GUI-software demonstration for MASSC application

D. Wang, D. H. Wilson, F. Yu
RTI International, NC 27709

Abstract

A Microsoft Windows based application has been developed for MASSC at RTI International. The methodology of the development was based on the idea that the MASSC statistical limitation process can be run on standard procedures and processes. By focusing on that approach, the backend MASSC engine was built on the standard SAS procedures and customized C programming based SAS callable packages. A GUI application was built to communicate between the MASSC users and the backend MASSC engine. The GUI software is very user friendly. The carefully designed user interface saves the user's interaction time, reduces level of confusion, and reduces the possibility of making errors. The GUI is also very flexible in terms of how each
step is executed, how the outcome for each step is examined, and the accessibility to the previous steps from the
current step. Some sample screens will be presented to show how the GUI works. A treated sample data set
result will also be presented with the GUI report in HTML format while the Graphics will be in PDF format.

1. Introduction

MASSC Graphical User Interface(GUI) software is a user friendly software developed on SAS software, customized C SAS callable procedures, and Visual Basic user interfaces. From the user stand point, the lower the system requirement (including hardware and software) the better the acceptance. For better system
execution, a Microsoft Windows system is required with 512 MB memory and 800MHz or above CPU. For
system execution software, the system requires SAS with SUDAAN embedded, Microsoft studio, and the text
editor (normally the Notepad from Windows).

The MASSC GUI takes SAS datasets as the data source and text files defining parameters, constraints, and other variables as execution flow control resources. The text files are usually generated directly from the GUI software. The communication between the users and the MASSC GUI software is managed by these text files. The text files are generated at each step when the users interface with the system, and the users have the ability to edit/modify the text files before running each step. The text files contain either standard SAS statements, variable recoding logics or constraint definitions.

The MASSC GUI produces text files for the purpose of executing the treatment steps, it also produces the treated datasets (in SAS dataset format) and the reports. The reports are in HTML report and PDF graphic formats. Each report links to the particular step or sub-step. The original dataset, the intermediate datasets which are produced during program execution and the final treated dataset are the sources of the reports and graphics. Some output datasets of certain steps can be treated as the input datasets of the following step(s). The datasets can be overwritten if one

chooses to re-run some of the treatment steps. This logic holds true for the text files.

In the following sections, we will break down the MASSC process function by function and show how the GUI
software works. A series of concepts and terms will also be introduced to demonstrate the logic flow of
the MASSC procedure and the
efficiency/effectiveness of the program. Finally, a series of screen shots,
tables and graphics will be presented to help people better understand MASSC.

2. GUI functions and user input information

MASSC GUI provides the screens for all processing needs. From IV(identifying variable) and SV(sensitive/secret variable) inputs to customized user variable recoding and constraint building, from treatment step execution to the carefully designed reports and graphics viewing, all screens reflect the philosophy of the MASSC GUI development in terms ease of use.

The MASSC GUI screens have two main categories: treatment step execution screens and recoding/defining
screens. For the treatment step execution screens like Micro-Agglomeration (Figure 2.1), Substitution Substrata
(Figure 2.2), Subsampling Substrata (Figure 2.3), and Weight Calibration (Figure 2.4), users will be able to
supply the parameters, call for the recoding screen and constraint definition screen, execute the treatment step,
and view the result reports. The screens are visible from the MASSC main control screen and are accessed
sequentially as the MASSC process is carried out. For instance, if the Subsampling step is incomplete, the users
will not be able to go to the next steps like Weight Calibration or Delta Calculation. Nevertheless, if the users
complete the MASSC process up to the Subsampling step, the screens for the previous steps will remain available. Users will likely repeat some of the steps to achieve acceptable results.

The recoding/definition screens will help the users to manage the data re-categorization and the constraint definitions. The recoding and definition screens were designed in such a way that the users can be guided throughout the process to either follow the standard method of recoding or try out the freestyle recoding/defining by typing the text in the provided text boxes. No matter what the user's choices are, the recoding/defining results will go into the same file for a given treatment step, and the users are always granted a last chance to modify the content. The ways of inputting or examining the parameters/variables/constraints are presented as follows.

IV stands for identifying variable. A set of IVs needs to be identified at the beginning of the MASSC treatment.
The user will pick the IVs from the list of variables that come out of the original dataset. For example, the IVs could be age, gender, and job status which could be known by intruders for a particular dataset.

SV stands for sensitive/secret variable. Similarly, a set of sensitive variables needs to be identified. For example, in a substance abuse survey , the SVs could be past month cigarette, alcohol, marijuana, and cocaine use. Note that a variable may not be deemed sensitive by the subject but nevertheless may be secret from the intruder (Singh, Yu, 2004).

Constraint variables are typically defined at each treatment step. Different constraints are defined for substitution, subsampling, and calibration. Constraints are defined via a specifically designed screen, the screen gives the users a flexible way to handle different scenarios.

α for controlling the information loss for substitution will be examined and carefully determined. The goal is to achieve the best treatment result when perturbing the records and control the information loss at the same time by assigning the appropriate α. From the provided screen a user can easily supply Alpha and adjust it when unsatisfied with the execution result.

Similarly, β for controlling the information loss for subsampling will be carefully determined.

δ's for uniques, doubles, triples, and others with respect to all core and noncore IVs will be computed to check for the disclosure risk of the final treated dataset. The GUI report will provide detailed information for all δ's.

Users can decide whether to go forward to continue treating the data or to go backward to repeat some steps of the process by adjusting input values.

3. User intervention for diagnostic checks

Before running MASSC treatment procedures, users need to determine if the categorization is sufficient for IVs and SVs. This step is critical since insufficient categorization will yield insufficient number of uniques and non-uniques. Such insufficiency will directly impact the treatment result. MASSC GUI provides a special recoding screen to help the users handling variable categorization issues.

Substitution and Subsampling rates will be determined by properly adjusting the Alpha and Beta values as well as properly clustering some non-core IVs. During these treatment procedures, the possibility of increasing or decreasing the number of SVs exists in order to achieve the best Substitution or Subsampling rates for balancing the disclosure risk and the information loss. Figure 3.1 demonstrates the MASSC process.

4. Remarks- user guidelines.

Here are some remarks for the MASSC process.

(1) The typical size of the data will be limited by the capacity of the computer. For example, if the user has a small PC with limited speed and memory, a small dataset is recommended, for instance, a dataset of 20 mega bytes or less. If the user has a powerful computer, then the size of the dataset is not necessarily an issue as long as the speed and the system resources are not compromised. Typically, MASSC treats a dataset with the size of 30 mega bytes or less, equal to one or two hundred thousand records or less depending on the size of each record. If the dataset is considerably large and the user does not have a powerful computer, partitioning of the dataset is needed. The user can apply the MASSC treatment to all subsets of the dataset, then combine the treated results together(Singh, Yu, 2003).

(2) The key IVs include core IVs which are most accessible to the intruder because of how they identify profiles, core + non-core group 1 which is the next most accessible group of variables after core IVs, core + non-core group 2 which is the next most accessible group of variables after core + non-core group 1, and so on (add next non-core variable, or group of variables, such that new uniques are produced).

(3) The purpose of grouping noncore IVs is to add a non-core IV or a group of non-core IVs to the core so that a new set of uniques may be identified.

(4) The number of key SVs will be determined by the user. SV has the value that the intruder wants to find out. The number of SVs for the given example is 4(Marijuana use, cocaine use, tobacco use, alcohol use).

(5) The constraints are study variables, such as marijuana use in some domain. The purpose of defining constraints is to combat the bias that MASSC process introduces. For example, Substitution introduces bias, users may want to constrain the substitution so that the relative bias of the CV(constraint variable) is less than Alpha. The number of constraints typically depends on the number of IVs and SVs.

(6) The running time for MASSC process varies from step to step. Some steps like substitution, subsampling, and calibration usually take longer time than others like Micro-Agglomeration. If the dataset is relatively small (for example, 20,000 records or less), the longest running time for a given step is usually within 20 minutes. The treatment for large datasets takes longer.

(7) If the data in the dataset is incomplete, an imputation process is needed before applying MASSC. The purpose of imputation is to replace null and negative values in the dataset with non-null and non-negative values.

(8) 7 ± 2 risk strata are recommended.

(9) 4 substrata for both substitution and subsampling are recommended. In combination with (8), these

recommended numbers of risk strata and substrata will give users a good starting point with the

experiences that were obtained from the past data disclosure treatments conducted at RTI.

(10) Approximate measure of disclosure risk for uniques is given by: $\delta = \pi(1-\psi)\phi(1-\chi)$

Where π – Proportion of uniques w.r.t all IVs

ψ – Overall substation probability

ϕ – Overall subsampling probability

χ - Overall proportion of misclassification

Typically,

$\psi = 0.15$, $\phi = 0.80$, $\chi = 0.10$, need $\pi <= 0.33$ to make $\delta <= 0.20$

5. Summary and Concluding Remarks

The MASSC GUI software is a very user-friendly software with a specific data disclosure treatment mechanism which is strictly guided by the MASSC methodology and has graphic tools that helps users to

turn their focus only on the treatment procedures. The MASSC GUI software limits user inputs by taking only the key words of the input function or sentence. The MASSC GUI provides users with multiple information input screens with the different styles which fit the user's own preference. The MASSC GUI applied the rule of simplicity to all screens, that is, for all the screens within GUI application, there are no extra buttons or lines or graphics or boxes that may complicate the process and

cause confusion to the users. Moreover, the MASSC GUI brought in some SAS callable C procedures

for some key steps of MASSC treatment. This effort dramatically reduces MASSC running time. Finally, the MASSC GUI software was exclusively designed for effectively executing the MASSC procedures. It has been used for some real data disclosure treatments.

In the near future, the different formats(styles) of MASSC GUI software will be developed to fit the needs of variety of real-world users.

References

- **Singh, A.C., Yu, F., and Dunteman, G.H. (2003). MASSC: A new data mask for limiting statistical information loss and disclosure.** ***Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*****, Luxembourg, April 7-9, 2003. (www.unece.org)**
- **Singh, A.C., and Yu, F. (2004), " Protecting Quality and Confidentiality of Micro Data by MASSC: Review with application",** ***European Conference on Quality and Methodology in Official Statistics*****, Mainz, Germany, May 24-26.**

Appendix A

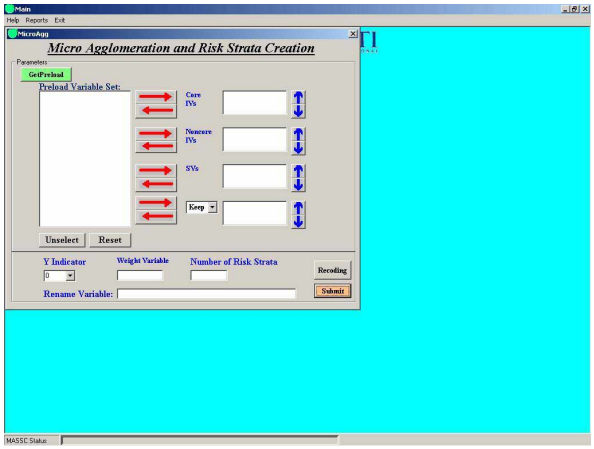

Figure 2.1 Micro-Agglomeration Execution Screen

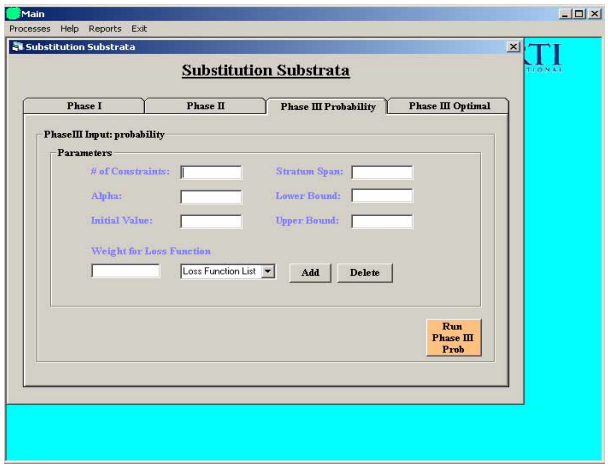

Figure 2.2 Substitution Substrata Execution Screen

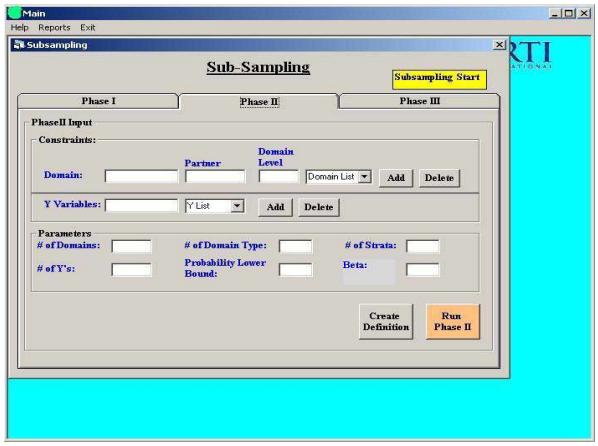

Figure 2.3 Subsampling Substrata Execution Screen

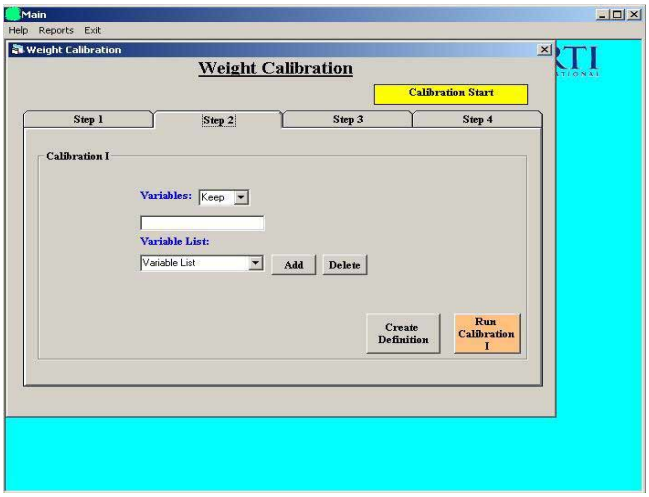

Figure 2.4 Weight Calibration execution screen

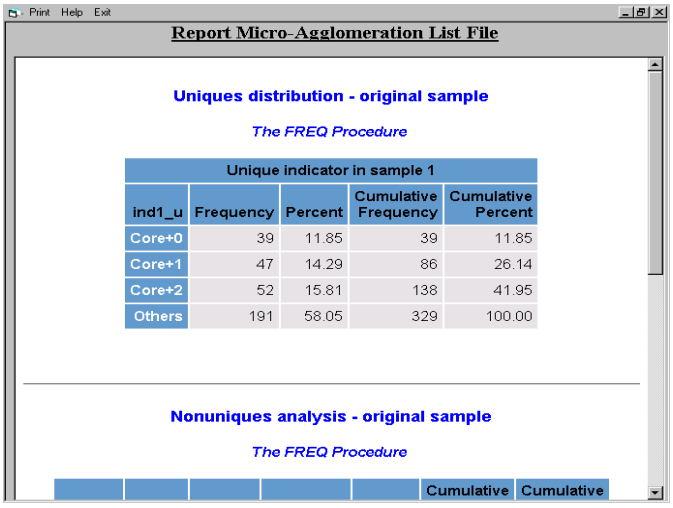

Figure 2.5 Example Report Screen

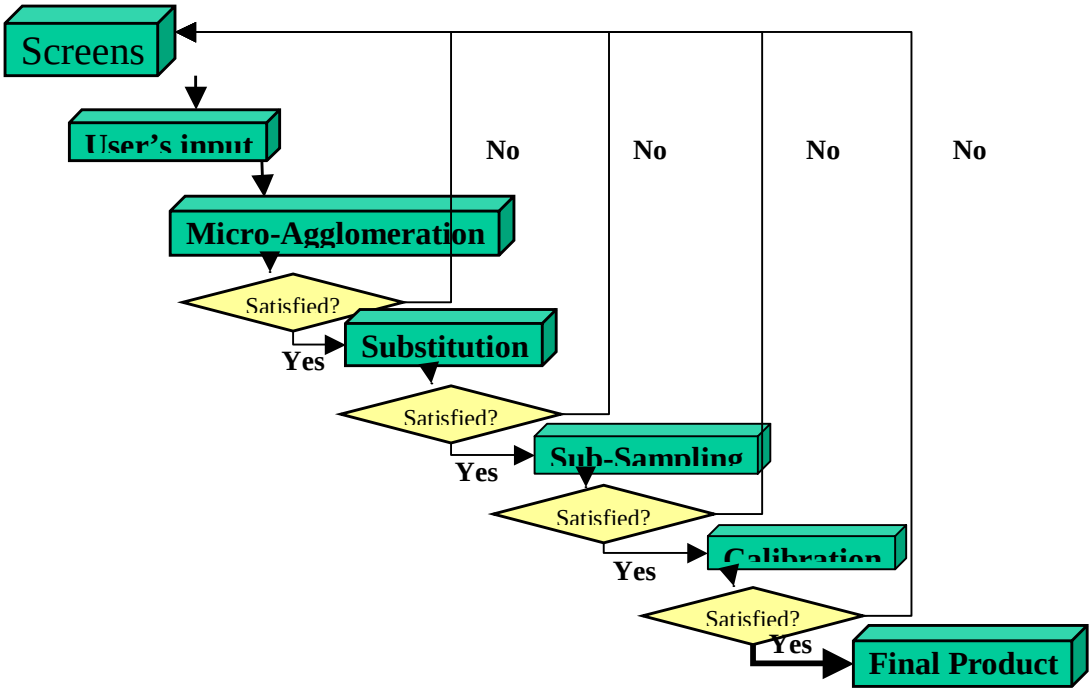

Figure 3.1 MASSC GUI Execution Flow