Swapping Approach and Process

Approach

The basic steps of the NCES controlled random swapping approach are to 1) select the target records, 2) select the partners, 3) swap the data for specified variables, and 4) generate output reports. With regards to selecting records for swapping, different random sampling approaches (simple random sampling, stratified, probability proportionate to size) may be used. Swapping partners are assigned based on a distance (or bias) measure. IES standard practice is to process the swapping five to seven times and review the global utility measures to recommend one of the runs.

For the controlled random swapping 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 should also understand the content of the data file as well as the purpose and focus of the study so user 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. For more details and illustrations of swapping cells and donor selection, please see the *DataSwap* user manual.

Targets

The first step is the selection of the records whose values will be swapped (i.e., target records). 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. Note that cases with a high risk of disclosure may be given a higher selection probability.

Selection of Swapping Partners

The second step in the algorithm is the selection of swapping partners (i.e., donors). The user must select variables to define the swapping cells. The swapping cells are formed by cross-classifying key categorical variables (i.e., identifiers such as age and education attainment categories). These variables are also the swapping variables – those variables whose values will be swapped. 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. That is, within each neighboring cell, the record with the closest sampling weight to the target record is selected as a potential swapping partner. The search process continues by comparing the swapping bias of the potential swapping partner. The record that results in the smallest swapping bias is chosen as the swapping partner. For example, if a user wants to swap values of occupation 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, but with values of occupation that are different (to ensure that swapping takes place) and with a minimum calculated swapping bias. It is also possible that 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.

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 alternative method will still be formed as a cross-classification of the three variables; however, each of the three 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 race and sex values will be swapped rather than having values of occupation swapped most often as in the former example.

The former (original) method tends to cause disproportionate swaps to the right-most variable. This is useful when the file contains only one or two highly identifying variables. The alternative (balanced) method provides a more balanced distribution of swapping among the 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. The balanced method provides data changes that are equitably distributed across the set of swapping variables on the data file so that no individual variable is adversely affected by the swapping.

If there are key variables that absolutely should not change value, but are critical in controlling the swapping partner search, then one or more variables may be specified as hard boundary variables. If hard boundaries are specified, only potential swapping partners within the same hard boundaries will be considered. If a swapping partner cannot be found within the hard boundaries, the algorithm will not proceed, and some adjustment to the hard boundary variables will be required. There may only be one neighboring cell if the target's cell comes at the beginning or end of the sort order within a hard boundary, or at the beginning or end of the file if no hard boundary is specified.

After the selection of the swapping partners, a check is made to determine that a swapping partner is used only once. If a partner is used more than once, then the partner is assigned to the target resulting in the smallest absolute bias. Ties are handled with a random selection. Next, a check is made to see if all target cases have final swapping partners. If any are found without final swapping partners, the partner search is repeated until all swaps have unique final swapping partners.

Swapping Values

In the third step, the data are swapped. By definition, the variables that are allowed to be swapped are referred to as "swapping variables." Therefore, hard boundary variables are not swapping variables. The swapping of data occurs as the values of the swapping variables are switched between each target 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). The user also has the option of specifying other variables to swap, apart from those used to form the swapping partner cells. These 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 a user wants to use age category as 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.

Reports

The last step is the generation of output reports for the user, the DRB chair, and the DRB members. The *DataSwap* software produces output that can be used to document and evaluate the swapping results. The output compares unweighted and weighted frequencies, means, correlations, and regression models before and after swapping. The output for the DRB can be pasted into a memorandum for the DRB.

Process

The data swapping process begins by conducting the risk assessment. The risk assessment will help inform the data swapping process. The results from statistical matching to external sources through probabilistic record linkage, and the risk assessment on combinations of indirect identifying variables through either the NCES *InitialRisk* software or the R program *SDCNway*, can help identify high risk data values. A two-stage sequential approach to swapping can be completed, where the first stage includes swapping records that has the highest risk and therefore are targeted with certainty, and the second stage is to give all records a chance of selection for random swapping.

In the first-stage, for the high risk data values that are found from therecord-linkage procedure, there are two options that are typically considered. One approach is to identify the records associated with high risk values and conduct deterministic swapping (select records to be swapped with certainty). That is, match the high risk records (among themselves if possible, or find other records to be swapping partners) and swap their data for specified variables. Another approach is to assign the high risk records to their own stratum with a selection rate of 1 for the *DataSwap* application. Note that the high risk records in the first stage can be identified by means other than probabilistic record linkage.

In the second stage, for all other data records, one can give a higher chance of selection to higher risk data values. For example, the *InitialRisk* risk measure assigns values 0 (low risk) to 4 (high risk) to records according to the number of table cell violations that the records were involved with. The risk measure can be used as a measure of size (or a function of the risk measure) to select the target records for the data swapping process. That is, in the first step of *DataSwap*, target records will be selected via probability proportionate to size sampling using the *InitialRisk* risk measure with higher chance of selection for higher risk data records.

Next, the values of the swapping variables will be exchanged between swapping targets and partners which reduces the disclosure risk, but at the same time incurs some bias. To reduce this bias, the swapping is set up to produce a higher probability for swapping for pairs with similar weights and/or close matches on specific characteristics; all records will be matched with each selected target case on characteristics that define swapping cells. The swapping cells are comprised of boundary variables (BOUNDARY) and swapping variables (SWAPVARS). The boundary variables are not allowed to be swapped. As discussed in the *DataSwap* manual, two potential swapping partners for the target record are initially selected, one from each neighboring (adjacent) cell. That is, within each neighboring cell, the record with the closest sampling weight to the target record is selected as a potential swapping partner. The search process continues by evaluating the swapping bias and the potential swapping partner. The record that results in the smallest swapping bias is chosen as the swapping partner.

To avoid the creation of logical inconsistencies, the swapping variables are selected carefully such that they are not structurally involved with several other variables. Therefore, a list of swapping variables will be specified to ensure that data consistency will be retained for other closely-related variables with different categorized versions. The parameter LINKSWAP will contain this list, and will ensure that when one of the SWAPVAR variables changes value during the swapping process, the linked variables will change as well.

The swapping is typically conducted five times (using different random seeds) before selecting the best swapped dataset based on the *DataSwap* impact measures, and according to the NCES Standards.

The weight variable that may be used in the process is the theoretical base weight, which is computed as the inverse of the overall selection probability of the person or entity. To help check the impact of the swapping, KEYOUT variables are chosen. Also, some categorical variables will also help to assess the impact of swapping through the KEYVARS parameter.