A Management and Technical Perspective on the use of alternative data sets at BLS

BLS Technical Advisory Committee
June 20, 2014

Michael W. Horrigan
Associate Commissioner
Office of Prices and Living Conditions

Using alternative data sets at BLS: A management and technical perspective

- BLS Senior Retreat
- Types of alternative data and BLS uses
- Cautions on the use of alternative data sets
- A ‘draft’ vision for the use of alternative data sets at BLS

BLS Senior Staff Retreat

- Theme was ‘Visioning for the future of BLS’
- Four initiatives

1. Respond to emerging data needs and rapidly fill data gaps
2. Constantly seek and incorporate alternative sources of data and techniques
3. Use consistent terminology and classifications for all BLS-released data
4. Develop a foundation for a highly skilled and flexible workforce to support BLS organizational excellence

BLS Senior Retreat

- Constantly seek and incorporate alternative sources of data and techniques
  - What should be the BLS's highest priorities in investing our scarce and declining real resources in terms of the uses of alternative data and techniques?
  - For each instance in which we use alternative data in the production of our economic statistics, what are the tradeoffs in terms of data quality and transparency, and are those tradeoffs worth making the investment?

Types of alternative data and BLS uses

- Webscraped data
- Internet search data
- Social network data
- Federal administrative data
- Private Vendor data
- Corporate data
- Private sector process data

Webscraped data

- Billion Prices project
  - My initial interest in big data
  - Daily CPIs in 22 countries
- Some BLS uses
  - Create data base of product characteristics for use in quality adjustment hedonic models
    - Televisions
    - Camcorders
    - Camera
    - Washing Machines
  - Research to expand use to collect prices for used and new books.

Internet Search Data

- Google
  - Tools to create large data files that combine publicly available data on social and economic activity stratified by geography, and social-demographic characteristics
  - Modeling form combines Google search index data in the current period with past values of an economic measure from the statistical system to predict a future value of the same concept.
  - $Y_t = f\ (\ \text{Search}_{\ t\text{-}1 \text{ to } t},\ Y_{t\text{-}1,\ t\text{-}2,\ \ldots})$
  - Example: Initial claims
  - No active BLS use of this alternative data source

Social network data

- Tweets – Matthew Shapiro et al., University of Michigan Study
  - Case study of job loss related tweets that examines the correlation with unemployment data to predict initial claims
  - No active BLS use of this alternative data source

Federal Administrative Data

- Sampling frames used by statistical agencies for drawing stratified probability samples and in the construction of estimation weights
  - Quarterly Census of Employment and Wages (QCEW)
  - Census Bureau Census of Manufacturing, etc
  - National Agricultural Statistics Service (NASS) Census of Agricultural Establishments
  - BLS Census of Fatal Occupations and Illnesses (CFOI)
  - Center for Medicare and Medicaid Services (CMS) Administrative Data

Federal Administrative Data

- These Federal sampling frames are also used for the production of statistics.

- A common primary purpose of administrative data bases is to administer tax and benefit programs:
  - Unemployment Insurance – QCEW
  - Social Security Administration records

BLS

BLS uses of Federal Administrative Data

- Drawing stratified probability samples and the construction of sampling weights.
  - Current Employment Statistics (CES)
  - Occupational Employment Survey (OES)
  - Job Openings and Labor Turnover Surveys (JOLTS)
  - Producer Price Index (PPI)
  - Imports and Exports – International Price Program (IPP)
  - National Compensation Survey (NCS)
  - Occupational Safety and Health (OSH) Survey

BLS uses of Federal Administrative Data

- Producer Price Index Program
  - Goal is to sample establishments by industry and product lines within establishments proportional to revenue
  - Lack of access to Census of industry data for drawing samples
  - Uses Census data for construction of weights
  - Uses various other alternative sampling frames for drawing samples in specific industries
    - Railroads
    - Hospitals (plus quality adjustment of hospitals)

BLS uses of Federal Administrative Data

- Imputation
  - CES uses QCEW data for imputation of key (highly weighted) non-respondents
- Modeling
  - CES uses QCEW data as an input into its birth-death model

BLS uses of Federal Administrative Data

- Uses of Administrative data for direct estimation
  - The International Price Program (IPP) uses data on crude petroleum from the Energy Information Agency for the construction of their import indexes.
  - The CPI Program uses SABRE data to construct their airline passenger fare index.
  - The PPI Program uses Department of Transportation data on baggage fees for that specific price index.

BLS uses of Federal Administrative Data

- Uses of Administrative data for direct estimation
  - The PPI Program uses State gambling commission reports to calculate the casino gambling portion of the casino hotels industry. While not what one may view as a traditional ‘administrative data set’ it is a comprehensive listing of fees and regulations across States.
  - The PPI Program uses administrative data from CMS on Medicare reimbursements for hospital and physician treatments in their health care indexes.

BLS uses of Federal Administrative Data

- Uses of Administrative data for direct estimation
  - The IPP indexes for export wheat, corn, and soybeans are based on an average over the first week of the month of daily market price quotes published by the U.S. Department of Agriculture (USDA).
  - The daily market prices that USDA collects are the universe of market prices at each port where a specific type of grain is exported from.

BLS uses of Federal Administrative Data

- Uses of Administrative data for direct estimation
  - In IPP, revised estimates for import petroleum are derived from the EIA-856 Monthly Crude Oil Acquisitions Report, a tape of price quotes that BLS receives from DOE.
  - These data are an administrative census of the actual price quotes. IPP also uses an average of daily spot market prices over the first week of the month to price some other petroleum products such as gasoline, diesel fuel, and heating oil.

BLS uses of Federal Administrative Data

- QCEW Hurricane maps
  - Combines detailed QCEW on total employment, total wages, and the count of establishments with flood zones (geographical areas) that have been created by the U.S. Corp of Engineers and State emergency management authorities. These maps are now on the BLS public web site. http://www.bls.gov/cew/hurricane_zones/home.htm

Private Vendor data: BLS uses

- Stock Exchange Security Trades
  - PPI receives a monthly census of all bid and ask prices and trading volume for all traded securities as of market close for 3 selected days of the month.
  - These data are used for index estimation

- JD Power
  - Used car frame for CPI
  - Researching use for CPI production of new car price indexes

Private Vendor data: BLS uses

- Scanner data: Homescan, Nielson
  - Actual sales transactions
  - Comparisons of national distribution of selected products using scanner data with results from CPI disaggregation process

Private Vendor data: BLS, BEA and Census uses

- Claims data
  - Validation of MEPS and CPI inflation rates
  - CPI constructs experimental disease based price indexes using annual weights from the MEPS household survey data
  - BEA and BLS have done a tremendous amount of joint work in this area
- Credit card data
  - BEA is looking into the use of credit card data for use in Personal Consumption Expenditures

Corporate data: BLS uses

- CES collects data from 88 corporations at their Electronic Data Interchange facility in Chicago, IL.
  - Accounts for nearly 10% of total weighted employment
  - Respondents submit electronic files in BLS formats
- More generally, corporate data may take the form of data extracts from company data systems that are not translated into BLS formats.
  - Example: OES collection

Corporate data: BLS uses

- Example: OES collection from large firms
  - NCS data collection at overlap establishments (reporting both for NCS and OES)
  - Electronic files with firms' occupation title, wage and employment information
  - Field staff converts company specific occupational formats into OES occupational codes
  - As needed, conversion of wages to OES wage intervals is also done.

Corporate data: BLS uses

- CPI is also examining the potential of using corporate data records.
  - Matched model requirements or some version of unit value pricing
  - Difficulty in capturing quality change
  - Actual recorded transactions, including all coupons and discounts
  - Processing challenges associated with large volumes of data
  - Potential for larger samples than from original sampling draw

Private sector process data

- UPS
  - Using telematic sensors in over 46,000 vehicles, big data on route selection, speed, and direction
  - Estimated savings of 8.4 million gallons of fuel by cutting off 85 million miles of route driven in 2011.
- GE
  - Use of real time monitoring of machines with big data analytic techniques to improve productivity of electricity generating machines, aviation, rail transportation, and health care.

Private sector process data

- GE
  - Power of 1% and the industrial internet
  - 1% savings in fuel consumption in aviation would generate savings of $30 billion
  - 1% efficiency improvement in GE’s global gas fire plant fleet would produce an estimated savings of $66 billion in 15 years.
- No active BLS use of these alternative data

Using alternative data sets at BLS: A management and technical perspective

- BLS Senior Retreat
- Types of alternative data and BLS uses
- Cautions on the use of alternative data sets
- A ‘draft’ vision for the use of alternative data sets at BLS

Some Cautions

- Groves, Washington Post, August 7, 2012
  - Costs and declining budgets make using big data in constructing blended estimates a reality
- Despite budget pressures, one of the principal reasons underlying the continued use of direct data collection from stratified probability samples is the goal of producing unbiased estimates of economic concepts with acceptable margins of error.

Some Cautions

- A natural question that arises in considering the use of alternative data sets is to ask, to what extent does the use of alternative data bring us into conflict with these goals?
- The previous section, however, shows that we have already made the choice of using blended data, a lot.
  - We must produce and maintain transparent methodological documentation in our use of blended data sources.

Some Cautions

- One of the biggest challenges in using alternative data is in knowing (or not knowing) the relationship between the scope of alternative data and how it relates to the target population under study.
  - For example, internet transactions, through web scraping or other sources, provide information on web transactions.
  - The types of transactions not captured and how they relate to the estimation goals of the program must be understood and documented, and may in some cases constitute a basis for rejecting their use.

Some Cautions

- In the cases where the alternative data does not represent a census or universe of units or transactions, do we have sufficient information to determine their weights or relative importances in the construction of estimates? Under what circumstances do we decide to use or not use such data?

Some Cautions

- At what level of aggregation do we use alternative data?
  - We already use alternative data for direct estimation of many top side estimates
  - The example of CPI’s use of Energy Information Agency data on gasoline revenue by type of gasoline to ratio allocate CE survey generated top side estimates of consumer gasoline expenditures is instructive
    - Trade lower Mean Squared Error for Bias

Some Cautions

- Finally, under what conditions is it not appropriate legally or by statistical principle to use alternative data sets?
  - For example, it is long-standing tradition on the part of BLS to not use data collected for the purpose of administrative enforcement. Does that standard still apply?
  - As another example, it is a principle in sampling not to use industry association membership lists as the sole source of information for sampling frames given the self-selection nature of who joins industry associations.

Some Cautions

  - Should a similar standard apply when choosing to use or not use data that are published by industry associations in our estimation processes? Or does a documentation of the sources (and possible biases in representing the target population) suffice?
- Finally, in the case of web scraping, does BLS need to seek permission from the web sites we scrape for the purposes of collecting data?
  - Census Bureau has concluded it does not.

Using alternative data sets at BLS: A management and technical perspective

- BLS Senior Retreat
- Types of alternative data and BLS uses
- Cautions on the use of alternative data sets
- A ‘draft’ vision for the use of alternative data sets at BLS

Draft ‘vision’

- Linking
- Electronic data collection
- Acquiring alternative data to replace direct data collection
- Webcraping

Vision: Linking

- Linking across BLS establishment data sets to the QCEW or other Federal administrative data bases has been underutilized
  - QCEW (9 million) and OES (1.2 million over 3 years)
    - Example: OES as a times series
    - Examination of occupations with rising wages and employment by industry employment growth and further stratification down to the MSA level (or lower using modelling
  - Similar linkages of QCEW to other BLS establishment data bases

Vision: Linking

- Linkages of QCEW to other establishment data bases
  - Custom Bureau sampling frame for exports match to the QCEW
  - Currently IPP gets export trade volumes from the Custom Bureau for sampled units – extend to all units?
- PPI use of Census establishment frames to draw samples based on product revenue
  - Current research using multi-establishments
  - Extension to small firms with CIPSEA amendments to allow access to IRS data

Vision: Electronic data collection

- A large share of collected information in our establishment surveys comes from a small share of total establishments owing to the size concentration of economic activity.
- In 2012, of the known value of U.S. exports that could be matched to specific companies:
  - the top 50 companies contributed nearly 31% of known value,
  - the top 100 nearly 40%,
  - the top 250 just over half,
  - and the top 2000 nearly 78%.

Electronic data collection

- Move beyond our current approach to collecting electronic records from firms using our survey forms through the EDI center or the BLS Internet Data Collection Facility
- Allow firms to report using their formats and data bases
- Using autocoding learning models, computational linguistics to convert firm based data and classifications to BLS concepts

Electronic data collection

- Alex Measure's TAC presentation at 2:00 pm on autocoding in the NCS is an example of such an approach.
  - Apply his research to large electronic OES files that are being converted by hand
- Advantages
  - Reduction of respondent burden
  - Enlarge the number of units reported for
  - Enter the world of global supply chains and global information management

Acquiring alternative data sets for use in estimation

- Acknowledging the need to develop statistical approaches to blending data, there are a lot of opportunities for acquiring alternative data sets that remain.
- The Big Data team John Eltinge and I chair provided numerous examples:
  - ACA related datasets
- Office of Employment and Unemployment
  - Supplement JOLTS data on vacancies with job openings data from private vendors (Snagajob, Burning Glass, Career Builder)

Acquiring alternative data sets for use in estimation

- Office of Productivity
  - Truven Health Analytics data for health care productivity measures
  - American Short Line and Regional Railroad Association data for the potential development of productivity measures for Short Line railroads (and complete coverage for Rail Transportation);
  - Data from Compustat to potentially produce State level productivity estimates;

Acquiring alternative data sets for use in estimation

- Office of Compensation and Working Conditions
  - IRS form 5500 series (working jointly with the Employee Benefits Statistics Administration) to improve imputation, improve data editing and validation, and supplement published National Compensation Survey publications;
  - Data from the Federal Mediation Reconciliation Service on work stoppages;

Acquiring alternative data sets for use in estimation

- Office of Compensation and Working Conditions
  - Data from National Institute for Occupational Safety and Health that combines fatal highway accident data with the Census of Fatal Occupational Injuries to obtain greater detail about these fatalities;
  - OSHA is proposing a regulation to require web-based reporting of injuries and illnesses, which would create a new web based source of administrative data on these concepts.

Acquiring alternative data sets for use in estimation

- Office of Prices and Living Conditions
  - Use of credit card data collected by BEA to potentially use to create travel and tourism price indexes.
  - Use of secondary source data on education to develop import and export education price indexes

BLS

Webscraping

- Determine whether or not we need permission to scrape web sites.
- Examining the most promising areas for webscraping (advice needed):
  - Food prices
  - Cable TV prices
  - Airline prices
  - Courier services

Concluding remarks

- Linking long overdue
  - High return on investment
- Electronic data collection
  - Significant investment in IT resources
- Alternative data sets
  - Allow programs to continue innovating
  - Share information, minimize stovepipe reinventing the wheel
- Webscraping
  - Fairly specific opportunity set
  - Need to overcome a number of issues

Concluding remarks

- Question for the TAC:
  - What are your views on this proposed prioritization of our future efforts to incorporate alternative data into BLS estimation systems?

Contact Information

Michael Horrigan

Associate Commissioner

Office of Prices and Living Conditions

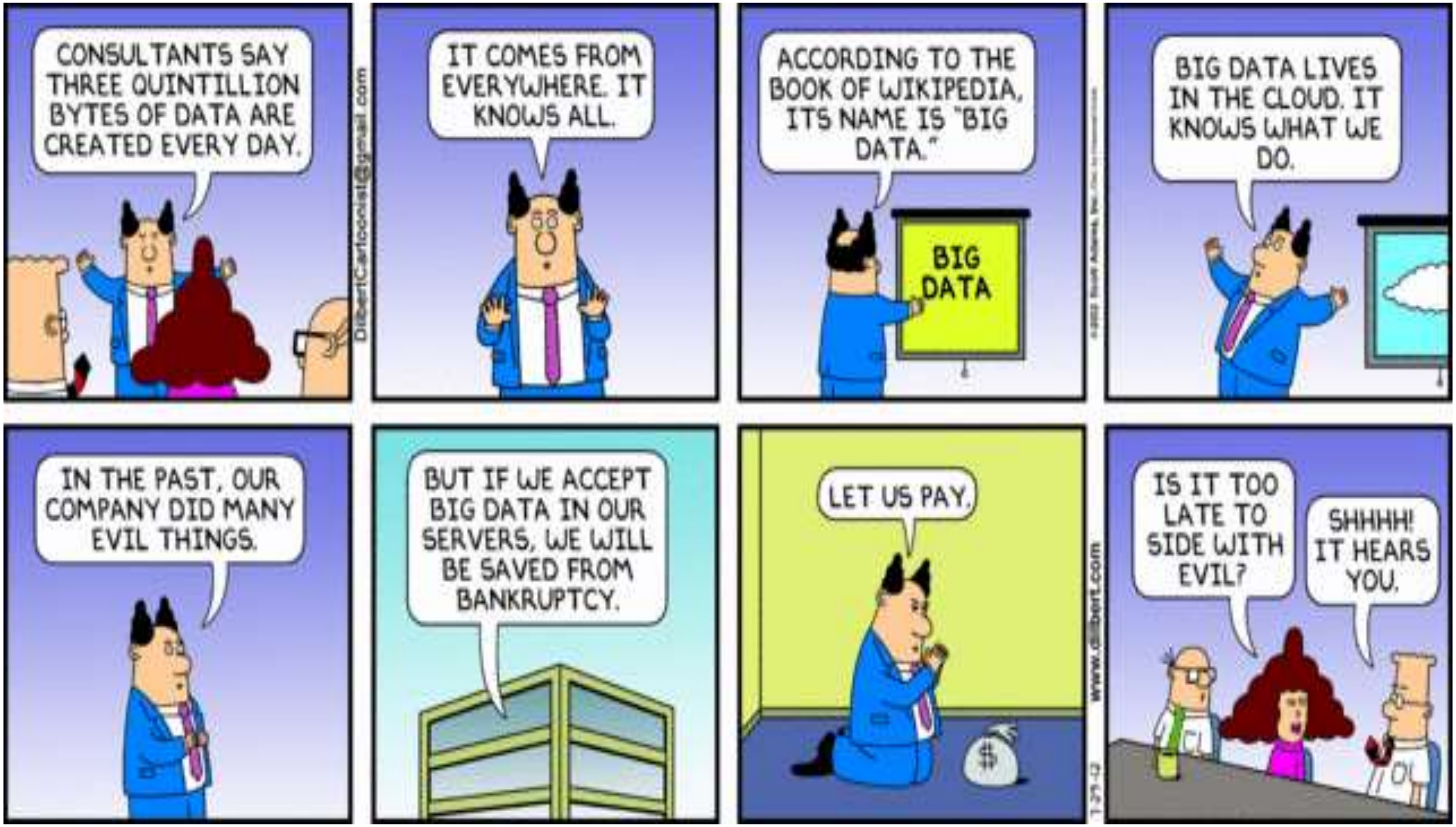
www.bls.gov

202-691-6960

horrigan.michael@bls.gov

until midnight tonight!

What are “Big Data”?


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.
DilbertCartoonist@gmail.com
www.dilbert.com
7-29-12