Journal of Empirical Research in Accounting & Auditing
ISSN (2384-4787)
J. Emp. Res. Acc. Aud. 1, No. 1 (Oct-2014)

Statistical Disclosure Risk: An Overview

Elsayed A. H. Elamir

Management & Marketing Department, College of Business, University of Bahrain, Kingdom of Bahrain and Department of Statistics and Mathematics, Benha University, Egypt

Received: 27 Nov. 2013, Revised: 8 Mar. 2014, Accepted: 10 Apr. 2014, Published (Oct) 2014

Abstract: National statistical institutes (NIS) publish a wide range of trusted, high quality statistical outputs. To achieve their objective of supplying society with rich statistical information these outputs are as detailed as possible. However, these objectives conflict with the obligation NSI that have to protect the confidentiality of the information provided by the respondents. Statistical disclosure control or statistical disclosure limitation seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities. In this paper we give an overview for the most important concepts of the statistical disclosure control tools.

Keywords: Confidentiality, Models, Risk measure, Rounding, Swapping.

1. Introduction

A disclosure occurs when a person or organization recognizes or learns something that they did know already about another person or organization, via released data. There are two types of disclosure risk; identity disclosure and attribute disclosure. Identity disclosure occurs with the association of a respondent's identity with a disseminated data record containing confidential information while attribute disclosure occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent; see, Willenborg and De waal (2000). Statistical disclosure control (SDC) techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organizations. SDC methods minimize the risk of disclosure to an acceptable level while releasing as much information as possible; see, Willenborg and De waal (2000).

There are two types of SDC methods; perturbative and non-perturbative methods. Perturbative methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. Non-perturbative methods reduce the amount of information released by suppression or aggregation of data. A wide range of different SDC methods are available for different types of outputs; see, An and Little (2007), Fienberg and Makov (2001) and Hundepool et.al (2010).

2. An approach to statistical disclosure control

A general framework for addressing the question of confidentiality protection for different statistical outputs is proposed based on the following; see, Rinott(2004, 2005) and Takemura(1999).

2.1 The need for confidentiality protection

There are three main reasons why confidentiality protection is needed for statistical outputs

- It is a fundamental principle for Official Statistics that the statistical records of individual person, businesses or events used to produce Official Statistics are strictly confidential, and are to be used only for statistical purposes.
- There may be legislation that protects individual business and personal data.

E-mail: qm.group70@gmail.com

• It is essential that the confidence and trust is maintained and that identifiable information is held securely, only used for statistical purposes and not revealed in published outputs.

2.2. The key characteristics and uses of the data

When considering confidentiality protection of a statistical output it is important to understand the key characteristics of the data since all of these factors influence both disclosure risks and appropriate disclosure control methods. This includes knowing the type of the data, e.g. full population or sample survey; sample design, an assessment of quality e.g. the level of non-response and coverage of the data; variables and whether they are categorical or continuous; type of outputs, e.g. micro-data, magnitude or frequency tables.

2.3. Disclosure risks protect against what

Disclosure risk assessment then combines the understanding gained above with a method to identify situations where there is a likelihood of disclosure. Risk is a function of likelihood (related to the design of the output), and impact of disclosure (related to the nature of the underlying data). In order to be explicit about the disclosure risks to be managed one consider a range of potentially disclosive situations or scenario and take action to prevent them. A disclosure scenario describes:

1. Which information is potentially available to an intruder and

2. How the intruder would use the information to identify an individual,

A range of intruder scenarios should be determined for different outputs to provide an explicit statement of what the disclosure risk are and what elements of the output pose an un-acceptable risk of disclosure; see, Hundepool et al. (2010).

2.4. Disclosure control methods

Once an assessment of risk has been undertaken an organization must then take steps to manage any identified risks. The risk within the data is not entirely eliminated but is reduced to an acceptable level, this can be achieved either through the application of statistical disclosure control methods or through the controlled use of outputs, or through a combination of both.

2.5. Implementation

The final stage in this approach to a disclosure control problem is implementation of the methods and dissemination of the statistics. This will include identification of the software to be used along with any options and parameters. Nowadays, there are mu-Argus and tau-Argus software. The most important consideration is maintaining confidentiality but these decisions will also accommodate the need for clear, consistent and practical solutions that can be implemented within reasonable time and using available resources; see, for example, Hundepool (2010).

3. Type of data in SDC

There are two ways in which data can be presented and each of these must be treated differently for the purposes of disclosure risk and control – micro-data and tabular data; see, Hundepool et al (2010), Forster and Webb (2007) and Fienberg and Slavkovic (2004).

3.1. Micro-data

Name	Class	Grade
Ibrahim	1	A
Gamal	1	C
Sana	1	A
Yehia	2	B
Nor	3	B

Micro-data are data held as individual records, such as a person or a data zone. In the past, micro-data was primarily used to construct aggregated tables which were subsequently released. Nowadays, the micro-data sets themselves should, wherever practically possible, be released even though there may be a higher risk of disclosure which consequently must be controlled. This disclosure risk is higher if the data includes an entire population, or all the units within an identifiable sub-population (for example all children).

A primary step to prevent disclosure is to provide micro-data only if all formal identifiers have been removed - this will protect against direct identification of statistical units included in the dataset. Formal identifiers are variables which are structurally unique for every unit in the population (e.g. identification codes such as National Insurance numbers). Other variables such as name, address and date of birth are described as pragmatic formal identifiers; as the probability of uniqueness is lower than for other sensitive variables, though they still usually represent an unacceptable disclosure risk. The greater the number of pragmatic formal identifiers that can be combined, the higher the disclosure risk. Removing these identifiers provides initial protection, then additional statistical disclosure control methods can subsequently be applied; see, Duncan et al. (2001) and Dobra and Fienberg (2000).

3.2. Tabular Data

	Grade			
Class	**A**	**B**	**C**	**Total**
1	3	0	3	6
2	0	1	3	4
Total	3	1	6	10

Tabular data is information which has been aggregated from micro-data sources, such as the Census, surveys and administrative data. This form of data may not appear to raise disclosure issues as the table does not contain individual level data and it is more difficult to identify specific individuals. There are however potential disclosure risks, including:

• Cells of **0** - particularly if they dominate a row or column. If all the respondents from a row or column are in one cell, then it is possible to attribute the characteristics of the cell to every person within that row or column; this is known as group disclosure.

• Cells of **1 and 2** - for sensitive variables, sometimes even larger numbers can be disclosed. This increases the likelihood of someone identifying them or individual and uncovering new information about them.

• Cells with **dominant** contributors to the cell total – this is especially relevant for financial information like turnover, profit etc. If, for example, a table shows the aggregate turnover for various industrial sectors, and a small number of companies account for the majority of one particular cell, this would be deemed unsafe as these large companies would be able to determine, with reasonable precision, the turnovers of the remaining, smaller companies.
• Cells which contain **population unique** - when there is only one person in an entire population who exhibits a certain characteristic, this individual is described as a population unique. For example, if there was only one elderly woman aged over 100 in a local authority then this woman would be a population unique and there would be a very high risk that information about her could be disclosed unless appropriate preventive measures were taken.

• **Linked tables** – these are tables which have been produced from the same micro-data and have at least one row/column heading in common. An apparently safe table may, when compared with a linked table, lead to the disclosure of confidential information about an individual or group.

Exceptions include:

• When the data does not refer to an individual, business or organization, for example animals or public buildings.

• When permission to release confidential information has been given by the individual, business or organization.

4. Types of Disclosure Risk

There are 3 different types of disclosure risk; see, Benedetti and Fraconi (1998), Bethlehem et al. (1990). These are:

• Identity Disclosure – if an intruder can identify a data subject (either themselves or someone else).

• Attribute Disclosure – where confidential information about a data subject or group is revealed and can be attributed to the subject or each person in the group.

• Residual Disclosure– where an intruder can combine outputs from the same source

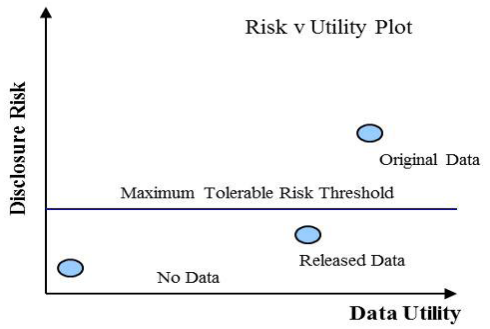
or different sources/databases to reveal information about an individual or group. Residual disclosure is becoming more of an issue. Consideration should be given to other data sources, previously released data, time series, overlapping geographies and possible interactions with other published data.

There is a high demand for data at small-area geographies. The release of such information increases the risk of disclosure because, in a small area, certain people/businesses/households are more easily identifiable than for larger geographical areas. However, even at higher geographies, statistical disclosure risk should be considered, especially for sensitive variables or minority/vulnerable groups.

5. Risk v Utility

One of the main issues involved in distributing statistics is striving to find an appropriate balance between protecting peoples' confidentiality and disseminating valid information. The value of a data set is known as 'data utility'. The 'Risk v Utility' plot shown below highlights the trade-off which underpins the entire process of statistical disclosure control – the challenge is to balance this trade-off effectively. Data utility is shown on the horizontal-axis and the risk of disclosure is shown on the vertical axis.

The original data is high in utility but also has some disclosure risk associated with it. Obviously if no data is published then it presents no disclosure risk but it is also useless to any external user and therefore has no utility. The desirable balance is to maintain maximum utility of data (as far to the right as possible in the risk/utility map) but reduce the risk below the maximum tolerable risk threshold. This threshold is not fixed but can change over time and be affected by public perception, intruder behavior or availability of other published data and should be reviewed regularly. The threshold is set by the data owner, who is best placed to take account of the sensitivities of the data and current issues/circumstances which will affect the tolerable level of risk.



To reduce the disclosure risk sufficiently it may be necessary to utilize several different methods on the same data. The disclosure control method(s) should take accounts the user and their needs. Consistency in disclosure control methods for data users is important to enable them to monitor real differences over time and not differences due to the use of different methods; see, Hundepool (2010), Aggarwal and yu (2008), Di Consiglio and Polettini (2006) and Feinberg and Makov (1998).

6. Methods of statistical disclosure control

For data which is going to be placed in the public domain there are two distinct groups of disclosure control methods: pre-tabular and post-tabular methods; see, Franconi and polettini (2004), Polettini (2004), Rinott (2004), Skinner and Elliot (2002) and Purcell and Kish (1980).

6.1 Pre-tabular

Pre-tabular methods are used on micro-data and can provide protection for micro-data outputs and tabular outputs (protecting the micro-data means that all tables subsequently produced from it are automatically safe). The methods within this family include record swapping, global recoding, PRAM, blurring and removal of sensitive variables.

6.1.1 Record swapping

Record swapping involves interchanging certain values of a sensitive variable between records. This involves swapping an individual

record with another record which is similar on most characteristics (for example age, gender or geographical area) but differs in terms of the sensitive variable (such as income or health data). This introduces some doubt about the identity of a person when the information is later displayed in a table.

6.1.2. Global recoding

Global recoding involves combining several categories of a variable into a single one. For example, imagine a dataset contained the following unique record:

Marital Status	Age
Widow	17

Global recoding could be applied to the variable 'Marital Status' by combining the categories 'divorced' and 'widow' to create the broader category of 'divorced/widow'. This would reduce the likelihood of the above record being unique in the dataset. Note, that global recoding must be applied to the entire dataset and not just to the part that is unsafe.

Top and bottom recoding is a special case of global recoding whereby top and bottom codes are set. Any value greater than the top code is set to the top code and any value less than the bottom code is set to the bottom code. For example, if the top code is set to 80+ then all people aged 80 and over are assigned to this category. Generally, for variables with normal distributions, smaller values are more likely to occur at the ends of the distribution, so top and bottom coding maintains the usefulness of the midrange values whilst protecting the small number of values at the ends of the distribution.

6.1.3. Post randomization method (PRAM)

This method applies protection to categorical variables in micro-data sets by randomly changing the category assigned to certain records. This is done according to a probability function and introduces ambiguity as to whether the category has changed or not. For example, imagine the categorical variable was Gender and the respondent is male. We may say that, according to the probability function, there is an 80% chance that the respondent's gender will remain male in the dataset and a 20% chance that the respondent will be reclassified as female. This introduces doubt about the true gender of the respondent for any individual record and makes it difficult to establish a link between a respondent in the survey and a real person in the population; see Ardo and Elamir (2006), Zhou et al (2007) and Zhang and Chambers (2004).

6.1.4. Blurring

This method replaces a reported value(s) for a certain variable with an average or median of that variable. The replacement value may be assigned to some or all records or may be assigned to those which have the same values of another variable.

6.2. Post tabular

Tabular outputs can be protected either before or after tabulation. Post-tabular methods require the desired table to be produced and any unsafe cells to be identified. The table can then be protected using one of the following methods: table redesign, cell suppression, and rounding; see, Buzzigoli and Giusti (1999) and Chen and Keller-McNulty (1998).

6.2.1. Table redesign

Table redesign should always be considered first when protecting tabular data as it is often the simplest method for the data owner to implement and for the user to understand. It involves combining rows and/or columns to increase the number of respondents in small cells. It is advisable to think about this method first because it doesn't actually change the data in any way – it simply minimises the level of detail that is published. Often, this straightforward technique removes low-frequency cells and produces a safe table without having to change the data or spend further time protecting it. However, if there are a large number of low-frequency cells then table redesign may fail to provide sufficient protection and other methods may have to be used alongside, or instead of, it. Groupings for variables should be standard to prevent disclosure by differencing from several tables in the public domain containing the same data. This is particularly important

for geographical variables where, unless there is good reason to do so.

6.2.2. Cell suppression

Suppression involves minimizing the level of detail that is released by 'hiding' the values of low-frequency cells or cells dominated by a small number of large contributors. High risk cells are not published and are called primary suppressions. All suppressed cells should be replaced with an asterisk (*) to show that they have been suppressed and not left blank.

Secondary suppressions are also required which means at least one other value in the row and/or column is also suppressed to ensure that disclosive cells cannot be deduced through subtraction. All other cells in the table are published. Secondary suppressions should be chosen in a way that tries to minimise the cost of disclosure control whilst also maximizing data utility. There should be no indication of whether the suppression is primary or secondary.

6.2.3. Rounding

There are various ways to apply rounding and each has its own advantages and disadvantages. The three methods are deterministic, random and controlled rounding.

6.2.3.1. Deterministic rounding

This method involves rounding the cell value to the nearest multiple of a rounding base, b. For example, if b = 5 then a 4 » 5, 6 » 5, 8 » 10. Hence a value of 5 in a rounded table can represent anything from 3 to 7. Each of the internal cells in the table are rounded first. The external cells (the totals) are then calculated from the unrounded components (internal cells) and rounded to the nearest multiple of 5 using the procedure described above. Because the internal and external cells are rounded separately, the rounded row and column totals may not correspond to the sum of the rounded values for their components. This is known as uncontrolled rounding and means that table additively is not preserved.

The protection level provided by deterministic rounding is b -1 because the original value lies in an interval of width (b - 1). For example, for the rounding base b=5: if the rounded value is 10, the original value lies in the interval [8;12] which is of width 4. This is the most basic form of rounding and can be unpicked/solved relatively easily therefore it is not recommended.

6.2.3.2. Random rounding

In this method each cell value is rounded up or down following a given probability. For example, if b = 5, an 8 is rounded to 10 with probability 0.8 (80% of the time) and 8 is rounded to 5 with probability 0.2 (20% of the time).

Again this is uncontrolled rounding because internal rounded cells do not necessarily add up to rounded row and column totals. It does give a little more protection than deterministic rounding however. The protection level is 2(b - 1) because the original value lies in an interval which is of width 2(b - 1). For example, b = 5: if the rounded value is 10, the original value lies in the interval [6;14] which is of width 2x(5-1)=8.

6.2.3.3. Controlled rounding

Users frequently prefer this method because internal rounded cells do add up to rounded totals (additively) and it affords the same protection level as random rounding, 2(b-1). For the basic form of controlled rounding, a cell value is rounded either up or down to the nearest multiple of the rounding base. Values of a table that are already multiples of the rounding base remain unchanged.

The nearest solution to the original data is sought but cannot always be found so the method may be extended so that the cell value is allowed to be rounded up or down to the next multiple of the rounding base. Software is available to carry out controlled rounding, as it requires a linear programming solver.

7. Cost of disclosure control

The cost of disclosure control is a way of measuring the damage inflicted on a table by protecting it, i.e. the information loss that has occurred. For example, when a table is rounded, the cost is the sum of the difference between every

cell in the protected table and its corresponding cell in the unprotected table. High costs indicate that the table has suffered a considerable amount of damage and has therefore lost lots of information (and hence utility).

There are two main ways of looking at cost when suppression is used to protect data: frequency and unity. The 'unity' approach minimizes the number of cells that are secondary suppressions, regardless of the number of units within those cells. Therefore if 20 cells in the table have been secondary suppressed then the cost of the disclosure control is 20. Within this approach cells with values of 1000 and of 1 are treated equally.

The 'frequency' approach however, minimizes the number of units contributing to the secondary suppressed cells. In this approach, a cell which contains 1000 contributors is valued much more highly than a cell with only 1 contributor – as the latter usually contains less valuable information. While adopting the frequency approach may result in an increased number of suppressed cells when compared with the unity approach, the summed values of the suppressed cells will be lower. This is extremely useful if you wish to retain high value cells.

8. Disclosure risk

Moreover in case of multiple release of the same survey coherence should be maintained also between different released files in the sense that releasing different files at the same time shouldn't allow the gaining of more information than for one file alone; see, Trottini et al. (2006). The principles apply also to the release of longitudinal or panel micro-data, where the differences between records pertaining to the same case in different waves will reflect events that have occurred to that case, as well as the attributes of the individuals.

Once the characteristics and uses of the survey data are clear, it is time to start the real analysis of the disclosure risk. This implies first a definition of possible situations at risk (disclosure scenarios) and second a proper definition of the risk in order to quantify the phenomenon (risk assessment).

A disclosure scenario is the definition of realistic assumptions about what an intruder might know about respondents and what information would be available to him to match against the micro-data to be released and potentially make an identification and disclosure. Again different types of releases may require different disclosure scenarios and different definitions of risk. Once a formal definition of risk has been chosen we need to measure and estimate it. It is also important to define when a unit or a file presents an acceptable risk and when it has to be considered at risk. This threshold depends of course on the type of the measure adopted. Choice of scenarios and level of acceptable risk are extremely dependent on different cultural situations in different member states, different policies applied by different institutes, different approaches to statistical analysis, different perceived risk.

8.1. Risk assessment

Micro-data has many analytical advantages over aggregated data but also poses more serious disclosure issues because of the many variables that are disseminated in one file. For micro-data disclosure occurs when there is a possibility that an individual can be re-identified by an intruder using information contained in the file, and when on the basis of that, confidential information is obtained. Micro-data are released only after taking out directly identifying variables, such as names, addresses, and identity numbers. However, other variables in the micro-data can be used as indirect identifying variables. For individual micro-data these are variables such as variables gender, age, occupation, place of residence, country of birth, etc., and for business micro-data variables such as economic activity, number of employees. These indirect identifying variables are mainly publicly available variables or variables that are present in public databases such as registers. If the identifying variables are categorical then the compounding- cross-classification- of these variables define a key. The disclosure risk is a function of such identifying variables either in the sample alone or in both the sample and the population. To assess the disclosure risk we need to make realistic assumptions about what an intruder might know about respondents

and what information will be available to him to match against the micro-data and make an identification and disclosure. Based on the disclosure risk scenario the identifying variables are determined. The other variables in the file are confidential or sensitive variables and represent the data not to be disclosed.

8.1.1. Disclosure risk scenarios

The definition of a disclosure scenario is a first step towards the development of a strategy for producing a safe micro-data file (MF). A scenario synthetically describes:

(i) Which is the information potentially available to the intruder, and

(ii) How the intruder would use such information to identify an individual i.e. the intruder's attack means and strategy.

Often defining more than one scenario might be convenient, because different sources of information might be alternatively or simultaneously available to the intruder. Moreover, re-identification risk can be assessed keeping into account different scenarios at the same time.

8.1.2. Risk assessment in micro-data

Consider adisclosure scenario defining q categorical key variables, denoted by Z_1,... Z_q with C_1,...C_q categories respectively. This scenario is appropriate for most population surveys, where identification can be based on variables such as place of residence, sex and age; see, Elamir and Skinner (2006). Records with the same key values are identical for re-identification and should have the same risk of disclosure. Cross-classification of the key variables generates a contingency table with a total number of

$$K = \prod_k C_k$$

cells at both the population and the sample level; cell frequencies in the population and sample table, respectively, are denoted by and . Intuitively, rare traits in the population are the ones that could lead to disclosure, but to be exposed to disclosure risk such rare records should also be included in the sample. The problem is therefore discriminating between the sample cells that are structurally small in the population and those that are small in the micro-data only because of sampling error; direct or indirect inference about the corresponding population size is required for these cells.

Analogously to the approach taken for tables, upper bounds on frequencies could be constructed, but usually comprehensive measures of risk are provided (Willenborg and de Waal, 2000), that may take into account other elements of the scenario such as the utility of the disclosed information, the probability of a disclosure, and so on (Polettini, 2004).

A large part of the literature has focused on estimating measures based on the frequency of sample unique cells that are also population unique (see Chen and Keller McNulty, 1998; Fienberg and Makov, 2001; Skinner and Elliot, 2002). These quantities can be used as global risk measures for the microdata file. However it is also important to be able to assess the disclosure risk associated with the release of individual records; if the population contingency table were known, a simple risk measure for each record in cell of the sample table could be defined using the corresponding population cell size

$$F_k : r_k = \frac{1}{F_k}$$

As F_K is unknown, the above definition is not usable. A solution is to specify a statistical model for

$$F = (F_1, \ldots, F_k)$$

and derive suitable risk estimates, such as

$$E(1/F_k | f)$$

Typical applications consider very large and sparse contingency tables, often with logical constraints inducing structural zeros. Estimating F_K is particularly difficult for high risk cells, having low sample and population sizes. Finite population theory cannot account for all the information about the population structure and would produce unreliable estimates, in particular when the sampling fraction is small. This justifies the introduction of super-population models. Further improvements

would be obtained from models that allow "borrowing strength" from larger cells while avoiding excessive shrinkage of the estimates; see, Elamir and Skinner (2006), Slavkovic and Fienberg (2004), Skinner and Holmes (1998), Skinner and Shlomo (2006), Rinott, Shlomo (2007a), Rinott and Shlomo (2007b), Trottini et al. (2006) and Wasserman (2007).

8.1.3 . Models for risk estimation with survey data

Part of the literature on disclosure risk assessment is related to the work of Bethlehem et al. (1990), which represents the first approach to defining a statistical model for samples where identifying variables form a contingency table. The model is a hierarchical Poisson-Gamma super-population model:

$$\pi_k \sim gamma(\alpha, K\alpha), \qquad k = 1, \ldots, K$$
$$F_k|\pi_k \sim Poisson(N\pi_k)$$

$f_k|F_k, \pi_k, p_k \sim binomial(F_k, p_k)$, independently acrros cells in which $\prod_k$ is the probability that a unit of the population falls into cell k and p_K (assumed constant across cells) is the probability that an individual in cell is sampled. The model was used to deduce the probability of population uniqueness given sample uniqueness.

The constraints

$$\sum_k F_k = N \text{ and } \sum_k \pi_k$$

are not exactly satisfied under the model, which can be seen as an approximation to the more coherent Dirichlet-multinomial model analysed by Takemura (1998). The above described Gamma-Poisson model was shown by Rinott (2004) to have as limit when ∞ » 0 the negative binomial model analysed in Benedetti and Franconi (1998)

9. Discussion

Statistical Institutes have so far adopted a release strategy that almost invariably consists of a set of tables (counts and intensities), often also published on official web pages, and one or more files of micro-data: files for the general public (public use micro-data files, PUF), and files for research.

Before release, risk is assessed through one of the models outlined in this paper, which, however, rely on a scenario where all the external information is contained in the key variables. On the other hand, data published by statistical Institutes may provide extra information that, if not already accounted for, should be included in the scenario for the evaluation of disclosure risk. Trottini et al. (2006) note that allowing for this extra information in risk estimation would be complex and propose adopting a hierarchical release strategy that excludes, for instance, the case that tables contain information not already present in the released micro-data.

When applying disclosure limitation techniques, account should be taken of the information that could be used to undo the protection, in a way similar to the assessment of the so called secondary confidentiality of tables. In this respect, methods that release the sufficient statistics of a "large" model or even simulation methods that preserve these approximately should be preferred, but apart from the case of contingency tables, the problem of defining such a model is far from being solved, especially for sparse data comprising many interrelated variables and subject to a variety of different analyses. An additional problem to be considered is that users might want to fit to the data a whole set of models, or even different classes of models. This makes it difficult to define a general purpose release strategy, and makes the development of remote access systems more appealing for granting access to the data to a larger variety of statistical users.

Finally, as suggested by some authors, the decision on the release strategy should also consider the trade-off between disclosure risk and the benefit deriving from an expanded access to the statistical information.

References

Aggarwal C.C., and P.S. Yu (eds.). 2008, 'Privacy-Preserving Data Mining: Models and Algorithms'. NY: Springer, pp.53-80.

An D., Little R.J.A. 2007, 'Multiple imputation: an alternative to top coding for statistical disclosure control' Journal of the Royal Statistical Society, A, 170, 923–940.

Ardo, V. and Elamir A.H.E 2006, 'Statistical disclosure control using post randomization: variants and measures for disclosure risk' Journal of Official Statistics, 22, pp. 711-731.

Benedetti R., Franconi L. 1998, 'Statistical and technological solutions for controlled data dissemination', in: Pre-Proceedings of New Techniques and Technologies for Statistics – Sorrento, 4-6 November 1998, Vol. 1, 225–232.

Bethlehem J., Keller W., Pannekoek J. 1990, 'Disclosure control of microdata', Journal of the American Statistical Association, 85, 38–45.

Buzzigoli L., Giusti A. 1999, 'An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals, in: Statistical Data Protection', Lisbon, 25 – 27 March 1998, Eurostat, Luxembourg, 131–147.

Chen G., Keller-McNulty S. 1998, 'Estimation of identification disclosure risk in microdata', Journal of Official Statistics, 79–95.

Di Consiglio L., Polettini S. 2006, 'Improving individual risk estimators, in: Privacy in Statistical Databases 2006', Domingo-Ferrer J. & Franconi L. (Eds.), Springer, Berlin, 243–256.

Dobra A., Fienberg S.E. 2000, 'Bounds for cell entries in contingency tables given marginal totals and decomposable graphs', in: Proceedings of the National Academy of Sciences, 97, No. 22, 11885–11892.

Duncan G.T., Keller-McNulty S., Stokes S.L. 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.A.H and Skinner, C.J. 2006, 'Record level measures of disclosure risk for survey micro-data' Journal of Official Statistics, 22, pp. 525-539.

Fienberg S.E., Makov U.E. 1998, 'Confidentiality, uniqueness and disclosure limitation for categorical data', Journal of Official Statistics, 14, 385–397.

Fienberg S.E., Makov U.E. 2001, 'Uniqueness and disclosure risk: urn models and simulation, in: Bayesian Methods With Applications to Science, Policy and Official Statistics', Monographs in Official Statistics, Eurostat, 135–144.

Fienberg S.E., Slavkovic A.B. 2004, 'Making the release of confidential data from multiway tables count', Chance, 17, 5–10.

Forster J.J., Webb E.L. 2007, 'Bayesian disclosure risk assessment: predicting small frequencies in contingency tables', Journal of the Royal Statistical Society, C, 56, 551–570.

Franconi L., Polettini S. 2004, 'Individual risk estimation in Mu-Argus: a review, in: Privacy in Statistical Databases 2004', Domingo-Ferrer J. & Torra V. (Eds.), Springer- Verlag, 262–272.

Hunhepool, Domingo-Ferrer, Franconi, Giessing, Lenz, Naylor, Nordholt, Seri, De Wolf 2010, 'Handbook on Statistical Disclosure Control', ESSNet SDC.

Polettini S. 2004, 'Some remarks on the individual risk methodology', in: Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7 – 9 April 2003, Eurostat, Luxembourg, 299–311.

Purcell N.J., Kish L. 1980, 'Postcensal estimates for local areas (small domains)', International Statistical Review, 48, 3–18.

Reiter J.P. 2005, 'Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study', Journal of the Royal Statistical Society, A, 168, 185–205.

Rinott Y. 2004, 'On models for statistical disclosure risk estimation', in: Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7 – 9 April 2003, Eurostat, Luxembourg, 275– 285.

Rinott Y., Shlomo N. 2007a, 'Variances and confidence intervals for sample disclosure risk measures', Proceedings of the 56th Session of the ISI, Lisbon 22-29 August 2007.

Rinott Y., Shlomo N. 2007b, 'A smoothing model for sample disclosure risk estimation, in: Complex Datasets and Inverse Problems: Tomography, Networks and Beyond', IMS Lecture Notes, Institute of Mathematical Statistics, 54, 161–171.

Skinner C.J., Elliot M.J. 2002, 'A measure of disclosure risk for microdata', Journal of the Royal Statistical Society, B, 64, 855–867.

Skinner C.J., Holmes D.J. 1998, 'Estimating the re-identification risk per record in microdata', Journal of Official Statistics, 14, 361–372.

Skinner C.J., Shlomo, N. 2006, 'Assessing identification risk in survey microdata using log-linear models', S3RI Methodology Working Papers, M06/14, University of Southampton, Southampton Statistical Sciences Research Institute.

Slavkovic A.B., Fienberg S.E. 2004, 'Bounds for cell entries in two-way tables given conditional relative frequencies, in: Privacy in Statistical Databases 2004', Domingo-Ferrer J. & Torra V. (Eds.), Springer-Verlag, 30–43.

Takemura A. 1999, 'Some superpopulation models for estimating the number of population uniques, in: Statistical Data Protection', Lisbon, 25 to 27 March 1998, Eurostat, Luxembourg, 45–58.

Trottini M., Franconi L., Polettini S. 2006, Italian household expenditure survey: a proposal for data dissemination, in: Privacy in Statistical Databases 2006', Domingo-Ferrer J. & Franconi L. (Eds.), Springer, Berlin, 318–333.

Willenborg L., deWaal T.D. 2000, 'Elements of Statistical Disclosure Control', Springer-Verlag, New York.

Zhang L., Chambers R.L. 2004, 'Small area estimates for cross-classifications', Journal of the Royal Statistical Society, B, 66, 479–496.

Zhou S., Lafferty J., Wasserman L. 2007, 'Multiple imputation: an alternative to top coding for statistical disclosure control', Journal of the Royal Statistical Society, A, 170, 923–940.