CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection

10-12 October 2017, Ottawa, Canada

WP.1-4

22 August 2017

Reducing Survey Burden Through Third-Party Data Sources

Rebecca J. Hutchinson (United States Census Bureau)

Rebecca.J.Hutchinson@census.gov

Abstract

Increased respondent burden has led to declining response rates for many of the United States Census Bureau's economic data products. Many high-burden or non-reporting companies already provide comparable date to private sector companies. Can that data be used by the Census Bureau to reduce respondent burden while maintaining or enhancing the quality of its published data?

This paper details a proof-of-concept project undertaken by a retail big data team within the Economic Directorate of United States Census Bureau along with the NPD Group, Inc., a privately held market information and business solutions company that captures point-of-sale transaction data from major retailers. To determine the data's quality and usability, the Team compared NPD' store and national-level data feeds to the Monthly Retail Trade Survey and to the 2012 Economic Census. Additionally, product data from the Economic Census was matched to the product-level data feeds provided by NPD.

The preliminary findings of the work have been positive for store-level and national-level comparisons. The product-level work identified mapping issues between the Census Bureau's product classification system and NPD's proprietary product classification system. The project has also highlighted concerns of financial sustainability when using third-party data sources.

Reducing Respondent Burden through Third-Party Data Sources

Rebecca J. Hutchinson [1]

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

Abstract: Increased respondent burden has led to declining response rates for many of the United States Census Bureau's economic data products. Many high-burden or non-reporting companies already provide comparable date to private sector companies. Can that data be used by the Census Bureau to reduce respondent burden while maintaining or enhancing the quality of its published data?

This paper details a proof-of-concept project undertaken by a retail big data team within the Economic Directorate of United States Census Bureau along with the NPD Group, Inc., a privately held market information and business solutions company that captures point-of-sale transaction data from major retailers. To determine the data's quality and usability, the Team compared NPD' store and national-level data feeds to the Monthly Retail Trade Survey and to the 2012 Economic Census. Additionally, product data from the Economic Census was matched to the product-level data feeds provided by NPD.

The preliminary findings of the work have been positive for store-level and national-level comparisons. The product-level work identified mapping issues between the Census Bureau's product classification system and NPD's proprietary product classification system. The project has also highlighted concerns of financial sustainability when using third-party data sources.

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 though, 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.

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

[1] Disclaimer: Any views expressed are those of the author and not necessarily those of the United States Census Bureau.

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?

This paper details a proof-of-concept project undertaken by a retail big data team within the Economic Directorate of United States Census Bureau to test the quality of third-party retail data purchased from a private company and to examine if this data could be used in place of data that would otherwise be collected on a survey or Economic Census form.

2 Respondent Burden

Through survey data collection, the Census Bureau obtains critical data to present accurate snapshots of the nation's economy. Sampled retailers can receive both the Monthly Retail Trade Survey (MRTS) and Annual Retail Trade Survey (ARTS); all retailers must also complete the Economic Census for each establishment or store location every five years. However, these same retailers are part of the Census Bureau's Business Register which is the sampling frame for many surveys conducted within the Economic Directorate. Larger, more diversified companies may be included in surveys in other industry trade areas including wholesale, manufacturing, and services.

Figure 1 displays percentage breakouts of respondent burden for multi-unit retailers—those retailers with more than one establishment or store location—in MRTS for 2015. Burden is captured here by number of survey forms received. Most MRTS multi-units receive no more than five forms.

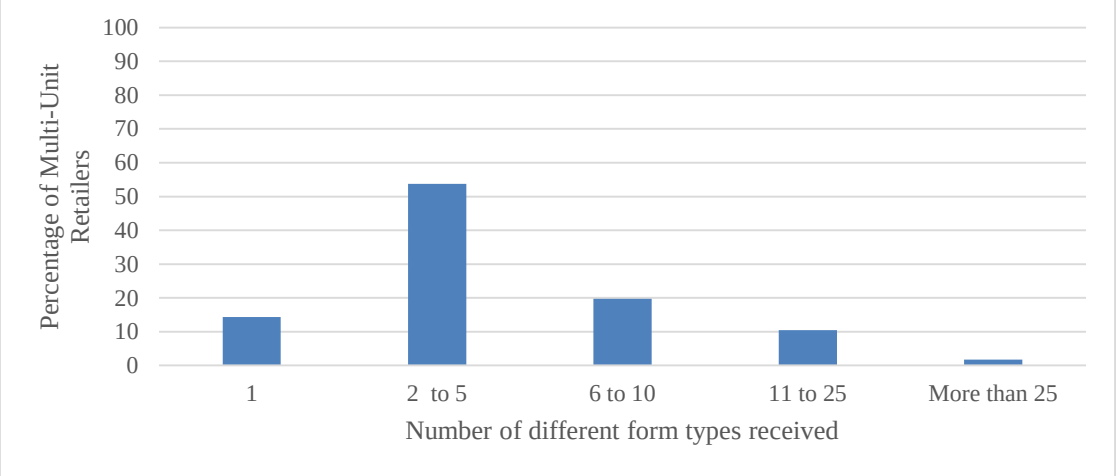

Figure 1: 2015 Form Burden for Multi-Unit MRTS Retailers
Source: U.S. Census Bureau data

However, it is also possible that these retailers are being sampled in surveys conducted by other government agencies as well.

The Census Bureau truly values the critical data that retailers are providing but is also quite cognizant of the burden survey response can place on retailers. To that end, the Census Bureau is dedicated to reducing respondent burden while maintaining the high quality and integrity of its official statistics.

3 Third- Party Data

Scanner data, also known as point-of-sale data, is one possible third-party source that may help reduce burden. Scanner 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 in retail stores or online. Scanner data can provide information about quantities, product characteristics, prices, and the total value of goods sold. Feenstra and Shapiro (2003) describe many potential benefits of using scanner data for improving economic measurement, specifically for estimating price indices. Benefits include reducing or eliminating sampling error, increasing the frequency of measurement, and providing detailed product-level information.

The NPD Group, Inc. (NPD) is a private-market research company that captures point-of-sale data from over 1,250 retailers representing 300,000 stores and e-commerce platforms worldwide. From each store location, NPD receives and processes data feeds containing aggregated scanner transactions by product. NPD edits, analyzes, and summarizes the point-of-sale data feeds at detailed product levels and creates market analysis reports for its retail partners.

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. Any price reductions or redeemed coupon values are filtered out before NPD receives the feeds; thus, the sales figures in the feed reflect the final amount that the customer paid, which aligns to the net revenue total for the company. Sales tax and shipping and handling are excluded. NPD does not receive data on individual transactions or purchasers.

NPD processes data for many industries including but not limited to apparel, small appliances, automotive, beauty, fashion accessories, books, consumer electronics, diamonds, footwear, office supplies, toys, and jewelry/watches. While NPD receives a feed of total store point-of-sale activity, NPD currently classifies only selected industries. Any sales on items that do not belong in these classified 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.

4 Proof-of-Concept Description

The retail big data team worked with NPD to formulate a proof-of-concept project to test the quality and feasibility of using NPD data feeds to reduce respondent burden. To uphold the confidentiality and privacy laws that the Census Bureau holds sacred and to facilitate a productive project, a small number of NPD staff working on this project had background investigations completed and were granted Special Sworn Status. With this Status, NPD staff must uphold the data stewardship practices and confidentiality laws put in place by United States Codes 13 and 26 for their lifetimes.

This proof-of-concept had two goals:

- Compare NPD data at the national and store levels to Monthly Retail Trade Survey (MRTS) data and 2012 Economic Census data.
- Determine feasibility of mapping NPD product lines to 2012 Economic Census product lines.

Three retailers were selected for this project that were good reporters to all Census retail data collections (monthly, annual, and census); good, reported data was necessary to perform a baseline quality check of the NPD data. NPD reached out to their contacts at these retailers and obtained formal approval for the retailer's NPD feeds to be used in this research project.

5 Data Description

The Census Bureau purchased datasets for three retailers from NPD for the first phase of this project. The retailer datasets contained monthly data by store and product level, i.e. for a given month, sales for Product Z in Store A. 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 United States and included values for the following variables: time period (month/year), retailer name, store number, zip code of the store location, channel type (brick & mortar or e-commerce), imputation flag, product classifications by industry, category, and subcategory, and sales figures. One observation for each month/year for each store includes a total sales value of the unclassified data.

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.

For this project, monthly data for 2012 through 2015 was obtained for the selected retailers. When doing these types of comparisons, data is needed for a year in which the Economic Census is conducted (years ending in "2" or "7") to have a baseline comparison of total retailer activity. The Economic Census also collects the most comprehensive store and product-level data of any of the Census Bureau's data collection vehicles.

6 Proof-of-Concept Results

To review the NPD data and achieve the goals of the proof-of-concept project, the following comparisons were made between:

- National-level monthly whole-store tabulations (brick & mortar sales plus e-commerce sales) between the NPD data and the MRTS data.
- E-commerce sales for NPD data and the MRTS data.
- Store-level location and sales between NPD data and 2012 Economic Census.
- Product-line categories in the NPD data and in the 2012 Economic Census.

The results of these comparisons follow below.

6.1 National Level Data Comparisons

At the national level, the NPD data for each of the retailers lined up well when compared to the data reported by the retailers to the MRTS. Figure 2 shows the comparison of NPD and MRTS data using an indexed sum of whole store sales (brick & mortar plus e-commerce) for the three retailers.

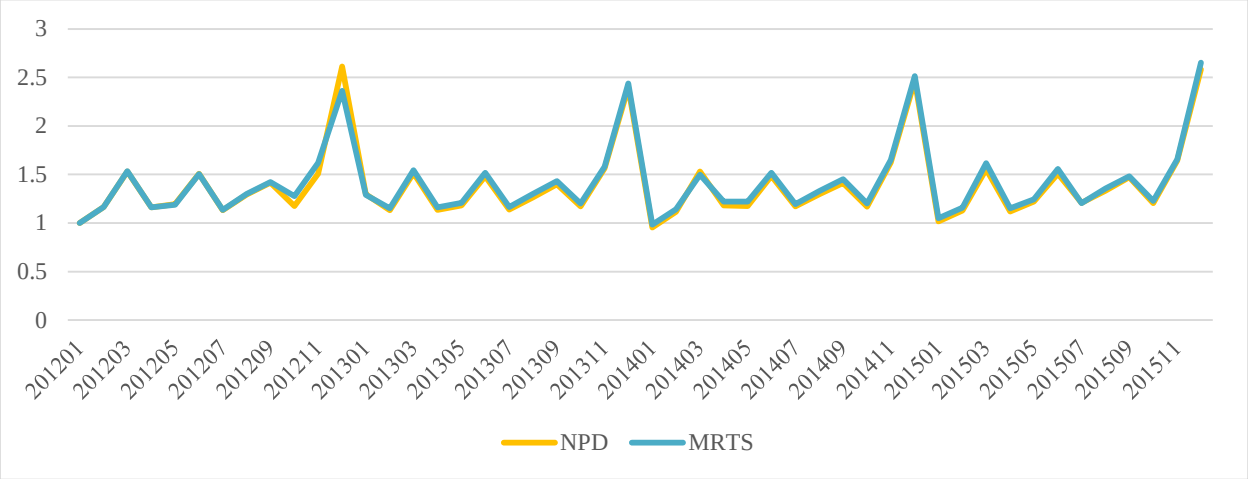

Figure 2: Whole Store Indexed Sales Comparison for Selected Retailers
Source: NPD and Monthly Retail Trade Survey data

As can be observed in Table 1, the differences between the NPD retailer data and the MRTS data were small and most months had absolute differences of less than one percent.

	Retailer 1	Retailer 2	Retailer 3
Average difference between NPD and MRTS data	-0.08%	-0.64%	0.56%
Average absolute difference between NPD data and MRTS data	2.20%	1.79%	1.87%
Median difference between NPD and MRTS data	-0.05%	-0.55%	-0.01%
No. of months with absolute difference between NPD and MRTS data <1%	32/48	35/48	24/48

Table 1: Descriptive Statistics for Differences between National Retailer NPD Data and MRTS Data
Source: NPD and Monthly Retail Trade Survey data

During this phase of the analysis work, 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. The team worked with NPD to identify the source of the issue and found the retailer had changed the format of its feed and an incorrect data feed had overwritten the original, correct data. This issue had gone undiscovered 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. This issue demonstrated one of the risks of relying on third-party data but also highlighted that the analysis work done in these types of collaborations have the potential to benefit the third-party data providers as much as they benefit the Census Bureau.

6.2 E-Commerce Comparison

The Monthly Retail Trade Survey also collects data measuring online sales from these three retailers. This reported MRTS data was compared to the NPD e-commerce data. The finding highlighted a common issue with measuring e-commerce sales: there is currently no industry standard or definition for what qualifies as an online sale. If a customer places an order online and picks it up in store, some retailers will classify that as an online sale; others will claim it as an in-store sale. And as a retailer included in this research demonstrated, some may change their e-commerce accounting at a

given point in time. In Figure 3, e-commerce data line up well until 2015 when NPD data drops below that of what was reported to MRTS. The reason for the difference was that this retailer likely began classifying in-store pick-ups of online purchases as in-store sales. Interestingly enough, it appears that that same retailer did not make the same change to the data it was reporting to MRTS.

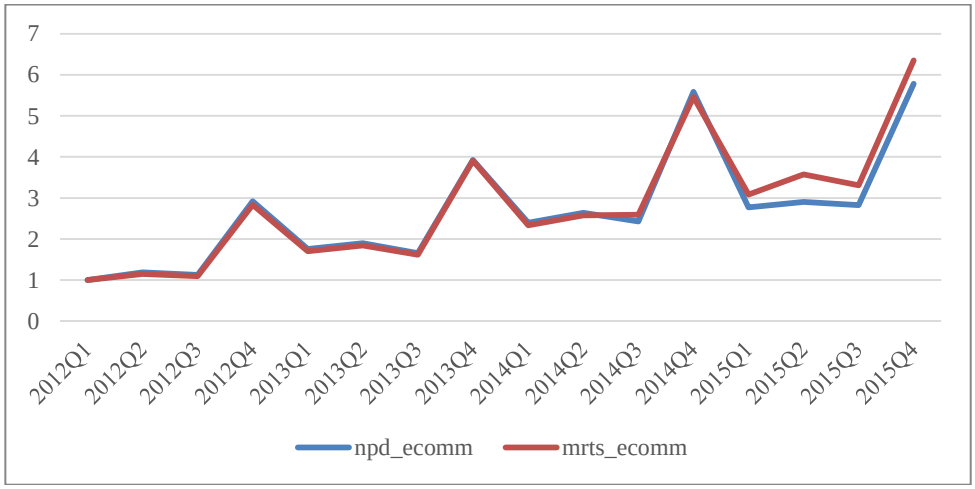

Figure 3: Indexed E-Commerce Quarterly Sales (First Quarter 2012 =1.000) for a single retailer
Source: NPD and Monthly Retail Trade Survey data

This is an example of how reconciling what is included in the NPD data feeds versus what is reported to MRTS is one of the measurement challenges involved with using third-party data.

6.3 Store-Level Data Comparisons

Unlike many of the Census Bureau surveys where data is collected at the company level, the Economic Census is conducted at the store or establishment level. In order to reduce the respondent burden involved, a third-party data source must have a similar, granular level of data collection. One of the great features of the NPD data feeds is the rich, store-level data that they capture.

The store-level data from 2012 NPD was compared to the store-level data that each retailer provided to the 2012 Economic Census. The inclusion of the store number in the NPD datasets allowed for a clean and logical match to the Economic Census database which includes a store number variable in each location record.

For the purposes of the proof-of-concept work, the analysis focused on how well the store locations and store sales lined up. Only one store location among all three

retailers did not match between the NPD and Economic Census data. As displayed in Table 2, the differences in sales at the store-level were on average under 3% and in most cases, lower than that.

	Retailer 1	Retailer 2	Retailer 3
NPD/2012 Econ Census Store Location Match Rate	100%	99.9%	100%
Average Percentage Difference in Store Sales Between NPD & 2012 Econ Census	-2.36%	-1.22%	0.96%
Median Percentage Difference in Store Sales Between NPD & 2012 Econ Census	-2.18%	-1.18%	0.31%

Table 2: Descriptive Statistics for Differences between Store-Level Retailer NPD Data and MRTS Data
Source: NPD and 2012 Economic Census data

6.4 Product Mapping

In its data feeds, NPD collects product data at a detailed SKU level. On the 2012 Economic Census, the Census Bureau collected data at a product line code level.[2] The methods of product categorization were developed by each organization to meet the differing needs of their data users.

Mapping the product lines to one another is the only way the NPD product-level data can be meaningfully used to reduce the burden of reporting product data on the Economic Census. This mapping exercise is not a simple one. During the proof-of-concept phase, a high-level product mapping of the apparel category was completed. Figure 4 displays the complexity of this mapping of what was perceived by the team to be the easiest product category to map.

Notable issues highlighted by this mapping exercise included:

- Level of comparison. In this example, the Economic Census breaks out apparel by men's, women's, and children's apparel items. NPD's breakouts are by apparel type. NPD may be able to provide additional attributes that would identify if the apparel product types as men's or women's.
- Limited one-to-one mappings. The lone perfect one-to-one product-line match between NPD and the Economic Census was "dresses." Smaller buckets could

[2] Beginning with the 2017 Economic Census, product-line data will be collected based on the North American Product Classification System (NAPCS).

be 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 Economic Census product lines.

7 Conclusion and Future Work

The results of the proof-of-concept work were a promising step towards using third-party data to reduce respondent burden. NPD is a collaborative partner that understands the goal of the work and can adapt its processes to meet the project's needs. Based on the success of the proof-of-concept, two additional projects have been proposed to further research the use of retailer data from NPD.

The first project addresses respondent burden and non-response in the Monthly Retail Trade Survey. For this project, twenty to forty retail data feeds would be used for companies that are either non-respondents to MRTS or have been identified as having a high response burden across Census Bureau economic surveys. The goals of this project would first be to simulate MRTS estimates using the alternative data source for those companies in place of the reported and/or imputed estimates previously used and determine the impact of its use on the industry estimates. Additionally, the NPD data would be used to conduct a validation exercise of current MRTS imputation methods to create sales figures for non-respondents.

The second project addresses respondent burden in the Economic Census. For this project, the focus would be on NPD feeds for retailers within a single NAICS code. NPD staff and Census Bureau product classification staff would map NPD products to the North American Product Classification System. The data would also be used to generate experimental Economic Census store and product-level estimates in early 2018 before the 2017 Economic Census forms are even mailed out. This work would then be compared to actual Economic Census figures later in the year to determine the feasibility of expanding this effort in future Economic Censuses and possibly creating more frequent product-level reports.

This work has shown great potential for using third-party data sources and the Census Bureau continues to explore ways to incorporate them into its products and methodologies. However, cost is one large limitation to expanding their use. Currently third-party retailer data is purchased on an a la-carte basis. This cost structure is not sustainable for large-scale implementation especially when faced with budget limitations.

Figure 4: NPD and Economic Census Product Mapping

References

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.