WP.25
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iv): Micro editing – methods and software

Innovative Use of Significance Editing in USDA's National Agricultural Statistics Service

Invited Paper

Prepared by Kay Turner, and Wendy Barboza, National Agricultural Statistics Service,
United States Department of Agriculture[1]

I. Introduction

1. The National Agricultural Statistics Service (NASS) is a statistical agency located under the United States Department of Agriculture (USDA). NASS' 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' Headquarters (HQ) which is located in Washington, DC, the agency's 46 Field Offices (FOs) also publish reports that target the specific interests of their local audiences. Some examples of areas covered in NASS' 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. In order to publish these reports, the size of the target population varies from fewer than 50 for a survey to nearly 3 million for the census of agriculture.

2. In order to understand the status of editing at NASS and the issues facing the agency as it plans for the future, it is important to be familiar with the physical structure of the agency. There are approximately 400 HQ employees. HQ staff are responsible for the overall survey methodology and processing systems including specifying the survey design, selecting the sample, creating the questionnaire, developing automated instruments for both data collection and editing/imputation, detailing and programming the estimation procedures, preparing manuals,

[1] The authors would like to thank Joe Parsons, who also works at the National Agricultural Statistics Service, for providing editorial comments on this paper.

conducting training (when necessary) for staff in the FOs, and providing support to FO staff during the survey proper. There are about 700 employees located in the 46 FOs. Most FO responsibilities focus on one state, but the New England FO is responsible for multiple states. FO staff are responsible for the day-to-day activities involved in conducting the surveys including training the telephone and field interviewers, collecting and editing the survey data, and submitting recommendations[2] to HQ for further review. HQ staff then analyze these recommendations along with the national-level estimate, and publish the final results after possibly revising the FO recommendations. In summary, HQ oversees the survey process while the FOs implement the survey proper.

3. 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 survey conducted by NASS changed from approximately 60,000 records to almost 3 million records. Historically, NASS' traditional approach to processing a survey was to perform a manual edit 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 identify the records that needed to be manually reviewed. This endeavor was the first step at changing the agency's culture.

4. 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.

II. BANFF software for edit and imputation

5. NASS is currently evaluating Banff software for edit and imputation, which is a system developed by Statistics Canada that consists of a collection of specialized SAS procedures. The agency is researching Banff to perform automated edits using Fellegi-Holt methodology (reference [2]), implement automated imputation using different alternatives, and identify outliers. Banff edits must 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. Prior to implementing Banff, it is assumed that some preliminary editing has been done during the data capture stage and respondent follow-up is complete.

6. 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

[2] Recommendations are submitted by FOs because published numbers are typically based on multiple sources, not just the survey indications. For example, more than one survey may be conducted or administrative data may be used in conjunction with the survey results in producing official published estimates.

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]).

7. 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.

III. Significance editing

8. 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 the FOs. In addition, outliers are identified using two methodologies and these records are also marked for manual review by the FOs. This approach reduces the number of records to be manually reviewed by the FOs while satisfying the cultural attitude to perform a manual edit review of all survey questionnaires.

9. Note that the significance editing process outlined in this paper is geared towards recurring surveys. 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. The selective editing process outlined below is only valid for recurring surveys because it uses previously reported data. In addition, previously reported data are being utilized during the statistical data editing phase. For a one-time survey or a new recurring survey, similar data could be used in lieu of previously reported data. For a new recurring survey, the survey could be conducted without selective editing and then updated once previously reported 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 previously reported data are more current.

A. Statistical Data Editing

10. 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 fixed by hand since the automated data edit cannot find a feasible solution.

11. 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 some questionnaire items are not a good candidate for Banff (e.g., county of residence). Also, Banff assumes the survey data are numeric and continuous. Thus, NASS is utilizing another system to edit the questionnaire items that cannot be edited using Banff.

12. 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. 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. Second, an imputation is attempted by using the record's previously reported data and applying an estimator function to impute the current value. This methodology is restricted to certain variables. Third, 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. Fourth, 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.

13. After imputation, the error localization procedure is run again to ensure the unchanged values and the newly imputed values passed 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. Records satisfying all of the edits are identified as clean records and eligible for selective editing.

B. Selective Editing

14. 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 the FOs. NASS' 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 were acceptable. Using

this approach, edit changes to records below the 50th percentile are considered to be of high quality.

15. 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, the FOs are focused on manually reviewing records with "large" statistical edit changes that also have a significant impact on the total survey estimates.

16. The record-level score is only calculated for records that are clean. An item-level score is calculated for each questionnaire item 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 $x_o(t)$ be the record's original response before the statistical data edit for item i at time t and $x_i(t)$ be the record's current response (i.e., after the statistical data edit) for item i at time t. For each item where $x_o(t) \neq 0$ and $x_i(t) \neq 0$, the absolute difference $d_i = | x_o(t)- x_i(t) |$ is first calculated for all items. Since the total survey estimate at time t is unknown at this point, information from time t-1 is utilized to approximate the record's impact on the total survey estimate. The record's weight from time t-1, denoted $w_i(t-1)$, is multiplied by the absolute difference, or d_i, and then divided by the total survey estimate from 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 $\max[(w_i(t-1) *d_i(t))/T_i(t-1)]$. Note that this implies that an item-level score (and thus a record-level score) can only be calculated for clean records that responded at both time t and time t-1.

C. Outlier Detection

17. 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[3], 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 the FOs, 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 score are marked for manual review by the FOs and the most extreme records are also excluded from the donor imputation process.

18. 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_M is then calculated across all eligible records.

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

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

$s_i = 1 - r_M/r_i$ when $0 < r_i < r_M$ or

$s_i = r_i/r_M - 1$ when $r_i \geq r_M$.

The effect of the record on the item of interest (i.e., 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 small deviations of large units. NASS is using a value of 1 for exp.

19. 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.

IV. Example of automated statistical editing

20. NASS' Research and Development Division (RDD) is testing automated statistical editing in a Windows XP operating system with a Windows XP version of Banff. In order to conduct this testing, RDD had to acquire survey data prior to the records being manually reviewed by the FOs (i.e., the raw data). In March 2009 and June 2009, both raw and manually edited data were obtained in two important hog producing states (Minnesota and Nebraska) for the hog survey, which is performed on a quarterly basis. An automated edit was programmed in Banff to perform linear edits and make imputation attempts in the following order of precedence: 1) deterministic, 2) imputation using previously reported data, 3) donor imputation using current data, and 4) mean imputation. The raw hog survey data from March 2009 and June 2009 were run through the automated edit and compared to the manually edited data. The results were not significantly different for a majority of questionnaire items.

21. Table 1 contains a modified example (actual record-level data are not shown due to confidentiality) that shows the raw data value, the value after the automated statistical edit, and the manually edited value made by the FOs. In this example, the total does not equal all of the sub-categories. The automated edit and the FO 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	Raw 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

22. 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 FO 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.

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

Item Description	Raw 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

23. Table 3 provides an example where donor imputation was used to satisfy the linear edits. In this example, the automated edit and FO 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. It should be noted that a very large item-level change made by the automated edit may result in a large record-level score, which means this record may be flagged for review during the selective editing phase.

Table 3: Similar Imputed Changes Made by the Automated and Manual Edits

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

V. Future direction and outstanding issues

24. NASS is currently in the process of moving its surveys to a centralized processing environment. The quarterly hog survey is scheduled to be migrated this year. Once the survey is migrated and immediately after the survey proper, the plan is to conduct a pseudo-operational test that implements the significance editing methodology. Given the outcome of this pseudo-operational test, further modifications and improvements will be made as needed and FOs will be phased into the new system. Then, the plan is to develop significance editing methodology for other surveys depending on when they are migrated to the centralized environment. To successfully implement this plan, a crucial issue is that all testing has been in the Windows XP operating environment, but the centralized processing environment will be in AIX UNIX 6.1. This version of Banff is not scheduled to be released by Statistics Canada until June 2011. Once

NASS obtains this software, the significance editing programs will first need to be updated and tested on the UNIX system prior to the pseudo-operational test.

25. Further work and testing is needed to move the significance editing programs from the development phase into production. The input and output files and corresponding variables need to be coordinated. Another editing system is being used in conjunction with Banff to identify errors that are unresolved by Banff and to edit the questionnaire items that cannot be edited using Banff. Various flags will need to be set and coordinated between the two edit systems to ensure that the process is working properly for all possibilities. During this process, the output data from significance editing will need to be reviewed and evaluated for the automated statistical edits, selective editing, and outlier detection. The centralized environment will make it possible to track the history of a record, which will be helpful in trouble-shooting and making improvements. Initial processing will be at the state-level; however, with centralization, the capability of editing at the regional-level or national-level is possible.

26. NASS has a tight time frame between the end of data collection and the publication date. Due to this time constraint, FO analysts currently edit records during the survey proper, rather than waiting until the end of data collection. The plan is to process records through the significance editing system during data collection. After a minimum number of records are available for use in a donor pool, the records will be processed through the significance editing system. With respect to selective editing, there are two methods for identifying the threshold level and processing the records. The first method is to use a record-level threshold value and to process the records on a daily basis, once the donor pool is available. However, previous survey raw data are needed in order to calculate this threshold value. If a threshold value is available, a score is calculated for each clean record and the record-level score is compared to the calculated threshold value. Records with a record-level score above the calculated threshold value are marked for manual review by the FOs and those below the calculated threshold value are finished with being processed. When previous survey raw data are not available and therefore a calculated threshold value is not available, the second method is to use a 50 percent threshold value and to process the records in batches. The records are not processed until a minimum number of records are obtained for a batch. Within the batch, the record-level scores are calculated for clean records and the records are sorted by score. The top half of the records with the largest record-level scores, which represents 50 percent of the batch, are marked for manual review by the FOs, while the lower half of the records are finished with being processed.

27. A selective editing threshold of 50 percent will be used initially. Currently, raw data are not captured for surveys, but it will be once the survey is migrated to the centralized processing environment. However, since raw data were captured for the hog survey for multiple quarters, the actual threshold value could be calculated for this survey. Regardless of which threshold method is used, the plan is to monitor the selective editing threshold by survey and adjust it upward when possible, so that, in the future, fewer records are flagged to be manually reviewed.

28. With respect to the H-B score, the plan is to exclude records from the donor imputation process that are below the 1% and above the 99% cut-off levels. Records below the 5% and above the 95% cut-off levels will be identified as outliers and marked for manual review by the FOs. These cutoff values, as well as the value of the exponent (i.e., exp), will be monitored and evaluated over time. It is possible to adjust these by survey if needed. For example, if too many or too few outliers are being flagged, then the cut-off levels can be updated accordingly.

29. By using significance editing, NASS expects large gains with respect to time, costs, and quality. By focusing the editing process on records that impact the overall survey results, staff

resources will be utilized much better. Furthermore, ensuring that records are processed consistently will also improve the quality of the results. The statistical edit will automatically correct records consistently and quickly, 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.