Enhancing the Foundation of Official Economic Statistics with Big Data

Special Session #34

Big Data and Official Statistics:

Challenges and Opportunities

Brian Dumbacher
Economic Directorate, U.S. Census Bureau, Washington DC

Rebecca Hutchinson
Economic Directorate, U.S. Census Bureau, Washington DC

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

Madrid, May 31 - June 3

Big Data Vision
## for the U.S. Census Bureau's Economic Directorate

Key goal is to leverage Big Data sources in conjunction with existing survey data to:

- Provide more timely data products

- Offer greater insight into the nation's economy through detailed geographic and industry-level estimates

- Improve efficiency and quality throughout the survey life cycle
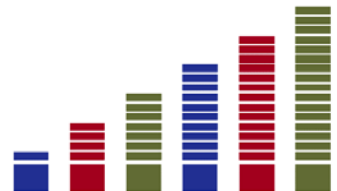
Big Data Concerns

- Methodological transparency
- Consistency of the data
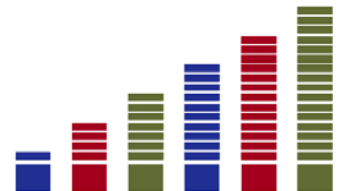- Information technology security
- Confidentiality
- General data quality
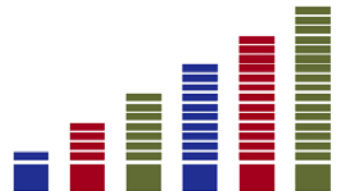
Use of third-party data to add detail to retail trade survey data

- Point of sale data: NPD
- Payment processor transaction data: Palantir/First Data

Application Programming Interfaces (APIs) for obtaining building permit data
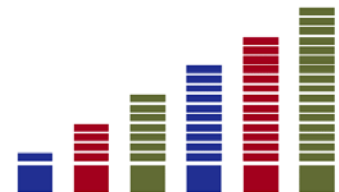
NPD collects point-of-sale scanner data from thousands of stores

Purchased point-of-sale transaction data from January 2012 through December 2014

Explored two datasets

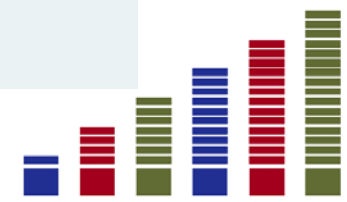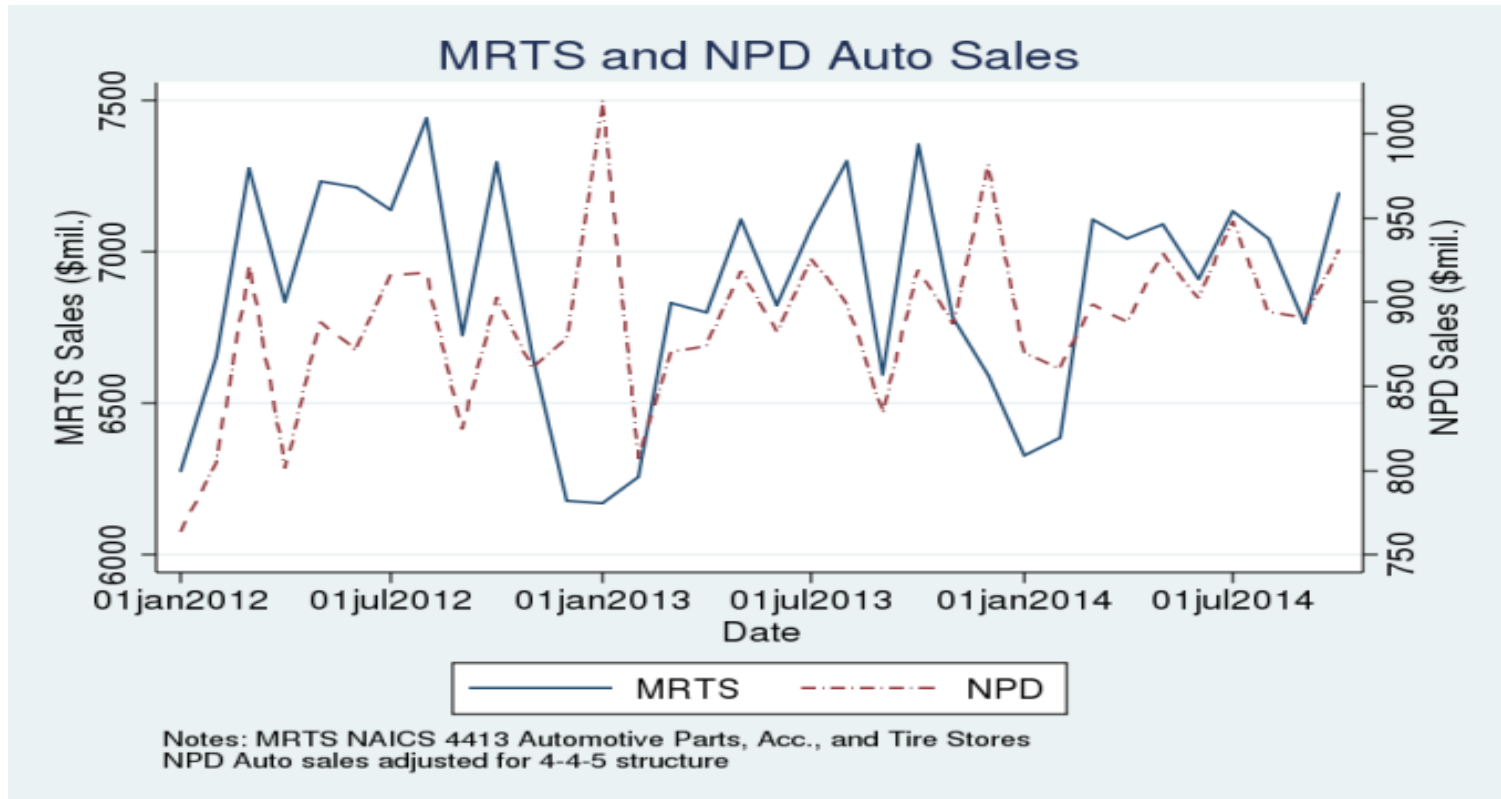- Auto parts
- Jewelry and watches

Obtained geographic detail at a Designated Market Area (DMA) level

NPD Auto Parts and Monthly Retail Trade Survey (MRTS) did not display similar trends

Notes: MRTS NAICS 4413 Automotive Parts, Acc., and Tire Stores
NPD Auto sales adjusted for 4-4-5 structure

NPD Jewelry and Watches and Monthly Retail Trade
Survey did display similar trends, but not levels
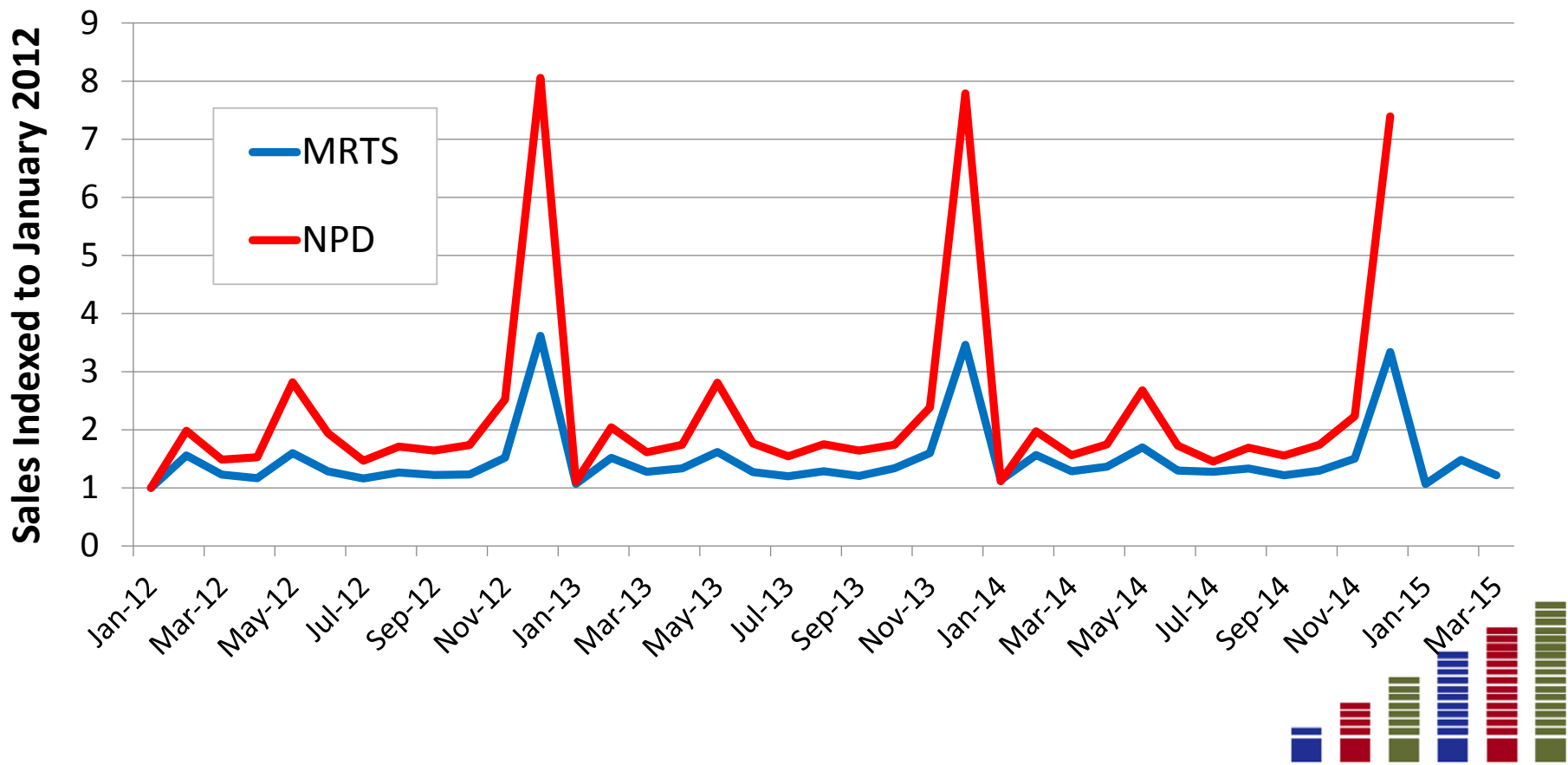
Recommendations for using NPD data:

- For geographic detail, Census Bureau needs data based on zip code, not Designated Market Area

- Data need to be standardized to the same level of geography and detail

- Product line data that align with Census Bureau product lines would be more useful than what was obtained
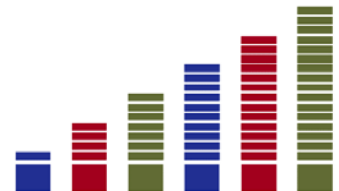
Collaborative short-term exploratory project involving Palantir, First Data, Bureau of Economic Analysis, and Census Bureau

Palantir's software tool

- Updated daily with retail transactions
- Custom dashboards
- Environment for using R and Python

First Data's consumer spending data

- Cover 58 billion transactions annually
- Capture about 45% of all point-of-sale transactions
- Capture credit, debit, and prepaid gift card transactions but not cash
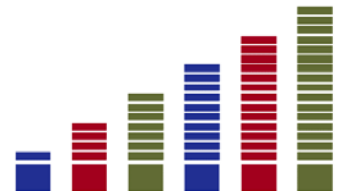- Cover five states for this pilot project
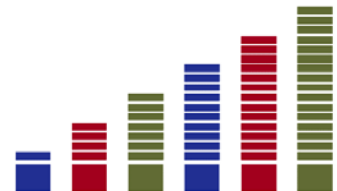
Current research projects

- Building Small Area Estimation models

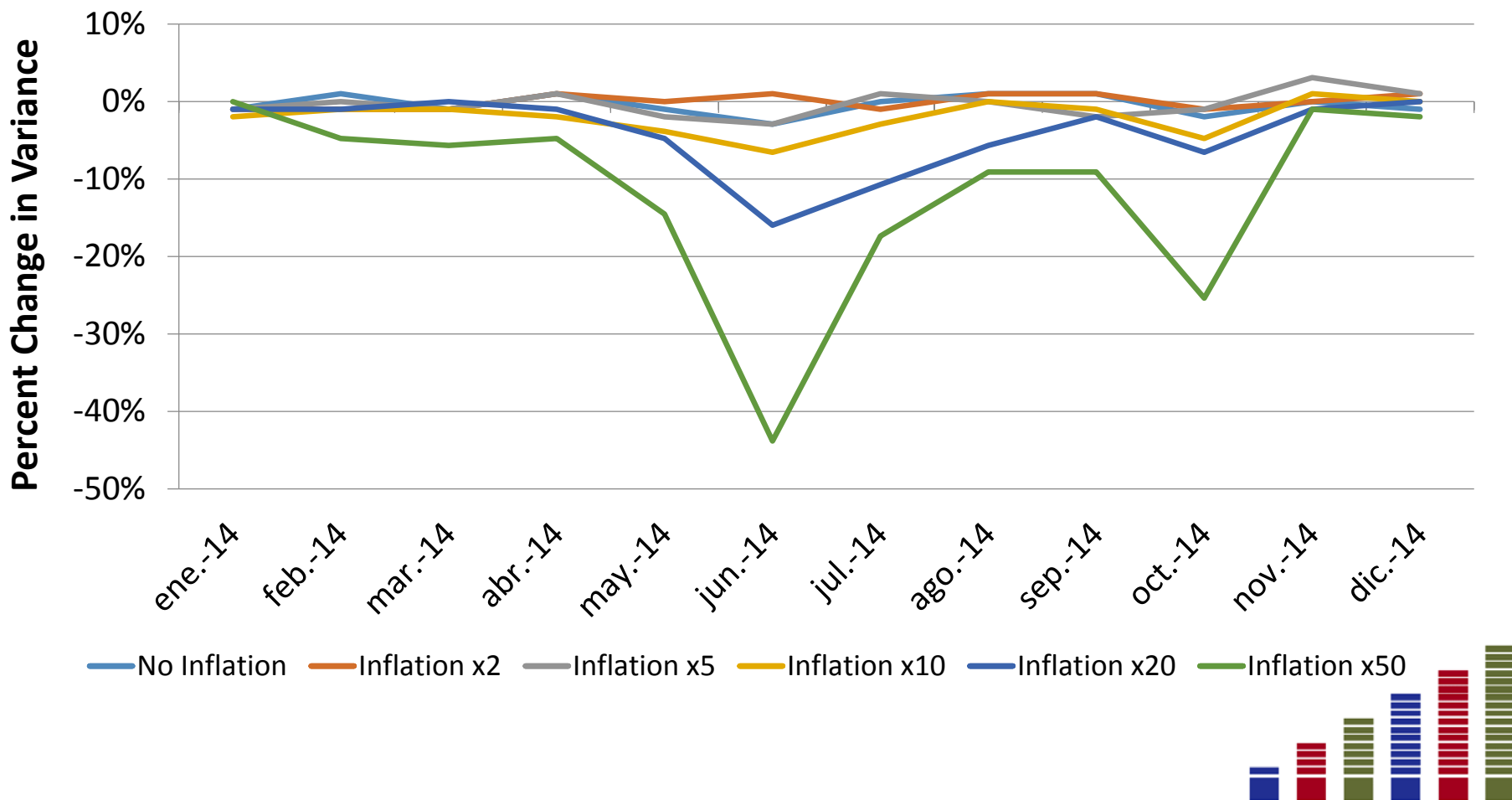- Examining trading day weight calculations and holiday adjustments

Building Small Area Estimation models

- Using Fay-Herriot models to model estimates of totals of sales directly
- Identifying limitations regarding input from MRTS at the state level
- Making the nation-level model resemble the state-level model by inflating variance inputs
- Showing usefulness of First Data aggregates as covariates

Economic Directorate Big Data Projects
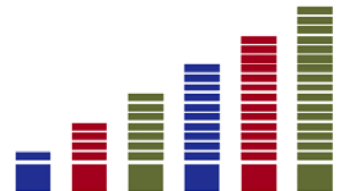## Retail Third Party Data (Palantir/First Data)
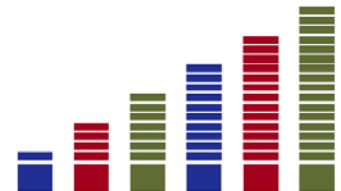
Percent change in model variance over original variance for national-level model

Examining trading day weight calculations and holiday adjustments

- Using daily data seasonal models developed by the US Census Bureau's Tucker McElroy and Brian Monsell
- 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
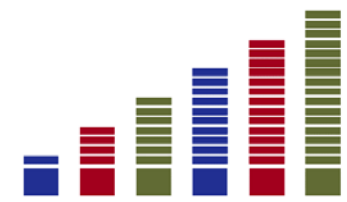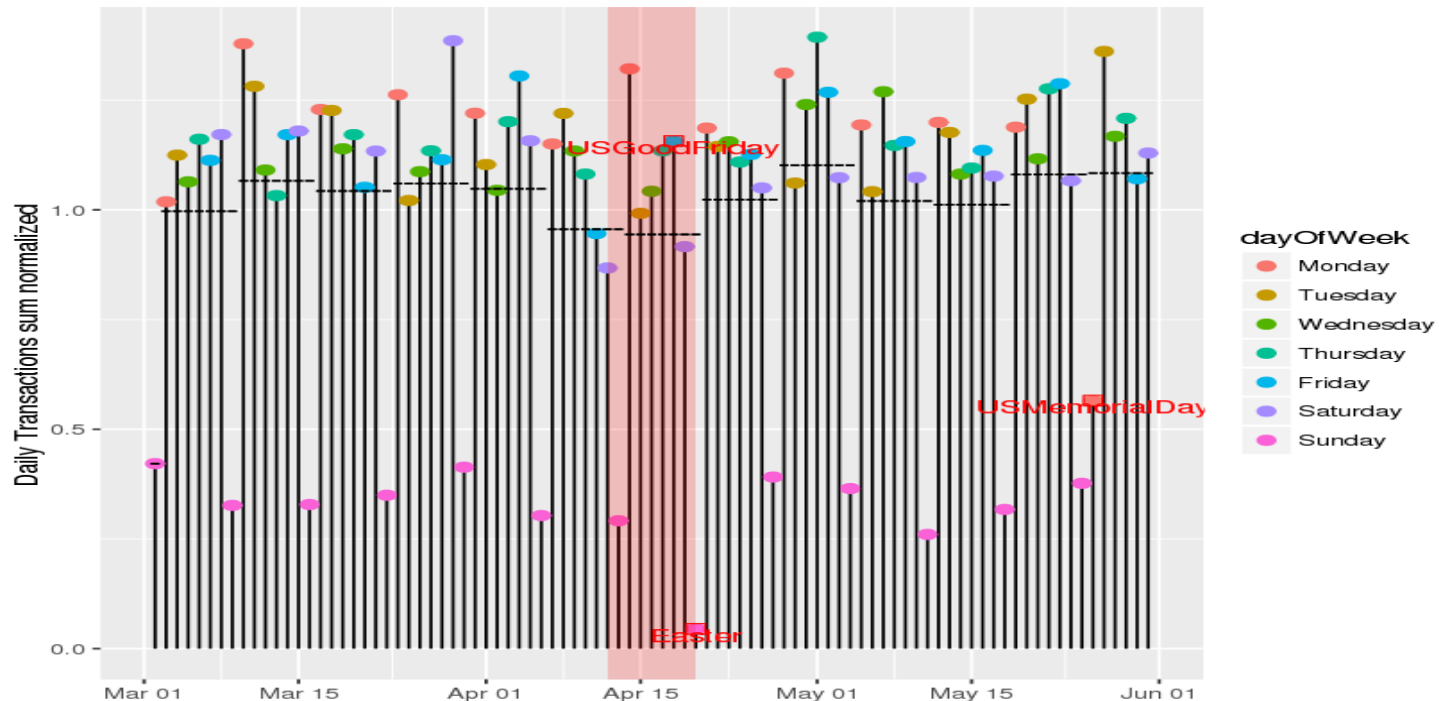
Economic Directorate Big Data Projects
Retail Third Party Data (Palantir/First Data)

Daily data has revealed an Easter Sunday holiday effect

Appliance, Television, and Other Electronics Store
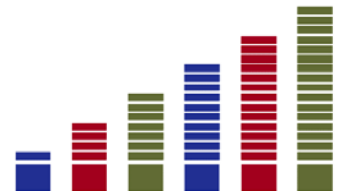(NAICS Code 44311) - 2014

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
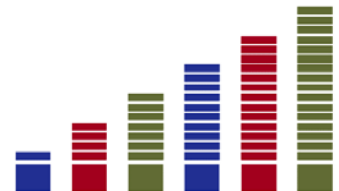- Low confidence in permit details

Retail Trade Data

- Continue to research outside data sources
- Complete work with Palantir and summarize findings

Passive Data Collection

- Explore publicly available API data feeds
- Possibly partner with a third-party source to determine whether their respondents will agree to share their data with the Census Bureau

Questions

Brian Dumbacher

Brian.Dumbacher@census.gov

Rebecca Hutchinson

Rebecca.J.Hutchinson@census.gov