From Checkout to Data Product: Using Retail Point-of-Sale Third-Party Data in U.S. Census Bureau Data Products

Rebecca Hutchinson, Shelby Plude, and Edward Watkins [1]
U.S. Census Bureau

Abstract

The demand for more timely data in an evolving retail economy coincides with falling survey response rates often due to respondents citing burden. As part of its big data vision, the United States Census Bureau is exploring where alternative data sources could be used to supplement existing data products and/or create new data products while maintaining the high quality of its official statistics. For the retail sector of the economy, point-of-sale data is one potential alternative data source. Point-of-sale data were acquired from the NPD Group, Inc. at the national, store, and product levels; these data were compared to data in the Monthly Retail Trade Survey, Annual Retail Trade Survey, and the 2012 Economic Census. Results from this effort were promising with good comparison results for both retailers that report to the survey and retailers that are nonrespondents. These data are now being used in the survey programs to mitigate nonresponse. Because data at the individual retailer level can be difficult to obtain due to cost and consent issues, the Census Bureau has also acquired point-of-sale data aggregated at the state level for groupings of retailers. Both the store-level data and the aggregated state-level data are being used as inputs to a new experimental data product, the Monthly State Retail Sales, and allowed for the creation of this data product without needing new data collections.[2]

Introduction

The U.S. Census Bureau has long produced high-quality official statistics for the retail trade sector.[3] These data are obtained through traditional survey data collection, such as the Monthly Retail Trade Survey (MRTS), the Annual Retail Trade Survey (ARTS), or the quinquennial Economic Census, and are a critical input to the calculation of the Gross Domestic Product (GDP) of which retail trade makes up nearly 25% of the 2019 GDP estimate (Bureau of Economic Analysis 2020). The retail data are also critical to Census Bureau data users as they analyze the current state of a retail sector that is facing store closures, industry disrupters, and e-commerce growth. To continue to meet this need for high-quality official statistics while also exploring opportunities for filling data product gaps, the Census Bureau's retail trade survey program is exploring the use of alternative data sources to produce higher-frequency and more geographically-detailed data products, to supplement traditional survey data collection, to ease

---

[1] Disclaimer: Any views expressed are those of the authors and not necessarily those of the United States Census Bureau. Census Bureau has reviewed Monthly State Retail Sales product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY21-ESMD002-033)

[2] Census Bureau experimental data products are statistical products that were created using new data sources or methodologies to benefit our data users that may not meet all of the Census Bureau's data quality standards. Experimental data products are clearly identified as such and the Census Bureau encourages data users to submit feedback on the data products. Experimental data products may or may not become official statistical products depending upon demand and data quality. More information on experimental data products is available at What are Experimental Data Products? (census.gov)

[3] The production of quality statistics is the principal goal of the U.S. Census Bureau. The Commerce Department (2014) lists the criteria government statistics must meet: comprehensive, consistent, confidential, credible, and relevant. The Census Bureau strives to meet these criteria.

respondent burden, and to assist with declining response rates (United States Census Bureau 2018).

Alternative data sources for retail may include point-of-sale data, credit card data, and payment processor data. In 2016, the Census Bureau conducted a pilot project to test if retailer point-of-sale data from The NPD Group, Inc. (NPD) could be used in place of the data reported by retailers to the monthly and annual retail surveys (Hutchinson 2020). The positive results of that project led to the acquisition of more third-party data.

Point-of-sale data, also known as scanner data, are detailed sales data for consumer goods that are obtained by scanning the bar codes or other readable codes on the products at electronic points-of-sale both in physical store locations and online (Organization for Economic Co-operation and Development 2005). Point-of-sale data offer important advantages relative to other types of third-party data. Point-of-sale data can provide information about quantities, product types, prices, and the total value of goods sold for all cash and card transactions in a store. These data are available at the retailer, store, and product levels. By contrast, credit card data or payment processor data are often only available at an aggregated level; due to confidentiality agreements, information about the retailer composition of these data is rarely available. Additionally, cash sales are excluded from both credit card data and payment processor data but are included in point-of-sale data.

The working hypothesis for the use of point-of-sale data is that if all items that a retailer sells are captured in a point-of-sale data feed, then the sum of those sales across products and store locations over a month or over a year should equal the total retail sales for a retailer for that same time period. If the hypothesis holds, the sales figure from the point-of-sale data should be comparable to what is provided by a retailer to Census Bureau retail surveys. When used for this purpose, a point-of-sale dataset needs to identify the data by retailer name, provide product-level sales for each retail store location, and have data available by month.

Retailer point-of-sale data feeds can be obtained either directly from a retailer or through a third-party vendor. While the raw data from either source should be identical, there are advantages and disadvantages to both (Boettcher 2014). A third-party vendor will clean and curate the data in a consistent format to meet its data users' needs but often at a high cost. These high costs pose a major challenge to the scalability of the effort as it can be difficult to find a third-party data source that both covers the scope of a survey program and can be obtained under budget constraints (Jarmin 2019).

Point-of-Sale Data

Point-of-sale data for this project were provided by NPD.[4] NPD is a private market research company that captures point-of-sale data for retailers around the world and creates market analysis reports at detailed product levels for its retail and manufacturing partners. [5]

[4] The NPD Group, Inc was selected as the vendor for this project through the official government acquisitions process.

[5] By providing the data to NPD, retailers have access to NPD-prepared reports that help retailers measure and forecast brand and product performance as well as identify areas for improved sales opportunities.

NPD receives, processes, edits, and analyzes weekly or monthly data feeds containing aggregated transactions by product for each individual store location of its retailers.[6] These data feeds include a product identifier, the number of units sold, product sales in dollars, and the week ending date.[7] Sales tax and shipping fees collected are excluded. Any price reductions or redeemed coupon values are adjusted for prior to the retailer sending the feed to NPD so the sales figures in the feed reflect the final amount that customers paid. Data from NPD are limited to stores located in the continental United States.

Because its market analysis reports are done at the product level, NPD's processing is driven by its product categories. NPD processes data for many product categories including apparel, small appliances, automotive, beauty, fashion accessories, consumer electronics, footwear, office supplies, toys, and jewelry and watches. NPD only classifies data for those products in the product categories listed above and sales from any items that do not belong in these categories are allocated to an unclassified category. For example, NPD currently does not provide market research on food items; all food sales data are tabulated as unclassified.

As part of the acquisition process, the Census Bureau provided dataset requirements to NPD and NPD curated datasets from their data feeds based on these requirements. Retailer datasets received by the Census Bureau from NPD contain monthly data at the store and product levels with monthly sales available for each product, store location, and retailer combination. The datasets include values for the following variables: time period (month/year), retailer name, store number, ZIP code of store location, channel type (brick-and-mortar or e-commerce), product classification categories, and sales figures. One observation for each month and each store location is the total sales value of the unclassified data.

The Census Bureau and NPD work together to onboard retailers to the project. From a list of retailers that provide data feeds to NPD, the Census Bureau selects retailers whose data would be most useful to this project. Retailers that consistently report to the MRTS, the ARTS, or the 2012 and/or 2017 Economic Census are useful for baseline comparisons. Priority is also given to selecting MRTS non-respondents as this voluntary survey is the most timely measure of retail sales and response is critical to survey quality.[8] High-burden retailers are also considered a priority.[9]

NPD needs to obtain signed agreements with retailers to share data with the Census Bureau. NPD utilizes its retailer client contacts to reach out to retailers. The Census Bureau provides a letter to the retailers detailing the goals of the project, including reducing respondent burden and improving data accuracy. The letter informs retailers that any data obtained from NPD is protected by the United States Code Title 13 such that it is kept confidential and is used only for statistical purposes.[10] Retailer participation in this effort is voluntary and some retailers do

[6] Some retailers do not provide individual store location feeds to NPD and just provide one national feed.
[7] NPD does not receive information about individual transactions or purchasers.
[8] Response rates to the Monthly Retail Trade Survey have fallen from 74.6% in 2013 to 66.5% in 2017.
[9] High-burden retailers are those retailers that receive a large number of survey forms from survey programs across the Census Bureau including the Annual and Monthly Retail Trade Surveys.
[10] Both to uphold the confidentiality and privacy laws that guide Census Bureau activities, a small number of NPD staff working on this project completed background investigations and were granted Special Sworn Status. These NPD staff are sworn to uphold the data stewardship practices and confidentiality laws put in place by United States Codes 13 and 26 for their lifetimes.

decline to participate. Declining retailers cited a variety of reasons including legal and privacy concerns; others stated that completing Census Bureau surveys is not burdensome.

Once a retailer agrees to share data, NPD delivers a historical data set of monthly data for the retailer back to 2012 or the earliest subsequent year available within 30 days from when the retailer, the Census Bureau, and NPD all sign the agreement of participation.[11] Subsequent monthly deliveries of retailer data are made 10-20 days after month's end. NPD datasets do not require much cleaning as the file formats, variables, and contents were specified in detail in the terms of the contract. Upon delivery, the Census Bureau first verifies contractual requirements are met. This process verifies that the product categories, store locations, retailer channels, and other categorical variables have remained consistent over time.

When working with third-party data providers like NPD, obtaining data for individual retailers can be difficult as the data are expensive and it is difficult to obtain permission from individual retailers to allow NPD to share their data with Census. During the creation of the Monthly State Retail Sales (MSRS) data product, the Census Bureau brainstormed what other more aggregated datasets could be curated by NPD without needing the permission of individual retailers and could be used as an input to the MSRS (Hutchinson, et al. In Press). The brainstorm led to the creation of aggregated state-level sales data for groupings of retailers. The Census Bureau provides NPD with a list of retailers for each retail North American Industry Classification System (NAICS) subsector and asks NPD to deliver a dataset containing one total monthly sales figure for each state for each grouping of retailers. For example, we would provide a list of retailers in retail NAICS subsector 444 that includes Retailer A, Retailer B, and Retailer C and NPD would deliver a monthly dataset with a value equal to the sum of sales for Retailers A, B, and C for each individual state.

Quality Review

Before the NPD data were deemed fit for use in any way, the data were run through a variety of quality checks. The quality review process focuses on determining how well the NPD data align with data collected or imputed by the Census Bureau's retail trade programs. For both the individual retailer and aggregated state-level data, national-level NPD sales for each retailer are compared against what the retailer reports to the MRTS and the ARTS. There are currently no official or standardized quality measures in place to deem a retail third-party data source's quality acceptable so developing a quality review process for third-party data sources is an important research goal. To date, the decision to use or not to use a retailer's data has relied heavily on retail subject matter expertise.

The review of a retailer's data begins with a simple visual review of the time series properties of the data, plotting the monthly NPD data against the MRTS data.[12] Issues with both the NPD data as well as the MRTS data have been identified during this visual review. To date, the issues identified were unique to the individual retailer and each issue required specific research.

[11] NPD will sometimes acquire data from other data providers. When these acquisitions occur, there is no guarantee that the full time series for the retailer will be available to NPD to process and share. In these scenarios, NPD provides data beginning with the earliest year available after 2012.

[12] Comparisons are done to the MRTS due to the large number of data points available (currently 60-84 monthly data points per retailer versus 5-7 annual data points).

The aggregated state-level data were reviewed in a similar manner. Because the MRTS collects data at the national level for retailers, we did not have an internal data source to check the state-level data in aggregated groupings against. Instead, we compared the aggregated groupings of the retailers at the national level against the aggregated national data for the same groupings of retailers in the MRTS.

Overall, the National-Level data align well between the NPD data and the MRTS data. Given the volume of data ingested, some data issues—particularly data points near the beginning of the time series—may be too far removed to be resolved. This is one challenge with committing to third-party data to replace a Census collection: determining its accuracy may not always be obvious from the exploration of time series properties.

Conclusion

This project has demonstrated potential for the use of point-of-sale data not only to reduce respondent burden and supplement existing Census Bureau retail surveys but also to create new data products. Beginning with the October 2018 MRTS estimates, NPD data for a small number of retailers who do not report to the survey were included in the estimates (United States Census Bureau 2019). The NPD data are also used by the Annual Retail Trade Survey and the Economic Census to assist with nonresponse. NPD data for the consistent reporters is used to verify reported survey data and we are developing retailer quality review profiles to guide the decision to use the NPD data and allow a retailer to stop reporting sales on Census Bureau retail surveys. The NPD data provide an opportunity not only to help with respondent burden and survey non-response but also to help produce more timely and more granular estimates. The Census Bureau started publishing the Monthly State Retail Sales data product, a blended and modeled data product that would not be possible without the use of third-party data sources like NPD. The cost and added respondent burden of obtaining state-level retail sales data on a monthly basis through traditional survey collection has always been prohibitive.

Despite the positive results in using these NPD data, there are numerous risks and challenges involved. Third-party data like the NPD data are expensive and the Census Bureau is subject to budget limitations which limits how much data can be purchased. Second, third-party data providers have coverage of only certain parts of the retail economy and improving coverage requires working with a diverse pool of retailers which allows the Census Bureau to not rely on a single data source but also increases the costs associated with the work. And last, retailers can stop agreeing to provide NPD with their data at any time which would end data flows to the Census Bureau.

References

Boettcher, Ingolf (2014). One size fits all? The need to cope with different levels of scanner data quality for CPI computation. Paper from the UNECE Expert Group Meeting on CPI. (26-28 May). Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/WS4/WS4_04_One_size_fits_all.pdf

Bureau of Economic Analysis (November 25, 2020). Gross Domestic Product, Third Quarter 2020 (Second Estimate), Table 3. Retrieved from: https://www.bea.gov/sites/default/files/2020-11/gdp3q20_2nd_0.xlsx

Department of Commerce (July 2014) Federal Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data [PDF File]. Washington, DC. Retrieved from: https://www.commerce.gov/sites/default/files/migrated/reports/revisedfosteringinnovationcreatingjobsdrivingbetterdecisions-thevalueofgovernmentdata.pdf

Hutchinson, Rebecca J. (2020). Investigating Alternative Data Sources to Reduce Respondent Burden in United States Census Bureau Retail Economic Data Products. In Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg. *Big Data Meets Survey Science: A Collection of Innovative Methods*, 359-385. Hoboken, New Jersey: John Wiley & Sons, Inc.

Hutchinson, Rebecca J., Scott Scheleur, Deanna Weidenhamer (In Press). Alternative Data Sources in the Census Bureau's Monthly State Retail Sales Data Product. In the Sixth International Conference on Establishment Statistics (ICES VI) edited volume. Hoboken, New Jersey: John Wiley & Sons, Inc.

Jarmin, Ron S. (2019). Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics. *Journal of Economic Perspectives* 33 (1): 165-184.

Organization for Economic Co-operation and Development (2005, January 11). OECD Glossary of Statistical Terms. Retrieved from: https://stats.oecd.org/glossary/detail.asp?ID=5755

United States Census Bureau (2019, February 5) *U.S. Census Bureau Streamlines Reporting for Retailers* [Press Release] Retrieved from: https://www.census.gov/newsroom/press-releases/2019/retailers.html.

United States Census Bureau. U.S. Census Bureau Strategic Plan- Fiscal Year 2018 Through Fiscal Year 2022 [PDF File]. Washington, DC: Author. Retrieved from: https://www2.census.gov/about/budget/strategicplan18-22.pdf