UNIVERSITY OF MICHIGAN

Big Data: Challenges and Opportunities for Central Banks

Matthew D. Shapiro
University of Michigan and NBER

IMFS Working Lunch
Goethe University Frankfurt

21 October 2019

Harness Naturally-Occurring Data

Key principles

- Collect data in the form that they are produced by businesses and households in their normal course of activity
- Shape data to measurement concepts, not vice versa

Harness Naturally-Occurring Data

Examples

- Prices and quantities from retailers' information systems (today's example)
- Income and transactions from household surveys (Gelman,Kariv,Shapiro,Silverman, Tadelis, *Science* 2014}
- Intergenerational inequality from tax returns (Chetty et al)

Continuing role for surveys

Subjective data

- Preferences
- Expectations
- Information
- Attitudes
- Wellbeing

Data for benchmarking of linked data

The Future is Now

From pilot to production

- Measuring quality change
- FRB: Currently using bog data indexs
  - Sales: Uses First Data Credit card processing
  - Employment: Uses ADP payroll processing

Opportunities for central banks

- Price and inflation

Re-Engineering Key National Economic Indicators

Project to build price and quantity from retail transactions

Joint work with Gabriel Ehrlich(Michigan), John C. Haltiwanger (Maryland), Ron Jarmin (Census), David Johnson (Michigan), Matthew D. Shapiro (Michigan)

- AEA P&P (2019)
- 2019 CRIW/NBER Conference on Big Data for 21st Century

Status quo:
Decentralized data collections

Real output

- Census collects the “numerator”: Revenue
- BLS collects the “denominator”: Prices
- BEA does the division: Q = P*Q/P

Non-simultaneous collection of price and quantity

- Stratified surveys from small and deteriorating samples
- Mismatch of price and revenue data
- High cost and burden
- Difficulty of accounting for changes in products

Re-engineering measuring sales and prices

Challenge: Tap the firehose of transactions level data now available from businesses on P and Q .

P&Q microdata

- Internet retailers
- Brick and mortar
- Aggregators

Agencies

Data products:

- GDP
- inflation

Data improvements:

- Quality change
- Timeliness
- Granularity
- Distributional statistics

Reengineered data for retail P and Q

Item-level transactions data

- Item-level data allows inferring price from sales and quantities
- Price, quantity and revenue measured
  - Simultaneously
  - At high frequency
  - Universe (or large sample) of transactions
  - With little lag
  - With reduced need for revisions
  - With granular information on location of sale (geography, store/online)
  - Immediate accounting for changes in goods

Re-engineering: Accept data as they come

Alternative modes of data collection should co-exist:

1. Direct collection of item-level transaction

   e.g., Australian Bureau of Statistics received transactions date from chain grocers

2. Firms aggregate transaction data with APIs

   Multiple APIs to accommodate different information systems

3. Aggregators

   Valued-added product: Prepare statistical reports (data feeds) from information already collected from firms

Re-engineering benefits to firms

- Data feed replaces multiple surveys and enumerations
- Data requests match information systems
- Official statistics better matched firm-specific metrics
- Better national economic indicators
- Better evidence on productivity and innovation

Re-engineering challenges

- Company buy-in for reporting item-level data
- Heterogeneity of company information systems
- Stability/consistency of data stream
- Re-engineering: Human, software, and computation/storage
- Organization and coordination of the statistical agencies
- Conceptual and measurement issues (this paper's topic)

Roadmap of analysis presented today

Using scanner data for P and Q

- Nielsen covers grocery stores and mass merchandisers
  - More than 100 product groups and 1000 product modules.
  - Classify into Food and NonFood items
    - Food nominal expenditures: Compare scanner data to Census surveys and Personal consumption expenditures for food (Scanner provides high frequency product detail)
    - Food and NonFood prices indices: Compare scanner price indices (with and without quality adjustment) to BLS CPI
- NPD covers general merchandise and online retailers
  - NPD data have rich product attributes
  - Explore hedonics vs. alternative methods (e.g., UPI) for quality adjustment

Price indices adjusted for quality

Key challenge/opportunity: Enormous Product Turnover

- 650,000 products per quarter from 35,000 stores
- Product entry and exit rates (quarterly)
  - 9.62% (entry) and 9.57% (exit)
- Sales-weighted entry and exit rates
  - 1.5% (entry) and 0.3% (exit)
  - Rates vary substantially across product groups
  - Asymmetry in sales-weighted: “slow death” of exiting products

Source: Nielsen scanner data (Food and NonFood)

Evaluating product turnover

Some product turnover is mainly packaging and marketing.
Product entry and exit rates
for soft drinks are both 7.1% per quarter.
Sales weighted: 0.3% (entry), 0.07% (exit)

Source: Nielsen scanner data

Some reflects substantial changes in product design.
Product entry and exit rates for video games 12.9%
and 13.5% per quarter.
Sales weighted: 30.3% (entry) 0.5% (exit).

Capturing product quality: Alternative approaches

UPI: Expenditure function approach using CES aggregators

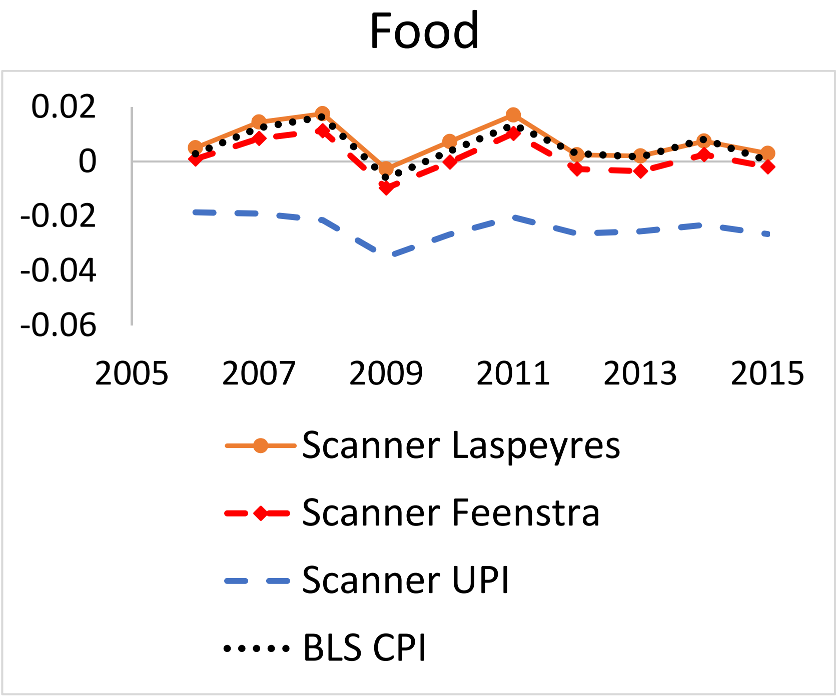
- Capture product turnover with changing expenditure shares of new vs. old goods PV_{adj} (Feenstra 1994)
- Extend to capturing quality/appeal change of existing goods CV_{adj} (Redding-Weinstein *QJE*)
- *Needs item classification/nesting + estimation of elasticity of substitution*

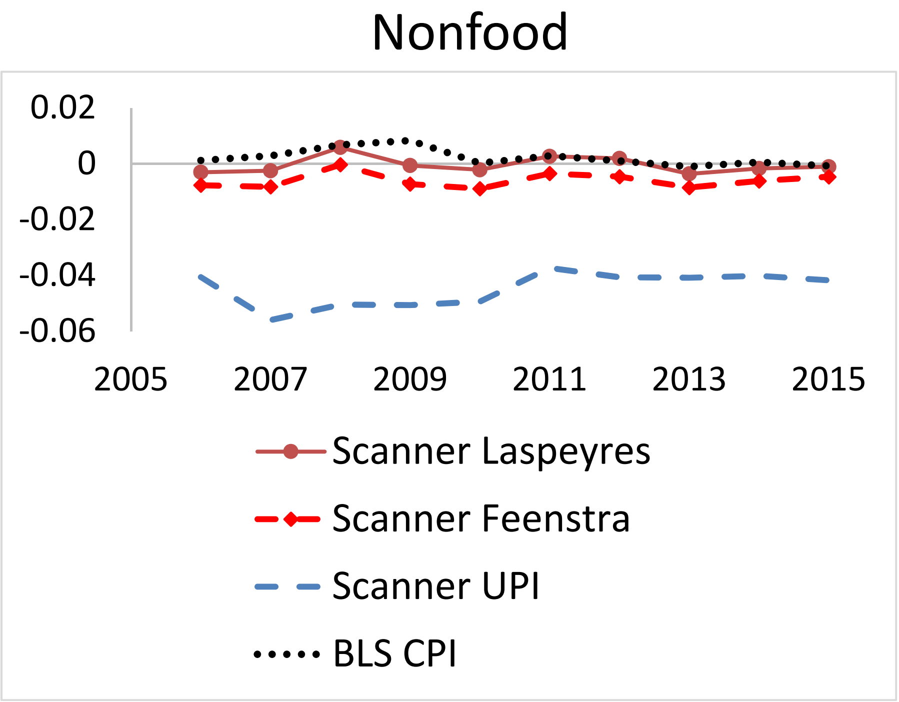
Capturing product quality: Alternative approaches

Hedonic approach

- Estimate hedonic function within product groups using relationship between P and attributes (Pakes 2003)
- Use chain-weighting to accommodate turnover (Bajari and Benkard 2005)
- *Needs item attributes*

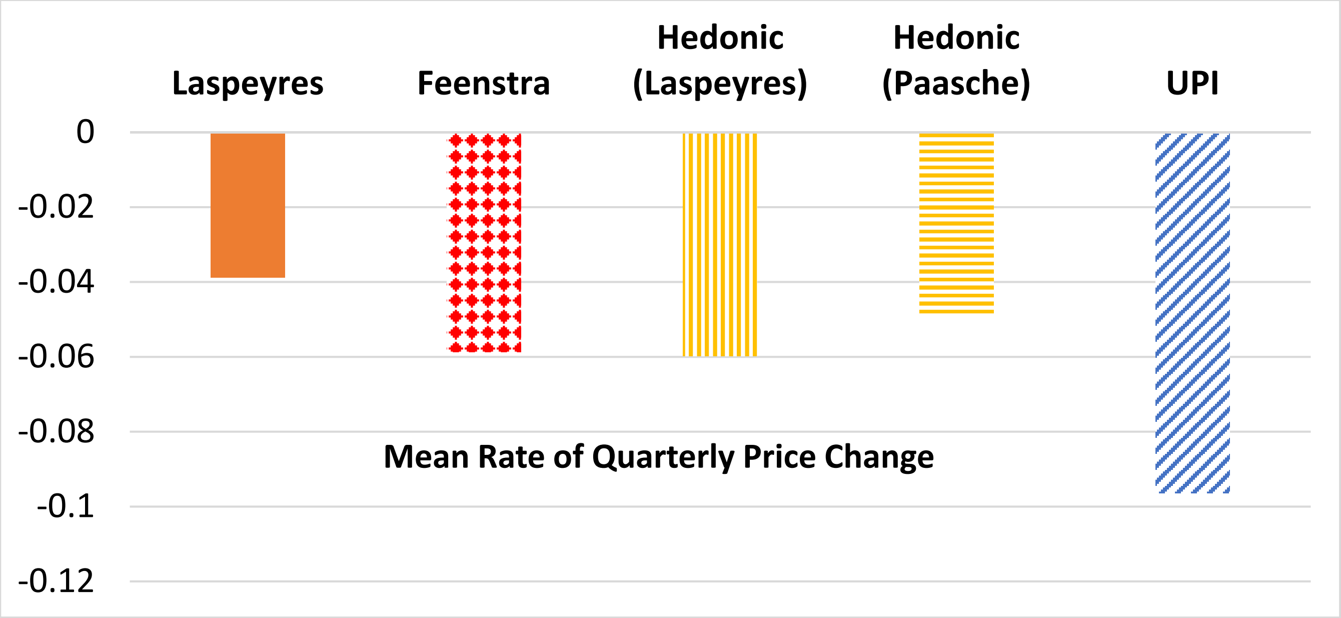
Laspeyres index using scanner is similar to BLS CPI (especially food)
UPI have much lower price change





UPI: Significant valuation of product variety

Alternative Price Indices Memory Cards



Key attributes for Memory Cards: Size and Speed, R-squared for Hedonics is about 0.8 each quarter

Summary of findings

- Possible to adjust for quality change/product turnover at scale
- Initial estimates imply substantial deflation from quality change
- Research questions
  - Reconciling hedonic and UPI approaches
  - How much to value product variety

How are central banks using big data?

Central banks have traditional used non-designed data

- Alan Greenspan said to follow 10,000 series
- Beige book (regional business conditions)

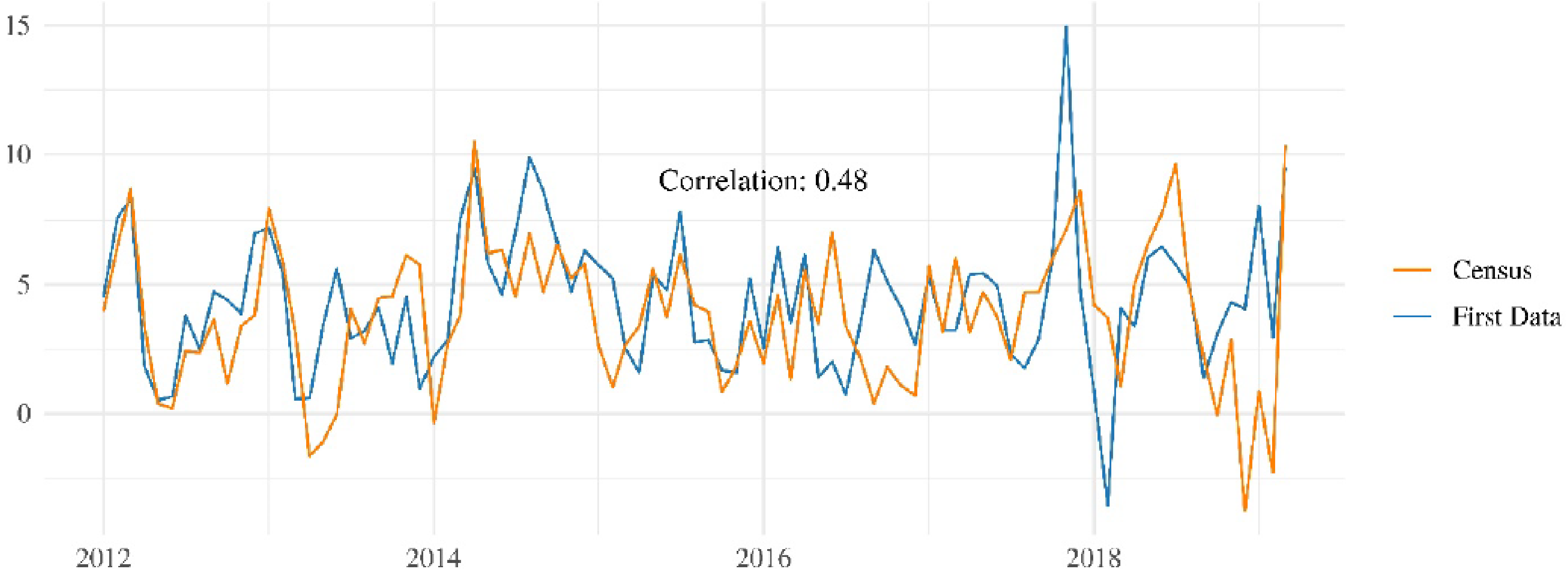
How are central banks using big data?

Current FRB projects monitor business conditions—High frequency activity

- First Data: Credit card transactions to measure spending
- ADP: Payroll data to measure employment

Credit card transactions vs retail sales

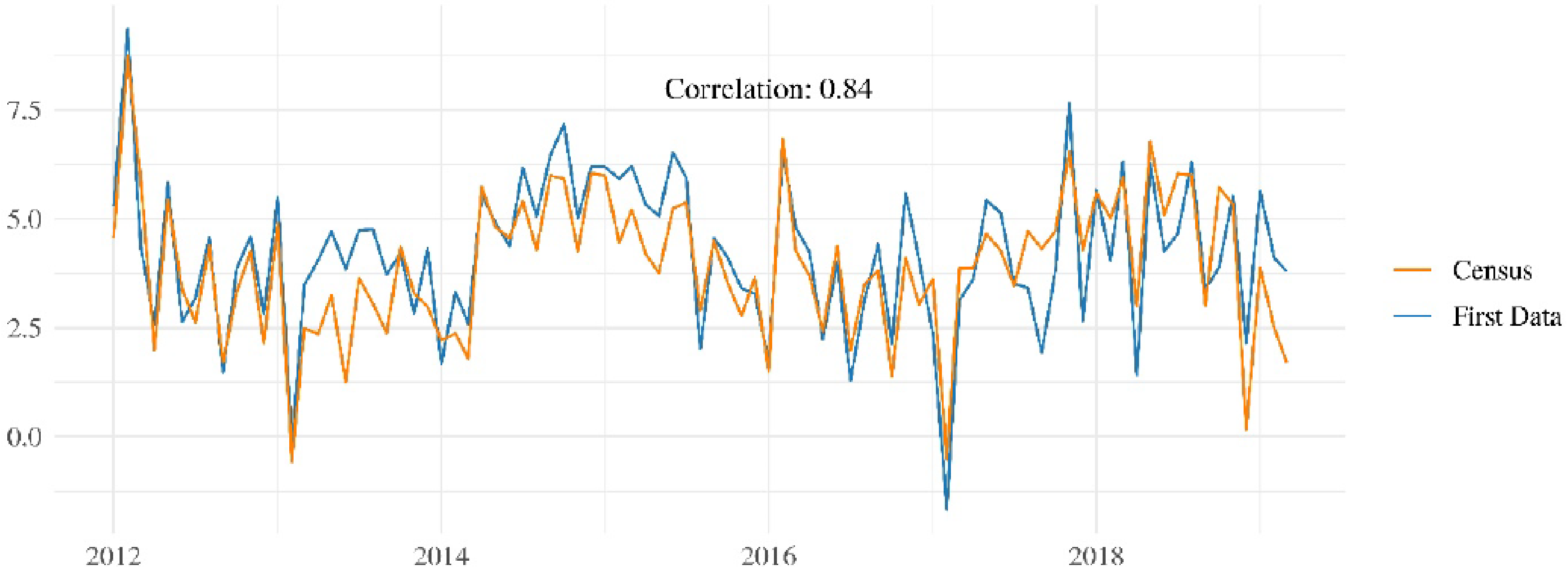
Figure 8. National Retail Sales Group, 3-Month Percent Change



Source: Aladangady, Aron-Dine, Dunn, Feiverson, Lengermann, and Sahm (2019)

Credit card transactions vs retail sales

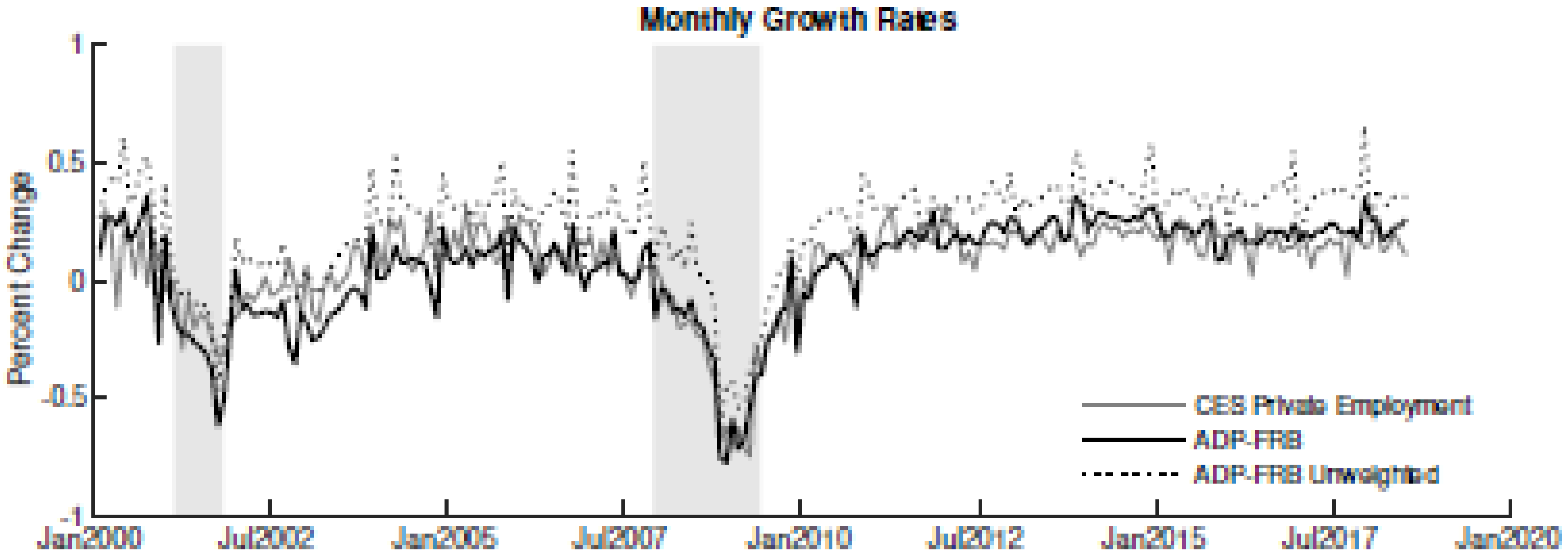
Figure 7. National Retail Sales Group, 12-Month Percent Change



Source: First Data and Census, not seasonally adjusted.

Source: Aladangady, Aron-Dine, Dunn, Feiverson, Lengermann, and Sahm (2019)

ADP vs CES Employment



Source: Cajner, Crane, Decker, Hamines-Puertolas (2019)

“While the incremental improvement in forecasting revisions in general is small, the First Data estimates are particularly helpful as an independent signal when Census preliminary estimates show an unusually large change in sales.”

[Aladangady et al. (2019), fn 18]

How are central banks using big data?
Big data on prices

Billion Prices Project (webscraping)

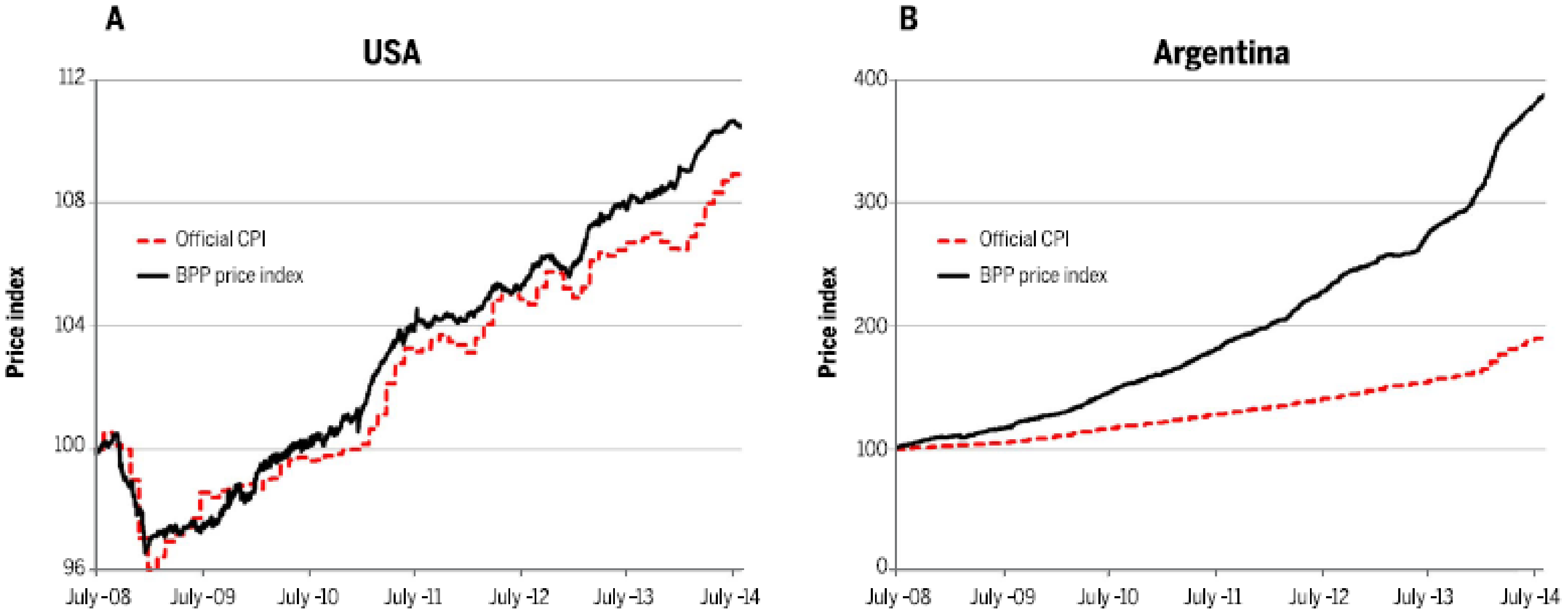


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of 1 July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (*18*), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP]

How are central banks using big data?
Big data on prices

Lessons from “Re-engineering” project

1. What is trend level of inflation
2. Short-term monitoring of inflation shocks
3. Understanding price dynamics and Phillips curve

EXTRA SLIDES

Product Variety and Consumer Valuation Bias Adjustments

$$PV_{adj} = \frac{\lambda_{t,t-1}}{\lambda_{t-1,t}}$$

$$CV_{adj} = \left(\frac{\tilde{S}_t^*}{\tilde{S}_{t-1}^*}\right)$$

$$\lambda_{t,t-1} \equiv \frac{\sum_{k\in\Omega_{t,t-1}} P_{kt}C_{kt}}{\sum_{k\in\Omega_t} P_{kt}C_{kt}}$$

$\Omega_{t,t-1}$ = Goods common to t-1 and t

$$\lambda_{t-1,t} \equiv \frac{\sum_{k\in\Omega_{t,t-1}} P_{kt-1}C_{kt-1}}{\sum_{k\in\Omega_{t-1}} P_{kt-1}C_{kt-1}}$$

Ω_t= All goods in period t

$$S_{lt}^* \equiv \frac{P_{lt}C_{lt}}{\sum_{k\in\Omega_{t,t-1}} P_{kt}C_{kt}}$$

PV_{adj} and CV_{adj} may be related by complex post-entry and pre-exit dynamics

$$\tilde{S}_t^* = \left(\prod_{k\in\Omega_{t,t-1}} S_{lt}^*\right)^{1/N_{t,t-1}}$$

Unified Price Index (UPI) (Redding and Weinstein 2018)

$$\mathrm{UPI} = PV_{adj}{}^{\frac{1}{\sigma-1}}\ CV_{adj}{}^{\frac{1}{\sigma-1}}\ \mathrm{RPI}$$

PV_{adj} = Product Variety Adjustment (Feenstra)

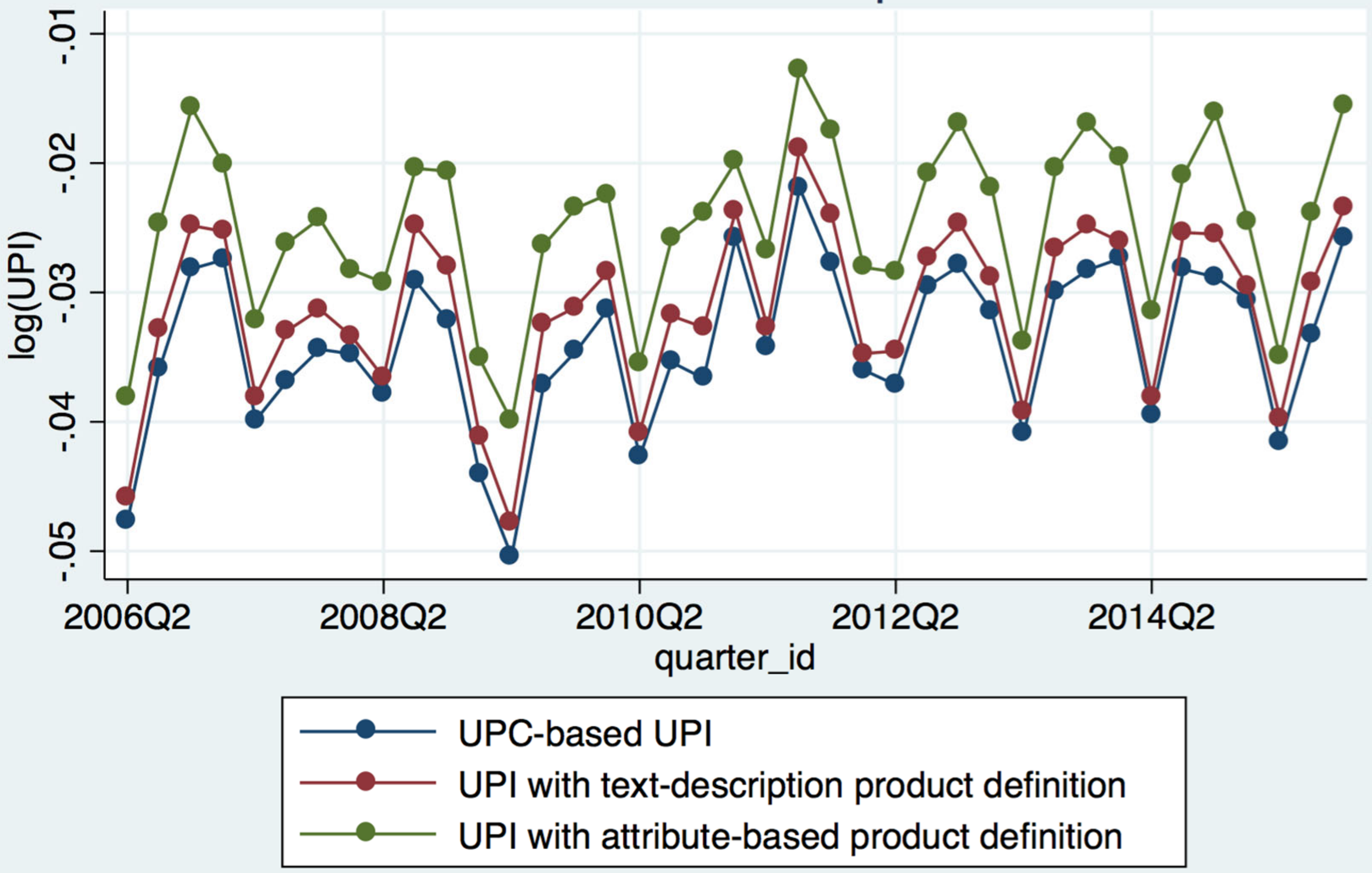
CV_{adj} = Consumer Valuation Adjustment (RW)

RPI = Continuing goods price index (Jevons)

σ = Elasticity of substitution

Applied to narrow product groups; requires estimate of elasticity of substitution

UPIs: UPC-based vs broader product definition



Notes: Quarter uses NRF definition, i.e. 2007Q1 refers to Feb 2007 - April 2007.

Item-level data shows that collapsing into broader product definitions increases UPI (towards Laspeyres)

Attributes here based on product module, brand, size and packaging.

This approach could be Modified to nested CES.

More generally, close substitutes in terms of grouping of goods based on attributes is worth considering.

If attributes used to nest, do we converge towards Hedonics?