Tactics for Reducing the Risk of Disclosure Using the NCES *DataSwap* Software

Thomas Krenzke [1], Stephen Roey [1], Sylvia Dohrmann [1], Leyla Mohadjer [1], Wen-Chau Haung [1],
Steve Kaufman[2], Marilyn Seastrom [2]
Westat, 1650 Research Blvd., Rockville, MD 20850[1]
National Center for Education Statistics, 1990 K St., N.W., Washington, D.C. 20006[2]

Keywords: Confidentiality, Statistical Disclosure Control

1. Introduction

The National Center for Education Statistics (NCES), part of the Institute of Education Sciences (IES), has developed statistical guidelines for reducing the risk of disclosure prior to release of microdata. Among the standards is the requirement for data swapping, which is used to perturb (or mask) the data in order to reduce the risk of disclosing the identity of individuals or entities without impacting aggregate data.

Recently Westat and NCES have coordinated efforts to implement and enhance the data swapping software originally developed at NCES. The software was approved by the IES Disclosure Review Board (DRB) for use in their data confidentiality procedures for IES projects.

Working with NCES, Westat addressed several issues relating to complex surveys while enhancing the NCES standardized swapping software, *DataSwap*. This paper discusses some data swapping tactics for reducing the risk of disclosure (in situations involving high-risk domains and variables, hierarchical data structures). It also describes practical approaches for maintaining data consistency (e.g., handling skip patterns and recodes) and reducing the swapping impact (addressing missing data in swapping variables, controlled swapping approaches).

1.1 Overview

We begin by providing an overview of the confidentiality procedures and “how” and “why” random data swapping, particularly with *DataSwap* software, has been incorporated into NCES confidentiality standards. Throughout this paper, we will provide some detail of the design and implementation of *DataSwap* software. The practical approaches to reducing disclosure risk in NCES public use survey data dissemination include the old accepted standbys - data suppression, data collapsing and coarsening. Data are also masked if they can be identified by matching the data against a secondary data source. The NCES standards have been broadened over the past few years to include random swapping and, whenever possible, data dissemination with the Data Analysis System (DAS). The DAS uses an internet interface for tabular and regression analysis. The DAS automatically suppresses small cells and provides disclosure controls not possible with a microdata release.

With such high levels of effort behind disclosure avoidance for public use data release, the concern is that the confidentialized data may no longer be analytically useful so that cost and effort of conducting a survey are wasted. We will focus on the capabilities and flexibilities of the data swapping component of the confidentiality procedures to provide an understanding of how data consistency can be maintained and how the impact of swapping can be reduced.

1.2 Disclosure Risk Analysis

Why is it Necessary to Include Random Data Swapping into the Confidentiality Procedures?

The NCES confidentiality standards generally require two separate procedures: (1) identifying and masking of potential disclosure-risk individuals/institutions by evaluating the risk of the study variables and comparing them with public databases (when applicable) and (2) implementing an additional layer of uncertainty with the random swapping of data elements. Random swapping creates doubt in the identification of respondents while still preserving relationships. Data swapping is used to reduce the risk of data disclosure for individuals or institutions by exchanging values for an identifying variable or set of variables for cases that are similar in other characteristics. Inasmuch as confidentiality in any data file cannot be absolutely assured, randomized swapping allows the agency to contend that no one can be certain if an individual unit is identified. To help meet the goal of creating uncertainty, the following swapping rule was established: The value of at least one variable has to change for the record to be considered swapped.

1.3 Which Data Files Require Swapping?

Generally all NCES files required swapping. In the case of hierarchical files, all levels must be swapped: the higher level files (e.g., school-level) and the detailed files (e.g., teacher- and/or student-level). Let's use the example of a survey that includes students, teachers of the students and schools of the students. It is generally accepted that you cannot identify a student if you can't identify the school. Nevertheless, should a data sleuth believe she has identified a school; there will be additional doubt in the identification of the student or teacher from the school.

1.4 Why Use *DataSwap* Methodology and Not Some Other Data Swapping Methodology?

DataSwap was reviewed and approved by the NCES DRB. The DRB must approve not only the approach used for confidentiality but the software as well.

It is generally better to use standardized software that can be used across various studies. Familiarity with one software package for both NCES staff and contractors facilitates the understanding of the process and the results. The time, effort, and review required for ad-hoc swapping software would be detrimental to the timeliness and cost of the studies.

The software uses controlled random swapping as the basic approach. "Controlled" means two things. First, it means that the user is responsible for identifying the data swapping variables and parameters. The user/data analyst should be familiar with which data would be the most identifiable and sensitive. The user understands the content of the data as well as the purpose, and focus, of the study and can therefore guide the data swapping procedures accordingly.

The second meaning of "controlled" is that, in *DataSwap*, once the target records are selected, the file is partitioned into swapping cells. The swapping methodology is designed to find a swapping partner that limits data distortion. The methodology includes the use of swapping cells to identify swapping partners in adjacent cells with similar (or identical) weights and close (with at least one or some different) variable values. The pair with the smallest swapping bias is selected as the swapping partner.

The variable for which the bias is computed (called BIASVAR in the software) is specified by the user and is used in selecting the swapping partner for each target case. The swapping bias is a function of the survey sampling weights and the bias variable:

$$\left(w_1 x_2 + w_2 x_1\right) - \left(w_1 x_1 + w_2 x_2\right)$$

where,

w_1 = the weight of the target record;
x_1 = the BIASVAR value for the target record;
w_2 = the weight of the partner record; and
x_2 = the BIASVAR value for the partner record.

Kaufman, Seastrom, and Roey (2005) evaluated the potential for bias due to data swapping. The bias was studied for weighted distributions, variances, and correlations. They concluded that their data swapping procedures did not appear to introduce any important biases into the estimates. However, high swap rates have the potential to result in largely underestimated standard errors.

2. Practical Approaches for Reducing Disclosure Risk

Next we will describe the practical approaches available with *DataSwap* software. The challenges that we face come from balancing the reduction of disclosure risks with retaining data quality. There are numerous approaches that attempt to measure the data disclosure risk including processing cross-tabulations and identifying small cell sizes (less than a predetermined threshold "*n*") – sometimes referred to as the *n*-rule. There are also numerous measures of data utility, including the Hellinger's distance measure (Le Cam and Yang 1990). We have used similar measures as a guide by thoroughly and carefully reviewing swapping results and then interactively fine-tuning swapping strategies to reach the risk-utility balance. The complexities of sample surveys add to the challenge of maintaining that balance.

In complex surveys there are usually numerous data items collected, some of which define key domains of interest. *DataSwap* has the flexibility for targeting records that are high-risk, or are part of a small domain that is high-risk. Clustering is often another result of conducting a sample survey. For example, suppose a sample of schools is selected and then a sample of students within those schools. This results in a hierarchical structure to the data in that once the identity of a school is known, then the disclosure risk of students increases considerably. Also, in complex studies, data collection may occur in waves for a sampled panel in the form of a longitudinal study. We address these aspects of complex surveys in the following sections.

2.1 What Controls Does the User Have?

The user can target records for data swapping. One way is by identifying high-risk domains and assigning parameters to increase the likelihood of their being included in the data swap. This can be done either by increasing the swapping rate in that domain, or by increasing the measure of size for these cases. The measure of size is used in probability proportionate to size sampling for the purpose of creating an unequal the probability of selection for individual cases.

The user can also identify the high-risk variables and include these variables in the swapping procedure. The software allows the user to review the data within the application. The user can run frequencies and cross-tabulations to identify small, and potentially risky cells. Variables considered to be high-risk for disclosure can be given a higher probability of being swapped.

2.2 Hierarchical Data

Many, if not most, NCES studies have hierarchical data; data with releases involving several files, all linked in their hierarchical structure. In a sense, the linked structure forms a system of funnels that the intruder could work through to find the identity of teachers or students.

Swapping procedures and rates are adjusted for these files to account for the hierarchical structure. Swapping is conducted at each level of data collection (as mentioned earlier). The impact of the data swapping is evaluated for each file. However, it is also important to review the cumulative affect of the swapping.

For example, a variable swapped at the school level needs to be evaluated at the teacher and student levels as well. Most analyses are conducted on flattened files, that is, at the student level. Thus the swapping impact should be evaluated by reviewing the means and correlations of all swapped variables at the student level. Since one school may impact 100 or more students, cumulative swapping rates should be considered when determining optimal swapping rates for each file.

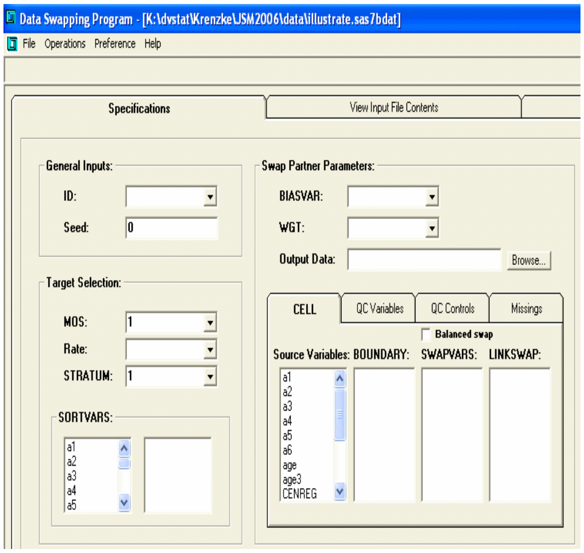
2.3 Longitudinal Studies

NCES also conducts many longitudinal surveys. The data analyst must be cautious when selecting variables for swapping since variables swapped in Wave 1 could have implications for new data collected in Wave 2 and in subsequent waves. A swapped variable that becomes a filter or key variable for data collection in subsequent waves could cause problems in the data. Any demographic identifying variables that are swapped in Wave 1 must be maintained in subsequent waves to keep the data consistent within the file and throughout the study. This consistency includes all related variables, including new ones collected in subsequent waves. For example, if in Wave 1, an individual’s educational status had been swapped from “high school dropout” to “high school graduate” and in Wave 2, the respondent received a GED, the data would not be consistent unless the education status in Wave 2 was modified as well. Often samples are added or freshened in subsequent waves; the new data should be swapped in a manner consistent with what had been conducted in Wave 1. This can be done by (1) swapping the new data items with the original respective swapping pairs, (2) selecting new random swaps for the new sample, or (3) assigning logical values to the new data.

3. *DataSwap*

We have described some practical approaches in applying data swapping to reduce disclosure risk. We now describe the software. The software discussion will help us further explain some other practical approaches to implementing the swapping task, specifically related to maintaining data consistency and reducing swapping impact.

The software is written as a SAS macro with a Windows-based interface that proves helpful in testing swapping strategies. Figure 1 provides a screen shot of the specification screen that we will refer to in our description of some of the main components of the software.



Data Swapping Program - [K:\dvstat\Krenzke\JSM2006\data\illustrate.sas7bdat]

General Inputs:

ID:

Seed: 0

Target Selection:

MOS: 1

Rate:

STRATUM: 1

SORTVARS: a1, a2, a3, a4, a5

Swap Partner Parameters:

BIASVAR:

WGT:

Output Data:

□ Balanced swap

Source Variables: a1, a2, a3, a4, a5, a6, age, age3, CENREG

BOUNDARY:

SWAPVARS:

LINKSWAP:

Figure 1. Screen shot from *DataSwap*

The first component of the software is the specification of the input parameters. In the upper left of the screen illustrated in Figure 1, the user can specify the ID variable, and a seed for the selection of swapping targets, which is useful if the user wants to duplicate swapping already processed. For example, if the user wants to evaluate the impact of swapping on an additional key outcome variable, then the program can be reprocessed without changes to the swaps, and the impact on the additional variable can be evaluated with the *DataSwap* output reports.

The second component, which is specified on the lower left hand side of the screen in Figure 1, is to select the target records for swapping. In doing so, the user needs to consider several facets of the swapping process: including the swapping rate, how the initial target records are to be selected (including stratification, probability proportionate to size, and simple random sampling), and whether or not some records will be targeted for swapping.

The third component, which is shown on the right hand side of Figure 1, involves finding swapping partners for the target records. The user needs to specify how the swapping cells are to be formed. The swapping cells are formed by cross-classifying boundary and swapping variables (i.e., identifiers such as age and education attainment categories). The swapping partner search is limited to within the boundaries defined by the concatenation of the boundary variables.

The fourth component is to actually swap the data. By definition, the variables that are allowed to be swapped are referred to as swapping variables. Since, by definition, the values of the boundary variables are the same between the targets and their partners, boundary variables are not considered “swapping variables.” The swapping of data occurs as the values of the swapping variables are switched between each target and their respective partners. In addition, there is a parameter that allows a list of secondary variables to be swapped whenever the primary swapping variables change values.

Lastly, output reports for the user (only), DRB chair, and the DRB members are generated to allow the user to evaluate the swapping impact. The reports can also be controlled through the software parameters. The user-only output report gives more details of the process for checking and quality assurance. This includes general results of the swapping, and detailed information on what was swapped for each swapping pair. The output also compares unweighted and weighted frequencies, means, and correlations before and after swapping.

The DRB chair report provides the actual swapping rates for all variables. The DRB members’ output report contains the same information discussed above, but for higher aggregate domains. Flags are generated to highlight potential bias areas after the swapping occurs.

3.1 *DataSwap* Methodology Illustration

We provide a simple illustration to describe the swapping process administered in the software. The example data, shown in Figure 2, has two swapping variables: RACE, which has 2 categories and AGE, which has 2 categories. The software requires a variable to be specified for computing the bias introduced by the swapping partners. AGE is the variable for which the bias is computed in this example. The sampling WEIGHT is also used in computing the bias. There are only seven cases in this example falling into three swapping cells formed by concatenating the RACE and AGE variables.

ID	Race	Age	Weight
1	1	2	140
2	1	2	540
3	2	1	790
4	2	1	495
5	2	1	590
6	2	2	500
7	2	2	955

Figure 2. Illustration of the *DataSwap* algorithm, prior to swapping

The first step is to select the target record at random. Suppose, as illustrated in Figure 3, case #4 is selected within the second cell. The target record has RACE = 2, AGE = 1 and WEIGHT = 495.

ID	Race	Age	Weight	Note
1	1	2	140	
2	1	2	540	
3	2	1	790	
4	2	1	495	Target
5	2	1	590	
6	2	2	500	Swapping partner
7	2	2	955	

Figure 3. Illustration of the *DataSwap* algorithm, after swapping

Next, the software searches for the swapping partner among the two adjacent cells. Here case #6, which has RACE = 2 and AGE = 2, is selected. Among the two adjacent cells, Case #6 contributes least to the overall swapping bias since it is closest to the target record in terms of AGE (which was the bias variable), and WEIGHT. Once the partner is identified, the values of the variables RACE and AGE are swapped between the two records.

Because of the way the cells are formed, and the limit of searching only within adjacent cells, it is certain that at least one value is changed among the variables. Another aspect of this process is that the right-most swapping variable (AGE in this example), is the variable that will be swapped the most often. As mentioned previously, the software can address high-risk variables; one way to do this is to include them toward the end of the list of swapping variables so that their values change most often.

4. Practical Approaches for Maintaining Data Consistency

Having described the *DataSwap* software, its features, and algorithm, we discuss practical applications. One of the issues with data swapping is the ability to produce data that are consistent between the swapping variables and other survey items. In the software this is handled by the parameter LINKSWAP (see Figure 1). The LINKSWAP parameter asks for a list of variables associated with each swapping variable. When the swapping variable changes value, all variables linked to it are changed as well.

One situation in which this would be advantageous is when public and restricted use files are to be released. For instance, the public use file may contain a recoded age variable, but the recode is not enough to protect one’s identity. As a result, it is specified as a swapping variable. The recoded age is swapped to ensure that a change in data value has occurred. Then through the LINKSWAP parameter, the restricted variable (detailed age) is swapped along with it to ensure consistency between public and restricted use files.

The LINKSWAP parameter can also be useful when handling the complexities of retaining skip patterns in the data and as well as in other situations, such as keeping a recoded race/ethnicity variable consistent with the multiple reporting categories from which it was derived.

Careful attention to maintaining data consistency is also needed when a swapping variable is used in poststratification or other calibration adjustments made during the weighting process. In this case it is best if the swapping occurs before poststratification, so that the weighting adjustments are conducted on the swapped data. This ensures that the sum of the final weights will match to known control totals. When swapping is conducted before the poststratification adjustment, base or preliminary weights should be computed and used to reduce the swapping bias.

If swapping must occur after a poststratification adjustment using the swapped variables, swapping should be restricted to cases with the same weights in order to ensure that the sum of the weights will still equal the control totals.

5. Practical Approaches for Reducing the Swapping Impact

Now we turn to some approaches that we have used in order to reduce the swapping impact on survey items.

5.1 Reducing the Impact on Key Outcome Variables

To minimize the impact on the key outcome variable, for example, test score, categories of score can be formed and the swapping specified so that pairs are within the same category of test score. This helps to reduce the impact on the associations with test score. Using the output reports, the swapping impact on the original score variable can be monitored through the pre- and post-swapping statistics.

The standard method disproportionately swaps the right-most variable in the definition of the cells. For example, if the variables *X*1, *X*2, and *X*3 are used to define the cells, then *X*3 is swapped the most. This is useful when the file contains only one or two highly identifying variables. The alternative method provides a more balanced distribution of swapping among the variables. If there is too much of an impact on a certain variable, this method offers a way of reducing the swapping impact by spreading the swapping across the variables, so that the right-most variables are not swapped as heavily. We call this “balanced swapping”, since it spreads the impact across variables more evenly than the standard approach.

Another way of reducing the impact is to ensure that there are good partners available. One way is to first run frequencies or summaries to determine the swapping cell sizes. If the cells are too small, the likelihood of a good match is slim, especially when there is a lot of variability in the weights. Categories of variables can then be collapsed to increase the swapping cell size so that there is a larger pool of donors. In the software, the more swapping variables there are, the more cells, the smaller number of swapping partners, and the less chance of a good match. Reducing the number of swapping variables can also increase the chance of a better target/partner match.

5.2 Practical Approaches for Reducing the Impact from Missing Data

NCES requires that every record should have a chance of being swapped. However, if imputed values exist among the swapping variables, they can be treated as masked (under certain conditions) and not eligible for swapping. The records having the imputed values within the swapping variables can be removed from the swapping process.

In general, missing data are coded as the smallest values or the largest values on a data file. Therefore, when defining cells, missing values will nearly always be sorted next to the same data value. For instance, missing values of “.” will always be sorted at the top of the list, next to values of 0. As a result, missing values will always be swapped with data values of the same magnitude. This will impact the analysis to the extent that cases with missing values are different in the key outcome variable from cases with values of 0.

There is an option in the software to temporarily impute values for the sole purpose of forming the swapping cells; therefore, the missing values will not always be adjacent to values of 0, but rather spread across other non-missing values. Once a swapping partner is found, then the original values are swapped, not the temporary imputed values.

6. Summary

Data swapping theory has been developed over recent decades. Dalenius and Reiss (1978) introduce the concept of protecting data through swapping procedures. Reiss, Post, and Dalenius (1982) extend data swapping to continuous variables. Reiss (1984) provides discussions on practical issues as well as theoretical discussions relating to the preservation of first- and second-order statistics. A summary of these and other theoretical discussions on data swapping is provided in Fienberg and McIntyre (2004).

While there are good theoretical discussions of data swapping in the literature, the focus of this paper is on practical issues, specifically when applying the swapping methodology in the context of sample survey data. The approaches presented here are the result of practical applications using NCES data and the need to reduce disclosure risk while maintaining data quality.

From our experience, swapping tasks need careful attention and the success of the process is in the hands of the operator of the software. We found that due to the complexities in questionnaires (such as skip patterns), different variable types and distributions, survey weights, and swapping impact, that it is important to develop a practical and adaptive interactive system that can be used to run and test multiple swapping strategies.

7. Acknowledgments

The authors wish to acknowledge the reviewers of this document, including Neil Russell (the IES DRB chair), and Westat reviewer, David Morganstein.

References

Dalenius, T. and Reiss, S. (1978). Data-swapping: A technique for disclosure control (extended abstract). *American Statistical Association, Proceedings of the Section on Survey Research Methods*, Washington, DC, 191-194.

Fienberg, S. and McIntyre, J. (2004). Data Swapping: Variations on a Theme by Dalenius and Reiss. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of Lecture Notes in Computer Science, pages 14-29, Heidelberg.

Kaufman, S., Seastrom, M. and Roey, S. (2005). Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? *American Statistical Association, Proceedings of the Section on Survey Research Methods*, Washington, DC, 1218-1225.

Le Cam, L. and Yang, G.L. (1990). *Asymptotics in Statistics*. Springer-Verlag. New York.

Reiss, S. (1984). Practical data-swapping: The first steps. *ACM Transactions on Database Systems*, 9, 20-37.

Reiss, S., Post, M. and Dalenius, T. (1982). Non-reversible privacy transformations. In *Proceedings of the ACM Symposium on Principles of Database Systems*, March 29-31, 1982, Los Angeles, California, pages 139-146.