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

Rebecca J. Hutchinson [1][2]

Economic Directorate, United States Census Bureau, rebecca.j.hutchinson@census.gov

[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)

1 INTRODUCTION

Retail store closures, mergers and acquisitions among major retailers, innovative industry disruptors, and the evolution of online shopping dominate business news feeds on a daily basis. Official statistics that accurately and consistently measure retail sales have long been closely watched economic indicators but in this dynamic retail environment, they are even more thoroughly monitored. At the same time, response rates are declining for many Census Bureau surveys, including the retail surveys. Respondents often cite the burden of completing multiple surveys on a monthly and/or annual basis as one reason for not responding (Haraldsen et al. 2013). Recognizing these challenges and the growing needs of its data users, the Census Bureau made a commitment to explore the use of alternative data sources to produce high-quality data products while reducing respondent burden (United States Census Bureau 2018).

One avenue that the Census Bureau has been exploring to reduce respondent burden is the use of third-party data. If a retailer is already providing another party with similar data to what the Census Bureau collects on surveys and censuses, can that data consistently be used in place of what a retailer would be asked to provide to Census Bureau surveys?

This paper details an effort undertaken by the Economic Directorate of the United States Census Bureau using retailer point-of-sale data to test if this alternative source of data could be used in place of the data reported by retailers to a survey. The paper provides background on the Economic Directorate and both the challenges facing the directorate as it tries to modernize its economic measurement as well as the opportunities for use of alternative data specifically in the retail survey programs. One of the challenges motivating the focus on retail is declining response rates due to increasing respondent burden. The paper provides an overview of respondent burden in business surveys and how it can impact the retail survey programs in particular.

The paper then focuses on one specific project using retailer point-of-sale data from The NPD Group, Inc. To validate the quality and use of this point-of-sale data in the retail survey, visual and regression analysis are conducted on the data at the national and store levels. To determine the data's quality and usability, the point-of-sale data are compared at the store and national-level against the Monthly and Annual Retail Trade Surveys and to the 2012 and 2017 Economic Censuses. The results show that the point-of-sale data align well with survey and Census data. Additionally, product data from the Economic Census were matched to the product-level data in the point-of-sale data sets. The paper also highlights the lessons learned and issues identified when using third-party data in official government statistics and closes with a discussion of the costs and benefits of using this data type of data in production.

Based on the success of this work, the project has now entered a production phase where data for up to 100 retailers will be purchased and evaluated over the next two years. Retailer data from NPD is now being included in the estimates for the Monthly Retail Trade Survey. Much of the initial focus of this effort has been placed on the use of the NPD data in our Monthly Retail Trade Survey where the NPD data can have an immediate and positive impact on response rates and data quality. In 2019, use of the NPD data will expand to include the Annual Retail Trade Survey and the Economic Census.

1.1 Overview of the Economic Directorate

The Economic Directorate of the Census Bureau is responsible for statistical programs that measure the economic activity of United States businesses and government organizations. The Economic Directorate's mission is to collect quality economic data and provide statistics that are critical to understanding current conditions in the economy. These data are important to the preparation of key measures of economic activity by other agencies, including Gross Domestic Product (GDP) estimates, benchmark input-output accounts, producer price indexes, and measures of industrial production and manufacturing capacity utilization.

Every five years, the Economic Directorate conducts an Economic Census and a Census of Governments. Together these Censuses cover all U.S. non-agricultural businesses: manufacturing, wholesale trade, retail trade, mining, utilities, construction, transportation, finance and insurance, real estate, healthcare, and other services sectors, as well as local, state, and federal governments.

On a monthly, quarterly, or annual basis, the Directorate conducts 70 separate surveys. These collections include twelve principal economic indicators that provide the most timely official measurement of the United States economy including housing starts; retail sales; wholesale sales; services revenue; manufacturers' shipments, inventories, and orders; new construction put in place; and corporate profits.

Additionally, the Economic Directorate is responsible for merchandise export and import statistics produced monthly, extensive compilations of administrative records, and numerous special research and technical studies.

1.2 Alternative data source vision

Official economic statistics produced by the Economic Directorate have long served as high-quality benchmarks for data users. However, demands for more timely and more granular data, a decline in respondent cooperation, increasing costs of traditional survey data collection, and a changing economic landscape are making it challenging for the Economic Directorate 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 Economic Directorate has initiated a number of exploratory projects using alternative data sources while remaining mindful of both the great possibilities and the great challenges that accompany the use of these sources. For measurement of the economy, these alternative data sources could include high-frequency and near real-time data such as point-of-sale data obtained from a retailer or a third party, building permit data obtained from an Application Programming Interface (API), or commodity flow data captured by sensors. The Economic Directorate envisions leveraging these alternative sources in conjunction with existing survey and administrative data to provide more timely data products, to offer greater insight into the nation's economy through 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 burden on respondents and Census Bureau analysts (Dumbacher and Hanna 2017). Rather than conducting costly follow-up operations with respondents, analysts could instead use the data gathered through alternative data collection methods.

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 the general quality of the data. Statisticians who set policy and quality standards for official government statistics are now faced with various issues surrounding third-party data. The United States Office of Management and Budget (OMB) and associations such as the American Association for Public Opinion Research (AAPOR) and the American Statistical Association (ASAS) have begun looking more closely at how to evaluate the quality of third-party data and statistics derived from them (American Association for Public Opinion Research 2015).

1.3 Overview of the Census Bureau retail programs

The retail trade program currently 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.

The Economic Directorate measures the retail economy every five years in the Economic Census and on a more frequent basis in monthly and annual surveys. In years ending in "2" and "7", the Economic Census--a mandatory survey--asks for detailed sales and product-level information as well as employment and payroll and business characteristics for each physical store location that a retailer operates. Data collected by the Economic Census are used to update the Census Bureau's Business Register, from which the sampling frames for many Economic Directorate surveys, including the annual and monthly retail trade surveys, are created. Each year, the Annual Retail Trade Survey (ARTS) collects data at the company or retailer level nationally; no store location data are collected. The ARTS collects annual sales, e-commerce sales, beginning and end-of-year inventories, and expenses data as well as some retailer characteristics; the annual data are released approximately 15 months after the data year ends.

Within the Economic Indicators Division of the Economic Directorate, two retail surveys are conducted. The Monthly Retail Trade Survey (MRTS) is a voluntary survey done at the retailer or company level and collects sales/receipts as well as end-of-month inventories and e-commerce sales from all retail industries. Estimates from this survey are released approximately six weeks after month's end. The MRTS is a subsample of the ARTS and a re-selection of the MRTS sample occurs approximately every five years to ensure the sample remains representative and to redistribute the burden for small- and mid-size businesses.

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). This survey measures only sales/receipts and estimates are published approximately two weeks after month's end. The MARTS is a subsample of the MRTS and this sample is selected every 2-3 years, again to ensure a representative sample and to redistribute burden.

Table 1.1 provides a summary of the retail trade programs at the Census Bureau.

	Economic Census	**Annual Retail Trade Survey (ARTS)**	**Monthly Retail Trade Survey (MRTS)**	**Advanced Monthly Retail Trade Survey (MARTS)**
Frequency	Conducted every five years (for 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 • Reporting period information • Sales • Inventories • E-commerce sales	• Limited business characteristics • Reporting period information • Sales • E-commerce sales

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

2 RESPONDENT BURDEN

The Office of Management and Budget (OMB) has authority over information collected by federal agencies, including surveys conducted by the Census Bureau. The Paperwork Reduction Act of 1995 grants OMB this authority with the goal of minimizing the burden on the public while simultaneously maximizing the public benefit of information collected and minimizing the costs of data and information collection to the federal government (United States Office of Personnel Management 2011). Census Bureau surveys must be authorized by OMB and are typically reauthorized every three years. As part of this reauthorization, each survey must estimate the number of burden hours that the survey will place on an individual respondent and on the potential set of respondents as a whole; this burden estimate appears on survey instruments or in a letter to the respondent. The Census Bureau obtains these burden hour estimates through cognitive and other pretesting activities conducted with survey respondents.

Respondent burden for business surveys is measured from two different perspectives: actual and perceived (Willeboordse 1997). There are actual, measurable units of burden that can be quantified by a dollar amount or by number of hours. Examples of these types of measurements include number of people involved in the survey completion, time spent to complete the survey, and salary cost of completing the survey (Haraldsen et al. 2013). This burden measurement can grow complex for larger businesses where multiple people are involved in the completion of one or more survey instruments. In addition to Census Bureau surveys, businesses may also be receiving surveys from other government agencies. The decentralized nature of the United States federal statistical programs makes it difficult to coordinate data collection and data sharing among over one hundred included agencies (United States Office of Management and Budget 2018). Each agency tends to have its own set of legal and organizational barriers in place that would prevent inter-agency data sharing activities that may reduce respondent burden.

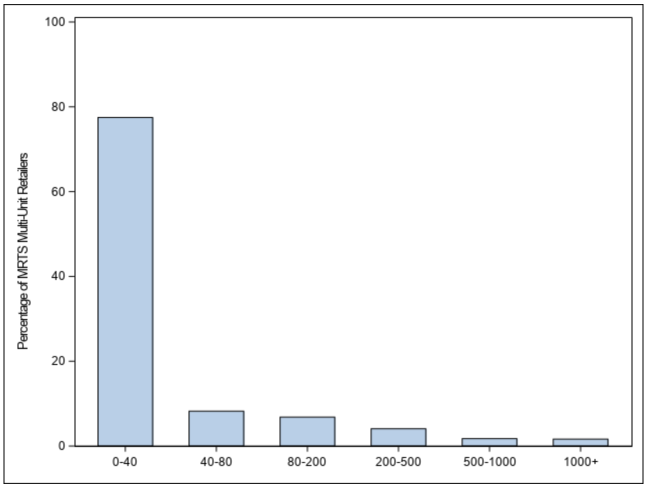

Figure 2.1: Estimated Total Burden Hours to Complete Retail Trade Surveys in 2017. This burden analysis was conducted on multi-unit establishments who receive the Monthly Retail Trade Survey. **Source:** Monthly Retail Trade Survey, 2017 Annual Retail Trade Survey, and 2017 Economic Census Initial Letter to Multi-Units

Figure 2.1 displays estimates of the total burden hours for multi-unit establishments for completing Census Bureau retail trade programs survey instruments in 2017. The survey instruments used in this measurement include the 2017 Monthly Retail Trade Survey, the 2017 Annual Retail Trade Survey, and the 2017 Economic Census. The Monthly Retail Trade Survey takes seven minutes to complete but retailers have to complete it 12 times each year and may also receive more than one survey instrument each month if its operations include more than one kind of retail business as defined by NAICS. The Annual Retail Trade Survey can take up to 3 hours and 19 minutes to complete and is completed only one time per year. The Economic Census is the most burdensome where a separate survey instrument needs to be completed for each store location. If a retailer has 100 store locations, then 100 surveys need to be completed. The estimated time burden for a single Economic Census survey ranges between 41 minutes and 5 hours and 36 minutes. This burden analysis discussed here was conducted on multi-unit establishments who receive the Monthly Retail Trade Survey. Multi-units are those retailers with two or more establishments or store locations and often have the bookkeeping practices or software and staff in place to streamline government reporting; therefore, for the purposes of this analysis, the minimum reporting time for the Economic Census was used. As seen in Figure 2.1, over 70 percent of these multi-unit retailers would have spent 40 hours or less completing all of their Census Bureau retail survey instruments. On the other hand, over seven percent of these businesses would have spent over 200 hours completing their survey instruments.

While the analysis above focused on the retail-specific burden, it is important to remember that these same retailers may be part of larger, more diversified companies may be included in surveys in other industry trade areas including wholesale, manufacturing, and services. These retailers may also be receiving business surveys conducted by other government agencies including the Bureau of Labor Statistics and the Bureau of Economic Analysis. In many cases, individual companies providing data about their performance is essential to the production of these data products. As respondents face increasing requests to provide data, respondent burden increases and the risk for lower quality estimates due to low response increases (Haraldsen et al. 2013).

Perceived burden is harder to measure in well-defined units. Perceived burden captures the willingness of the respondents to participate in the survey (Willeboordse 1997). Perceived burden can include the usefulness of providing the data both to the respondent and to society (Haraldsen et al. 2013). Interviews conducted by Census Bureau researchers with 30 large businesses—both in retail and in other industries—found that work that benefits the business will take priority over work related to reporting to government surveys, especially those surveys that are voluntary and do not carry penalty for non-response (Willimack and Nichols 2010). Some businesses will even refuse to participate in any voluntary surveys, including the voluntary Monthly Retail Trade Survey (the Annual Retail Trade Survey and the Economic Census are mandatory and required by law).

Ease of access to data and the ability to retrieve the requested data from across a company and across information systems are also perceived factors that play into a business's decision to complete a survey instrument (Willimack and Nichols 2010). This is especially relevant for instruments requesting more than just top-line national sales estimates. For example, the Economic Census asks retailers for employment and payroll figures, class of customer, physical characteristics including square footage of an establishment, and sales by product types for each physical store location that a retailer is operating. Product-line sales information was cited by large businesses as being particularly difficult to provide (Willimack and Nichols 2010).

3 POINT-OF-SALE DATA

3.1 Background on point-of-sale data

There are different approaches to reducing respondent burden. If the data needs to be collected via surveys, then improvements to the questionnaire can be made or the collecting agency or party can better communicate the importance of the data to the respondent (Haraldsen et al. 2013). However, if the respondent is already providing another party with data similar to what the Census Bureau collects on surveys and censuses, there is the potential that the survey may no longer need to be completed.

Point-of-sale data, also known as scanner data, is one possible third-party source that may help reduce burden. Point-of-sale data are 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 & mortar retail stores and online. Point-of-sale data can provide information about quantities, product characteristics, prices, and the total value of goods sold and has the advantage that the data are available at the retailer, store, and product levels. Other third-party data sources, including credit card data or data from payment processors, are often only available at an aggregated level. Due to confidentiality agreements, vendors of this data often cannot reveal which retailers are included in the aggregates. Additionally, point-of-sale data are more complete, capturing all purchases in a store whereas credit card data would only capture purchases made with a credit card.

A large body of work has been done exploring the use of scanner data in producing price indices. Eurostat (2017) identified a number of advantages of using scanner data when working with price indices including that scanner data provides information on expenditures for all items sold over a continuous period. Scanner data also can capture new product offerings faster than traditional price collection methods. The detailed product attribute data available in scanner data is also useful when working with price indices (Bird et al. 2014). The United States Bureau of Labor Statistics has researched using scanner 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). There are also creative non-price index uses of the data. IBM has used grocery store data to locate the source of food-borne illnesses (IBM 2016). The National Cattleman's Beef Associations uses scanner data to understand better the changing meat preferences of its customers (Krebs 2016).

The Economic Directorate project also explored the use of point-of-sale data in a manner where the focus is not on prices. Rather the focus is on the theory that if all items that a retailer sells are captured in a point-of-sale data feed, then summing those sales up across products and store locations over a month or a year should reflect total retail sales for a retailer. If the theory holds, the sales figure from the point-of-sale data should be comparable to what is provided by a retailer to the Monthly Retail Trade Survey, the Annual Retail Trade Survey, and the Economic Census. To successfully test this theory, a point-of-sale dataset needed to meet the following requirements:

- Identify the data by retailer name
- Provide product-level sales for each of the retailer's store locations
- Have data available by month at a minimum.

Obtaining retailer scanner feeds can be done either directly through a retailer or through a third-party vendor. While the raw scanner data from either source should be identical, there are advantages and disadvantages to both (Boettcher 2014). First, the third-party vendor will clean the data and aggregate and curate the data in a consistent format to meet its data users' needs. This service does come 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, third-party data may also be expensive to purchase or the price of the data may limit the amount of data that can be purchased. The United States Congress appropriates the budget of the Economic Directorate of the Census Bureau and there are limited funds available for alternative data source purchases in a given fiscal year.

One other concern with using third-party data is that control of the raw data is relinquished. In traditional data collection through survey instruments, the Economic Directorate controls the full data collection and processing life cycle. Statistical agencies must be transparent in their methodologies so if third-party data is used, the third party must also be transparent with its methodologies.

Though potentially cheaper in terms of data buy costs, obtaining data directly from a retailer can require extensive physical IT resources as well as staffing resources to clean and understand the data. Monthly datasets with sales by store location and product type are large datasets that can cause server slowness when processed. While the Economic Directorate would be interested in obtaining data feeds directly from retailers in the future, staffing and IT limitations currently limit this effort from being implemented on a large scale. For the future of this project, cloud storage is being pursued. Additionally, the Census Bureau is currently developing an enterprise data lake. Data lakes are scalable storage repositories that can bring together datasets of all sizes where data can be stored in a way that allows for easy access and sharing of the data (Miloslavskaya et al. 2016). At the time of this project, point-of-sale data from a third party are the more feasible option.

3.2 Background on NPD

Through the official government acquisitions process, third-party data sources that would potentially be useful to this effort 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. To put that store count in perspective, the Census Bureau's 2016 County Business Patterns identified 1,069,096 establishments (or stores) in the Retail NAICS 44-45. Thus, one limitation of the NPD dataset is that it is not scalable to the entire economy as the NPD data represent less than one-third of the retail universe.

From each store location, NPD receives and processes weekly or monthly data feeds containing aggregated scanner transactions by product. 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.

At a minimum, each data feed includes a product identifier, the number of units sold, product sales in dollars, the average price sold, total store sales in dollars, and the week ending date. Sales tax and shipping and handling are excluded. Any price reductions or redeemed coupons values are adjusted for before NPD receives the feeds; thus, the sales figures in the feed reflect the final amount that the customer paid; this should align to the net revenue total 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. NPD also curates datasets for customers like the Census Bureau. It is important to note that NPD's business model is not focused on delivering data for individual retailers rather it is based on aggregated product reports; thus, NPD has had to modify existing processes and create new ones in order to deliver the curated retailer datasets to the Census Bureau.

NPD processes data for many product categories including apparel, small appliances, automotive, beauty, fashion accessories, consumer electronics, footwear, office supplies, toys, video games, 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 groceries sales data are tabulated as unclassified. This is another limitation of the data: that there is not a whole store picture available at the product level unless detailed information from the unclassified bucket can be provided as well.

Retailer datasets purchased 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 were limited to stores located in the continental United States and included values for the following variables:

- Time period (month/year)
- Retailer name
- Store number
- ZIP code of store location
- Channel type (brick & mortar or e-commerce)
- Imputation flag (indicates if sales figure is tabulated directly from a feed or derived using imputation methodology)
- Product classifications by industry, product category, and product subcategory
- Sales figures.

One observation for each month and year for each store location includes a total sales value of the unclassified data. At this time, NPD only collects brick & mortar sales and e-commerce sales. If a retailer has business-to-business or catalog operations, that data can also be broken out. Other data items that are collected by the Census Bureau surveys including inventories and expenses are currently not collected by NPD. This means while the respondent burden can be reduced on reporting sales, retailers may still have to provide other data items via traditional surveys unless these other data items can also be captured in the NPD data feeds or by other data sources.

Additionally, NPD provided national-level datasets by month for each of the retailers. These national-level estimates could also be obtained by summing the sales data in the store-level datasets by month.

4 PROJECT DESCRIPTION

With a point-of-sale data provider selected, the project's initial focus is to make a preliminary determination of the viability of point-of-sale data as a replacement for retail survey data. The project plan favors an incremental approach where data from a small number of retailers are purchased over a two-year research period (Hutchinson 2017). The research phase allows for an exploratory review of a small amount of data for quality concerns and for exploring potential uses for the data. The goal of this research phase is determine if the point-of-sale data could be used in a production environment where point-of-sale data were included in published estimates for the monthly and annual surveys as well as the Economic Census without sacrificing the quality of the estimates. Additionally, there are currently no official or standardized quality measures in place within the Economic Directorate to deem a retail third-party data source's quality acceptable; thus, this research phase is also useful for developing a quality review process for third-party data sources that continues to evolve as the project grows. This review process is detailed in Section 5.

Recall that the driving question of this project was "If a retailer is already providing another party with similar data to what the Census Bureau collects on surveys and censuses, can that data be used in place of what a retailer would be asked to provide on a Census Bureau form?" For each type of data, this project seeks to answer that question by also answering these questions:

- **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? If the data aligned well for retailers who reported and survey data was available for comparison, how is the quality of the point-of-sale data for those retailers who do not report to the survey determined?
- **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 and 2017 Economic Censuses?
- **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? If the mapping is possible, how well do the product sales compare between the NPD data and Economic Census product data?

4.1 Selection of retailers

At the beginning of the project, NPD provided a list of retailers that provide data feeds to NPD. From the list, the retailers are selected according to the following criteria. The selection of retailers that are good reporters (i.e., retailers that consistently report to the survey) to the Monthly Retail Trade Survey, the Annual Retail Trade Survey, and the 2012 and/or 2017 Economic Census allows for an initial baseline quality comparison to the NPD data. Priority is also given to selecting Monthly Retail Trade Survey nonrespondents as this voluntary survey is one of our most timely measures of retail sales and response is critical to data quality. High burden-retailers identified by the burden calculation in Section 2 are also considered a priority as they have the potential to benefit the project by reducing respondent burden.

Using these selection criteria, the Economic Directorate provides NPD with prioritized lists of retailers to be onboarded to the project. While NPD has access to data from over 1,300 retailers, they need to obtain signed agreements with the retailers to share these data with the Census Bureau. Because NPD has the client contacts for these retailers, NPD leads the onboarding process. The Associate Director of the Economic Directorate 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 Codes Title 13 which requires that data are kept confidential and only used for statistical purposes.

Both to uphold the confidentiality and privacy laws that guide Census Bureau activities and also to facilitate productive work with specific retailers, a small number of NPD staff working on this project completed background investigations and were granted Special Sworn Status. With this Status, 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.

Retailer participation in this effort is voluntary. NPD takes great strides to provide background to the retailers on the benefits of this project including reducing burden but it is ultimately the retailers' decisions to participate or not. For the most part, retailers are enthusiastic about participating, often stating "We've been waiting for something like this to happen!", that "something" being a way to use data that is already being captured and aggregated to eliminate the need to complete a survey instrument. However, some retailers do decline to participate. Those that have declined have cited a variety of reasons ranging from legal and privacy concerns to not understanding the purpose of the project. Others acknowledge that completing Census Bureau surveys was not that difficult for the retailer.

As this project has progressed, NPD has grown efficient in its outreach to retailers about the project.. They can better identify the specific people within a retail company who can make the decision to participate in the effort. To help with the outreach, the Census Bureau provides names of survey or outreach contacts for NPD's use when possible. As part of the outreach process, NPD has also been asking for the number of hours each retailer spent working on Census Bureau retail survey instruments. This information has the potential to understand benefits to retailers from having their respondent burden reduced.

4.2 Data Ingest

Once a retailer is onboarded to the project, NPD has to deliver 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.[3] After this initial delivery, monthly deliveries of all onboarded retailers are made 10-20 days after month's end. Because of security constraints for both NPD and the Census Bureau, the current data delivery vehicle is not the most efficient. The data are delivered through NPD's secure FTP site. NPD's security protocols currently mandate that the each individual data file (one file per retailer per year at the national level and one at the store level) be compressed and zipped in a password-protected file. This process is manageable for a small number of retailers but as this project grows to include a larger number of retailers, a more efficient but equally secure delivery mechanism will need to be implemented.

Unlike many big data type datasets, NPD datasets do not require much cleaning. The file formats, variables, and contents were specified in detail in the terms of the contract. One issue that has caused delays during the data ingest process is inconsistent file names varying from retailer to retailer or even from month to month. Again, this issue was easy to resolve manually in the early phases of the project but as the number of retailer files ingested has increased, it was necessary to have NPD implement and automate a consistent file naming convention.

As part of the data ingest, characteristics of the dataset are verified and check for inconsistencies over time. This process verifies that the product categories, store locations, retailer channels, and other categorical variables have remained consistent over time. This part of the review also verifies that the correct number of months of data have been delivered. In a recent delivery of 2018 data, a retailer flagged

[3] NPD will sometimes acquire industries 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 with Census. In these scenarios, NPD provides data beginning with the earliest year available after 2012.

as only having eight months of data instead of the expected twelve. NPD was alerted immediately and delivered a corrected file.

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%.

5 DATA QUALITY REVIEW

Once the ingest process is complete, the quality review of the data begins. The quality review process has evolved as the project has grown. Access to more retailer data has allowed for the formation of better answers to the project objectives set forth in Section 4, specifically determining how well the NPD data aligns with or can explain the data collected by or imputed for by the Census Bureau's retail trade programs on a consistent basis. To date, the decision to use or not to use a retailer's data has relied heavily on retail subject matter expertise. This expertise is crucial and will always play a role in the decision-making process but the project is currently at a point where it would benefit from the development of consistent quality metrics that determine if individual retailer data from NPD should be used in Census Bureau data products.

The full quality review of a retailer's data always begins with a visual review of the data, plotting the monthly NPD data against the MRTS data. Comparisons are done initially 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). Any issues with the NPD data have been identified during this visual review. Some examples of issues identified during this phase of the review include:

- One year's worth of NPD data for one of the retailers was markedly different from the other three years of NPD data and exhibited a large deviation from the MRTS data that was not present in other years. Working with NPD, the source of the issue was identified as a discrepancy that occurred when the retailer changed the format of its feed and an incorrect data feed overwrote the original, correct data. This issue had gone undiscovered by NPD until this project. The retailer and NPD actively worked together to recover the data but there was no way to restore the historic feed. As a substitute, the retailer was able to provide store-level totals for that year but no product-level information.
- A recent delivery of data for three retailers flagged immediately during the visual review as can be seen in Figure 5.1. This graph contains monthly sales data (presented as indexed values) for three retailers who are good, consistent reporters to the MRTS.[4] While the data in the later part of the time series lined up relatively well, there are large deviations in the early part of the time series where the NPD data are substantially higher than the MRTS data. For one of the retailers, sales for the months of February and March were almost the same as the sales for the holiday shopping month of December, which would be highly unusual for most retailers. Exploring the store and product-level data for these retailers revealed that many product categories and stores had monthly sales 40-50 percent higher than other years. NPD was notified of the issue. Using information the Census Bureau provided them, the NPD technical team found that additional files were being stored in the same place that the usual data feed files were stored and the additional files were being included in the tabulations; as a result, the data in those duplicated files were double counted. NPD is now enacting safeguards to ensure that this will not happen in the future.

[4] Retailer data is always graphed and reviewed on its own during the quality review process. For the purposes of this paper and giving examples of this work while not disclosing information about the individual retailers, the quality review will be demonstrated using aggregations of retailer data.

- Last, the plot of NPD and MRTS data for a retailer aligned well from 2012-2015 but the data deviated from one another in more recent years, with the MRTS data being noticeably higher than the NPD data. The size of the difference was roughly equivalent to the retailer's reported MRTS e-commerce value. MRTS staff investigated the issue and discovered that e-commerce was being double counted in MRTS for that retailer. The retailer's data has since been corrected in both the monthly and annual retail surveys and the NPD and MRTS data now align well.

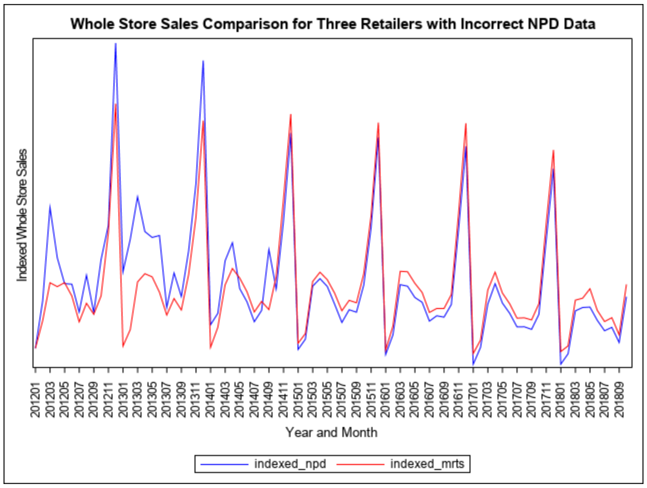

Figure 5.1 Indexed whole store sales comparisons between incorrect NPD data and MRTS data for three good reporters to the MRTS for January 2012-October 2018. (January 2012=1.000)
Source: NPD and MRTS data

These issues highlight the value gained from this project by both the Census Bureau and NPD. Issues that may have otherwise gone undetected are identified by having another source of data against which to validate. However, these issues also highlight that currently the data require close monitoring. The sources of these issues were all unique and their resolutions required research. As this project grows in size, automation will be the goal but it must be implemented in a way that these types of issues can be identified immediately and then resolved efficiently by both NPD and Census Bureau staff. Keeping records of the issues and how they are resolved will be helpful in developing best practices for the future.

In conjunction with the visual review, descriptive statistics like mean and median differences between the NPD data and MRTS data have also been standard elements of the quality review process and have served as a good initial attempt at making a determination of data quality. These descriptive statistics are calculated for both the differences in the levels and the month-to-month changes between the NPD data and MRTS data.

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 first attempt at developing these metrics are regression models that show how much of the variation in the MRTS data can be explained by the NPD data and has expanded to examine how well the store-level NPD data can explain store-level data reported to the Economic Census. These models are run not only on individual retailer data and all of the retailers' combined data but also on data for different groups of retailers including those that are good reporters, those that are not good reporters, and those that operate a similar kind of retail business.

A valuable takeaways from the review process is that all retailer data and comparison results cannot be viewed the same. For data quality review purposes, retailers have been classified as either a "good reporter" or a "non-reporter." If a retailer is a good, consistent reporter to the Monthly Retail Trade Survey, there is a baseline value to compare the NPD data against to make a quality determination. If a retailer does not report to the monthly survey, finding a comparison point is trickier. If the retailer reports to the Annual Retail Trade Survey or the Economic Census, that data can be used. If the retailer does not report to any retail survey program, outside information like public financial information (e.g., SEC filings) can be used. That said, comparisons to the monthly survey for non-reporters are still of value because they can either validate or reject current imputation methodologies. Once the number of retailers onboarded increases, the review process can be expanded to include comparisons among retailers in similar kinds of business (e.g., sporting goods stores, clothing stores, department stores, etc.).

This section of the paper details the results of these data quality review elements (visual, descriptive, and modeling) for data at the national level and at the store level. The data are further broken out to look at the brick & mortar sales, e-commerce sales, and the sum of those two sales figures - referred to as whole store sales. Data is flowing into this project on a regular basis and to create a consistent base for the analysis presented here, data for ten retailers from the project is used. These retailers represent a mix of different types of retail businesses. Most, but not all, have an e-commerce component to their retail operations. Six of the retailers are good and consistent reporters to the Monthly Retail Trade Survey. The remaining four are sporadic reporters or non-reporters to the survey. Starting dates for retailer participation with NPD varies: six retailer time series begin in 2012; the remaining four begin in either 2013, 2014, or 2015. The analysis end point for all of the retailer time series is October 2018. 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. Additionally, any errors in the data identified above were corrected for in these analyses.

5.1 National-Level Data

Time series plots of the NPD and MRTS data have been the best way to identify issues visually at the national level. At a glance, views of the data can easily identify any issues or unusual patterns in the data. Consider the results displayed in Figure 5.2 for 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 notably higher than the MRTS; this data point has been investigated but at this point, a cause has not been identified. Given the volume of data ingested, some data issues—particularly data points at the beginning of the time series—may not be resolved. This is one of the risks of using an alternative data source: Committing to its use means that some months may not be what the survey would have obtained and it will be difficult to know which value is the more accurate one.

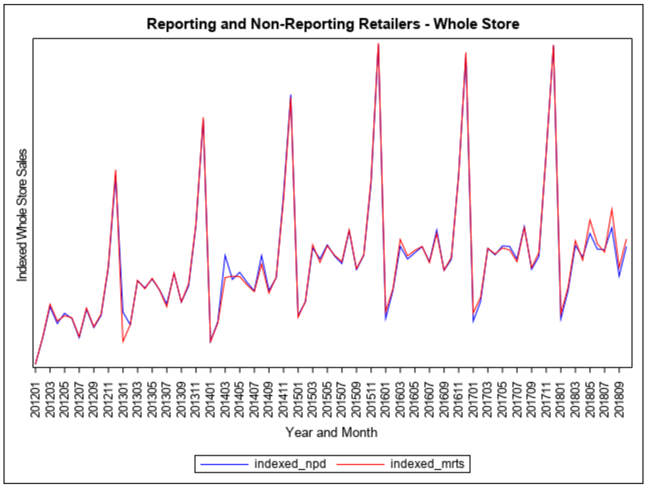

Figure 5.2 Indexed whole store sales comparisons between NPD data and MRTS data for ten retailers. (January 2012=1.000)
Source: NPD and MRTS data

Figure 5.3 displays the graph when the data are separated into groups of those that consistently report to the MRTS and those that do not report or only report occasionally. The plot of the good reporters reflects the similar good alignment observed in Figure 5.2 for all retailers. However, the graph of the non-reporters shows deviations between the NPD data and the imputed MRTS data over the time series.

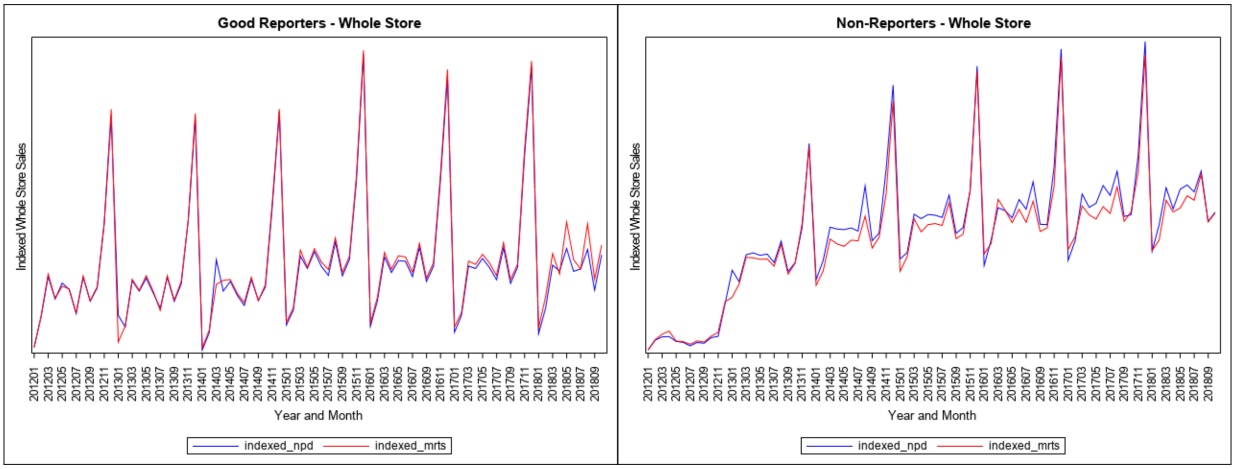

Figure 5.3 Indexed whole store sales comparisons between NPD data and MRTS data for six good reporters (left) and four non-reporters (right) to the MRTS for January 2012-October 2018. (January 2012=1.000). MRTS data for non-reporters are imputed values.
Source: NPD and MRTS data

The time series plots for those retailers that do not report to our Monthly Retail Trade Survey are not particularly useful on their own in determining data quality. Imputation methodology for the MRTS is based on having a consistent and repeatable process that reflects past information about the retailer as well as industry behavior from reporting companies each month. Thus, survey imputation will often not be successful in capturing retailer activity that is unusual including seasonal spikes in sales that are unique to a retailer. Point-of-sale data will capture this unusual movement so differences between the NPD data and the imputed MRTS data are expected. The primary goal of the MRTS is to generate the best estimates of month-to-month change possible. There are little to no controls over the levels and level changes over time for a company that is consistently imputed. The Annual Retail Trade Survey data are used to adjust the levels annually using a benchmark operation at both the industry level and often at a company level. Since this benchmarking is only done annually, imputed data levels can deviate from actual levels over time, so the most important quality metric for any consistently non-reporting retailer is how well the month-to-month changes compare between the NPD data and the MRTS data. This is exactly the case as the levels for the imputed cases are on average 5.71 percent different between the NPD data and the MRTS data and the month-to-month changes differ by just over 0.69 percentage points on average (Table 5.1).

Descriptive Statistics for Level and Month-to-Month Change Comparisons

	(Whole Store)			(E-Commerce)			(Brick & Mortar)		
	All Retailers	Good Reporters	Non-Reporters	All Retailers	Good Reporters	Non-Reporters	All Retailers	Good Reporters	Non-Reporters
Levels									
Average percentage difference between NPD and MRTS retail sales data	2.68	0.73	5.71	2.90	2.91	2.87	2.14	0.27	5.07
Average absolute percentage difference between NPD and MRTS retail sales data	5.39	3.25	8.72	22.48	12.90	44.46	6.11	3.30	10.50
Median percentage difference between NPD and MRTS retail sales data	0.00	-0.02	0.51	-0.28	-0.79	13.82	-0.04	-0.10	0.00
Month-to-Month Changes									
Average percentage point difference between NPD and MRTS retail sales data	0.03	-0.27	0.69	-0.46	-0.45	-0.51	0.10	-0.24	0.84
Average absolute percentage point difference between NPD and MRTS retail sales data	4.48	3.22	7.20	11.67	10.80	16.05	4.40	3.01	7.40
Median percentage point difference between NPD and MRTS retail sales data	-0.02	0.00	-0.10	-0.02	-0.01	-0.21	0.00	0.02	-0.09
Total number of months	748	456	292	491	342	149	748	452	292

Table 5.1 Descriptive statistics for level and month-to-month comparisons between NPD and MRTS data. **Source**: NPD and MRTS data

Brick & mortar sales make up the largest share of the whole store sales. But given the growth and evolution of e-commerce in recent years, the comparisons of the e-commerce sales captured by NPD and the MRTS are worth reviewing. In Figure 5.4, the e-commerce sales data do not align as well as the whole store data. One potential reason for this variation is that there is currently no standard definition for e-commerce across the retail industry. Some retailers will tabulate sales made online and picked up in store as e-commerce sales; others will classify that transaction as a brick and mortar sale. Thus, it is possible that e-commerce sales are being reported differently between the MRTS and the NPD feeds. This is a topic that will continue to be researched as part of both using alternative data sources as well as improving the measurement of e-commerce.

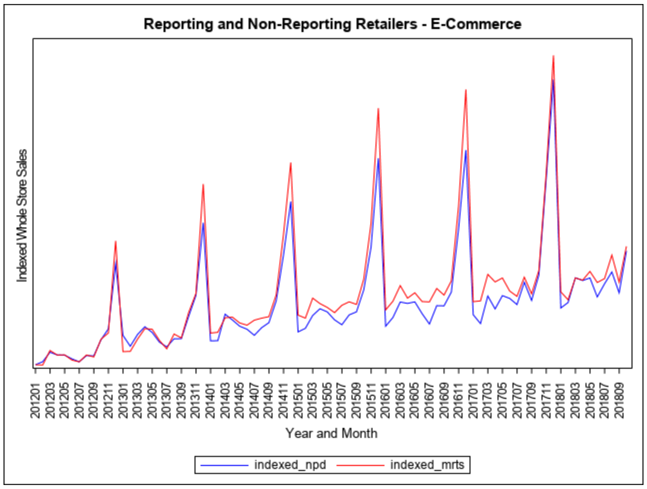

Figure 5.4 Indexed e-commerce sales comparisons between NPD data and MRTS data for ten retailers. (January 2012=1.000)
Source: NPD and Monthly Retail Trade Survey data

Another possible explanation for the lack of alignment between NPD and MRTS e-commerce sales is the current MRTS imputation methodology. Figure 5.5 displays the e-commerce sales comparison for good reporters and non-reporters to the MRTS. The non-reporting retailers demonstrate a large difference between the NPD and MRTS data. One possible explanation for this is the current imputation methodology for e-commerce sales. According to NAICS, 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. That non-store retailer category includes companies from across various retail businesses. The current imputation methodology estimates e-commerce sales for nonrespondents 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. Having an alternative data source like NPD to validate imputation methodology improvements against is valuable.

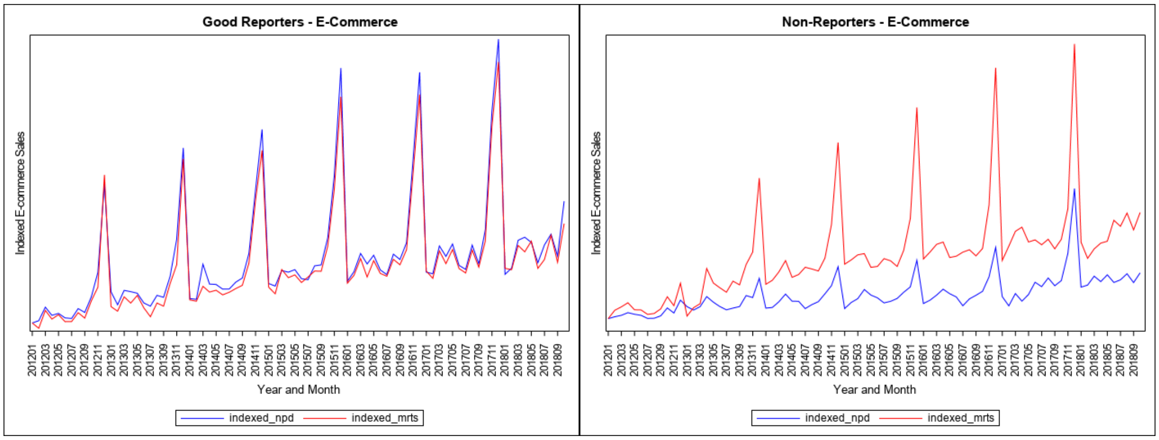

Figure 5.5 Indexed e-commerce sales comparisons between NPD data and MRTS data for six good reporters (left) and four non-reporters (right) to the MRTS for January 2012-October 2018. (January 2012=1.000). MRTS data for non-reporters are imputed values.
Source: NPD and MRTS data

As stated earlier, there are no official metrics in place for determining if the quality of a third-party data is good enough to be used in the retail surveys. As part of this project, work has begun on establishing such 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. The R-squared value from this regression indicates how well the NPD data for an individual retailer can explain the variation in the retailer's MRTS data. Consider again the distinction between the "good reporter" retailers and the "non-reporter" retailers. A higher valued R-square could be a statistical diagnostic that the NPD data are good enough to use in place of MRTS data. However, as shown in the examples above (Figures 5.3 and 5.5), NPD data for non-reporters may not align as well with the imputed MRTS and the R-squared value might be low but that does not mean the NPD data are not of good quality. But if there were other retailers who were good reporters with similar characteristics (e.g., kind of retail business, size) to the non-reporter, individual and pooled regression results for the similar good reporters could be used to make a determination for or against using the NPD data for the non-reporter. This logic is similar to current monthly retail imputation practices where the ratios of the reported data for retailers in similar kinds of retail business and of similar size are used to impute for non-reporters.

At this time, there are not enough retailers onboarded to the project to fully explore this idea and determine what R-squared values and other diagnostic values should be established. However, some initial work can be done. Regressions were performed using all retailers, good reporters, and non-reporters using seasonal and firm fixed effects as well as controls for months when the MRTS data were imputed. When seasonal and firm effects are included, the NPD data explain over 99.7% of the variation in the MRTS data. Table 5.2 shows the results from these regressions for the whole store, brick and mortar, and e-commerce sales for all ten retailers, good reporters, and non-reporters. The results validate the findings from the earlier visual reviews. The model for e-commerce sales for those retailers who do not report to the MRTS has the lowest R-squared with the NPD data explaining 62.6% of the MRS data. In all other models, the NPD data explain over 96% of the MRTS data.

Dependent Variable: Natural Log of Monthly Retail Trade Sales

	(Whole Store)			(E-Commerce)			(Brick & Mortar)		
	All Retailers	Good Reporters	Non-reporters	All Retailers	Good Reporters	Non-reporters	All Retailers	Good Reporters	Non-reporters
Natural Log NPD monthly Sales	0.951***	1.009***	0.854***	0.826***	1.026***	0.514***	0.944***	1.008***	0.831***
	(0.011)	(0.009)	(0.024)	(0.025)	(0.024)	(0.051)	(0.011)	(0.008)	(0.026)
MRTS_impute	-0.006	-0.013	0.014	-0.020	0.027	0.139**	-0.013	-0.035***	0.005
	(0.009)	(0.008)	(0.016)	(0.033)	(0.039)	(0.057)	(0.01)	(0.011)	(0.016)
q1	0.009	0.009	0.016	-0.061*	0.010	-0.137*	0.009	0.008	0.015
	(0.009)	(0.007)	(0.018)	(0.034)	(0.03)	(0.074)	(0.009)	(0.006)	(0.019)
q2	0.007	0.011	0.002	-0.078**	0.036	-0.250***	0.006	0.007	0.001
	(0.008)	(0.007)	(0.018)	(0.034)	(0.03)	(0.072)	(0.009)	(0.006)	(0.019)
q3	0.005	0.009	0.004	-0.059*	0.030	-0.179**	0.005	0.006	0.007
	(0.008)	(0.007)	(0.018)	(0.033)	(0.029)	(0.071)	(0.009)	(0.006)	(0.019)
Constant	1.048	-0.194	2.894	3.351	-0.479	9.092	1.186	-0.172	3.296
	(0.24)	(0.192)	(0.495)	(0.487)	(0.462)	(0.883)	(0.243)	(0.178)	(0.517)
Observations	748	456	292	491	342	149	748	456	292
R-squared	0.997	0.998	0.987	0.960	0.983	0.626	0.995	0.999	0.985

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

Table 5.2 Ordinary least squares regression results for regression of MRTS whole store sales data on NPD whole store data with fixed effects by quarter. Firm effects are included for each retailer but not displayed. **Source**: NPD and MRTS data

5.2 Store-level data

The rich store-level data in the NPD dataset also has the potential to reduce respondent burden in the Economic Census. The fundamental unit of the Economic Census is the establishment or store location. The inclusion of the store number in the NPD datasets allowed for a clean and logical match to the 2012 and 2017 Economic Census databases, which include a store number variable in each store location record. Thus far in the project, there has only been one retailer with store numbers that did not align between the two datasets. In order to conduct a comparison between NPD and Economic Census data for that retailer, a store location ZIP code in the NPD dataset allowed for a successful secondary matching methodology. This issue again highlights that while an automated process will work for most retailers, manual intervention may be required for others.

Of the ten retailers considered in this paper, seven reported store-level information to the 2012 Economic Census and also had 2012 NPD data available. The 2017 Economic Census is still collecting data but at the time of this analysis, six retailers had reported store-level data. The NPD store location match rates for the 2012 Economic Census and 2017 Economic Census are displayed in Table 5.3.[5] Potential causes for mismatches include opening and closures of store locations that were captured by one and not the other or possibly store number discrepancies.

	2012 Economic Census	2017 Economic Census
Total Retailers	7	6
In both NPD and Economic Census data	98.98%	95.01%
In only NPD data	0.64%	3.52%
In only Economic Census data	0.38%	1.47%

Table 5.3 Store location match rates between the NPD data and the 2012/2017 Economic Censuses.
Source: NPD and the 2012/2017 Economic Censuses

Retailers can have a large number of store locations and analyzing this data for outliers and inconsistencies can be difficult. The use of scatterplots has been extremely useful for this exercise. Figure 5.6 displays the ratio of NPD sales to 2012 Economic Census sales for each individual store location. Values on or near the 45 degree line indicate that the NPD data for a store location were close to the sales that the retailer reported to the 2012 Economic Census for that particular store location.[6]

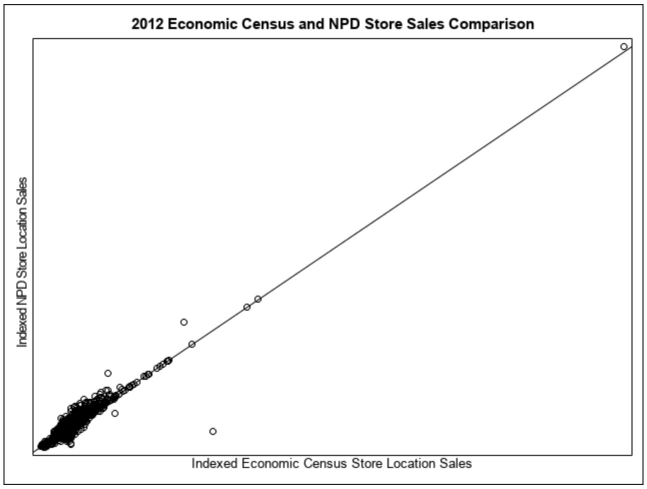

Figure 5.6 Ratio of 2012 NPD store location sales to 2012 Economic Census store location sales. The NPD values for each individual retailer are indexed to the median store location NPD sales value. The 2012 Economic Census values values for each individual retailer are indexed to the median store location 2012 Economic Census sales value.
Source: NPD and 2012 Economic Census data

[5] Data for the 2017 Economic Census is still being collected; these numbers will be updated as more data is made available.

Much like the regression analysis done for retailers at the national level detailed in section 5.1, similar work is done at the store level. These regressions use an ordinary least squares regression 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, the NPD sales explain over 97% of the variation in the store sales figures tabulated in the 2012 Economic Census (Table 5.4).

	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 5.4 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 Monthly Retail Trade Survey data

As of the writing of this paper, data are still being collected for the 2017 Economic Census and currently only six of the ten retailers discussed in this paper have reported. However, the early store sales comparison results are promising. This section will be updated as more data are available. The next step in this effort is to work with Economic Census staff so they can take advantage of this data in their editing and analysis work.

6 PRODUCT DATA

Every five years, the Economic Census collects detailed product line sales, receipts, or revenues from all large retailers and a sample of smaller retailers. Approximately three years after the end of the Economic Census year, product-level reports are made available to the public. The product-level data is valuable to Census Bureau data users but comes at the cost of high response burden for retailers and of a delivery with a nearly three-year lag. Product-level data is one area where alternative data sources--including point-of-sale data--could help with not only reducing respondent burden but also creating more timely product reports since data would be available more frequently than every five years.

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. By collecting data at the SKU level, NPD is able to assign detailed product attributes to each of these SKUs and place them into broader categories including apparel, small appliances, automotive, beauty, fashion accessories, consumer electronics, footwear, office supplies, toys, video games, and jewelry. These categories are defined differently than the product-level categories defined by the Census Bureau, which makes sense as the two organizations are serving two different--though likely overlapping--data user groups. For this reason, the product-line work focuses on whether or not a mapping between the NPD product lines and the Census Bureau product lines is even feasible.

The timing of this project fell during a time of transition for product-line data collection at the Census Bureau. Through the 2012 Economic Census, the Census Bureau collected product-level data at product lines defined for each industry (retail, wholesale, manufacturing, etc.) by the Census Bureau. With the 2017 Economic Census, the North American Product Classification System (NAPCS) was fully implemented. NAPCS moves product classification away from an industry-based system and into a demand-based, hierarchical classification system. The new system is consistent across the three North American countries of Canada, Mexico, and the United States.

Because this project began in 2016, before the start of the 2017 Economic Census data collection, early datasets from NPD contained 2012 data. In order to keep the project moving along, the classification system used in the 2012 Economic Census is used to gain an early understanding of the NPD product classification system with the goal of applying this knowledge to mapping the NPD product catalog to NAPCS. The large number of product categories in the NPD data and in the Economic Census data requires narrowing the scope of the effort to make the mapping exercise feasible while still being able to gain useful information. Apparel is used as the product category to focus on, as it is perceived to be the easiest category to map. However, this mapping exercise is not a simple one.

Notable issues highlighted by this mapping exercise include:

- **Level of comparison.** The 2012 Economic Census product lines are broken out by men’s, women’s, and children’s apparel items. NPD’s breakouts are by apparel type. By far, the biggest takeaway of this effort is that NPD would need to provide additional attributes that would identify the apparel product types as men’s or women’s. This likely extends to other product categories as well.
- **Limited one-to-one mappings**. The lone perfect one-to-one product-line match between NPD and the 2012 Economic Census product lines is dresses. Smaller buckets are created to allow for more specific matching. For example, pants, jeans, shorts, shorts/skirts, and other bottoms product-line categories in the NPD data could be combined and compared to an aggregation of the men’s tailored and dress slacks, men’s casual slacks, jeans, shorts, etc., and women’s slacks/pants, jeans, shorts, skirts from the 2012 Economic Census product lines. These buckets would not be problematic provided that enough product-level detail is available to ensure the correct products are being assigned to the correct buckets.

Product-level data from the 2012 Economic Census is available for those retailers that were onboarded with 2012 NPD product-data. However, the complexities of the apparel product matching above make it impossible to draw any meaningful conclusions from comparing the 2012 Economic Census apparel sales data to the NPD apparel sales data. Instead, the focus of the effort shifted to what additional information is needed from NPD to map its full product catalog to NAPCS. This would allow for meaningful comparisons of the NPD data to 2017 Economic Census product data.

NPD has been able to provide additional product-level data. With this additional information, the strategy for the mapping is to assign a NAPCS code to each item in the NPD product catalog with assistance from classification staff at the Census Bureau as well as product-line experts at NPD. With this additional information from NPD, a NAPCS code has been successfully assigned to every item in the NPD product catalog.

With this mapping successfully completed, sales in the NPD dataset could be tabulated by NAPCS code. As 2017 Economic Census data are collected from retailers onboarded to the NPD project, comparisons between the NPD product level data and the 2017 Economic Census data by NAPCS code will be completed. However, this is also the first Economic Census year that retailers were required to report data to the Economic Census using NAPCS classifications so this may introduce uncertainty to any comparison results.

7 COSTS

The upfront cost for a third-party data source like NPD can be substantial. These costs cover the overhead expenses of onboarding retailers to the project and curating the retailer datasets. After the initial onboarding of retailers and data delivery, the process becomes much more streamlined and costs are expected to diminish over time. At the same time--like other official statistical agencies, the Census Bureau is experiencing increasing survey collection costs and declining response rates. The unit response rates for our monthly retail programs are between 50 and 55 percent, meaning incurred collection costs result in the successful collection of data only about half of the time. 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).

At this time, the NPD project costs do not lead to a comparable savings on the part of the retail programs. However, many benefits of the NPD data cannot be measured in dollars. First, the NPD data will improve data quality. It is a consistent source of data that has the potential to improve the response rate to the closely-monitored Advanced Monthly Retail Trade Survey. Second, the NPD data contains detailed store- and product-level data on a monthly basis that is cost-prohibitive for the retail areas to collect more frequently than every five years in Economic Census. This opens the door for potential new data products, especially as data users have expressed interest in more timely product-level information. If access to a third-party data source frees up time that analysts would have spent conducting nonresponse follow-up operations, analysts may have time to pursue the development of new data products. And finally, the value of the reduction in respondent burden is difficult to quantify. As discussed in Section 2, retailers can spend a substantial amount of time and money completing surveys. Use of third-party data sources has the potential to reduce the burden and cost that retailers incur providing valuable data and information.

8 NEXT STEPS

This project demonstrated great potential for the use of point-of-sale data in our retail programs and the project continues to be a work in progress. NPD is onboarding new retailers on a weekly basis; additional data from the 2017 Economic Census is flowing in for comparisons. What began with just three retailers and a theory that maybe point-of-sale data could be used to help Census Bureau retail programs has the potential to evolve into something much bigger.

In November 2018, the project moved to a production phase for the Monthly Retail Trade Survey. NPD demonstrated the ability to deliver most of the retailer files in time to be incorporated into the Advanced Monthly Retail Trade Survey. Two days before the survey is finalized, retailer data is delivered by NPD, processed, reviewed, and transferred to the Retail Indicator Staff via an analysis portal in under an hour from time of delivery. The analysis portal allows the retail staff to have full time series views—both graphical and tabular--of each individual retailer's NPD sales at the whole store, brick and mortar, and e-commerce breakouts and how these values compare to what is being tabulated in the monthly and annual retail trade survey. Work has begun to implement similar tools for our annual retail staff and retail Economic Census staff to use.

The work done to date has motivated NPD to begin the process of onboarding a data scientist to this project. This data scientist would work as a counterpart to the analysis work being done by the Census Bureau. While NPD has technical staff dedicated to the project, there is no one dedicated to helping with retailer data issues both proactively and reactively. With NPD data now being used in a production environment rather than research, time is critical when using the NPD data for the Monthly Retail Trade Survey and questions regarding the data need to be resolved quickly. When a retailer provides a dollar value on a survey that seems incorrect or questionable, retail analysts often follow-up with the retailer via a phone call or secure communication. When a questionable data value appears in a third-party dataset, the process for communicating questions is unclear. Are questions directed through the third-party vendor? How quickly can answers be obtained for indicator programs when time is limited? With NPD placing a dedicated staff member in this new role, the Census Bureau will work to determine the process to answer to these questions.

The alternative data source vision of the Economic Directorate goes beyond using this type of data to reduce burden; it also seeks to leverage data sources like point-of-sale data in conjunction with existing survey and administrative data to provide more timely data products, to offer greater insight into the nation's economy through detailed geographic and industry-level estimates, and to improve efficiency and quality throughout the survey life cycle. With the NPD data, there is an exciting opportunity to use the data to help fulfill this vision. Of particular interest are the product-level data. The Census Bureau currently only publishes product-level data every five years using data from the Economic Census. The NPD data has monthly product-level information that could be utilized to create more timely product-level data products. Additionally, the monthly datasets include store-level information, which means the NPD data can identify store openings and closures quicker than current Census Bureau survey operations. Developing a pipeline to use the NPD 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.

As exciting as this move to production has been, the project still faces challenges to being fully useful in a production environment. First, only sales data are currently available through the NPD data feeds. The retail surveys collect a number of other items including inventories and expenses. NPD is currently working to check the feasibility of collecting other data items through their feeds and other non-NPD data sources that capture business operations data may also be able to provide additional data items. Until then, traditional survey collection instruments will need to collect the remaining items and respondents will still take on burden. Additionally, even if every retailer that works with NPD is onboarded to the project, this still only captures about 30 percent of the retail estimate. How the remaining 70 percent of the retail universe can benefit from a similar effort is a question that needs exploring.

In addition to the actual implementation and data product work, the Economic Directorate continues to tackle the difficult issues and questions that accompany the use of third-party data sources not just for retail but across trade areas. One question being considered is how can the risk of a change in availability of a retailer's data from a third party be mitigated. If the contract ends or if a retailer's data is no longer available through a third party, how that retailer onboarded back onto the survey and how that messaging handled are decisions that will need to be made. Additionally, there is also the possibility that a third-party vendor use could create its own data product comparable to an existing Census Bureau data product. Finally, there is not an official set of guidelines in place to determine if the quality of a third-party data source is good enough to use in place of survey data. While determining quality will always have a subjective element, a checklist of objective measurements that could be followed and recorded for third-party data review would allow for a consistent review across the Economic Directorate.

6 REFERENCES

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

Bavdaz, Mojca, Giesen, Deidre, Cerne, Simona Korenjak, et al. (2015). Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes. *Journal of Official Statistics* 31 (4): 559-588.

Bird, Derek, Breton, Robert, Payne, Chris, et al. (2014). Initial Report on Experiences with Scanner Data in ONS [PDF File]. Retrieved from https://www.ons.gov.uk/ons/guide-method/user-guidance/prices/cpi-and-rpi/initial-report-on-experiences-with-scanner-data-in-ons.pdf.

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

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.

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data [PDF File]. Retrieved from https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf.

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

Hutchinson, Rebecca J. (2017). Reducing Survey Burden Through Third-Party Data Sources. Paper from the Conference of European Statisticians, Workshop on Statistical Data Collection, October 10–12, 2017, Ottawa, Canada, Working paper 1-4 [PDF File]. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ ge.44/2017/mtg3/DC2017_1-4_Hutchinson__USA_AD.pdf.

IBM (2016). IBM Researchers Use Grocery Scanner Data to Speed Investigations During Early Foodborne Illness Outbreaks. Press Release (12 August).

Krebs, Alliston (2016). What’s In a Number? Leveraging Scanner Data to Understand the Retail Beef Consumer [PDF File]. *Beef Issues Quarterly* (13 October). Retrieved from http://www.beefissuesquarterly.com/CMDocs/BeefResearch/BIQ/Fall%202016%20Full%20BIQ.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.

Lohr, Steve (2010). A Data Explosion Remakes Retailing. New York Times (January 2), p. BU3.

Miloslavskaya, Natalia, Tolstoy, Alexander (2016). Big Data, Fast Data, and Data Lake Concepts. *Procedia Computer Science* 88: 300-305.

United State 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

United States Office of Personnel Management (2011). Paperwork Reduction Act (PRA) Guide Version 2.0 [PDF File]. Washington, DC: Author. Retrieved from https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf

Willeboordse, A. (1997). Minimizing Response Burden. In A. Willeboordse *Handbook on Design and Implementation of Business Surveys*: 111-118 [PDF File]. Luxembourg: Eurostat. Retrieved from http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook%20on%20surveys.pdf.

Willimack, Diane K. and Nichols, Elizabeth (2010). A Hybrid Response Process Model for Business Surveys. *Journal of Official Statistics* 25 (1): 3-24.