Enhancing the Foundation of Official Economic Statistics with Big Data

Brian Dumbacher[1], Rebecca Hutchinson[2]

[1] *U.S. Census Bureau, Washington, DC; Brian.Dumbacher@census.gov*
[2] *U.S. Census Bureau, Washington, DC; Rebecca.J.Hutchinson@census.gov*

Abstract
Official economic statistics produced by the United States Census Bureau have long served as a high-quality benchmark for data users. To maintain this quality and enhance the foundation of its economic programs, the Census Bureau has begun exploring the potential of Big Data sources such as credit card transaction data, point-of-sale data, and publicly available building permit data. While this type of data may allow the Census Bureau to improve the timeliness, geographic detail, and product-line coverage of its economic data products, there are concerns such as methodological transparency and consistency of the data. This paper covers the Big Data findings that the Economic Directorate of the Census Bureau has discovered so far as well as the Directorate's Big Data vision for the future.

Keywords: Big Data, Official statistics, Economic statistics

1. Introduction

Official economic statistics produced by the United States Census Bureau have long served as a high-quality benchmark for data users. However, demands for timely and detailed data, a decline in respondent cooperation, an increasingly costly way of collecting data through traditional surveys, and a changing economic landscape are making it challenging for the Economic Directorate of the Census Bureau to meet its data users' needs. Using third-party data that are large, real-time, and granular, i.e. Big Data, could help address granularity and timeliness, but the Census Bureau must still produce high-quality statistics.

1.1 Big Data Vision

To enhance the foundation of its economic programs, the Census Bureau has begun exploring the potential of using Big Data sources such as point-of-sale scanner data, payment processor transaction data, and publicly available building permit data. The Economic Directorate envisions leveraging Big Data sources such as these in conjunction with existing survey 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 passive data collection and web scraping could also play a large role in reducing burden on respondents and Census Bureau analysts.

1.2. Big Data Concerns

Incorporating Big Data 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 uses of third-party data. The United States Office of Management and Budget and professional associations such as the American Association for Public Opinion Research (AAPOR) and the American Statistical Association have begun looking more closely at how to evaluate the quality of third-party data and statistics derived from them (AAPOR, 2015).

1.3. Goals of the Economic Big Data Team

The priority of the Economic Directorate has been to examine the potential of using third-party data to enhance its retail programs, specifically the Monthly Retail Trade Survey (MRTS), the Annual Retail Trade Survey (ARTS), and the retail component of the Economic Census, as well as its construction programs, specifically the Building Permit Survey (BPS), the Survey of Construction (SOC), and the Nonresidential Coverage Evaluation (NCE). In 2015, the Big Data Team was formed and tasked with exploring the following:

- **Adding geographic and industry detail**: Currently, MRTS and ARTS yield estimates of totals at the national level for broad categories of industry. There is interest in producing estimates and economic indicators at more detailed levels of geography such as states and metropolitan statistical areas and at more detailed levels of industry.
- **Reducing respondent burden**: Data on retail sales for respondents may already exist in the form of transaction data, and data on construction may already exist in public records. Using these data in lieu of having respondents complete a questionnaire could reduce or eliminate respondent burden.
- **Improving timeliness**: Economic indicators need to be released in a timely manner to be useful to decision makers and data consumers. Using real-time data has the potential to improve the timeliness and frequency of economic data products.
- **Extending coverage to product lines**: The Census Bureau classifies business establishments by North American Industrial Classification System (NAICS) code for the purposes of collecting and analyzing data and publishing estimates. In 2017, a complementary product-based classification system called the North American Product Classification System (NAPCS) will be implemented. NAPCS focuses on the goods and services themselves and not on their industry of origin. Big Data sources that are product-based could help produce new estimates for product lines.

2. Progress

This section highlights findings from three exploratory projects that the Big Data Team has worked on and identifies areas of concern and promise. These projects involve point-of-sale scanner data, payment processor transaction data, and publicly available building permit data.

2.1. Scanner Data

In January 2015, the Big Data Team kicked off with a project to determine whether point-of-sale data, also known as scanner data, have the potential to supplement retail programs. Scanner data are detailed data on sales of consumer goods obtained by scanning the bar codes of products at electronic points of sale in retail stores. 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.

For this project, the Census Bureau purchased scanner data from NPD Group, Inc. NPD collects scanner data from 1,200 retail partners with 165,000 stores worldwide. From each store location, NPD receives and processes data feeds containing aggregated scanner transactions by product. 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. NPD does not receive data on individual transactions or purchasers. However, all forms of payment are captured, which is one advantage of scanner data. Sales tax and shipping and handling are excluded. NPD processes data for many industries including apparel, appliances, automotive, beauty, consumer electronics, footwear, office supplies, toys, video games, and jewelry and watches. NPD edits, analyzes, and summarizes the point-of-sale data at detailed product levels and creates market analysis reports for its retail partners.

The Census Bureau purchased scanner datasets covering auto parts and jewelry and watches for January 2012 through December 2014. These datasets were selected because their NPD industry definitions best aligned with the NAICS codes used by the Census Bureau for the budget that the Big Data Team had. The NPD data were provided geographically at the Designated Market Area (DMA) level. DMAs are useful for marketing purposes because they are roughly defined as areas that receive similar television news stations and newspapers.

2.1.1. Auto Parts Research

The NPD auto parts data accounted for only a small fraction of the total MRTS auto parts sales. Compared to the annualized 2012 MRTS total of $82.9 billion in NAICS 4412, the NPD data covered $10.4 billion. This is likely due to two factors. First, the auto parts data covered a relatively small portion of the overall market in automotive parts, accessories, and tire stores.

Second, the auto parts data excluded a large number of purchases that are included in MRTS estimates. Despite this, the month-to-month trends tracked each other fairly well, aside from significant increases in NPD sales around January.

The team also explored using third-party data to supplement the product-level Economic Census questionnaire. Because of differences in NPD and Census Bureau product-line classifications, the only auto parts category that was a one-to-one match was the batteries category. In 2012, the Economic Census estimated $4.7 billion in auto batteries, whereas NPD captured $2.4 billion in sales. All other NPD categories were a many-to-many match, making it difficult for comparisons.

2.1.2. Jewelry and Watches Research

The NPD jewelry and watches data consisted of two datasets. The jewelry dataset captured transactions only on branded jewelry, which is about 20 percent of the jewelry market. The NPD watches dataset was comprised of aggregated watch sales from three types of stores: chain jewelry stores, discount stores, and department stores. Watch sales from the chain stores would be the only sales captured in the jewelry category for MRTS (NAICS 44831), and any sales at the discount or department stores would be captured in other MRTS NAICS codes. Thus, over the period January 2012 through December 2014, only between 16 and 27 percent of NPD watch data each month would contribute to the MRTS jewelry estimates.

With those exclusions, the two datasets combined accounted for only a small piece of the jewelry industry. Annualized MRTS data estimated the jewelry industry to have approximately $30.8 billion in sales in 2012 whereas NPD total jewelry and watches data captured $2.0 billion in sales. NPD sales were about five percent of the MRTS sales. The month-to-month trends tracked each other well, with the NPD month-to-month trends having similar shape and magnitude to those of the MRTS data. Moreover, the NPD and MRTS data had the same seasonal peaks in February, May, June, and December, which correspond to holidays when jewelry purchases are common: Valentine's Day, Mother's Day, Father's Day/graduations/weddings, and Christmas, respectively. The team theorized that on special holidays, consumers may be more inclined to purchase branded jewelry.

2.1.3. NPD Summary of Findings

The Big Data Team noted concerns with coverage, product definitions, and geographic definitions. Retailers represented by the NPD data are systematically different from the full universe of retailers in the Economic Census data: they are large, multi-unit companies. The underrepresentation of small firms and establishments in the NPD data introduces biases towards the characteristics and trends of large firms.

Results for the auto parts data were not promising. Front-store sales and the omission of oil products, car care products, etc., made comparisons with Census Bureau data tenuous at best. The fact that MRTS and NPD cover different types of store is problematic for jewelry because in the time span of the NPD data, only between 16 and 27 percent of the NPD data each month is in scope to MRTS. Despite this, based on simple linear regression models, the NPD data explained 90 percent of the MRTS monthly trend in jewelry.

The team struggled to create direct comparisons between NPD's product data tabulations and the Census Bureau's sales and product-line sales data. This was due to the differences in product detail and fundamental differences in classification. NPD product categories are a many-to-many match to NAPCS codes, and there are very few direct links. Census Bureau product lines are often at a much higher level of aggregation. NPD's product lines, on the other hand, are at an arbitrary level of aggregation because the base unit is so detailed. For instance, NPD may have product groupings at such detailed levels as "premium silicone wiper blades." These could, in principle, be aggregated to match the Census Bureau's product lines, but the concordance would require a significant amount of work. Also, NPD's defined geographic unit is the DMA. For statistical purposes at the Census Bureau, this unit is of no value as DMAs can overlap or exclude entire areas.

The Big Data Team identified several points of risk related to using these data. First, this project highlighted the importance of transparency. Confidentiality agreements in place with NPD would inhibit the Census Bureau as a federal statistical agency to be transparent about its methodology. The third-party data source and the Census Bureau must agree on a level of transparency satisfactory to both parties. Additionally, NPD makes case-by-case agreements with large retailers. Those agreements define the coverage of NPD data. The Census Bureau

would have very little control over those agreements and the impact that changes in those agreements would have on the data. Lastly, the NPD data are not a probability sample and do not represent the entire universe of establishments engaged in retail trade as defined by the Census Bureau. The NPD and MRTS trends can be similar, but one would not expect the NPD totals to be near the MRTS estimates.

Further work involves having NPD explore obtaining explicit permission from the retailers to share store-level retail data feeds with the Census Bureau. The focus would be on product information for the 2017 Economic Census and sales for MRTS. Rather than having companies fill out a questionnaire, the Census Bureau might be able to obtain some of the necessary information from data feeds. This would reduce respondent burden. Given concerns about product-line alignment, the Big Data Team is currently working on a proof of concept that would allow them to study not only how well the data align, but also to identify any issues with definitions, items collected, and overall usefulness of the NPD data.

2.2. Payment Processor Point-of-Sale Data

In November 2015, the Big Data Team joined the Bureau of Economic Analysis to work with a payment processing company, First Data (FD), and a software developer, Palantir, on an exploratory project to analyze transaction data. FD serves as an intermediary between a customer swiping a debit or credit card and the card's financial institution and captures about 45 percent of all non-cash transactions in the United States. Palantir has taken FD data stored on different acquisition platforms and integrated them into a single platform.

This project involved analyzing FD aggregates of sales via a data analytics tool developed by Palantir. The tool is a secure website consisting of custom dashboards, a data management system, and an environment for performing analyses using R and Python. The Big Data Team was given access to daily, weekly, and monthly aggregates for five states across many sectors of the economy, including retail and services, for October 2012 through April 2016. To assign industry to transactions, Palantir mapped the Merchant Category Code (MCC) to NAICS code.

2.2.1. Small Area Estimation Research

The FD data have good coverage for many industries. National aggregates of sales are predictive of published MRTS estimates. Aggregates at more granular levels can serve as useful covariates for area-level small area estimation models such as the Fay-Herriot model (Fay and Herriot, 1979). The small area framework offers a nice setting in which to balance the properties of direct survey estimates against those of model-based estimates with FD aggregates as covariates. Using Big Data covariates in such a way is mentioned in Capps and Wright (2013).

One limitation of this research is that, because of the way data are collected for MRTS, direct survey estimates for low geographic levels cannot be calculated. Also, the FD aggregates are available as indexed values within industry as opposed to dollar values, so comparing these values across industries posed a challenge. The Big Data Team tried various approaches based on alternative inputs and transformations to put everything on the same scale. Preliminary model results showed that the Fay-Herriot model with FD aggregates of sales as covariates helped reduce variability for state-by-industry-level estimates.

2.2.2. Trading Day Weight Research

The Big Data Team also looked at using daily FD transaction data to calculate national trading day weights for retail NAICS codes to reflect varying levels of activity during the week. Models were fit to estimate trading day weights based upon a weekly effect modeled by McElroy (2016) and account for the effects of holidays on retail sales. The team compared these modeled weights with the trading day weights currently calculated for MRTS using the X-13 autoregressive integrated moving average seasonal adjustment program and looked for areas of improvement.

2.2.3. Summary of First Data and Palantir Findings

The FD data are rich and offer opportunities to improve methodology at different stages in the survey life cycle such as weighting and estimation. Advantages of payment processor data over data from a single credit card network include greater coverage and less sensitivity to the

variations in a single network's behavior. Aspects out of the Big Data Team's control include the MCC-to-NAICS conversion and calibration of the underlying microdata to account for the changing nature of merchants in FD's pool of clients. Palantir has sought the Big Data Team's input at every step and has documented its methodology well.

2.3. Publicly Available Building Permit Data

New construction data collected by the Census Bureau are used by government agencies, policy analysts, and others to measure and evaluate size, composition, and change occurring within the construction sector. To measure new construction, the Census Bureau conducts the BPS, the SOC, and the NCE. Issues regarding respondent burden and data collection for these surveys are similar to that of establishment surveys such as MRTS and ARTS. This is especially true for respondents that may receive requests for all three construction surveys. Additionally, survey costs are rising, and response rates are falling. Taking all these issues into consideration, the prospect of expending more resources to obtain less data from increasingly burdened respondents takes on more importance.

In October 2015, a research project addressing these issues was conducted. This research examined incorporating publicly available building permit data (currently in the form of Application Programming Interfaces from Chicago, IL and Seattle, WA) into new construction surveys. Survey validity is currently being tested prior to formal incorporation to ensure new data sources account for all new construction. During the initial research phase, validation of these potential new data sources was conducted against estimates from the BPS. In March 2016, additional research began where validation was conducted against estimates from the SOC and NCE surveys.

There are several risks involved with this Big Data approach. First, it is reasonable to assume publicly available building permit data will not be obtainable for all areas in the United States. Building permit data will likely be available for areas where new construction activity is large or increasing. Areas where new construction is minimal or limited may not be willing to invest necessary resources to incorporate this technology. Frequency of new data source updates is also important to address. Infrequent updates or potential missed updates

due to human or computer error will introduce uncertainty and error into survey estimates. Another risk is that publicly available building permit data will not provide complete information on new construction. For example, information on housing units and specific physical characteristics is generally lacking at the level of detail needed for estimation. In many cases, these new data sources will only provide broad construction information.

3. Conclusion

The Big Data Team's work so far with scanner data, payment processor transaction data, and building permit data has shown promise for helping the Census Bureau meet its challenges and enhance the foundation of its official economic statistics. In its research, the team also has identified concerns using third-party data that touch on the following data quality attributes (Wang and Strong, 1996):

- **Accuracy**: The data are error free; outliers and possible errors can be identified.
- **Consistency**: The data and the format in which they are provided are consistent over time.
- **Transparency**: Methodologies and data processing procedures are transparent and well documented.
- **Representativeness**: The data represent the target population; coverage of the data, differences in classification, and biases can be understood.
- **Completeness**: The data are complete; patterns of missingness can be understood.
- **Access security**: Access to the data can be restricted.
- **Continuity**: The company providing the data may stop collecting data.

These qualities are important for the Big Data Team to keep in mind and attempt to measure as it explores new sources of Big Data and potential applications. Next steps for the team include continuing research with Palantir and First Data and developing a new economic data product based on Big Data.

4. References

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

Capps, C. and Wright, T. (2013). Toward a Vision: Official Statistics and Big Data. *AMSTAT News (August)*, Alexandria, VA: American Statistical Association, pp. 9-13.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, *74*(366), pp. 269-277.

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.

McElroy, T.S. (2016). "Modeling and Seasonal Adjustment of Weekly and Daily Time Series." Working Paper.

Wang, R.Y. and Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, *12*(4), pp. 5-33.