Challenges in Moving Research into the Operational Production Environment

Wendy J. Barboza[1]

Abstract

The mission of the National Agricultural Statistics Service (NASS), a statistical agency under the United States Department of Agriculture (USDA), is to publish statistics in service to U.S. agriculture. NASS accomplishes this mission by collecting data from farmers, ranchers, and agri-businesses across the U.S. and then publishing hundreds of reports covering virtually every aspect of agriculture. Within NASS, the agency is separated into divisions based on functionality; one of the divisions focuses on research and development. The primary goals of the Research and Development Division (RDD) are (1) to improve the data collection and statistical estimation methodology for the agency's surveys and censuses, and (2) to maintain, develop, and improve NASS's operational data products. Thus, the projects investigated by researchers within RDD concentrate on making improvements to statistical processes and procedures related to the production cycle of surveys and censuses. Once the research project is deemed to be complete, it needs to be incorporated into the operational production environment (a.k.a., development). The development phase is not a trivial task. This paper discusses the challenges of incorporating three different projects into the operational production environment.

Key Words: research, development, production environment

1. Introduction

The National Agricultural Statistics Service (NASS) is a statistical agency within the United States Department of Agriculture (USDA). NASS's mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. NASS accomplishes this mission by collecting data from farmers, ranchers, and agri-businesses across the U.S. and then publishing hundreds of reports covering virtually every aspect of agriculture. 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 demographics of U.S. producers. 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 mushrooms and flowers; from agricultural prices to land in farms; from once-a-week publication of cheddar cheese prices to detailed census of agriculture reports every five years.

Within NASS, the agency is separated into divisions based on functionality; one of the divisions focuses on research and development. The primary goals of the Research and Development Division (RDD) are (1) to improve the data collection and statistical estimation methodology for the agency's surveys and censuses, and (2) to maintain,

[1] Wendy Barboza, USDA/NASS, 1400 Independence Avenue, SW, Washington, DC 20250, email: wendy.barboza@nass.usda.gov.

develop, and improve NASS's operational data products. Thus, the projects investigated by researchers within RDD concentrate on making improvements to statistical processes and procedures related to the production cycle of surveys and censuses. Once the research project is deemed to be complete, it needs to be incorporated into the operational production environment (a.k.a., development).

2. Definition of Research and Development

The definition of research and development can vary. For purposes of this paper, the author will utilize the definition according to the Office of Management and Budget (OMB). Within the United States, the Federal statistical system is decentralized and OMB provides oversight of the different Federal statistical agencies. OMB reports directly to the President and helps a wide range of executive departments and agencies across the Federal Government to implement the commitments and priorities of the President. Within the statistics arena, OMB is charged with developing and overseeing the implementation of Government-wide principles, policies, standards, and guidelines concerning the development, presentation, and dissemination of statistical information.

According to OMB, the definition of research and development is:
"**Research and development (R&D) activities** comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.

Basic research is defined as systematic study directed toward fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts without specific applications towards processes or products in mind.

Applied research is defined as systematic study to gain knowledge or understanding necessary to determine the means by which a recognized and specific need may be met.

Development is defined as systematic application of knowledge or understanding, directed toward the production of useful materials, devices, and systems or methods, including design, development, and improvement of prototypes and new processes to meet specific requirements."

Within NASS, RDD concentrates on both applied research and development. The development phase is not a trivial task. The division's largest challenge is moving research into production.

This paper discusses the challenges of incorporating three different projects into the operational production environment. These projects were specifically chosen to represent a variety of different issues. An overview of each project, the new research methodology, and the major hurdles associated with moving the research into development is provided.

3. Modeling Corn and Soybean Yields

During the crop season, NASS publishes forecasts and estimates of crop yield on a monthly basis. The first forecast for corn and soybeans is published in August. The forecasts are updated monthly until November, and the final yield estimates are published at the beginning of January. NASS conducts several surveys to obtain data for forecasting and estimating crop yields: the Agricultural Yield Survey (AYS); the Objective Yield Survey (OYS); and the Quarterly Crops Acreage, Production, and Stocks (APS) Survey. The Crop Progress and Conditions Survey is also utilized, but in a different manner.

The AYS is a farmer-interview survey conducted monthly from May through November. Each month from August through November, enumerators ask questions on corn yield in 41 states and soybean yield in 31 states. The OYS is a field-measurement survey conducted monthly from May through December. Each month from August through December, enumerators collect data on corn in 10 states and soybeans in 11 states. Enumerators measure crop characteristics in the sampled plot, such as number of plants, number of fruit, and fruit measurements. The Quarterly Crops APS Survey is a farmer-interview survey conducted quarterly in March, June, September, and December. In December, enumerators ask questions on corn and soybean yields in 48 states. The Crop Progress and Conditions Survey is an interview survey conducted monthly from early April until late November. Enumerators ask two types of questions, crop progress and crop condition, in 48 states. Crop progress questions ask respondents to estimate the percent of a particular crop that is at or beyond a specified stage of development, while crop condition questions ask respondents to estimate the percent of a particular crop that is in each of five condition categories ranging from very poor to excellent.

The Crop Agricultural Statistics Board (ASB), whose members are commodity experts, uses the survey results along with weather data to determine yield forecasts and estimates. The Crop ASB compares the current data for a particular month to historical results for the same month and synthesizes the information to obtain a forecast/estimate for the geographic area represented by the OYS states. This process is not easily repeatable and does not result in an associated measure of uncertainty.

3.1 New Methodology

RDD developed a model that mimics the Crop ASB's decision-making process (Nandram et al. 2013). For the geographic area represented by the OYS states, a hierarchical Bayesian model is used to combine both current and historical corn and soybean yields from the AYS, OYS, and Quarterly Crops APS Survey (when available) for a particular month to obtain a composite forecast. The model also incorporates three covariates, although they differ somewhat for each crop. The first covariate is the state's percentage of the crop rated good or excellent for a particular week; this information comes from the Crop Progress and Conditions Survey. The particular week is different for corn and soybeans, week 30 for corn and week 34 for soybeans. The second covariate is the state's average monthly temperature in July for corn and in August for soybeans, although July is used for soybeans until August is available. The third covariate is the state's average monthly precipitation in July for corn and in August for soybeans, although again July precipitation is used for soybeans until August is available. Finally, the model includes a linear trend, which accounts for the fact that corn and soybean yields are increasing over time. In addition to the forecasts/estimates, standard errors are obtained.

3.2 Challenges

One challenge was updating and loading the database used by the Crop ASB. Only certain personnel within NASS have access to this confidential database and RDD personnel were not granted access to it for this project. In order to update and load the database, RDD staff first needed to understand it. To overcome this hurdle, RDD staff developed specifications for updating and loading the database after spending a significant amount of time with personnel who had access to it. In the end, the modeled estimates and standard errors were saved to a secure directory and approved personnel loaded them into the database.

Another challenge was transferring this new methodology to operational staff. As of today, this transfer has not happened. RDD staff documented the new methodology, outlined the programs utilized, and personally trained staff in the production environment. Rotation of operational staff is common within NASS, and the person responsible for running the programs changed three years in a row. Rather than retraining another person, RDD decided to produce the modeled estimates until a stabilization occurred.

There was an occasion where the modeling program failed during the survey proper due to circumstances beyond NASS's control. Two of the covariates, temperature and precipitation, are obtained from data files produced by an external source. Due to new technology and methodology, the format and data contained in these external files were updated. RDD was not aware of this update until the modeling program failed. Although changes were quickly made to the program, this is a concern when using an external source.

4. Incorporating Automated Editing/Imputation

Since the early 1990s, NASS has used Statistics Netherlands' Blaise software to process many of the agency's surveys. In 2010, NASS took advantage of new Blaise functionality and started processing these surveys in a centralized environment, which was a tremendous improvement. Blaise amply supports computer-assisted telephone interviewing (CATI) in the agency's data collection centers. In addition, after data collection, analysts within the agency can review edit failures identified by the Blaise edits. Two drawbacks with this capability are that there is not an automated editing feature available and all edit failures are manually edited by the agency's analysts.

In 1997, responsibility for conducting the agricultural census was transferred from the United States Department of Commerce's Census Bureau to NASS. With this transfer of ownership, the largest sample size for any national-level survey conducted by NASS changed from 75,000 records to over 2 million records. Although NASS's traditional approach had been to manually address edit failures for surveys, the agency adopted the computer edit logic and donor imputation previously utilized by the Census Bureau. The agency realized this paradigm shift was necessary in order to process the census of agriculture in a timely manner.

The census of agriculture was the first step toward changing the agency's culture away from manually handling edit failures. To address analysts' concerns that automated edit corrections may be unacceptable, the agency incorporated methodology that allowed analysts to perform a manual review of data changes made during the automated process. Unfortunately, this methodology and the editing and imputation processing system used for the agricultural census was not easily portable to NASS's surveys.

4.1 New Methodology

RDD is currently evaluating Statistics Canada's Banff software to perform the editing and imputation for surveys. Banff is a system that consists of a collection of specialized SAS procedures (Banff Support Team 2008). It requires the edits to be expressed in linear form and assumes the survey data are numeric and continuous. Banff performs automated statistical edits using Fellegi-Holt methodology (Felligi and Holt 1976), which attempts to satisfy all edits by changing the fewest possible values. Banff verifies that the edits in a group of edits are consistent with each other. A group of edits involving n variables defines the feasible region, or acceptance region, in the n-dimensional space. If a record falls within this feasible region, it has satisfied all of the edits within the group. If a record falls outside the feasible region, Banff's error localization procedure identifies the minimal number of variables that must be changed in order for the record to pass all of the edits. The original data are not changed at this point. The values that will replace the original values for these variables are determined during the imputation phase. Note that since Banff assumes the survey data are numeric and continuous, some questionnaire items are not good candidates for Banff (e.g., gender).

Within Banff, NASS is utilizing multiple options for performing the imputation. By employing several alternatives, it decreases the probability of manual intervention. The ordering of the alternatives depends on the survey and is specified by the subject matter experts. For NASS's quarterly hog survey, deterministic imputation is used first to determine if there is only one possible value that would satisfy the original edits. If not, donor imputation is then evaluated to see if there is a nearest neighbor available to provide current data that would allow the record to pass the edits. This procedure requires a minimum number of ten donors. Next, an imputation is attempted by using the record's previous survey data and applying an estimator function to impute the current value. This methodology is restricted to certain variables. Finally, an imputation is attempted by using the mean based on current data within a specified group and applying an estimator function to impute the current value. At the end of the imputation phase, a prorating procedure is implemented to round imputed fields to ensure the record passes the edits. After imputation, the error localization procedure is run again to ensure the unchanged values and the newly imputed values pass all of the edits. If a record does not pass all of the edits, the imputed values are returned to the original values and the record requires manual intervention.

4.2 Challenges

One challenge was obtaining approval from the NASS Enterprise Architecture Council (NEAC). This council is responsible for "supporting the business initiatives of NASS by applying Enterprise Architecture concepts, principles, standards, defined processes, and chosen technology in a consistent fashion across all Agency projects." All new systems are reviewed by NEAC. Documentation is submitted to the council via a standard template and they provide feedback on each layer (e.g., application, database, infrastructure, and security). NEAC required an enterprise database structure rather than SAS flat files. In order to satisfy NEAC's goal of moving NASS systems into an enterprise environment, an IT specialist was brought in to create a centralized database.

As part of the system, RDD had created an application tool in SAS AF to review the original data, the Banff edited data, and the final data (which could be different if an analyst decided to override the change made by Banff). This new tool was included in the system documentation to NEAC. Although SAS AF was the standard software used by current NASS systems, NEAC specified using .NET as the standard software for new applications. In order to satisfy NEAC's goal of using standard software, another IT specialist was engaged to convert the new tool to .NET.

Banff needed to be incorporated with the current Blaise processing environment. To accomplish this, RDD worked closely with the Blaise programmers. The Blaise shell (i.e., introductory code used for all surveys) was updated so that an analyst could not open a survey response prior to performing the Banff edit. Two procedures were established to transfer data back and forth between Blaise and Banff as well as quality control checks so that records were not lost during this process. Unlike many agencies, NASS analysts edit the data during the survey proper. An IT specialist was recruited to write script code to interactively transfer the data between the two systems.

5. Improving Imputation Methodology

The Agricultural Resource Management Survey (ARMS) is conducted by NASS and cosponsored by USDA's Economic Research Service (ERS). The ARMS provides an annual snapshot of the financial health of the farm sector and farm household finances and is administered in three phases. Based on data collected during the third phase (a.k.a. ARMS III), NASS publishes estimates of farm production expenditures for the U.S. (except Alaska and Hawaii) in addition to five regions. The regional estimates are broken down by the fifteen leading cash receipt states and then all other states within the region. Farm production expenditures are also estimated for eight economic sales classes and two farm type categories. In addition to farm production expenditures, the ARMS III also collects data on production practices and costs of production for one to three targeted crop and livestock commodities each year, selected on a rotational basis. The production practices and cost of production data for these designated commodities are collected in the top producing states; the farm production expenditures data are collected in all states (except Alaska and Hawaii).

Prior to imputation, analysts perform a cursory manual edit of the ARMS III questionnaire. The questionnaire is then processed through a computer edit that checks the consistency of the data and verifies data values fall within a certain range. After this, analysts review all questionnaire items that fail any of the edits. Analysts have the option of manually imputing the data item or letting the computer program impute it. A manual imputation is typically performed over a machine imputation when an analyst has knowledge about the questionnaire item for that farm.

For the ARMS III, missing data items are calculated through an automated imputation algorithm that calculates an unweighted mean for an imputation group based on locality, farm type, and value of sales class. These groups of homogeneous farms exclude extreme outliers (both high and low) so that the imputed values are not biased as a result of a few large/small or unique operations. An imputation group must have a minimum of ten or more positive responses. When a group lacks a sufficient number of responses, groups are collapsed by value of sales class, locality, and farm type according to a pre-defined hierarchy, preserving as much of the homogeneity as possible.

5.1 New Methodology

There are some disadvantages with the mean imputation methodology, 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 will typically be 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 between the variables imputed and all variables that might be included as related variables in a multivariate analysis.

RDD developed a multivariate imputation approach to preserve important relationships and the distribution of the respondents' data as well as to provide a better estimate of uncertainty. The new methodology incorporates more information (covariates) and requires data to be transformed marginally and 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 are used to select the covariates, which allow 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 sampling method (Robbins et al. 2013). This new methodology is referred to as iterative sequential regression (ISR). NASS performed an analysis on several years of data; although there was a significant difference in some estimates between the two methodologies, the differences were caused by outliers that would have been identified and corrected during the survey proper (Barboza et al. 2014).

5.2 Challenges

A major challenge was to modify the program to be useable in the operational environment. An RDD researcher with proficient programming knowledge was assigned to document the program and make it more user-friendly. The code was generalized so that variables and items that changed on an annual basis were not embedded in the program. The program's speed was too slow for the operational environment so the researcher recoded the iteration methodology in another software language.

Both NASS and ERS were concerned about how this new methodology would affect the survey estimates for farm production expenditures, which is the primary variable of interest for the ARMS III. ERS asked NASS to run a parallel test to provide a measurement of change due to the new methodology versus the actual change in farm production expenditures. This request was not possible to fulfill during the survey proper due to the short timeframe between data collection and publication. However, RDD was able to provide comparable data for the previous three years.

Another challenge was explaining ISR, rather than mean imputation, to a data user. The new methodology is statistical in nature, even to a seasoned mathematical statistician. NASS had to provide data users with understandable, non-technical documentation of the imputation methodology. RDD worked closely with operational staff and ERS to develop this documentation in addition to assisting in developing workshops to educate the data users within ERS who utilized the final dataset to perform detailed analysis on issues of interest.

6. Conclusions

In conclusion, the primary goals of the Research and Development Division (RDD) are to (1) improve the data collection and statistical estimation methodology for the agency's surveys and censuses, and (2) maintain, develop, and improve NASS's operational data products. Once the research project is deemed to be complete, it needs to be incorporated into the operational production environment. The development phase is not a trivial task, and the division's largest challenge is moving research into production. This paper summarizes three different research projects and some major challenges encountered when moving the research into development.

Acknowledgements

I would like to thank my boss, Dr. Linda Young, for encouraging me to write on this topic. Originally, I thought it unwise to write a paper that was not statistical in nature. But, after finishing this paper, I believe the importance of this topic is worthy of discussion. Mark Harris, who was my previous boss, used to say, “The research is easy; it is the development that is hard.” Although I would not insinuate the research is easy, a majority of the time, the research is easier than the development.

References

Nandram, B., Berg, E., & Barboza, W. (2013). A hierarchical Bayesian model for forecasting state-level corn yields. *Environmental and Ecological Statistics*, 21(3):507-530.

Banff Support Team. (2008). Functional description of the Banff system for edit and imputation - version 2.03, Statistics Canada, Canada.

Fellegi, I.P., Holt, D. (1976). A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association,* 71: 17–35.

Barboza, W., Miller, D., & Cruze, N. (2014). Assessing the impact of a new methodology for the Agricultural Resource Management Survey, UNECE Work Session on Statistical Data Editing.

Robbins, M.W., Ghosh, S.K., & Habiger, J.D. (2013). Imputation in high dimensional economic data as applied to the Agricultural Resource Management Survey, *Journal of the American Statistical Association*, 108: 81-95.