Edit Reduction Research in the U.S. Census Bureau's Economic Directorate

L. Kaili Diamond
Brian Dumbacher

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Outline

- Purpose of Research
- Defining an Edit
- Survey Overview
- Research Data
- Experiments
- Results
- Summary
- Future Research

Purpose of the Research

- Prior research indicates over-editing, which can introduce bias and delay data dissemination.
- Identify areas of improvement in our editing processes in order to improve the timeliness and quality of our estimates while reducing our cost.
- Are there clear “stopping points”?
- What steps can be taken to allow the development of a model to signal when to discontinue edits?

What is an Edit?

- Changes made to variable (item) values not created by General Imputation Programs (GIMP).

- 2 Ways to Edit:
  - Automated Edits
    - Example: 65 + 37 = 100
  - Analyst Edits

Annual Capital Expenditures Survey (ACES) Overview

- Provides data on capital spending for new and used structures and equipment by U.S. nonfarm businesses.
- Capital expenditures is not highly correlated with any other variable, therefore general imputation methods are not used.
- Securities and Exchange Commission (SEC) filings are available for public companies and can be used for missing data.
- Estimation Results Files (ERFs) are saved on a bi-weekly (Tuesday and Thursday) basis during the collection and editing process.
- ACES uses a flagging system to identify possible items that need to be edited ("edit failures").

Quarterly Services Survey (QSS) Overview

- QSS is a Principal Economic Indicator Survey.
- QSS is a subsample of the Services Annual Survey (SAS).
- It produces quarterly estimates of total operating revenue, the percentage of revenue by class of customer (government, business, and households), and total operating expenses for selected service industries by tax status.
- It also produces estimates for the number of inpatient days and discharges for hospital services.
- "Snapshots" spreadsheets contain current estimates and coefficient of variation for the revenue variable over the editing cycle.

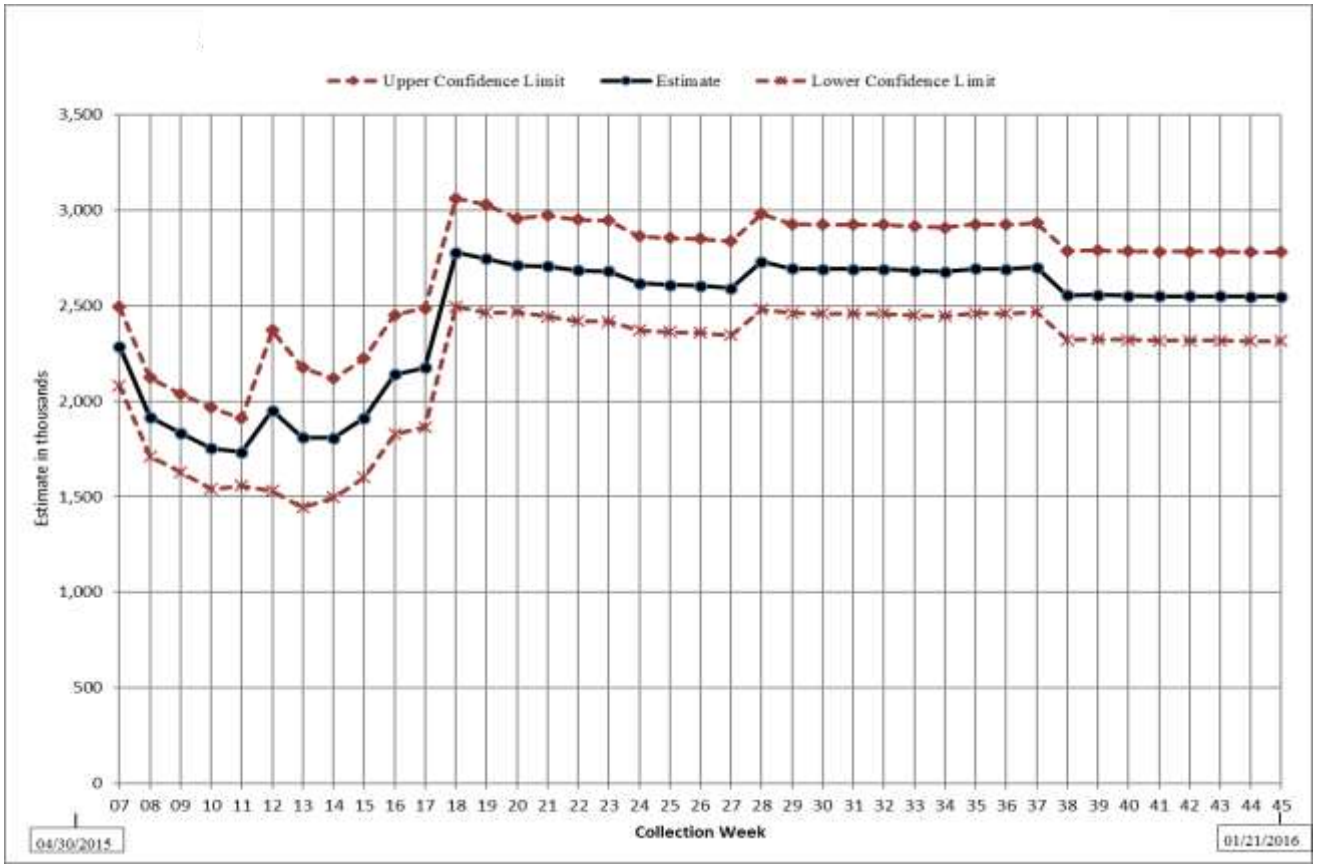
Research Methodology – Data

- ACES 2014 Survey Year
  - Audit trails for the survey's control and item data
  - ERFs
- QSS 2016 quarters 1, 2, and 3
  - Audit trails for the survey's control and item data
  - Snapshots

ACES Edit Reduction – Experiments

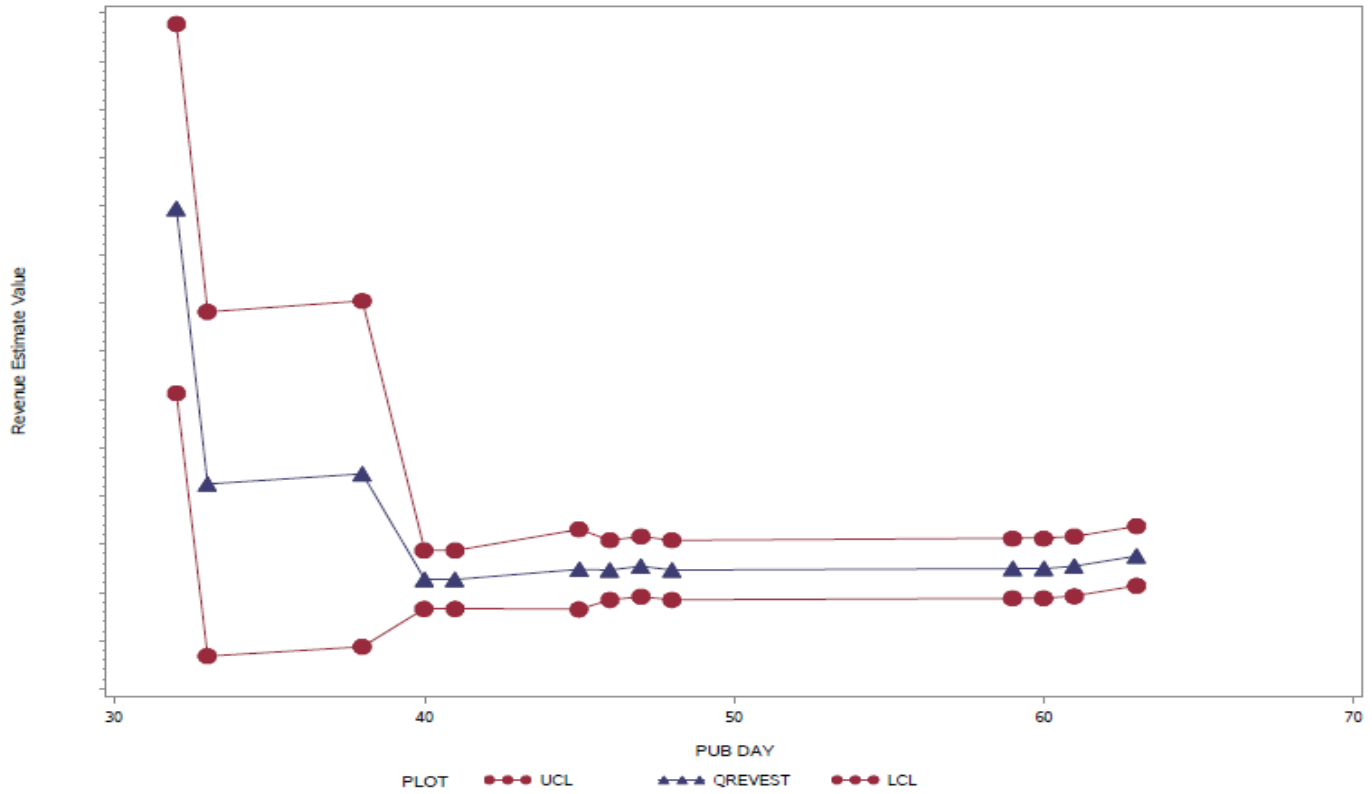
Experiment	Purpose
Examining Quantities Over Time	To examine estimates, standard errors, and the number of edit failures over time
Impact of Editing	To quantify the impact of editing on estimates by NAICS and edit type
Modeling Stopping Points	To model when the stop editing certain NAICS codes and switch resources to other NAICS codes

2014 ACES Estimates and 90 Percent Confidence Interval Limits



Data obtained from the 2014 ACES survey year Estimation Results Files.
Technical documentation can be located at:
https://www.census.gov/programs-surveys/aces/technical-documentation.html

2016Q3 QSS Revenue Estimate Over Time for NAICS 6214T (taxable outpatient care centers)

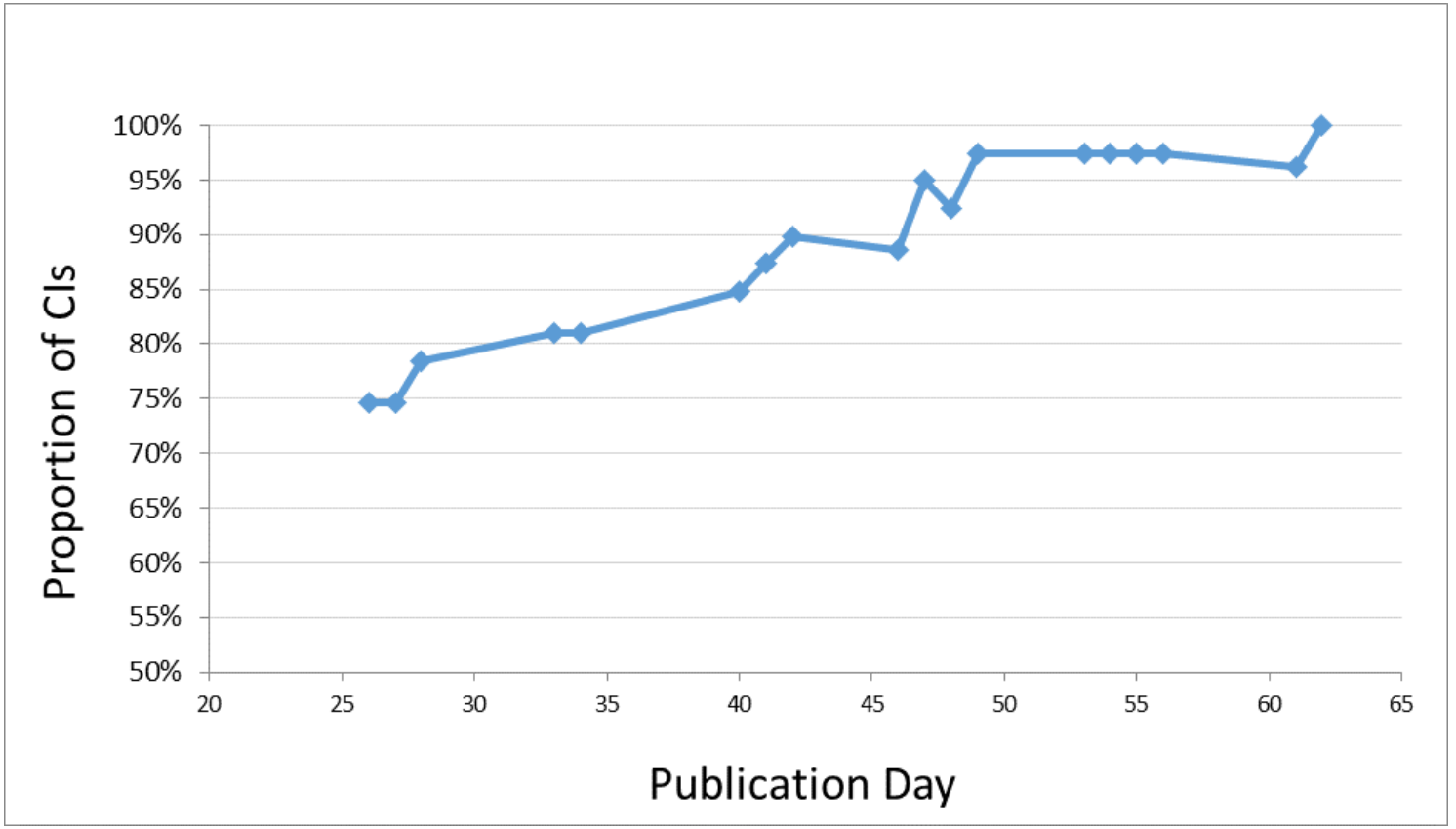


Data obtained from the 2016 Quarter 3 QSS Survey Snapshot File.
Technical documentation can be located at:
https://www.census.gov/services/qss/how_the_data_are_collected.html

2014 ACES Proportion of Total Capital Expenditures CIs Covering Final Estimates

Data obtained from the 2014 ACES survey year Estimation Results Files.
Technical documentation can be located at:
https://www.census.gov/programs-surveys/aces/technical-documentation.html

2016Q1 QSS Proportion of CIs Covering Final Estimates-Revenue



Data obtained from the 2016 Quarter 1 QSS Survey Snapshot File.
Technical documentation can be located at:
https://www.census.gov/services/qss/how_the_data_are_collected.html

QSS Data Flags and Counts

Detailed Source Flags			
Data Flag	Data Flag Description	EFLG2	EFLG2 Description
R	Data provided by respondent	X	Approximation
		F	Exact Value
		K	Improperly keyed data corrected to reported value
		R	Imputed data corrected to reported value
		N	Instructions to use annual report, 10K, or 10Q
		B	Instructions to use company website
A	Analyst deduced value that reverses reporting error	U	Wrong units (e.g. in gallons instead of barrels)
		M	Summing Error
		O	Other reporting errors
O	Obtained data from other source; quality equivalent to reported data	W	Administrative data
		I	Another survey or census
		A	Annual report, 10K, or 10Q. Needs validation
		C	Company website. Needs validation
		S	Some other source, NEC
I	Obtained data from another source; quality NOT equivalent to reported data	Y	Administrative data
		D	Analyst-derived value
		I	Another survey or census
		A	Annual report, 10K, or 10Q
		C	Company website
		T	Third-party website
		S	Some other source, NEC

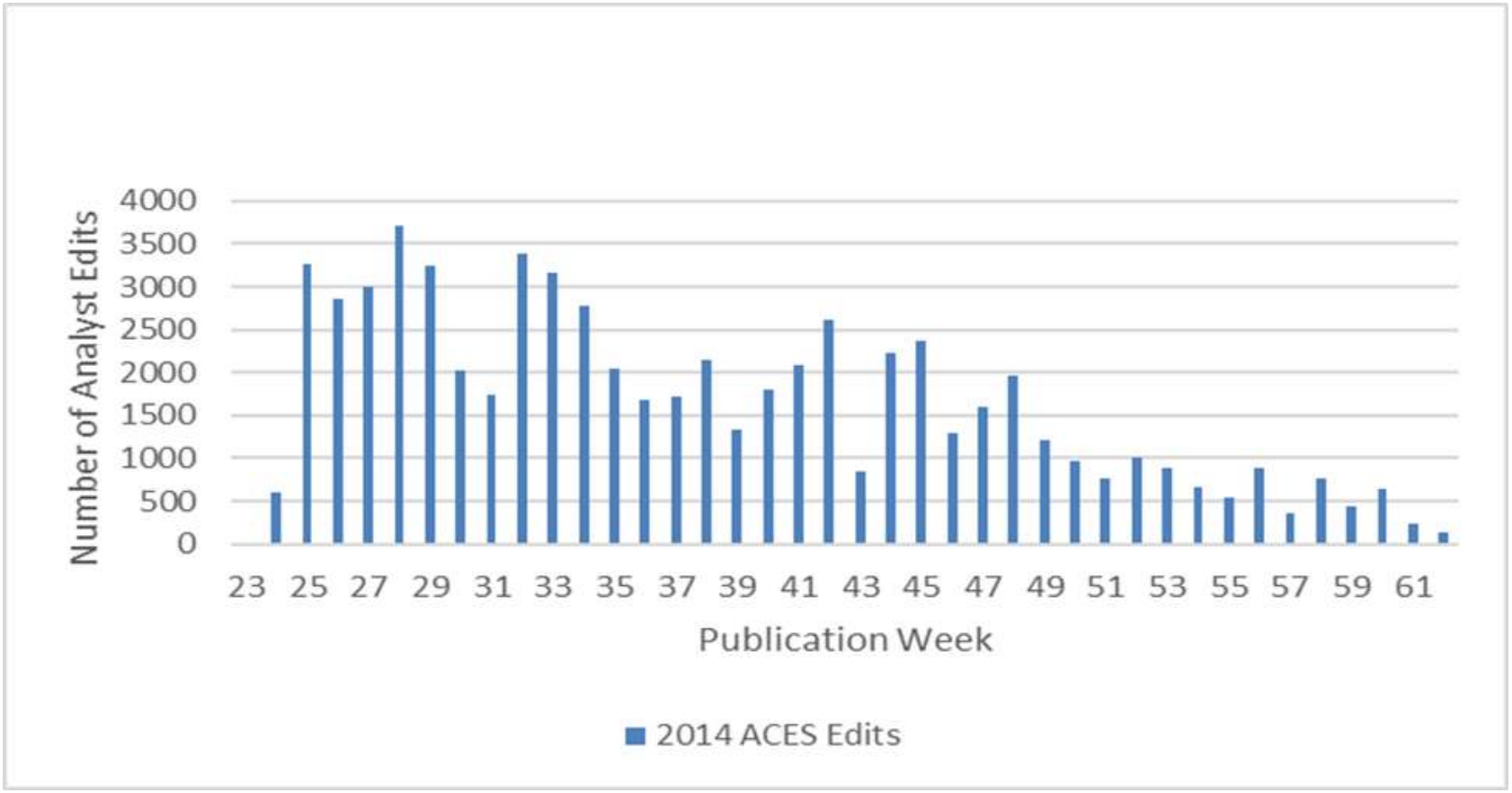
2016q1		2016q2		2016q3	
Flag	Count	Flag	Count	Flag	Count
AO	332	AO	440	AO	396
OA	322	OA	345	OA	356
ID	91	ID	112	ID	130
RF	40	RF	45	RF	58
OS	10	RX	13	OS	27
RX	6	OS	9	RK	6
OC	6	OC	9	OC	6
RK	4	AU	1	RX	6
OW	3			OI	2
AU	1				
RR	1				
IY	1				
OI	1				

Impact of Analyst Edits

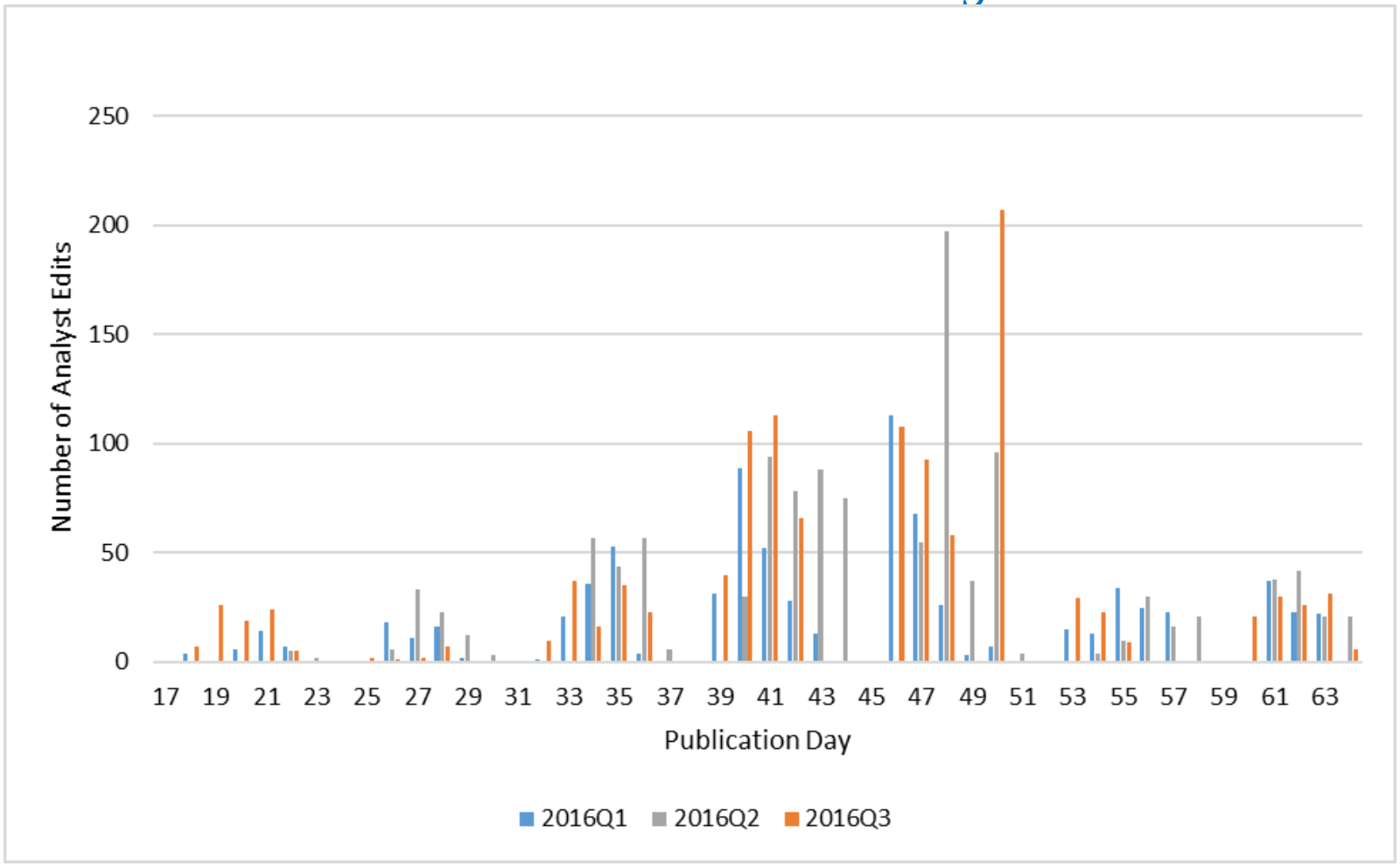
$$\boldsymbol{impact} = \frac{|\boldsymbol{wgt}*\boldsymbol{newval}-\boldsymbol{wgt}*\boldsymbol{oldval}|}{\boldsymbol{wgt}*\boldsymbol{oldval}} \times \boldsymbol{ln}|\boldsymbol{wgt} * \boldsymbol{newval} - \boldsymbol{wgt} * \boldsymbol{oldval}|$$

- Impact is a ranking measure only
- 2016Q1 flag OI (obtained data from another survey or census) was only used once but had the highest impact
- Data flag AO (analyst deduced value that reverses a reporting error, other reporting errors) was the most used flag in all 3 quarters and had a relatively large impact on estimates

Number of 2014 ACES Analyst Edits by Publication Week



Number of QSS Analyst Edits by Publication Day



Methods for Flagging Estimates For Review Weighted Robust Regression - ACES

$$estimate_t = \beta_0 + \beta_1 date_t, \quad t = 1, \ldots, T,$$

- T is the number of dates under consideration
- Regression weight set equal to the inverse square of the estimate's standard error
- The parameter T is limited to how often the estimates are created
- T is relatively small and seemed arbitrary, along with some of the other parameters used

Estimate Review Criteria - QSS

- Snapshot-to-Snapshot percent difference in the estimate is greater than 10 percent

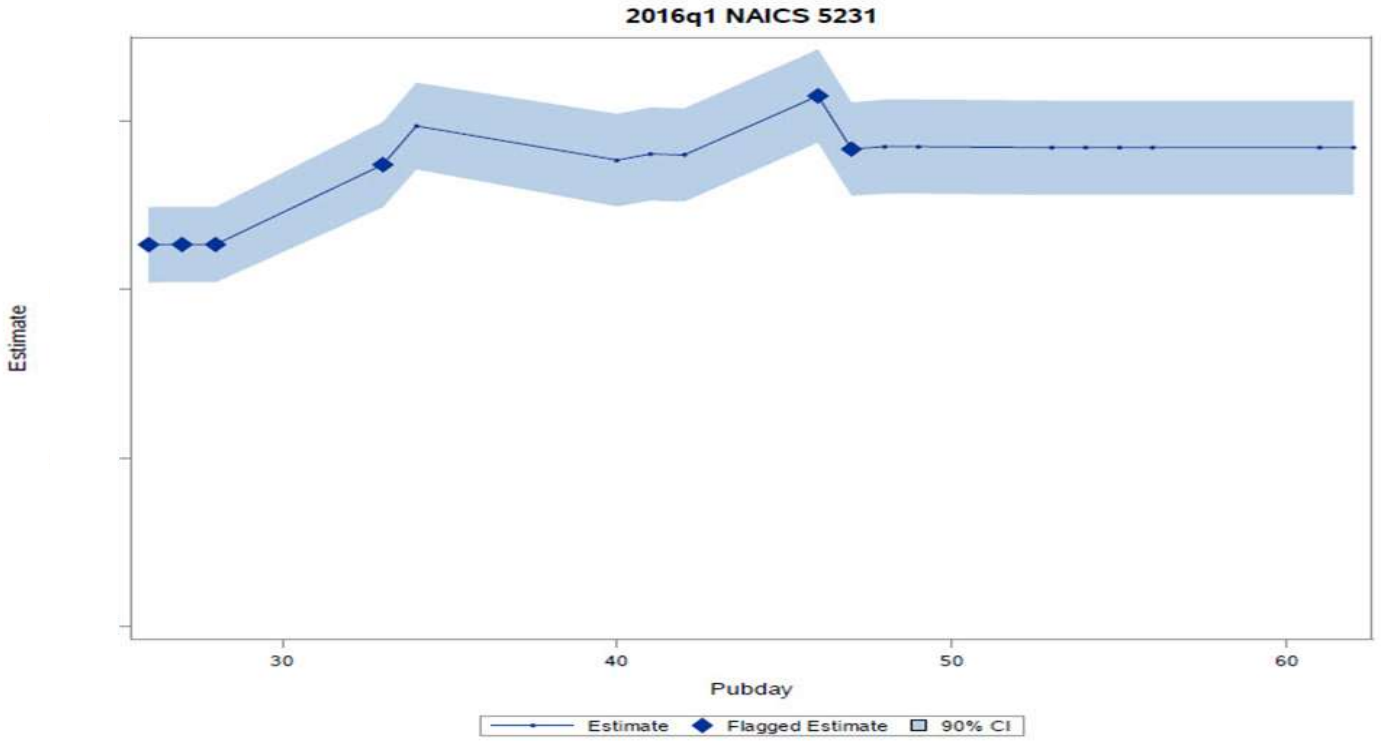
and

- Snapshot-to-Snapshot percent difference in the standard error is greater than 10 percent

and

- Percent difference between the estimate and previous year's final estimate is less than -33 percent and greater than 50 percent

2016Q1 QSS Estimates are Flagged that Violate Percent Difference Criteria



Data obtained from the 2016 Quarter 1 QSS Survey Snapshot File.
Technical documentation can be located at:
https://www.census.gov/services/qss/how_the_data_are_collected.html

Setting Priority of NAICS Codes

- Key NAICS: the corresponding percent of the overall revenue for the same quarter the previous year was greater than 1%
- discrepancy measure $=$

$$\max\left(\frac{estimate\_CQ}{estimate\_final\_PQ}, \frac{estimate\_final\_PQ}{estimate\_CQ}\right).$$

Setting Priority of NAICS Codes QSS 2016Q3

Rank	NAICS	Key NAICS	Discrepancy Measure	Criterion 1 Violation	Criterion 2 Violation	Criterion 3 Violation
1	5242	1	1.776	0	0	1
2	5221	1	1.164	0	1	0
3	5172	1	1.057	1	1	0
4	6242	0	2.149	0	0	1
5	5414	0	1.360	1	1	0
6	5182	0	1.313	1	0	0
7	5611	0	1.308	0	1	0
8	5312	0	1.188	1	0	0
9	6214	0	1.113	0	1	0
10	4853	0	1.092	0	1	0
11	6243	0	1.084	0	1	0
12	8111	0	1.037	0	1	0
13	8129	0	1.030	0	1	0

Summary of Results for QSS

- Many of the estimates are stable by publication day 50.
- Prioritization of edits could ensure that the most impactful edits are given priority.
- Automating the integration of publicly available information (example: SEC filings) would cut down on analyst burden and editing time.
- Analysis of when the less descriptive data flags in the Standard Economic Processing System (StEPS) are used could allow for the creation of better and more appropriate data flags.

Future Research

- Similarly investigate other Economic Area Surveys.
- Investigate utilizing machine learning methods to automate certain types of edits.
- Research the use of Big Data editing techniques for larger surveys and censuses.
- Continue researching stopping point models so that editing can become more adaptive.
- Investigate what circumstances lead to the use of the less descriptive data flags (AO).

Thank You

Contact information:

L. Kaili Diamond

lisa.kaili.diamond@census.gov

Brian Dumbacher

brian.dumbacher@census.gov

Thank you for your attendance and attention!