Developing Fitness for Purpose Guidelines for Alternative Data Sources

Sarah Konya

U.S. Census Bureau

2024 FCSM Research and Policy Conference

October 23, 2024

Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau.

Fitness for Purpose Team Overview

- Survey programs are looking to alternative data to supplement or replace their traditionally collected survey data
- Produce guidelines that provide a consistent approach to evaluating alternative data sources
- Reviewed:
  - Federal Committee on Statistical Methodology. 2020. *A Framework for Data Quality.*
  - U.S. Census Bureau. 2023. *Data Quality Assessment Tool for Administrative Data.* U.S. Census Bureau Statistical Quality Standards.
  - Mannshardt, E., Banks, J., Breidt, J., Finamore, J., Mirel, L., Seeskin, Z., Rice, K. 2023. *A Data Quality Scorecard to Assess a Data Source's Fitness for Use*. FCSM and IEEE Xplore.
  - Hutchinson, R. In progress. *5 Cs of Comparison*. U.S. Census Bureau.

Fitness for Purpose Guidelines

- A set of questions that survey teams should answer
- Not a pass/fail
- Asked when data have been acquired
- 7 categories

Use Cases

1. Monthly State Retail Sales (MSRS)
   - Purchased retailer point of sale data

2. Construction Re-engineering
   - Satellite imagery data to detect housing construction starts

Construction

- Understand the methodology used for the alternative data

1. Are the data raw? Or have they been edited, weighted, etc.?
2. Are the data seasonally adjusted?
3. Are the data adjusted for real dollars using an inflator/deflator index?
4. Are the data a result of modeling?
5. Are the data benchmarked to any other data?
6. How often are the data updated and revised?

Classification

- Finding and matching established categories in both the alternative data source and the survey data.

1. Can the data be mapped to the category/geography that the survey data use?

Connection

- Linkage

1. Can the data be linked to survey data or estimates at the record level?
2. Is the linkage done probabilistically or deterministically?
3. Can the data be mapped to survey data or estimates at the item level?

Coverage

Unit level

Match Rate

Geography	Match Rate
Region 1	
Region 2	
Region 3	
Region 4	

Sex	Match Rate
Female	
Male	

Item Level

Survey Missing Rate	Alternative Data Missing Rate

Geography	Survey Missing Rate	Alternative Data Missing Rate
Region 1		
Region 2		
Region 3		
Region 4		

Sex	Survey Missing Rate	Alternative Data Missing Rate
Female		
Male		

Comparability

- Micro level
- Macro level
- Conducted on the unit level matches for every item

Comparability: Micro-Level Analysis

Numeric variable

$$d = \left| \frac{\text{Alternative data value} - \text{Survey data value}}{\text{Survey data value}} \right| * 100$$

Absolute Percentage Diff. (d)	Number of Records	% of Records
0%		
≤10%		
10% < d ≤ 20%		
20% < d ≤ 30%		
30% < d ≤ 40%		
40% < d ≤ 50%		
d > 50%		
Missing in Alternative Source and not in Survey Data		
Missing in Survey Data and not in Alternative Source		
Missing in Both Sources		

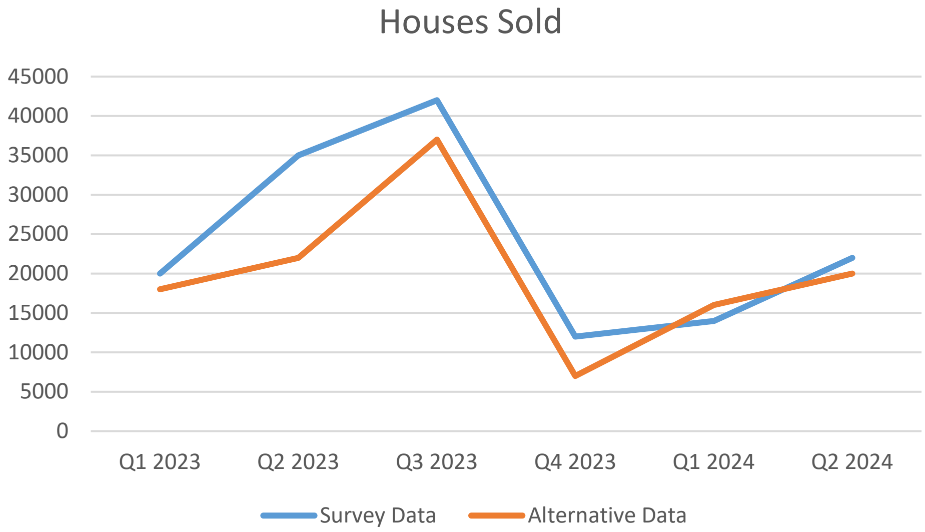
Categorical variable

Match Status	Number of Records	% of Records
Matches		
Non-matches		
Missing in Alternative Source and not in Survey Data		
Missing in Survey Data and not in Alternative Source		
Missing in both Sources		

Comparability: Macro-Level Analysis

Data Source	Total Revenue
Survey Data	$123 million
Alternative Data	$102 million

Race	Survey Data Distribution	Alternative Data Distribution
White	56%	54%
Black	7%	10%
Asian	8%	4%
NHPI	1%	1%
AIAN	1%	0%
SOR	2%	0%
Two or More	5%	3%
Missing	20%	28%



Notional data

Comparability: Statistical Analysis

- Confidence intervals
- T-tests to compare two point estimates
- Chi-square test to compare distributions
- Patterns of Missingness
- Sign directions
- Regression analysis
- Sensitivity and specificity analysis

Consistency

1. Is the time series consistent over time? Are extreme changes explainable?
2. Is the coverage consistent over time?
3. How long have these data been available and do we think these data are going to be available long into the future?
4. How often is a new sample selected? Is there an overlap sample analysis conducted between samples?

Continuous Evaluation

- The evaluation should be repeated regularly to determine if the quality of the alternative data is acceptable.

1. Over time, do the results of this evaluation remain acceptable?
2. Are the users (Census employees and data users) satisfied with the alternative data source and resulting data product?
3. Is the data easy to use in practice?
4. Does the cost of acquisition remain acceptable?

Final Determination of Fitness

- Showstoppers along the way?
- Quantitative analysis may result in only a subset of the data being usable
- Save results so they can be revisited during Continuous Evaluation and shared with other teams

Next Steps

- Creating a template to store responses to questions
- Share throughout the Census Bureau
  - Continuously update as survey teams provide feedback
- Coordinating reviews for multiple survey teams with the same data
- Automate some of these checks?
- Blended products methodology standards

Questions?

sarah.konya@census.gov