply the methods described in Section 3 according to the variable type (i.e. categorical or continuous), and re-measure the risk after each iteration. As a final component of this step, data curator will check to ensure that no illogical relationships, or changes to edit rules occurred as a result of SDC.

Step 4: Evaluate protected dataset and document

Finally, data curator will do a final evaluation of the dataset considering data utility and information loss. The first method of evaluation will be to assess the coherence of published statistics provided in Step 2. Then, information loss measures will be applied to categorical and continuous variables. When data curator is satisfied that an appropriate point on the Risk/Utility Map (Figure 1) has been reached, the MPR will be finalized and a .pdf will be generated and saved to the internal archive.

The MPR and the protected dataset provided to the data provider for review. If based on the SDC process, data curator determines that the preferred terms of access are not appropriate, and will provide a justification. The data provider will have 1 month to provide feedback.

Step 5: Approval by the Chief Statistician and release of the anonymized microdata file

Data curator in OCS will provide the MPR to the Chief Statistician for approval. Once approved, the micro dataset and metadata will be uploaded into the microdata dissemination platform.

References

Agresti, A. (2002) Categorical Data Analysis, John Wiley & Sons, Inc., Hoboken, New Jersey.

Abowd, L., Schmutte, I.M. (2015) Economic Analysis and Statistical Disclosure Limitation, Brookings Papers on Economic Activity.

Brand, R. (2002). Microdata protection through noise addition. In J. Domingo-Ferrer, editor, Inference Control in Statistical Databases, volume 2316 of LNCS, pages 97-116, Berlin Heidelberg, 2002. Springer.

Burgert, C.R., Colston, J., Roy, T. and Zachary, B. (2013). "DHS Spatial Analysis Report No. 7 - Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys" (USAID). http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf

Dalenius T., Reiss, S. P. (1978). Data-swapping: a technique for disclosure control (extended abstract). In Proc. of the ASA Section on Survey Research Methods, pages 191–194, Washington DC, 1978. American Statistical Association.

Domingo-Ferrer, J., Mateo-Sanz, J. and Torra, V. (2001). Comparing SDL Methods for Micro-Data on the Basis of Information Loss and Disclosure Risk. ETK-NTTS Proceedings of the Conference, Crete, .

Domingo-Ferrer, J., Torra, V. (2001) A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111-134. North-Holland (2001).

Domingo-Ferrer, J. and Torra, V. (2003). Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage, Statistics and Computing, Vol. 13, No. 4, 343-354.

Domingo-Ferrer, J., and Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 11(2):195-212, 2005.

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001). Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory.

Elamir, E. and Skinner, C.J. (2006). Record-Level Measures of Disclosure Risk for Survey Micro-data. Journal of Official Statistics, 22, 525-539.

Elliot, M. J., Manning, A., Mayes, K., Gurd, J., and Bane, M. (2005). SUDA: A Program for Detecting Special Uniques. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Geneva.

Fuller, W. A. (1993). Masking Procedures for Micro-data Disclosure Limitation. Journal of Official Statistics, 9, 383-406.

Gibbons, J. D., Chakraborti, S. (1992). Nonparametric Statistical Inference. Marcel Dekker, New York, 3rd edition.

Gomatam, S. and Karr, A. (2003). Distortion Measures for Categorical Data Swapping. Technical Report Number 131, National Institute of Statistical Sciences.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998). Post Randomisation for Statistical Disclosure limitation: Theory and Implementation. Journal of Official Statistics, 14, 463-478.

Greenberg, B. (1987). Rank swapping for ordinal data, Washington, DC: U. S. Bureau of the Census (unpublished manuscript).

Hillesland M. & Mwaniki P.M. (2018). Measuring inadequate employment in Kenya: field test report for Decent Work within an agricultural context in developing countries. Technical Report no. 35. Global Strategy Technical Report: Rome.

Hundepool et at. (2010). Handbook on Statistical Disclosure Control. Essnet SDC, https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and Wolf, P.-P. (2012) Statistical Dislcosure Control. Chichester: Wiley and Sons.

Hundepool et al. (2014). μ-Argus manual, http://neon.vb.cbs.nl/casc/.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-Diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1(1), 3.

Muralidhar, K., & Sarathy, R. (2006). Data Shuffling- A New Masking Approach for Numerical Data. Management Science, 658-670.

Rinott, Y. and Shlomo, N. (2007). Variances and Confidence Intervals for Sample Disclosure Risk Measures. 56th Session of the International Statistical Institute Invited Paper, Lisbon 2007.

Shlomo, N. and De Waal T. (2008). Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. Journal of Official Statistics, 24, No. 2, 1-26.

Shlomo, N., Skinner, C.J. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Annals of Applied Statistics, Vol. 4, No. 3, 1291-1310.

Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Micro-data Using Log-linear Models. Journal of American Statistical Association, Vol. 103, Number 483, 989-1001.

Sweeney.L. (2002). k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10, 2002; 557-570.

Trent, A. J., Davern, M., Stevenson, B.. (2010). Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications. Public Opinion Quarterly 74, no. 3: 551–69.

Yancey, W. W., Winkler, W. E., & Creecy, R. H. (2002). Disclosure Risk Assessment in Perturbative Microdata Protection. Research Report Series, Statistics 2002-01.

Willenborg, L. and De Waal, T. (2001). Elements of Statistical Disclosure limitation in Practice. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Winkler, W., (1998), Re-identification methods for evaluating the confidentiality of analytically valid microdata, in Statistical Data Protection, Luxembourg: Office for Official Publications of the European Communities, 1999.

United Nations Statistical Commission. (2014) “Microdata dissemination best practices.”. Accessed at: https://unstats.un.org/unsd/accsub-public/microdata.pdf

---