Utilizing Automated Statistical Edit Changes in Significance Editing

Wendy Barboza[1], Kay Turner[1]
[1]USDA, NASS, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030

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 perform a manual edit and review of all questionnaires for most surveys. As staff resources become more constrained, the agency has embraced technological advances. The goal of significance editing, defined as statistical data editing; selective editing; and outlier detection, is to (1) reduce the time and effort spent manually reviewing/correcting survey questionnaires, without damaging the quality of the resulting data, and (2) focus the manual effort on the accuracy of the survey questionnaires that strongly impact the overall results. During the survey process, the most influential records are identified by calculating a unit-level score based on the changes made by the automated statistical data edit. This paper provides details on these unit scores as well as the implementation of the significance editing concepts.

Key Words: significance editing, statistical data editing, selective editing, outlier detection, NASS

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 publishes numerous reports covering virtually every aspect of U.S. agriculture. Although most of the reports are published by personnel at NASS's Headquarters which is located in Washington, DC, the agency's 46 Field Offices also publish reports that target the specific interests of their local audiences. 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 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 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.

The census of agriculture was previously conducted by the Bureau of the Census, United States Department of Commerce. In 1997, responsibility for conducting the agricultural census was transferred 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 almost

3 million records. Historically, NASS’s traditional approach to processing a survey was to perform a manual edit and review of all questionnaires for most surveys. The agency quickly realized a paradigm shift was necessary in order to process the census of agriculture in a timely manner. New strategies were utilized to perform the edit and imputation and identify records that needed to be manually reviewed. This endeavor was the first step at changing the agency’s culture.

In the past few years, staff resources have been more constrained and the agency has been researching ways to improve the editing/imputation methodology used for surveys while satisfying the cultural attitudes. NASS is investigating significance editing to (1) reduce the time and effort spent manually reviewing/correcting survey questionnaires, without damaging the quality of the resulting data, and (2) focus the manual effort on the accuracy of the survey questionnaires that strongly impact the overall results. This endeavor is supported by the fact that editing too much can have a negative effect on the survey results (reference [1]). This paper discusses the research initiative to incorporate significance editing concepts into the agency’s surveys.

2. Banff Software for Edit and Imputation

NASS is currently evaluating Banff software to perform the edit and imputation for surveys. Banff is a system developed by Statistics Canada that consists of a collection of specialized SAS procedures. It performs automated edits using Fellegi-Holt methodology (reference [2]), 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 a survey. 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 their original data (reference [3]).

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. Significance Editing

Significance editing is defined as statistical data editing, selective editing, and outlier detection. As stated earlier, the goal of significance editing is to (1) reduce the time and effort spent manually reviewing/correcting survey questionnaires, without damaging the quality of the resulting data, and (2) focus the manual effort on the accuracy of the survey questionnaires that strongly impact the overall results. NASS is currently evaluating Banff to perform the statistical data edit and imputation for surveys performed by the agency. After the statistical data editing phase, selective editing identifies the records to be manually reviewed by an analyst. In addition, outliers are identified using two methodologies and these records are also marked for manual review by an analyst. This approach reduces the number of records to be manually reviewed by an analyst while satisfying the cultural attitude to perform a manual edit and review of all survey questionnaires.

Note that the significance editing process outlined in this paper is geared towards recurring surveys because it uses previous survey data. This statement is not being made to suggest that significance editing cannot be performed for one-time surveys. The point is that significance editing is different for recurring surveys. In this paper, all three phases utilize previous survey data. For a one-time survey or a new recurring survey, similar data could be used in lieu of previous survey data. For a new recurring survey, the survey could also be conducted once and then updated after previous survey data are available. Most of the surveys at NASS are performed on at least an annual basis, with the exception of the census of agriculture which is performed every five years. The expectation is that the significance editing process would perform better for surveys conducted more frequently. Therefore, significance editing should yield better results for a survey conducted on a quarterly basis, rather than an annual survey, since the previous survey data are more current.

3.1 Statistical Data Editing

The term statistical data editing refers to automatically changing reported data values that do not meet specified edit checks and imputing missing data values. After the statistical data editing phase, a record is classified as either clean or dirty. If all values within the record pass all of the editing criteria, the record is clean; if any value does not pass the editing criteria, the record is dirty. Clean records do not need to be manually edited and are eligible for the donor imputation process if such an imputation technique is utilized. However, clean records that are identified as outliers are excluded from the donor imputation process (see III.C. for more information). Dirty records need to be manually fixed by an analyst since the automated data edit cannot find a feasible solution.

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. This methodology preserves as much of the reported data as possible. 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 so that the record passes 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).

For the imputation phase, NASS is utilizing several alternatives for performing automated imputation in Banff. By employing several alternatives, it increases the chance of ending up with a clean record. Deterministic imputation is used first to determine if there is only one possible value which would satisfy the original edits. If so, the value is imputed. The order of the next three methods depends on the survey. Donor imputation is 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, 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 an edit, the changed values are returned to their original, unedited value. When any record does not satisfy all of the editing criteria, it is defined to be a dirty record and flagged to be manually reviewed by an analyst. Records satisfying all of the edits are identified as clean records and eligible for selective editing.

3.2 Selective Editing

The selective editing process applies only to records that are clean after the statistical data editing phase is performed. The purpose of selective editing is to identify records that have a significant impact on the total survey estimates and to manually review these records to ensure the integrity of the data. To accomplish this process, a record-level score is assigned to every clean record, the records are sorted by their score, and all records above the 50th percentile are marked for manual review by an analyst. NASS's selective editing process is unique in that the difference between the original value and the Banff edited/imputed value is utilized to calculate the record-level score. Thus, records with "large" statistical edit changes (i.e., records above the 50th percentile) are manually reviewed to ensure the automated changes are acceptable. Using this approach, edit changes to records below the 50th percentile are considered to be of high quality.

The threshold level of 50% is somewhat arbitrary but supported by the statistical literature on selective editing. The optimal threshold level is probably much higher than 50%, but it is clear that the best threshold level also varies by survey depending on the subject matter. Regardless, the 50% cutoff is advantageous to NASS since it is much lower than the cultural attitude of performing a 100% manual review for most surveys. With the selective editing approach, an analyst is focused on manually reviewing records with "large" statistical edit changes that also have a significant impact on the total survey estimates.

Again, the record-level score is only calculated for records that are clean. 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. The record's maximum item-level score is then 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-1, denoted w(t-1), 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 $\max[(w(t-1) *d_i(t))/T_i(t-1)]$.

3.3 Outlier Detection

Outliers are identified using two methodologies. The first method focuses on the clean record's data at time t and the second method uses the H-B score[1], which compares the clean record's data at both time t and time t-1. For the first method, a record is identified as an outlier if any of the items for the record are extremely large relative to the corresponding items for other records. In addition to being marked for manual review by an analyst, these records are also excluded from the donor imputation process. For the second method, outliers are identified based on how much the record changed over time. An extreme positive or negative H-B score means that there is a potential for the record to have a significant impact on the total survey estimate. Records above or below a specified percentile are marked for manual review by an analyst and the most extreme records are also excluded from the donor imputation process.

The H-B score is only calculated for clean records that have responded to the current survey (i.e., time t) and a previous survey (i.e., time t-1). In order to specify the formula for calculating the H-B score, some notation is necessary. Let $x_i(t)$ be the record's response after the statistical data edit for item i at time t and $x_i(t-1)$ be the record's response after the statistical data edit for item i at time t-1. For each item where $x_i(t) > 0$ and $x_i(t-1) > 0$, the ratio $r_i = x_i(t)/x_i(t-1)$ is first calculated for all items and the median ratio r_{Mi} is then calculated across all eligible records. The ratios are then transformed so the difference between $x_i(t)$ and $x_i(t-1)$ are the same on either side of the median difference. In other words, define the size, denoted s_i, as $s_i = 1- r_{Mi}/r_i$ when $0< r_i < r_{Mi}$ or $s_i = r_i/r_{Mi} -1$ when $r_i \geq r_{Mi}$. The H-B score is then calculated as $s_i[\max(x_i(t),x_i(t-1))]^{exp}$ where exp is between 0 and 1. An exponent of 0 treats all relative differences the same, regardless of the size, while an exponent of 1 gives greater importance to deviations of larger units. NASS is using a value of 1 for exp.

By using H-B scores for items of interest, the idea is to identify problem records that would not be marked for review by other procedures previously discussed. The expectation is to identify large-sized records with a significant change over time and median-sized records with a significant change over time. Small-sized records with a significant change or small changes over time for records of any size should not have extreme positive or negative H-B scores.

[1] The H-B score was developed by Mike Hidiroglou and Jean-Marie Berthelot who work for Statistics Canada. The methodology discussed here is based on their work, which is documented in reference [3].

4. Example of Automated Statistical Editing

NASS is testing automated statistical editing using the Windows XP version of Banff. The Hog Survey was selected to conduct the testing. 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. The testing was conducted using original survey data for several months in a couple of the top hog producing states. The current edits used for the survey were programmed as linear edits in Banff and the imputation methodologies were specified. The results from processing the data using Banff were then compared to the manually edited survey results. 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 edit, 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 edit and the analyst corrected this error in the same way. This correction is categorized as deterministic. The corresponding linear edit is sows and gilts for breeding + boars and young males for breeding + hogs and pigs for market and home use by the weight categories under 60 pounds, 60-119 pounds, 120-179 pounds, and over 180 pounds = total hogs and pigs owned by the operation.

Table 1: Similar Deterministic Changes Made by the Automated and Manual Edits

Item Description	Original Data	Automated Edit Value	Manual Edit Value
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 edit changed the total to equal the sum of the sub-categories, whereas the analyst changed one of the sub-categories so the sub-categories sum to the total. An advantage that the analyst has over the automated edit 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 edit 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 edit should result in a large record-level score, which means this record would be manually reviewed during the selective editing phase.

Table 2: Dissimilar Deterministic Changes Made by the Automated and Manual Edits

Item Description	Original Data	Automated Edit Value	Manual Edit Value
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, the automated edit and analyst made similar changes. Death loss refers to the number of weaned and older pigs owned by the operation that died. It is assumed that the death loss for a hog operation cannot be equal to zero and is within a specific range of the percentage of total hogs and pigs owned by the operation. 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 Made by the Automated and Manual Edits

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

5. Conclusion

By using significance editing, NASS expects large gains with respect to time, costs, and quality. Since selective editing focuses on certain records, significant time will be saved during the manual review process. Staff resources are better utilized, which results in considerable cost savings. The automated statistical edit will correct records consistently and improve the quality of the results, in addition to mitigating problems related to over-editing survey data. Since the edited value is integrated into the record-level score for selective editing, it provides a safety net so that analysts can review large changes. The two outlier detection procedures also add another safeguard to catch extreme values from current data and identify large changes between current and previous survey data.

References

[1] Granquist, Leopold and Kovar, John G. Editing of Survey Data: How Much Is Enough? Survey Management and Process Quality (1997).

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

[3] Statistics Canada’s Banff Support Team, Functional Description of the Banff System for Edit and Imputation, Version 2.03, July 2008.