UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

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

Imputation at the National Agricultural Statistics Service

Prepared by Darcy Miller and Linda J. Young, PhD, 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 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 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. Information concerning all areas of farming and ranching operations, including production expenses, market value of products, and operator characteristics is collected. Some census data are also used in frame-building activities for NASS census follow-on surveys such as the Farm and Ranch Irrigation Survey.

3. As part of its agricultural estimates program, NASS conducts hundreds of surveys every year and publishes more than 400 reports. Some examples of areas covered in NASS's reports are production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm income and finances, chemical use, and rural development. A wide variety of topics are covered within these different areas. The subject matter ranges from traditional crops, such as corn and wheat, to specialty commodities, such as mink; from agricultural prices to land in farms. The size of the target population varies from fewer than 50 to all U.S. farms (approximately 2.1 million). The sampling design, data collection mode, processing, estimators, and publication schedule can differ from survey to survey. Depending on the survey, estimates are generally produced at a national, regional, and state level, but not at the county level as with the census of agriculture. Exceptions are the cash rents estimate and county estimates of yield and acreage programs, which publish limited data for some, but not all, counties. For those counties for which data are published, the farm characteristics collected are not as detailed as on the census of agriculture. Data are collected and published at different time intervals: weekly, monthly, and annually. Published estimates include totals and coefficients of variation. Other estimates such as indices and ratio estimates are also produced by NASS. In some cases, economic models are applied to the data and estimates produced by other agencies, such as the Economic Research Service and the Bureau of Economic Analysis.

The survey publications supply information on a more frequent basis than the census's five-year intervals.

II. Historical NASS Perspective and Practices of Imputation

A. Census of Agriculture

4. 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. Software was used primarily to evaluate records and provide lists of data errors and warnings of potential errors. Subject-matter experts used their discretion to make changes to a survey report until the report was internally consistent. They used administrative data, previously reported data from the operation, or industry trends to make changes. Manual imputation on the scale of the census of agriculture is impractical. As part of developing a system to handle the census of agriculture, NASS for the first time gave editing software the authority to change data without an analyst's oversight. Final authority for decisions regarding editing of specific data records remained with analysts. Edit staff were provided tools to conduct manual imputation and to review and override automated changes made by the edit. The tools required access to databases of current survey data, original raw responses, historical values, and frame lists. The edit software ran both as an efficient batch process and as a responsive interactive process invoked upon demand during and after the data collection process. The goal of utilizing both capabilities was to allow edit processing to be balanced between large-scale automated imputation and manual analyst review.

B. Agricultural Resource Management Survey III (ARMS III)

5. ARMS III is part of the agricultural estimates program and is a large annual survey conducted collaboratively with ERS. For the ARMS III, item-level nonresponse is accounted for by imputing data where there are missing values. Some items are manually imputed while others are eligible for statistical imputation. Formerly, imputed values were calculated through an automated imputation system that calculated an unweighted mean for an imputation group based on locality, farm type, and value of sales class. These groups of homogeneous farms excluded extreme outliers (both high and low) so that the imputed values were not biased as a result of a few large/small or unique operations. An imputation group had to have a minimum of ten or more positive responses. When a group lacked a sufficient number of responses, groups were collapsed by value of sales class, locality, and farm type according to a defined hierarchy preserving as much of the homogeneity as possible.

6. The mean imputation methodology has some disadvantages, especially in the ARMS III. The methodology relies on the use of conditional means as estimates of missing values. For survey estimates of univariate-level statistics or statistics cross-classified by several variables, this methodology should be adequate in general. However, estimates of variability in the data are typically artificially reduced. When more complex multivariate relationships are estimated, conditional mean imputation generally cannot condition on a sufficiently large set of variables to maintain relationships among the variables imputed and all variables that might be included as related variables in a multivariate analysis.

C. Other NASS Surveys

7. Other NASS surveys primarily used manual imputation to rectify cells that had been flagged as erroneous through the edit process or cells that had missing values. Analysts could have used previously reported data, administrative data, industry trends, or a combination of methods to derive the imputed value. The exception was a SAS program that imputed appropriately-weighted strata means as part of the NASS Survey Processing System (SPS), in limited instances where analysts

deferred to this automated capability.

III. Editing and Imputation Methodologies

8. For most NASS surveys and the census of agriculture, the editing process is integrated with the imputation process. Hence, discussing the two methodologies together forms a clearer picture of the application of methodology. Some exceptions do allow for statistical imputation to be done separately from the editing process, which is noted where appropriate.

A. Interactive Data Analysis System (IDAS)

9. Survey-based estimators can be impacted by "outliers" – individual reports that have "excessive influence" on the results due to either improper classification, erroneous data, or extremely unusual data for a given operation (i.e. operation is not representative of other operations). NASS conducts a macro review of most of its survey data. An outlier analysis can be performed within commodity and geographical domains at the unexpanded (no weight applied to cell value) and expanded (weight applied to cell value) for many surveys using the Interactive Data Analysis System (IDAS). Graphical displays illuminate unusual observations relative to other units of similar type within domains; unusual observations are reviewed and then changed or confirmed by an analyst.

B. Census of Agriculture

10. 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. Initial plans for a Fellegi-Holt system to process the census of agriculture in 2002 were dropped due to the unrealistic computing demands at the time. Instead, what became 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.

11. 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.

12. 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 consistent records are used to incrementally update the donor pool. Donor data are maintained separately for each "module". In practice, many distinct donor pools function together to provide imputation for the editing of an entire Census record, across all modules. 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.

13. During editing, each recipient is classified into an appropriate stratum, and the ensuing search is limited to donors in its stratum. Handling of each census recipient includes several dynamic choices, which provide alternatives laid out by subject-matter experts while they are composing the DLTs. As

these steps are carried out, the reasoning of the DLT author dictates how a donor is selected and how its values are used. Each imputed field may be limited to positive values, while the donor value may be scaled by the recipient-vs-donor ratio computed for a designated auxiliary variable. Eligibility may be further refined by requiring that potential donors satisfy an expression appropriate for the recipient's circumstances. Finally, the DLT author may anticipate which variables should contribute to the definition of similarity for a given donor search. 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 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.

C. Agricultural Resource Management Survey III (ARMS III)

14. Prior to the 2014 survey year, the ARMS III utilized the SPS system, which is a batch edit system that utilizes linear edits to flag inconsistencies and questionable relationships within a questionnaire report. Cells were flagged at the level of required intervention from warning flags to critical flags. For warning flags, analysts could check the value but did not need to resolve the issue. For critical flags, analysts had to check and resolve the issue or have an explanation approved. The edit was not interactive as it was run in a batch mode. The ARMS III survey now utilizes the Prism system and DLTs to do editing; however, the edit system is only integrated with manual imputation, not statistical imputation. So, statistical imputation is done as a completely separate process from editing. Statistical imputation is executed when all of the data are collected and cleaned through the editing and manual imputation process.

15. Because ARMS III has many complex multivariate relationships the conditional mean imputation generally cannot condition on a sufficiently large set of variables to maintain relationships among the variables imputed and all variables that might be included as related variables in a multivariate analysis. To develop methodology that would incorporate more information when conducting imputation, NASS collaborated with the National Institute of Statistical Sciences (NISS). Iterative sequential regression (ISR) was adapted to ARMS III and implemented for the 2014 survey year.

16. ISR is founded on the normal distribution. Thus, the semi-continuous nature of the ARMS III dataset requires special handling. To handle the probability mass at zero, an indicator variable is constructed for each item to denote whether a value of the item is non-zero or zero. Marginal transformations of the non-zero, continuous portion of each variable are then joined to form a multivariate normal joint density. The multivariate joint density is decomposed into a series of conditional linear models, and a regression-based technique is used. Various criteria utilized by subject-matter experts are used to select the covariates, which allows for flexibility in the selection of the covariates while still providing a valid joint distribution. Parameter estimates for the sequence of linear models and imputations are obtained in an iterative fashion using a Markov-chain-Monte-Carlo (MCMC) sampling method. The ISR method is described as a blend of data augmentation (DA) and fully conditionally specified (FCS) models, having the covariate choice flexibility of the FCS methods but the theoretical background of the DA methods (See Robbins, et al. 2013 for more details).

D. Significance Editing (SignEdit)

17. The SignEdit system is designed to recognize data inconsistencies in NASS survey data and to provide automated imputation to correct them. Statistics Canada's Banff system is used to identify problems and to make changes. The criteria defining a consistent data record are written as a set of linear expressions, which Banff applies to judge each record. If a record has inconsistencies, a mathematically-driven error localization process determines which fields should be imputed to make the record consistent. Banff attempts to minimize the number of changes made to a record. It also incorporates weights assigned by subject-matter experts to determine preferences to change one value

in the record versus another if a decision is needed.

18. The imputation strategy in SignEdit is based on Fellegi-Holt principles. The SignEdit processing system includes the application of Banff to locate inconsistent records, identify cell values to be changed, and implement an imputation method. The Fellegi-Holt methodology anticipates that all survey records will be processed as a batch. However, to make maximum use of all available information, the SignEdit processing system is organized to allow incremental applications of Banff as survey records are received. So, like the census edit and imputation process, editing and imputation is conducted during the data collection process.

19. For fields identified for imputation by Banff, whenever possible, deterministic imputation rectifies those constrained to a unique value by the logic of the linear expressions. Banff techniques appropriate for each survey may then be used for other imputation. According to a hierarchy determined by NASS staff, these include donor imputation, cell averages of current data, and previously reported data. For donor imputation purposes, Banff classifies all available survey records as either donors or recipients, based on their success in satisfying all the linear expressions. When donor imputation is needed, recipients are updated using data from a nearest neighbor chosen from the donors. Data in fields involved in measuring similarity between farms are transformed to scaled ranks, making use of all known values across donors and recipients. This group of similarity indications may be augmented by other fields at the discretion of subject-matter experts. A metric defining separation among recipients and donors is calculated. The metric is the maximum, taken across all relevant variables, of the absolute difference between donor and recipient values in the scaled ranks. The donor chosen by Banff is the one that minimizes this distance.

20. After automated Banff imputation, the SignEdit viewer may be used to follow up on the imputations that have a significant impact on estimates, while the Blaise system must be used to make further changes. The viewer not only shows the Banff changes and their relative influence, but also gauges which records collectively account for a targeted portion of the total magnitude of the accumulated changes. This feature allows priorities to be assigned to individual records, promoting efficient use of time for manual review. Following SignEdit processing, the data are checked by another editing system, Blaise, which may also be used by the analyst to make those manual adjustments deemed necessary. Automated Banff imputation is applicable only to data fields with continuous data; discrete values and administrative data must be handled separately. The SignEdit system is slated for implementation in December of 2015.

E. Tenure, Ownership, and Transition of Agricultural Land (TOTAL)

21. The Tenure, Ownership, and Transition of Agricultural Land (TOTAL) survey is a survey conducted of land owners. A sample of agricultural land owners who are not farm operators and rent out their land are surveyed using a TOTAL survey specifically designed for this purpose. The farm operators who rent out part of their agricultural land and surveyed through TOTAL relevant questions in the ARMS III questionnaire. As part of the integration of the TOTAL survey and the TOTAL portion of the ARMS III, NASS implemented an imputation methodology for TOTAL that was similar to ISR, which was used for ARMS III. ISR could not be applied to TOTAL directly and maintain the statistical theoretical justification for its use due to the nature of many of the variables being imputed. Some of the variables imputed were categorical in nature, such as sex of the landowner. NASS utilized IVEware, which is a flexible imputation program developed by the University of Michigan and based on the FCS method described in Ragunathan (2001). Although the FCS method does not have the ISR properties, such as ensuring a coherent joint distribution or convergence, it has empirically performed well in studies and is utilized in organizations inside and outside of the United States (see Van Buuren, S. et al.). Outcomes from imputing the TOTAL data utilizing IVEware were assessed by NASS's operational units and deemed successful.

IV. Conclusions

22. NASS continues to evaluate methods that may improve its editing and imputation processes and to apply these improved methods to more surveys. The implementation of more modern statistical imputation methods such as ISR, IVEware, and SignEdit reflect a cultural shift from manual handling every record to modelling distributions to maintain the integrity of the distributions and the estimates. Currently, several surveys are in the pipeline to utilize SignEdit. Methods utilized for the Census and other major surveys are under review for potential improvements to their editing imputation methodologies. NASS's vision for editing and imputation is to reduce the extent of analyst intervention to increase consistency and processing efficiency, to account for variability in the editing and imputation processes, and to improve overall data quality.

V. References

Barboza, W., Miller, D. and Cruze, N. (2014). "Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey". United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Paris, France, 28-30, April 2014.

Fellegi, I., Holt, D. T. (1976). "A systematic approach to automatic edit and imputation". Journal of the American Statistical Association, vol. 71, ppg. 17-35.

Hidiroglou, M., and Berthelot, J. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". Survey Methodology, Vol. 12, No. 1, ppg. 73-83.

Johanson, J. M. (2012). "Banff Automated Edit and Imputation on a Hog Survey". Proceedings of the Fourth International Conference of Establishment Surveys, June 11-14, 2012, Montréal, Canada [CD-ROM]: American Statistical Association.

Kosler, J. (2012) "Survey Process Control with Significance Editing: Foundations, Perspectives, and Plans for Development". Proceedings of the 2012 Joint Statistical Meetings.

Kozak, R. (2005). "The Banff System for Automated Editing and Imputation". Statistical Society of Canada Annual Meeting. Saskatoon, Canada, 12-15 June 2005.

Lawrence, D., and McKenzie, R. (2000). "The General Application of Significance Editing". Journal of Official Statistics, Vol. 16, No. 3, ppg. 243-253, 2000.

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., Robbins, M., and Habiger, J. (2010). "Examining the Challenges of Missing Data Analysis in Phase Three of the Agricultural Resource Management Survey". Proceedings of the 2010 Joint Statistical Meetings, pages 816-829.

National Agricultural Statistics Service (2014). "Farm Production Expenditures Methodology and Quality Measures".
http://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Production_Expenditures/08_2013/fpxq0813.pdf

National Agricultural Statistics Service (2014). "Farm Production Expenditures 2012 Summary".
http://usda01.library.cornell.edu/usda/current/FarmProdEx/FarmProdEx-08-02-2013.pdf

National Agricultural Statistics Service (2013). "Quarterly Hogs and Pigs Methodology and Quality Measures".
http://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Hogs_and_Pigs/12_2013/hgpgqm1213.pdf

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.

Robbins, M., Ghosh, S., and Habiger, J. (2010). "Innovative Imputation Techniques Designed for the Agricultural Resource Management Survey". Proceedings of the 2010 Joint Statistical Meetings, pages 634-641.

Robbins, M., Gosh, S., and Habiger, J. (2013). "Imputation in high-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey". Journal of the American Statistical Association, 108:501, 81-95, DOI: 10.1080/01621459.2012.734158.

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.