Investigating Alternative Data Sources to Reduce Respondent Burden in United States Census Bureau Retail Economic Data Products

Rebecca J. Hutchinson [1][2][3]
Economic Directorate, United States Census Bureau, rebecca.j.hutchinson@census.gov

1 INTRODUCTION

Retail store closures, innovative industry disruptors, and the evolution of online shopping dominate business news feeds on a daily basis. In this dynamic retail environment, official statistics that measure retail sales are closely watched economic indicators. At the same time, response rates are declining for many United States Census Bureau surveys, including the retail surveys. Respondents often cite the burden of completing multiple surveys as one reason for not responding (Haraldsen et al. 2013). The Census Bureau has been exploring the use of alternative source data, including data acquired from third parties, to reduce respondent burden while ensuring production of high quality statistics (United States Census Bureau 2018). For example, if retailers are providing their data to a third party, could those data consistently be used in place of a Census Bureau survey collection?

This paper details an effort undertaken by the Census Bureau testing if retailer point-of-sale data could be used in place of the data reported by retailers to a survey. Section 1 provides background on Census Bureau economic programs, specifically retail, as well as modernization efforts currently underway. Section 2 discusses point-of-sale data. Section 3 provides a project plan for using retailer point-of-sale data from The NPD Group, Inc. Section 4 has a discussion on the initial quality review of the point-of-sale data including visual and regression analysis conducted at national and store levels. Section 5 provides an overview on the product category mapping exercise done between the NPD and the

[1] Disclaimer: Any views expressed are those of the author and not necessarily those of the United States Census Bureau.
[2] 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)
[3] The author would like to thank Catherine Buffington, William Davie, Lucia Foster, Xijian Liu, Javier Miranda, Nick Orsini, Scott Scheleur, and Stephanie Studds 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.

Economic Census product categories. Section 6 provides a discussion of the challenges and costs of using this type of data in official government statistics. Section 7 lays out the next steps for this project.

1.1 Overview of the Census Bureau Retail Programs

The Census Bureau is responsible for the measurement of the economic activity of United States businesses and government organizations. As part of its mission, the Census Bureau collects quality economic data and provides statistics that are critical to understanding current conditions in the economy. This data is important to the preparation of key measures of economic activity by other agencies, including Gross Domestic Product (GDP) estimates and producer price indexes.[4]

Every five years, the Census Bureau conducts an Economic Census and a Census of Governments. Together these Censuses provide comprehensive coverage of all U.S. non-agricultural businesses. On a monthly, quarterly, or annual basis, the Census Bureau conducts 70 separate economic surveys. These collections include twelve principal economic indicators that provide the most timely official measurement of the United States economy including housing starts, corporate profits, and retail sales.

The Census Bureau's retail trade program covers retail companies as defined by the North American Industry Classification System (NAICS) and represents all retail companies (NAICS Sector 44-45) with and without paid employees. These retail businesses may be large retailers with many store locations, single-unit retailers with only one location, or retailers operating solely as an e-commerce business. Table 1.1 provides a summary of the Census Bureau 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 Census Bureau's Business Register, the sampling frame for many surveys, including the annual and monthly retail trade surveys. Each year, the Annual Retail Trade Survey

[4] 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.

(ARTS) collects data at the company or retailer level nationally. The ARTS collects annual sales, e-commerce sales, inventories, and expenses data as well as some retailer characteristics.

The Monthly Retail Trade Survey (MRTS)—a subsample of the ARTS—is a voluntary survey done at the retailer level and collects sales as well as inventories. The timeliest measurement of the retail economy and earliest indication of nominal consumer spending produced by the government is the Advanced Monthly Retail Trade Survey (MARTS), a subsample of the MRTS. This survey measures only sales and estimates are published approximately two weeks after month's end.

	Economic Census	**Annual Retail Trade Survey (ARTS)**	**Monthly Retail Trade Survey (MRTS)**	**Advanced 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	Establishment or store level	Company level	Company level	Company level
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

Table 1.1: Overview of the Census Bureau's Retail Trade Programs

1.2 Alternative data source vision

Official economic statistics produced by the Census Bureau have long served as high-quality benchmarks for data users. However, demands for more timely and more granular data, a decline in respondent cooperation, and increasing costs of traditional survey data collection are making it challenging for the Census Bureau to meet its data users' needs. To meet these needs, a growing emphasis has been placed on exploring nontraditional means of collecting and obtaining data (Jarmin 2019).

The Census Bureau has initiated a number of exploratory projects using alternative data sources while remaining mindful of the possibilities and the challenges that accompany the use of these data. Alternative data sources of interest include high-frequency and near real-time data such as point-of-sale

retailer data to measure retail sales or satellite imagery to detect new construction. The Census Bureau envisions leveraging these data sources in conjunction with existing survey and administrative data to provide more timely data products, to provide more granularity with detailed geographic and industry-level estimates, and to improve efficiency and quality throughout the survey life cycle. Alternative data collection methods such as system-to-system data collection and web scraping could also play a large role in reducing respondent burden (Dumbacher and Hanna 2017).

Incorporating these types of alternative data sources into official government statistics has promise but also raises concerns related to methodological transparency, consistency of the data, information technology security, public-private partnerships, confidentiality, and data quality. 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 the unknowns in determining the quality of these data sources as well as the challenges when combining data sources (Groves & Harris-Kojetin 2017).

2 POINT-OF-SALE DATA

2.1 Background

Reducing respondent burden has been an initial focus of the alternative source data work. Point-of-sale data, also known as scanner data, may help reduce burden for retail surveys. Point-of-sale data is detailed data on sales of consumer goods obtained by scanning the bar codes or other readable codes of products at electronic points-of-sale both in brick and mortar retail stores and online.

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. This data is available at the retailer, store, and product levels. By contrast, credit card data or payment processor data is often only available at an aggregated level; due to confidentiality agreements, vendors of these data often cannot reveal which retailers are included in the aggregates. Additionally, point-of-sale data is more complete, capturing all purchases in a store whereas credit card data only captures purchases made with a credit card and excludes cash transactions.

A large body of work has been done on the use of point-of-sale data in producing price indices. Feenstra and Shapiro (2003) highlighted the benefits of point-of-sale data including its comprehensiveness, capturing all products over a continuous period. Point-of-sale data also captures 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 calculations and cited the potential of using alternative data sources to validate data collected through traditional operations (Horrigan 2013).

This paper also explores the use of point-of-sale data but focuses on the sales value rather than on 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 a year should reflect total retail sales for a retailer. 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. To test this hypothesis, at a minimum 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. One of the primary challenges facing this type of effort is finding a solution that is scalable to the scope of a survey while operating within budget limitations (Jarmin 2019). While survey operations—specifically non-response follow-up—are themselves expensive to conduct, the cost of third-party data may limit the amount of data that can be purchased.

Another concern with using third-party data is that the statistical agency no longer controls the raw data and cannot ensure its quality. In traditional survey data collection, the Census Bureau controls the full

data collection and processing life cycle. Statistical agencies strive to be transparent in their methodologies and it is unclear how adopting third-party data will impact that transparency.

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 understand the data. 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 feasible option from a resource perspective at this time.

2.2 Background on NPD

Through the official government acquisitions process, third-party data sources were researched and the NPD Group, Inc. (NPD) was selected as the third-party data source vendor for this project. NPD is a private market research company that captures point-of-sale data from over 1,300 retailers representing 300,000 stores and e-commerce platforms worldwide. In comparison, the Census Bureau's 2016 County Business Patterns identified 1,069,096 retail establishments (or stores). Thus, the NPD dataset is not scalable to the entire sector as the NPD data represent less than one-third of the retail universe.

From each store location, NPD processes weekly or monthly data feeds containing aggregated transactions by product. Each data feed includes a product identifier, the number of units sold, product sales in dollars, total store sales in dollars, and the week ending date. Sales tax and shipping and handling are excluded. Any price reductions or redeemed coupon values are adjusted for before NPD receives the feeds. The sales figures in the feed reflect the final amount that the customer paid which should align to the total net revenue for the company. NPD does not receive data on individual transactions or purchasers.

NPD edits, analyzes, and summarizes the point-of-sale data feeds at detailed product levels and creates market analysis reports for its retail and manufacturing partners. [5] NPD processes data for many product categories including apparel, small appliances, automotive, beauty, fashion accessories, consumer

[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.

electronics, footwear, office supplies, toys, and jewelry and watches. While NPD receives a feed of total store point-of-sale activity that includes all purchased items, NPD only classifies data for those products in the product categories listed above. Any sales on items that do not belong in these categories are placed in an unclassified bucket. For example, NPD currently does not provide market research on grocery items; all grocery sales data is tabulated as unclassified. Therefore, a whole store picture is not available at the product level unless detailed information from the unclassified bucket can be provided.

Retailer datasets from NPD contain monthly data by store and product level (i.e., for a given month, sales for Product Z in Store Location Y for Retailer X). As part of the acquisition process, the Census Bureau provided dataset requirements to NPD and NPD curated the datasets from their data feeds. The datasets are limited to stores located in the continental United States and 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 for each store location includes a total sales value of the unclassified data.

3 PROJECT DESCRIPTION

Determining the viability of point-of-sale data as a replacement for retail survey data is the initial focus of the research phase. A small amount of data are reviewed for quality concerns and other potential uses for the data are explored. During this phase, the following questions need to be answered:

- **National-Level Data**: How well do national-level sales data tabulated from the point-of-sale data compare to data that retailers reported to the monthly and annual retail surveys? How is the quality of the point-of-sale data determined for those retailers who do not report to the survey?
- **Store-Level Data**: How well do store-level sales and location data tabulated from the point-of-sale data compare to data that retailers reported to the 2012 Economic Census?
- **Product-Level Data**: How well do the product categories in the point-of-sale data align to the North American Product Classification System (NAPCS) used 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 another research goal. This review process is detailed in Section 4.

3.1 Selection of retailers

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 from NPD would be protected by United States Code Title 13 and be kept confidential and used only for statistical purposes.[6] Retailer participation in this effort is voluntary. Retailers are mostly enthusiastic about participating but some retailers decline to participate. Declining retailers cited a variety of reasons including legal and privacy concerns, while others stated that completing Census Bureau surveys is not a difficult task.

From a list of NPD retailers that agree to provide data feeds to the Census Bureau, we select retailers whose data would be useful for our analysis. 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 one of the most timely measures of retail sales and response is critical to survey quality.[7] High-burden retailers are also considered a priority.[8]

3.2 Data Ingest

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

[6] 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.
[7] Response rates to the Monthly Retail Trade Survey have fallen from 74.6% in 2013 to 66.5% in 2017.
[8] 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.

Bureau, and NPD all sign the agreement of participation.[9] 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. 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]

4 DATA QUALITY REVIEW

The quality review process has evolved as the project has evolved. Access to more retailer data has allowed for the formation of better answers to the project objectives set forth in Section 3, specifically determining how well the NPD data align with data collected or imputed by the Census Bureau's retail trade programs. To date, the decision to use or not to use a retailer's data has relied heavily on retail subject matter expertise. This expertise will always be crucial to the decision-making process.

A quality review of a retailer's data begins with a 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 data as well as the MRTS data have been identified during this visual review. This highlights that the data may require close monitoring. To date, the sources of these types of issues were each unique and their resolutions required research. As this project grows in size, automation will be the goal but it must allow these types of issues to be identified immediately and then resolved efficiently by both NPD and Census Bureau staff.

As the project has grown, the need for more definitive quality metrics in determining if a retailer's data is of good enough quality to be used in official government statistics has also grown. The long-term goal is to develop quality review profiles for each individual retailer that would dictate the decision to use the

[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%.

[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).

NPD data and allow a retailer to stop reporting sales to Census Bureau retail surveys. This profile would include metrics including but not limited to variation in levels and month-to-month changes between the NPD 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. One approach is to summarize differences through regression models that show how much of the variation in the MRTS data can be explained by the NPD data. The models are run at the national and store levels for individual retailers and groupings of retailers.

Sections 4.1 and 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 which 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, data for ten retailers is presented here. These retailers represent a mix of different types of retail businesses. Most have an e-commerce component to their MRTS. Six of the retailers are consistent reporters to the MRTS. The remaining four are sporadic reporters or non-reporters to the survey. Starting dates for NPD data vary: six retailer time series begin in 2012; the remaining four begin in 2013, 2014, or 2015. The analysis end point for each of the retailer time series is October 2018. [12]

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 to develop intuition about the type of issues that might arise. Figure 4.1 displays the whole store sales aggregated for all ten retailers. Overall, the data align well between the NPD data and the MRTS data. The most notable deviation is in March 2014 where the NPD sales are higher than the MRTS; this data point has been investigated but a cause has not been identified. Given the volume of data ingested, some

[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.

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 be always obvious from the exploration of time series properties.

Figure 4.1 Indexed whole store sales comparisons between NPD data and MRTS data for ten retailers. (January 2012=1.000)
Source: NPD and MRTS data

An important use of the NPD data is to validate Census Bureau tabulated data. Figure 4.2 displays the comparisons for two groups: consistent reporters to the MRTS and the sporadic or non-reporters whose data is imputed by the Census Bureau. The consistent reporters are responsible for the tight alignment observed in Figure 4.1. MRTS non-reporters drive the deviations between the NPD data and the imputed MRTS data over the time series.

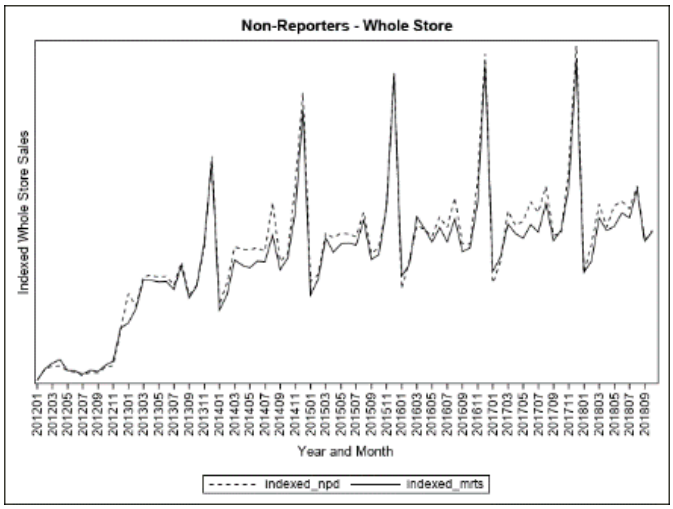

Figure 4.2 Indexed whole store sales comparisons between NPD data and MRTS data for six consistent reporters (top) and four non-reporters (bottom) to the MRTS for January 2012-October 2018. (January 2012=1.000). MRTS data for non-reporters are imputed values.
Source: NPD and MRTS data

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 often not be successful in capturing idiosyncratic retailer activity outside of industry trends and seasonal patterns. Point-of-sale data will capture firm-specific movements so differences between the NPD 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-squared 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-valued R-square could be one statistical diagnostic to determine whether the NPD data is good enough to use in place of MRTS data.

Figure 4.2, however, indicates that because the NPD data for non-reporters may not align as well with the imputed MRTS, the use of R-squared to evaluate the quality of the NPD data may be less useful in those situations. If future data feeds include a large enough number and types of retailers, such that there are

other retailers with similar characteristics (e.g., kind of retail business, size) to the non-reporter retailers, more sophisticated models and their resulting statistics could be used to evaluate the use of NPD data for non-reporters.

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 4.1 presents results from regressions performed using data from the ten retailers. This specification explains over 99.3% of the variation in the MRTS data. The model for e-commerce sales for MRTS non-reporters has the lowest R-squared of 24.0%. One possible explanation for this is the current imputation methodology for e-commerce sales. The e-commerce component of companies with a separate online division are captured in a separate NAICS code (NAICS 4541) from their brick and mortar sales. The current imputation methodology estimates e-commerce sales for non-respondents within this non-store retailer grouping with no differentiation between the primary types of business conducted. That is, e-commerce sales for sporting goods stores, department stores, clothing stores, etc. within the non-store retailer component are imputed using the same imputation ratio. Research is underway to determine if this imputation should take into account the primary kind of retail business.

Dependent Variable: Natural Log of Monthly Retail Trade Sales									
	(Whole Store)			(E-Commerce)			(Brick and Mortar)		
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)
	All Retailers	Consistent Reporters	Non-Reporters	All Retailers	Consistent Reporters	Non-reporters	All Retailers	Consistent Reporters	Non-reporters
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-squared	0.993	0.998	0.982	0.898	0.978	0.240	0.992	0.998	0.974

Standard errors in parentheses.
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 4.1 Ordinary least squares regression results for regression of MRTS sales data on NPD sales data.
Source: NPD and MRTS data

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 store number in the NPD datasets allows for a clean match to the Economic Census databases, which also include the same store number variable in each store location record, allowing for comparing store-level sales data. These store numbers are assigned and provided by retailers.

Of the ten 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%. Potential causes for mismatches include store number differences and store opening and closures that are captured by one and not the other. The ratio of the natural log 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 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 difference 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 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

[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.

Census and had NPD data available for 2012, this specification explains over 98% of the variation in the store sales figures tabulated in the 2012 Economic Census (Table 4.2).

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	2601
R-squared	0.984

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4.2 Ordinary least squares regression results for regression of 2012 Economic Census Store Sales on NPD Annualized 2012 store sales. Firm effects are included for each retailer but not displayed.
Source: NPD and 2012 Economic Census data

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 is a level that allows retailers to track product inventories. NPD assigns detailed product attributes to each of these SKUs and places them into 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, which makes sense as the two organizations are serving two different—though likely overlapping—data user groups. For this reason, the NPD product-line research focuses on whether a mapping between the NPD and Census Bureau product lines is feasible. With the 2017 Economic Census, the North American Product Classification System (NAPCS), a demand-based, hierarchical product classification system, was fully implemented. 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. However, this is the first Economic Census where retailers were required to report data using NAPCS classifications so this may introduce uncertainty to any comparison results.

6 CHALLENGES

While the initial findings of this project have been promising, there are a number of challenges. There are substantial upfront costs for 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).

In addition to cost being a challenge, only sales data is currently available through the NPD data feeds. The retail surveys collect a number of other items including inventories and expenses. NPD is currently exploring the feasibility of collecting other data items through data feeds. Non-NPD 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. 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 non-participating retailers. Mitigating these risks requires careful selection of a diversified pool of data sources.

7 NEXT STEPS

What began with data from just a few retailers to test the hypothesis that point-of-sale data could be used to help Census Bureau retail programs has the potential to evolve into something much bigger. This project has demonstrated potential for the use of point-of-sale data not only to reduce respondent burden

but also to supplement existing surveys or to create new data products. Currently, a conservative approach is being taken to use the data in survey estimates based 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 (United States 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 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 non-response 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 using Economic Census data. The NPD data has monthly product-level information that could be utilized to create more timely product-level data products. Additionally, the monthly NPD datasets include store-level information which can identify store openings and closures more quickly than current Census Bureau survey operations. Developing a pipeline to use this 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 all of these concepts is currently underway.

8 REFERENCES

American Association for Public Opinion Research. (2015). AAPOR Report on Big Data, AAPOR Big Data Task Force.

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

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

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

Feenstra, R.C. and Shapiro, M.D. (2003). "Introduction to Scanner Data and Price Indexes." In Scanner Data and Price Indexes, Chicago, IL: University of Chicago Press, pp. 1-14.

Groves, Robert M. and Harris-Kojetin, Brian A (eds). (2017). *Innovation in Federal Statistics*. Washington, D.C.: The National Academies Press.

Haraldsen, Gustav, Jones, Jacqui, Giesen, Deirdre, et al. (2013). Understanding and Coping with Response Burden. In Ger Snijkers, Gustav Haraldsen, Jacqui Jones, and Diane K. Willamck, *Designing and Conducting Business Surveys*, 219-252. Hoboken, New Jersey: John Wiley & Sons, Inc.

Horrigan, Michael (2013). *Big Data and Official Statistics* [PDF File]. Washington, DC: Author. Retrieved from https://www.bls.gov/osmr/symp2013_horrigan.pdf

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.

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://www.census.gov/content/dam/Census/about/about-the-bureau/PlansAndBudget/strategicplan18-22.pdf

United States Office of Management and Budget (2017). Analytical Perspectives [PDF File]. Washington, DC: Author. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2018/02/ap_15_statistics-fy2019.pdf