# Use of Significance Editing for Agricultural Surveys

Wendy J. Barboza[1,2], James M. Harris[1]
[1] National Agricultural Statistics Service, Fairfax, VA, USA
[2] Corresponding author: Wendy Barboza, email: wendy.barboza@nass.usda.gov

## Abstract

The National Agricultural Statistics Service (NASS) is a statistical agency within the U.S. Department of Agriculture (USDA) that conducts hundreds of surveys every year and prepares reports covering virtually every facet of U.S. agriculture. NASS's traditional approach has been to manually fix edit failures for their surveys. As staff resources have become more constrained, the agency is attempting to embrace technological advances. NASS is currently evaluating Statistics Canada's Banff software to perform the editing and imputation for their agricultural surveys. A manual review of "large" data changes would then be performed after the survey is processed through this automated procedure. Significance editing can be used to prioritize the manual review of records after the editing/imputation process is complete. The purpose of significance editing is to identify records with imputed values that have a significant impact on the total survey estimate and to manually review these records to ensure the integrity of the imputed data. A record-level score is assigned based on changes between the original and edited/imputed data and records with the "largest scores" are identified for manual review. For all records, a score is first calculated for each item on the questionnaire based on the weighted absolute difference of the original and edited/imputed values divided by the estimated total. The record's maximum item-level score is then assigned as the record-level score. Records with scores above a pre-specified threshold value are manually reviewed by an analyst. This paper discusses the research initiative to incorporate the automated editing/imputation procedure and significance editing into the agency's surveys.

Key Words: Banff, manual review, NASS, record-level score, significance editing

## 1. Introduction

The National Agricultural Statistics Service (NASS) is a statistical agency located under the United States Department of Agriculture (USDA). NASS's mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. In order to successfully accomplish the agency's mission, NASS conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Some examples of areas covered in 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 changes in the demographics of U.S. producers.

The census of agriculture is the largest data collection effort performed by NASS. In 1997, responsibility for conducting the agricultural census was transferred from the Bureau of the Census, United States Department of Commerce, 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 has been to manually fix edit failures for their 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. This endeavor was the first step at changing the agency's culture away from manually fixing the edit failures.

Unfortunately, the editing and imputation processing system used for the census of agriculture is not easily portable to NASS's surveys. However, as staff resources have been more constrained, the agency is attempting to embrace technological advances. NASS is currently evaluating Statistics Canada's Banff software to perform the editing and imputation for their agricultural surveys. To suffice the agency's cultural attitudes, a manual review of "large" data changes would then be performed after the survey is processed through this automated procedure. Significance editing can be used to prioritize the manual review of records. The purpose of significance editing is to identify records with imputed values that have a significant impact on the total survey estimate and to manually review these records to ensure the integrity of the imputed data. This paper discusses the research initiative to incorporate the automated statistical editing and imputation procedure as well as significance editing into the agency's surveys.

## 2. Banff Software for Editing and Imputation

NASS is currently evaluating Banff software to perform the editing and imputation for surveys. Banff is a system developed by Statistics Canada that consists of a collection of specialized SAS procedures (Banff Support Team 2008). It performs automated statistical edits using Fellegi-Holt methodology, carries out imputation using different methodologies, and identifies outliers in the data. Banff requires the edits be expressed in linear form and it assumes the survey data are numeric and continuous. In most SAS procedures, negative data can be accepted or rejected as invalid.

The SAS procedures in Banff can be used independently or put together in order to satisfy the edit and imputation requirements of the survey data. This independence provides the user with a great deal of flexibility, but also entails more responsibility in ensuring that the inputs are of good quality and the outputs are interpreted and applied correctly. In Banff, each of the procedures accepts independent inputs provided by either the user or another Banff procedure. In the case of inputs being supplied by the user from outside the system, the user has the responsibility of guaranteeing the quality of the input since Banff will attempt to process whatever it is provided. In addition, each of the procedures provides its own unique outputs. The data records output from Banff procedures contain only those data which have been changed from the input data. Thus, the user has the responsibility of incorporating these changes into the original dataset.

Similar to regular SAS procedures, Banff procedures are able to process data in BY groups. To explain further, rather than process separate datasets for each individual group, a user may include all groups in a single dataset and Banff will process each of these groups independently according to the BY variable which identifies the groups.

## 3. Overview of Fellegi-Holt Methodology

The concept of Fellegi-Holt methodology was introduced in a paper published in the Journal of the American Statistical Association (Felligi and Holt 1976). There were three main criteria that the authors focused on, with the first being of primary importance.

(1) The data in each record should satisfy all edits by changing the fewest reported data values. This philosophy supports the idea of keeping the most amount of original

data as possible while satisfying all of the edit constraints (i.e., generate as little data as possible).

(2) The imputation rules should be automatically derived from the edit rules, which ensures the imputed questionnaire items will satisfy all of the edits. In addition, this simplifies the specification and implementation processes.

(3) As much as possible, the imputed values should maintain the marginal, and preferably the joint, frequency distributions of the questionnaires items of the records that satisfy all of the edit constraints.

The authors define logical edits as edits involving qualitative (coded) data that are not subject to a meaningful metric. With this methodology, the logical edits need to be expressed in normal form. Since logical edits are typically created by subject matter experts in the form of a description, flowchart, or decision logic table, this information needs to be translated into normal form.

## 4. Automated Statistical Data Editing and Imputation

As stated earlier, NASS is researching Banff to perform the automated linear edits using Fellegi-Holt methodology, 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., county of residence).

Automated statistical data editing and imputation refers to automatically changing reported data values that do not meet specified criteria and ensuring missing data values are filled. After this process is performed, a record can be classified as either "clean" or "dirty". If the data values within the record pass all of the edits, the record is clean; if any value fails an edit after the automated procedure is complete, the record is dirty. Clean records are eligible for the donor imputation process if such an imputation technique is utilized. However, if desired, clean records that are identified as outliers can be excluded from the donor imputation process. Dirty records need to be manually fixed by an analyst since the automated procedure cannot find a feasible solution. After the manual review, the record is passed through the automated procedure again.

Within Banff, NASS is utilizing multiple options for performing the imputation. By employing several alternatives, it increases the probability of obtaining a clean record. The ordering of the alternatives depends on the survey and is specified by the subject matter experts. Here, the ordering for NASS's quarterly hog survey is described. First, deterministic imputation is used to determine if there is only one possible value which would satisfy the original edits. If so, the value is imputed. Donor imputation is then evaluated to see if there is a nearest neighbor available to provide current data that will allow the record to pass the edits. This procedure requires a minimum number of donors. Next, since the hog survey is performed on a quarterly basis, 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 is classified as dirty.

## 5. Significance Editing

Records satisfying all of the edits and classified as clean are eligible for significance editing. The purpose of significance editing is to identify records with imputed valued that have a significant impact on the total survey estimates and to manually review these records to ensure the integrity of the imputed data. During significance editing, a manual review of "large" data changes is performed to validate large changes made by the automated edit/imputation procedure. For the manual review process, records can be prioritized by the magnitude of the change. A record-level score is assigned based on changes between the original and edited/imputed data and scores for the records are sorted in descending order. A score is first calculated for each item on the questionnaire and the record's maximum item-level score is then assigned as the record-level score. Records with scores above a pre-specified threshold value are then manually reviewed by an analyst.

Again, the record-level score is only calculated for records that are clean. NASS's significance editing process is somewhat unique in that the difference between the original value and the edited/imputed value is utilized to calculate the record-level score. First, an item-level score is calculated for specified questionnaire items based on the weighted absolute difference of the original and edited/imputed values divided by the estimated total. Then, the record's maximum item-level score is used to identify the most influential records to review. In order to specify the formula for calculating the record-level score, some notation is necessary. Let $x_{oi}(t)$ be the record's original response for item i at time t and $x_{ei}(t)$ be the record's edited/imputed response for item i at time t. The absolute difference $d_i = | x_{oi}(t)- x_{ei}(t) |$ is first calculated for all specified items. Since the total survey estimate from time t is unknown at this point, information at time t-1 is utilized to approximate the record's impact on the total survey estimate. The record's weight at time t, denoted w(t), is multiplied by the absolute difference, or $d_i$, and then divided by the total survey estimate for item i at time t-1, denoted $T_i(t-1)$. The record-level score is then the maximum of the item-level scores. In other words, the record level score is equal to $\max[(w(t) *d_i(t))/T_i(t-1)]$. In literature, the record's weight at time t-1, denoted w(t-1), may be used instead of w(t). For reasons beyond the scope of this paper, NASS is utilizing w(t) to calculate the record-level score.

The pre-specified threshold value is subject to debate and still being researched. NASS's research with the quarterly hog survey has suggested a threshold level of 20 percent. One could argue using a threshold level of 50 percent, which is supported by the statistical literature. However, literature also states that the optimal threshold level varies by survey depending on the subject matter. Regardless, the 50 percent cutoff would be advantageous to NASS since it is much lower than the traditional approach of manually reviewing all edit failures. Furthermore, with the automated procedure and significance editing, the analyst is focused on manually reviewing records with imputed values that have a significant impact on the total survey estimates, rather than all records.

## 6. Comparison of Automated Procedure to Manual Edit

NASS is in the process of evaluating the automated statistical data editing and imputation using the Windows XP version of Banff. In the production environment, the programs will be migrated to the UNIX platform. The Hog Survey was the first survey chosen to conduct the research. This survey provides detailed inventory of breeding and marketing hogs and the future supply of market hogs. The Hog Survey is performed on a quarterly basis (December, March, June, and September); December is the base month and performed in all states and the survey is conducted in the 29 most important hog producing states for the remaining months. Original survey data (i.e., prior to an analyst manually reviewing the records) was captured in order to investigate the performance of Banff. In addition, the current edits used for the survey were programmed as linear edits in Banff and the imputation methodology was specified. The original survey data was processed using Banff and then compared to the manually edited survey results. This research has been conducted for all states since December 2011 and the results have been favorable. The macro-level results were not significantly different for a majority of the questionnaire items and the micro-level results were comparable for the most part.

Table 1 contains a modified example (actual record-level data are not shown due to confidentiality) that shows the original data value, the value after the automated statistical data editing/imputation, and the value after the manual edit by an analyst. In this example, the total does not equal all of the sub-categories and the automated procedure and the analyst corrected this error in the same way. This correction is categorized as deterministic. The corresponding linear edit is breeding sows + breeding boars + market hogs for all weight categories (under 60 pounds, 60-119 pounds, 120-179 pounds, and over 180 pounds) = total hogs owned.

Table 1: Similar Deterministic Changes - Automated Procedure Versus Manual Edit

| Item Description | Original Data | Automated Procedure | Manual Edit |
|---|---|---|---|
| Breeding Sows | 1,800 | 1,800 | 1,800 |
| Breeding Boars | 0 | 0 | 0 |
| Market Hogs < 60 | 5,400 | 5,400 | 5,400 |
| Market Hogs 60-119 | 2,200 | 2,200 | 2,200 |
| Market Hogs 120-179 | 2,000 | 2,000 | 2,000 |
| Market Hogs 180+ | 2,100 | 2,100 | 2,100 |
| Total Hogs Owned | 11,700 | 13,500 | 13,500 |

Table 2 contains a similar example as above. However, in this example, the automated procedure changed the total to equal the sum of the sub-categories, whereas the analyst changed one of the sub-categories rather than the total. An advantage that the analyst has over the automated procedure is that a questionnaire can be reviewed for notes if any exist on the paper questionnaire or are captured electronically. It should be noted that the automated procedure is flexible and can be programmed based on criteria specified by the user. For example, the user can associate weights with various questionnaire items, which make it more or less likely that an item will be changed. In this example, it is likely that the analyst's change is correct but the item-level change made by the automated procedure should result in a large record-level score, which means this record would be manually reviewed during significance editing.

Table 2: Dissimilar Deterministic Changes - Automated Procedure Versus Manual Edit

| Item Description | Original Data | Automated Procedure | Manual Edit |
|---|---|---|---|
| Breeding Sows | 55,000 | 55,000 | 55,000 |
| Breeding Boars | 500 | 500 | 500 |
| Market Hogs < 60 | 120,000 | 120,000 | 120,000 |
| Market Hogs 60-119 | 45,000 | 45,000 | 45,000 |
| Market Hogs 120-179 | 0 | 0 | 45,000 |
| Market Hogs 180+ | 45,000 | 45,000 | 45,000 |
| Total Hogs Owned | 310,500 | 265,500 | 310,500 |

Table 3 provides an example where donor imputation was used to satisfy the linear edits. In this example, both methods made similar changes. Death loss refers to the number of hogs owned by the operation that died. The edit specifies that death loss cannot be equal to zero and is within a specific range of the total hogs owned. The linear edits are death loss <= 0.2 x total hogs owned and death loss >= 0.005 x total hogs owned.

Table 3: Similar Imputed Changes - Automated Procedure Versus Manual Edit

| Item Description | Original Data | Automated Procedure | Manual Edit |
|---|---|---|---|
| Total Hogs Owned | 50,000 | 50,000 | 50,000 |
| Death Loss | 0 | 5,810 | 6,350 |

One important lesson learned while researching the automated statistical data editing/imputation on the quarterly hog survey has to do with keying errors. The automated procedure does not perform well when a particular data value is miskeyed. However, during the testing, these records received a large record-level score and would be identified and fixed during significance editing.

## 7. Conclusion

By using the automated statistical data editing/imputation and significance editing, NASS expects large gains with respect to time, costs, and quality. Since significance editing focuses on certain records, considerable time will be saved during the manual review. Staff resources are better utilized, which results in substantial cost savings. The automated procedure will correct records consistently and improve the quality of the results. Significance editing also provides a "safety net" since analysts verify large changes to the survey data that were made by the automated procedure.

## References

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, pg.17–35.