Census Scientific Advisory Committee

Economic Directorate's Big Data and Passive Data Collection Overview

September 15, 2016

Modernizing Economic Statistics
Major Activities

Activity
Conduct research on big data sources, methods, procedures
Implement new data sources into retail data products
Re-engineer the Business Register infrastructure to absorb and process alternative source data Harmonize item and unit content to streamline survey collections and enable passive collection

Economic Directorate Big Data Projects
Currently Active Projects

- Explore Third-party passive collection, system to system collection, and accounting/payroll software collaborations
- Continue development of a web crawler/scraper
- Purchase more third-party data
- Wrap up payment processing company pilot projects
- Research machine learning to improve imputations

Passive Collection

System-to-system Collection

- Plan to explore the possibility of having a company respond once for multiple Economic surveys collecting the same data
  - Requires harmonization of our definitions and terms across our intercensal surveys
  - Requires exploration of how products would be collected for the Economic Census

- Contacted three companies who are interested in participating in this research
  - Held initial teleconferences
  - Personal visits in August/September

Passive Collection
Third-party Collection

- Signed a contract with NPD Group
- Plan to compare individual company data supplied to NPD to data reported to Census
  - Informed consent from up to 5 companies
    - Regional and national companies
    - Different geographies
  - General approach of helping Census research ways of reducing respondent burden

Passive Collection

Web-crawling/scraping Data Collection

Developed a web crawler with machine learning that learns the location of relevant data on public websites, scrapes unstructured data, and organizes it

- Started with scraping PDFs of Comprehensive Annual Financial Reports (CAFR) to obtain tax data from state and local governments
- Used Apache Nutch and Python to build a combination web crawler/classifier that finds PDFs containing relevant data
- Presented at the August Joint Statistical Meetings and through a Research Matters blog on www.census.gov
- Obtained a server in the CAT lab to build a prototype to expand the crawler's usage

Passive Collection
Web-scraping Data Collection

- Researching the availability of online building permit information for jurisdictions in the Survey of Construction sample
  - Focus on jurisdictions with large 2015 residential construction
  - Looking to complete the top 30% of the sample to estimate coverage
- Incorporating cost data to investigate cost savings
- Results thus far
  - New data sources are available, mainly for large jurisdictions
  - Numerous formats: APIs, reports, database queries
  - Inconsistencies in classification of single-family and multi-family and square footage
  - High confidence in timeliness and validity of data
  - Low confidence in coverage of building characteristic details

Passive Collection
Software Partnership Collection

- This effort is in its infancy.

- American Payroll Association (APA) representative is interested in helping us pursue this work.
  - Contacts with the major financial and payroll software companies
  - Retailers of all sizes use financial and payroll software

Economic Directorate Big Data Projects
Purchase of Retail Third Party Data

- Request for Information (RFI) in Spring to assess what data and/or tools are available
- Request for Proposal (RFP) soon
- Purchase in November
  - Will research the quality of the data using the established procedures
  - Will determine what to purchase

Economic Directorate Big Data Projects
Retail Third Party Data

What are we doing?

- Building Small Area Estimation models
  - To improve estimates at the nation-by-industry level
  - To estimate totals at the state-by-industry level
  - Covariates are aggregates of sales from credit card transaction data

- Examining Trading Day weight calculations and holiday adjustments
  - Using daily data seasonal models
  - Comparing daily data modeled from credit card output to current X-13 generated trading day weights and looking for areas for improvement

Economic Directorate Big Data Projects
Retail Third Party Data

- Collaborative short-term exploratory project involving a software company, a payment processing company, BEA, and Census
- Software tool updates with retail transactions within a day
  - Can use R and Python with this tool
  - Can research trading day adjustments and small area estimation models
- Consumer spending data
  - Cover 58 billion transactions annually
  - Captures about 45% of all point-of-sale transactions
  - Captures credit, debit, and prepaid gift card transactions but not cash transactions
  - Cover five states for this pilot project

Economic Directorate Big Data Projects
Small Area Estimation Models

Model state-by-industry estimates of sales totals directly using Fay-Herriot models. Evaluate model fit by examining reduction in variance of estimates

Findings:

- Limitations regarding input from Monthly Retail Trade Survey (MRTS) at the state level
- Consider a nation-level model instead and make it resemble the state-level model by inflating variance inputs
- This proof of concept shows the usefulness of First Data aggregates as covariates

Economic Directorate Big Data Projects

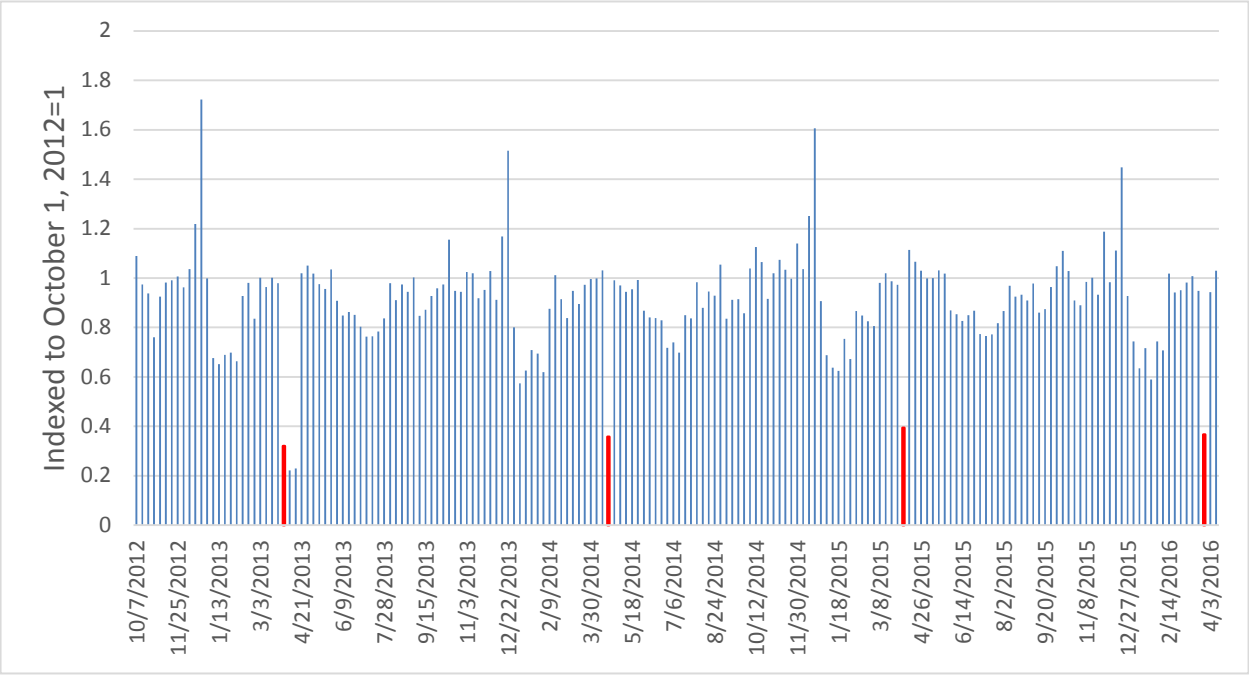
Trading Day Weight Calculations and Holiday Adjustments

- Using daily data seasonal models developed by Tucker McElroy and Brian Monsell at the Census Bureau
- Comparing daily data modeled from credit card output to current X-13 generated trading day weights and looking for areas of improvement
- Modeling holiday effects for Super Bowl Sunday, Chinese New Year, Easter Sunday, Ramadan, Labor Day, and Cyber Monday

Economic Directorate Big Data Projects

Example of a Holiday Effect Revealed by Daily Data

- How do Easter Sundays for Shoe Stores (NAICS Code 44820) compare to all other Sundays in the series?



Machine Learning

Advanced Monthly Retail Trade Survey (MARTS) Imputation

Estimates for MARTS can be affected greatly by missing data for certainties/large units. The current manual imputations are based on historical trends.

- Research if machine learning methods can take advantage of a large number of variables and observations from MARTS and the Monthly Retail Trade Survey (MRTS)
- Consider regression trees and neural networks
- Received a server dedicated to research

Modernizing Economic Statistics
Outreach Activities

- Conferences and Seminars
  - European Conference on Quality in Official Statistics 2016
  - 2016 Joint Statistical Meetings
  - Joint Census and BEA internal methodology seminar on Big Data
  - 2017 Joint Statistical Meetings Invited Session proposal submission
  - E-Poster submissions to the 2017 Conference on Statistical Practice
- Publications
  - Proceedings of 2016 Joint Statistical Meetings
  - Proceedings of the 2016 European Conference on Quality in Official Statistics
  - Spanish Statistical Journal

Modernizing Economic Statistics
Outreach Activities (continued)

- Consultations/Collaborations
  - Big Data Video Conference with Eurostat
  - Frauke Kreuter
  - Roberto Rigobon (Full day at Summer at Census)
  - Susie Fortier, Chief of Seasonal Adjustment at Statistics Canada

Economic Directorate Big Data Projects
Next Steps

- Retail Trade Data
  - Award contracts to continue to research third party data sources
  - Continue development of more detailed data products
- Passive Data Collection
  - Continue our efforts to harmonize our questionnaires and work towards system-to-system collection
  - Compare the data from NPD companies to the data from Census
- Web crawler
  - Collaborate with others to expand the uses for the crawler beyond public sector
  - Begin phase 2 to crosswalk the scraped data to the questionnaire
- Presentations and publications of our results in various external venues.