Evaluating the Impact of Data Swapping Using Global Utility Measures

Sylvia Dohrmann[1], Tom Krenzke[1], Shep Roey[1], and J. Neil Russell[2]

[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]National Center for Education Statistics, 1990 K Street NW, Washington, DC 20006

Federal and state agencies continue to balance the need to release usable survey data while protecting the confidentiality of the individuals or institutions included in these surveys. The confidentiality standards of the Institute of Education Sciences (IES) require (1) the identification and masking of sensitive variables and records and (2) introducing an additional measure of uncertainty with random swapping through the use of the IES *DataSwap* software. *DataSwap* includes three sets of global utility measures to help evaluate the impact of swapping on the weighted distribution of the swapped variables, selected pairwise associations, and coefficients in regression models. The objective of having these measures built into the software is to help users select the best swapping result from several randomly replicated runs generated by using different swapping targets and parameters.

In addition to a description of the swapping methodology, in this paper we present each of the global utility measures, explain their characteristics, and their interpretation. We will further demonstrate how the measures may be used to evaluate various swapping scenarios and display the variability among replicated runs. We address how swapping scenarios may be modified to reduce the swapping impact, and present strategies for determining the best swapped dataset for a study.

1.0 Software

The basic idea of data swapping is to transform a database by interchanging, or "swapping" values of one or more variables between records. The benefit of using swapping as a statistical perturbation technique is that it maintains the unweighted univariate distribution of each variable while still introducing uncertainty about the identity of records since the data intruder does not know which variables or records contain swapped information. Inasmuch as confidentiality in any data file cannot be absolutely assured, the randomized swapping allows the agency to contend that no one can be certain if an individual unit has been identified.

The former chair of the IES Disclosure Review Board (DRB), Steve Kaufman, developed the underlying methodology for IES data swapping and designed a series of SAS macros automating this methodology in *DataSwap* which is described by Kaufman, Seastrom, and Roey (2005). *DataSwap* is the only DRB approved swapping software package and has been commended by the National Center for Education Statistics/National Institute of Statistical Sciences Data Confidentiality Task Force as a powerful software tool.[3]

The standardized, parameter driven software facilitates the review and revision of the process in order to help the user limit and reduce the swapping impact on data utility. *DataSwap* also allows for a consistency in swapping methodology and systematic interpretation of swapping results across studies.

1.1 Method of Sampling Records for Swapping

IES uses a controlled random swapping approach on restricted-use and public-use microdata files. "Controlled," in this sense, means two things. First, the user identifies the data swapping variables and parameters and selects the target records (the records whose values will be swapped). Target records are selected systematically with probabilities proportionate to a measure of size. The user may specify a stratified design and/or use variables to sort the data before selection. The sampling rate is predetermined by the user and entered as a parameter. Records with a high risk of disclosure may be given a higher selection probability for swapping.

---

[3] NCES/NISS Data Confidentiality Task Force final report dated February 5, 2008.

Second, the swapping methodology is designed to find a swapping partner that limits data distortion. Swapping partners are selected for each target within swapping cells. The swapping cells are formed by cross-classifying key categorical variables (i.e., identifiers such as age and education attainment categories), henceforth referred to as swapping variables. The search for swapping partners proceeds as follows. Consider a selected target record in a given cell. Two potential swapping partners for the target record are initially selected, one from each neighboring (adjacent) cell where each record has the closest sampling weight to the target record.

The search process continues by comparing the swapping bias and the potential swapping partner. The swapping bias is a function of the survey sampling weights (w) and a variable selected by the user for this purpose (x):

$$(w_1x_2 + w_2x_1) - (w_1x_1 + w_2x_2)$$

where,

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

The record that results in the smallest swapping bias is chosen as the swapping partner. The swapping of data occurs as the values of the designated swapping variables are switched between all targets and their respective partners (i.e., the values of the target's variables identified for swapping are assigned to the respective partner and the partner's values are assigned to the respective target case). Other variables can be linked to the swapping variables so that, as the value of a particular swapping variable changes, the linked variable(s) will also be changed. For example, if age is categorical and a swapping variable, the detailed age variable should be specified as a linked variable so that the two variables remain consistent on the final file. The software supports two methods for how the variables values are swapped between the selected records targeted for swapping.

1.2 Standard Method

With the standard approach, if a user wants to swap values of occupation, for example, among cases with the same age, sex, and race, then the swapping cells will be formed by the cross-classification of those three variables with occupation as the last (right-most) variable. For most cases, a swapping partner will be chosen such that the values of the first three variables are as similar as possible, if not identical, but with values of occupation that are different (to ensure that swapping takes place) and with a minimum calculated swapping bias. In some cases a record resulting in the lowest bias is one that has the same value of occupation, but differing levels of one or more of the other variables. In general the standard method tends to cause disproportionate swaps to the right-most variable which is useful when the file contains only one or two highly identifying variables. If the variables used for swapping are all important and useful for analysis, it is preferable to lessen the impact of change on an individual variable.

1.3 Balanced Method

As an alternative, the user may vary the order in which the variables are cross-classified to form the swapping cells across all the cases. When specified, the records in the input file are randomly allocated to groups such that each swapping variable is used as the right-most variable an equal number of times. In the example above, swapping cells using this balanced method will still be formed as a cross-classification of the four variables; however, each of the four variables will be the right-most variable in the cross-classification an equal number of times (with the other variables ordered in a random fashion). Using this method will result in a more balanced distribution of swapping across the swapping variables since more age, race, and sex values will be swapped rather than having values of occupation swapped most often as in the former example.

The balanced method provides data changes that are equitably distributed in expectation across the set of swapping variables on the data file so that no individual variable is adversely affected by the swapping. In practice, this procedure will result in some deviation from an equitable distribution of swapping across the variables due to the random ordering of the variables. The standard and balanced methods are evaluated in terms of their impact on the data utility in section 3.

2.0 Global Data Utility Measures

When using statistical disclosure control approaches, there is a dual objective of reducing disclosure risk while maintaining the usefulness of the data. The more distortion introduced into a dataset via swapping or other method, the greater the reduction in disclosure risk and data utility. Conversely, the less distortion, the greater the increase in disclosure risk and data utility.

When using *DataSwap* and its data utility measures, it is assumed that the first objective -- disclosure risk reduction -- is satisfied through the swapping rate assigned by the DRB and the manner at which selection probabilities are assigned to records with higher risk. Since the data utility measures in *DataSwap* measure the level of distortion between the original data and the swapped data, they can be used to address the second objective -- maintaining the usefulness of the data.

Three sets of global data utility measures are discussed below. The term "global" implies that an overall value is used to summarize the impact of swapping on data utility. These are designed to help the user evaluate the impact of swapping on the following:

1. Tables,
2. Selected pairwise associations, and
3. Coefficients in regression models of the key output variables on the swapping variables.

For all measures, large values imply less data utility.

2.1 Global Data Utility Measures for Tables

Multiple calculations of Hellinger's Distance (HD) are used to determine the data utility of a swapped dataset in terms of the change in weight distributions across swapping variables. *DataSwap* calculates this measure for the weight distribution over the cross-tabulation of all the swapping variables and also separately for each individual swapping variable.

Gomatam et al (2005) give the HD formula as:

$$HD(\hat{N}_{orig}, \hat{N}_{swapped}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{c} (\sqrt{\hat{N}_{orig}(c)} - \sqrt{\hat{N}_{swapped}(c)})^2} \qquad (2\text{-}1)$$

where,

$\hat{N}_{orig}(c)$ = sum of weights for cell *c* on the original data; and,

$\hat{N}_{swapped}(c)$ = sum of weights for cell *c* on the swapped data.

The full application (i.e. including all swapping cells) emphasizes differences in small cells. If the cross-tabulation of the swapping variables results in only a few cases in some cells, the differences between the weighted sum in those cells before and after swapping dominates the calculation; therefore, the HD measure may underestimate the data utility. Given the impact of small cells, four possible applications of the HD measure have been implemented into *DataSwap* as follows:

HD1: All cells, across all variables. The weighted original and swapped data are tabulated by crossing all swapping variables. HD1 is then computed according to equation 2-1 above to provide a single value of global data utility.

HD2: Excluding small cells, across all variables. The weighted original and swapped data are tabulated by crossing all swapping variables. HD2 is then computed according to equation 2-1 using only those cells meeting a user-specified minimum size requirement.

HD3: All cells, for each individual swapping variable. The weighted original and swapped data are tabulated for each swapping variable separately. HD3 is then computed for each swapping variable according to equation 2-1 to provide one value of data utility for each variable.

HD4: Excluding small cells, for each individual swapping variable. The weighted original and swapped data are tabulated for each swapping variable separately. HD4 is then computed for each swapping variable using only those cells meeting a user-specified minimum size requirement to provide one value of data utility for each variable.

If a swapping scenario has a large number of cells relative to the sample size, then it may be more likely that there are cells with sizes below the user's tolerance level. In that event, the restricted HD measure (excluding small cells), may be based on relatively few cells, and would be less stable.

2.2 Global Data Utility Measures for Pairwise Associations

While the HD measures focus on the change in the weight distributions between the original and swapped datasets among the swapping variables only, the measures based on the pairwise associations may be used to evaluate how swapping influences the relationship between these and other important key variables specified by the user. The three data utility measures of pairwise association in *DataSwap* are based on Pearson Product correlations, the Pearson contingency coefficient and Cramer's V (as given in Gomatam et al (2005)).

Measure Based on the Pearson Product Correlation. To obtain a global utility measure, we consider the average deviations relative to the before swapping standard error (using the formula from normal theory presented in Wolter (1985)).

$$R\_ASED(Y_{original}, Y_{swapped}) = \frac{\sum_{i \neq j} | r_{w,original}(Y_i, Y_j) - rw_{w,swapped}(Y_i, Y_j) | / SE(r_{w,original}(Y_i, Y_j))}{n'}$$

where,

$r_{w,<dataset>}$ = weighted Pearson product correlation as computed for the original or swapped datasets;

n' = the total number of pairwise comparisons with pairwise comparison differences before and after swapping greater than 0; and,

$$SE(r_{w,original}) = \frac{(1 - r_{w,original})^2}{\sqrt{n}}$$

Where n = the number of non-missing values used in the correlation.

Since the Pearson correlation is appropriate only when both variables lie on an ordinal scale, any nominal variables are converted to dummy variables. The impact on the resulting data utility measure is that more correlations are computed when there are more nominal variables. Further, differences in correlations for individual levels from the same nominal variable may influence the R_ASED measure. More discussion is given in section 3.

Measure based on the Pearson Contingency Coefficient. Let the Pearson contingency coefficient (C) between two variables be equal to:

$$C(Y_i, Y_j) = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

where,

$$\chi^2 = \sum_c \frac{(n_c - e_c)^2}{e_c};$$

n_c = actual weighted frequency in cell c; and,

e_c = expected weighted frequency in cell c.

The average absolute relative deviation (ARD) of the C measure is computed as follows:

$$C\_ARD(Y_{original}, Y_{swapped}) = \frac{\sum_{i \neq j} | C_{orig}(Y_i, Y_j) - C_{swapped}(Y_i, Y_j) | / C_{original}(Y_i, Y_j)}{n'}$$

where n' = the total number of pairwise C computations with differences before and after swapping greater than 0.

Measure based on Cramer's V. Let the Cramer's V statistic (*V*) between two variables be equal to:

$$V(Y_i, Y_j) = \sqrt{\frac{\chi^2 / n}{\min(k-1, l-1)}}$$

where,

k = number of categories for variable Y_i; and,

l = number of categories for variable Y_j.

The range is $0 \leq V \leq 1$. The Cramer's V is defined slightly differently for 2×2 tables in which case the range is $-1 \leq V \leq 1$. The computation is equal to:

$$V(Y_i, Y_j) = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 2}n_{\cdot 1}}}$$

The average absolute relative deviation (*ARD*) of the *V* measure is computed as follows:

$$V\_ARD(Y_{original}, Y_{swapped}) = \frac{\sum_{i \neq j} |V_{original}(Y_i, Y_j) - V_{swapped}(Y_i, Y_j) / V_{original}(Y_i, Y_j)|}{n'}$$

where n' = the total number of pairwise *V* computations with differences before and after swapping greater than 0.

The data utility measures based on the Pearson contingency coefficient and Cramer's V are both calculated as the average absolute deviation of the estimates between the original and swapped data, relative to the values of the correlations prior to swapping. Since these measures treat all variables as nominal, the variables are used in their original form (i.e., dummy variables are not needed for this computation of the Pearson contingency coefficient and Cramer's V).

2.3 Global Data Utility Measures Based on Regression Coefficients

Weighted regression models are produced with each key output variable (such as test score) as a dependent variable and all the swapping variables together as the independent variables. Similar to the measure based on the correlations, the data utility measure based on regression coefficients is calculated as the average absolute deviation between the weighted before and after swapping regression coefficients, relative to the standard errors of the coefficients prior to swapping.

$$ASED_i(\beta_{original}, \beta_{swapped}) = \frac{\sum_j |\beta_{original}(i, j) - \beta_{swapped}(i, j)| / SE(\beta_{original}(i, j))}{k_i}$$

where, $\beta(i, j)$ = beta coefficient for weighted regression model *i* and term *j*;

k_i = the total number of beta coefficients for weighted regression model i; and,

SE = standard error.

A global utility measure is computed as the unweighted average of the model-level measures:

$$ASED\_REG = \frac{\sum_i ASED_i(\beta_{original}, \beta_{swapped})}{m}$$

where *m* = number of models.

In addition to models using all swapping variables as independent variables, the user may also specify other models. However, the user is cautioned that when specifying a model to ascertain data utility, misspecification of the model will produce possible erroneous results. An extreme example of this is a model consisting of variables for which no values were swapped. The models used to ascertain data utility should be meaningful and involve some variables with swapped values.

Also similar to the measure based on the Pearson correlation, the regression coefficient utility measure makes use of the dummy variables created for any nominal swapping variables since linear regression is appropriate only when all variables lie on an ordinal scale. Again, the impact on the resulting data utility measure is that more coefficients will be included in a calculation for a swapping scenario with more nominal variables. Further, differences in coefficients for individual levels associated with a nominal variable will have more influence on the ASED_REG measure than if it were an ordinal variable. More discussion is given in section 3.

3.0 Evaluating the Characteristics of Utility Measures

All the utility measures are affected by changes in the number of swapping cells, the sampling rate, the swapping method (standard or balanced), the random sample of targets, and may be affected differently depending on the size of the dataset under consideration. Some of the measures are further affected by the number and types of variables used in their calculations. In order to understand the impact of the swapping parameters on the utility measures, we conducted systematic variations of the swapping parameters on four related, but varying sized datasets. The sections below describe the methods and results of this evaluation.

3.1 Evaluation Input Data

The household and prison public-use data files for the 2003 National Assessment of Adult Literacy (NAAL), sponsored by the National Center for Education Statistics (NCES), were used for the evaluation. NAAL is a nationally representative assessment of English literacy among American adults ages 16 and older.[4] Even though these data already experienced swapping before release, as per the IES standards, the files were treated as unperturbed data in need of swapping for the purpose of this evaluation.

Four datasets were created from the NAAL files: three household datasets including all the household records (18,102 observations), the southern region only (8,153 observations), and the northeastern region only (3,648 observations). The data from the prison component were used as the fourth dataset (1,156 observations). Even though two of the files are the subset of the larger file, we assume the files are independent of each other.

For each file, twelve scenarios were created with varying numbers of swapping variables (ranging from two to five) and potential swapping cells (ranging from 12 to 960). For the three household datasets, the same swapping variables were used in all scenarios. However, given that some variables were not available on the prison data, alternate scenarios were created where necessary. Table 3-1 shows the twelve scenarios used as well as the number of swapping variables, potential number of swapping cells (i.e., the product of the number of categories in each of the swapping variables), and the total number of nominal categories (the total number of levels contained in non-ordinal swapping variables).

Table 3-1. Scenarios used in evaluation

Scenario	Number of variables	Number of cells (household/prison)	Number of nominal categories (household/prison)
1	4	192	6
2	4	960 / 224	37 / 6
3	4	96 / 64	9 / 8
4	5	192	11
5	5	576	9
6	5	384	8
7	3	32	6
8	3	144 / 96	4
9	2	72 / 148	2
10	2	180 / 42	30 / 0
11	2	12	2
12	2	36 / 24	0

[4] http://nces.ed.gov/naal/datafiles.asp

For each of the 12 scenarios, the four datasets were swapped using two possible swapping rates (.03 and .15) and using each of the two swapping methods (standard and balanced). Each scenario/dataset/swapping rate/swapping method were replicated 10 times using different swapping targets (randomly selected), so that the replicate variability of the utility measures could be incorporated into the evaluation. In total, 1,920 *DataSwap* runs were conducted. In all the *DataSwap* runs, the NAAL participants' proficiency from the public-use microdata file on the prose, document, and quantitative scales were specified as key outcome variables. The impact of swapping on these variables will be exhibited in the global data utility measures for multivariate associations (based on pairwise comparisons and regression coefficients).

For the purpose of this evaluation, only the global data utility measures, in other words, not the Hellinger's Distance (HD) measures on single variables, were considered for this analysis. The utility measure based on Cramer's V was also not considered since the algorithm for calculating this measure was not fully implemented in *DataSwap* at the time of the evaluation.

3.2 Effect of swapping on utility

In order to examine how each of the swapping parameters affects the utility measures, multivariate regression models were fit with each of the utility measures as the dependent variable and the swapping parameters, along with their two-way interactions, as the independent variables. Final models were determined using backward elimination. Given the scale of the utility measures for multivariate associations, these measures were modeled using a log transformation.

The following parameters were considered as independent variables in the regression models:

- Swapping rate;
- Swapping method;
- Number of possible swapping cells;
- Size of dataset;
- Number of swapping variables (considered only for the data utility measures for pairwise and multivariate associations); and,
- Number of nominal categories among swapping variables (considered only for the data utility measure for pairwise association based on the Pearson Product Correlation and the data utility measure based on regression coefficients)

The number of possible swapping cells was used in the models rather than the actual number of swapping cells since the latter is highly correlated with sample size. Also, considering the number of possible cells and their affect on utility is more useful for planning purposes.

The number of swapping variables was considered for the data utility measures for multivariate associations since the calculations of the measures depend on these values. Since dummy variables are produced for each nominal variable level and used in R_ASED and ASED_REG, the number of nominal levels was included in these models.

Results. Not surprisingly, several interactions were significant in each of the models; if an interaction term was highly significant, the corresponding all main effects were left in the model even if they appeared not to be significant. For all the models, the interaction between the rate and dataset size was significant as was the number of possible swapping cells and the dataset size. The independent variables considered in each of the models, the significant beta coefficients, and model R-square values are presented in tables A-1 and A-3 of the appendix.

The effects of all swapping scenario characteristics on the data utility measures, as found by reviewing the regression output, are illustrated in Table 3-2. For all the measures, as the swapping rate increases, the data utility decreases for most samples. In many cases, the magnitude of this effect is impacted by other swapping variables.

In addition to rate increases, increasing the number of swapping cells will result in loss of data utility in tables produced from the swapped data for most samples. Small samples, with already high swapping rates and/or large numbers of swapping cells relative to the sample size, may see the reverse effects. However, this is likely an artifact of the measure calculation itself. If the sample size of some cells is below the specified tolerance (set to the default of 45 in this evaluation), it is likely that the small sizes are dominating the HD1 calculations and thus overestimating the data utility. However, if there are many small cells, once they are removed for the HD2 calculation, the measure may be based on too few cells to be stable. In the extreme

case, if all the cells with the swapped values are below the tolerance level, the HD2 calculation will be zero, indicating no loss of data utility.

Increasing the number of swapping cells, but keeping the number of variables constant can increase the utility of pairwise associations, but not so for the R_ASED measure if the swapping variables have nominal levels. If there are nominal categories in the swapping cell definitions, the overall utility as measured by R_ASED may be increased by increasing the number of swapping variables without increasing the overall number of nominal categories in the process. Since each nominal category has the same weight as entire ordinal variables in the R_ASED and ASED_REG calculations, it may be that the overall impact of the nominal variables are outweighing the impact of the other variables.

In most cases, the standard swapping method results in more data utility overall, even though swapping disproportionately impacts an individual variable with this method. For smaller datasets, the balanced method may produce more data utility in terms of the measures based on tables.

Table 3-2 Results of the modeling of the data utility measures:
Impact of actions indicated in the columns on the measures indicated in the rows.

	Effect on data utility as a result of…				
	Increasing swap rate	**Increasing number of swapping cells**	**Increasing number of swapping variables**	**Increasing number of nominal categories**	**Using the standard swapping method over the balanced**
HD1 Utility for tables (all cells)	⇧ Small samples if number of cells is large relative to sample size* ⇩ Most samples	⇧ Smaller samples with larger sampling rates** ⇩ Most samples			⇧ Most samples ⇩ Number of cells is large relative to sample size*
HD2 Utility for tables (excluding small cells)	⇧ Small samples if number of cells is large relative to sample size* ⇩ Most samples	⇧ Smaller samples with standard method ⇩ All other samples			⇧ Most samples ⇩ Number of cells is large relative to sample size*
C_ARD Utility for pairwise associations based on Cramer's V	⇩	⇧ Most samples ⇩ Larger samples with many swapping variables and/or high rates.	⇩		⇧
R_ASED Utility for pairwise associations based on Pearson's Product Correlation	⇩	⇧ No nominal categories ⇩ Large samples with many swapping variables (with or without nominal categories) or high rate	⇧ Small samples with many nominal categories ⇩ Most samples	⇩	⇕ Dependent on all other parameters
ASED_REG Utility based on regression coefficients	⇩	⇧ No nominal categories ⇩ Nominal categories	⇧ High swapping rate and/or large number of nominal categories ⇩ No nominal categories	⇕ Dependent on all other parameters	⇧ Small samples with no nominal categories; larger samples with many swapping cells ⇩ Larger samples and scenarios fewer swapping cells

⇧ Data utility improves (indicated by a decrease in the row measure) as a result of the action described in the column header.
⇩ Data utility is reduced (indicated by an increase in the row measure) as a result of the action described in the column header.
* In our data, there was an increase in utility when the ratio of cells to sample size was ~10%.
** In our data, there was an increase in utility when the rate was > 4% for the standard method, and > 10% for the balanced.

3.3 Replicate variability of the utility measures

The most fundamental way to impact the data utility of a swapped dataset is to simply change the records being swapped. Since the records are selected for swapping in a controlled, but random, manner, this may be done easily in *DataSwap* by simply re-running the swapping process with a different random seed. However the variability of the utility measures can be quite large across different target sample selections; this variability may make it difficult to determine which swapping parameters are optimal for a particular dataset since multiple runs may yield seemingly contradictory messages regarding the resulting data utility.

Consider the scenario using the balanced method with a swapping rate of 3%, five swapping variables, 384 possible swapping cells and 8,153 observations. The values of the five data utility measures across the 10 replicate runs are shown in the figure 1 below.

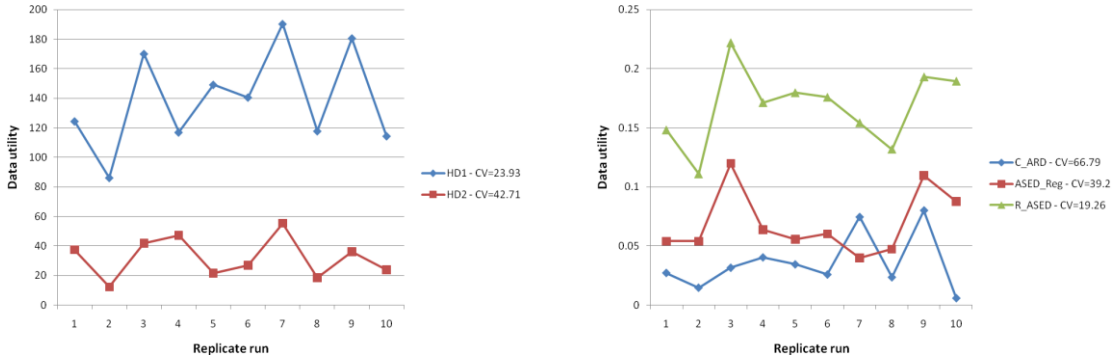

Figure 1 –Utility measures over 10 replicates for the scenario based on the balanced method with a swapping rate of 3%, five swapping variables, 384 possible swapping cells and 8,153 observations

Replicate 7 has the smallest value of ASED_REG, but the second highest value of C_ARD. The HD measures for this replicate are also the highest of all the replicates. The coefficients of variation of the data utility measures over the replicate runs displayed in figure 1 are at about the median levels across all the scenarios evaluated. The average coefficients of variation across all 192 scenarios evaluated and the ranges of those averages across all the scenarios are shown in table 3-3 below.

Table 3-3. Distributions of the utility measure coefficients of variation (CV*) across the 10 replicate runs for each scenario/dataset/swapping rate/swapping method combination.

Utility Measure	Average CV*	Minimum CV*	Maximum CV*
HD1	34.1938	5.2215	186.3426
HD2	42.4313	5.2550	182.4358
R_ASED	24.9027	4.3132	68.3395
C_ARD	60.9735	9.6106	178.3409
ASED_REG	29.0998	6.0285	91.3749

* For each scenario, the CV was calculated as the standard deviation of the utility measures across the 10 replicate runs divided by the average utility across the 10 runs.

From the table above, it is clear that some swapping scenarios may result in more replicate variability than others and may require more replication to find an acceptable swapped dataset. The scenario using the balanced method with a swapping rate of 15%, five swapping variables, 576 possible swapping cells and 8,153 observations resulted in the smallest variability in the HD1 measure, while the scenario using the standard method with a swapping rate of 3%, two swapping variables, 12 possible swapping cells and 1,156 observations resulted in the largest variability. Since some scenarios produce more variability than others, it is important that users are aware of when more replicates might be necessary to determine an acceptable swapped dataset.

In order to determine which swapping characteristics yielded the most variability, and thus might require more replication, multivariate regression models were fit with the log of the coefficients of variation for the utility measure as the dependent variable and the swapping characteristics, along with all two-way interactions, as the independent variables. Final models were determined using backward elimination. The same characteristics used in the modeling each of the respective utility measures were used to model their coefficients of variation.

Table 3-4. Results of the modeling of the coefficients of variation of the data utility measures: Impact of actions indicated in the columns on the measures indicated in the rows.

	Effect on data utility variability as a result of…				
	Increasing swap rate	**Increasing number of swapping cells**	**Increasing number of swapping variables**	**Increasing number of nominal categories**	**Using the standard swapping method over the balanced**
HD1 Utility for tables (all cells)	⇩	⇩			
HD2 Utility for tables (excluding small cells)	⇩	⇧ Smaller samples ⇩ Larger samples			⇧ Lower rates* ⇩ Higher rates*
C_ARD Utility for pairwise associations based on Cramer's V	⇩		⇩		
R_ASED Utility for pairwise associations based on Pearson's Product Correlation	⇩	⇩		⇧	⇧ Most samples ⇩ Larger samples with relatively higher rates**
ASED_REG Utility based on regression coefficients	⇩	⇧ Fewer swapping variables† ⇩ More swapping variables†	⇧ Fewer swapping cells*** ⇩ More swapping cells***	⇩	⇧ If no nominal categories ⇩ Large sample with high rate and/or many nominal categories

⇧Data utility variation increases (indicated by an increase in the CV of the row measure) as a result of the action described in the column header.

⇩Data utility variation decreases (indicated by an increase in the CV of the row measure) as a result of the action described in the column header.

* Our data indicated that the variation increased when the rate was 10% or larger.

** Our data indicated that the variation increased when the rate was 6% or larger.

*** In our data the threshold was 108 swapping cells.

† In our data the threshold was 4 swapping variables.

Results. The variables considered in each of the models, the significant beta coefficients, and model R-square values are presented in tables A-2 and A-4 of the appendix. The effects of all swapping scenario characteristics on the variability of the data utility measures, as found by reviewing the regression output, are illustrated in Table 3-4.

As with the results of the utility measure modeling, all the utility measures experience less variability as the rate increases. This is not surprising since by increasing the swap rate, the number of possible unique target samples decreases. Hence, scenarios with very low swapping rates would likely require more replicate runs to improve the data utility of the swapped dataset.

The variability of the measures will increase in scenarios for small samples as the number of swapping cells increase (HD2) and scenarios with few swapping variables (ASED_REG). Similarly, the more variables in a scenario, the higher the variability of the ASED_REG measure in scenarios with few swapping cells. The R_ASED measure will be more variable in scenarios many nominal categories in the swapping variables. Under any of these conditions, again, more replicate runs will be necessary in order to improve the data utility.

The standard method seems to result in utility measures with more variability across runs if there are no, or very few, nominal categories in the swapping cells (ASED_REG) or lower swapping rates (HD2).

4.0 Discussion and Conclusion

The data utility measures in *DataSwap* are designed to aid users in determining how the data perturbation has affected the overall usefulness of the resulting data files. Several tools are implemented in *DataSwap* to help measure the utility in terms of weighted frequencies and multivariate associations. For each of these measures, the baseline, or the value indicating the most utility, is zero. However, if any swapping of data values has occurred, this value is not obtainable. Instead the ideal value is the lowest possible that may be obtained while still introducing uncertainty through swapping. Even this value may only be realized after viewing multiple replicate runs and understanding the limitation on utility as a result of the specified swapping parameters.

Swapping scenarios with lower swapping rates result in greater data utility. Our evaluation further demonstrated that using fewer variables with many cells most often results in higher data utility for the multivariate associations provided that the resulting swapping cells are of acceptable size. Doing so results in more swapping cells, better swapping partners in adjacent cells, and may result in a lower bias. For example, if detailed age is swapped along with a six-level collapsed age (instead of a three-level variable), this will result in detailed age values closer together being swapped. However, if the variables in question are nominal, the utility measured by the ASED_REG and R_ASED measures may not be clearly stating the utility if there are many nominal categories in the swapping cells. The data utility of multivariate associations may be best measured in these instances by the C_ARD variable in scenarios with many nominal variables.

In contrast, using more detailed swapping variables (that is, increasing the number of cells) may produce less utility in tables of the swapping variables, as measured by the HD1 and HD2 variables. Less detail in swapping variables results in fewer, but larger swapping cells and, thus, more potential swapping partners for a selected target, which could in turn improve the data utility of tables. Users are cautioned against relying on these measures for tables, however, when the sample sizes in the swapping cells are very small.

The standard method proved to provide more data utility over the balanced in most scenarios. This may seem contradictory since this method concentrates the amount of data perturbation on one variable. However, since the number of variables affected by the swapping is limited when using the standard method, fewer multivariate associations and table cells are affected.

While not shown in this evaluation, the order of the variables forming the swapping cells will also impact the data utility. If a *DataSwap* run results in unacceptable data utility, a logical reordering of the variables may have a positive impact. It may also be the case that unacceptable utility is resulting from particular records not having acceptable swapping partners. In this event the problematic cases should be given a lower chance of selection as swapping targets or a different random seed could be given to select different target records.

It is imperative that with each change in parameter specification, utility measures from several replicate *DataSwap* runs (each with different target samples) be compared with each other. Our evaluation found that scenarios involving small datasets, low swapping rates, few swapping cells or variables, and using the standard method may see great improvement from several replicate runs. Our suggested strategy is to select two or three of the best replicated runs with the smallest multivariate data utility measures (those based on pairwise and regression associations). Then subsequently select the replicate run resulting in

the best data utility among the HD measures. However, since the HD measures are sensitive to small cells, if there are many small swapping cells it is best to choose the replicate based solely on the multivariate data utility measures (those related to pairwise and regression coefficients).

DataSwap employs several data utility measures to assist researchers in their search for a file acceptable for public use since, as the results from the evaluation have shown, relying on only one measure may not fully represent the utility of the data. Some measures may not be fully descriptive of the data if there are many small cells, while others may overstate the impact of swapping if there are many nominal levels. Researchers should not let the value of any one of these measures replace good judgment. The swapping rate, and all other parameters used to instruct the data swapping, must be considered in tandem with the replicated measures to fully understand the impact of data swapping, on the final data file.

References

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? In Proceedings of the American Statistical Association Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Gomatam, S., Karr, A., and Sanil, P. (2005). Data swapping as a decision problem. *Journal of Official Statistics* **21(4)** 635-656.

Wolter, , K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Appendix

Table A-1. Data utility measure evaluation regression model results for tables: Beta coefficients and model R-square

Variables/interactions	HD1	HD2
Model R-Square	0.8251	0.4858
Intercept	11.9307	-2.2368
Number of possible swapping cells	-0.0562	0.0169
Swapping rate	-160.9149	6.9084
Swapping method (Reference: Balanced)	10.9444	4.8847
Size of dataset (Reference: 1,156)		
Size:18,102	-21.7572	-7.9574
Size: 8,153	-14.8579	2.5009
Size: 3,648	1.3689	5.3424
Possible swapping cells × swapping rate	1.3672	n.s.
Possible swapping cells × swapping method	-0.0868	-0.0348
Possible swapping cells × size of dataset		
Size:18,102	0.2728	0.0313
Size: 8,153	0.3010	0.0201
Size: 3,648	0.2378	-0.0150
Rate × swapping method	n.s.	n.s.
Rate × size of dataset		
Size:18,102	978.2369	803.4187
Size: 8,153	747.4320	397.5493
Size: 3,648	704.1250	321.6811
Swapping method × size of dataset		
Size:18,102	-12.8278	-11.4233
Size: 8,153	-12.4992	-6.5276
Size: 3,648	-10.7728	0.6899

n.s. = not found significant in model

Table A-2. Data utility variability evaluation regression model results for tables: Beta coefficients and model R-square

Variables/interactions	log(CV of HD1)	log(CV of HD2)
Model R-Square	0.5798	0.5688
Intercept	1.8763	1.7943
Number of possible swapping cells	-0.0007	0.0005
Swapping rate	-2.7019	-1.3984
Swapping method (Reference: Balance)	n.s.	0.1693
Size of dataset (Reference: 1,156)		
Size:18,102	-0.1162	0.1631
Size:8,153	-0.0853	0.1646
Size:3,648	-0.0680	0.1567
Possible swapping cells × swapping rate	n.s.	n.s.
Possible swapping cells × swapping method	n.s.	n.s.
Possible swapping cells × size of dataset		
Size:18,102	n.s.	-0.0010
Size:8,153		-0.0007
Size:3,648		-0.0003
Swapping rate × swapping method	n.s.	-1.5251
Swapping rate × size of dataset	n.s.	n.s.
Swapping method × size of dataset	n.s.	n.s.

n.s. = not found significant in model

Table A-3. Data utility measure evaluation regression model results for multivariate associations: Beta coefficients and model R-square

Variables/interactions	log(R_ASED)	log(C_ARD)	log(ASED_REG)
Model R-Square	0.7900	0.6432	0.7811
Intercept	-1.0900	-3.8421	-1.8777
Number of possible swapping cells	-0.0078	-0.0003	-0.0037
Swapping rate	2.9528	3.3592	4.5020
Swapping method (Reference: Balanced)	-0.3210	-0.7136	-0.4144
Size of dataset (Reference: 1,156)			
Size:18,102	-0.1213	-0.6867	0.0963
Size:8,153	-0.1711	-0.3366	0.0061
Size:3,648	-0.0566	-0.1359	-0.0523
Number of swapping variables	0.1251	0.5356	0.1259
Number of nominal categories among swapping variables	0.0649		0.0813
Possible swapping cells × swapping rate	n.s.	-0.0025	0.0015
Possible swapping cells × swapping method	-0.0003	n.s	-0.0003
Possible swapping cells × size of dataset			
Size:18,102	-0.0004	0.0012	-0.0004
Size:8,153	-0.0003	0.0012	-0.0003
Size:3,648	-0.0003	0.0010	-0.0004
Possible swapping cells × number of swapping variables	0.0013	-0.0002	0.0006
Possible swapping cells × number of nominal categories	0.0001		0.0001
Swapping rate × swapping method	0.4396	1.5267	0.6488
Swapping rate × size of dataset			
Size:18,102	2.1578	2.7072	1.2911
Size:8,153	1.7625	1.4733	1.0060
Size:3,648	1.7184	1.0912	1.1394
Swapping method × size of dataset			
Size:18,102	0.1174		0.2691
Size:8,153	0.0978		0.2028
Size:3,648	0.1042		0.2264
Swapping rate × number of swapping variables	n.s	n.s	-0.3958
Swapping rate × number of nominal categories among swapping variables	-0.0333		-0.0571
Number of nominal categories among swapping variables × swapping method	0.0084		0.0066
Number of nominal categories among swapping variables × size of dataset			
Size:18,102	-0.0247	n.s	-0.0322
Size:8,153	-0.0243		-0.0333
Size:3,648	-0.0244		-0.0295
Number of swapping variables × swapping method	0.0413	0.0727	0.0567
Number of swapping variables × size of dataset			
Size:18,102	0.1358	n.s	0.0968
Size:8,153	0.1349		0.1026
Size:3,648	0.1001		0.0957
Number of swapping variables × Number of nominal categories among swapping variables	-0.0251		-0.0176

n.s = not found significant in model

■ = not considered for the model

Table A-4. Data utility variability evaluation regression model results for multivariate associations: Beta coefficients and model R-square

Variables/interactions	log(CV of R_ASED)	log(CV of C_ARD)	log(CV of ASED_REG)
Model R-Square	0.7232	0.4100	0.7328
Intercept	1.6456	2.2016	1.5938
Number of possible swapping cells	-0.0001	n.s.	0.0015
Swapping rate	-0.5770	-1.1686	-1.5954
Swapping method (Reference: Balanced)	0.2137	n.s.	0.1683
Size of dataset (Reference: 1,156)			
Size:18,102	0.1660	n.s.	-0.1784
Size:8,153	0.0924		-0.0581
Size:3,648	0.0366		-0.0633
Number of swapping variables	n.s.	-0.1028	0.0340
Number of nominal categories among swapping variables	-0.0297		-0.0111
Possible swapping cells × swapping rate	n.s.	n.s.	n.s.
Possible swapping cells × swapping method	n.s.	n.s.	n.s.
Possible swapping cells × number of swapping variables	n.s.	n.s.	-0.0003
Possible swapping cells × number of nominal categories	n.s.		n.s.
Swapping rate × swapping method	-0.8271	n.s.	-0.6184
Swapping rate × size of dataset			
Size:18,102	-1.1440	n.s.	n.s.
Size:8,153	-1.0397		
Size:3,648	-1.1645		
Swapping method × size of dataset			
Size:18,102	-0.1626	n.s.	-0.1379
Size:8,153	-0.0738		-0.1038
Size:3,648	-0.1153		-0.0649
Swapping rate × number of swapping variables	n.s.	n.s.	n.s.
Swapping rate × number of nominal categories among swapping variables	n.s.		n.s.
Number of nominal categories among swapping variables × swapping method	-0.0054		-0.0065
Number of nominal categories among swapping variables × size of dataset			
Size:18,102	0.0219		n.s.
Size:8,153	0.0244		
Size:3,648	0.0242		
Number of swapping variables × swapping method	n.s.	n.s.	n.s.
Number of swapping variables × size of dataset	n.s.	n.s.	n.s.
Number of swapping variables × number of nominal categories among swapping variables	n.s.		n.s.

n.s. = not found significant in model

■ = not considered for the model