WP. 23
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (v): Successful strategies for implementing new E/I methods

THE DEVELOPMENT OF AN EDIT IMPROVEMENT PROCESS

Supporting Paper

Prepared by Paula Mason and Amerine Woodyard, U.S. Energy Information Administration,
Jason Marley, SAIC, United States[1]

I. INTRODUCTION

1. In 2007, a project was initiated to improve the efficiency of the editing process in a family of eights surveys that contribute to monthly estimates of product supply components and resulting product supplied. In addition to improving edit efficiency, the project was intended to establish a process that was consistent, systematic, reproducible, transparent, and would remove subjectivity wherever possible.

2. The approach developed consisted of four main parts. The first part of the improvement process was to fully document the edit rules and business processes that were currently in place. The second part was to conduct and tape record an interview with each of the key production persons of some of the surveys, using a protocol of questions designed to determine how the analysts examined edit failures, and the processes used to resolve them, while looking at the production system. The third part entailed research and analysis of the production system's audit trails to determine how the cumulative transactions compared to what the data analysts had said, as well as to provide a basis for improvement for the fourth part. The fourth part of the project involved a review by a team of higher level analysts, industry specialists, and statisticians of the edit performances and the associated edit resolutions for each rule to determine which edit rules to drop, modify and to explore what new rules should be added. This project resulted in improvements to the edit process through the modification of business rules, edit rules, edit failure documentation, system's edit module, and analysts' perception of the edit process.

[1] This report is released to inform parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily of the U.S. Energy Information Administration.

3. The surveys involved were processed by the standardized processing system, StEPS, developed by the U.S. Census Bureau as modified by the EIA. In general, the data processing flow included: a) "rollover" which initialized surveys for a new reference period; b) simple imputation (autocorrect totals through balance complex); c) edit; d) general imputation; and, e) estimation. Through edit specifications in the edit module, analysts could select from seven different edit types which included: required item, range, list specific, skip pattern validation, balance, free form survey rule, and negative test. When EIA migrated these surveys from the legacy production system to StEPS, industry specialists recoded many of the edits that were in place at the time to StEPS free form survey rule edit type which allowed the user to define a precondition for the edit and the error condition in SAS code. Some edits were not translated and some were modified for computation and processing ease. Over time, more edits were added by survey with the intent of improving data quality. By 2007, for one survey alone, there were over 35 subtypes of edits with the free form survey rule type. These edits were applied in StEPS, cell by cell, totaling over 1000 different rules, each with one or more preconditions as well as the actual error condition. The only documentation of the edit rules was the survey specifications within the system; edit rule code testing and edit rule performance reviews were non-existent.

II. THE EDIT IMPROVEMENT TEAM AND PROJECT OBJECTIVES

4. An edit improvement team comprised of the survey management team leader, industry analysts, data processing staff and survey managers, statisticians, and a SAS/StEPS programmer dedicated to these surveys was formed. The three-fold objectives of the team were laid out and agreed to in advance. The first objective was the efficiency of the editing process. Efficiency was defined in terms of the "hit rate" (% of edit failures that resulted in a change to a data item), miss rate (% of non-failed items identified as incorrect), as well as consideration of the amount/cost of staff resources required to review and resolve edit failures. Efficiency could be tested using current failure counts by edit rule and simulating the effect on the counts of a modified or new rule. The second objective was to establish a process that was consistent across surveys and data, inclusive of the automated system as well as associated business rules. This included a consistent process for identification and flagging of edit failures, the process and actions for resolving the edit failures, the coding of the resolutions, the recording and accessibility of edit performance to improve the process based on results. The third objective was to establish a process that was systematic, reproducible and transparent, removing subjectivity wherever possible. Both short term and long term goals were considered as the team worked through the project. Short term goals included actions that could be taken with no change to the system, but could be implemented by changing user selected specifications or business rules without changes to the generalized system. Long term goals included actions that required changes to the generalized system.

III. DOCUMENTING THE EDIT RULES AND BUSINESS PROCESSES

5. The edits were first mapped to the individual survey form data elements to provide the team a visual representation of what was being implemented for each cell. Each cell's edit was then captured in a spreadsheet along with the pre-conditions, conditions, edit type, and error message. Through filters and sorts, the team was able to quickly identify some inconsistencies in pre-conditions, conditions, and parameters, and identify some coding errors. This also revealed that a couple of the edits were duplicative and could be eliminated. The edits were then classified as to which StEPS edit type best accommodated the edit rule. The edit type differentiated the edit rules, determined the order in which the edits were executed and edit failures displayed. It was found that almost one-half of the edits written under the survey rule type could be implemented

under one of the other StEPS edit types, as shown in Table 1. It was determined that proper use of the other edit types would allow better management of the edit rules.

Table 1. Count of Edits by Type

EDIT TYPE	
BALANCE	3
NEGATIVE	4
RANGE	10
SURVEY RULE	19
TOTAL	**36**

6. In addition to documenting the systems edit types, the individual surveys' business processes were documented. This process identified specific rules for data analysts to follow. This identification determined that certain edit failures were systematically and manually fixed without respondent contact, such as for balance items off by 5 or less. It also documented which data flag was to be applied by the analyst according to how the edit failure was resolved, such as "If you adjust data without contact with a respondent, change the data flag to an 'I' " to indicate an imputed value or "Do not override 'I' data flags." The business rules also provided guidance to data analysts about using alternate data sources such as other survey data or industry publications in resolving edit failures and/or preparing for follow-up respondent calls. This part of the project revealed inconsistencies across surveys and analysts as to when data analysts associated a data flag of "A" indicating analyst's correction versus "I" indicating an impute value.

IV. INTERVIEWING THE DATA ANALYSTS

7. The second part of the improvement process was observing and interviewing individually the key production persons for the surveys, using a protocol of questions. The purpose of these observations was to:

- determine how edit failures were resolved,
- note if and when there is interaction and/or guidance between/from survey staff and industry analysts,
- document the editing procedure conducted by the survey staff,
- determine and document related activities that affect the edit, edit resolution or data flags, or impute values, and,
- document the sequence and timing of the edit and edit resolution process.

8. These observations also revealed inconsistencies in how analysts resolved edit failures and their choice of data flag associated with the resolution. It was also noted how the analysts' determined when an edit failure should be overridden. In particular, a number of edit failures were generated by the presence of data from the previous period but not reported in the current period (referred to as prior-no-current, or PNC) or vice versa (current-no-prior, CNP). The data analysts would review how often the item was or wasn't reported over a 12 month period. The exact number of periods for which the item was or wasn't reported did not seem to be the basis of the edit override, but rather that the data item was sporadically reported over the previous 12 months.

9. Various suggestions were made by the data analysts during the observation process. Most of these suggestions related to observed changes in the electronic collection process that resulted in more edit failures, or changes recommended for StEPS to make it easier to record and view notes recorded by the analysts, or changes for StEPS review and correction screens that would be helpful to the data analysts. It was commented though by analysts that the edits relied too heavily on the previous month's data with no regard to historical patterns, particular last year's data for the same month, resulting in edits being "tripped" too easily. In general though, the analysts felt that the edit rules were good; the edit rules helped the data analysts to focus on the failed edits and confirm that the data were reported correctly.

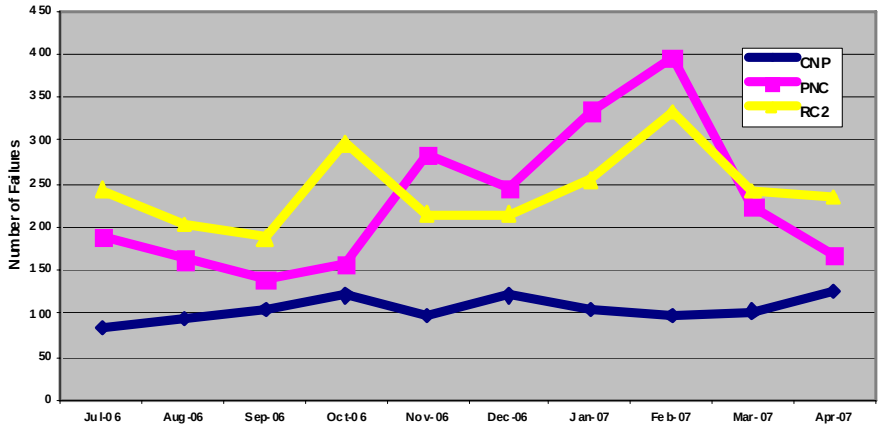
V. ANALYZING AUDIT TRAILS ON EDITS AND EDIT RESOLUTIONS

10. One team member was tasked with analyzing the StEPS audit trails and edit reject files to determine edit failure counts by survey and edit rule by reference period, and map failed cells to resulting data changes recorded by the system. Summaries were prepared and presented to the team for review. This process identified the three out of 36 edit rules with the highest failure rates but the lowest hit rate. These three as shown in Table 2 were: a) current but no prior period value (CNP); b) prior but no current period value (PNC); and c) one of three range check edits known as the RC2. These three edit failures represented roughly 44% of all edit failures for an average month but varied across the original time period examined as illustrated in Figure 1.

Table 2. Percent Failures of Top Three Rules

Edit Rule	Average per month (July 06-April 07)	Percent of all Edit Failures
CNP	136.8	9.2
PNC	261.9	17.7
RC2	246.4	16.6

Figure 1. Count of Failures over Ten Months



11. The RC2 was a key range edit used for most of the monthly surveys. This edit was used to identify volumes that deviate from a respondent's historical reporting pattern. The edit

required that all three conditions be satisfied. For some cells, if a fourth condition was satisfied, the other three conditions did not have to be satisfied. Specifically,

Condition 1: ABS ($X_t - X_{t-1)} > K$ (where K was set to 30, 50 or 100 depend on cell) and,
Condition 2: (1.2 * MAX (X(t)) < X_t) OR (X_t <.8 * MIN ($X_{(t)}$)) over 12 month history, and,
Condition 3: ABS (X_t - μ_x) > K (where K was set to 30, 50 or 100 depend on cell)
OR
Condition 4: ABS (X_t - X_{t-1}) > M; (M> K).

12. The RC2 edit was revised with the intent of preserving the spirit of the old edit in flagging items that changed significantly relative to the previous period or relative to the respondent-cell's mean value. However, the parameters had been defined arbitrarily and did not vary by the size of the respondent, the region of the country, or other factors that might contribute to determining unusual changes, other than the three different parameter values. After exploratory analysis of the data over a five year history and simulation of possible options, the recommended revised edit flag a value if two conditions were met. The first condition required that the absolute value of the difference between respondent-cell's mean change from period-to-period over five years[2] and the change from the previous period to the current period was greater than the respondent-cell's standard deviation of the differences over that time period (formula 1 below). The second condition required that the absolute difference between the respondent's mean value and the current value was greater than two standard deviations over that time period (formula 2 below).

$$(1) \quad \text{failure} = \left|\mu_{\text{diff}} - \left(\text{x}_{\text{diff}}\right)_t\right| > \sigma_{\text{diff}}$$

and, $$(2) \quad \left|\mu - \text{x}_t\right| > 2 * \sigma$$

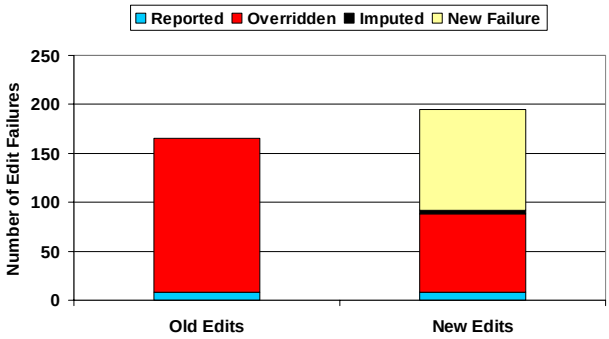
and, $$(3) \quad \frac{\sum_{t=1}^{12} x_t}{\sum_{j=1}^{n}\sum_{t=1}^{12} x_{jt}} > M_i$$

13. In addition, in order to eliminate small values that are more volatile but contribute little to aggregate values, the effect of a market share threshold was tested and included as a third condition (formula 3). The threshold was adjusted by survey to accommodate the differences in distributions among the surveys. It was expected that the threshold would later also be tailored to individual products after the large differences were resolved. Given the deviations also regionally also among the surveys, the market share was calculated at regional levels when appropriate. Concerns regarding the possibility of missing a very large value reported by a respondent normally reporting small values, and therefore not satisfying the third condition, were addressed through other edits rules already in place. Results of the revised edit were simulated using historical data for five years. The comparison of the results using the revised RC2 and the old RC2 edit were examined systematically from top to bottom showing the overall difference in performance graphically and through tables, and then more closely, separating out specific

[2] If a respondent-cell did not have five years of history, whatever history was available was used as long as at least 12 months history was available.

products and supply types, and performing in-depth analysis of the differences and resolution codes to alleviate fears of not identifying values that should have been identified. This process continued until consensus was reached within the team. Figure 2 shows the edit failures generated by old and new range edit segmented by the resolution to the failure. "Reported" indicates a change was made to the value, "overridden" indicates there was no contact with the respondent but the analyst over rode the failure, "imputed" indicates the failed value was replaced with an impute value, and "new failure" indicates a newly identified failure so the resolution is unknown.

Figure 2. New and Old Range Edit, Comparison of Failures



14. The process for developing new edit rules was similar. Two possible new edit rules were considered. The first was a range check edit that took into account the seasonality of the item. Work on this edit though was not fruitful because of the variability in the seasonality itself. While it was known that certain elements are larger during certain seasons (such as heating oil in the winter), the specific reporting period varied. Simulations were performed based on aggregating data to form quarters and moving quarters to get around the problem, but it was clear that an edit would require a parameter that depended on information outside of the survey processing system and work in this area was postponed for later research. The second possible edit also required information outside the system, but that information was much more defined. In particular, an edit was examined that compared those respondents' who reported also on a corresponding weekly survey, for those comparable data items. Weekly data transformed into monthly values (MFW) were available in an Oracle database. These data (y_t) could be viewed and brought into StEPS through an auxiliary file and compared to reported data (x_t) in StEPS where counterparts existed (formula 4). Similar to the RC2 edit, the MFW edit was designed to not fail small values or small relative differences (formula 5).

$$(4) \quad failure = \left| \frac{x_t - y_t}{x_t} \right| > Z_i$$

and,

$$(5) \quad \frac{\sum_{t=1}^{12} x_t}{\sum_{j=1}^{n} \sum_{t=1}^{12} x_{jt}} > M_i$$

15. This new edit accounted for some of the activities identified in part two of the improvement process, those activities being performed by survey staff outside of the processing

system. By bringing those activities into the edit process within StEPS, those edit activities were made consistent and systematic across analysts and survey elements. The values of the parameters Z_i and M_i were decided based on simulation and detailed review of historical data and their distributions by survey and set to reflect what analysts would consider necessary identification of values needing respondent follow-up while not producing an unrealistic workload relative to the survey production cycle.

16. It was planned that the MFW edit would be expanded once the new version of StEPS was implemented. The new version of StEPS incorporates a new imputation methodology. The old methodology used the previously reported monthly value as the impute value. The new methodology uses either an MFW value, if it exists, or for respondents and items with no MFW value, the impute value is based on the smoothed historical value adjusted by the trend of aggregated MFW values where they exist. The impute values will replace the y_t as the expected value in the above formula for identifying edit failures.

VI. OTHER FINDINGS

17. Throughout the edit improvement process changes to the StEPS system were identified. As a result of those, StEPS has undergone a major modification that provides analysts information previously available only through audit trail in the modified review and correction screens with extended user interfaces. Impute values will be shown along with reported data to assist the analysts. Data flags were expanded to distinguish if a change was made with respondent contact or not, and the usage of the data flags was standardized across surveys for consistency in process and documentation of that process.

18. To further manage and maintain the edit process, wildcard features have been implemented for edit specifications. These features allow the user to create an edit rule and apply the rule to multiple data elements, ensuring that edit rules are written consistently and reducing the potential of coding error. This approach was useful that all edits were re-specified using the wildcard features. Edit rule changes were implemented in production in phases with some of the changes implemented once the team decided on the change. Other changes, such as those involving more than 12 months of data (such as the calculation of means and standard deviations), were delayed until views from the Oracle database could be constructed in order to simplify operations. These views will prevent StEPS from having to store an extended history and update the values needed for the edits.

19. Business rules have also been modified. Balance items that had been corrected manually were included in simple imputation in StEPS that performs balances based on specifications defined by the user. Again this approach applies balances systematically and consistently without concern, and flags imbalances with beyond the specifications as edit failures.

VII. SUMMARY

20. The project initiated to improve the efficiency of the editing process in a family of eights surveys combined the skills of statisticians, analysts, programmers, and managers to provide the perspectives of sound statistical approaches, industry knowledge, and production process limitations. The mapping of audit trails to edit reject files provided the basis for examining edit performance measures on failures and resolutions. Observation of the edit process along with documentation of the business rules unveiled inconsistencies across analysts and surveys about the treatment of edit failures. This information provided a foundation that enabled the team to share a common belief that the editing process was not efficient. The most inefficient edit rules

were examined and reconstructed to account for variation among and between individual respondents and data elements. New rules were constructed to include activities that were being performed outside the edit module. Performance of all proposed changes was simulated on historical data and compared to old edit rules. Both system's and business processes were modified to better enable data analysts to resolve edit failures and record their actions. As a result, in addition to improving edit efficiency, the project established processes and procedures that were consistent, systematic, reproducible, transparent, and removed subjectivity wherever possible.