Banff Automated Edit and Imputation Applied to the US Hog Inventory Survey

James M. Johanson
United States Department of Agriculture, National Agricultural Statistics Service
1400 Independence Avenue SW, Washington, DC 20250

Introduction

The National Agricultural Statistics Service (NASS) is a statistical agency located under the U.S. Department of Agriculture (USDA). The mission of NASS is to provide timely, accurate, and useful statistics in service to U.S. agriculture. To fulfill this mission, NASS conducts hundreds of surveys every year and publishes numerous reports covering virtually every aspect of U.S. agriculture.

Recently, NASS has been investigating ways to be more efficient, yet reduce costs. Edit review and imputation (E&I) have been identified as possible ways to increase efficiency. Traditionally, E&I within NASS have been labor intensive, especially for large diverse surveys. Manual review and updating via Blaise are the primary methods of E&I. Blaise provides the analyst a direct interface to the data, which is particularly useful when analysts touch many records, which is common in NASS. However, touching numerous records may not be necessary, or even beneficial. Additionally, E&I may not be uniform across this agency as various analysts review data.

The goals of this study are: 1) to reduce the amount of effort spent manually reviewing and updating survey questionnaires, without damaging the quality of the summarized estimates, 2) have a more consistent correction of errors found in questionnaires, and 3) to focus the manual effort on the accuracy of the survey questionnaires that strongly impact the overall estimates.

Analysis

NASS is evaluating the Banff System for Edit and Imputation, a Statistics Canada product (Banff Support Team, 2003). The edit portion of the system utilizes Fellegi-Holt (FH) methodology (Fellegi and Holt, 1976). The FH method requires linear edits to define a feasible region for potential values for imputation. In this study, edit equations were written to process commodity specific data, not administrative code data. For example, commodity data are the number of market hogs under 50 pounds or the number of hogs that died. While, administrative code data are the date, interviewer's identification number, or non-response codes.

The FH method attempts to satisfy all edit equations by changing the fewest possible values. The Banff system enables the user to use a variety of imputation methods. This study utilized deterministic imputation, donor imputation, mean imputation, and imputing with previously reported data. The Banff system interfaces directly with SAS software, which is commonly used within NASS.

Applied Banff to Hog Survey

NASS chose to begin the evaluation of Banff on the hog survey. The hog survey provides a detailed inventory of breeding and marketing hogs and the future supply of market hogs. The base month for the hog survey is each December where all states are included. In subsequent quarters through the year, only the top 29 hog inventory states are surveyed. This survey uses a stratified sample design, where the largest operations have a sample probability of one and smaller operations have a probability of less than one.

At this time, NASS does not routinely save original data, only final edited data. Therefore, original data were captured from the 29 quarterly states in December 2012. (The remaining annual states were analyzed as well, but the results are not shown in this report.) The quarterly states had a sample size of 8892. Of that sample, only 3011 had positive hog data. This report expands a previous study done on 2 states (Johanson, 2012) to a national level.

This data set was processed 4 ways resulting in four analyses. Each analysis gets more similar to the way we intend to implement this research in practice. The first analysis (ALL) processed all the data through Banff as one large batch at the conclusion of data collection. There was no additional review of the data after Banff imputation. The data were then sent to a summary that generated estimates.

The second analysis (BATCH) processed data as cumulative daily batches. Again, there was no additional review of the resulting data after Banff imputation. Then the data were summarized and estimates were compared. This analysis simulated how ongoing editing occurs in NASS during data collection. Due to the quick turnaround time between data collection and publication, NASS must begin editing before data collection is complete.

The third analysis (REVIEW) took the daily batched results from the second analysis, then used selective editing to review Banff changes in a results viewer. The results viewer was programmed in-house, it is not a part of the Banff system. An analyst reviewed the top 20% of the most influential Banff imputations within each state (or at least one in each state). The subsequent results of the review were summarized and estimates were compared.

The fourth analysis (CONTROL) served as the control data. After traditional editing in the production setting was complete, the clean data set was summarized and estimates compared. This is how NASS currently processes the hog survey.

Imputation Results

At the start of the analysis, Banff identified 889 records having at least one error, which yielded 1864 fields to impute.

The first round of imputation resulted in deterministic imputation changing 180 fields. Then donor imputation changed 1296 fields. At this point, error localization was run again. If records did not pass the edit equations, then all field in a record reverted back to its original values. In the second round of imputation, deterministic imputation changed 7 fields. Next Banff applied imputation using valid current observations. Often this was a mean imputation that resulted in changes to 219 fields. Again, error localization was run to verify that records pass edits. If a record didn't pass edits, it was returned to its original values. The third and final round of imputation yielded 6 changed fields via deterministic imputation. Then Banff imputed using the previous report from that same operation, which provided 123 changed fields. The total number of changes attempted to fields was 1831. Fields, which failed to pass edits in a previous round of imputation, underwent multiple rounds of imputation.

At the end of the analysis, only 72 records remained with errors and only 283 fields to impute. That is a reduction of 91.9% in the number of records with errors and a reduction of 84.8% in the number of fields to impute. Similar analyses have been processed on other quarters of data collection (not shown in this report). All other analyses have successfully reduced the number of fields to impute by more than 90%. It should be noted here that there were 10 questionnaires from the same interviewer where every question was refused. This resulted in 18 fields to impute per questionnaire. The interviewer should have coded the whole questionnaire as a refusal, which would not be processed by Banff. Therefore 180 of the 283 remaining field to impute are not true fields to impute. The true reduction is still greater than 90%. Any records remaining with errors were given the values from the clean data set from the fourth analysis (control data). This should make certain that differences found between Banff and control estimates were a result of Banff imputation and not confounded with other factors. Also, records that Banff cannot fix will be edited by the same process used to obtain the clean data.

Selective Editing Review

The third analysis utilized a selective editing review in order to allow analysts to locate and review the most influential changes made by Banff, selective editing has been incorporated (Silva et al, 2009). The Banff value was inserted into the following selective editing equation to obtain an item score:

$$item\ score = \frac{\text{weight} * |\text{original} - \text{Banff}|}{\text{Total}}$$

where original is the originally reported value and Banff is the value that Banff imputed. (If Banff didn't change the value, then the item score is zero.) We used the sampling weight as the weight (so a small operation may still have a large impact on the estimate due to the weight). Typically selective editing texts use a final weight, but we want to review during data collection instead of waiting until the end of data collection. Therefore, final weights are not yet available. Finally, the total is the estimated total from the previous quarter for the estimate of interest (i.e. total number of boars).

The above equation results in the item score for one variable. This calculation is repeated for all variables on the questionnaire. We were investigating 12 items of interest, so each report would have 12 item scores. There are several ways to calculate a unit score from the item scores. We chose to use the maximum item score as the unit score. The maximum unit score eliminates double counting that would occur with our data. If the item score on 'market hogs over 180 pounds' is large, then the item score for 'all market hogs' may be large due to the previous category's influence. Therefore, if a sum were used for the unit score, it would count 'market hogs over 180 pounds' twice. Now that we have one unit score from each report, we can sort the reports by largest impact.

We developed a small software application that allows the analyst to view the data. This 'viewer' displays four columns of data. Column 1 is originally reported data. Column 2 is the Banff version of the data. Column 3 is the most recent version of the data. Finally, column 4 is the previous quarter's data. If a value differs across any of the first three columns of data, that value is highlighted in yellow to draw the attention of the analyst. At this point, the analyst decides either to keep the Banff value or to update the value via Blaise. The benefit of this new review process is that the analyst can go directly to the edit changes with the largest impact on the total, instead of reviewing numerous changes with small impact. It also adds transparency to the automated imputation.

Comparison of Estimates

After Banff was run on the data, the resulting four data sets were summarized to obtain estimates. Twelve key estimates from the traditionally edited data and the 3 Banff edited data sets were compared. Recall, traditionally edited data used labor intensive review and updating via Blaise, where analysts touch many records during the editing process. Percent relative change and 95% confidence intervals (CI) were calculated comparing the 3 Banff edited data sets versus the control data set.

National Level Estimates

In the ALL analysis, 11 of the 12 national level estimates fell within the 95% CIs when compared to the CONTROL estimates. The boars for breeding estimate was the lone estimate that fell outside the 95% CI. In Table 1, the range for the percent relative difference was -0.02 to 0.38% for the 11 estimates. The percent relative difference of boars for breeding estimate was 24.62% higher than the CONTROL estimate.

Table 1. Percent Relative Difference (%) of 12 Key National Estimates

Variable Description	ALL	BATCH	REVIEW
Sows for breeding	0.05	0.04	0.01
Boars for breeding	24.62	25.22	2.42
Expected farrowings next quarter	-0.01	0.13	0.08
Expected farrowings in 2 quarters	0.11	-0.02	-0.06
Hogs under 50 pounds	-0.02	0.05	0.09
Hogs between 50-119 pounds	0.03	0.03	0.01
Hogs between 120-179 pounds	-0.02	0.01	0.01
Hogs over 180 pounds	0.18	0.17	0.12
Total hog inventory	0.08	0.10	0.06
Farrowings last quarter	0.35	0.41	0.24
Pigs born last quarter	0.32	0.27	0.21
Death loss	0.38	0.46	0.24

Similarly in the BATCH analysis, of the 12 estimates, the boars for breeding estimate was the lone estimate that fell outside the 95% CI. Aside from boars, the percent relative difference in Table 1 ranged from -0.02 to 0.46%, whereas the boar estimate was 25.22% higher than the CONTROL estimate.

Finally in the REVIEW analysis, all of the estimates fell within the 95% CIs. In Table 1, the percent relative difference ranged from -0.06 to 0.24%, except for boars for breeding, which was 2.42%. The boars for breeding estimate routinely has a larger coefficient of variation than the other estimates which explains how an increase of 2.42% of boars over the CONTROL estimate still falls within the 95% CI. The review process utilized in this analysis worked well at finding and correcting issues related to the boars for breeding estimate.

State Level Estimates

There were 29 states in the analysis and 12 key estimates compared in each state resulting in 348 estimates total. In the ALL analysis, 7 of the 348 CONTROL estimates did not fall within the ALL 95% CIs. Similarly, 9 of the 348 ALL estimates did not fall within the CONTROL 95% CIs. There were 8 occasions where the absolute percent relative change of the ALL estimates differed by more than 5% from the CONTROL estimates. Three of those 8 occasions were changes of greater than 10%.

In the BATCH analysis, 8 of the 348 CONTROL estimates did not fall within the BATCH 95% CIs. Whereas, 11 of the 348 BATCH estimates did not fall within the CONTROL 95% CIs. There were 8 occasions where the absolute percent relative change of the BATCH estimates differed by more than 5% from the CONTROL estimates. On only one occasion was the change greater than 10%.

Finally, in the REVIEW analysis, 8 of the 348 CONTROL estimates did not fall within the REVIEW 95% CIs. On the other hand, 10 of the 348 REVIEW estimates did not fall within the CONTROL 95% CIs. This analysis reduced down to 3 the number of occasions where the absolute percent relative change of the REVIEW estimates differed by more than 5% from the CONTROL estimates. On only one of those occasions did the change exceed 10%.

Table 2 shows the percent relative difference of the state estimates that were deemed problematic out of the total 348 state level estimates. The results of the confidence interval tests are only shown for the REVIEW analysis. In general, the REVIEW analysis tended to reduce the absolute percent relative difference in the shown estimates. In some of the instances where the REVIEW analysis did not reduce the absolute percent relative difference, the change was very small (less than 1% in the respective CONTROL estimate). Boars for breeding in Texas had the largest percent relative change at 23.23%. However, that difference is not statistically significant, since the respective estimates fell within the specified 95% CI. This is directly impacted by the fact that the coefficient of variation is large for the boars for breeding estimate. Also recall, the results of the national estimates identified the estimate of boars for breeding to be problematic. The REVIEW analysis was beneficial in reducing the percent relative difference to a reasonable level. However, that reduction was not associated with the problematic estimate in Texas, due to the fact that the estimate was the same in all three analyses.

Conclusions

Banff routinely reduces the number of records with errors by more than 90%. This shows strong potential for a reduction in manual effort required for editing, which addresses the first goal from the introduction. The fact that Banff treats every report the same demonstrates consistency in error correction, which relates to the second goal. The third and final goal is met when the selective editing methods used in this study provided the opportunity to review and ensure the quality of the data, as well as, directing the analyst to edit changes that have the largest impact on the estimated totals. So the analyst can quickly and efficiently clean up inconsistencies in the data. As a result high quality data can be delivered with less effort. Overall the estimates show little impact of the automated impuatation with few exceptions as described previously.These favorable results motivate implementation into the operational program.

Table 2. Percent Relative Difference (%) of Problematic State Estimates

State	Variable Description	ALL	BATCH	REVIEW
AL	Expected farrowings in 2 quarters	1.34	-1.34	-1.34[#]
CO	Pigs born last quarter	0.37	0.32	0.32[^]
CO	Farrowings last quarter	0.38	0.30	0.30[^]
IL	Death loss	3.41	3.66	3.28[*]
MI	Hogs over 180 pounds	7.11	7.11	7.11[^]
MI	Death loss	10.65	5.03	5.03[*]
MN	Death loss	-2.18	-2.52	-2.52[*]
MO	Death loss	0.47	0.80	0.80[*]
MT	Death loss	4.67	3.63	3.63[^]
OK	Expected farrowings next quarter	0.00	0.15	0.15[^]
TX	Boars for breeding	23.23	23.23	23.23
TX	Market hogs under 50 pounds	5.25	5.25	-0.33
TX	Pigs born last quarter	6.33	6.16	2.68
TX	Farrowings last quarter	5.06	5.06	1.61
VA	Death loss	-0.66	-0.66	-0.66[#]
VA	Pigs born last quarter	6.99	6.02	0.33
VA	Farrowings last quarter	12.28	7.33	2.37
WY	Pigs born last quarter	0.10	0.10	0.10[#]
WY	Farrowings last quarter	0.07	0.12	0.12[*]

[#]The REVIEW estimate fell outside the CONTROL 95% confidence interval.
[^]The CONTROL estimate fell outside the REVIEW 95% confidence interval.
[*]Both estimates fell outside their respective 95% confidence intervals.

References

Banff Support Team (2003). Functional Description of the Banff System for Edit and Imputation. Statistics Canada, Quality Assurance and Generalized Systems Section technical report.

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

Johanson, J. (2012). Banff Automated Edit and Imputation on a Hog Survey. Paper presented at the Fourth International Conference on Establishment Surveys. Montreal, Canada. June 2012.

Silva, P.L.N, Lewis, D., Al-Hamad, A., and Zong, P. (2009). Investigating selective editing ideas towards improving editing in the UK Retail Sales Inquiry. Paper presented at the 2009 European Establishment Statistics Workshop – EESW09. Stockholm, Sweden. September 2009.