Tips for Applying DataSwap to your Project

Before implementing *DataSwap*, the user must consider several facets of the swapping process, including how the targets are to be selected, if particular records will be targeted for swapping, how the swapping cells are to be formed, and which variables will be swapped. Output reports to help the user evaluate the swapping impact can also be controlled through the software parameters.

For data swapping, there are several important points to make in terms of guidance toward reducing the impact on data utility, which include:

1. Conduct a preliminary evaluation. When choosing hard boundary variables and key variables for output review of the swapping impact, for each of the target variables for swapping, find the variables that have the highest correlation with each swapping variable. Create a list of all variables that should be considered for the process. If the sets of associated variables are much different between the target variables, then a separate process for each target variable may need to be considered.

2. Identify hard boundary variables. The hard boundary variables need to be related to the target variables.

3. Protect against an over-constrained swapping process. For the *DataSwap* algorithm, which forms cells based on the hard boundary and swapping variables, very sparse cells can occur due to too many hard boundary variables and swapping variables, and therefore the swapping partners may not be good matches. The following options may help: 1) reduce the number of variables, 2) switch a hard boundary variable to an evaluation variable (KEYOUT, KEYVARS), 3) switch a swapping variable to a LINKSWAP variable.

4. Collapse categories. Collapsing of categories will support the swapping process by reducing the number of swapping cells.

5. Select good and appropriate utility comparisons. Include variables used in swapping and not used in swapping.

6. Select an appropriate swapping rate. The swapping rate is assigned by the DRB chair. Feedback to the DRB chair on trial runs of the process may help with this assignment. Swapping can be used when the perturbation rate is low and a small number of records need to be targeted.

Some risk reducing factors to consider may include item nonresponse rates and imputation. Taking into account the item missingness may be helpful when assigning the swapping rates. Imputation for missing data may be considered a statistical confidentiality treatment because the reported value is not disseminated, and further risk reduction may occur when imputation flags are not disseminated.

Another note is to limit the amount of targeted (deterministic) swapping due to the potential bias it may introduce. This is similar to non-ignorable nonresponse, that is, the high risk values are fairly unique, which is why they have relatively high risk. Therefore, finding a good swapping partner with similar characteristics is less likely.

Accessing the *DataSwap* software

DataSwap is available at the GitHub here: https://github.com/Westat-Stats/DataSwap. The user manual can also be accessed at that location, which also provides an example.