Guidelines for Requesting Results of Analysis on the Confidential Data

The purpose of this document is to give guidance to SIPP Synthetic Beta (SSB) users who wish to have their computer programs from the Cornell Synthetic Data Server (SDS) run on internal, confidential Census data for validation purposes. In order to facilitate this process and enable Census staff to help as many users as possible, we present some programming requirements and describe the disclosure process for releasing results using confidential data.

1. What is the confidential data?
    a. When we refer to the confidential data we mean the Completed SIPP Gold Standard files. The completed files contain all the originally non-missing data, and fill in the originally missing elements with multiple imputations based on models estimated from the non-missing data. The completed data sets are the confidential data sets that have universes and variable structures comparable to the synthetic data. The incomplete SIPP Gold Standard has a larger universe and more variables, but is not available for validation requests to the Synthetic Data Server users.
2. What kind of results on the confidential data will be approved by the Census Disclosure Review Board for release to users?
    a. For guidance on what kind of results can and can not be released, it is very important to carefully read chapter 3 of the *RDC Researcher Handbook*, found on our website. Some pertinent things to keep in mind for the SIPP Gold Standard data are:
        i. Geography is one of the most sensitive stratifying variables that causes disclosure concerns. In your memo, please explain if and how you use the state of residence variable.
        ii. MAKE SURE TO SUPPRESS MINs and MAXs OF ANY VARIABLE IN YOUR REQUESTED OUTPUT.
        iii. For every statistic you calculate, you need to provide a person-level sample size. Your goal should be to do all your calculations on cells that contain at least 10 people. If you are taking a mean of a 0/1 indicator, than both cell sizes implied by that mean should be at least 10. If you have 0/1 indicator variables in a discrete regression, you need to show a cross-tab

of the discrete independent and dependent variables. The cells in the cross-tab should also abide by the same cell size rules. Keep in mind that a cell close to 10 in the synthetic data might actually be smaller in the confidential data.

iv. Round your point estimates. For most statistics (means, regression coefficients, etc.) round to 3 significant digits.

b. Spend some time thinking about what results you really need. We will reject requests that we deem too large and cumbersome for review and require the user to simplify logs and tables and clarify descriptions. Ideally log files should be no longer than 20 pages.

3. What do you need to provide Census staff when making a request?
   a. You will need to provide SAS or STATA code and the error-free logs of this code in a user-specific subdirectory of sds.vrdc.cornell.edu:/rdcprojects/co00517/SSB/programs/users
      i. This code should run without error on a single implicate (synthetic or completed) of data. The code can have multiple parts, but, if so, please have a single code file that launches the others and clear instructions on which file should be launched. We will want to see the logs and output files for all 16 synthetic data implicates in the aforementioned directory, but you only need to provide code to work on a single implicate.
      ii. At the very top of this code file there should be a variable(s) specifying the name and location of the input data set, the name and location of any output files, and, if your programs use it, the name of the person ID variable. These lines of code should be the only thing the Census employee should have to edit when adapting your code to run on the confidential data on the internal server. The Census employee can then easily edit these variables to run the code internally on each completed implicate.
      iii. Keep in mind that the Census employee running your code for you might not be an expert in the coding language you are using. Do not expect the

Census employee to debug your code for you. Only send code that has successfully run on the synthetic implicates.

iv. Keep in mind that the synthetic data implicates have a slightly different naming convention than the completed data implicates since there are R=4 synthetic data implicates for each of the M=4 completed data implicates. So if you use variables in your code to refer to the two synthetic indices (m and r), design your code so that it will not break when the input data is only referenced with one index (m).

v. The convergence of many optimization routines is very sensitive, so keep in mind that just because a routine converged on the synthetic data does not mean it will definitely converge on every completed implicate. Make sure your code can continue to do everything you want even if convergence fails.

vi. The output of this code should be as simple as possible, containing just the results you want released to you. Limit your output to results that should be disclosable based on the instructions in section 2.

b. Documentation of the output of your code.

i. For disclosure, we follow the system used by the Center for Economic Studies (CES) in all Census RDCs and outlined in *The Researcher Handbook.* This handbook gives guidance to researchers working in a confidential environment on how to have results approved and released. Since an SDS user is not necessarily an RDC user, we can submit your disclosure request internally for you, but you should provide all the documentation necessary for us to make that request. Your documentation and code output should conform to the guidelines laid out in chapter 3 of *The Researcher Handbook*. The handbook also includes a template for this documentation in the appendices, which we require that SSB researchers use as well. You can also find this template on our website (see RDC Disclosure Request Memo).

ii. We strongly recommend that your code produce an output data set for each implicate containing the statistics (i.e. coefficients, means, etc.) you

want released. These output data sets (one for each implicate) can then be the input to a program you use to combine the results using the proper multiple imputation formulae. But make sure the documentation you provide explains every element of every field in these output data sets. We have provided a Stata ado file on the SDS server called estmi which can be used to combine regression results from both synthetic and completed implicates and produce coefficients and confidence intervals that can be properly compared. This routine can be called by a Stata do file and information is available by typing "help estmi" within Stata.

iii. If you prefer not to have results for each implicate and instead only want results combined across all four completed implicates according to the Rubin formulae, you must provide a separate piece of code that takes the output from (a.ii) and combines it across implicates according to Rubin's rules, either using the estmi command in Stata or using code you have written. Again, this code should have lines at the top defining directory names and file names that should be the only things the Census employees have to edit.

4. The Output has been released to you. Now what?
    a. The Census Bureau would like a copy of your paper or research product. An inventory of end user products allows Census to track the topics customers integrate into successful research products. This allows Census to better tailor our products for superior customer service while providing indications of customer interest for future areas of data collection.
    b. We request that researchers who publish results from analyses done using these data cite the SSB as their data source and acknowledge the use of the SDS server at Cornell and the support of Census staff in running any validation programs. These citations will help ensure continued funding for the SDS server and the creation of the Gold Standard File and the SSB.
    c. Suggested acknowledgement: "This analysis was first performed using the SIPP Synthetic Beta (SSB) on the Synthetic Data Server housed at Cornell University which is funded by NSF Grant #SES-1042181. Final results for this paper were obtained from a validation analysis conducted by Census Bureau staff using the

SIPP Completed Gold Standard Files and the programs written by this author and originally run on the SSB. The validation analysis does not imply endorsement by the Census Bureau of any methods, results, opinions, or views presented in this paper. These data are public use and may be accessed by researchers outside secure Census facilities. For more information, visit http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html."

d. Data Citation: U.S. Census Bureau. SIPP Synthetic Beta: Version 6.0 [Computer file].Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2015.