Wiley StatsRef:
Statistics Reference Online

Data Masking for Disclosure Limitation

By *Paul B. Massell, Michael H. Freiman, and Laura V. McKenna*

Keywords: *disclosure limitation, statistical disclosure control, data swapping, rank swapping, multiplicative noise, synthetic data, cell suppression*

Abstract: Governmental agencies that conduct surveys and censuses collect data from respondents with the purpose of releasing it in the form of statistical summaries. The more detailed the summary is, the more likely a data intruder will be able to extract confidential data about individual respondents from the released data. However, there are various ways of redesigning the data product and/or modifying the data themselves to protect the data while preserving their usefulness. We discuss methods that achieve these two goals: (i) a data intruder will not be able to extract, with high confidence, confidential data directly from the data product or derive confidential microdata from several data products; and (ii) the released data are still quite detailed and useful to most data users, including researchers. Such "data-masking" methods comprise a fast growing field often called *statistical disclosure control*.

We discuss some simpler methods that have been used for decades, such as detail reduction, cell suppression, and data swapping; some methods developed in the 1990s, such as rank swapping, data shuffling, and multiplicative noise; and some methods developed in recent decade, such as randomization of microdata with constraints (PRAM) and synthetic data.

Government statistical agencies are in the business of collecting and disseminating data. Most agencies assure census andsurvey respondents that their responses will be kept confidential. Unfortunately, the more information provided to data users is, the greater will be the possibility that a data user can determine data values belonging to a particular respondent (one type of disclosure). Disclosure avoidance techniques used by the agencies to protect confidentiality must be continually improved to keep pace with the growing demand for an ever wider variety of data products. This article provides an overview of recently developed disclosure avoidance techniques with respect to both microdata files and tabular data that may facilitate broader dissemination of data.

U.S. Census Bureau, Washington, DC, USA

Update based on original article by Paul B. Massell, Wiley StatsRef: Statistics Reference Online, © 2014, John Wiley & Sons, Ltd

DOI: 10.1002/9781118445112.stat00064.pub2

1 Traditional Disclosure Avoidance Techniques

In general, disclosure avoidance can be done in two ways: one can limit the amount of data that is given out and the other can mask the data by adding noise, swapping values, and so on. An excellent overview of the methods used by federal statistical agencies to protect data is given in Ref. 1. There are now several books that give an even broader overview of this subject, sometimes referred to as *statistical disclosure control*, for example, Refs 2–4. Government statistical agencies have traditionally opted for the approach of limiting the amount of information given out, but they are now developing methods that will allow more data to be released, albeit in masked form.

Microdata are unit-record data. Each record corresponds to an individual respondent. Agencies have released microdata files from demographic censuses and surveys since the 1960s. For microdata, all identifiers – such as name and address – are removed from the file. Agencies usually limit the geographic detail on the data file. They often categorize continuous variables or combine sparse categories in an attempt to prevent matching the microdata file with other data files containing identifiers. Extreme values of numerical variables often increase the risk of disclosure for the records in which they appear. Suppose that only about 1% of the records in some database have income that is above $120 000. The agency may choose to do a simple type of topcoding, in which all incomes equal to or above $120 000 are set equal to $120 000. There are other types of topcoding (i.e., recoding values above a certain level) that convey more information about the upper 1% of the income distribution but still provide protection against disclosure. For example, the agency could set the income for the top 1% of records to the median of all values in that group.

Establishment tabular data are data collected from establishments and published in aggregated form in tables. For such data, agencies have used cell suppression since the 1950s. They remove from publication all cell values (called *primary suppressions*) that pose too much disclosure risk. If these values were published, users could closely estimate a respondent's value. Because most tables are additive, agencies then have to remove a sufficient number of additional cells (called *complementary suppressions*) to make sure that the primary cell values cannot be derived or estimated closely through addition and subtraction of cells that are published. By suppressing the complementary suppressions as well as the primary ones, agencies are again limiting the amount of information released.

For demographic data (about people or households) that are published in tabular form from a census, cell suppression may be used. However, for the decennial US Census of Housing and Population, a swapping routine is used instead[5]. This method selects pairs of similar households in different geographical areas that agree on certain key variables, and then swaps the geographic identifiers within each pair before tabulation. The American Community Survey, the ongoing US Census Bureau survey that replaced the decennial long form, is also protected by swapping microdata records[5].

2 Why New Methods Are Needed

The need for new protection methods stems from the technological revolution and the ever-increasing demand for data. In addition, an ever-increasing pool of identified data is available to universities, governmental (e.g., state) agencies, and the general public. Some of these data could be matched to a statistical agency's survey-based microdata files, which in turn might lead to disclosures.

In addition, there is increasing interest in the idea of statistical agencies incorporating data from other government agencies in many surveys and censuses. This involves using a name or other identifier for each respondent in order to match the two sets of data and then releasing the linked data with the identifiers removed. This greatly increases the risk for microdata files because the agency supplying the additional data has great matching capability. It could use the data it supplied to link respondents to

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

responses and thus obtain the statistical agency's confidential information. Today's matching software is very powerful, and simply reducing the detail of variables on the files would not be enough to prevent such matching. Section 3 discusses ways to protect microdata when the microdata are released to the public.

For establishment tabular data, agencies want an alternative to cell suppression. Cell suppression must be performed on each table separately; suppression patterns must be coordinated among all tables; the technique greatly limits a user's ability to aggregate data; and much information is lost in the form of complementary suppressions. The worst of these disadvantages is probably the required coordination of cell suppression patterns among all tables. When the number of tables to be published is large, this coordination can require a major effort, involving both bookkeeping and computation. This coordination requirement arises not just when the agency releases the initial set of tables for some survey/census but later on when special tables, requested by data users, are released. Section 6 discusses cell suppression and other ways to protect magnitude data tables. It includes ways to protect linked tables.

When masking methods are difficult or too time-consuming to apply to certain data needed by a researcher, another possibility is to allow access to a secure site housing the data, sometimes referred to as a *data enclave*[6]. The US Census Bureau, the National Center for Health Statistics (NCHS), and the Agency for Healthcare Research and Quality (AHRQ) have all established Research Data Centers (RDCs), where researchers may perform analyses in a secure environment, with the NCHS and AHRQ RDCs located at the agencies' respective headquarters, while Census RDCs are located mostly at research universities around the country. Any output at any of the RDCs must be approved by RDC staff before it can be removed from the RDC to ensure that no confidential data are revealed or can be inferred from the released research. In addition, a researcher at the Census RDCs must submit a proposal explaining a need for access to confidential data, and if the application is accepted, the researcher must be given "special sworn status" before accessing the data. Similar procedures are in place with the NCHS and AHRQ RDCs[7]. A variant of this is to allow online access to authorized users who access the database via secure computer lines. NORC at the University of Chicago also manages data enclaves, either physical or online.

As the administrative processes associated with using an RDC—either physical or virtual—can be rather onerous, other options are desirable. One possible approach to releasing data is a confidentiality-preserving remote access system, a type of computer system well suited to the World Wide Web. An interface allows a user to request a custom tabulation or other analysis of confidential data and to receive a result without ever having access to the underlying microdata. All computations occur behind a firewall to protect the confidentiality of the data. Such a system may sometimes refuse to give output for a particular query if doing so is deemed overly risky. For example, comparison of two tables on data universes that are very similar but not identical to each other can lead to disclosures on individual respondents, so certain checks are usually in place to prevent output from being given when this is a concern. If an analysis does not satisfy these checks, output will not be provided to the user. Results may also be computed on slightly perturbed data, rather than the precise raw data, but the intention in most cases is to keep such perturbation to a minimum. Systems of this sort include the Microdata Analysis System (MAS) that is currently in development at the US Census Bureau. Similar systems are being created by statistical agencies around the world; see, for example, Schouten and Cigrang[8].

Some remote access systems – such as the MAS and Privacy Preserving Analytics[9], developed by CSIRO in Australia – allow off-the-shelf analyses (e.g., tables, regressions, and descriptive statistics). Other systems – such as ANDRE (run by the United States National Center for Health Statistics)[10], Real Time Remote Access (Statistics Canada)[11], the Data Analysis System (National Center for Education Statistics), and LISSY (the LIS Cross-National Data Center in Luxembourg)[12] – are hybrids between the RDC and remote access system models, allowing the submission of code in SAS or another statistical language, with the code and results subject to disclosure checks and in some cases partial censoring before release. These systems may also require a fee or application process before they can be used.

DOI: 10.1002/9781118445112.stat00064.pub2

3 Methods for Protecting Microdata Released as a Data Product

3.1 Protection by Sampling and Reduction of Geographic Detail

If an agency plans to release microdata from a survey with a sampling fraction less than one, the very fact that the data represent only a fraction of the total population typically provides great protection. This is because when a data intruder finds a match to a unique record in the released agency microdata using some key, he or she cannot be certain that there is a unique record in the population with that key. In other words, uniqueness in the population of a set of variables is not revealed in the microdata. Because of this protection, when microdata are released from a census, typically only a fraction of the full set of records is released.

One variable whose level of detail is very important to a data intruder is geography. For a microdata file that consists of household records, if a geography variable identifies the block in a residential area in which there are only about 10 households per block, this geographic datum will likely be quite helpful to a data intruder. On the other hand, if the geography variable identifies only a large area, which contains, say, 100 000 households, then the data intruder is likely to have a much greater challenge.

3.2 Protection by Reducing Detail of Demographic Variables

Just as reducing the geographic detail increases the challenge to a data intruder, so reducing the detail of other variables that are often used by data intruders, the agency can make the disclosure challenge much greater for them. For person or household records, variables such as age, sex, and race are especially useful to data intruders. For age, an obvious way to reduce the detail is to release age intervals rather than specific year; for example, 5-year intervals are often used. For race, the agency might decide to reveal the race of a person only if there are many others (e.g., 100) of the same race in the geographic area specified in the record. Of course, as the level of detail is decreased to increase the protection of the data, the utility of the data to users is diminished.

3.3 Protection by Imputation

Imputation has been used for several decades to create values for missing data (i.e., data values not supplied on a respondent's form) or to replace data values that are blanked during editing of the raw data. "Hot deck" is a method that has been used at the Census Bureau since 1947. It involves defining a notion of "similarity" among a set of records, based on the values of a few basic variables. If, say, five records in a given dataset are "similar" to the record for which there is a missing value for variable X, and some of the other records *do* have a value for X, one can define various ways of deciding which of those records should become the donor record and donate its value to the record in need of a value.

The imputation method may be extended so that it can be applied to records in which several variables, perhaps all, are missing. In those cases, one can view imputation as a method for creating new records that come from the same population or sampling frame as the original set of records. However, records formed in this way will be realistic, but not real; that is, they do not correspond to a real member of the population. These microdata may be viewed as an elementary type of synthetic data (see Section 3.5). Synthetic microdata have nice protection qualities. The challenging task is for dataset creators to evaluate the data quality of the new records to see that they reflect closely the original data[13].

It is possible to combine two or more methods to generate a single microdata file that has some values that are not in the original data but are based on it. A concrete example of this hybrid approach was used by Larsen and Huckett[14] to create microdata of person-level records that contain data about state tax

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

burden and standard demographic variables. In their method, hot deck imputation was combined with rank swapping (see Section 4.2) to create values for certain variables and quantile regression is used for other variables. Their method may be viewed as a simpler approach to the generation of safe releasable microdata than synthetic data methods (discussed in Section 3.5), which use complex models.

3.4 Protection by Multiplicative Noise

At the Census Bureau, multiplicative noise has been used for modifying microdata that are used to generate magnitude data tables, mainly economic tables. However, to date, the Bureau has not considered this method sufficiently protective of microdata to allow for release of microdata itself. This decision may be related to certain features of economic data; for example, skewness. For microdata without that feature, for example, certain demographic data, multiplicative noise may be suitable as a protection method for microdata. Anticipating the future use of multiplicative noise to protect microdata, Klein, Mathew, and Sinha have studied the effect of this type of noise on statistical inference in recent papers. In those papers, a comparison is made of the noise-based method for creating synthetic data with synthetic data created via multiple imputations[15].

3.5 Releasing a Synthetic Version of the Actual Microdata

Another option for protecting data is to create *synthetic data* that mimic the real data. This provides disclosure protection because any given observation in the synthetic data need not be directly associated with an observation in the real data. One can synthesize all variables for all records (full synthesis) or a subset of variables for a subset of records (partial synthesis). If doing partial synthesis, we target records and/or variables that have a potential disclosure risk and those variables that are causing this risk. We can synthesize demographic data and establishment data, although demographic data are easier to model and synthesize because establishment data are typically quite skewed. We can synthesize data with a goal of releasing the synthetic microdata or some tabulation or other type of product (such as a map) generated from the synthetic microdata. Finally, we can generate one implicate (one synthetic dataset) that looks similar to the original file, but with synthetic data; or we can generate several implicates (several different synthetic datasets) that could be released together. Multiple synthetic implicates can be analyzed using multiple imputation analysis techniques. Zayatz[16] provides an overview of this method. Reiter[17] finds that as long as the number of copies of the dataset that are created has only a small effect on the quality of inferences, as long as that number is "reasonably large."

Generally, creating synthetic data involves fitting a model that is posited to describe the true data and then drawing new data from the model. Synthetic data can be created by treating the data to be synthesized as missing and using something analogous to a multiple imputation approach[18]. Synthetic data are often generated by sequential regression imputation, one variable for each record at a time. Using all of the original data, we develop a regression model for a given variable; see Raghunathan *et al.*[19]. Then, for each record, we remove the value of that variable and use the model to impute it. We then go to the next variable and repeat the process until all variables have been imputed and none of the original data remain[18]. This imputation method is somewhat different from a pure missing data approach, as it results in a dataset where every observation is based on a single observation from the original dataset.

Although synthetic data have the benefit of not being "real data" for disclosure purposes, such data are not without risk if they are not created carefully. If the model used to generate the synthetic data is overfit, then individual records in the synthetic dataset may be too similar to individual records in the original dataset, and it may be possible to infer information about these individual records, especially when a record is an outlier. If the model is underfit, then the synthetic data may be insufficiently similar to the original data and is likely to lose important analytic properties. These concerns are discussed in Ref. 4, which suggests

DOI: 10.1002/9781118445112.stat00064.pub2

that overfitting is unlikely to lead to a disclosure but could more easily lead to a perception of data being at risk. It is also very difficult to make synthetic data that are like the real data in every respect that could be analyzed, so sometimes the best that can be hoped for is that the data maintain their properties with respect to the types of analyses that seem most important or most likely to be attempted. Because of the delicacy with which synthetic data must be constructed, methods of synthesizing such data are an active research area.

Another useful feature of fully synthetic data is that the data created are akin to a simple random sample of the population of possible realizations of the model, and thus the analysis of the data need not consider any complexities in the survey design[20].

Synthetic data have been used for several purposes, including the Census Bureau mapping application OnTheMap[21,22].

4 Methods for Protecting Count Tables with Categorical Spanning Variables

The expression "spanning variable" is used in Ref. 2 to denote a variable whose categories are used to define the row, column, level, or higher dimension of a statistical table. Some methods are designed to work with categorical spanning variables; others with ordered spanning variables (see Section 5).

Government agencies are presently investigating (and sometimes using) masking techniques that are variants of data swapping. One frequently used technique that is conceptually simple and easy to implement is called *rank swapping* (formerly called *rank-based proximity swapping*)[23]. A more recent variant is data shuffling, which uses some of the ideas of swapping to come up with a perturbed dataset that – like swapping – preserves the marginal distributions of all variables.

4.1 Data Swapping

Data swapping – wherein pairs of households are chosen and their geographic information exchanged – was originally proposed in 1982[24] and has been used for the decennial US Census of Housing and Population since 1990. In response to concerns about effectiveness and preservation of data quality, a targeted swapping procedure was developed. Under this method, households that were considered particularly vulnerable to disclosure were identified, and disclosure control efforts focused primarily on swapping these households. Nonetheless, to create doubt in the mind of the intruder regarding all records, even low-risk households had a possibility of being swapped. A pair of households could be swapped with each other only if they had the same values for the match key, a set of critical variables (such as the numbers of people under and over age 18 within the household) that were kept fixed throughout the procedure. Through this method, all census counts mandated by law were preserved. Swapping is less useful in cases where certain areas are characterized by high demographic homogeneity. For census or survey data, this could occur in group quarters – people living in a nonhousehold arrangement such as a college dormitory or nursing home. As these types of quarters may consist mostly or entirely of people whose age falls into a particular range (or who tend to share some other set of demographic characteristics), a swap that placed someone outside that range into the group quarter would have the potential to be overly conspicuous, and thus another method of disclosure protection, such as synthetic data, might be used for group quarters.

The Census Bureau is currently developing a new form of swapping known as *n-cycle swapping*, in which, in addition to pairwise switching of geographic identifiers between two records, it is allowable to cyclically permute geographic data across n records, for a convenient value of $n > 2$. This is believed to be more effective, as it will lessen the possibility of a record's not being chosen to be swapped because a suitable swapping partner could not be found[25].

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

Table 1. Data swapping.

	Original tables					Tables after applying swapping				
	Geographic Area A									
	White	Black	Asian	Other		White	Black	Asian	Other	
Married	1	2	3	4	10	0	2	3	3	8
Single	5	6	7	8	26	5	6	7	8	26
Divorced	9	1	2	3	15	10	1	2	4	17
Widowed	4	5	6	7	22	4	5	6	7	22
	19	14	18	22	73	19	14	18	22	73
	Geographic Area B									
	White	Black	Asian	Other		White	Black	Asian	Other	
Married	9	8	7	6	30	10	8	7	7	32
Single	5	4	3	2	14	5	4	3	2	14
Divorced	1	9	8	7	25	0	9	8	6	23
Widowed	6	5	4	3	18	6	5	4	3	18
	21	26	22	18	87	21	26	22	18	87

Table 1 gives a small example of how (ordinary pairwise) swapping might affect tabular counts, assuming that race is part of the match key and marital status is not. Some households have been swapped across the two geographic areas; households could be swapped with each other only if they had identical numbers of people of each race, but counts by marital status were allowed to change as a result of the swapping. One nice feature of geography-only swapping is that whenever the data user does analysis on a region that is "closed with respect to swapping" (i.e., all swaps occur with the region), there is no effect on data quality. Note that in each geography area in the example, the marginal totals for each race remain the same after swapping, whereas the other totals (the marginal totals for marital status, and the cross-tabulations of the two variables) change. In addition, the sum of any cell in the Geographic Area A table and the corresponding cell in the table for Geographic Area B remains constant when swapping is performed.

In addition to prohibiting swaps between certain types of regions (e.g., between households in different states), the swapping procedure may be designed to prioritize swaps in other smaller regions when possible, without making an absolute requirement that swaps stay within that region. For example, swaps within a metropolitan area may be preferred but not required in cases where a record has been targeted for swapping but no suitable partner exists within the metropolitan area. As such a method minimizes the number of swaps across metropolitan area boundaries, any analysis of the swapped data at the metropolitan area level will generally differ only minimally from the same analysis of the unswapped data.

4.2 PRAM : a Post-Randomization Method

Data swapping, at least the type described earlier, may be viewed as changing at most one value in a record; namely, the geography value. PRAM (which stands for Post-Randomization Method) is a much more general approach to changing values of variables in a set of records. An agency can create a histogram for each variable in a set of microdata records. Suppose that for a variable X, there are k values, $x_1, x_2, \ldots, x_i, \ldots, x_k$ represented in the microdata, with relative frequencies $rf_1, rf_2, \ldots, rf_k$. One can create a "k by k" left stochastic matrix that allows a change of value of X for some or all values of X. A stochastic matrix is one in which the entries are nonnegative and represent probabilities. A *left* stochastic matrix has columns that sum to one. It can operate on the vector of values $(x_1, x_2, \ldots, x_i, \ldots, x_{k)}$. Note that stochastic matrices are sometimes called *transition probability matrices* or *Markov matrices*. PRAM authors sometimes use a

DOI: 10.1002/9781118445112.stat00064.pub2

right (i.e., rows summing to one) stochastic matrix. There are several special types of left stochastic matrices. One is the identity matrix, that is, the matrix with a 1 for the *i*th entry on the main diagonal and a zero for all other values in the *i*th column. It preserves the values of any input vector; thus, it produces no mixing. Now, consider a matrix that requires that a value x_i *must* be changed. Then, the *i*th entry on the main diagonal would be zero; and the sum of the other (positive) values in the *i*th column would be one. In the general case, a given value is allowed to change to a few other values or to maintain the current one. The PRAM matrix entry for the *i*th diagonal would be close to one if one wanted to change only a small percentage of values. If one wishes to change two variables, X_1 and X_2, with k_1, k_2 values, respectively, one can form a $(k_1 \cdot k_2)$ by $(k_1 \cdot k_2)$ matrix. One can choose to make the PRAM transformations of X_1 and X_2 independent or to impose various relationships.

An important aspect of PRAM is that it introduces randomness into the process of choosing values for the modified records. The code uses a random number generator. The draws from record to record are independent. However, even if the matrix is designed so that the *expected* histogram for each variable is identical to the observed histogram in the original data, the randomness in the *generated* histogram will somewhat different from the original one. The same idea applies to any correlations in the original data that one tries to carry over to the modified records. If the number of records is large, it is likely that the correlations in the modified microdata will be close to those in the original data, but not identical. Although this method does not exactly preserve the global distributions of each variable as is done in data swapping, the distortion of the original data in PRAM may be viewed as acceptable for certain protection scenarios. On the positive side, PRAM is a simpler process, implementable with a simpler program and it can be applied to small geographies; that is, ones with a small number of sampled units. Its performance is more predictable than the swapping program because the latter requires matching constraints that may not be satisfied for units in need of swapping. Original papers on this method are Refs 26, 27. A recent paper is by Nayak and Adesheyan[28].

4.2.1 Example of a PRAM matrix *M*

Suppose that some categorical variable X has three values *a,b,c*, and we wish to modify the values in each of 100 records in such a way that the observed frequency distribution is preserved. Say, there are 20 *a*'s, 50 *b*'s, and 30 *c*'s in observed records. The identity matrix has this property but, of course, we are interesting in other matrices that have at least a moderate level of "mixing." Here is another such matrix

$$M = \begin{pmatrix} 1/4 & 1/10 & 1/3 \\ 1/4 & 1/2 & 2/3 \\ 1/2 & 4/10 & 0 \end{pmatrix} \text{that satisfies } M \cdot \begin{pmatrix} 20 \\ 50 \\ 30 \end{pmatrix} = \begin{pmatrix} 20 \\ 50 \\ 30 \end{pmatrix}$$

where the entries in a given column represent conditional probabilities; for example, column 1 are the probabilities, $P_{a|a}$, $P_{b|a}$, $P_{c|a}$, of changing an "*a*" (or preserving it). Note that these probabilities add to 1. For column 2, entries are conditioned on *b*; and for column 3, they are conditioned on *c*. Note "0" in column 3 implies that all the *c*'s are changed; this may not be desirable in practice.

4.3 Constrained Hot Deck

Constrained hot deck[29] is a new method for protecting microdata; it has some features in common with the synthetic data approach. Like all hot deck approaches, it avoids creating new values, and no distribution of values is created explicitly. The data are modeled, and new imputed values are generated. For each imputed value, the method then finds the closest actual value and replaces the imputed value with the actual value. If this is not enough protection, then rank swapping is also performed (see Section 5.1). The method has been used recently to allow release of much more transportation data in

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

Table 2. Rank swapping.

Income (rank)		Income (rank)	
Original	Masked	Original	Masked
\$1 000 (1)	\$4 000 (2)	\$14 000 (7)	\$10 000 (4)
\$4 000 (2)	\$1 000 (1)	\$16 000 (8)	\$13 000 (6)
\$8 000 (3)	\$12 000 (5)	\$17 000 (9)	\$40 000 (10)
\$10 000 (4)	\$14 000 (7)	\$40 000 (10)	\$17 000 (9)
\$12 000 (5)	\$8 000 (3)	\$48 000 (11)	\$52 000 (12)
\$13 000 (6)	\$16 000 (8)	\$52 000 (12)	\$48 000 (11)

frequency tables from the American Community Survey (ACS) 5-year sample. Previously, when the tables were constructed without this form of microdata protection, Census Bureau disclosure rules required that a large percentage of cells in these "thin" tables, that is, tables with small counts, be suppressed. The tables based on this perturbed data are of value to transportation analysts because the data quality is kept high.

5 Methods for Protecting Count Tables with Ordered Spanning Variables

5.1 Rank Swapping

The method of rank-based proximity swapping[10,23] involves sorting the values for each continuous variable and swapping values pairwise, so that the ranks of exchanged values differ by less than a prescribed amount. Although Brian Greenberg, developer of the idea, indicated that such a swap might retain analytic utility, he did not actually prove it. Moore's research in this area focused on deriving the "prescribed swapping difference" for each continuous variable subject to certain constraints. A small example appears in Table 2. The original data have already been sorted. Each value was swapped with another value whose rank differed by no more than 3 (the prescribed swapping distance). Note that the swap that most substantially affects the ranks of the income records is the swap between the records with incomes of \$43 000 and \$51 000, which have a difference of 3 in income rank and – less relevant for the mechanics of the algorithm – a difference of \$8 000 in absolute income. However, the swap that leads to the greatest perturbation in income numbers in absolute terms is the swap between the records with income \$45 000 and \$57 000.

5.2 Data Shuffling

Another approach to masking numerical data is the *data shuffling* method of Muralidhar and Sarathy[30], which is somewhat similar to data swapping. In this method, the user begins with some dataset where each observation possesses values for a given set of numerical variables, some of which may be public and some of which are confidential. For each observation, the public variables, if any, are to remain the same, whereas the confidential variables are to be modified. To make this modification, the user begins by estimating the distribution of the vector of confidential variables conditional on the public variables, although no particular formulation of these conditional distributions is prescribed. For each observation, this distribution is used to find preliminary perturbed data.

For a given confidential variable, the data shuffling method ranks the values of the perturbed data. Then, for each observation, the perturbed value is replaced with the value with the same rank with respect to

DOI: 10.1002/9781118445112.stat00064.pub2

the true confidential values of that variable. This is repeated for all confidential variables. Data shuffling is similar to data swapping in that the set (or multiset) of masked values of a given variable is the same as the set (or multiset) of unmasked values. This ensures that the data quality is preserved globally but restricts its use to cases in which the empirical distribution of the confidential variable is not itself considered sensitive. Data shuffling cannot currently be implemented by a data user independently, as it is covered by US Patent 7 200 757.

6 Methods for Protecting Magnitude Data Tabular Data

6.1 Determining Whether a Cell Is Sensitive

A cell is deemed sensitive if the contributions to the cell value can be well estimated by a data intruder. Typically, cells are sensitive if the cell value is dominated by one or two contributions. The p% rule is one precise way of measuring two contributor dominance. It can be generalized to cases in which companies share data; that is, forms of collusion (Ref. 1, p. 61).

6.2 Secondary Cell Suppression

The traditional method used by most government agencies to implement disclosure avoidance with establishment tabular data has been cell suppression. Cell suppression has the nice feature that the cell values that can be published do not contain any noise added solely to prevent disclosures. However, for published tables that are additive, that is, a sum row and a sum column are published (which is typically the case), it does have some drawbacks. These are as follows: (i) there is no information published about the sensitive cells, not even a noisy estimate (although rough estimates can be derived by users who have numerical skill); (ii) there are generally many cells, called *secondary* suppressions, that must be suppressed, in addition to the sensitive cells, in order to prevent recovery of the sensitive cell values; and (iii) finding that these secondary suppressions may require use of a long, complicated program developed by the agency if the agency has special requirements regarding the way in which secondary suppressions need to be selected.

Another problem occurs if there are two or more linked (i.e., overlapping) tables (i.e., tables with cells in common). The program that finds the optimal suppression pattern should ideally process simultaneously all tables that are linked to a given table. (Here, "optimal" refers to finding a suppression pattern that minimizes the information loss due to suppression of the secondaries.) This was not possible in previous years owing to the limitations of computer memory and processing speed. Recent advances in computer hardware as well as great improvement in linear programming software now make it possible to process large sets of linked tables, except when the tables are huge. There have also been computational advances in the cell suppression algorithm itself, for example, parallel protection of sensitive cells. Many of these new approaches have been used in the new cell suppression software developed at the US Census Bureau[31].

During the past decade, several variants of traditional cell suppression have been developed. One such method is called *controlled tabular adjustment*. This method produces a set of new cell values for the table being protected that (i) are close to the initial (true) values and (ii) produce the same amount of uncertainty about the true values as does cell suppression. Thus, in this method, the table user must be told the accuracy of the cell values. One nice feature is that the user does not have to try to estimate the value of a suppressed cell; he or she is supplied a noisy estimate already. However, for linked tables, and for tables with specialized types of protection, this method may be challenging to implement[32].

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

6.3 Multiplicative Noise

Cell suppression and its variants operate at the "table level"; that is, one suppresses or modifies cell values but one does not change the underlying microdata. Since 1995, noise methods have been developed that add noise directly to establishment microdata before tabulation. The resulting tables will then be protected and no further "table-level" adjustments are required; see Evans *et al.*[33]. Specifically, one can perturb each responding establishment's data by a small amount, say 10%. Then, if a single establishment dominates a cell, the value in the cell will not be a close approximation to the dominant establishment's value because that value has had noise added to it. By adding noise, one will avoid disclosing the dominant establishment's true value.

Noise would be added to an establishment's data by means of a multiplier. For our 10% example, the multiplier would be near either 0.9 or 1.1 and would be applied to all of the establishment's data items before tabulation. Because the same multiplier would be used with an establishment wherever that establishment was tabulated, values would be consistent from one table to another. Thus, if the same cell were to appear in more than one table, all its microdata contributions would be the same in each table, so the cell (i.e., the sum of the contributions) would have the same value in each table.

We could use a variety of distributions to generate the multipliers, provided that they were centered at or near 1.1 and 0.9. It is a key requirement, however, that the overall distribution be symmetric about 1.0.

In the case where a cell contains only a single establishment, the cell value will contain about 10% noise. Other sensitive cells, in which one large establishment dominates the cell value, will also contain large amounts of noise, because the amount of noise in the cell total will resemble the amount of noise in the dominant establishment (roughly 10%). The more dominant the largest establishment is, the more closely the cell resembles the single-contributor case. On the other hand, cells with only two contributors might have little noise if the noise for one nearly cancels the noise for the other. In that case, the agency needs to decide if publishing a cell value for a sensitive cell is risky even if the data intruder cannot know how much noise has been assigned to each contributor.

There is a weakness to the random noise method. There is a chance that the noise associated with all (or almost all) of the contributions to a cell will be in the same direction. This might lead to a cell value that is almost as distorted as the most distorted of the underlying contributions. One can do simple probability calculations to find the probability that this scenario will arise. In cells with a large number of contributions all of the same magnitude, the random noise cancellation may be adequate to produce cell values that are not greatly distorted. However, for some surveys, an agency may not be willing to depend on this random type of noise cancellation. In that case, the agency may choose a different form of noise cancellation. "Noise balancing" starts with a "table of interest" and selects multipliers for the contributing company values in a way that produces low noise in most cells. This method works best when most of the companies in the microdata contribute to at most one interior cell in the table; see Massell and Funk[34].

The percentage of noise in a cell would be defined as the percentage by which the noise-added value differed from the true, noise-free value. One would have to calculate both values for each cell in order to quantify the amount of noise each cell contained. All cells exceeding a certain noise threshold, say 7%, would be flagged. In this way, users would be alerted to cells whose values contain a lot of noise and hence might be unreliable. In addition, the description of the flag would explain how and why the noise was added, thus assuring users (who might be surprised not to see any suppressed cells) that disclosure avoidance had indeed been performed.

This noise technique is currently being used for protecting tabular data produced in various Census surveys, for example, County Business Patterns survey, the Non-Employer Statistics program (which uses administrative data), and the Commodity Flow Survey. There are some issues related to longitudinal analysis of noisy data that are still being researched.

DOI: 10.1002/9781118445112.stat00064.pub2

Table 3. Tabular protection: suppression versus noise.

Original table					Table after cell suppression					Table after noise addition				
10	20	30	40	100	D	20	30	D	100	D	21	30	38	97
50	60	70	80	260	50	60	70	80	260	52	60	69	81	262
90	10	20	30	150	D	10	20	D	150	90	9	21	D	154
40	50	60	70	220	40	50	60	70	220	40	48	61	70	219
190	140	180	220	730	190	140	180	220	730	190	138	181	223	732

A small example that compares cell suppression with the noise approach is given in Table 3. In the original data, the underlined values are the primary suppressions. In the post-disclosure-avoidance tables, *D* means that the cell value is withheld owing to disclosure avoidance.

Related Articles

Coding: Statistical Data Masking Techniques; **Masking and Swamping**; **Privacy Protection in an Era of Data Mining and Record Keeping**; **Statistical Confidentiality**; **Confidentiality and Computers**.

References

[1] Federal Committee on Statistical Methodology (2005) *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22. Federal Committee on Statistical Methodology, Washington DC, *http://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf* (accessed 25 February 2015).

[2] Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Springer, New York.

[3] Duncan, G.T., Elliot, M., and Salazar-Gonzalez, J.J. (2011) *Statistical Confidentiality: Principles and Practice*, Springer, New York.

[4] Hundepool, A., Domingo-Ferrer, J., and Franconi, L., *et al.* (2012) *Statistical Disclosure Control*, John Wiley & Sons Inc., Chichester, West Sussex, United Kingdom.

[5] Zayatz, L., Lucero, J., Massell, P., and Ramanayake, A. (2009) *Disclosure Avoidance for Census 2010 and American Community Survey Five-Year Tabular Data Products*, Statistical Research Division Report (Statistics #2009-10), *http://www.census.gov/srd/papers/pdf/rrs2009-10.pdf* (accessed 25 February 2015).

[6] NORC at the University of Chicago. *Data Enclave*, *http://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx* (accessed 25 February 2015).

[7] Confidentiality and Data Access Committee (2002) *Restricted Access Procedures, Report, Federal Committee on Statistical Methodology*, Washington DC, *https://fcsm.sites.usa.gov/files/2014/04/CDAC-RAP.pdf* (accessed 25 February 2015).

[8] Schouten, B. and Cigrang, M. (2003) Remote access systems for statistical analysis of microdata. *Statist Comput.*, **13**, 381–389.

[9] Sparks, R., Carter, C., and Donnelly, J.B., *et al.* (2008) Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics®. *Comput. Methods Programs Biomed.*, **91**, 208–222.

[10] Harris, K.W. and Gambhir, V. (2003) *CDC/NCHS Research Data Center, presentation to the Bureau of Transportation Statistics Confidentiality Seminar Series*, *http://www.bts.gov/programs/confidentiality_policy/2003_07_harris/ppt/2003_07_harris.ppt* (accessed 25 February 2015).

[11] Price, D. (2014) *The Real Time Remote Access (RTRA). System*, *http://www.statcan.gc.ca/rdc-cdr/rtra-adtr/rtra-adtr-eng.htm* (accessed 25 February 2015).

[12] LIS Cross-National Data Center in Luxembourg. *http://www.lisdatacenter.org/* (accessed 25 February 2015).

[13] U.S. Census Bureau (2014) *Survey of Income and Program Participation: Data Editing and Imputation*, *http://www.census.gov/programs-surveys/sipp/methodology/data-editing-and-imputation.html* (accessed 25 February 2015).

[14] Larsen, M.D. and Huckett, J. (2009) *Synthetic Data Methods Using Quantile Regression and Hot Deck With Rank Swapping*, presentation to the NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, Suitland, MD. *http://www2.vrdc.cornell.edu/news/wp-content/uploads/2009/08/1-4-Larsen.pdf* (accessed 25 February 2015).

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat00064.pub2

[15] Klein, M., Mathew, T., and Sinha, B. (2013) *A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication*, Center for Statistical Research and Methodology Report, Statistics #2013-02, US Census Bureau, Washington DC. *http://www.census.gov/srd/papers/pdf/rrs2013-02.pdf* (accessed 25 February 2015).

[16] Zayatz, L. (2007) *New Implementations of Noise for Tabular Magnitude Data, Synthetic Tabular Frequency and Microdata, and a Remote Microdata Analysis System*, Statistical Research Division Report, Statistics #2007-17, US Census Bureau, Washington DC. *https://www.census.gov/srd/papers/pdf/rrs2007-17.pdf* (accessed 25 February 2015).

[17] Reiter, J.P. (2002) Satisfying disclosure restrictions with synthetic data sets. *J. Off. Statist.*, **18**, 531–544.

[18] Rubin, D.B. (1993) Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *J. Off. Statist.*, **9**, 461–468.

[19] Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003) Multiple imputation for statistical disclosure limitation. *J. Off. Statist.*, **19**, 1–16 *http://www.jos.nu/Articles/abstract.asp?article=191001*.

[20] Reiter, J.P. (2004) Simultaneous use of multiple imputation for missing data and disclosure limitation. *Surv. Methodol.*, **30**, 235–242 *http://www.stat.duke.edu/~jerry/Papers/sm04.pdf*.

[21] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008) *Privacy: Theory Meets Practice on the Map*, in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference*, pp. 277–286, *https://courses.cit.cornell.edu/jma7/ICDE08_conference_0768.pdf* (accessed 25 February 2015).

[22] Abowd, J. M., Andersson, F., Graham, M., Vilhuber, L., and Wu, J. (2010) *Privacy Guarantees and Analytical Validity of OnTheMap Public-Use Data*, Presentation to IPAM Workshop, Statistical and Learning-Theoretic Challenges in Data Privacy, University of California at Los Angeles. *http://helper.ipam.ucla.edu/publications/data2010/data2010_8556.pdf* (accessed 25 February 2015).

[23] Moore, R. A. (1996) *Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets*, Statistical Research Division Report Series, RR 96–04, US Census Bureau, *http://www.census.gov/srd/papers/pdf/rr96-4.pdf* (accessed 25 February 2015).

[24] Dalenius, T. and Reiss, S.P. (1982) Data-swapping: a technique for disclosure control. *J. Statist. Plann. Inference*, **6**, 73–85.

[25] DePersio, M., Lemons, M., Ramanayake, K. A., Tsay, J., and Zayatz, L. (2012) n-cycle swapping for the American Community Survey, in *Privacy in Statistical Databases, Lecture Notes in Computer Science, Vol. 7556,* Springer, Berlin, pp. 143–164.

[26] Gouweleeuw, J.M., Kooiman, P., Willenborg, L., and de Wolf, P.-P. (1998) Post randomisation for statistical disclosure control: theory and implementation. *J. Off. Statist.*, **14**, 463–478.

[27] de Wolf, P.P., Gouweleeuw, J., Kooiman, P., and Willenbourg, L. (1999) *Reflections on PRAM*. *http://neon.vb.cbs.nl/casc/related/Sdp_98_2.pdf* (accessed 25 February 2015).

[28] Nayak, T. and Adeshiyan, S. (2015) *On Invariant Post-randomization for Statistical Disclosure Control*. International Statistical Review: *http://onlinelibrary.wiley.com/doi/10.1111/insr.12092/epdf*.

[29] Krenzke, T., Li, J., Judkins, D., and Larsen, M. D (2011) *Evaluating a Constrained Hotdeck to Perturb American Community Survey Data for the Census Transportation Planning Products*, Proceedings of the 2011 Joint Statistical Meetings, Miami, FL.

[30] Muralidhar, K. and Sarathy, R. (2006) Data shuffling – a new masking approach for numerical data. *Manage. Sci.*, **52**, 658–670.

[31] Steel, P., Fagan, J., Massell, P., Moore J. r.R, Slanta, J., and Wang, B. (2013) *Re-development of the Cell Suppression Methodology at the US Census Bureau*, Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality, October 28–30, 2013, *http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_USA.pdf* (accessed 25 February 2015).

[32] Cox, L.H., Kelly, J.P., and Patil, R. (2004) Balancing quality and confidentiality for multivariate tabular data, in *Privacy in Statistical Databases, Lecture Notes in Computer Science*, vol. 3050 (eds J. Domingo-Ferrer and V. Torra), Springer, New York, pp. 87–98 *http://sisla06.samsi.info/ndhs/dc/Papers/Cox-Kelly-Patil-LNCS3050-2004.pdf*.

[33] Evans, T., Zayatz, L., and Slanta, J. (1998) Using noise for disclosure limitation of establishment tabular data. *J. Off. Statist.*, **14**, 537–551 *http://www.jos.nu/Contents/issue.asp?vol=14&no=4*.

[34] Massell, P. and Funk, J. (2007) *Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding That Preserves Protection*, Proceedings of the Research Conference of the Federal Committee on Statistical Methodology, Arlington, VA, November 5–7, 2007, *http://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Massell-IX-B.pdf* (accessed 25 February 2015).

DOI: 10.1002/9781118445112.stat00064.pub2