Working Paper No. 23
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 ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

MASSC: A NEW DATA MASK FOR LIMITING STATISTICAL INFORMATION LOSS AND DISCLOSURE

Invited paper

Submitted by Research Triangle Institute International, United States[1]

Acknowledgments: This paper is planned for an invited session at the joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, April 7-9, 2003. RTI International owns the intellectual property rights of the MASSC (patent pending) system for statistical disclosure limitation presented in this paper. Views expressed in the paper are those of the authors and not necessarily of RTI. The authors are grateful to J. Lessler, D. Kulka, K. Boyle, and D. Camburn for their support and encouragement, and to D. Wright (SAMHSA), L. Cox (NCHS), P. Biemer, J. Chromy, and R. Folsom for useful comments.

[1] Prepared by A.C. Singh (asingh@rti.org), F. Yu, and G.H. Dunteman.

MASSC: A new data mask for limiting statistical information loss and disclosure

(Draft Version, March 31, 2003)

Abstract

We propose a method termed 'MASSC' for ensuring statistical disclosure limitation (SDL) of categorical or continuous micro data, while maintaining the analytical quality of the micro data. The new SDL methodology exploits the analogy between (1) taking a sample (instead of a census,) along with some adjustments, including imputation, for missing information, and (2) releasing a subset, instead of the original data set, along with some adjustments for records still at disclosure risk. Survey sampling reduces monetary cost in comparison to a census, but entails some loss of information. Similarly, releasing a subset reduces disclosure cost in comparison to the full database, but entails some loss of information. Thus, optimal survey sampling methods can be used for statistical disclosure limitation. The method consists of Micro Agglomeration signifying a partition of the database into risk strata, optimal probabilistic Substitution, optimal probabilistic Subsampling, and optimal sampling weight Calibration.

The proposed method uses a paradigm shift in the practice of disclosure limitation in that the original database itself is viewed as the population and the problem of disclosure by inside intruders is considered. (Inside intruders know the presence of their targets in the database in contrast to outside intruders.) This new framework has two main features: one, it focuses on the more difficult problem of protecting from inside intruders and as a result also protects against outside intruders, and second, it provides model-free measures of both information loss and disclosure risk when disclosure treatment is performed by employing known random selection mechanisms for substitution and subsampling. Empirical results will be presented to illustrate computation of measures of information loss and the associated disclosure risk for a small data set.

Key words: disclosure risk; information loss; substitution; subsampling; calibration; optimal selection

1. Introduction

The disclosure problem arises if an individual in the population can be associated with a record in the database containing sensitive values. This is an old problem started in 1960's, see e.g., the two reports by the Federal Committee on statistical methodology (1978, 1994). There are mainly two types of disclosure that are of concern to the data producer; first, Identity disclosure in which using a combination of direct identifying variables (IVs such as name, address and SSN), an individual's identity can be associated with a record, and second, attribute disclosure in which using a combination of indirect IVs, an individual's attributes can be associated with a record, which, in turn, leads to the disclosure of sensitive values. Since direct IVs (which typically have little or no analytic value) are generally suppressed globally in any database for public use, identity disclosure is not really of concern. However, it is the attribute disclosure problem that is of concern in public use files (PUFs), which has led to new HIPAA regulations (2000) and a lot of discussions in recent times for protecting personal health information.

It is of interest to note that the new HIPAA regulations provide two options for protecting personal health information so that an individual's identity can be protected beyond a reasonable doubt. The first option is based on a statistical method, and requires a statistician's opinion to judge whether there is adequate protection. The second option, relatively much easier to implement if deemed adequate, involves global recoding of 18 direct and indirect identifiers, but requires an additional condition that the data producer must not be aware of any individuals that may be at risk; this condition may be hard to satisfy in practice because it depends on the type of data intrusion test one employs. Note that global suppression can be viewed as a special case of global recoding by recoding the variable into very broad categories.

In this paper, we first review in Section 2, types of data intrusion and the corresponding types of intruders. We next consider in Section 3 the problem of measuring disclosure risk and information loss after disclosure treatment of a dataset. It is a common practice to use some forms of perturbation and/or suppression of data for disclosure treatment. Such treatment affects analytical quality of the resulting data, and leads to a tension between data quality and data confidentiality. In other words, as disclosure risk goes down, loss of information goes up. In practice, it is, therefore, desirable to have a simultaneous control on information loss and disclosure risk. It is observed that the existing methods of measuring disclosure risk involve strong modeling assumptions. Although innovative frameworks for the objective of simultaneous control on information loss and disclosure risk have been recently developed (see e.g., Zaslavsky and Horton, 1998; Duncan and Keller-McNulty, 2001), more research is needed to make them suitable for general applications. In the history of SDL research, most research efforts have been directed toward developing disclosure control methods, but very little on analyzing the treated dataset as standard analysis tools become no longer applicable, see e.g., Fuller (1998), and Little (1998).

The main purpose of this paper is to propose, in Section 4, a new framework for measuring disclosure risk and information loss without relying on modeling assumptions, and then propose, in section 5, a practical method of disclosure treatment that provides a simultaneous control on information loss and disclosure risk in a suitable sense. The framework represents a paradigm shift in that the database itself is viewed as the population so that the disclosure problems becomes that of inside intrusion as defined below. The method relies on known randomization schemes for selecting a portion of the database for perturbation via substitution, and suppression via subsampling, and can be motivated by analogy with survey sampling of a finite population. Because of known randomization schemes that govern disclosure treatment, and hence the link between the treated database and the original one, it is possible to compute suitable measures of disclosure risk and information loss respectively without modeling assumptions.

The new approach may seem at first sight overly conservative because by regarding the database as the population, it tries to protect against inside intrusion in which the intruder essentially knows his target's presence in the database (and thus has a known target), and as a result also protects against outside intrusion in which the intruder doesn't know his target's presence in the database, and attempts to identify an unusual record (corresponding to an unknown target) by matching to an external file. Although the outside intrusion is the commonly considered problem, the inside intrusion problem always exists with any database because of the threat from coalition intruders. (The terms inside and outside intrusion,

although not standard in the literature, are used here for convenience in presenting ideas.) The threat from coalition intrusion is widely discussed in the literature on statistical disclosure limitation especially in the context of tabular data, and is a more serious and difficult problem than the threat from outside intrusion. It is, therefore, believed that the proposed framework, although seemingly more stringent, is quite realistic as the threat from inside intrusion (even if it is remote) cannot be ignored in practice.

The proposed method, termed MASSC, is described in section 5. MASSC signifies an acronym for four components of the treatment process: Micro Agglomeration of records for defining risk strata, Substitution of a random subset of records using an optimal sampling design, Subsampling of a random subset of substitution-treated database again using an optimal sampling design, and finally Calibration of sampling weights of the treated database. Since MASSC is based on survey sampling techniques, an important byproduct of the proposed method is that it lends itself to use of standard analysis softwares for survey data such as SUDAAN. The MASSC process has four features: (i) it provides measures of disclosure risk and information loss in a certain sense without modeling assumptions; (ii) it provides a simultaneous control on both information loss and disclosure risk; (iii) it provides a unified approach to both tabular and micro data; and (iv) standard survey data analysis tools can be used to analyze MASSC-treated dataset. A version of MASSC has been successfully used to create PUFs for the National survey on Drug Use and Health (1999- present) conducted annually by RTI. In Section 6, a simple hypothetical example is presented to explain the computation of disclosure risk after MASSC treatment. Results from a simulation study on a small dataset are presented in Section 7 to illustrate how disclosure risk varies with information loss, and to present various diagnostics for data confidentiality and quality. Finally, concluding remarks are given in Section 8.

2. Types of data intrusion

In data disclosure problem, it is perceived that an intruder always exists and so the data should be protected against this perception. The situation is a bit similar to taking a fire insurance even though the event may be very unlikely. The reason is simple. If the event does occur, it is devastating and has serious consequences. Even if the data has restricted access, but if it is believed that there is the possibility of intrusion, then there is a need of protection. In the literature on data disclosure, as mentioned in the introduction, there are two types of intrusion which we refer to as outside and inside. The problem of coalition intrusion can be viewed as a special case of inside intrusion. In outside intrusion, presence of a target in the database is not known to the intruder because the database is regarded as a subset of the population. For example, in the case of microdata, when the intruder sees some records which have unique profiles in terms of identifying variables (IVs), then he may render them as targets and may want to use an external file to attempt to identify those individuals. Note that the amount of known IVs varies from intruder to intruder. It is the values of the sensitive variables (SVs) that the intruder wants to expose for his targets. Note that since the database unique need not be a population unique, the intruder may not be able to make a successful match even if the matching variables in the external file are free from error.

There are mainly three types of disclosure in PUF: one, reidentification of (apparently) unique records with respect to a set of IVs leading to disclosure of SVs; two, disclosure of SVs of (apparently) nonunique (i.e., a cluster of records with common values of IVs) records because of common values; and three, disclosure of (apparently) nonunique records because of coalition. Note that under outside intrusion, the risk of disclosure of nonuniques is generally considered negligible, see e.g., Skinner and Holmes (1998).

An example of outside intrusion in the case of tabular data of counts (see e.g., Willenborg and de Waal, 2001) arises if the intruder finds a cell with a single count, and then he may be tempted to try to identify that individual using an external matching file. (Here it is implied, although not stated explicitly in the literature, that the SV is also an IV and is part of table's cross-classification.) Consequently, it is best to avoid tables with unique cells, and often a rule of 3 or 5 is used to discourage the intruder from trying to predict an individual's association with the record even though he may not have a unique match because the dataset is assumed to be subset of the population.

In the case of inside intrusion, the database is regarded as census of a certain domain, and target's presence is essentially known to the intruder. If an intruder already knows the presence of a target, he doesn't need a matching file to identify his target. Even if he doesn't have a known target in mind, he can choose one after looking at the data. For example, if he sees a unique record worth targeting, he can use a matching file to know this target. Alternatively, if he sees a cluster of records with common IVs (i.e., nonuniques) worth targeting, then he can choose one of them from the matching file arbitrarily as his target. The problem of inside intrusion arises in the literature in the context of tabular data. With count tabular data, if a cell count is small (such as 3 or so), then through a coalition intrusion, one can disclose the remaining member of the cell. With magnitude count data (often in the context of economic data), along with the cell count, information about the SV is provided as an aggregate value for the cell. Here again, through a coalition intrusion, the contribution to the aggregate value from the remaining member can be disclosed. In the above scenario, coalition intruders act as inside ones. In the case of household survey data, members of the household may act as inside intruders.

As mentioned earlier, the problem of inside intrusion, although not emphasized in the literature, is more serious and should be addressed for adequate protection as it subsumes the problem of outside intrusion. This is the approach taken in this paper.

3. Problems of measuring disclosure risk and information loss

In computing disclosure risk, a conservative approach is generally taken whereby the probability of intrusion is taken as one. Clearly, it is the perceived risk that we are trying to reduce assuming that there is an intruder attack. Note that it is always the individual record that needs to be protected which, in turn, protects the database. So disclosure risk, ideally, should be record-specific. However, due to theoretical limitations, often the disclosure risk is common for a set of records sharing certain values of IVs, or it may even be common for the whole dataset. If the number of risky records is relatively small in the dataset, it is easier to protect confidentiality of risky records. Even if a single record is known to be potentially at risk, some treatment is needed.

For a given database untreated for disclosure, disclosure risks from an outside intruder for unique records may be computed using different sets of assumptions. This is an exercise in probability theory. The risk may be common for all records or for a group of records that are at risk by being uniques. A simple-minded approach is (based on fairly strong assumptions) is as follows. Cells of a table cross-classified by demographic and geographic IVs are first defined, and using census data, approximate population counts for all cells are obtained, and then the risk is computed as inverse of the cell count. This risk is the probability that the intruder would be able to pick at random an individual from the population in a given cell and associate him with the single (unique) record in that cell obtained from the database. Here, it is assumed that the records are selected at random from the population with equal probability, and that the intruder is able to identify individuals in the population that belong to a particular cell with probability one. If a cell count is small, then the risk for individuals in that cell would be high, and some treatment would be needed. In practice, often the census does not provide counts at the desired cell level, but only at the marginal level such as at the geography level. Then the assumption of uniform distribution is used to allocate counts to each cell. Clearly, assumptions used in the above approach are not quite realistic. Moreover, it doesn't provide any guidance on computing disclosure risk when the data requires some treatment.

A more general approach for computing risk from the outside intrusion mentioned above is based on relatively more realistic but still strong modeling assumptions. Instead of assuming that the population count, say k, for a given cell sharing the IVs of an individual under attack by the intruder, is known, and a model is used to compute the probability that the count is k. The disclosure risk is then defined as the average of the random variable 1/k, and is given by $\sum_k P(k)/k$. A related and probably more useful measure is defined as the conditional probability that a record is a population unique given that it is a sample unique (in other words, the probability that there is a population unique sharing the values of IVs

of a record in the database), see e.g., Skinner and Holmes (1998), and Fienberg and Makov (1998). Often assumptions of Poisson sampling from the super-population, and Bernoulli sampling from the finite population are used to compute the above conditional probability. In fact, this probability is only an upper bound as the disclosure risk is given by the product,

Pr(the sample unique is a population unique)x Pr(the sample unique is matched correctly with an individual in the population). (3.1)

In practice, however, the second probability is generally assumed to be one for convenience. If the probability (3.1) is not small or equal to 1 as in the case of inside intrusion, then the data needs some disclosure treatment, and a different approach is needed for computing risk. The reason for this is that the conditional probability is meaningful under outside intrusion because target's presence is not known in the database. In fact, here IVs of a potential target is selected after looking at the database. Moreover, it is assumed that the data didn't undergo any treatment. It follows that the above framework for measuring risk is not well defined if disclosure treatment is performed because of uncertainty in the IVs of a potential target based on a record in the treated database. However, it is meaningful to define unconditional probability of a record to be population unique with a given set of values of IVs as considered by Bethlehem et al. (1990). Thus, one can define risk from outside intrusion for a record in the treated database by a sandwich formula as follows:

Pr(potential target is a population unique and is selected in the database)x **Pr(target record survives treatment and looks unique)** x Pr(the sample unique is matched correctly with the individual in the population) (3.2)

Notice that in addition to modeling for computing the outer probabilities in the above sandwich formula, a separate modeling is needed to compute the middle part as it depends on the nature of treatment performed by the data producer. This, however, may be difficult to model because of the finiteness of the database and if the treatment is done on an ad hoc basis, i.e., not governed by a random mechanism. In the case of inside intrusion, the first and the last probabilities are one because the database is the population and the intruder knows the target and its presence, and so the disclosure risk is given by the middle probability only. It is interesting to note from the above sandwich formula that the risk from outside intrusion is at most equal to the risk from inside intrusion. A model-free approach to compute risk from inside intrusion can be used if known randomization mechanisms are used to introduce some form of perturbation and suppression for treatment. This forms the basis of the proposed method (which is motivated in the next section) using survey sampling techniques when the database is regarded as the population. It is interesting to note that since the data producer is not expected to collaborate with the intruder to reveal the treatment process, he would not revise his risk for a record even if the inside intruder gives him the extra information about his target's values of IVs that are shared by a record in the treated database. Thus, with treated database, the producer, in the interest of confidentiality, may not want to compute risk that varies with target's IVs. However, it would be useful to discourage the intruder by computing separate risk for the targets that are unique-looking or nonunique-looking in the treated database.

In the literature, there exists alternative ways of modeling to define disclosure risk from inside intrusion. Duncan et al. (2001) consider perturbation of small cells in tabular data by random rounding, and define disclosure risk by the expected value of the risk associated with intruder' s prediction of the true value for a given perturbed cell count. The expectation is computed with respect to the probability model that the intruder might use in describing his belief in possible values of the true cell count. As in any modeling method of measuring risk, its success depends on the validity of underlying assumptions.

Unlike disclosure risk, standard measures such as bias, variance, MSE (mean squared error), and multivariate association defined suitably for discrete or continuous data can be used for assessing information loss. However, as in the case of disclosure risk, a suitable randomization mechanism (known or modeled) that governs the relationship between the treated database and the original database is needed to compute the information loss.

4. Motivation of the proposed method

As mentioned in the introduction, there are four desirable features of any disclosure treatment. It turns out that the proposed method can have all these features due to problem simplification induced by the paradigm shift in which the database is itself regarded as the population, and consequently the problem reduces to that of inside intrusion. Moreover, besides measuring risk of disclosure from sample uniques, it can measure other types of disclosure risk as well such as that of nonuniques with common values of at least one SV (sensitive variable).

To motivate the proposed method, we first note that different types of methods of disclosure treatment for creating (nonsynthetic) PUFs essentially use some form of substitution (or data perturbation) and/or subsampling (or data suppression). Using the taxonomy of Cox (1996), methods under data aggregation (recoding and micro-averaging), data modification (random rounding and random noise addition), and data fabrication (swapping and imputation) can be classified under the broad category of substitution, and methods under data abbreviation (suppression and subsampling) can be simply referred as subsampling. Observe that if known random mechanisms are used for introducing substitution and subsampling for disclosure treatment, then both disclosure risk and information loss as defined earlier can be computed. In practice, we need to design the random mechanisms optimally so that an objective function (related to disclosure risk) can be minimized subject to constraints on information loss.

The above optimization problem is analogous to defining an optimal sampling design for a finite population and so suitable solutions can be obtained using survey sampling techniques. To see this analogy, we make the following observations: (i) Conducting a census of a finite population is expensive but entails no loss of information. Similarly, releasing a database without any treatment has high disclosure cost (defined in the following section) but entails no loss of information. (ii) In surveys, item nonresponse is common, and a suitable (superpopulation) model is used for imputation. In disclosure treatment, using a randomization design to select a sample for substitution induces item nonresponse for imputation. To obtain an optimal design, disclosure cost can be minimized subject to bias constraints. (iii) In survey sampling, a randomization design is used to take a sample which entails some loss of information but reduces monetary cost. The design can be made optimal by minimizing monetary cost subject to precision constraints. Similarly, in disclosure treatment, use a randomization design to select a subsample, and make it optimal by minimizing disclosure cost subject to precision constraints. (iv) In survey sampling, sampling weight calibration is used to improve estimation, and software tools such as SUDAAN can be used for data analysis. Similarly, in disclosure treatment, calibration can be used to improve estimation from treated data, and standard survey data analysis tools can be used for analysis as treated data is rendered into a survey data.

With the above motivation, it seems almost natural to propose a treatment method as defined in the next section.

5. MASSC: the proposed method for disclosure treatment

Under the inside intrusion framework, the database is the population, and one can find records at risk that are uniques with respect to IVs. Nonunique records (i.e., clusters of records that share the same values of IVs) are also at risk if they have common values of at least one SV. If no treatment is done, then these records will be at risk with probability 1. The proportion of records at risk can be reduced by global recoding of IVs. Moreover, by local suppression of values of SVs (Willenborg and de Waal, 2001) for the records at risk, the disclosure risk can be reduced to zero. However, it may introduce considerable bias (and hence loss of information) in the estimates based on the treated data because of treatment of all the risky records. A good compromise between risk of 1 and 0 is to allow risk to be positive but small such as being less than 0.2 which corresponds to requiring a cell count to be at least 5 in publication of tabular data. This can be achieved by introducing uncertainty about a record being at risk by treating a random subset of records via substitution and subsampling. Substitution of values of

IVs introduces uncertainty about a record being truly unique, or truly being part of a nonunique cluster. Subsampling introduces uncertainty about the presence of a target in the database.

The MASSC process provides simultaneous control on both disclosure risk and information loss in a suitable sense by meeting the following two goals for PUF. For suitable choices of small positive numbers e and d, the goals are

1. Max of RRMSE over a set of key Outcome Variables is at most e, (5.1)
2. Pr(sensitive information about a record is disclosed) is at most d, (5.2)

where RRMSE denotes relative root MSE with respect to the true domain level total parameters as obtained from the original database. The upper bound d on disclosure risk is generally a composite of four measures of risk corresponding to whether the target looks like a unique, or a nonunique-double, or a nonunique-triple, or a nonunique-four-plus in the treated database. Here 'four-plus' denotes that the nonunique cluster size is four or more. For each of these four types of records in the treated database, MASSC computes a measure of disclosure risk, and then an overall single measure can be obtained by taking a weighted average of the four measures, the weights being the relative proportion of each type of record in the treated database. This overall measure is useful as a single summary measure but is an oversimplification and not as meaningful as the individual ones because it allows for ignorance on the part of the intruder about which of the four types of record the target belongs.

The MASSC process can be defined by the following four steps:

Step I. Micro Agglomeration It consists of defining risk strata with respect to core and noncore IVs. Core IVs signify those IVs which will be easily available to the intruder. noncore IVs are those IVs which are not easily accessible to the intruder. Noncore IVs are ranked with respect to anticipated difficulty in obtaining the information about a target. Risk stratum 0 is a group of micro records that are unique with respect to core IVs. Risk stratum 1 is a group of micro records that are new uniques when the first noncore as per the ranking is added to the core IVs. Similarly, other risk strata are defined. The last risk stratum consists of all nonuniques with respect to core and noncore IVs. The above formation of risk strata can be seen to be also applicable to tabular data by representing them as a rectangular file with columns corresponding to cross-classifying variables. Note that risk strata are termed agglomerates of micro records because the records may be quite disparate with respect to the values of IVs. The only reason they belong to a stratum is that they happen to be unique or nonunique with respect to a given set of IVs.

Step II. Substitution In this step, substitution partners for all records are assigned using the nearest neighbor imputation idea of survey sampling such that the donor record closest to the recipient in terms of a distance function based on IVs and possibly SVs. Only values of IVs are typically substituted if the record is selected for substitution. Next each risk stratum is partitioned further into substitution substrata such that contribution to absolute bias over key outcome variables is small from each substratum. To find optimum selection probabilities, y_{hk}, for the hkth substratum, the disclosure cost function is minimized subject to bias constraints for a number of key outcomes. That is

$$\text{Min } \sum_{hk} C_{hk}(1) N_{hk}(1-y_{hk}) \text{ subject to } E_y\left(bias(q_y^*)\right)^2 \le a q_y^2, \qquad (5.3)$$

where $C_{hk}(1)$ is the disclosure cost function chosen as a decreasing function of y_{hk}, N_{hk} is the stratum size, $N_{hk}(1-y_{hk})$ is the expected number of records not substituted, q_y is the true population total of an outcome variable y from the original database, q_y^* is estimated totals after random substitution, a is an upper bound on the bias squared relative to squared population total, and E_y is the expectation operator with respect to the random substitution mechanism. The selection probabilities y can be restricted to lie

strictly between 0 and 0.5. In other words, there is a positive probability for each record to be substituted, but in any substratum, proportion of substituted records is no more than 50%. It can be shown that as y increases between 0 and an upper bound below 0.5, the expected value of bias squared increases but the disclosure cost decreases, i.e., the two functions move in opposite directions as required for the optimization problem.

Step III. Subsampling In this step, first the risk strata are partitioned into substrata such that variability of observations within each substratum with respect to a set of key outcome variables is small. Next optimum selection probabilities, f_{hk}, the disclosure cost function is minimized subject to variance constraints on a set of key outcomes. That is

$$\text{Min } \sum_{hk} C_{hk}(2) N_{hk} f_{hk} \quad \text{subject to } E_y\,(V_{f/y}(\hat{q}_y^*))^2 \le b q_y^2, \qquad (5.4)$$

where $C_{hk}(2)$ is the disclosure cost function chosen as an increasing function of f_{hk}, $N_{hk} f_{hk}$ is the expected number of records sampled-in, $\hat{q}_y^*$ is an estimate based on the sample from the substituted database, $V_{f/y}$ is the variance operator under the random subsampling mechanism given the substituted database, and b is the upper bound on variance relative to the squared population total. The selection probabilities f_{hk} can be restricted to lie strictly between 0.5 and 1. In other words, every record has a positive chance of being sampled out, but the proportion of records sampled out from given substratum is no more than 50%. It can be shown that the two functions, variance and the disclosure cost function move in opposite directions as f_{hk} increases, a necessary condition for optimization. The sample is then selected using a design similar to a two-phase nested design where the original database is regarded as a first phase design. By nesting, we mean that the second phase units are selected within first phase PSUs using appropriate selection probabilities. If the original database is not based on PSUs, then pseudo-PSUs can be created for this purpose. This modification allows one to use standard analysis tools for single phase designs while analyzing MASSC-treated data.

Step IV Calibration In this step the sampling weights obtained from the previous step as inverse of selection probabilities are calibrated so that estimates based on a set of key outcome variables match exactly those obtained from the original database. This is a sort of poststratification, and helps reduce the bias caused by substitution, see Folsom and Singh (2000).

It is important to note that under MASSC all records whether at risk or not have a positive chance of treatment by substitution and/or subsampling. It may be mentioned that treating a random portion of not at risk records may be a small price to pay in terms of information loss, as it tends to introduce sufficient uncertainty in risk status of records to achieve an adequate level of disclosure limitation without treating too many records. The information loss is measured by max RRMSE over a set of key outcome variables with an upper bound of e, and has the property that

$$\max_y RRMSE(\hat{q}_y^*) \le e$$

because

$$
\begin{aligned}
E_{yf}\left(\hat{q}_y^* - q_y\right)^2 &= E_y E_{f|y}\left(\hat{q}_y^* - E_{f|y}\hat{q}_y^*\right)^2 + E_y\left(E_{f|y}\hat{q}_y^* - q_y\right)^2 \\
&= E_y V_{f|y}\left(\hat{q}_y^*\right) + E_y\left(q_y^* - q_y\right)^2 \\
&= E_y V_{f|y}\left(\hat{q}_y^*\right) + V_y\left(q_y^*\right) + \left(E_y\left(q_y^*\right) - q_y\right)^2 \qquad (5.5) \\
&= E_y V_{f|y}\left(\hat{q}_y^*\right) + \left[V_y\left(q_y^*\right) + Bias^2\left(q_y^*\right)\right] \\
&\le (b + a) q_y^2 = e^2 q_y^2
\end{aligned}
$$

The four measures of disclosure risk can be computed as follows. For simplicity we assume there are only two risk strata, uniques (U), and nonuniques (NU). Let 'U' denote a true unique and 'u'a pseudo unique; pseudo in the sense that the record looks unique but with substituted values, or that it is a nonunique but classified as unique after its or other's treatment. Similarly 'NU' and 'nu' denote respectively true and pseudo nonuniques.

Now, for a record that looks unique in the treated data set, we have the probability of its disclosure by an inside intruder,

$$d_u = p_U \times (1 - y_U) \times f_U \times (1 - c_U) + p_{NU} \times (1 - y_{NU}) \times f_{NU} \times c_{NU}, \qquad (5.6)$$

where in the first term, the first component is the probability that it comes from the U stratum, the second component is the probability it survives substitution given that it comes from U-stratum, the third component is the probability that it survives being sampled out given that it comes from the U-stratum (note that the f -probabilities are stratum-specific and not record-specific, although they depend on the set of records substituted), and the fourth component is the probability that the record is not misclassified as a nonunique given that it survived both the treatment of substitution and subsampling; the second term is defined similarly in which the last component is the probability the record is misclassified as a unique given that it survived both substitution and subsampling.

Similarly, if a record looks like a nonunique double in the treated database, the probability of its disclosure is given by

$$
\begin{aligned}
d_{nu(d)} &= p_U \times (1 - y_U) \times f_U \times c_U \times h_{U(d)} \times (1 - V_d) \\
&\quad + p_{NU} \times (1 - y_{NU}) \times f_{NU} \times (1 - c_{NU}) \times h_{NU(d)} \times (1 - V_d) \qquad (5.7)
\end{aligned}
$$

where $h_{U(d)}$ denotes the probability that a unique is classified as a double given that it survives treatment but gets misclassified as a nonunique, $h_{NU(d)}$ denotes the probability that a nonunique is classified as a double given that it survives treatment, and V_d is the probability that a nonunique double in the treated database has no common value for all sensitive variables; it is assumed to be same for records from both unique and nonunique strata in the interest of stability of its estimates.

Similarly, $d_{nu(t)}$ for records that look like a nonunique-triple, and $d_{nu(o)}$ for records that look like nonunique-others (i.e., four-plus) can be defined. A single measure of disclosure risk d , although not as meaningful, can be obtained by taking a weighted average of the four measures as mentioned earlier. After the treatment, diagnostics for disclosure and analytical quality are computed to check for adequacy. If not, then the process is repeated with revised values of parameters e and d , recoding (or coarsening) of IVs, if necessary, to reduce the number of records at risk.

6. Calculation of Disclosure risk using MASSC: an illustration

We present a hypothetical example (table 6.1 below) to illustrate computation of four types of disclosure risks after MASSC has been applied. The IVs are age (four categories: 1= 12, 2=17, 3=21, 4=25), and gender (1=M, 0=F). The SV is marijuana use (1=use, 0=nonuse). Observe that record 2 survived subsampling but not substitution, and looks unique. Record 3 survived the treatment but gets misclassified as a nonunique-double. Record 4 survives both treatment, but looks unique. Record 6 survives treatment, was double but now looks unique. Record 7 was substituted, was double but now looks triple. Records 8 and 9 survive treatment, still triple but with a different member, and record 10 was substituted, was triple, but now looks double.

It follows from the original and treated data sets that

$$p_U = 3/10; y_U = 1/3; f_U = 2/3; c_U = 1/1$$

$$p_{NU} = 7/10; y_{NU} = 2/7; f_{NU} = 6/7; c_{NU} = 2/4$$

therefore,

$$\begin{aligned} d_u &= (3/10)\times(1-1/3)\times(2/3)\times(1-1) + (7/10)\times(1-2/7)\times(6/7)\times(1-2/4) \\ &= 0.2143 \end{aligned}$$

Similarly,

$$\begin{aligned} d_{nu(d)} &= (3/10)\times(1-1/3)\times(2/3)\times(1/1)\times(1/1)\times(1-0/1) \\ &\quad + (7/10)\times(1-2/7)\times(6/7)\times 91-2/4)\times(0/2) \\ &= 0.1333+0 \end{aligned}$$

It is easy to see that $d_{nu(t)}$ is zero because $h_{U(t)}$ is 0 and V_t is 1 where 't' stands for nonunique triples.

Similarly, it can be seen that $d_{nu(o)}$ is also zero where'o' stands for nonunique others, i.e., 4+.

Thus we have computed four measures of delta after treatment. In the original database, there are seven records that are at risk, three among uniques and four among nonuniques, i.e., a total of 70% records at risk before treatment. Here probability of disclosure is 100% for these records. However, after treatment, no record is at risk with 100% chance. The probabilities of disclosure for various records in the treated database are given by delta values as each record is either a unique or nonunique looking , and then within nonunique it is either double, triple, or four plus.

7. Simulation Results

8. Concluding Remarks

Table 6.1 Example of Risk Calculation

Data before Treatment Data after Treatment

Obs	Age	Gender	MRJ use	Comment	Obs	Age	gender	MRJ use	Comment
1	1	1	0	unique	1	1	1	0	Sampled out
2	1	0	1	unique	2	1	1	1	Pseudo-unique
3	2	1	1	unique	3	2	1	1	Pseudo-nonunique
4	2	0	1	Nonunique double	4	2	0	1	Pseudo-unique
5	2	0	1	Nonunique double	5	2	0	1	Sampled out
6	4	0	0	Nonunique double	6	4	0	0	Pseudo unique
7	4	0	0	Nonuniue double	7	3	1	0	Pseudo-triple
8	3	1	1	Nonunique triple	8	3	1	1	Nonunique triple
9	3	1	0	Nonunique triple	9	3	1	0	Nonunique triple
10	3	1	1	Nonunique triple	10	2	1	1	Pseudo double

References

Bethlehem et al. (1990)
Cox , L (1996)
Duncan and Keller-McNulty(2001)
Duncan et al. (2001)
FCSM (1978)
FCSM (1994)
Fienberg and Makov (1998)
Folsom and Singh (2000)
Fuller (1998)
HIPAA (2000)
Little (1998)
Skinner and Holmes (1998)
Skinner and Elliott (2002)
Willenborg and de Waal (2001)
Zaslavsky and Horton (1998).