

Data Quality Assessment of Alternative Data

Brendan Williams

Senior Economist, Branch of Consumer Prices

Consumer Price Index Division

prepared for
National Academies of Sciences, Engineering, and Medicine
Committee on National Statistics
Virtual Meeting with BLS
October 30, 2020

Overview

- Data Source Selection
- Data Source Evaluation
- Aggregate Metrics



Data Source Selection



Source Selection

■ Initially priorities

- ▶ Problem items: Low response rate, measurement complexity, etc.
- ▶ Maintain respondents

■ Relevant Factors

- ▶ Size (relative importance, number of observations)
- ▶ Industry concentration
- ▶ Methodological simplicity

■ New data source selection process

- ▶ Proposed projects
- ▶ Data on relevant factors
- ▶ Field assessment of cooperativeness of top respondents
- ▶ Run biannually for list of priority respondents

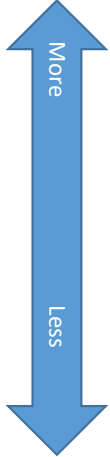
Data Source Evaluation



Preliminary Considerations

- Granularity
- Geographic Coverage
- List vs. Transaction Prices
- Timeliness and Frequency
- Sample vs. Convenience
- Census vs. Subset
- Processing: Data Cleaning, Outlier Removal

Corporate Data Preferences

		Sales data (price and quantity sold, in preference order)	Item Coverage (in preference order)	Outlet Coverage (in preference order)	Time Coverage (in preference order)
Data Granularity 	A	unique item (UI) by price point by outlet	1. All items sold 2. Sample of Items > CPI Sample 3. Items in the CPI Sample 4. Sample of Items < CPI Sample	1. All U.S. outlets in the chain 2. All outlets in the CPI PSU sample and non-self-representing PSUs not selected for the CPI sample 3. All outlets in CPI PSUs 4. All outlets in the CPI Sample	1. Pricing period averages 2. Monthly averages 3. One day in each of 3 pricing periods in the month
	B	UI by specific outlet			
	C	UI by city/PSU			
	D	Item category by specific outlet			
	E	Item category by city			
	F	Unique Item by region or national data			
	G	Item category by region or national data			

Empirical Assessments

- Market shares in data compared to CPI and industry sources
- Price matching to survey data
- Quote replacement indexes
- Basic index construction
- Multilateral index tests



Aggregate Metrics



Overview

■ Primary Quality Metrics:

▶ Response Rate

- Quote level and outlet level

▶ Precision

- Standard errors

■ Alternative Data Options:

▶ Separate Metrics for Survey Sources

▶ Combine with Existing Metrics

- Response rates: Transaction vs. list price, corporate vs. aggregator
- Precision: Estimation assumes survey data. Blending survey, census, and non-sampled estimates.

▶ Change Framework

- Total survey error

Survey Variance Methodology

- Stratified Random Groups
- Focus on variation due to item and geographic sampling, not unique product level
- Indexes differing from the item structure (“Special Relative Calculations”) use a jackknife variance estimate where each item-area component index is systematically omitted



Treatment to Date

■ Response Rates

- ▶ Dependent on quote structure

■ Variance

- ▶ National pricing (CorpX and CorpY)
 - No geographic price variation
 - Variation from sampling and differences from other respondents
 - Unclear net effect
- ▶ New Vehicles and Gas: Create replicates and use existing system

New Vehicle/Gasoline Replicates

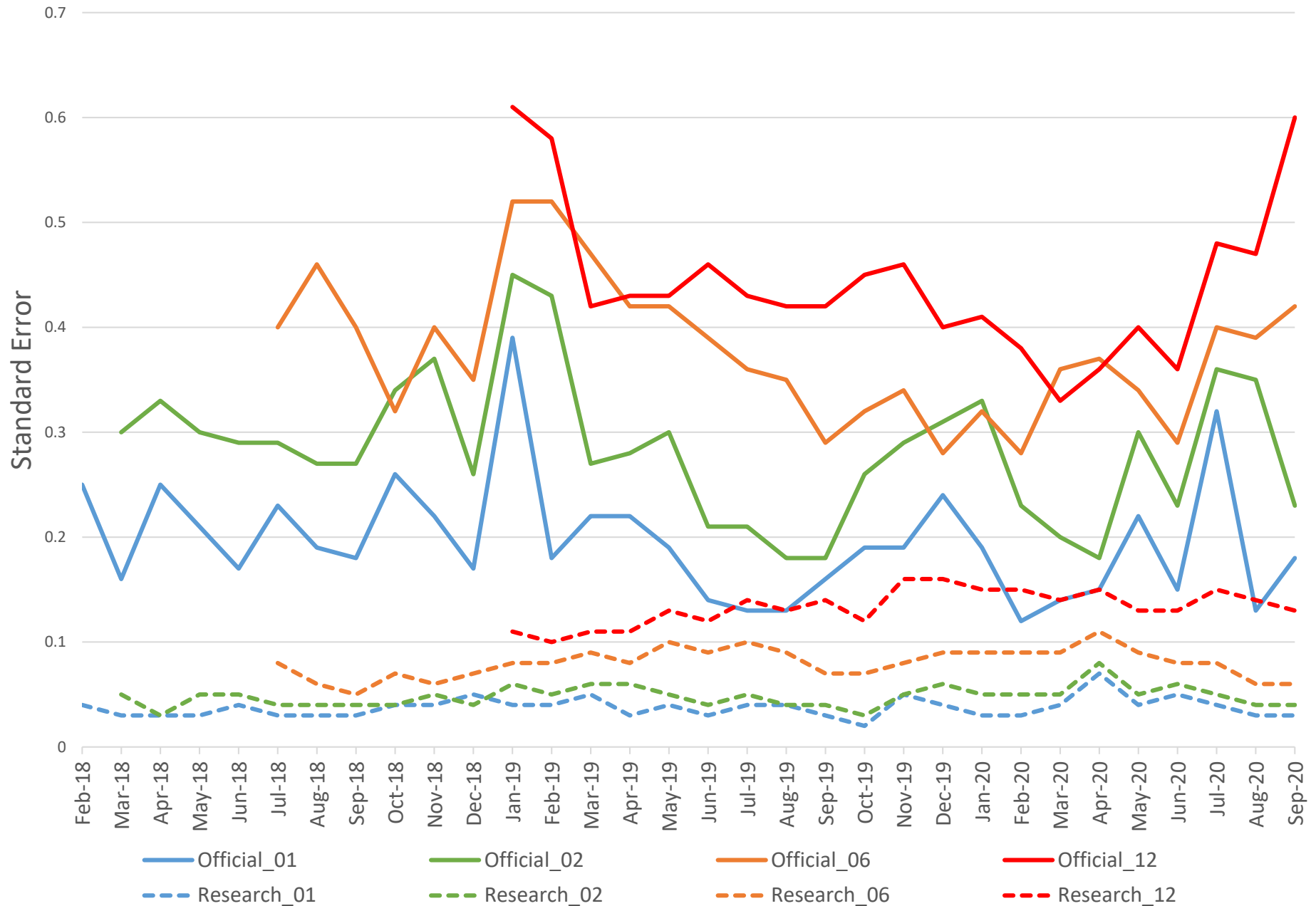
■ Self-Representing Areas

- ▶ Randomly split unique items or outlets into replicates
- ▶ Keep same observations in same replicate from month-to-month

■ Non-Self-Representing Areas

- ▶ Match to PSU to get replicate

New Vehicle Standard Errors: Official vs. Research Indexes



Problems with Replicates and Alternative Data

- Aggregation across replicates vs. observations
- Implies geographic sampling for non-sampled source
- Variation below strata level not represented
 - ▶ Transaction prices for same unique item
 - ▶ Variation among goods
- Replicate sub-samples vs. repeated resampling

Contact Information

Brendan Williams

Williams.brendan@bls.gov

