# Using Passive Data Collection, System-to-System Data Collection, and Machine Learning to Improve Economic Surveys

Brian Dumbacher[1], Demetria Hanna[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
Brian.Dumbacher@census.gov, Demetria.V.Hanna@census.gov

**Abstract**
As part of the ongoing effort to improve its economic surveys, the U.S. Census Bureau is exploring alternative data collection methods with the goal of reducing respondent burden and enhancing the efficiency of data processing. Some of these methods belong to the category of passive data collection, in which the respondent either has little awareness of the data collection effort or does not need to take any explicit actions. Examples include scraping data from respondents' websites and obtaining respondent data from third parties that have already collected it. Other methods belong to the category of system-to-system data collection, which involves respondents transferring data directly from their computer systems to the Census Bureau's systems. In this paper, we outline the Census Bureau's data collection vision for its economic programs and describe recent work on exploring the potential of alternative methods. We also explain how machine learning can be used to assist in collecting and processing data, especially data scraped from websites. Lastly, we describe concerns and challenges associated with all of these methods.

**Key Words:** Big Data, official statistics, economic statistics, data collection, public-private partnerships

## 1. Introduction

### 1.1 Challenges

Official economic statistics produced by the U.S. Census Bureau have long served as a high-quality benchmark. However, the Census Bureau faces many challenges in producing official economic statistics that continue to meet data users' needs. First, data users are demanding data that are more timely and granular. External data sources produce data faster and offer insights into the economy that are more detailed. At the same time, the Census Bureau faces fiscal pressures and possibly fewer resources. The economic landscape is constantly changing as well, and companies empowered by new internet tools, social media, and start-up funding are making it increasingly difficult to measure today's economy accurately.

*Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.*

Another key challenge, which is the focus of this paper, is declining respondent cooperation. The Census Bureau must find and adopt new ways of gaining cooperation from respondents and streamlining data collection. Associated with this challenge are the costs of current data collection via traditional paper questionnaires and electronic instruments and certain aspects of data processing that are manually intensive. To help address these challenges, the Census Bureau is exploring alternative data collection methods and data sources. The goals are to increase respondent cooperation, improve response rates, reduce respondent and analyst burden, save costs, and enhance the efficiency of data collection operations while maintaining the quality of data products.

## 1.2 Data Collection Vision

The data collection vision for the Census Bureau's economic programs is to maximize the use of alternative data collection methods, alternative data sources such as Big Data, and machine learning, which can help automate certain aspects of data collection and processing. The following are definitions and examples of the alternative data collection methods that are a part of this vision:

- **Passive data collection**: This is a type of data collection in which the respondent either has little awareness of the data collection effort or does not need to take any explicit actions. Collecting data passively has the potential to reduce burden and costs significantly.

- **Web scraping**: Collecting, or scraping, data from web sources automatically is an example of passive data collection. For economic surveys conducted by the Census Bureau, respondent data or equivalent-quality data can sometimes be found online on respondent websites, in public filings with the Securities Exchange Commission (SEC), or through Application Programming Interfaces (APIs). There are opportunities to go directly to these types of web sources for economic data. For an excellent overview of web scraping, see Mitchell (2015).

- **Informed consent data collection**: This is a specific type of passive data collection involving a third party that has already collected data from the Census Bureau's respondents. An example of such a third party is a private transaction aggregator that collects sales data from retailers for market analysis research. The third party then provides the data to the Census Bureau with the informed consent of the respondents.

- **System-to-system data collection**: System-to-system data collection involves respondents transferring data directly from their computer systems to the Census Bureau's systems. Data obtained in this way probably would come in large data dumps and cover a variety of economic variables such as expenditures, revenue, and inventory. This type of collection would allow companies to provide information to multiple surveys at the same time using a single portal, which is another way to reduce respondent burden (Snijkers *et al.*, 2013, p. 242).

Many of the data sources mentioned above can be considered Big Data, and associated machine learning methods can be used in various ways. The following are definitions and examples related to alternative data sources and machine learning:

- **Big Data**: Big Data generally refers to data sources characterized by three V's (Kreuter and Peng, 2014): volume (a large number of observations or variables), velocity (real-time or frequently generated data), and variety (various data formats and structures). Big Data are also known as "found," "organic," and "undesigned" data, adjectives that convey the notion that the data are being used for a purpose other than the one for which they were created. Sources such as electronic transaction data, data dumps from private companies, and administrative records have the potential to supplement or even replace survey data and can offer greater geographic, industry, and product-level granularity. Evaluating the representativeness, consistency, and general quality of data from these sources is a time-consuming but very important first task.

- **Machine learning**: This interdisciplinary field covers topics such as predictive modeling, association analysis, and clustering. Supervised learning, in particular, involves predicting an outcome measurement using a set of predictors, or features, and a training set of data (Hastie *et al.*, 2009, chap. 1). In the context of data collection, machine learning can be used to help scrape data from respondent websites, classify data to match Census Bureau definitions, and automate coding processes that are largely manual.

### 1.3 Outline

The rest of this paper describes various projects that the Census Bureau is undertaking to explore the potential of using alternative data collection methods and sources to improve its economic surveys. Sections 2 and 3 describe passive data collection projects on scraping public sector data and building permit data, respectively, from the web. Section 4 describes a pilot project with a private company, The NPD Group, Inc., on informed consent data collection from retailers. Next, Section 5 covers a pilot project on system-to-system data collection involving large private sector companies. Section 6 describes research on using machine learning to automate the assignment of North American Industry Classification System (NAICS) codes to business establishments. Each of these sections outlines associated challenges and future work. Finally, Section 7 summarizes all of these efforts.

## 2. Public Sector Web Scraping

### 2.1 Background

The Census Bureau conducts many public sector surveys that collect data on public employment and finance from state and local governments. One such survey is the Quarterly Summary of State and Local Government Tax Revenue (QTax), which collects data on tax revenue collections. The taxes in scope to QTax include general sales and

gross receipts tax, individual income tax, and corporate net income tax. Much of this data is publicly available on government websites. In fact, instead of responding via questionnaire, some respondents direct QTax analysts to their websites to collect the data. The public sector area is an ideal setting in which to explore web scraping methods because the data on state and local government websites are meant for public consumption and are not confidential.

An automated process for finding useful data sources on tax revenue collections and then scraping the data is ideal but challenging to develop. There are thousands of government websites with very little standardization in terms of structure and publications, so a long-term solution to scraping data needs to be able to adapt to new situations. Also, a large majority of useful documents on state and local government websites are in Portable Document Format (PDF), a file type that does not lend itself to analysis right away.

## 2.2 Project

In September 2015, Census Bureau researchers began developing tools for scraping tax revenue data from state and local government websites. This collection of tools is known as SABLE, which stands for Scraping Assisted By LEarning. Elements of SABLE involve machine learning for performing classification (Tan *et al*., 2006, chaps. 4 and 5). Table 1 below describes the three main tasks that SABLE is being developed to perform in its application to public sector surveys: crawling, scraping, and classifying.

Table 1. Three Main Tasks of SABLE for Public Sector Surveys

| **Crawling** |
|---|
| Given a website,<br>• Crawl the website<br>• Discover all documents on the website (most likely in PDF format)<br>• Apply a classification model to predict whether the document contains useful data on tax revenue collections |
| **Scraping** |
| Given a document classified as useful,<br>• Find the useful data in the document<br>  o Apply a model based on frequencies and locations of important sequences of words, or<br>  o Apply a hard-coded template<br>• Extract numerical values and contextual information such as data labels |
| **Classifying** |
| Given scraped data,<br>• Put scraped data into a normalized data structure<br>• Map scraped data to the Census Bureau's tax classification codes<br>  o Apply a classification model to predict the tax code based on data labels associated with the scraped data, or<br>  o Apply a hard-coded template |

SABLE is based on two main pieces of software: Nutch, an open-source web crawler (Apache, 2017), and Python. Nutch is used to crawl websites, discover PDFs, and compile a training set of documents for model building. Python is used in the rest of the analysis to extract information from PDFs, preprocess the data, apply text analytics techniques, and fit classification models. Researchers are currently considering models such as Naïve Bayes, support vector machines, decision trees, and random forests. These models are based on the one-word and two-word sequences in the documents that are most highly associated with the class labels.

As described in Dumbacher and Capps (2016), this methodology was applied to state government websites with positive results. New data sources of monthly tax revenue collections were discovered, and the classification models used to predict the usefulness of a PDF based on its text achieved high accuracy. Recently, researchers have begun developing templates that can be applied to specific documents to scrape desired tax revenue figures automatically. Lastly, classification models are being developed to aid QTax analysts in assigning tax codes (U.S. Census Bureau, 2017a) to scraped data based on associated data labels and descriptions.

## 2.3 Challenges

Machine learning has been shown to help identify useful documents and assign tax codes, but developing accurate models requires compiling a large, representative, and good-quality training set. This is a manual and often very time-consuming task. As Census Bureau staff learn web crawling and web scraping skills, acquiring the data for training sets will become easier. Another challenge involves the multitude of document formats in current use. PDF is commonly used by state and local governments, but useful data on tax revenue collections also have been found in Microsoft Excel, CSV, TXT, and HTML formats. Developing a manageable and unified approach to scraping data from all documents seems like a challenging task. A possible solution may involve converting documents to PDF format and then applying current methodology to extract the data.

## 2.4 Future Work

The ultimate goal of this project is to create a public sector data product based on data scraped from government websites. One potential product is a monthly summary of state government tax revenue. Because not all state governments publish monthly tax revenue reports, this data product would be based on a panel of state governments. Using a combination of web crawling, internet searches, and tax policy resources, researchers working on SABLE have identified over 30 states that could be a part of this panel. QTax subject matter experts are currently evaluating the usability and general quality of these sources. Researchers also plan to refine the classification models used to assign tax codes to scraped data based on associated data labels and descriptions.

# 3. Building Permit Web Scraping

## 3.1 Background

New construction data collected by the Census Bureau are used by government agencies and policy analysts to measure and evaluate size, composition, and change occurring within the construction sector. To measure new construction, the Census Bureau conducts the Building Permit Survey (BPS), the Survey of Construction (SOC), and the Nonresidential Coverage Evaluation (NCE). As with many economic surveys, survey costs are increasing, response rates are decreasing, and respondents are feeling burdened. This is especially true for respondents that receive requests from all three construction surveys. Information on new, privately owned construction is available online for some building permit jurisdictions, and, as with the public sector data project described in Section 2, it makes sense to explore the feasibility of scraping these data from the web.

## 3.2 Project

In October 2015, research began on examining issues regarding incorporating publicly available building permit data into construction surveys. The initial stage of research focused on two building permit jurisdictions, Chicago and Seattle, whose data are publicly available through APIs. During this stage, data from these two jurisdictions were analyzed to determine advantages, limitations, and implications surrounding incorporation of these new potential data sources. The result of this initial research was a promising first step as the new sources appeared to provide timely and valid data with respect to corresponding BPS data.

In mid-2016, work continued on the project along two fronts. The first front consisted of researching publicly available data for building permit jurisdictions across the U.S., focusing on jurisdictions that issued large numbers of residential permits in 2015. Here information was discovered in different formats. Other than APIs, publicly available building permit data were also obtainable via downloadable reports, Excel files, database queries, and other media.

The second front consisted of additional research into the Chicago and Seattle data sources. Through validation, researchers noticed differences in classifications and definitions from one jurisdiction to another. For example, the term “living space” versus “finished floor space” when reporting residential square footage data. Also, publicly available building permit data do not seem to provide complete information on new construction. 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. More recently in 2017, building permit jurisdictions for Nashville and Boston were included in the research because they appear to provide information found lacking above.

### 3.3 Challenges

Many challenges in incorporating publicly available building permit data from the web into construction surveys are related to the Big Data concerns of representativeness and consistency of the data source. 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 make their information available online. Lastly, as with the public sector data project in Section 2, these data are available in many different formats. A viable solution might have to be able to extract information from APIs, reports, and databases alike.

### 3.4 Future Work

Prior to formally incorporating publicly available building permit data into new construction surveys, a number of consistency issues must be addressed. Important issues involve dealing with certain key characteristics such as construction classification that use various terms and definitions across jurisdictions. Researchers hope detailed text analysis and machine learning will be beneficial. Finally, successful ongoing validation against corresponding BPS, SOC, and NCE data will also be required to ensure appropriate coverage and completeness of building permit information.

## 4. Informed Consent Data Collection Via The NPD Group

### 4.1 Background

Point-of-sale data, or 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 establishments. The NPD Group, Inc. (NPD) is a private company that collects scanner data from hundreds of retail partners and thousands of establishments worldwide. From each establishment, NPD receives and processes data feeds containing aggregated scanner transactions by product. NPD edits, analyzes, and summarizes the data at detailed product levels and creates market analysis reports for its retail partners. NPD collects all forms of payment and processes data for a variety of industries including apparel, appliances, automotive, beauty, consumer electronics, footwear, housewares, office supplies, toys, video games, and jewelry and watches.

These data cover key parts of the retail sector and could be used to supplement or replace survey data from the Census Bureau's Monthly Retail Trade Survey (MRTS), Annual Retail Trade Survey (ARTS), and retail component of the Economic Census. Collecting establishment-level data through informed consent data collection via NPD's data feeds could significantly reduce respondent burden and costs.

### 4.2 Project

To explore the feasibility of informed consent data collection, the Census Bureau recently purchased company-level data from NPD for three private companies. NPD and the Census Bureau selected companies to contact for this study based on their size, the geographic distribution of their establishments, their MRTS, ARTS, and Economic

Census reporting history, and their relationship with both the Census Bureau and NPD. For this pilot project, the Census Bureau is interested in large retailers representing various geographies. Good reporting history would allow better comparisons between the NPD data and survey values reported to the Census Bureau. Good relationships with the Census Bureau and NPD would indicate a more cooperative company for this pilot project. The data consist of sales aggregates broken down by month, industry, channel, and establishment and cover January 2012 through December 2015. Channel refers to either brick-and-mortar or e-commerce. This e-commerce level of detail could offer the Census Bureau new and useful retail insights.

The Census Bureau has developed a plan for evaluating the quality of the NPD data by comparing the values with reported values from MRTS, ARTS, and the Economic Census. During the analysis, the Census Bureau plans to identify issues with definitions and classifications. Preliminary comparisons suggest NPD data are of good quality. The data already have been used to validate reported survey values. For details of the data evaluation, see Hutchinson and Scheleur (2017).

### 4.3 Challenges

The main challenge was obtaining cooperation from companies. NPD tried different strategies, but in some cases it was not clear that the right people at the companies were involved. To help, the Census Bureau wrote a letter to the companies explaining how their participation in this research would benefit them and the Census Bureau. Successful informed consent data collection would reduce respondent burden and costs for both parties. Additionally, some companies had information technology concerns about allowing NPD to provide the Census Bureau with their data.

### 4.4 Future Work

If the results are promising for the three initial companies, then the Census Bureau would like NPD to reach out to additional companies to continue the study. Future work with NPD may also involve looking at the detailed product-level information from the NPD data feeds. This work would entail studying how well NPD's product data align with product data from the Economic Census and identifying issues with definitions, products collected, and the overall usefulness of the data. These data could help the Census Bureau produce new estimates for product lines.

## 5. System-to-System Data Collection

### 5.1 Background

To address decreasing response rates and respondent cooperation, a team was formed to begin discussions with companies on establishing an alternate system-to-system method of collection that would be suitable for multiple surveys. This would ease respondent burden, increase response, and streamline processes. With the ease of the current transfer of data (for example, sales, inventory, etc.) through the internet and computer systems, this method of collection appears technologically plausible.

## 5.2 Project

The Census Bureau selected companies to contact for this study based on discussions with retail trade subject matter experts. The selection criteria were similar to the ones for the NPD project in Section 4: company size, structure, public or private status, reporting history, and relationship with the Census Bureau. The size of the company was important because a small company might not have the resources to devote to such a project whereas a mid-size or large company most likely already had tools in place that would facilitate system-to-system collection. The structure of the company was critical since the team wanted to work with homogeneous companies, i.e. companies engaged primarily in one industry, instead of multi-industry companies. A homogeneous company's financial records would be less complicated and more focused on end-product tracking. With public companies, the team would be able to compare the transferred data with public SEC data. Lastly, the company's reporting history and relationship with the Census Bureau are important because a good reporting history and relationship would indicate a more cooperative and responsive company for a pilot study.

The team contacted a group of companies, and three agreed to be interviewed for the study. The team scheduled conference calls, prepared a draft protocol, and made initial contact. During the initial conference call, the team discussed the concept of the study, the protocol for the formal interview, and the company's willingness to participate in the study. The team requested and arranged a second "formal interview" to follow the protocol. During the formal interview, the team went through the protocol and discussed their accounting systems, their different modes of transferring data, obstacles the company might face with such a data transfer, and questions related to computer software and systems. The team also asked how likely it was to use a single source of data transfer with the Census Bureau, and all three companies felt this could be done.

The team is currently at different stages with the three companies. One company, which appeared to be promising during the conference call, did not grant the team a face-to-face interview and has since declined to participate in the study. Company visits were conducted with the other two companies. The team met with various staffs such as internal accounting, human resources, information technology, and payroll to discuss the proposed study and their systems. For one company, the meeting at their headquarters was promising, but it was clear after a few minutes that the people in the room would not be able to discuss fully all of the items necessary as far as timing and availability. The team was in the process of setting up further conference calls when it was informed that the contact was no longer with the company. The team will be contacting one of the other people who attended the meeting to see if the study can proceed. For the last company, the team conducted follow-up telephone conversations and developed a template for the data collection. However, during conversations on how the company would arrive at their data submission, it became apparent that they were not willing to release their data before internal reconciliation. The team concluded that its efforts would only increase burden on the company due to duplication of reporting.

### 5.3 Challenges

Companies today are involved in many industries, something that poses a key challenge when collecting data. Most companies do not have accounting systems that track their activities by industry. Rather, they track their activities by product. Another challenge is asking the right questions in order to develop a system that will work for each respondent as well as the Census Bureau. The team was hoping to meet with the right people and at the right level. It is a challenge to determine the organizational level for obtaining authorization and for obtaining the necessary information. Organizational structure varies from company to company, thus requiring customization. System-to-system data collection is an intensive individually tailored effort, and it may be better to have conversations with software makers to tailor software for data collection.

### 5.4 Future Work

Future work will involve further discussion on harmonizing the data to be collected and developing a standard data dictionary. This could include examining methods of collecting product data in a manner synchronized with the way businesses keep their records. Storage for housing the collected data will need to be established, and the collection will need to be designed so that multiple surveys can access the data. Accordingly, the team needs to address maintenance and security issues.

## 6. Autocoding and Machine Learning

### 6.1 Background

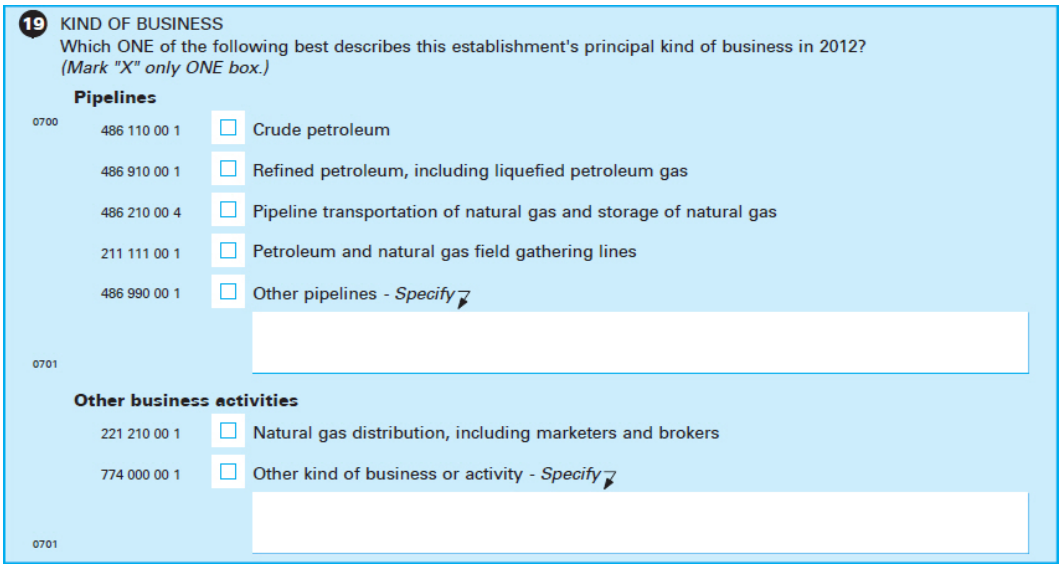
The Census Bureau classifies business establishments according to the North American Industry Classification System (NAICS). NAICS groups establishments into industries based on the activities in which they are primarily engaged and where revenue is generated[1]. The Census Bureau uses the NAICS classification for a variety of purposes such as stratifying establishments for sample selection and tailoring survey questionnaires to respondents. For more information about NAICS, see U.S. Census Bureau (2017b).

To assign NAICS codes to business establishments, the Census Bureau uses information from different sources such as the Economic Census, the Internal Revenue Service (IRS), and the Social Security Administration (SSA). Aspects of NAICS coding can be manually intensive. According to Snijkers *et al.* (2013, p. 478), manual coding has three key disadvantages: (1) it is expensive, (2) it is time-consuming, and (3) it can introduce systematic errors. Using machine learning to assign NAICS codes automatically can help address these disadvantages and make it easier to diagnose errors.

[1] A NAICS code is made up of six digits. The first two digits indicate the industry sector, and subsequent non-zero digits add industry detail. NAICS codes are updated every five years. As an example, the 2017 code 440000 refers to an establishment primarily engaged in the retail sector. The code 445000 indicates food and beverage stores, 445200 indicates specialty food stores, 445290 indicates other specialty food stores, and 445292 indicates confectionary and nut stores.

Kornbau (2016, sec. 2) and Kearney and Kornbau (2005) describe how Census Bureau staff, in collaboration with the IRS and the SSA, developed a NAICS autocoder for new businesses. The autocoder assigns a NAICS code to a new business using write-in text and other variables from the IRS's SS-4 form that businesses use to apply for an Employer Identification Number. The methodology uses dictionaries of one-word and two-word sequences from the SS-4 business name and description fields that occur frequently and that map a large percentage of the time to a particular NAICS code. A logistic regression model with dictionary frequencies as the main predictors is used to assign the NAICS code. In 2015, 79 percent of 3.6 million new business records were autocoded using this methodology, and about 69 percent of these coded records were classified to a complete 6-digit NAICS level (Kornbau, 2016, p. 3). Continual improvements and a robust quality control process have helped ensure quality autocoding over time.

A similar NAICS autocoding problem involves responses from the Economic Census. The Census Bureau sends forms to business establishments based on the most recent estimate of the establishment's NAICS code at the time of mail-out. The self-designated kind of business (SDKB) question asks respondents to describe their kind of business. This question contains a list of checkboxes, and the respondent is asked to mark one box. The respondent also has the option to write in a description. Figure 1 is a screenshot of the SDKB question from the 2012 Economic Census Pipelines form. For the 2012 Economic Census, there were hundreds of thousands of write-in cases. Clerks currently process and assign NAICS codes manually for these cases, so it would be helpful to develop a NAICS autocoder in this setting.


19 KIND OF BUSINESS
Which ONE of the following best describes this establishment's principal kind of business in 2012?
*(Mark "X" only ONE box.)*

**Pipelines**

0700
- 486 110 00 1 ☐ Crude petroleum
- 486 910 00 1 ☐ Refined petroleum, including liquefied petroleum gas
- 486 210 00 4 ☐ Pipeline transportation of natural gas and storage of natural gas
- 211 111 00 1 ☐ Petroleum and natural gas field gathering lines
- 486 990 00 1 ☐ Other pipelines - *Specify*

0701

**Other business activities**
- 221 210 00 1 ☐ Natural gas distribution, including marketers and brokers
- 774 000 00 1 ☐ Other kind of business or activity - *Specify*

0701

**Figure 1**. Self-designated kind of business question from the 2012 Economic Census Pipelines form (TW-48601). Respondents can write in their own description of their establishment's principal kind of business. Example write-ins for this form include "ammonia pipeline station" and "asphalt terminal."

### 6.2 Project

Researchers in the Economic Directorate have started a research project on using machine learning to assign a NAICS code to an SDKB write-in from the Economic Census based on the write-in text and other information from the Economic Census form such as company name and form number. The plan is to use the hundreds of thousands of SDKB write-ins from the 2002, 2007, and 2012 Economic Census as a training set to build and evaluate classification models.

The proposed modeling approach borrows elements from the new business NAICS autocoder (Kornbau, 2016, sec. 2) and the classification models used in the application of SABLE to state government websites to identify useful PDFs (Dumbacher and Capps, 2016). Many write-ins consist of text such as "not applicable" or "none" that do not provide any useful information. These write-ins will be removed from the training set prior to model building. The write-in text will be normalized by removing common words, punctuation, and extraneous whitespace. Features will be created based on one-word and two-word sequences appearing in the write-in text and business name. Researchers plan to consider models besides logistic regression such as support vector machines and decision trees. As part of model evaluation, the researchers would like to see how well the models perform as the level of detail of the prediction increases from 2-digit NAICS to 6-digit NAICS.

### 6.3 Challenges

One challenge involves how to use the best estimate of the NAICS code at the time the Economic Census forms are mailed. This estimate is known as the mailed NAICS. One reason respondents may be writing in a description is that the mailed NAICS is inaccurate and the respondent does not receive the appropriate form, and hence does not see the appropriate checklist. However, at the same time, the mailed NAICS does have some predictive power. As a compromise, the 2-digit mailed NAICS corresponding to industry sector could be used as a model feature.

### 6.4 Future Work

Future work on the methodology could involve investigating feature dimensionality methods such as stemming. Stemming is the process of identifying word roots and removing suffixes and prefixes. For example, the two words "manufacturing" and "manufactured" can be stemmed to the common root "manufacture." Using only these roots results in a smaller set of features. An autocoder for the SDKB write-ins will not be production-ready for the 2017 Economic Census. Instead, researchers plan on using write-in data received from the 2017 Economic Census to test the methodology.

## 7. Summary

The U.S. Census Bureau is undertaking a variety of projects in support of its vision of using alternative data collection methods and data sources to improve its economic surveys. For many respondents, data of adequate quality are available online on their

websites and through APIs, for example. Researchers are studying the feasibility of scraping public sector data and building permit data from these types of web sources. Passive data collection such as web scraping has the potential to reduce burden and costs significantly. Likewise, an informed consent data collection pilot project is underway with NPD and is aimed at making it easier for companies to respond to the Census Bureau's retail trade surveys, namely MRTS, ARTS, and the retail component of the Economic Census. Sales data from NPD also offer the opportunity to add e-commerce and product-level detail to the Census Bureau's data products.

Another pilot project underway is studying the feasibility of system-to-system collection from large companies. This type of collection involves transfers of large data files from the companies' computer systems to the Census Bureau's system. It would also allow companies to provide information to multiple surveys at the same time using a single portal.

Many aspects of data collection and processing are manually intensive, and machine learning can help automate certain tasks such as coding. Using classification models to assign NAICS codes and tax codes, for example, has shown very positive results. For the projects that have elements involving machine learning, an important but time-consuming task is creating a large, representative, and good-quality training set to build and evaluate models. When working with unstructured text as a source of model features, it is also important to think about, given the application, how best to normalize the text.

## 8. Acknowledgments

The authors would like to thank Carma Hogue, Diane Willimack, Justin Nguyen, Rebecca Hutchinson, Angela Delano, and Michael Kornbau of the U.S. Census Bureau for their helpful comments and insight.

## References

The Apache Software Foundation. (2014). Apache Nutch. <http://nutch.apache.org>. Accessed April 27, 2017.

Dumbacher, B. and Capps, C. (2016). Big Data Methods for Scraping Government Tax Revenue from the Web. *2016 Proceedings of the American Statistical Association, Section on Statistical Learning and Data Science*. Alexandria, VA: American Statistical Association, 2940–2954.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition). Berlin, Germany: Springer.

Hutchinson, R. and Scheleur, S. (2017). Using Big Data to Enhance US Census Bureau Economic Data Products. *2017 Proceedings of the American Statistical Association, Business and Economic Statistics Section*. Alexandria, VA: American Statistical Association.

Kearney, A.T. and Kornbau, M.E. (2005). An Automated Industry Coding Application for New U.S. Business Establishments. *2005 Proceedings of the American Statistical Association, Business and Economic Statistics Section*. Alexandria, VA: American Statistical Association, 867–874.

Kornbau, M.E. (2016). Automating Processes for the U.S. Census Business Register. *25th Meeting of the Wiesbaden Group on Business Registers*.

Kreuter, F. and Peng, R.D. (2014). Extracting Information from Big Data: Issues of Measurement, Inference and Linkage. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Eds. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, New York, NY: Cambridge University Press, 257–275.

Mitchell, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA: O'Reilly Media, Inc.

Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D.K. (2013). *Designing and Conducting Business Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Tan, P.N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. New York, NY: Pearson.

U.S. Census Bureau. (2017a). Federal, State, & Local Governments. <https://www.census.gov/govs/classification/>. Accessed April 28, 2017.

U.S. Census Bureau. (2017b). North American Industry Classification System. <https://www.census.gov/eos/www/naics/>. Accessed April 28, 2017.