UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

Topic (ii): Managing and supporting changes related to editing and imputation

Creating an Initial Donor Pool for New Questions in the Census of Agriculture

Prepared by Darcy Miller, National Agricultural Statistics Service, United States Department of Agriculture, USA

I. Introduction

1. The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) provides timely, accurate, and useful statistics in service to U.S. agriculture. NASS has two primary programs: the census of agriculture and the agricultural estimates program. The census is conducted every five years, in years ending in 2 and 7. Census data provide a foundation for farm policy. They are used to make decisions about community planning, company locations, availability of operational loans, staffing at service centers, and farm programs and policies. The agricultural estimates program provides reports on virtually every aspect of U.S. agriculture. Many estimates provide market-sensitive information. Both the census and agricultural estimates reports simultaneously provide all market participants accurate supply/demand information for the agricultural sector, which promotes efficiency and fairness in competitive markets.

2. The census of agriculture (COA) is the only source of uniform, comprehensive agricultural data for every state and county in the United States. The census has a list frame with approximately 3 million records. The COA is the leading source for information on the characteristics of people operating U.S. farms and ranches. Understanding changes in farm structure and the demographics of farm operators over time is important in assessing how well USDA programs serve the farm population.

3. After a panel of experts recommended that the COA update information collected about women and new or beginning farmers, NASS redesigned the demographics section and added more than a dozen detailed farm operation decision-making questions, referred to as the decision-making matrix, to the 2017 COA. Data collected in the new questions are unique and similar data are not collected elsewhere in the hundreds of surveys NASS conducts. This major redesign to the questionnaire require changes/updates to downstream processes such as editing and imputation.

4. The COA is imputed using a nearest neighbor methodology. Before each COA, an initial donor pool is formed based on records from previous COAs as well as a content test conducted for the upcoming COA. As data are collected, edited and imputed for the current census, those records are added to the donor pool, with a preference for records to be used from the current census. For the decision-making matrix, data from the 2012 COA or similar data from other NASS survey programs do not exist to use in the initial donor pool. This necessitated the development and implementation of a different method to create values for the initial donor pool for the decision-making matrix. To manage this change in content, different units within an organization worked in concert to effectively to manage editing and imputation processes.

II. History of Editing and Imputation in the Census of Agriculture

5. The census of agriculture moved from the U.S. Census Bureau to NASS in 1997, though the Census Bureau collaborated with NASS for the 1997 census. Until it accepted full responsibility for the data editing of the 2002 Census of Agriculture, NASS handled nearly all of its imputations manually. The size of the census of agriculture brought the need for automated (statistical) imputation to NASS and introduced NASS to a broader understanding of statistical data editing. The NASS Prism system was developed in-house to continue the use of decision logic tables (DLTs) for Census processing, as had been done previously at the Census Bureau. However, the Census Bureau's imputation strategy was modified in the NASS implementation of DLTs. Editing and imputation systems are integrated for both manual imputation and statistical imputation so that editing and imputation happen as data are collected and entered into the system. The imputation does not occur at the end of the process, after all of the records are collected.

6. Edit logic is written by subject-matter experts and are applied in coherent “modules” of the census of agriculture report. The "conditions" portion of DLT processing identifies each data inconsistency, allowing an "action" chosen from a hierarchy of three imputation strategies. First any value that can be determined through DLT evaluation of relevant responses, such as a missing total, is imputed. As its next choice for imputation, DLT logic makes use of previously-reported data. For Census purposes, previously-reported data are assembled from a variety of NASS surveys, as well as the previous Census, and are maintained in their own database. Donor imputation is invoked as the third option.

7. Donor imputation requires a pool of donors who provide values to recipients needing imputation. The donor pool membership begins with a mixture of data from the previous census and preliminary census test data. As editing proceeds over a period of several months, recently-edited records that have passed all of the edits are used to incrementally update the donor pool. Donor data are maintained separately for each "module" which roughly correspond to sections of the COA questionnaire. Many of the distinct donor pools function together to provide imputation during the editing of an entire COA record. Each time donor records are added or updated, all donor records are stratified using a data-driven algorithm that groups farms by type, size and income, according to a strategy developed for each edit module and its respective donor pool. Early in the editing schedule, newer donors are favored over similar donors with older data since the initial donor pool is composed of records from the previous census and preliminary census test data.

8. During editing, each recipient is classified into an appropriate stratum, and the ensuing search is limited to donors in its stratum. Donor selection employs Euclidean distance computations, which are normalized across values within each stratum. The distance computation during the donor search always includes an estimated mileage between the respective county centroids. When appropriate, the donor value may be scaled before imputing the value into the recipient’s record. When a recipient falls outside all current strata definitions, or when none of the donors in the recipient's stratum meet the DLT selection criteria, a backup automated strategy using donor averages may be applied, or the record may be referred to an analyst for manual resolution.

III. Creating an Initial Donor Pool for the New Demographic Questions

9. In 2015, a panel of experts reviewed the COA to determine improvements that could be made to allow data users to better understand the role and effectiveness of USDA programs directed at women and beginning farmers. The panel recommended collecting additional data on the types of decisions made by individuals contributing to decisions on the farm. With this new content in mind, NASS designed a new set of questions to add to the 2017 COA, referred to as the decision-making matrix. These data are not available from previous COAs or other surveys at NASS. Hence, an initial donor pool for the decision-making matrix needed to be constructed differently than the other 2017 COA questions.

10. NASS forms a team each census cycle to lead the effort to implement the editing and imputation system for the COA. The team includes individuals from several divisions at NASS. This team discussed options to create an initial donor pool for the decision matrix. They decided to make the decision-making matrix a separate module from the other demographic module and impute the census content test using a multivariate model to create an initial donor pool for the decision-making matrix module. The PRISM system will continue to flag cells which require values. After the initial donor pool is formed using the imputed census content test data, the nearest neighbour methodology will be utilized for the remainder of the COA records that are collected.

11. To implement the imputation, the team selected IVEware, an iterative multivariate imputation approach recently implemented in other NASS surveys. IVEware is a flexible imputation program developed by the University of Michigan and based on the Fully Conditional Specification (FCS) method described in Ragunathan (2001). The joint distribution is induced from a conditional specification. Parameter estimates and deviates used for imputation are generated through a Gibbs sampling routine (Geman and Geman 1984; Gelfand and Smith 1990). After initialization of this routine, sets of parameter values are drawn iteratively and, for each set of parameter values, missing data are imputed based on a conditional model, where each conditional model may be linear or non-linear (e.g. generalized logit) in nature and a diffuse prior is used for the parameters. IVEware is available as a stand-alone program, or it can be run in SAS (SAS callable). It is easy to implement and NASS's familiarity with SAS led to the decision to run it in SAS.

12. IVEware has several modules are available to perform imputation and to conduct analysis of the data. For the purpose of creating an initial donor pool using census content test data, the IMPUTE module will be used. The IMPUTE module not only defines the model but also contains a host of other features that are appealing to NASS. Within the IMPUTE module, the type of regression used can be determined by defining the variable type. Variable types that can be imputed include continuous, binary, categorical (polytomous with more than two categories), counts, and semi-continuous. All variables in the dataset are potentially used in each conditional model, unless indicated in the transfer statement. The imputation programmer has options to utilize statements for model selection, such as for step-wise regression. The user also has the option to incorporate some types of edits, such as restrictions on variables to be imputed based on the value of other variables and bounded imputations.

13. IVEware is free, user-friendly, and easy to apply on a variety of data sources. Empirically, FCS methods, like those implemented in IVEware, have produced reasonable results (see Ragunathan, et al., 2001; Van Buuren et al., 2006; White and Reiter, 2008) with a high degree of variable flexibility and other desirable features for implementation by a statistical agency. However, the user accepts that convergence may not be reached due to a potential lack of a valid joint distribution. NASS has implemented IVEware for the 2014 Tenure, Ownership, and Transition of Agricultural Land (TOTAL) survey, 2016 Local Food Marketing Practices Survey, and plans to implement IVEware in the 2017 Organic Food Survey as well as development of the initial donor pool for the 2017 COA decision-making questions.

14. Outcomes from imputing NASS surveys in the past along with data types similar to the type of data found in the decision matrix using IVEware were assessed by NASS's operational units and research division and deemed successful. Additional stress testing of the software included running IVEware for 20 iterations on up to 500,000 units with a similar number of variables and data types as the decision-matrix. Up to 50% missing values were missing in the test data. Given that the number of records to be imputed to form the donor pool is only approximately 15,000 records, IVEware can handle the task.

15. NASS is developing imputation models are being developed through a combination of statistical analysis in the research and methodology divisions as well as the farm operator expertise of a NASS subject matter expert. The 2017 COA edit and imputation team meets bi-monthly and additional individual communications occur between those directly involved with creating this initial donor pool. Both verbal communication as well as visual tools such as spreadsheet grids describing the

models are being used to create a successful imputation model to produce the initial donor pool for the decision-making matrix.

IV. Conclusions

16. Organizations that produce official statistics continually review the needs of the data users so that data maintain quality, relevance and timeliness. When changes, especially significant changes, are made to content and survey design, changes may need to be made to methodology or systems in downstream processing. For editing and imputation methods and processes, the changes are often reflexive at NASS, meaning that the changes in content are not often considered in conjunction with changes that will need to be made to editing and imputation. Changes to content are made and edit and imputation methods/systems are fitted to the updated content while simultaneously considering summarization process selected for the data. Hence, large changes to content, as we see in the 2017 COA, require a high degree of collaboration between the divisions within NASS to develop an effective edit and imputation strategy that incorporates the changes made while considering processes that follow the current edit and imputation process. Timely communication between research and operational units has been imperative to create an initial donor pool for the new content added to the 2017 Census of Agriculture. At NASS, the joint effort to develop a strategy has been fruitful and the implementation is expected to be successful.

V. References

Manning, A. and Atkinson, D. (2009). "Toward a Comprehensive Editing and Imputation Structure for NASS – Integrating the Parts". *USDA NASS RDD*. United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Neuchatel, Switzerland, 5-7 October 2009.

Miller, D., Dau, A., and Lisic, J. (2016). "Imputation's Reaction to Data: Exploring the Boundaries and Utility of IVEware and Iterative Sequential Regression (ISR)". Fifth International Conference on Establishment Surveys. Geneva, Switzerland, 20-23 June 2016.

Miller, D. and Young, Linda (2015). "Imputation at the National Agricultural Statistics Service". United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Budapest, Hungary, 14-16, September 2015.

Miller, D., Ridolfo, H., Harris, V., McCarthy, J., and Young, L. (2015). "Expert Panel on Federal Statistics on Women and Beginning Farmers in U.S. Agriculture". Documentation for the Expert Panel on Federal Statistics on Women and Beginning Farmers in U.S. Agriculture. Washington, DC. 2-3, April 2015. Unpublished Report.

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V. and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models". Survey Methodology, 27, 85-95.

Ridolfo, H., Harris, V., McCarthy, J., Miller, D., Sedransk, N., and Young, L. (2016). "Developing and Testing New Survey Questions: The Example of New Questions on the Role of Women and New/Beginning Farm Operators". Fifth International Conference on Establishment Surveys. Geneva, Switzerland, 20-23 June 2016.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC.

Van Buuren , S., Brand , J. P.L., Groothuis-Oudshoorn, C. G.M., and Rubin, D.B. (2006). "Fully conditional specification in multivariate imputation". Journal of Statistical Computation and Simulation, 76:12, 1049-1064, DOI: 10.1080/10629360600810434

de Waal, T., Pannekoek, J., and Scholtus, S. (2011). "Handbook of Statistical Data Editing and Imputation". Wiley Handbooks in Survey Methodology. John Wiley & Sons, Inc.

Report of the Expert Panel on Statistics on Women and Beginning Farmers in the USDA Census of Agriculture. 2015. Unpublished report.