Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Virtual, 3-5 October 2022)

Session 1b: Modernisation of data editing and statistical production (Part 2)

Growing a Modern Edit and Imputation System

Prepared by Megan Lipke, Darcy Miller, Vito Wagner, Karl Brown, and Vikas Agnihotri of 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 (COA) and the agricultural estimates program. The census is conducted every five years, in years ending in 2 and 7. COA 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 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. 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 COA. Exceptions are the cash rents county estimates and crops county estimates 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 COA. 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.

4. In many cases, the data collected contain missing or erroneous values. To mitigate bias due to these factors, data collected are edited, imputed, and analysed prior to summarization to produce estimates. NASS continually looks to improve its processes to ensure the highest quality of estimates and efficient allocation of staff and technological resources. In October 2016, NASS commissioned a review of its editing and imputation processes. Survey processes from data collection to data dissemination were documented and recommendations were made in an official report in June 2017. Major recommendations focused around automating imputation using statistical methods in surveys where primarily manual corrections are made for missing and erroneous values.

5. Over the past 5 years, NASS has taken steps to modernize and generalize its processes to include automation as well as improved imputation methodology to increase overall efficiency and data quality. This paper discusses the technical, administrative, and cultural growth moving NASS towards a more modern editing and imputation system and the challenges encountered.

II. Planting the Seed: Editing and Imputation Review

A. Background

6. NASS needed to build a united vision regarding editing and imputation methodology and systems for the agency. Different editing and imputation methodologies are used for NASS surveys, the agricultural census, and census follow-on surveys. In turn, several different systems are located on various platforms that are utilized to perform editing and imputation. For most surveys conducted by the agency, the corrections of edit failures and item imputations are a manual process. In many cases, the corrections of unit imputations are performed manually rather than by weighting. The agency needed to revisit the current philosophy and then reevaluate the systems that support it.

7. NASS commissioned an independent review of the editing and imputation methodologies used by NASS and the systems that support them. This review was conducted by Westat, a contractor with knowledge of this particular statistical area. The evaluation included an overview of editing and imputation methodologies used by other governmental statistical agencies within and outside the U.S.; independent, non-profit institutions; and well-known private survey organizations. In addition, any new innovative editing and imputation methodologies applicable to NASS were outlined. After performing a thorough assessment of the methodologies and systems used within NASS, Westat identified any issues with editing and imputation methodologies and/or systems. The final report included recommendations on how to proceed with a unified automated statistical editing and imputation approach.

B. Editing and Imputation in the COA

8. The COA moved from the U.S. Census Bureau to NASS in 1997, though the Census Bureau collaborated with NASS for the 1997 COA. Until it accepted full responsibility for the data editing of the 2002 COA, 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 COA

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 were 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.

9. Edit logic is written by subject-matter experts and is applied in coherent “modules” of the COA 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 COA purposes, previously reported data are assembled from a variety of NASS surveys, as well as the previous COA, and are maintained in their own database. Donor imputation is invoked as the third option and is carried out using a custom script.

10. Donor imputation requires a pool of donors that 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 corresponds to a section 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.

11. 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.

C. Editing and Imputation in PRISM Surveys

12. Some of the larger surveys also use the PRISM system for editing and imputation processing. Like the COA, DLTs may be used to make some deterministic imputations or use previously reported data. Some manual corrections are also allowed during this editing phase. Unlike the COA, nearest neighbor imputation is not implemented through a call from the DLT. Statistical imputation is conducted as a separate phase, outside of PRISM, after data have been collected. Depending on the survey, the statistical imputation method changes. The statistical imputation methods used for some of the surveys have already been updated to modern multivariate statistical imputation methods. Some of these modern methods are implemented using commercial-off-the-shelf (COTS) software. The remainder are primarily imputed based on a conditional mean approach using an in-house developed script.

D. Editing and Imputation in Non-PRISM Surveys

13. Most of the surveys NASS administers do not use the PRISM system for editing and imputation. The primary editing and imputation instrument for these surveys is the interactive edit in Blaise. Blaise is also used as a data collection and processing system for most of these surveys. Using the interactive edit in Blaise, errors are flagged in an edit and analysts use multiple external resources as information to manually make corrections. It is a record-by-record interactive

process and is completed during data collection. A separate phase for statistical imputation is not conducted for most of these surveys. Automated corrections of any kind are limited in each survey, reducing consistency.

14. An exception to the predominant use of manual corrections in surveys that utilize the Blaise interactive edit is in the June Area Survey. The June Area Survey is sampled from NASS's area frame and provides direct estimates of acreage. It is an important component in coverage adjustments for many other surveys that are list frame based and the COA. In this case, an in-house script was developed that uses administrative sources to impute missing values.

E. Final Report Recommendations

15. After Westat reviewed and documented the editing and imputation processes at NASS in detail, a final report was provided to NASS. Recommendations within the report include, but are not limited to:

- Using COTS software where possible and adding the standard packages and procedures to the NASS editing and imputation toolbox
- Not replacing current procedures with Banff
- Adopting an evolving approach to updating systems
- Limiting manual editing and imputation
- Avoiding reliance on the multivariate normal distribution and moving towards fully conditional specification (FCS) using predictive mean matching and other models within the FCS paradigm
- Standardizing code in searchable libraries and leveraging current repositories (metadata, etc.)
- Using Optical Character Recognition (OCR) and Intelligent Character Recognition (ICR)
- Emphasizing more modularized systems (separate phases for editing and statistical imputation)

III. Tending to the Fields: Updating Methods and Practices

16. NASS plans to build a generalized system for hundreds of surveys, called IDEAL (Imputation, Deterministic Edits, Automation and Logic). Because this project is a multi-year effort, a NASS team identified areas where efficiencies could be created while increasing data quality relatively quickly. Those three efforts are Warning Pass-Through, Estimation Tools, and Automated Administrative Code Editing. These 3 efforts and the progress on the IDEAL system are described next. Then some of the ways that common challenges in large, modernization projects were overcome are described.

A. Warning Pass-Through in Blaise

17. The first internal effort realized by NASS was warning pass through. In the Blaise interactive edit, there are two types of errors: warnings (soft) and critical errors (hard). A critical error requires a change in the cell's value to make the record clean and to enable the record to be processed downstream (e.g. enter analysis system and estimation system). A warning error does not require the value to be changed, but it does require manual intervention and keystrokes for each warning error to 'suppress' the warnings. This manual intervention for each warning error consumes time but does not improve data quality. A record is not able to be processed downstream until all critical errors are resolved and all warning errors are resolved or suppressed. The idea behind warning pass-through is to not require warning errors to be suppressed to continue processing a record. Note that warning errors can still be viewed in a list (as they are now) and values related to that warning error still may or may not be changed. Many warning errors can also be viewed and addressed in the NASS analysis system where impactful records can also be identified. This change in policy increases efficiency and provides more time for reviewing records that impact estimates.

18. The mechanics have been established, allowing records that have warning errors (but not critical errors) to continue processing. Staff were trained, and survey documentation and communications were updated. Then a schedule that relied on NASS's schedule to move surveys from Version 4 of Blaise to Version 5 of Blaise was set. It was anticipated that this would help staff remember which surveys had warning pass through available. So far, over twenty-five Blaise surveys have been converted, providing staff more time to focus on complicated records and records that impact estimates.

B. Extreme Operator Record Estimation Tool

19. This tool provides estimated values for records that are (1) sampled with probability one (must strata records), which are primarily the largest producers, and (2) are also unit nonresponses. This estimation process was automated quickly because these records are already handled separately from other records and some code developed by staff in the NASSS regional field office could easily be generalized, updated, and launched in a space for all regional field offices to use. Utilizing the program to automate this process reduced some clerical work associated with moving data from other NASS databases manually and provided tracking and repeatability to the process.

20. This tool was developed for two surveys: the Off Farm Grain Stocks Survey and Hog Inventory Survey. The estimated values are based on a ratio of the current to previous quarter where data were collected. An interactive user interface allows analysts to remove records used to calculate this ratio or apply a subject matter expert ratio developed outside of the tool. The tool also provides some tracking abilities so that the process can be replicated, if needed. The tool was tested and documented, and NASS regional field office staff were trained to use it. The tool has reduced the time staff spend interacting with other databases and manually moving data, which allows analysts to devote more effort on other, higher value-added activities.

C. Blaise Automation Manipula (BAM!)

21. Many NASS administrative code changes have consistent business rules across surveys, and NASS could implement these changes in a generalized format for all surveys processed in Blaise (and some in PRISM too). BAM automates the administrative code changes, providing staff more time to review administrative codes associated with global business rules.

22. In collaboration with operational units, the BAM code was developed, tested, and implemented for all Blaise surveys and three PRISM surveys that have come online this year. Documentation of the purpose and logic used was created, and NASS staff in the regional field offices were trained. Between January and September 2021, this process saved over 1000 hours of staff time.

D. Imputation, Deterministic, Edits, Automation and Logic (IDEAL)

23. And last, but not least, IDEAL is the large, longer-term project to modularize and automate editing and imputation processes for Blaise surveys in a generalized system. The goal here is to decrease staff time devoted to editing and increase data quality.

24. An automated parser was developed to extract, document, and organize edit logic in the Blaise code for multiple surveys. This was required because detailed specifications for editing and imputation in Blaise surveys are not documented. The parsed logic is provided in an organized document that can be reviewed by a NASS panel of experts for approval. The parsed code, once approved, can be converted into logic in R. R is the software used to validate the values, provide imputations, and correct invalid values.

25. The IDEAL system is composed of three components, working in tandem. The first component is the "R Engine" that applies edit logic, conducts imputation, and automates corrections. The second component is a user interface that is used to (1) manage survey instances, user access, and edit logic; (2) observe and track outcomes from the R Engine; (3) make any additional manual interventions needed; and (4) monitor clean status (all critical/hard edit logic met). The R Engine utilizes several packages including *validate*, *error_locate, validatetools*, *dcmodify*, *deductive* and *simputation* from Statistics Netherlands. There is also a database to hold and maintain data being created.

26. Cleaning the data occurs in several modules: (1) Validation: Validate data and identify where values require imputation; (2) Imputation; (3) Error Identification: identify where errors remain; and (4) Automated Correction: correct remaining errors utilizing a Felligi-Holt paradigm. The IDEAL system is also generalized so that logic for variables on one survey can be applied to another survey.

27. A cloud-based environment (Microsoft Azure) will be used in production to run IDEAL. As of August, the team is working with the NASS IT staff to create a sandbox for testing pieces of the newly developed system.

IV. Conclusions: Future of Editing and Imputation at NASS

28. Organizations that produce official statistics continually review survey processes to increase data quality, consistency, and efficiency of resources. NASS has completed the review phase and is currently modernizing its processes. A major priority is to automate the manual imputation done by analysts. A generalized system for hundreds of surveys is being developed, which is a large undertaking. Because the system development is a multi-year effort, the NASS team identified areas where efficiencies could be created relatively quickly while increasing data quality. NASS looks forward to a phased implementation of IDEAL beginning later this Winter and building out to a fully developed system next year.

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.

Dau, A. and Miller, D. (2018). "Dancing with the Software". 2018 Joint Statistical Meetings. Vancouver, BC, Canada, 28 July – 2 August, 2018.

E.de Jonge and M. van der Loo, "errorlocate: Locate Errors with Validation Rules," 2018.

E. de Jonge and M. van der Loo, "validatetools: Checking and Simplifying Validation Rule Sets," 2019.

Miller, D. (2021). "Growing a Modern Editing and Imputation System". 2021 Federal Committee on Statistical Methodology Conference.

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.

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.

SAS. (2011). “SAS/STAT 9.3 User’s Guide” https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect003.htm

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

M. van der Loo and E. de Jonge, "dcmodify: Modify Data Using Externally Defined Modification Rules," 2018.

M. van der Loo and E. de Jonge, "deductive: Data Correction and Imputation Using Deductive Methods," 2017.

M. van der Loo, "simputation: Simple Imputation," 2017.

M. van der Loo and E. de Jonge, "validate: Data Validation Infrastructure," 2018.

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.

Fellegi, I. P. and D. Holt (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association 71, pp. 17–35.