

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14 - 17 October 1997)

Item 4 of the provisional agenda

DATA EDITING AND PERFORMANCE MEASURES

Submitted by the U.S. Department of Energy¹

¹ Prepared by Paula Weir.

I. INTRODUCTION

1. Surveys, and in particular business surveys, have become increasingly costly and burdensome. The resources required by government to collect and process the data, as well as the resources required by the respondents to file the surveys are an ongoing concern. The resources required for data editing are a major portion of this burden. While the costs of editing data have been estimated to be 40% of business survey costs, no estimates have been provided for the respondent side. Performance measures on the entire survey processing system, and most particularly the editing portion, are the key for monitoring, evaluating, and reducing the costs and burden while improving the quality of the processed data. Performance measures ideally should track the amount of work done, the areas in which the work is done, how long the work takes, and the effect of that work on the released data over time. More work does not translate into better data. These measures enable managers to allocate resources more efficiently in both the broad/long run sense and the more specific/short run sense. They provide information from the process that contributes to a preventative approach, i.e. prevents data errors in future periods or surveys. Performance measures may indicate that less resources should be dedicated to editing and more resources to survey instrument design and respondent training. Performance measures may indicate more resources are needed on specific survey edits and less on others. Performance measures may indicate deterioration in the process through time or that a process is out of control.

2. Performance measures are needed by survey managers throughout the survey processing cycle. They aid in assessing the overall workload and the workload to individual data analysts. They summarize performance at the macro or aggregate level, as well as the micro or reported level, reflecting the top down approach to editing. They should be maintained and analyzed through time for periodic surveys. Performance measures should be calculated not just as a total, but for groups of respondents, and reporting categories to reveal any differences in performance between groups or categories. These measures even provide feedback on the editing rules and the parameters themselves. They track the correction rate, as well as the detection rate. Regardless of whether data failing an edit are automatically imputed (correction rate = 100% of failed data) or manually reviewed and examined for followup and recontact, performance measures are equally necessary.

3. Associated with the work done are type I and type II errors for both detection and correction. A response flagged by the edits but not changed could be considered a type I error in terms of detection (figure 1). A record not flagged but changed could be considered a type II error in

Figure 1: type I and II errors for detection

	FLAG	NO FLAG
CHANGE	(OK)	Type II Error
NO CHANGE	Type I Error	(OK)

Figure 2: type I and II errors for correction

	CHANGED	NOT CHANGED
RESPONSE WAS RIGHT	Type I Error	(OK)
RESPONSE WAS WRONG	(OK)	Type II Error

terms of detection. A record that is changed but should not have been changed is a type I error in terms of correction (figure 2). A record that is not changed but should have been changed, is a type II error in terms of correction. Type I and II errors for detection can be directly measured. Some systems may not have a type II error for detection if the system only allows change to responses that are directly flagged (i.e., changes can not be the indirect result of another response flagged or, in the case of a micro edit, result from a macro edit). Some systems may not have a type I error for correction if corrections are only made based on the respondent's confirmation. Both type I and II errors for correction, however, require information about truth that usually is not available to the process itself. If changes are followed up on later by, for example, sampling the changed responses and recontacting the respondent, a type I correction error could be calculated. Similarly, if unchanged responses are sampled and the respondents recontacted, a type II correction error could be calculated.

4. In order to understand the information contained in the performance measures summaries of the measures should be constructed in a meaningful manner. At the macro and micro level, the summaries should reflect the frequency of changes, the extent of changes, and the distribution of changes by edit type for groups of related aggregates. At the micro level, the summaries should also reflect respondent type, data analyst, or other grouping relevant to the process. Graphics are particularly helpful in conveying a cohesive picture of the overall process, as well as the individual pieces.

II. MACRO EDITING PERFORMANCE MEASURES

5. The process of verifying data at the aggregate level is referred to as data validation or macro editing. These aggregates most often refer to the level at which the data are released. There may be multiple levels of aggregates that are released. The performance measures should reflect those levels, drilling down through the aggregates. Performance measures for macro editing should include measures of the work performed and the effect of that work on the aggregates. The work performed is determined through the detection and correction rates. The detection rate is calculated as the number of aggregates flagged for review divided by the number of aggregates eligible for flagging. The correction rate is determined as the number of aggregates changed divided by the number of aggregates flagged. The effect of the work is calculated as the difference in each of the aggregates before and after resolution of the edit.

6. Summaries across the changes in the aggregates should be calculated. The summary should include the number of aggregates changed, the mean value of the changes in the aggregates, the median, min and max change of the aggregates, and even the distribution of

the changes. The summary of aggregate changes by edit type should represent not only the total, but also the groups of relational aggregates in order to reveal possible differences in the effect of the edits.

7. Additional measures should be calculated for processes that make use of scores in the macro editing to prioritize or rank the aggregates flagged by the edits. For aggregates that were changed, the mean score, the median score, the min and max score, and even the distribution of the scores provide information that is useful in assessing the performance of the score. Depending on how the score values are used in the system, the same information should be provided for aggregates that were flagged but not changed and compared to the results for the aggregates that were changed. This information is useful in determining the effectiveness of the score, as well as possible optimal score values for macro editing.

III. MICRO EDITING PERFORMANCE MEASURES

8. The process of verifying the data at the reported level is referred to as micro editing. In parallel with performance measures for macro editing, measures for micro editing should include measures of work performed and the effect of that work on the aggregates. The effect of that work on that particular response variable is only important in so far as it effects the aggregate, unless the data are released at the micro level. Here again, the work performed is determined through the detection and correction rates. The detection rate is calculated as the number of reported values flagged for review divided by the number of reported values eligible for flagging. The correction rate is determined as the number of reported values changed divided by the number of reported values flagged.

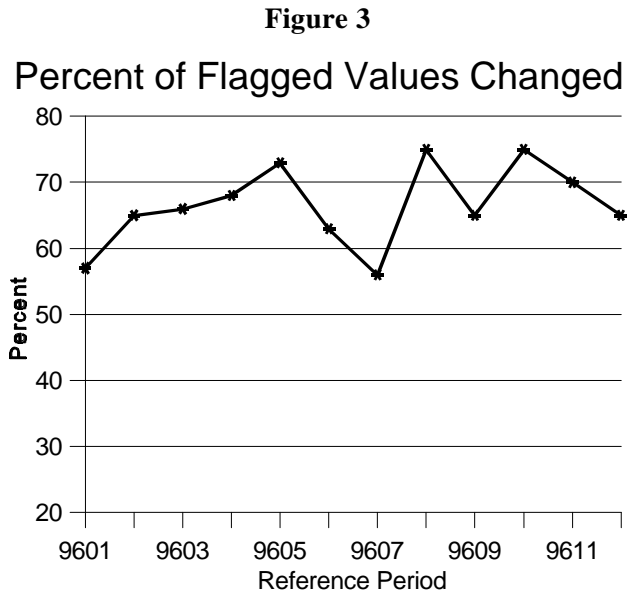
9. Some surveys may have an additional work component that occurs between flagging and final resolution of an edit. If it is the procedure of the survey to examine flagged values and determine if further follow-up is required before resolving, such as contacting the respondent, assuming not all failed data require contact, a contact or called rate should be calculated as the number of responses for which calls were made divided by the number of responses flagged. This work performed represents work by the processors of the data, as well as the providers of the data. The effect of the work (detection, calling, and correction) is calculated as the difference in the aggregate before and after resolution of the edits for each variable reported.

10. As in macro edits, a summary of aggregate changes by edit type (count, mean, median, min and max, and even distribution) should be calculated for the total, and also for the groups of relational aggregates in order to reveal possible differences in the effect of the edits. Additional measures should be calculated for processes that make use of scores in the micro editing to prioritize or rank the flagged responses. For reported or micro level data that were flagged and changed, the mean score, the median score, the min and max score, and even the distribution of the scores provide information that is useful in determining the performance of the score. Depending on how the score values are used in the system, the same information should be provided for reported data which were flagged but not changed and compared to the results for the reported data that were changed (as well as data that are changed but not flagged if that possibility exists in the system). This information is useful in determining the effectiveness of the score, as well as possible optimal score values for micro editing. These

measures should be calculated at the edit variable level for each type of edit that exists, if more than one type exists. The summary of the scores previously described should be provided in total and for the relational aggregates, each variable reported, as well as, the respondent type, the individual analysts reviewing the edits, etc., that make sense for that particular survey.

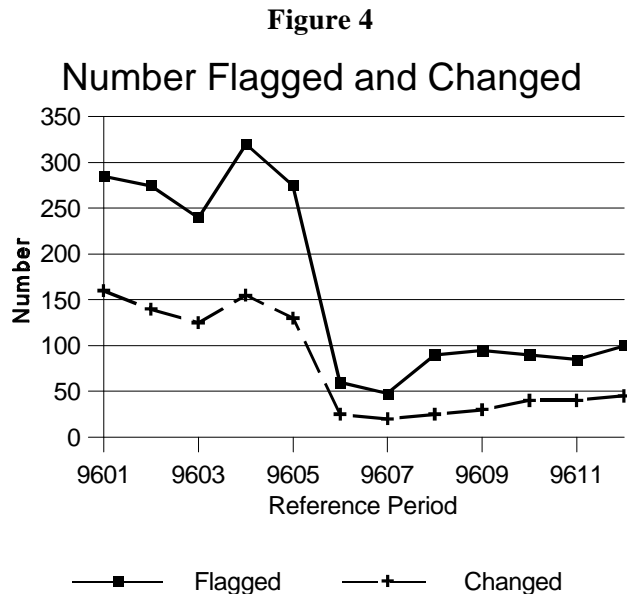
IV. PRESENTATION

11. Today's technology provides the tools for ease in recording information necessary for producing performance measures. The editing process should record the edit failure type, the edited response before and after correction, if corrected. It should also record the aggregates before and after macro editing. While tabular display of counts and rates as performance measures is common, graphic presentation enables visualization of the process results. A simple example



of such a summary graph of a performance measure would be a time series plot with each survey reference period along the x axis and the percent of flagged responses (or aggregates) that were changed along the y axis (figure 3).

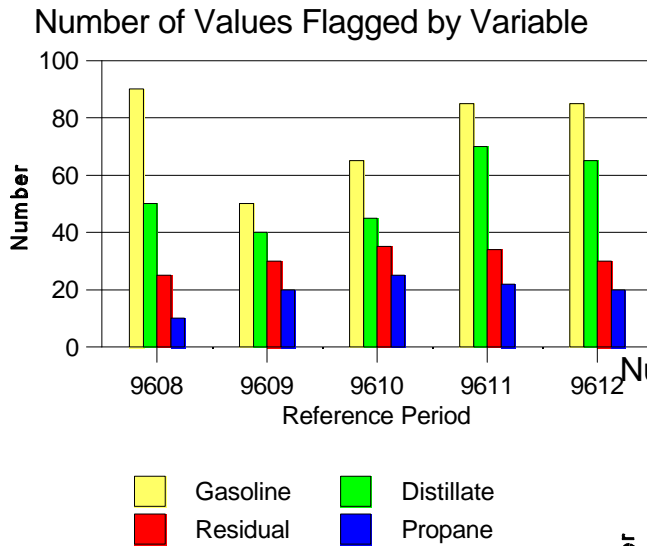
This would provide an understanding of: 1) how the flagged and changed rate varies over time, 2) whether there are cyclical patterns, and, 3) when the process seems to be out of control. An alternative time series graph would be a plot with each reference period along the x axis and the response (or aggregate) count along the y axis. Two lines could be plotted--one for the number of responses (or aggregates) flagged, and one for the number of responses (or aggregates) changed (figure 4).



While this graph provides information similar to the changed rate graph, it further shows how the number of responses flagged tracks with the number of responses changed, and also emphasizes the total amount of work. To highlight or compare different response variables, a summary of performance measures could be depicted using multiple bars, one for each variable. The set of bars could be repeated along the x axis for the relevant periods. The y axis would represent the count of responses (or aggregates) flagged or changed (figure 5). Furthermore, individual bars could be stacked to reflect the number flagged with the number

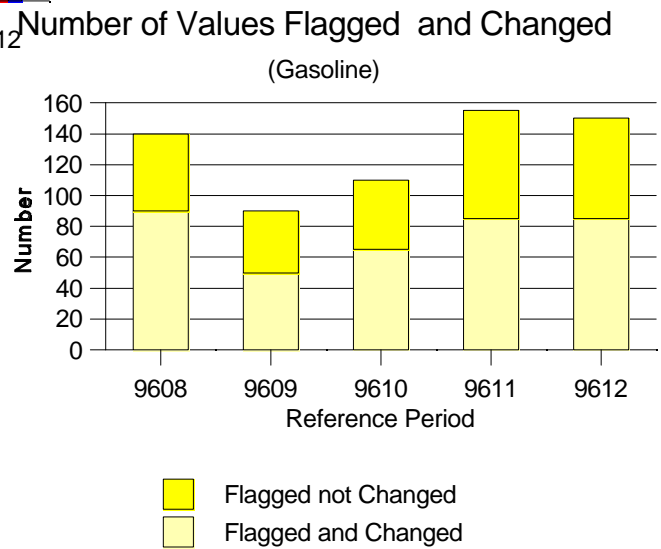
changed shaded within the bar (figure 6). Graphical displays provide more insight than tabular displays by making use of the techniques of exploratory data analysis. More

Figure 5



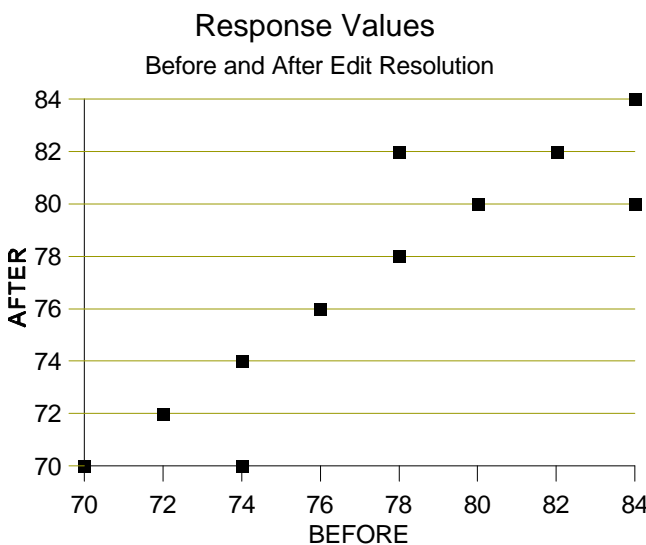
analytical graphs provide information on distribution and importance of flagged and changed records. For example, a graph of the number of records flagged and changed does not provide information on what, if anything, is different about the changed responses as compared to the flagged but unchanged responses. However, a scatter plot of after edit

Figure 6



resolution responses plotted against before edit resolution responses would visually depict the frequency of values that were changed, and how much they changed (figure 7). The majority of the data would lie on a straight 45 degree line. Responses could be plotted using three different symbols to represent values not flagged and

Figure 7

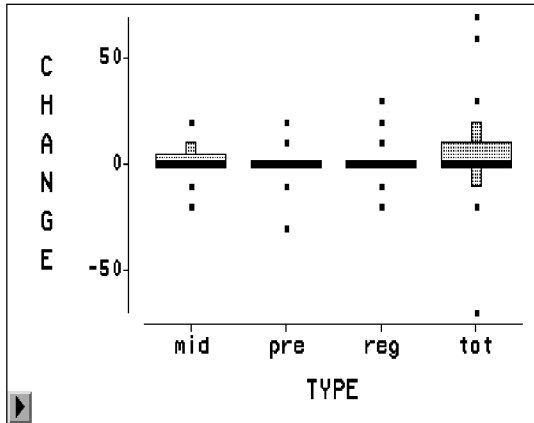


not changed, values flagged but not changed, and values not flagged but changed. The data could be plotted using color to represent the edit score range or measure of the response's contribution to the aggregate. This graph would show the distribution of the changed responses and their relative importance. Another analytic graph of performance measures would present multiple Box Whiskers plots for the various response groups or response variables along the x axis. The y axis would represent the amount of change in the data (figure 8). For micro editing, the individual points plotted would represent the difference in the individual responses before and after edit resolution. For macro editing, the individual points would represent the difference in the aggregates before and after the edit resolution. Related

aggregates such as different geographic regions for a particular variable, would form the box. Side by side boxes could be shown for each response variable. On line graphics would have point and click functionality to further identify the particular data points. Points between boxes could even be connected when clicked on to show how an individual respondent tracks across variables or how a particular geographic area tracks across variables.

Figure 8

Change in Responses by Response Variable



V. SUMMARY

12. Performance measures are the key to a successful editing process that minimize resources and maximizes quality. They provide the tool for analyzing the process determining where improvement is needed, and evaluating alternatives based on quantitative information. If these measures are tied into the processing system, they can provide real time information for decisions and actions during the production cycle, as well as longer term preventative actions and alternatives. The measures should start at the top with macro editing and flow down to the micro level and provide some historical perspective. Micro level performance measures mirror the macro level performance measures. Both track the work done in terms of correction and detection, and the effect of that work on the data at the level that it is released. Both provide information on scores used to prioritize or rank data for editing. The measures are summarized to an overall total and summarized by edit type respondent groups, response variables, etc. Graphic presentation of the performance measures enables visualization of the process results and allows for exploration and insight into the process for continuous improvement.