Initial Evaluation of Selective Editing at the U.S. Energy Information Administration (EIA)

Elizabeth E. Panarelli (Presenter), Mathematical Statistician

Additional Team Members: Ruey-Pyng Lu (Project Leader) Mathematical Statistician; Jason Worrall, Mathematical Statistician

EIA, Office of Survey Statistics, Office of Survey Development and Statistical Integration

March 28, 2012

Overview

- Why Use Selective/Significance Editing?
- Identifying Likely Error Candidates
- Evaluating Selective Editing at EIA
- Lessons Learned
- Conclusions
- Next Steps

Why Use Selective/Significance Editing?

- Selective/Significance Editing involves identifying a smaller set of survey responses that are error candidates, with the following goals:
  - Utilizing statistical methods to identify the survey response data most likely to be in error by a disproportionate amount
  - Lowering the resources required in the process of identifying and manually investigating error candidates
  - Accomplishing the above while maintaining the quality of aggregated survey data

Why Use Selective/Significance Editing? (Cont'd)

- **Score Function:** a numerical indicator used to prioritize micro data review
  - Example of ranking scores to create error candidate selection sets:

ID	Score	Rank
17	5.5826	1
62	4.4582	2
54	4.1098	3
79	3.9075	4
33	3.6713	5
92	3.2847	6
46	3.0731	7
81	3.0159	8
99	2.8976	9
25	2.6244	10

80% Threshold: Top 20% are selected/"cutoff" for manual investigation of potential errors

- For EIA, Selective Editing provides an opportunity to use score functions as a proxy for some data validation rules (i.e., in some instances, scoring all responses rather than scoring only those responses that fail certain types of data validation rules may identify similar sets of influential error suspects)

Identifying Likely Error Candidates

- For simplicity, a single, non-zero response item (vs. a unit response comprised of multiple items) across EIA-defined geographic regions is being validated in the initial survey used for the study
  - Separate processes are being used to validate zero and non-response items
- The percentage of ranked responses included in the proposed selection sets were chosen to approximate the quantity of error candidates investigated by current survey response data validation techniques

Identifying Likely Error Candidates (Cont'd)

- Score Functions developed by Latouche and Berthelot (1992) were adapted for evaluating single response items

- **Adapted Ratio Score Function:**
    - Intended to provide a more uniform distribution of score values, based on the ratio of the current reported value and final value from the completion of the previous survey cycle
    - "Raw"/initially reported data required: current and previous two survey cycles/collection periods
    - "Final"/published data required: prior three survey cycles/collection periods

Identifying Likely Error Candidates (Cont'd)

- Score Functions developed by Latouche and Berthelot (1992) were adapted for evaluating single response items
  - **Adapted Diff Score Function:**
    - Emphasizes the absolute discrepancy between the current reported value and the final edited values of the previous cycle/collection period, as weighted by the sum of all response values from the previous cycle/collection period.
    - "Raw" data required: current survey cycle
    - "Final" data required: prior cycle

Identifying Likely Error Candidates (Cont'd)

- A third score function was created to emulate current data validation methods on single response items for this survey:
  - **TenP (10%) Score Method:**
    - The current survey cycle "raw" data value is at least +/– 10% different from the prior survey cycle "final" data value,

      AND

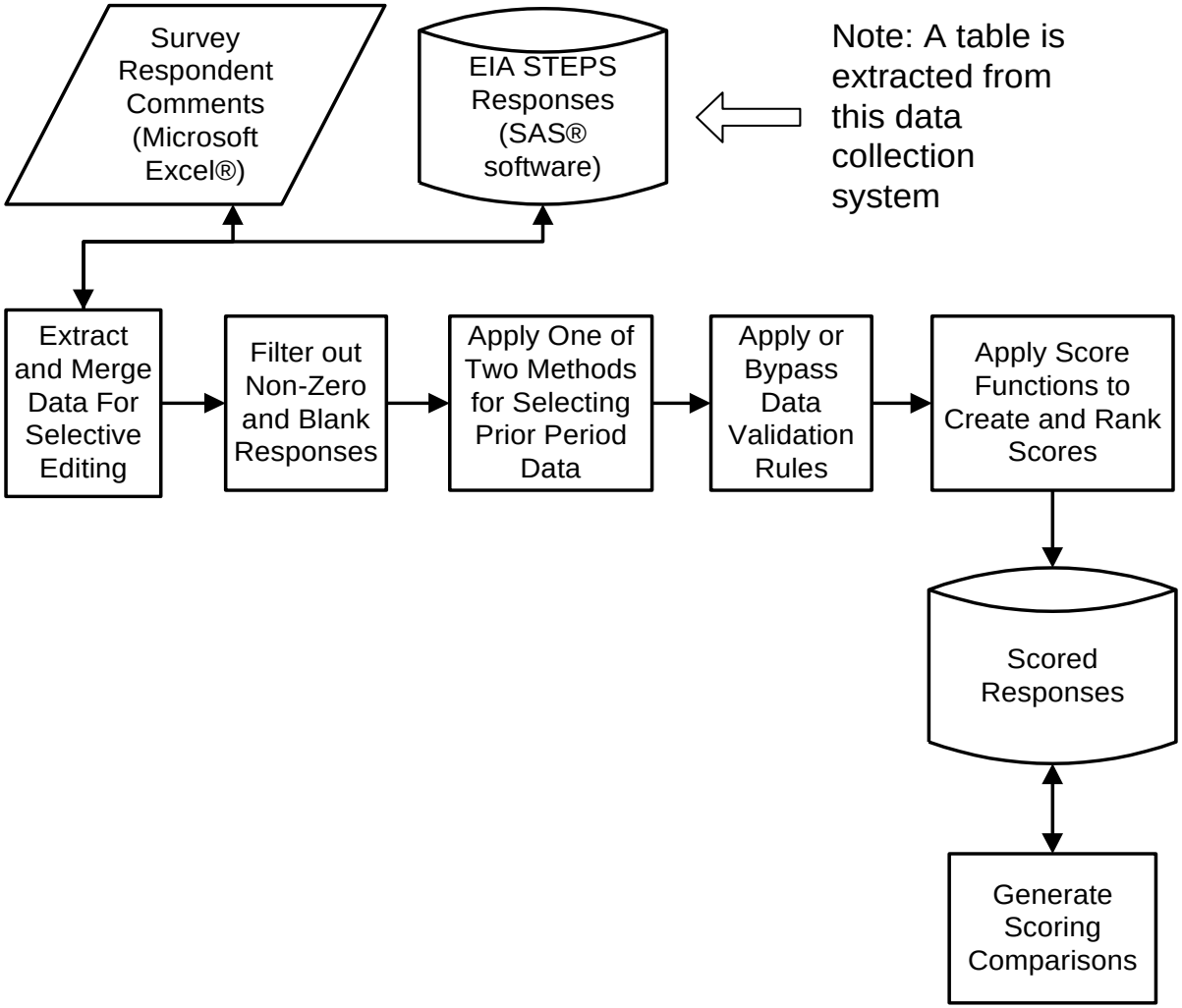
    - The size of this difference is greater than 4 units of measure per day

- Note that current validation techniques for this survey are practical (i.e., do not involve statistical significance): potential errors are investigated as time permits

Identifying Likely Error Candidates (Cont'd)

- A SAS® software application with an Enterprise Guide® software interface was developed to apply score functions and create and evaluate error candidate selection sets
- More than two calendar years of non-zero historical monthly survey cycle data were scored
- The application scored and ranked the data for each survey cycle by EIA geographic region and enabled adjusting the percentage threshold cutoffs to arrive at selection sets of different sizes

Identifying Likely Error Candidates (Cont'd)



Evaluating Selective Editing at EIA

- In the survey studied, "final" data can be changed from "raw" data. Some reasons for this include:

Validation Rule/Edit?	Re-Contact Respondent?	Reason for Data Change
Yes	Yes	Respondent agrees that the data is in error and resubmits the survey form
Yes	No	EIA identifies a unit-of-measure error and corrects the data in the system
Yes	No	Respondent resubmits corrected data prior to being contacted by EIA
No	No	EIA knows due to a market change (e.g., a merger or acquisition) to adjust the data in the system
No	No	Respondent later determines the data needs to be changed and resubmits the survey form

Evaluating Selective Editing at EIA

- To evaluate any data validation method, potential data errors that were flagged and the results of any corresponding manual data investigation must be noted
- Once the study was started, the study team realized that business practices for this survey provided only some of the indicators of the reasons for changes between “final” and “raw” data
- The survey team agreed to utilize additional reason codes available in EIA Standard Energy Processing System (STEPS) to categorize data changes for survey cycles
- Efforts to apply these additional reason codes are now underway

Evaluating Selective Editing at EIA

- To incorporate historical data in this initial study for the evaluation techniques, assumptions had to be made about changes between “final” and “raw” data:

  “True errors” captured by Selective Editing are assumed to be all responses included in selection sets where there is a change between “final” and “raw” data

- The magnitude of changes between “raw” and “final” data in the selection sets were summarized by EIA geographic region and compared across the different score function methods

Evaluating Selective Editing at EIA (Cont'd)

The composition of selection sets for different score functions by survey cycle and region are compared by survey response

An "X" indicates that respondent was included in the selection set for the particular score function method

	Scoring Method			Avg. Daily	Final Minus
ID	RATIO	DIFF	TENP	Raw Value (MMcf)	Raw Value (MMcf)
	X	X	X	216	38.51613
	X	X	X	209.03226	-73.709677
	X	X		180.64516	0.225806
	X	X	X	165.03226	0
	X	X	X	137.32258	0
	X	X	X	126.6129	0
	X		X	118.22581	0
	X	X	X	62.225806	38.83871
	.	.	.	.	.

Evaluating Selective Editing at EIA (Cont'd)

RATIO vs. DIFF			RATIO vs. 10%			DIFF vs. 10%		
Qty. Same Suspects	Qty. Suspects	%	Qty. Same Suspects	Qty. Suspects	%	Qty. Same Suspects	Qty. Suspects	%
10	16	63	12	16	75	8	16	50

- The similarities between selection sets by survey cycle and region are compared in this example
- Note that the size of the selection sets in this example is 16 observations
- The Ratio and Diff score function selection sets share 10 suspects, which means they have roughly 63% of their observations in common

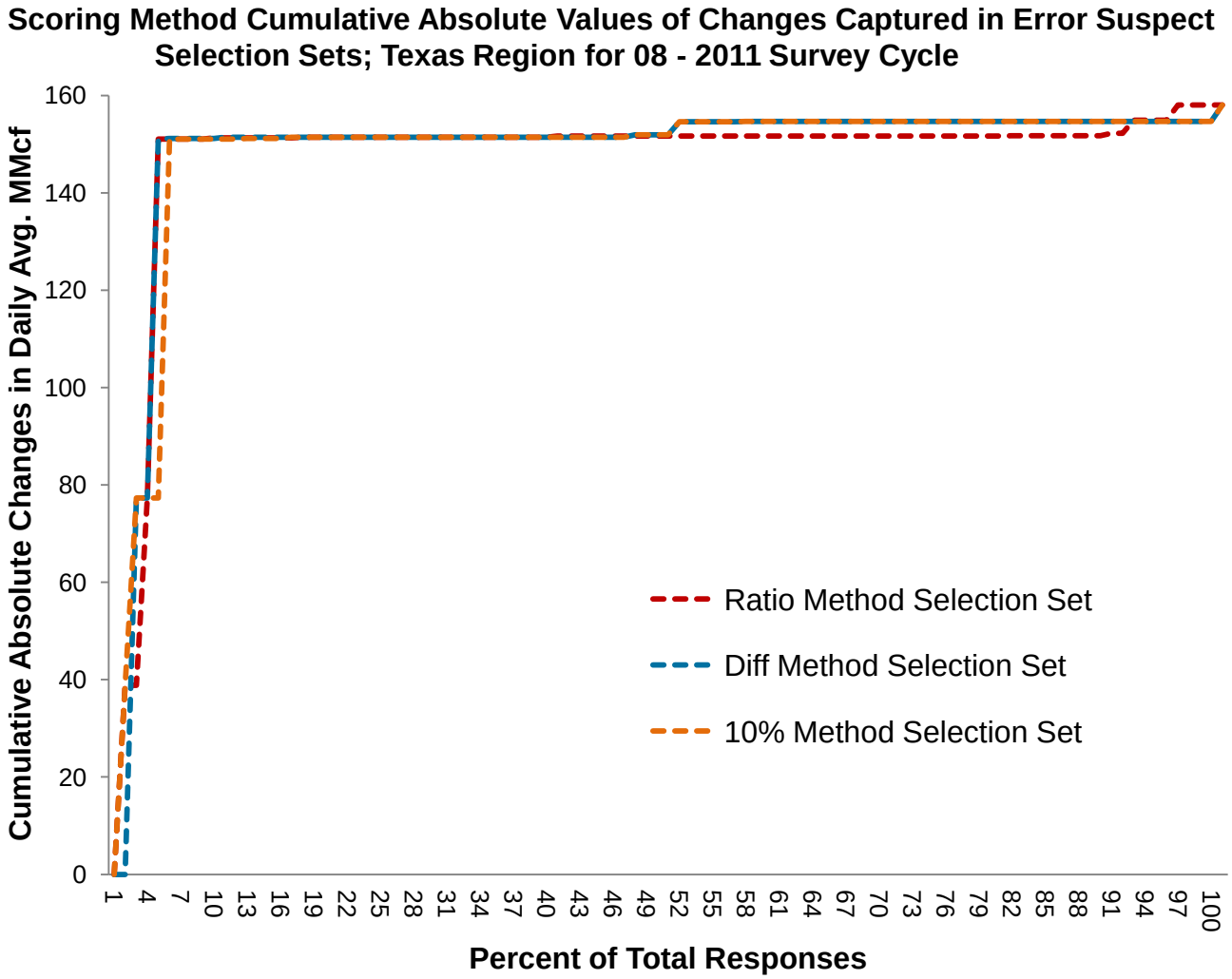
Evaluating Selective Editing at EIA (Cont'd)

Net Score Changes (Avg. Daily MMcf)				Score Change As a share of Total Final Responses (%)			
RATIO	DIFF	10%	All Changes	RATIO	DIFF	10%	All Changes
3.870968	4	3.645161	10.03226	0.02	0.02	0.02	0.05

- When considering only the changes between "final" and "raw" historical data in the selection sets for this cycle and geographic region, selection set response value changes represent approximately 40% of all historical data changes
- The remaining 60% of data changes not in these selection sets have a minimal impact on estimated summary-level data
- Note that this example shows a bias in summary-level data corrections for this particular cycle and geographic region (i.e., "final" values are increased over "raw" values)

Evaluating Selective Editing at EIA (Cont'd)

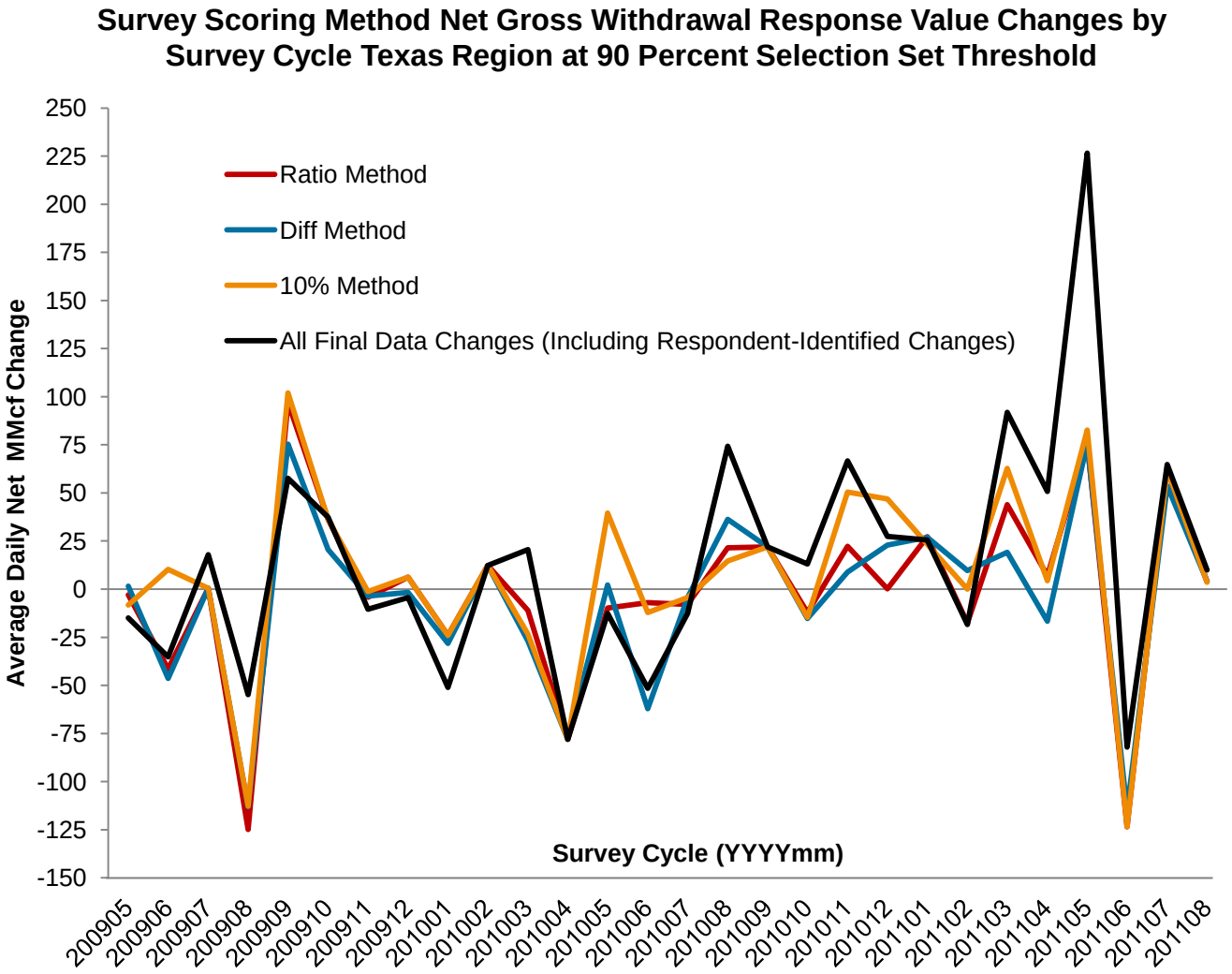
This graph provides information to help select a cutoff threshold from which to create the selection sets from the scored observations



Evaluating Selective Editing at EIA (Cont'd)

This graph is useful for determining if different scoring methods are directionally correct vs. all historical "final" vs. "raw" data changes

The selection set data series include only the data changes in observations contained in the selection sets for the cycle and region



Lessons Learned

- Evaluating any editing method requires an “audit trail” of changes to data and associated reasons for these changes
- Different selection set cutoffs and different score functions may be appropriate for different survey strata
- Applying scoring of survey responses as tool in data validation/editing can enforce a repeatable process that, in turn, can be evaluated over time

Conclusions and Next Steps

Conclusions

- For this initial survey, differences in revised historical data cannot definitively be used to evaluate Selective Editing
- When assuming all data revisions are due to manual investigation, Selective Editing appears to provide for this survey similar quality for summary-level published data as does the current editing practice

Next Steps

- System enhancements for this survey and parallel testing of editing methods should permit better evaluation of the use of Selective Editing for this particular survey

Thank You!

Contact Information

Liz Panarelli, Mathematical Statistician
Energy Information Administration,
Office of Energy Statistics,
Office of Survey Development and Statistical Integration
(202) 586-2234
elizabeth.panarelli@eia.gov

Appendix: Adapted LaTouch & Berthelot (1992) Score Functions

These scores were adapted to apply to a single item vs. a unit response and to remove weights for survey sample inclusion (as a quasi-cutoff sampling methodology is used for the survey in the study)

$$\boldsymbol{RATIO}_{k,t} = \frac{|g_{k,t} - MDG_{,t-1}|}{IRG_{,t-1}}$$

where
$MDG_{,t-1}$ is the median of the $g_{k,t-1}$ computed using previous cycle data; and
$IRG_{,t-1}$ is the interquartile range of the $g_{k,t-1}$.
where

$$g_{k,t} = s_{k,t} \times \sqrt{MAX(y_{k,t}, \acute{y}_{k,t-1})}$$

where

$$s_{k,t} = \begin{cases} \left| \frac{r_{k,t}}{MDR_{,t-1}} - 1 \right| & if\ r_{k,t} > MDR_{,t-1} \\ \left| 1 - \frac{MDR_{,t-1}}{r_{k,t}} \right| & otherwise \end{cases}$$

where
$r_{k,t}$ is an estimate of error given by $r_{k,t} = \frac{y_{k,t}}{\acute{y}_{k,t-1}}$;

$y_{k,t}$ is the value reported by respondent $k\ (k = 1, 2, \dots, K)$ at time t ;

$\acute{y}_{k,t-1}$ is the value reported by respondent $k\ (k = 1, 2, \dots, K)$ at time *t-1* ;

$MDR_{,t-1}$ is the median of the $r_{k,t-1}$ computed using data from time *t-1* and time *t-2*

$$\boldsymbol{DIFF}_{k,t} = \frac{|y_{k,t} - \acute{y}_{k,t-1}|}{\hat{Y}_{,t-1}}$$

where

$\hat{Y}_{,t-1}$ is the sum total of all response items from the previous survey cycle