Multi-Source Hierarchical Models for Geographically Granular Retail Sales Estimates

Darcy Steeg Morris*
U.S. Census Bureau

Joint Statistical Meetings

August 4, 2024

* with Census Bureau colleagues Brian Dumbacher, Carma Hogue, Stephen Kaputa, and Jenny Thompson.

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Disclaimer

This presentation is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not those of the U.S. Census Bureau.

Challenges in Survey Estimation

Survey programs are facing challenges in producing official estimates, due in part to:

- declining response rates,
- increasing operational costs,
- demand for more timely/geographically granular data products, and
- data collection interruptions.

Alternative Data Sources at Census

Concurrently, continual advancements in availability and linkability of

- third party data (e.g. data aggregators, business transactions), and
- administrative records (AR), e.g. IRS income tax records.

Sources complicated by *big data* nature (AAPOR 2015, CNSTAT 2023):

- not designed by researchers,
- generated for a different purpose, and
- arise organically through administrative/business processes.

They offer relevant information, but proceed with caution.

Statistical Modeling to Integrate Sources

Caution with direct/exclusive reliance on "big data" (bias!) – instead indirect use via familiar statistical modeling techniques.

Such methods can be used to combine information across surveys and third party data.

These case studies adopt two traditional small area estimation and missing data methods to integrate data sources via statistical modeling:

- **ratio-synthetic estimation** – allocate domain estimate to a subdomain by a fixed proportion, and
- **imputation modeling** – draw plausible value from statistical model at the unit level and aggregate to subdomain.

Third Party Data for Retail Sales Estimates

Monthly Retail Trade Survey (MRTS) is designed to produce retail sales estimates

- monthly
- at the national level
- by industry.

Production Goal: Produce monthly retail sales estimates at the *state-level* by industry, but the survey is designed for national-level!

A Solution: Use sub-national third party purchase transaction data to achieve more geographic granularity via statistical modeling.

Question: How to use third party "big data" even though it is likely not representative of the target population?

Ratio-Synthetic: Aggregated First Data

Imputation-Based: Establishment-Level NPD Data

Ratio-Synthetic + Imputation-Based: Aggregated and Establishment-Level NPD Data

Ratio-Synthetic: First Data Case Study

Dumbacher, Morris, and Hogue (2019), Journal of Big Data.

First Data (FD) is a large payment processor: 72 billion transactions & \$1.9 trillion per year, 10% US GDP in 2016.

Provided **state-level aggregated** sales $y_g^{(FD)}$ & merchant counts $n_g^{(FD)}$

- by industry,
- by month (January 2012 – March 2018),
- protected merchant identities,
- some suppressed data for confidentiality,
- may reflect business activity more than economic activity.

Ratio-Synthetic: Exploratory Data Checks

Understanding uncertain and imprecise nature of the data is critical, but there is no gold standard!

Devise quality criteria to understand *representativeness*

- coverage compared to sub-national 2012 Economic Census estimates,
- coverage compared to national, monthly MRTS estimates,

and *consistency*

- region-to-region FD trends,
- FD region-to-national MRTS trends.

Geographic departures from MRTS are precisely the value of FD data, but don't want to mistake business for economic activity.

Ratio-Synthetic: Exploratory Findings

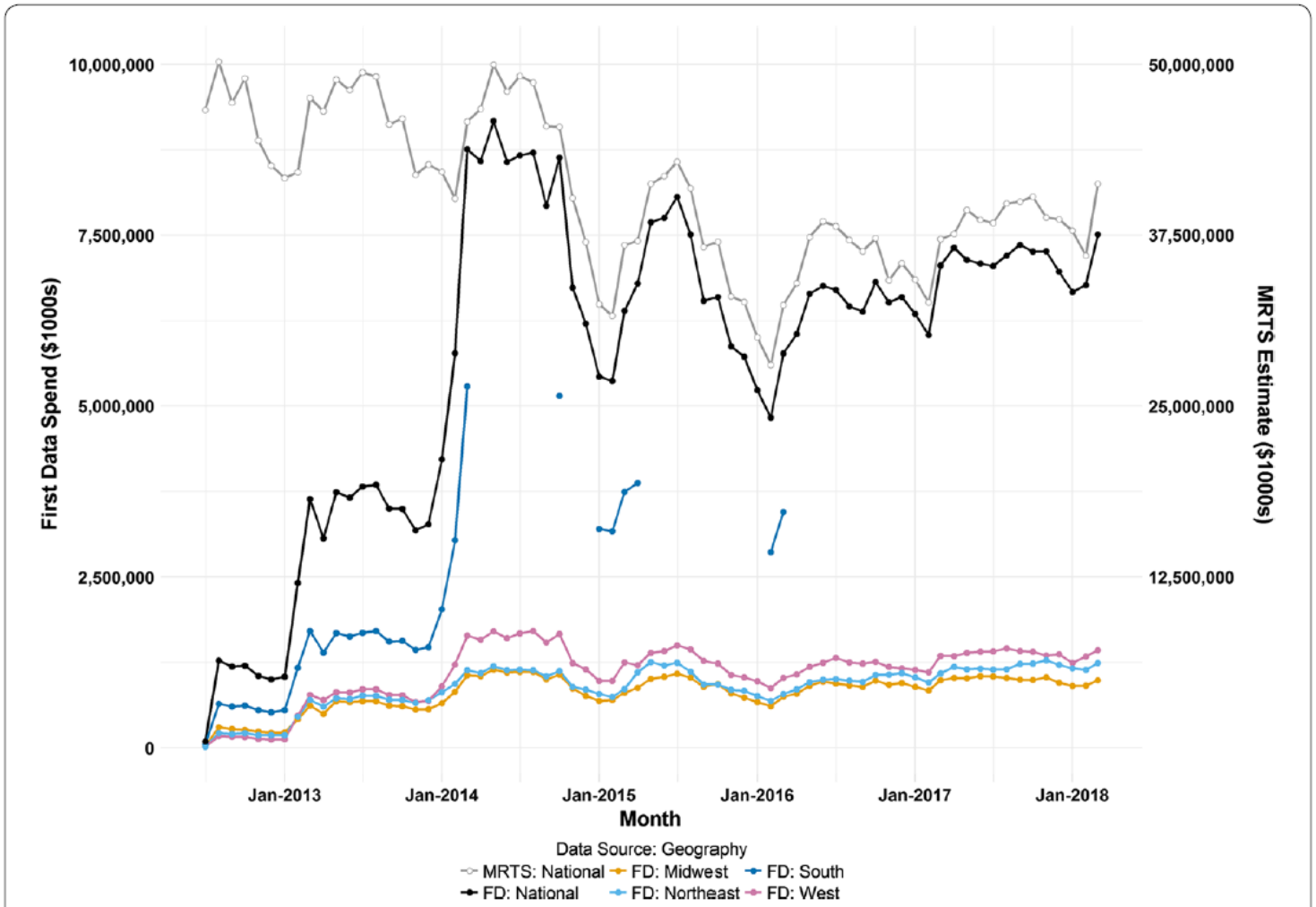


Fig. 3 National Monthly Retail Trade Survey (MRTS) estimates with regional and national First Data (FD) aggregate spend over time for NAICS 447 (gasoline stations)

Ratio-Synthetic: Experimental Estimates

Model only 5 of 13 retail industries that have *acceptable* quality.

State-level linear mixed models within each industry & month with

- *Dependent Variable* $Y_g^{Ratio-Synthetic} \rightarrow$ national MRTS estimate allocated to state according to FD sales data,

$$\frac{n_g^{(BR)}}{n_g^{(FD)}} \times y_g^{(FD)} \times \frac{1}{\sum_g \left(n_g^{(BR)} / n_g^{(FD)} \right) y_g^{(FD)}} \times Y^{(MRTS)}$$

where n represents number of establishments, BR is the business register, g indexes state, and Y & y are measures of total sales.

Ratio-Synthetic: Experimental Estimates

- *Independent Variables* $X_g \rightarrow$ quarterly earnings (BEA), population (Census), retail sales (2012 Economic Census).
- *Random Effects* $u_{g'} \rightarrow$ geographic division (region).
- *The Model* $\rightarrow$

$$\ln\left(Y_g^{Ratio-Synthetic}\right) = X_g\beta + u_{g'} + \epsilon_g$$

which yields model-based state-level estimate $\hat{Y}_g^{(Ratio-Synthetic)}$ that smooths artifacts of FD business activity & take advantage of timeliness of transaction data.

Ratio-Synthetic: Experimental Estimates

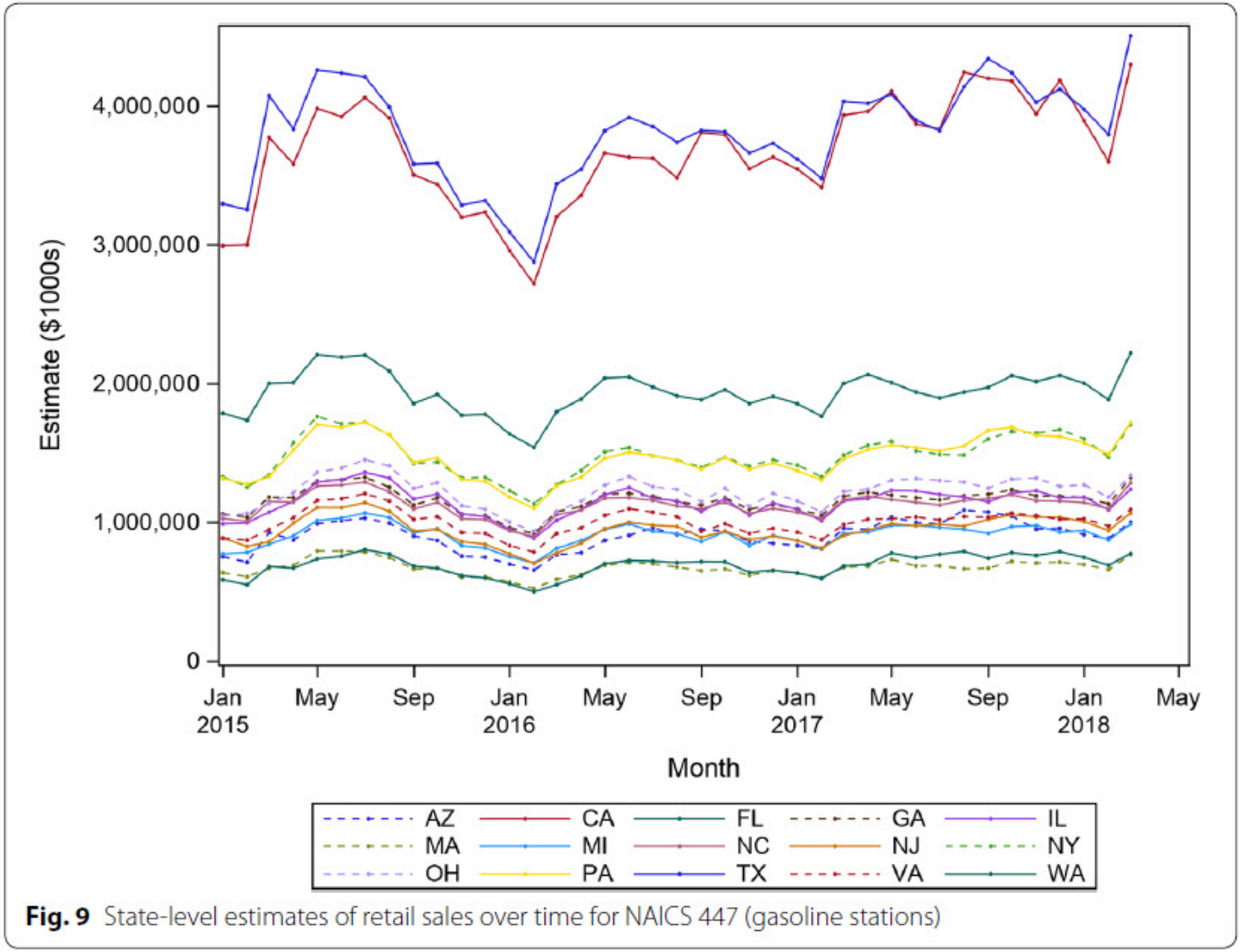


Fig. 9 State-level estimates of retail sales over time for NAICS 447 (gasoline stations)

Ratio-Synthetic: Summary

- Comparisons (e.g. simple correlations) with proxy and available external estimates (e.g. number of employees) are reasonable.
- Many features of experimental estimates seem reasonable, but there are caution flags.
- FD transaction data offers insight into sub-national economic activity.
- But the use for enhancing official estimates is challenging because of limitations in representativeness.
- Can only indirectly assess bias when business identities are unknown...

Modeling Approach and Case Study

Ratio-Synthetic: Aggregated First Data

Imputation-Based: Establishment-Level NPD Data

Ratio-Synthetic + Imputation-Based: Aggregated and Establishment-Level NPD Data

Imputation-Based: NPD Case Study

Kaputa, Morris, and Holan (2024), Journal of Survey Stats. & Method.

The market research company NPD Group (now Circana) provided **establishment-level aggregated** point-of-sale transaction data

- by industry,
- by month,
- for all establishments in a known set of 22 multi-unit companies,
- with unique establishment identifiers and precise geographic information.

See Hutchinson et al. (2023) for details on the data.

Imputation-Based: Imputation Model

Model establishment-level NPD data to mass (multiply) impute sales for establishments in the entire MRTS sampling frame.

National linear mixed models within each industry & month with

- *Dependent Variable* $y_{j(g)}^{(NPD)} \rightarrow$ establishment-level j sales from NPD.
- *Independent Variables* $X_{j(g)} \rightarrow$ payroll (BR), detailed industry.
- *Random Effects* $u_g \rightarrow$ state with ICAR structure.

Imputation-Based: Imputation Model

- *The Model* $\rightarrow$

$$\ln\left(y_{j(g)}^{(NPD)}\right) = X_{j(g)}\beta + u_g + \epsilon_{j(g)},$$

  which yields an imputed value $Y_{j(g)}^{(Impute)}$ for all establishments in the MRTS sampling frame.

Establishments without NPD data are imputed based on relationships from establishments with NPD data.

Imputation-Based: Estimate

The state-level estimate is obtained by aggregating establishment-level third party and imputed sales values by state:

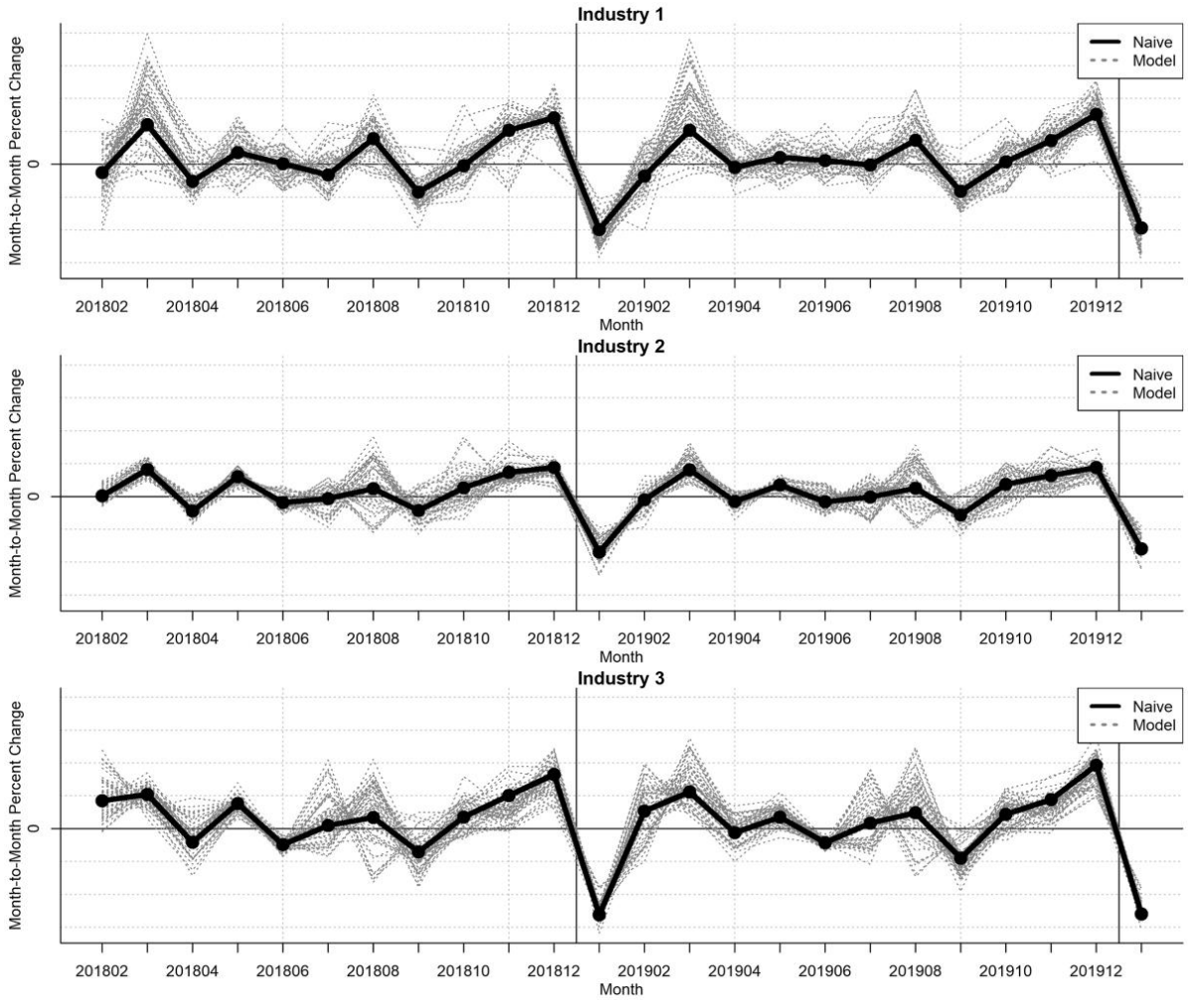
$$\hat{Y}_g^{(Impute)} = \sum_{j \text{ in } g} y_{j(g)}^{(NPD)} \mathrm{I}_{(\text{has NPD data})} + \sum_{j \text{ in } g} Y_{j(g)}^{(Impute)} \mathrm{I}_{(\text{no NPD data})}.$$

For MRTS respondents, MRTS sales $Y_{j(g)}^{(MRTS)}$ are used for:

- single unit companies, and
- single-establishment multi unit companies,

instead of NPD-based sales values $y_{j(g)}^{(NPD)}$ and $Y_{j(g)}^{(Impute)}$

Imputation-Based: Month-to-Month Change



Payroll Ratio-Synthetic (Black) vs Imputation-Based (Grey), Kaputa et al. (2024)

Imputation-Based: Summary

- No official benchmark – accuracy conclusions are exploratory.
- NPD data offers insight into sub-national economic activity.
- Establishment-level data allows survey frame checks of representativeness $\rightarrow$ not for company-deidentified aggregated data.
- As such, can indirectly enhance official estimates through a variety of modeling procedures from the unit-level on up.

Modeling Approach and Case Study

Ratio-Synthetic: Aggregated First Data

Imputation-Based: Establishment-Level NPD Data

Ratio-Synthetic + Imputation-Based: Aggregated and Establishment-Level NPD Data

Ratio-Synthetic + Imputation-Based: MSRS

The Monthly State Retail Sales (MSRS) experimental estimates are our first blended data product measuring the rapidly evolving economy.

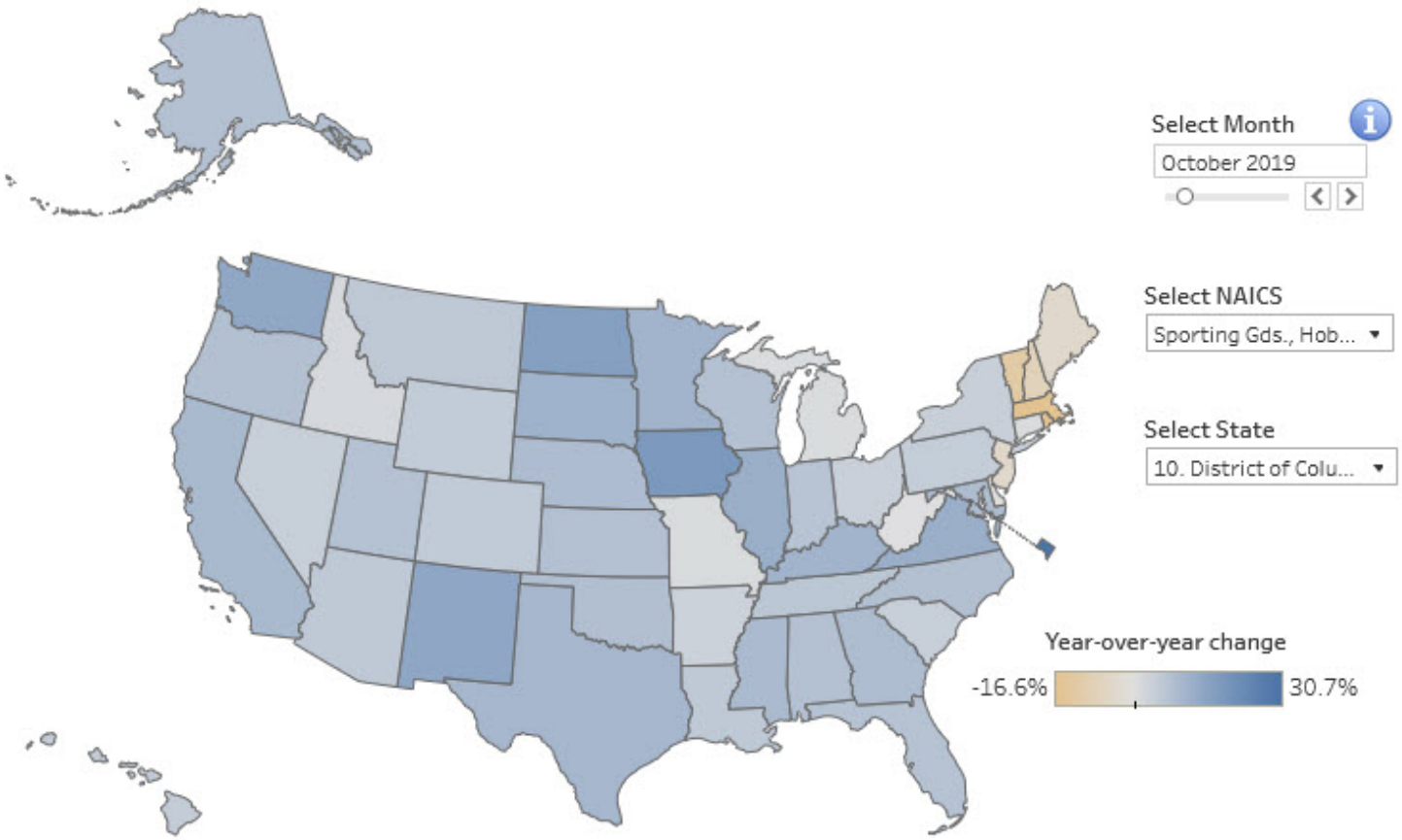
Based on *composite estimate* combining payroll-adjusted (not third party) ratio-synthetic estimate and imputation-based estimate

$$\hat{Y}_g^{(Composite)} = \phi \hat{Y}_g^{(Ratio-Synthetic)} + (1-\phi)\hat{Y}_g^{(Impute)},$$

weighted by relative variance ϕ of estimator [$\nearrow$ variance $\Rightarrow$ $\searrow$ weight].

Composite estimates are further adjusted to national MRTS totals.

Ratio-Synthetic + Imputation-Based: MSRS Year-over-Year



https://www.census.gov/retail/state_retail_sales.html

2018 World Series: Boston Red Sox* vs. LA Dodgers, 2019 World Series: Washington Nationals* vs. Houston Astros

Summary

- Blending survey and third party data yields opportunity to produce detailed, timely, and geographically granular estimates.
- Each case requires understanding/assessing the fitness for use of third party data $\rightarrow$ there is no one-size-fits-all standard for quality and accuracy checks (AAPOR 2015, CNSTAT 2023).
- Statistical modeling offers principled, indirect way to extract geographic signal for producing subdomain official statistics.
- Geographic random effects models fit with off-the-shelf Bayesian software.
- Composite estimates add layer of protection – can help balance advantages and disadvantages of different blended data methods.

Select References

- Committee on National Statistics (2023) "Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources." https://nap.nationalacademies.org/catalog/26804.
- Dumbacher, B., Morris, D.S., and Hogue, C. (2019) "Using Electronic Transaction Data to add Geographic Granularity to Official Estimates of Retail Sales." *Journal of Big Data*.
- Hutchinson, R., Scheleur, S., and Weidenhamer, D. (2023) "Alternative Data Sources in the Census Bureauâs Monthly State Retail Sales Data Product", Chapter 26 pp. 593â611.
- Kaputa, S.J., Morris, D.S., and Holan, S.H. (2024) "Bayesian Multisource Hierarchical Models with Applications to the Monthly Retail Trade Survey." *Journal of Survey Statistics and Methodology*.
- Kim, J.K., Park, S., Chen, Y., and Wu, C. (2021) "Combining Non-Probability and Probability Survey Samples through Mass Imputation." Journal of the Royal Statistical Society: Series A.
- The American Association for Public Opinion Research (2015) "AAPOR Report on Big Data." Available at https://aapor.org/wp-content/uploads/2022/11/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.
- U.S. Census Bureau (2020) "Monthly State Retail Sales Methodology." https://www.census.gov/data/experimental-data-products/monthly-state-retail-sales.html

Thank you!

darcy.steeg.morris@census.gov