Volume Title: Big Data for Twenty-First-Century Economic Statistics

Volume Authors/Editors: Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-80125-4 (cloth), 978-0-226-80139-1 (electronic)

Volume URL: https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics

Conference Date: March 15-16, 2019

Publication Date: Februrary 2022

Chapter Title: Improving Retail Trade Data Products Using Alternative Data Sources

Chapter Author(s): Rebecca J. Hutchinson

Chapter URL: https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics/improving-retail-trade-data-products-using-alternative-data-sources

Chapter pages in book: p. 99 – 114

3

Improving Retail Trade Data Products Using Alternative Data Sources

Rebecca J. Hutchinson

3.1 Introduction

The US Census Bureau has long produced high-quality official statistics for the retail trade sector.[1] These data are obtained through traditional survey data collection and are a critical input to the calculation of the Gross Domestic Product (GDP), of which retail trade made up nearly 25 percent of the 2019 estimate (Bureau of Economic Analysis 2020). The retail data are also critical to Census Bureau data users because they analyze the current state of a retail sector 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 improvement, 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

Rebecca J. Hutchinson is the Big Data Lead in the Economic Indicator Division of the US Census Bureau.

The author would like to thank Catherine Buffington, William Davie, Nicole Davis, Lucia Foster, Xijian Liu, Javier Miranda, Nick Orsini, Scott Scheleur, Stephanie Studds, and Deanna Weidenhamer, as well as the CRIW committee of Katharine Abraham, Ron Jarmin, Brian Moyer, and Matthew Shapiro for their thoughtful comments on previous versions of this paper. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied (Approval ID: CBDRB-FY19-EID-B00001). Disclaimer: Any views expressed are those of the author and not necessarily those of the United States Census Bureau. For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see https://www.nber.org/books-and-chapters/big-data-21st-century-economic-statistics/improving-retail-trade-data-products-using-alternative-data-sources.

1. The production of quality statistics is the principal goal of the US 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.

products, to supplement traditional survey data collection, to ease respondent burden, and to assist with declining response rates (US 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. Here I expand that initial work by examining the viability of using point-of-sale data as a replacement for retail survey data more broadly. During the pilot project and in preliminary analysis, I used data for five retailers. During this expanded effort, I review a larger purchase of this third-party retailer data for quality issues and explore additional uses. I document the use of point-of-sale data for a small number of retailers in the production of the Monthly Retail Trade Survey (MRTS) estimates (US Census Bureau 2019).

The rest of the paper proceeds as follows. Section 3.2 provides background on the Census Bureau's retail survey programs as well as modernization efforts currently underway. Section 3.3 discusses point-of-sale data broadly and provides details on the point-of-sale data from NPD used in this project. Section 3.4 discusses the results from a review of the point-of-sale data, including visual and regression analysis conducted at national and store levels. Section 3.5 provides an overview on the product category mapping exercise done between NPD and the Census Bureau's Economic Census product categories. Section 3.6 provides a discussion of the challenges and costs of using these data in official government statistics. Section 3.7 lays out the next steps for this project.

3.2 Retail Data Collection and Modernization Efforts

Retail trade is currently measured by the Census Bureau through monthly and annual surveys, as well as through a quinquennial Economic Census, and covers retail companies as defined by the North American Industry Classification System (NAICS). Retail businesses (NAICS Sector 44-45) may be chain retailers with many store locations, retailers with only one store location, or retailers operating solely online as e-commerce businesses. The retail businesses represented may or may not have paid employees. Table 3.1 provides a summary of the Census Bureau's retail trade programs. In years ending in "2" and "7," the Economic Census collects detailed sales and product-level information as well as employment, payroll, and business characteristics for each physical store location that a retailer operates. Data collected by the Economic Census is used to update the sampling frame for the annual and monthly retail trade surveys. Each year, the Annual Retail Trade Survey (ARTS) collects annual sales, e-commerce sales, inventories, and expenses data as well as some retailer characteristics at the retailer

Table 3.1 **Overview of the Census Bureau's retail trade programs**

	Economic Census	Annual Retail Trade Survey (ARTS)	Monthly Retail Trade Survey (MRTS)	Advance Monthly Retail Trade Survey (MARTS)
Frequency	Conducted every five years in years ending in '2' and '7'	Conducted annually	Conducted monthly	Conducted monthly
Response	Required by law	Required by law	Voluntary	Voluntary
Sample source	N/A	Sampled from frame created by the Economic Census	Subsampled from the Annual Retail Trade Survey	Subsampled from the Monthly Retail Trade Survey
Data collection level	Individual store location	Total retailer by NAICS	Total retailer by NAICS	Total retailer by NAICS
Data items captured	• Business characteristics • Employment and payroll • Detailed product-level sales	• Business characteristics • E-commerce sales • Sales • Inventories • Expenses	• Limited business characteristics • Sales • Inventories • E-commerce sales	• Limited business characteristics • Sales • E-commerce sales

level nationally by NAICS. The MRTS—a subsample of the ARTS—is a voluntary monthly survey done at the retailer level and collects sales and inventories. The timeliest measurement of the retail economy is the Advance Monthly Retail Trade Survey (MARTS), a subsample of the MRTS. This survey's estimates are published approximately two weeks after month's end and measure only sales.

In recent years, the Census Bureau has placed a growing emphasis on the use of nontraditional means to collect and obtain data (Jarmin 2019). These nontraditional means have the potential to help the Census Bureau continue producing high-quality data while also addressing data user demands for higher-frequency and more granular data. They can also address both declines in survey response and increases in the cost of traditional survey operations. Alternative data sources are one such nontraditional avenue of exploration. Data sources of interest to the retail programs include high-frequency and near real-time data that can be used to measure retail sales, including point-of-sale retailer data. Additionally, system-to-system data collection and web scraping are two alternative data collection methods that could be utilized to collect and obtain data (Dumbacher and Hanna 2017). These alternative data sources and collection methods could be used in conjunction with existing survey and administrative data to create new data products while improving the efficiency and quality of the survey lifecycle.

The improvements and benefits that may be achieved through these alternative data sources and collection methods are coupled with concerns. These concerns include the transparency in the methodology as well as issues related to the quality, consistency, and confidentiality of the data. The Census Bureau strives to be transparent in its methodologies and it is unclear how adopting third-party data use will impact that transparency. A study done by the National Academy of Sciences recommends that federal statistical agencies explore the benefits of using third-party data sources but remain mindful of both the unknowns in determining the quality of these data sources and the challenges when combining data sources (Groves and Harris-Kojetin 2017).

3.3 Point-of-Sale Data

Point-of-sale data, also known as scanner data, are detailed sales data for consumer goods that are obtained by scanning the barcodes or other readable codes on the products at electronic points-of-sale both in physical store locations and online (Organisation 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.

Much work has been done on the use of point-of-sale data in producing price indices. Feenstra and Shapiro (2003) highlight the benefits of point-of-sale data including the comprehensiveness of the data and capturing all products over a continuous period. Point-of-sale data also capture new product offerings faster than traditional price collection methods. The United States Bureau of Labor Statistics has researched using point-of-sale data to supplement the Consumer Price Index (CPI) calculations (Horrigan 2013).

This paper explores the use of point-of-sale data with a focus on the sales value rather than the prices. The working hypothesis 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 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). Though potentially cheaper in terms of data costs, obtaining point-of-sale data directly from a retailer can require extensive IT and staffing resources to store, clean, and process. The Census Bureau is interested in obtaining data feeds directly from retailers in the future but point-of-sale data from a third party are the more reasonable option from a resource perspective at this time.

Point-of-sale data for this project were provided by NPD.[2] 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.[3] NPD currently has data that

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

3. 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.

are potentially useful for this project for over 500 retailers. In comparison, the 2017 Economic Census identified over 600,000 retail firms. Thus, the NPD dataset is not scalable to the entire retail sector.

NPD receives, processes, edits, and analyzes weekly or monthly data feeds containing aggregated transactions by product for each individual store location of its retailers.[4] These data feeds include a product identifier, the number of units sold, product sales in dollars, and the week ending date.[5] Sales tax and shipping fees 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 nonrespondents because this voluntary survey is the timeliest measure of retail sales and response is critical to survey quality.[6] High-burden retailers are also considered a priority.[7]

4. Some retailers do not provide individual store location feeds to NPD and just provide one national feed.

5. NPD does not receive information about individual transactions or purchasers.

6. Response rates to the Monthly Retail Trade Survey have fallen from 74.6 percent in 2013 to 66.5 percent in 2017.

7. 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.

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 United States Code Title 13, such that it is kept confidential and used only for statistical purposes.[8] Retailer participation in this effort is voluntary and some retailers do 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.[9] Subsequent monthly deliveries of retailer data are made 10 to 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.[10]

3.4 Data Quality Review

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. National-level NPD sales for each retailer are compared against what the retailer reports to the MRTS and the ARTS. NPD store-level retail sales for each retailer are compared against the retailer's reported store-level sales in the Economic Census. NPD product-level sales for each retailer are compared to the retailer's reported product-level sales in the Economic Census. 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

8. To uphold both 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.

9. 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.

10. In this early part of the review, the imputation rate of the NPD data is also checked. For the vast majority of months, the imputation rate is zero for retailers. However, NPD will impute a small amount of data if the retailer could not provide all values in its data feed for a given month. The average imputation rate for the data provided by NPD across all retailers and all months is 0.15 percent.

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.[11] Issues with both the NPD and 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. As this project grows, a process including automated algorithms must be developed so these types of issues can be identified in a timely manner and then resolved efficiently by both NPD and Census Bureau staff.

With the project expansion, the need for more definitive quality metrics has grown more urgent. The long-term goal is to develop quality review profiles for each individual retailer that can dictate the decision to use the NPD data and allow a retailer to stop reporting sales to Census Bureau retail surveys. This profile might include metrics that show variation in levels and month-to-month changes between the NPD point-of-sale data and Census Bureau survey data. Included in this profile will be an algorithm that identifies cases for analyst review based on the size of the anomalies detected.

In developing these metrics, a method to identify discrepancies between NPD and Census data is needed. Here we summarize differences through the use of regression models that show how much of the variation in the MRTS data is explained by the NPD data. The models are run for aggregated national and store levels, individual retailers, and groupings of retailers.

Sections 3.4.1 and 3.4.2 detail the results of this initial data quality review for data at the national level and at the store level. The review includes breakouts of the brick-and-mortar sales, e-commerce sales, and the sum of these two sales figures, also known as whole store sales. New retailers agree to share their NPD data with the Census Bureau on a regular basis. To create a consistent base for the analysis, I report results for 10 retailers. These retailers represent a mix of different types of retail businesses. Most have an e-commerce component to their MRTS data. Six of the retailers are consistent reporters to the MRTS. The remaining four are sporadic reporters or nonreporters to the survey. Starting dates for NPD data vary between 2012 and 2015. The analysis end point for each retailer's time series is October 2018.[12]

3.4.1 National-Level Data

Visual inspections of time series plots of the NPD and MRTS data are a good way to identify issues early and develop intuition about the type of issues that might arise. Figure 3.1 displays whole store sales aggregated for

11. 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).

12. Because most retailers operate on a fiscal calendar that runs from February to January, any annualized NPD figures referenced below are for that fiscal year.

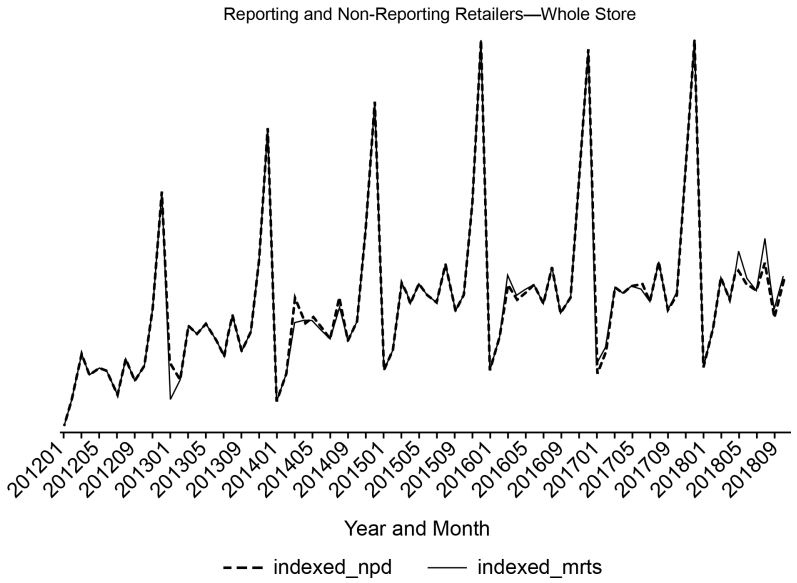

Fig. 3.1 Indexed whole store sales comparisons between NPD data and MRTS data for 10 retailers

Source: NPD and MRTS data.

Note: January 2012 = 1.000.

all 10 retailers. Overall, the data align well between the NPD and MRTS data. The most notable deviation is in March 2014, where the NPD sales are higher than the MRTS sales; this data point has been investigated but a cause has not been identified. 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.

Another important use of the NPD data is to validate Census Bureau tabulated data. Figure 3.2 displays the comparisons for two groups: consistent reporters to the MRTS and the sporadic or nonreporters whose data are imputed by the Census Bureau. The consistent reporters are responsible for the tight alignment observed in figure 3.1. MRTS nonreporters drive the deviations between the NPD data and the imputed MRTS data over the time series.

Imputation methodology for the MRTS reflects a retailer's past information as well as industry behavior from reporting companies each month. Thus, survey imputation will not be successful in capturing idiosyncratic retailer activity outside of industry trends and seasonal patterns.

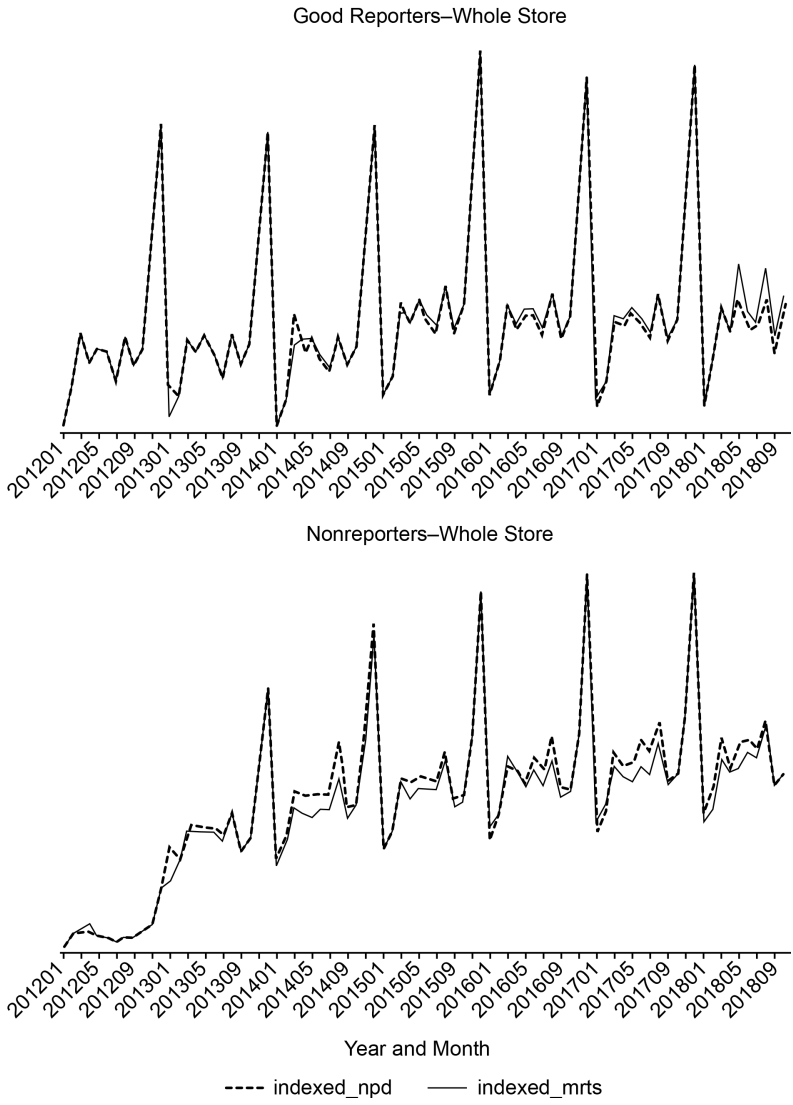

Fig. 3.2 Indexed whole store sales comparisons between NPD data and MRTS data for six consistent reporters (top) and four nonreporters (bottom) to the MRTS for January 2012–October 2018

Source: NPD and MRTS data.

Note: January 2012 = 1.000. MRTS data for nonreporters are imputed values.

Point-of-sale data will capture firm-specific movements so differences between the NPD data and imputed MRTS data are expected.

As part of this project, work has begun on establishing more sophisticated quality metrics for the NPD data. The first attempt at this utilizes an ordinary least squares regression with the natural log of the NPD monthly sales data as an independent variable and the natural log of the MRTS sales data as the dependent variable. A coefficient of one suggests that a change in the NPD data results in an equal change in the MRTS data. The R^2 value from this regression indicates how well the change in a retailer's NPD data can explain the change in variation in the retailer's MRTS data. A higher R^2 value could be one statistical diagnostic to determine whether the NPD data are good enough to use in place of MRTS data.

Figure 3.2, however, indicates that because the NPD data for nonreporters may not align as well with the imputed MRTS data, the use of R^2 to evaluate the quality of the NPD data may be less useful for retailers who do not report to the MRTS. If future data feeds include a large enough number of retailers such that there are other retailers with similar characteristics (e.g., kind of retail business, size) to the nonreporting retailers, more sophisticated models that include local, industry, and time-specific shocks could be used to evaluate the use of NPD data for nonreporters. At this time, the Census Bureau is not receiving enough retailer data to fully explore this idea and determine what other diagnostic values should be established but some initial work has been done. Table 3.2 presents results from regressions performed using data from the 10 retailers. This specification explains over 99.3 percent of the variation in the MRTS data. The model for e-commerce sales for MRTS nonreporters has the lowest R^2 of 24.0 percent. One possible explanation for this is the current imputation methodology for e-commerce sales. The e-commerce component of retailers with a separate online division is captured in a NAICS code (NAICS 4541, Nonstore Retailers) that is different from their brick-and-mortar sales NAICS code. The current imputation methodology estimates e-commerce sales for nonrespondents within this nonstore retailer grouping with no differentiation among the primary types of business conducted. That is, e-commerce sales for sporting goods stores, department stores, clothing stores, etc. within the nonstore retailer component are imputed using the same imputation ratio. Research is planned to determine if this imputation should consider the primary kind of retail business.

3.4.2 Store-Level Data

The store-level data in the NPD dataset has the potential to reduce respondent burden in the Economic Census where the reporting unit is the establishment or store location. The inclusion of a retailer-provided store number in the NPD datasets allows for a direct match to the Economic Census database, which also includes the same retailer-provided store

Table 3.2 **Ordinary least squares regression results for regression of MRTS sales data on NPD sales data**

Dependent variable: Natural log of monthly retail trade sales									
	(Whole store)			(E-commerce)			(Brick and mortar)		
	All retailers (A)	Consistent reporters (B)	Non-reporters (C)	All retailers (D)	Consistent reporters (E)	Non-reporters (F)	All retailers (G)	Consistent reporters (H)	Non-reporters (I)
Natural log NPD monthly sales	1.008***	1.008***	1.005***	0.984***	1.047***	0.335***	0.999***	1.002***	0.992***
	(0.003)	(0.002)	(0.008)	(0.015)	(0.009)	(0.049)	(0.003)	(0.002)	(0.010)
Constant	–0.180	–0.166	–0.149	0.306	–0.828	11.844	–0.016	–0.036	0.119
Observations	748	456	292	491	342	149	748	456	292
R^2	0.993	0.998	0.982	0.898	0.978	0.240	0.992	0.998	0.974

Source: NPD and MRTS data. Standard errors in parentheses.

Notes: An F-test for the null hypothesis that the coefficient for the natural log of NPD monthly sales is equal to 1 is not rejected for columns C, D, G, H, and I. Coefficients are statistically different from 1 otherwise. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 3.3 **Ordinary least squares regression results for regression of 2012 Economic Census store sales on NPD annualized 2012 store sales**

	Dependent variable: Natural log of 2012 economic sales by store location
Natural log annualized 2012 NPD sales by store location	0.871***
	(0.007)
Constant	2.075
	(0.126)
Observations	2,601
R^2	0.984

Source: NPD and 2012 Economic Census data. Standard errors in parentheses.
Notes: Firm effects are included for each retailer but not displayed. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

number variable in each store location record. As a result, store-level sales data comparisons are possible.

Of the 10 retailers considered in this paper, seven reported store-level information to the 2012 Economic Census and had 2012 NPD data available.[13] The store-location match rate between the two data sets was over 98 percent. Potential causes for mismatches include store number differences and store openings and closures that are captured by one source but not the other. The ratio of the natural log of 2012 NPD sales to the natural log of 2012 Economic Census sales for each matched store location were plotted.[14] In this plot, there is a large cluster of values around the 45-degree angle line, indicating that the 2012 NPD data for a store location is close to the sales data that the retailer reported to the 2012 Economic Census for that particular store location. There are also some outliers. Store-level data can be more burdensome for retailers to report and retailers may report estimates rather than actual figures. Store openings and closures may also affect the precision of the data. Thus, store-level data can be noisier than the national-level data where small differences across store sales may cancel out.

Store-level regression analysis is done for retailers using an ordinary least squares regression similar to the national-level regressions in section 3.4.1 but with the natural log of the NPD annualized sales for each store as an independent variable and the natural log of 2012 Economic Census store sales as the dependent variable. At the individual store locations for retailers that reported to the 2012 Economic Census and had NPD data available for 2012, this specification explains over 98 percent of the variation in the store sales figures tabulated in the 2012 Economic Census (table 3.3).

13. A complete analysis of the data in the 2017 Economic Census is underway. For the purposes of this paper, the focus is on the 2012 Economic Census store-level data.

14. This graphic could not be displayed due to disclosure concerns.

3.5 Product Data

The Economic Census collects detailed product-line sales information from all large retailers and a sample of smaller retailers. Product-level reports are made available to the public approximately three years after the end of the Economic Census year. Alternative product-level data sources could help with not only reducing respondent burden but also creating more timely and higher-frequency product reports.

Point-of-sale data from NPD is collected at the stock-keeping unit level (SKU), which allows retailers to track product inventories. NPD assigns detailed product attributes to each of these SKUs and assigns them to product categories including but not limited to apparel, small appliances, and toys. These categories are defined differently than the Census Bureau's product-level categories. For this reason, the NPD product-line research focuses on whether a mapping between the NPD product lines and the Census Bureau product lines is feasible. The 2017 Economic Census was the first Economic Census to use the North American Product Classification System (NAPCS), a demand-based, hierarchical product classification system. With assistance from Census Bureau classification staff and NPD product-line experts, a NAPCS code has been assigned to each item in the NPD product catalog.

With this mapping successfully completed, sales in the NPD dataset can be tabulated by NAPCS code. Work is underway to compare the NPD product-level data and the 2017 Economic Census data by NAPCS code.

3.6 Challenges

While the findings of this project have been promising, there are several challenges. There are substantial upfront costs associated with a third-party data source like NPD. These costs cover the overhead expenses of working with retailers to obtain consent to share NPD data with the Census Bureau and curating the retailer datasets. This process becomes more streamlined over time and costs may diminish. Any arrangement that would reduce Census Bureau costs while still benefiting the Census Bureau, NPD, and the retailers would likely require a change in government policy regarding third-party vendors' ability to collect fees from retailers and provide the data to official statistical agencies (Jarmin 2019).

Another challenge is that only sales data are currently available through the NPD data feeds. The retail surveys collect other items including inventories and expenses. NPD is exploring the feasibility of collecting other data items through its data feeds. Other third-party data sources that capture business operations data may be able to provide additional data items.

There are several risks associated with the use of third-party data. Concerns with transparency and coverage were highlighted earlier in the paper. Other risks include a vendor going out of business or being acquired by

another entity, a decline in the vendor's share of the market, or an increase in the price of the data. Additionally, a third-party vendor could create its own data product comparable to an existing Census Bureau data product, reverse engineer Census Bureau estimates for financial benefit, or recover confidential information about other nonparticipating retailers. Mitigating these risks requires careful selection of a diversified pool of data sources.

3.7 Next Steps

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. Currently, a conservative approach is being taken to use the data in survey estimates based on a case-by-case review of the differences between the NPD and MRTS data by retail subject matter experts. 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 (US Census Bureau 2019). 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. We continue to analyze the data at the store and product levels, comparing against the newly collected 2017 Economic Census data. The NPD data provide an opportunity not only to help with respondent burden and survey nonresponse but also to help produce more timely and more granular estimates. Of particular interest are the product-level data. The Census Bureau currently only publishes product-level data every five years, making use of data collected in the Economic Census. The NPD data have monthly product-level information that could be utilized to create timelier product-level data products. Additionally, the monthly NPD datasets include store-level information that can identify store openings and closures more quickly than current Census Bureau survey operations. Developing a pipeline to use these data to create a more up-to-date picture of retail economic turnover would be valuable both at the national level and at more granular geographies. Exploratory work on these concepts is currently underway.

References

Boettcher, Ingolf. 2014. "One Size Fits All? The Need to Cope with Different Levels of Scanner Data Quality for CPI Computation." Paper presented at the UNECE Expert Group Meeting on CPI, Geneva, Switzerland, May 26–28, 2014. 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. 2020. Gross Domestic Product, Third Quarter 2020 (Second Estimate), table 3, November 25, 2020. https://www.bea.gov/sites/default/files/2020-11/gdp3q20_2nd_0.xlsx.

Department of Commerce. 2014. "Federal Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data." July 2014. Washington, DC. https://www.commerce.gov/sites/default/files/migrated/reports/revisedfosteringinnovationcreatingjobsdrivingbetterdecisions-thevalueofgovernmentdata.pdf.

Dumbacher, Brian Arthur, and Demetria Hanna. 2017. "Using Passive Data Collection, System-to-System Data Collection, and Machine Learning to Improve Economic Surveys." Paper presented at the 2017 Joint Statistical Meetings, Baltimore, MD. http://ww2.amstat.org/meetings/jsm/2017/onlineprogram/AbstractDetails.cfm?abstractid=322018.

Feenstra, R. C., and M. D. Shapiro. 2003. Introduction to *Scanner Data and Price Indexes*, edited by Robert C. Feenstra and Matthew D. Shapiro, 1–14. Chicago: University of Chicago Press.

Groves, Robert M., and Brian A. Harris-Kojetin, eds. 2017. *Innovation in Federal Statistics*. Washington, DC: National Academies Press.

Horrigan, Michael. 2013. *Big Data and Official Statistics*. Washington, DC: Author. https://www.bls.gov/osmr/symp2013_horrigan.pdf.

Hutchinson, Rebecca J. 2020. "Investigating Alternative Data Sources to Reduce Respondent Burden in United States Census Bureau Retail Economic Data Products." In *Big Data Meets Survey Science: A Collection of Innovative Methods*, edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg, 359–85. Hoboken, NJ: John Wiley and Sons.

Jarmin, Ron S. 2019. "Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics." *Journal of Economic Perspectives* 33 (1): 165–84.

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

US Census Bureau. 2018. *U.S. Census Bureau Strategic Plan—Fiscal Year 2018 through Fiscal Year 2022*. Washington, DC: US Census Bureau. https://www2.census.gov/about/budget/strategicplan18-22.pdf.

———. 2019. "U.S. Census Bureau Streamlines Reporting for Retailers." Press release, February 5. https://www.census.gov/newsroom/press-releases/2019/retailers.html.