

Census Scientific Advisory Council

# Economic Directorate's Big Data Overview

April 14, 2016

# Modernizing Economic Statistics

## FY17 Budget Initiative

- Customers need more frequent data with more detail and greater geographic granularity
- A major transformational strategy is to fully leverage administrative records and other external data sources, including “big data,” to supplement and possibly supplant direct data collection, support new data products, and expand existing data products
- Intended outcome is more timely, more detailed economic statistics through the incorporation of alternative data sources

# Modernizing Economic Statistics

## Major Activities

### Activity

Conduct research on big data sources, methods, procedures

Implementation of new data sources into retail data products

Business Register infrastructure re-engineering to absorb and process alternative source data

Data harmonization and coherence – Item and unit harmonization to streamline survey collections

# Economic Directorate Big Data Projects

## Currently Active Projects

- Use of third-party data to add detail to Retail Trade survey data
  - NPD
  - Credit card data
  - Software tool
- Development of a Public Sector Statistics web scraper
- Text mining of respondent comments
- APIs for yielding Building Permit data

# Economic Directorate Big Data Projects

## Retail Third Party Data

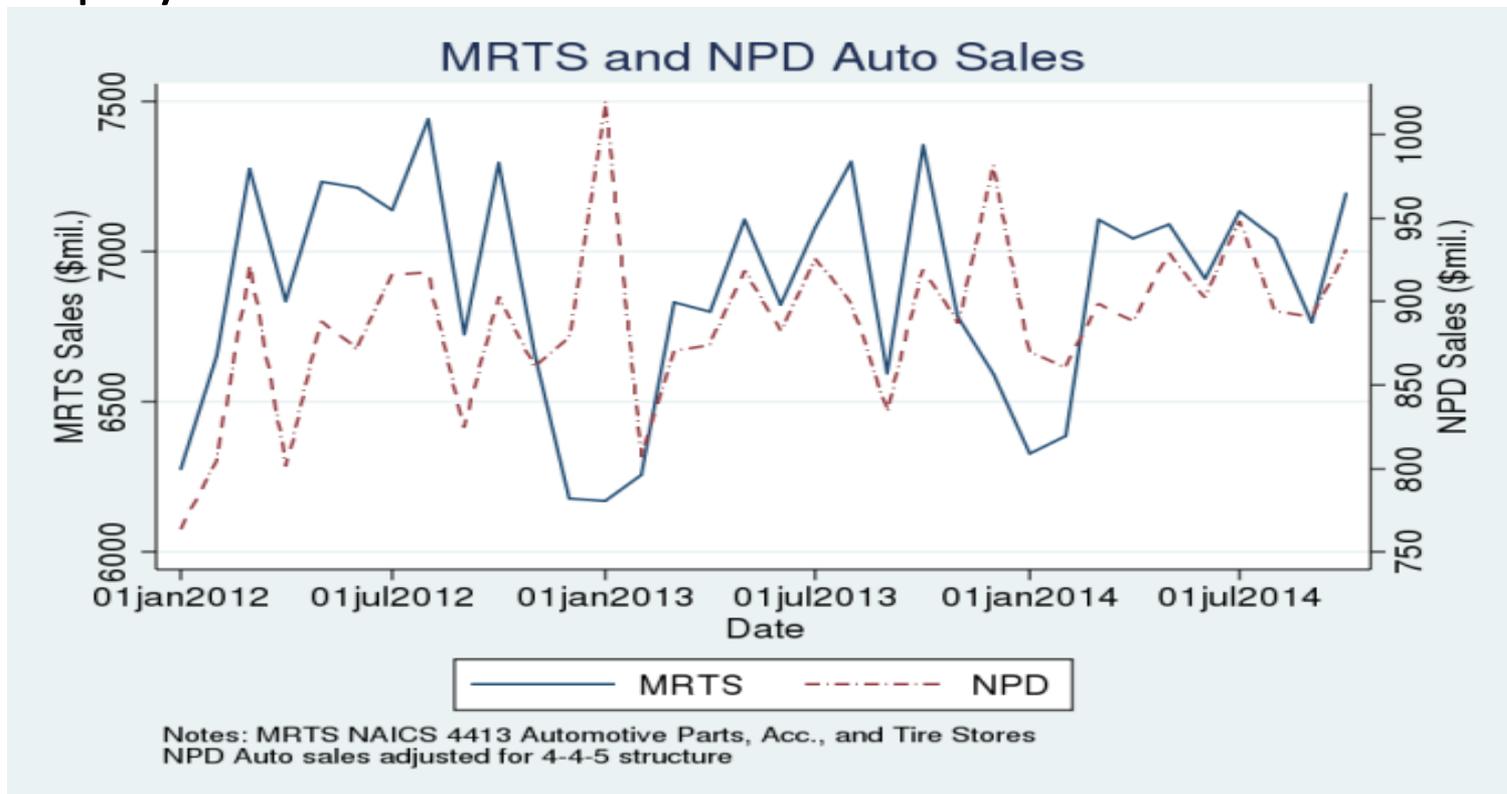
### NPD

- Purchased a dataset of point-of-sales transactions from January 2012 through December 2014
- Explored two datasets
  - Auto Parts
  - Jewelry and Watch data
- Obtained geographic detail at the Nielsen Designated Market Area (DMA) level

# Economic Directorate Big Data Projects

## Retail Third Party Data

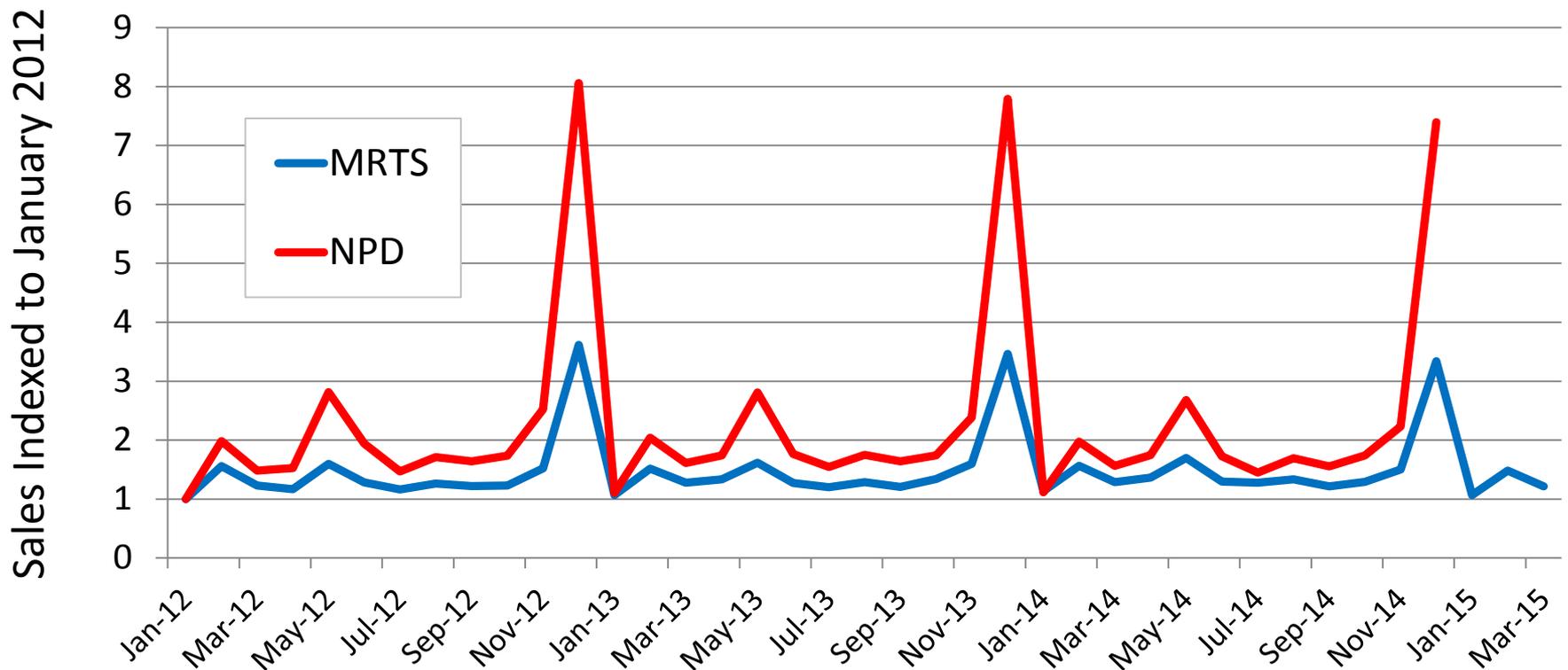
NPD Auto Parts and Monthly Retail Trade Survey did not display similar trends.



# Economic Directorate Big Data Projects

## Retail Third Party Data

NPD Jewelry & Watches and Monthly Retail Trade Survey did display similar trends, but not levels.



# Economic Directorate Big Data Projects

## Retail Third Party Data

### Recommendations for the use of NPD data:

- For geographic detail, Census needs data based on zip code, not Designated Market Area.
- Data sets need to be standardized to the same level of geography and detail.
- Having product line data that aligns with Census product lines would be more useful than what was obtained.

# Economic Directorate Big Data Projects

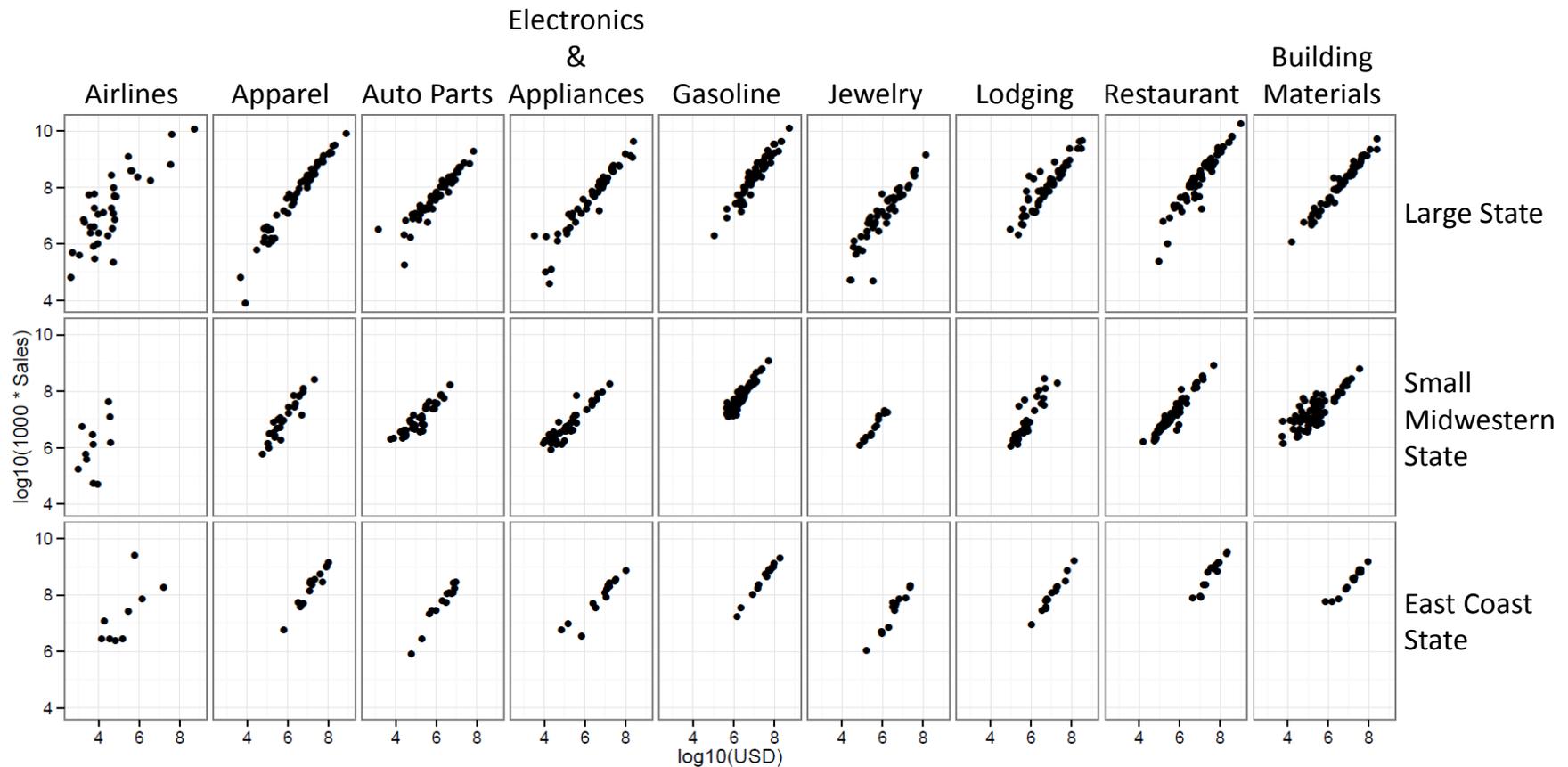
## Retail Third Party Data

### Credit Card Data

- Joint project with BEA
- Data provided for a limited six-month project
- Dataset included seven variables for 75 months (January 2009 through March 2015) of county level industry aggregates for three states.
  - Total Retail excluding auto and gas
  - Total Retail excluding auto, gas, and building materials
  - Airlines
  - Apparel
  - Auto Parts
  - E-Commerce
  - Electronics and Appliances
  - Jewelry
  - Lodging

# Economic Directorate Big Data Projects

## Retail Third Party Data



Log10 dollars to shrink scale; only counties that appear in both Credit Card and Economic Census and are not 0 are included (due to logging)

# Economic Directorate Big Data Projects

## Retail Third Party Data

- Collaborative short-term six-month 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
  - Researching trading day adjustments and small area estimation models
- Consumer spending data
  - Cover 58 billion transactions annually
  - Captures over 50% 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

## Retail Third Party Data

What are we doing?

- Analyzing transaction data
- Building Small Area Estimation models
  - Fay-Herriot models to improve estimates at the nation-by-industry level
  - Fay-Herriot models for estimating 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
  - Looking at daily data seasonal models developed by Tucker McElroy and Brian Monsell in Center for Statistical Research and Methodology
  - Comparing daily data modeled from credit card output to current X-13 generated trading day weights and looking for areas of improvement
- Making recommendations for future use of data

# Economic Directorate Big Data Projects

## Public Sector Statistics Web Scraper

Data collected by public sector statistics surveys are publicly available on government web sites

- Goal: Build web scraper that crawls government web sites and scrapes relevant data
- Benefits: Potential to reduce respondent burden and cost and increase timeliness of data products
- Challenges:
  - No standardization in structure of web sites, documents, and data products
  - Relevant documents such as tables of tax revenue collections and Comprehensive Annual Financial Reports (CAFRs) are PDFs, a format not readily amenable to analysis

# Economic Directorate Big Data Projects

## Public Sector Statistics Web Scraper

### Using machine learning methods

- Document classification
  - Build models for predicting whether a document has relevant data based on word combinations and frequencies
  - Apply models to new PDFs that the web crawler comes across
- Data extraction
  - Extract relevant unstructured data from document
  - Store data in a normalized data structure
- Software
  - R, Python, and Nutch

# Economic Directorate Big Data Projects

## APIs for Yielding Building Permit Data

### Survey of Construction (SOC) background

- Each month new construction sampled via
  - Monthly sample of building permits from predefined permit-issuing jurisdictions
  - Canvassing operation for new construction in predefined non permit-issuing jurisdictions
- Once sampled, survey returns to same housing unit approximately every month until completion and/or sale

# Economic Directorate Big Data Projects

## APIs for Yielding Building Permit Data

Examined publicly available building permit APIs

- Seattle, WA
- Chicago, IL

In summary thus far

- New building permit data sources are available
- Mainly for larger permit-issuing jurisdictions
- Different definitions to contend with
- High confidence in timely, valid data
- Low confidence in permit details

# Economic Directorate Big Data Projects

## Next Steps

- Retail Trade Data
  - Continue to research third party data sources
  - Complete our work with the software tool and write the summary document
  - Complete the final presentations on the credit card project
- Passive Data Collection
  - Searching for opportunities to receive data feeds directly from large companies that we can use across all of our surveys
  - Possibly partner with a third-party source to determine if their respondents will agree to share their data with Census
- Web crawler
  - Will present our results at the Joint Statistical Meetings
  - Will look to expand the uses for the crawler beyond public sector
- We are presenting our research at the European Conference on Quality in Official Statistics in Madrid.

# Questions for the Committee

- In building our Fay-Herriot models, we are challenged with not having direct Monthly Retail Trade Survey estimates at a state-by-industry level. We are considering alternative inputs with known limitations; the variances associated with the alternative inputs are too small and the models are not offering as much benefit as they should. Does the Committee have any suggestions?
- Does the Committee have any recommendations on additional data sources for us to examine, particularly for our Retail Trade research?
- Does the Committee have any recommendations on the direction that we should follow in looking to reduce respondent burden – passive collection or other suggestions?
- Are there concerns with how we should blend official statistics and third-party data?