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

Brian Dumbacher

Demetria Hanna

U.S. Census Bureau

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Outline

- Data Collection Vision
- Research Projects
  - Public Sector Web Scraping
  - Building Permit Web Scraping
  - Informed Consent Data Collection Via The NPD Group
  - System-to-System Data Collection
  - Autocoding and Machine Learning
- Summary

Challenges in Producing Official Economic Statistics

- The U.S. Census Bureau faces many challenges
  - Data users are demanding data that are more timely and granular
  - The Census Bureau faces fiscal pressures
  - The economic landscape is constantly changing
  - Respondent cooperation is declining
- Related to the challenge of declining response rates are:
  - Costs of current data collection methods
  - Aspects of data processing that are manually intensive

Data Collection Vision

Maximize the use of alternative data collection methods, sources, and techniques to increase respondent cooperation, reduce burden, save costs, and enhance the efficiency of data collection operations while maintaining the quality of data products

- Passive data collection
  - The respondent either has little awareness of the data collection effort or does not need to take any explicit actions
  - Examples include web scraping and informed consent data collection

Data Collection Vision (cont.)

- System-to-system data collection
  - Respondents transfer data directly from their computer systems to the Census Bureau's systems
  - Data are used for multiple surveys
- Big Data
  - Point-of-sale scanner data
  - Data dumps from private companies
- Machine learning
  - Classification
  - Autocoding

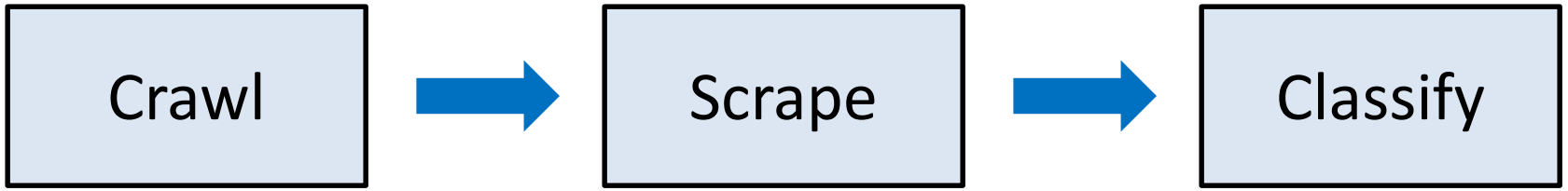
Project 1: Public Sector Web Scraping

- For many public sector surveys, respondent data are available online
- Respondents sometimes direct Census Bureau analysts to their websites to obtain the data
- Data are often in Portable Document Format
- Automate the process of finding, scraping, and organizing data from government websites
- Focus on Quarterly Summary of State and Local Government Tax Revenue (QTax)

SABLE

- Scraping Assisted by Learning (SABLE)
- Collection of tools for
  - Crawling websites
  - Scraping documents and data
  - Classifying data
- Models based on text analysis and machine learning methods
- Implemented using free, open-source software
  - Apache Nutch
  - Python

Three Main Tasks



Given a website,

- Crawl website
- Find documents (in PDF format)
- Apply model to predict whether document contains useful data

Given a document classified as useful,

- Apply model to learn the location of useful data
- Extract numerical values and contextual information

Given scraped data,

- Put scraped data in a normalized data structure
- Apply model to map terminology to the Census Bureau's tax classification codes

State Of New Hampshire
Monthly Revenue Focus

Department of Administrative Services

Vicki V. Quiram, Commissioner

Monthly Revenue Summary

(for month)

	FY 17 Actual	FY 17 Plan	Actual vs. Plan
Gen & Educ	$645.4	$645.8	$(0.4)
Highway	$17.9	$17.2	$0.7
Fish & Game	$0.6	$0.7	$(0.1)

Current Month

GENERAL & EDUCATION FUNDS	*FY 17 Actuals*	*FY 17 Plan*	*Actual vs. Plan*
Business Profits Tax	$52.2	$60.4	$(8.2)
Business Enterprise Tax	34.9	38.7	(3.8)
Subtotal Business Taxes	87.1	99.1	(12.0)
Meals & Rentals Tax	21.5	23.0	(1.5)
Tobacco Tax	18.7	17.4	1.3
Transfer from Liquor Commission	8.5	8.7	(0.2)
Interest & Dividends Tax	4.3	3.9	0.4
Insurance Tax	102.8	95.0	7.8
Communications Tax	4.2	4.9	(0.7)
Real Estate Transfer Tax	7.4	6.6	0.8
Court Fines & Fees	1.3	1.2	0.1
Securities Revenue	6.9	8.6	(1.7)
Utility Consumption Tax	0.5	0.5	-
Beer Tax	0.8	0.9	(0.1)
Other	5.5	6.2	(0.7)

Analysis

Unrestricted revenue for the General and Education Funds received during March totaled $645.4 million, which was below the plan by $0.4 million (0.1%) and below the prior year by $22.0 million (3.3%). YTD unrestricted revenue totaled $1,756.4 million, which was above plan by $29.2 million (1.7%) but below prior year by $31.9 million (1.8%).

Business Taxes for March totaled $87.1 million, which were $12.0 million (12.1%) below plan and $24.9 million (22.2%) below prior year. YTD business tax collections are above plan by $23.0 million (6.0%) but $19.1 million (4.5%) below the prior year. While comparing YTD collections to the prior year, it should be noted that fiscal 2016 included $19.0 in tax amnesty receipts, which do not repeat in fiscal 2017. The tax amnesty program, as set forth in Chapter 276:242, Laws of 2015, was in place from December 1, 2015 through February 15, 2016. According to the Dept. of Revenue Administration (DRA), the decrease in monthly revenue for March 2017 was largely due to the timing of payments as a result of 2016 tax law changes which altered the due dates for certain types of returns.

Meals and Rentals Tax (M&R) receipts for March came in below plan by $1.5 million (6.5%) and below prior year by $0.8 million (3.6%). YTD collections were $4.9 million (2.1%) above plan and $10.5 million (4.6%) above prior year. According to DRA, March collections (February activity) from hotels were down 10% while full service restaurants were down 3% as compared to the same month last year.

Tobacco Tax receipts for the month were $18.7 million, or $1.3 million (7.5%) above plan and $1.2 million (6.9%) above prior year. YTD collections were $4.2 million (2.5%) below plan and $5.1 million (3.1%) below the same YTD period last year. According to the DRA, stamp sales were up 4% in March as compared to the same month of the prior year.

Interest and Dividends Tax (I&D) collections for the month were reported at $4.3 million, which were $0.4 million (10.3%) above plan and $0.8 million (22.9%) above prior year. YTD collections through March were $42.2 million, or $3.0 million (6.6%) below plan, but $0.4 million (1.0%) above the prior year. DRA has reported that the slight increases in March interest and dividend collections compared to prior year were attributable to all categories of payments and largely the result of the timing of payments.

- Meals & Rentals Tax
- Tobacco Tax
- Transfer from Liquor Commission
- Interest & Dividends Tax
- Insurance Tax
- Communications Tax
- Real Estate Transfer Tax
- Court Fines & Fees
- Securities Revenue
- Utility Consumption Tax
- Beer Tax
- Other
- Transfer from Lottery Commission
- Transfer from Racing & Charitable Gaming
- Tobacco Settlement
- Utility Property Tax
- State Property Tax

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Source: New Hampshire Department of Administrative Services.
https://das.nh.gov/accounting/FY%2017/Monthly%20Rev%20March.pdf

Potential Data Product

- Monthly version of QTax based on a panel of state governments that produce monthly reports such as the New Hampshire example
- Possible approach
  - Use SABLE crawler, search engines, and tax policy resources to find monthly reports
  - Apply hard-coded template to scrape data from monthly reports
  - Apply model to map definitions in monthly reports to Census Bureau tax classification codes

Project 2: Building Permit Web Scraping

- Data on new construction
  - Used to measure and evaluate size, composition, and change in the construction sector
  - Building Permit Survey (BPS)
  - Survey of Construction (SOC)
  - Nonresidential Coverage Evaluation (NCE)
- Information on new, privately owned construction is available for some building jurisdictions
- Investigate feasibility of using publicly available building permit data to supplement new construction surveys

Research and Findings

- Chicago and Seattle building permit jurisdictions
  - Data available through Application Programming Interfaces (APIs)
  - Initial research indicated that these sources provide timely and valid data with respect to BPS
  - Additional research uncovered definitional differences
  - Data may not provide enough detail to aid estimation
- Other jurisdictions
  - Data come in other formats such as reports and Excel files
  - Nashville and Boston jurisdictions were recently included in the research

Challenges and Future Work

- Challenges of using online building permit data
  - Representativeness
  - Consistency of data formats
- Future work
  - Use text analysis and machine learning to deal with differences in terminology
  - Continue validation and compare data to survey data from BPS, SOC, and NCE

Project 3: Informed Consent Data Collection Via The NPD Group

- The NPD Group
  - Collects point-of-sale scanner data from thousands of retail establishments
  - Receives and processes data feeds containing aggregated scanner transactions by product
  - Edits, analyzes, and summarizes data at detailed product levels and creates market analysis reports for its retail partners
- Investigate feasibility of using these data to supplement or replace survey data from the Census Bureau's retail surveys

Pilot Project

- Census Bureau purchased data from three companies with the companies' consent
- Data consist of sales aggregates broken down by month, industry, channel, and establishment
- Companies contacted for this study based on
  - Size
  - Geographic distribution
  - Reporting history to the Monthly Retail Trade Survey, Annual Retail Trade Survey, and Economic Census

Evaluation and Challenges

- Evaluation of data
  - Identify issues with definitions and classifications
  - Comparisons suggest NPD data are of good quality
- Challenges of informed consent data collection
  - Obtaining cooperation from companies
  - Explaining how informed consent data collection is mutually beneficial to companies and the Census Bureau

Project 4: System-to-System Data Collection

- Team was formed to investigate feasibility of system-to-system collection that would be suitable for multiple surveys
- Companies contacted for this study based on
  - Size
  - Structure
  - Public or private status
  - Reporting history

Contact with Companies

- Three companies agreed to participate
- Initial conference call
  - Discuss concept of system-to-system data collection
- Formal interview
  - Discuss accounting systems and computer software
  - Potential obstacles with transfers of large data files
- Company visits
  - Meetings with accounting and human resources staff
  - Further discussions on accounting systems

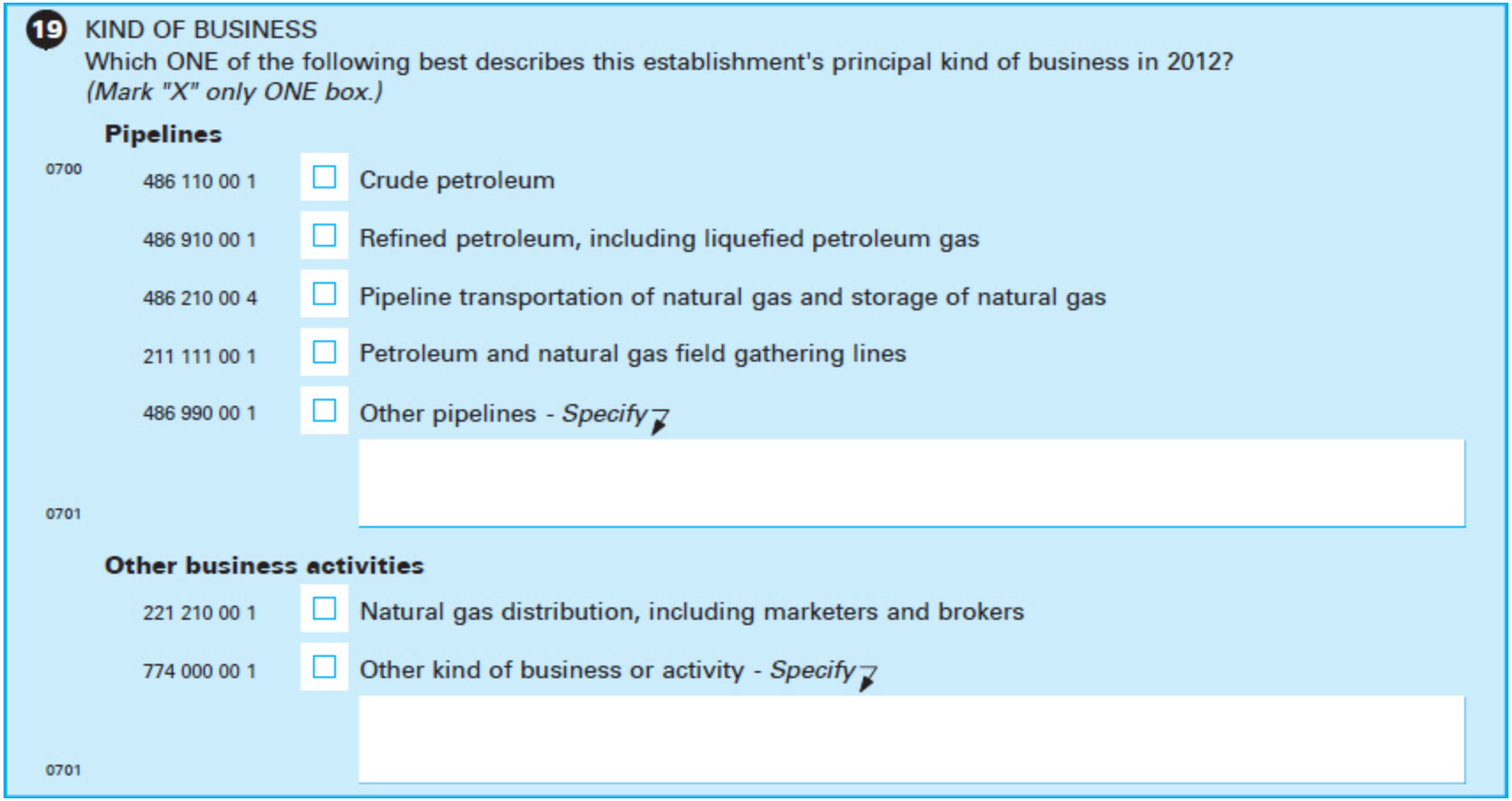
Challenges

- Accounting systems may not track activities by industry
- Asking the right questions to develop a system that will work for each respondent as well as the Census Bureau
- System-to-system data collection is an intensive individually tailored effort
- Designing a collection instrument that will work with multiple systems
- Harmonizing terminology so common terms and concepts are used

Project 5: Autocoding and Machine Learning

- The Census Bureau classifies business establishments according to the North American Industry Classification System (NAICS)
- Information for classification comes from:
  - Economic Census
  - Internal Revenue Service
  - Social Security Administration
- Disadvantages of assigning NAICS codes manually
  - Expensive
  - Time-consuming
  - Introduce systematic errors

Self-Designated Kind of Business (SDKB) Question



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

Source: U.S. Census Bureau. https://www2.census.gov/programs-surveys/economic-census/2012/questionnaires/forms/tw48601.pdf

Economic Census
Write-in NAICS Autocoder

- Use machine learning to assign a NAICS code to an SDKB write-in based on the text and other information from the Economic Census form
- Over 1.5 million write-ins from 2002, 2007, and 2012 Economic Census make up the training set
- Modeling approach
  - Remove throw-away write-ins such as “None” or “NA”
  - Remove stop words, punctuation, and whitespace
  - Create features based on occurrence of word sequences

Example Write-in

Write-in Text: Paintball Field, Supplies, & Games

Standardized Text: paintball field supplies games

1-Word Sequences: "paintball", "field", **"supplies"**, **"games"**

2-Word Sequences: **"paintball field"**, "field supplies", "supplies games"

Associations with certain NAICS codes

Summary

- For many respondents, equivalent quality data are available online or from third parties
  - Web scraping and informed consent data collection show promise and can reduce burden and costs
- System-to-system collection would allow companies to provide information to multiple surveys
  - Data harmonization is a key challenge
- Many aspects of data collection and processing are manually intensive
  - Machine learning can help automate tasks such as assigning classification codes

Contact Information

- Brian.Dumbacher@census.gov
- Demetria.V.Hanna@census.gov