WP. 45
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vii): General statistical confidentiality issues

GLOSSARY ON STATISTICAL DISCLOSURE CONTROL

Invited Paper

Submitted by Statistics Netherlands, Statistics Canada, Destatis Germany and University of Manchester, United Kingdom[1]

[1] Prepared by Anco Hundepool and Eric Schulte Nordholt, Statistics Netherlands; Jean-Louis Tambay, Statistics Canada; Thomas Wende, Destatis Germany and Mark Elliot, University of Manchester, United Kingdom.

GLOSSARY ON STATISTICAL DISCLOSURE CONTROL

Mark Elliot (University of Manchester)
Anco Hundepool (Statistics Netherlands)
Eric Schulte Nordholt (Statistics Netherlands)
Jean-Louis Tambay (Statistics Canada)
Thomas Wende (Destatis, Germany)

Version August 2005

Introduction

At the Joint ECE / Eurostat Work Session on Statistical Data Confidentiality (7-9 April 2003) in Luxembourg the idea for a glossary on Statistical Disclosure Control was launched. The five people who produced this glossary were present at that Work Session and also met on 18 August 2003 at the ISI Session in Berlin. This new glossary on Statistical Disclosure Control will be presented at the next Joint ECE / Eurostat Work Session on Statistical Data Confidentiality (9-11 November, 2005) in Geneva. In the meantime preliminary versions have been presented so that experts in the field from all over the world could comment on these versions. The aim of this glossary is twofold: firstly, it should help people who are new in the field to get acquainted with the terminology used in Statistical Disclosure Control and secondly it can be used in courses on Statistical Disclosure Control as a back-up facility. We hope that this glossary will be useful and the two aims will be reached. If you have any comments or questions, please forward them to Eric Schulte Nordholt (e-mail: ESLE@CBS.NL) so that they can be taken into account for future versions.

Acknowledgements

The authors acknowledge the input of many people during the drafting of this document, especially: Paul Feuvrier (INSEE), Kingsley Purdam (University of Manchester), Barry Schouten (Statistics Netherlands), Duncan Smith (University of Manchester) and Peter-Paul de Wolf (Statistics Netherlands).

A

Ambiguity rule: Synonym of (p,q) rule.

Analysis server: A form of **remote data laboratory** designed to run analysis on data stored on a safe server. The user sees the results of their analysis but not the data.

Anonymised data: Data containing only anonymised records.

Anonymised record: A record from which direct identifiers have been removed.

Approximate disclosure: Approximate disclosure happens if a user is able to determine an estimate of a respondent value that is close to the real value. If the estimator is exactly the real value the disclosure is exact.

Argus: Two software packages for Statistical Disclosure Control are called Argus. µ-Argus is a specialized software tool for the protection of **microdata**. The two main techniques used for this are **global recoding** and **local suppression**. In the case of **global recoding** several categories of a variable are collapsed into a single one. The effect of local suppression is that one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. Both **global recoding** and **local suppression** lead to a loss of information, because either less detailed information is provided or some information is not given at all. τ-Argus is a specialized software tool for the protection of tabular data. τ-Argus is used to produce safe tables. τ-Argus uses the same two main techniques as µ-Argus: global recoding and local suppression. For τ-Argus the latter consists of suppression of cells in a table.

Attribute disclosure: Attribute disclosure is **attribution** independent of identification. This form of disclosure is of primary concern to **NSIs** involved in **tabular data** release and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place. Minimally, the presence of a single zero within a table means that an intruder may infer from mere knowledge that a population unit is represented in the table and that the intruder does not possess the combination of attributes within the cell containing the zero.

Attribution: Attribution is the association or disassociation of a particular attribute with a particular population unit.

B

Barnardisation: A method of disclosure control for tables of counts that involves randomly adding or subtracting 1 from some cells in the table.

Blurring: Blurring replaces a reported value by an average. There are many possible ways to implement blurring. Groups of records for averaging may be formed by matching on other variables or by sorting on the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or to the "middle" member (as in a moving average). It may be performed on more than one variable with different groupings for each variable.

Bottom coding: See **top and bottom coding**.

Bounds: The range of possible values of a cell in a table of frequency counts where the cell value has been perturbed or suppressed. Where only margins of tables are released it is possible to infer bounds for the unreleased joint distribution. One method for inferring the bounds across a table is known as the **Shuttle algorithm**.

C

Calculated interval: The interval containing possible values for a suppressed cell in a table, given the table structure and the values published.

Cell suppression: In tabular data the cell suppression SDC method consists of **primary and complementary (secondary)**

suppression. Primary suppression can be characterised as withholding the values of all risky cells from publication, which means that their value is not shown in the table but replaced by a symbol such as '×' to indicate the suppression. According to the definition of risky cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or presenting a case of **dominance** have to be primary suppressed. To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called **complementary (secondary) suppression**. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the risky cells with the least amount of suppressed information.

Complementary suppression: Synonym of **secondary suppression**.

Complete disclosure: Synonym of **exact disclosure**.

Concentration rule: Synonym of **(n,k) rule**.

Confidentiality edit: The confidentiality edit is a procedure developed by the U.S. Census Bureau to provide protection in data tables prepared from the 1990 Census. There are two different approaches: one was used for the regular Census data; the other was used for the long-form data, which were filled by a sample of the population. Both techniques apply statistical disclosure limitation techniques to the **microdata** files before they are used to prepare tables. The adjusted files themselves are not released; they are used only to prepare tables. For the regular Census microdata file, the confidentiality edit involves "data swapping" or "switching" of attributes between matched records from different geographical units. For small blocks, the Census Bureau increases the sampling fraction. After the microdata file has been treated in this way, it can be used directly to prepare tables and no further disclosure analysis is needed. For long form data, sampling provides sufficient confidentiality protection, except in small geographic regions. To provide additional protection in small geographic regions, one household is randomly selected and a sample of its data fields are blanked and replaced by imputed values.

Controlled rounding: To solve the additivity problem, a procedure called controlled rounding was developed. It is a form of **random rounding**, but it is constrained to have the sum of the published entries in each row and column equal to the appropriate published marginal totals. Linear programming methods are used to identify a controlled rounding pattern for a table.

Controlled Tabular Adjustment (CTA): A method to protect tabular data based on the selective adjustment of cell values. Sensitive cell values are replaced by either of their closest safe values and small adjustments are made to other cells to restore the table additivity. Controlled tabular adjustment has been developed as an alternative to **cell suppression**.

Conventional rounding: A disclosure control method for tables of counts. When using conventional rounding, each count is rounded to the nearest multiple of a fixed base. For example, using a base of 5, counts ending in 1 or 2 are rounded down and replaced by counts ending in 0 and counts ending in 3 or 4 are rounded up and replaced by counts ending in 5. Counts ending between 6 and 9 are treated similarly. Counts with a last digit of 0 or 5 are kept unchanged. When rounding to base 10, a count ending in 5 may always be rounded up, or it may be rounded up or down based on a rounding convention.

D

Data divergence: The sum of all differences between two datasets (data-data divergence) or between a single dataset and reality (data-world divergence). Sources of data divergence include: data ageing, response errors, coding or data entry errors, differences in coding and the effect of disclosure control.

Data intruder: A data user who attempts to disclose information about a population unit through **identification** or **attribution**.

Data intrusion detection. The detection of a **data intruder** through their behaviour. This is most likely to occur through analysis of a pattern of requests submitted to a **remote data laboratory**. At present this is only a theoretical possibility, but it is likely to become more relevant as **virtual safe settings** become more prevalent.

Data Intrusion Simulation (DIS). A method of estimating the probability that a **data intruder** who has matched an arbitrary population unit against a sample unique in a target microdata file has done so correctly.

Data protection: Data protection refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into respondents' privacy caused by the collection, storage and dissemination of personal data.

Data swapping: A disclosure control method for microdata that involves swapping the values of variables for records that match on a representative key. In the literature this technique is also sometimes referred to as "multidimensional transformation". It is a transformation technique that guarantees (under certain conditions) the maintenance of a set of statistics, such as means, variances and univariate distributions.

Data utility: A summary term describing the value of a given data release as an analytical resource. This comprises the data's **analytical completeness** and its **analytical validity**. Disclosure control methods usually have an adverse effect on data utility. Ideally, the goal of any disclosure control regime should be to maximise data utility whilst minimising disclosure risk. In practice disclosure control decisions are a trade-off between utility and **disclosure risk**.

Deterministic rounding: Synonym of **conventional rounding**.

Direct identification: Identification of a statistical unit from its **formal identifiers**.

Disclosive cells: Synonym of **risky cells**.

Disclosure: Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: **identification** and **attribution**.

Disclosure by fishing: This is an attack method where an intruder identifies risky records within a target data set and then attempts to find population units corresponding to those records. It is the type of disclosure that can be assessed through a **special uniques analysis**.

Disclosure by matching: Disclosure by the linking of records within an identification dataset with those in an anonymised dataset.

Disclosure by response knowledge: This is disclosure resulting from the knowledge that a person was participating in a particular survey. If an intruder knows that a specific individual has participated in the survey, and that consequently his or her data are in the data set, identification and disclosure can be accomplished more easily.

Disclosure by spontaneous recognition: This means the recognition of an individual within the dataset. This may occur by accident or because a data intruder is searching for a particular individual. This is more likely to be successful if the individual has a rare combination of characteristics which is known to the intruder.

Disclosure control methods: There are two main approaches to control the disclosure of confidential data. The first is to reduce the information content of the data provided to the external user. For the release of tabular data this type of technique is called **restriction-based disclosure control method** and for the release of microdata the expression disclosure control by data reduction is used. The second is to change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased, but the information content is retained as much as possible. These are called **perturbation based disclosure control methods**.

Disclosure from analytical outputs: The use of output to make attributions about individual population units. This situation might arise to users that can interrogate data but do not have direct access to them such as in a **remote data laboratory**. One particular concern is the publication of residuals.

Disclosure limitation methods: Synonym of **disclosure control methods**.

Disclosure risk: A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.

Disclosure scenarios: Depending on the intention of the intruder, his or her type of a priori knowledge and the microdata available, three different types of disclosure or disclosure scenarios are possible for microdata: **disclosure by matching, disclosure by response knowledge** and **disclosure by spontaneous recognition**.

Dissemination: Supply of data in any form whatever: publications, access to databases, microfiches, telephone communications, etc.

Disturbing the data: This process involves changing the data in some systematic fashion, with the result that the figures are insufficiently precise to disclose information about individual cases.

Dominance rule: Synonym of **(n,k) rule**.

E

Exact disclosure: Exact disclosure occurs if a user is able to determine the exact attribute for an individual entity from released information.

F

Formal identifier: Any variable or set of variables which is structurally unique for every population unit, for example a population registration number. If the formal identifier is known to the intruder, identification of a target individual is directly possible for him or her, without the necessity to have additional knowledge before studying the microdata. Some combinations of variables such as name and address are pragmatic formal identifiers, where non-unique instances are empirically possible, but with negligible probability.

G

Global recoding: Problems of confidentiality can be tackled by changing the structure of data. Thus, rows or columns in tables can be combined into larger class intervals or new groupings of characteristics. This may be a simpler solution than the suppression of individual items, but it tends to reduce the descriptive and analytical value of the table. This protection technique may also be used to protect microdata.

H

HITAS: A heuristic approach to **cell suppression** in hierarchical tables.

I

Identification: Identification is the association of a particular record within a set of data with a particular population unit.

Identification dataset: A dataset that contains formal identifiers.

Identification data: Those personal data that allow direct identification of the data subject, and which are needed for the collection, checking and matching of the data, but are not subsequently used for drawing up statistical results.

Identification key: Synonym of **key**.

Identification risk: This risk is defined as the probability that an intruder identifies at least one respondent in the disseminated microdata. This identification may lead to the disclosure of (sensitive) information about the respondent. The risk of identification depends on the number and nature of **quasi-identifiers** in the microdata and in the a priori knowledge of the intruder.

Identifying variable: A variable that either is a formal identifier or forms part of a formal identifier.

Indirect identification: Inferring the identity of a population unit within a microdata release other than from **direct identification**.

Inferential disclosure: Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject. In general, NSIs are not concerned with inferential disclosure for two reasons. First, a major purpose of statistical data is to enable users to infer and understand relationships between variables. If NSIs equated disclosure with inference, no data could be released. Second, inferences are designed to predict aggregate behaviour, not individual attributes, and thus often poor predictors of individual data values.

Informed consent: Basic ethical tenet of scientific research on human populations. Sociologists do not involve a human being as a subject in research without the informed consent of the subject or the subject's legally authorized representative, except as otherwise specified. Informed consent refers to a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including awareness of any risks involved, of uses and users of the data, and of alternatives to providing the data.

Intruder: A data user who attempts to link a respondent to a microdata record or make attributions about particular population units from aggregate data. Intruders may be motivated by a wish to discredit or otherwise harm the NSI, the survey or the government in general, to gain notoriety or publicity, or to gain profitable knowledge about particular respondents.

J

K

Key: A set of **key variables**.

Key variable: A variable in common between two datasets, which may therefore be used for linking records between them. A key variable can either be a **formal identifier** or a **quasi-identifier**.

L

Licensing agreement: A permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information. These penalties can vary from withdrawal of the license and denial of access to additional data sets to the forfeiting of a deposit paid prior to the release of a **microdata** file. A licensing agreement is almost always combined with the signing of a contract. This contract includes a number of requirements: specification of the intended use of the data; instruction not to release the **microdata** file to another recipient; prior review and approval by the releasing agency for all user outputs to be published or disseminated; terms and location of access and enforceable penalties.

Local recoding: A disclosure control technique for microdata where two (or more) different versions of a variable are used dependent on some other variable. The different versions will have different levels of coding. This will depend on the distribution of the first variable conditional on the second. A typical example occurs where the distribution of a variable is heavily skewed in some geographical areas. In the areas where the distribution is skewed minor categories may be combined to produce a courser variable.

Local suppression: Protection technique that diminishes the risk of recognition of information about individuals or enterprises by suppressing individual scores on **identifying variables**.

Lower bound: The lowest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

M

Macrodata: Synonym of **tabular data**.

Microaggregation: Records are grouped based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates for those variables. The aggregates are released instead of the individual record values.

Microdata: A microdata set consists of a set of records containing information on individual respondents or on economic entities.

Minimal unique: A combination of variable values that are unique in the **microdata** set at hand and contain no proper subset with this property (so it is a minimal set with the uniqueness property).

N

NSI(s): Abbreviation for National Statistical Institute(s).

(n,k) rule: A cell is regarded as confidential, if the n largest units contribute more than k % to the cell total, e.g. n=2 and k=85 means that a cell is defined as risky if the two largest units contribute more than 85 % to the cell total. The n and k are given by the statistical authority. In some **NSIs** the values of n and k are confidential.

O

On-site facility: A facility that has been established on the premises of several NSIs. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a 'safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.

Ordinary rounding: Synonym of **conventional rounding**.

Oversuppression: A situation that may occur during the application of the technique of cell suppression. This denotes the fact that more information has been suppressed than strictly necessary to maintain confidentiality.

P

Partial disclosure: Synonym of **approximate disclosure**.

Passive confidentiality: For foreign trade statistics, EU countries generally apply the principle of "passive confidentiality", that is they take appropriate measures only at the request of importers or exporters who feel that their interests would be harmed by the dissemination of data.

Personal data: Any information relating to an identified or identifiable natural person ('data subject'). An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous.

Perturbation based disclosure control methods: Techniques for the release of data that change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased but the information content is retained as far as possible. Perturbation based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. For example, an error can be inserted in the cell values after a table is created, which means that the error is introduced to the output of the data and will therefore be referred to as output perturbation. The error can also be

inserted in the original data on the **microdata** level, which is the input of the tables one wants to create; the method will then be referred to as data perturbation - input perturbation being the better but uncommonly used expression. Possible perturbation methods are:

- rounding;
- perturbation, for example, by the addition of random noise or by the **Post Randomisation Method**;
- disclosure control methods for microdata applied to tabular data.

Population unique: A record within a dataset which is unique within the population on a given **key**.

P-percent rule: A **(p,q) rule** where q is 100 %, meaning that from general knowledge any respondent can estimate the contribution of another respondent to within 100 % (i.e., knows the value to be nonnegative and less than a certain value which can be up to twice the actual value).

(p,q) rule: It is assumed that out of publicly available information the contribution of one individual to the cell total can be estimated to within q per cent (q=error before publication); after the publication of the statistic the value can be estimated to within p percent (p=error after publication). In the (p,q) rule the ratio p/q represents the information gain through publication. If the information gain is unacceptable the cell is declared as confidential. The parameter values p and q are determined by the statistical authority and thus define the acceptable level of information gain. In some **NSIs** the values of p and q are confidential.

Post Randomisation Method (PRAM): Protection method for microdata in which the scores of a categorial variable are changed with certain probabilities into other scores. It is thus intentional misclassification with known misclassification probabilities.

Primary confidentiality: It concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are:

- too few units in a cell;
- dominance of one or two units in a cell.

The limits of what constitutes "too few" or "dominance" vary between statistical domains.

Primary protection: Protection using disclosure control methods for all cells containing small counts or cases of dominance.

Primary suppression: This technique can be characterized as withholding all disclosive cells from publication, which means that their value is not shown in the table, but replaced by a symbol such as '×' to indicate the suppression. According to the definition of disclosive cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or representing cases of dominance have to be primary suppressed.

Prior-posterior rule: Synonym of the **(p,q) rule**.

Privacy: Privacy is a concept that applies to data subjects while confidentiality applies to data. The concept is defined as follows: "It is the status accorded to data which has been agreed upon between the person or organisation furnishing the data and the organisation receiving it and which describes the degree of protection which will be provided." There is a definite relationship between confidentiality and privacy. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used. Informational privacy encompasses an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

Probability based disclosures (approximate or exact): Sometimes although a fact is not disclosed with certainty, the published data can be used to make a statement that has a high probability of being correct.

Q

Quasi-identifier: Variable values or combinations of variable values within a dataset

that are not structural uniques but might be empirically unique and therefore in principle uniquely identify a population unit.

R

Randomized response: Randomized response is a technique used to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered.

Random perturbation: This is a disclosure control method according to which a noise, in the form of a random value is added to the true value or, in the case of categorical variables, where another value is randomly substituted for the true value.

Random rounding: In order to reduce the amount of data loss that occurs with suppression, alternative methods have been investigated to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. In random rounding cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down. The rounding mechanism can be set up to produce unbiased rounded results.

Rank swapping: Rank swapping provides a way of using continuous variables to define pairs of records for swapping. Instead of insisting that variables match (agree exactly), they are defined to be close based on their proximity to each other on a list sorted on the continuous variable. Records which are close in rank on the sorted variable are designated as pairs for swapping. Frequently in rank swapping the variable used in the sort is the one that will be swapped.

Record linkage process: Process attempting to classify pairs of matches in a product space A×B from two files A and B into M, the set of true links, and U, the set of non-true links.

Record swapping: A special case of **data swapping,** where the geographical codes of records are swapped.

Remote access: On-line access to protected microdata.

Remote data laboratory: A virtual environment providing remote execution facilities.

Remote execution: Submitting scripts on-line for execution on disclosive microdata stored within an institute's protected network. If the results are regarded as **safe data**, they are sent to the submitter of the script. Otherwise, the submitter is informed that the request cannot be acquiesced. Remote execution may either work through submitting scripts for a particular statistical package such as SAS, SPSS or STATA which runs on the remote server or via a tailor made client system which sits on the user's desk top.

Residual disclosure: Disclosure that occurs by combining released information with previously released or publicly available information. For example, tables for nonoverlapping areas can be subtracted from a larger region, leaving confidential residual information for small areas.

Restricted access: Imposing conditions on access to the **microdata**. Users can either have access to the whole range of raw protected data and process individually the information they are interested in - which is the ideal situation for them - or their access to the protected data is restricted and they can only have a certain number of outputs (e.g. tables) or maybe only outputs of a certain structure. Restricted access is sometimes necessary to ensure that linkage between tables cannot happen.

Restricted data: Synonym of **safe data**.

Restriction based disclosure control method: Method for the release of **tabular data**, which consists in reducing access to the data provided to the external user. This method reduces the content of information provided to the user of the **tabular data**. This is implemented by not publishing all the figures derived from the collected data or by not

publishing the information in as detailed a form as would be possible.

Risky cells: The cells of a table which are non-publishable due to the risk of statistical disclosure are referred to as risky cells. By definition there are three types of risky cells: small counts, dominance and complementary suppression cells.

Risky data: Data are considered to be disclosive when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit.

Rounding: Rounding belongs to the group of disclosure control methods based on output-perturbation. It is used to protect small counts in **tabular data** against disclosure. The basic idea behind this disclosure control method is to round each count up or down either deterministically or probabilistically to the nearest integer multiple of a rounding base. The additive nature of the table is generally destroyed by this process. Rounding can also serve as a **recoding** method for microdata.

R-U map: A graphical representation of the trade off between disclosure risk and data utility.

S

Safe data: Microdata or **macrodata** that have been protected by suitable **Statistical Disclosure Control** methods.

Safe setting: An environment such as a **microdata** lab whereby access to a disclosive dataset can be controlled.

Safety interval: The minimal **calculated interval** that is required for the value of a cell that does not satisfy the primary suppression rule.

Sample unique: A record within a dataset which is unique within that dataset on a given **key**.

Sampling: In the context of disclosure control, this refers to releasing only a proportion of the original data records on a **microdata** file.

Sampling fraction: The proportion of the population contained within a data release. With simple random sampling, the sample fraction represents the proportion of population units that are selected in the sample. With more complex sampling methods, this is usually the ratio of the number of units in the sample to the number of units in the population from which the sample is selected.

Scenario analysis: A set of pseudo-criminological methods for analysing and classifying the plausible risk channels for a data intrusion. The methods are based around first delineating the means, motives and opportunity that an intruder may have for conducting the attack. The output of such an analysis is a specification of a set of **keys** likely to be held by **data intruders**.

Secondary data intrusion: After an attempt to match between identification and target datasets an intruder may discriminate between non-unique matches by further direct investigations using additional variables.

Secondary disclosure risk: It concerns data which is not primary disclosive, but whose dissemination, when combined with other data permits the identification of a microdata unit or the disclosure of a unit's attribute.

Secondary suppression: To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called secondary suppression or complementary suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the disclosive cells at the highest level of information contained in the released statistics.

Security: An efficient disclosure control method provides protection against exact disclosure or unwanted narrow estimation of the attributes of an individual entity, in other words, a useful technique prevents exact or partial disclosure. The security level is accordingly high. In the

case of disclosure control methods for the release of **microdata** this protection is ensured if the identification of a respondent is not possible, because the identification is the prerequisite for disclosure.

Sensitive cell: Cell for which knowledge of the value would permit an unduly accurate estimate of the contribution of an individual respondent. Sensitive cells are identified by the application of a dominance rule such as the (n,k) rule or the (p,q) rule to their microdata.

Sensitive variables: Variables contained in a data record apart from the key variables, that belong to the private domain of respondents who would not like them to be disclosed. There is no exact definition given for what a 'sensitive variable' is and therefore, the division into key and sensitive variables is somehow arbitrary. Some data are clearly sensitive such as the possession of a criminal record, one's medical condition or credit record, but there are other cases where the distinction depends on the circumstances, e.g. the income of a person might be regarded as a sensitive variable in some countries and as quasi-identifier in others, or in some societies the religion of an individual might count as a key and a sensitive variable at the same time. All variables that contain one or more sensitive categories, i.e. categories that contain sensitive information about an individual or enterprise, are called sensitive variables.

Shuttle algorithm: A method for finding lower and upper cell bounds by iterating through dependencies between cell counts. There exist many dependencies between individual counts and aggregations of counts in contingency tables. Where not all individual counts are known, but some aggregated counts are known, the dependencies can be used to make inferences about the missing counts. The Shuttle algorithm constructs a specific subset of the many possible dependencies and recursively iterates through them in order to find bounds on missing counts. As many dependencies will involve unknown counts, the dependencies need to be expressed in terms of inequalities involving lower and upper bounds, rather than simple equalities. The algorithm ends when a complete iteration fails to tighten the bounds on any cell counts.

Special uniques analysis: A method of analysing the per-record risk of **microdata**.

Statistical confidentiality: The protection of data that relate to single statistical units and are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of unlawful disclosure.

Statistical Data Protection (**SDP):** Statistical Data Protection is a more general concept which takes into account all steps of production. SDP is multidisciplinary and draws on computer science (data security), statistics and operations research.

Statistical disclosure: Statistical disclosure is said to take place if the dissemination of a statistic enables the external user of the data to obtain a better estimate for a confidential piece of information than would be possible without it.

Statistical Disclosure Control (SDC): Statistical Disclosure Control techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released.

Statistical Disclosure Limitation (**SDL):** Synonym of **Statistical Disclosure Control**.

Subadditivity: One of the properties of the (n,k) rule or (p,q) rule that assists in the search for complementary cells. The property means that the sensitivity of a union of disjoint cells cannot be greater than the sum of the cells' individual sensitivities (triangle inequality). Subadditivity is an important property because it means that aggregates of cells that are not sensitive are not sensitive either and do not need to be tested.

Subtraction: The principle whereby an intruder may attack a table of population counts by removing known individuals from the table. If this leads to the presence of certain zeroes in the table then that table is vulnerable to **attribute disclosure**.

Suppression: One of the most commonly used ways of protecting sensitive cells in a table is via suppression. It is obvious that in a row or column with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by **subtraction** from the marginal total. For this reason, certain other cells must also be suppressed. These are referred to as **secondary suppressions**. While it is possible to select cells for secondary suppression manually, it is difficult to guarantee that the result provides adequate protection.

SUDA: A software system for conducting analyses on population uniques and special sample uniques. The **special uniques analysis** method implemented in SUDA for measuring and assessing disclosure risk is based on resampling methods and used by the ONS.

Swapping (or switching): Swapping (or switching) involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all or some of the other variables between the matched records. Swapping (or switching) was illustrated as part of the confidentiality edit for tables of frequency data.

Synthetic data: An approach to confidentiality where instead of disseminating real data, synthetic data that have been generated from one or more population models are released.

Synthetic substitution: See **Controlled Tabular Adjustment**.

T

Table server: A form of **remote data laboratory** designed to release safe tables.

Tables of frequency (count) data: These tables present the number of units of analysis in a cell. When data are from a sample, the cells may contain weighted counts, where weights are used to bring sample results to the population levels. Frequencies may also be represented as percentages.

Tables of magnitude data: Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. When data are from a sample, the cells may contain weighted aggregates, where quantities are multiplied by units' weights to bring sample results up to population levels. The data may be presented as averages by dividing the aggregates by the number of units in their cells.

Tabular data: Aggregate information on entities presented in tables.

Target dataset: An anonymised dataset in which an intruder attempts to identify particular population units.

Threshold rule: Usually, with the threshold rule, a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number. Some agencies require at least five respondents in a cell, others require three. When thresholds are not respected, an agency may restructure tables and combine categories or use cell suppression, rounding or the confidentiality edit, or provide other additional protection in order to satisfy the rule.

Top and bottom coding: It consists in setting top-codes or bottom-codes on quantitative variables. A top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is replaced by the upper limit or is not published on the **microdata** file at all. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be used for different quantitative variables, or for different subpopulations.

U

Union unique A sample unique that is also population unique. The proportion of sample uniques that are union uniques is one measure of file level disclosure risk.

Uniqueness: The term is used to characterise the situation where an individual can be distinguished from all other members in a population or sample in terms of information

available on **microdata** records (or within a given **key**). The existence of uniqueness is determined by the size of the population or sample and the degree to which it is segmented by geographic information and the number and detail of characteristics provided for each unit in the dataset (or within the key).

Upper bound: The highest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

V

Virtual safe setting: Synonym of **remote data laboratory**.

W

Waiver approach: Instead of suppressing tabular data, some agencies ask respondents for permission to publish cells even though doing so may cause these respondents' sensitive information to be estimated accurately. This is referred to as the waiver approach. Waivers are signed records of the respondents' granting permission to publish such cells. This method is most useful with small surveys or sets of tables involving only a few cases of dominance, where only a few waivers are needed. Of course, respondents must believe that their data are not particularly sensitive before they will sign waivers.

X

Y

Z