Using New forms of Information for Official Economic Statistics
-- Examining *The Commodity Flow Survey:*
Executive Summary from the 1rst Workshop in the Census-MIT Big Data Workshop Series

Oct 1, 2015

Micah Altman, MIT; Cavan Capps & Ronald Prevost, U.S. Census Bureau[1]

Workshop Series Overview

Trends and opportunities. In today's increasingly data-driven economy, technology changes data uses, and stakeholder expectations increase, while statistical agency budgets and staffing remain flat. Big Data offers both tremendous promise and new challenges for official statistics. New forms and scales of data may offer opportunities to enhance and strengthen official statistics by improving estimates; lowering costs and helping agencies improve the frequency and timeliness of data releases. Achieving this promise requires innovative integration of methods from many disciplines and the expertise of many sectors.

A number of new initiatives illustrate the breadth of the new evidence available for understanding economic behavior, business, and the economy, and the range of methods available for using it.[2] For example, the Billion Prices project @ MIT is a research initiative that collects prices from hundreds of online retailers around the world on a daily basis to estimate real-time measures of inflation. Google searches have been used to track a variety of economic activity: the Google Unemployment Index "nowcasts" unemployment estimates, while Google Domestic Trends tracks search traffic across individual sectors of the economy as a proxy for activity.

Workshop approach. The workshops series brings together select groups of experts in universities, industry, and the U.S. government. Each of the first three workshops focus on a different set of issues relates to big data -- potential sources; data privacy and security; and barriers to statistical inference. During the workshop, the experts are guided to *explore* the challenges involved in building the next generation of official statistics; *identify* new opportunities

---

[1] Authors are listed in alphabetical order. We describe contributions to the paper using a standard taxonomy. (Allen, Liz, et al. "Credit where credit is due." *Nature* 508.7496 (2014): 312-313.) All authors take equal responsibility for the article in its current form. MA and CC authored early versions of the manuscript; all authors contributed to review and revision; and lin the conception of the article (including core ideas, analytical framework, and statement of research questions). All authors contributed to the project administration and to the writing process through direct writing, critical review, and commentary.

[2] Einav, Liran, and Jonathan Levin. "Economics in the age of big data." Science 346.6210 (2014): 1243089.

to use big data in statistical agencies, and synergistic work in the discipline; and to *examine* broader questions through their application to an exemplar use case.

Workshop One Discussion: New forms of Information for Economic Statistics

Workshop questions and use case. The first workshop in this series explored focused on how big data could be used to accurately augment, extend, and inform official estimates of economic behavior and performance. The workshop viewed these issues through the examination of an exemplar use case "The Commodity Flow Survey".

The CFS, a flagship of Department of Transportation (DOT) statistics, provides critical information to guide long-term changes in the nation's transportation infrastructure. The Commodity Flow Survey (CFS) is the primary source of national and state-level data on domestic freight shipments by American establishments in mining, manufacturing, wholesale, auxiliaries, and selected retail and services trade industries. Data are provided on the types, origins and destinations, values, weights, modes of transport, distance shipped, and ton-miles of commodities shipped. The CFS is a shipper-based survey and is conducted every five years as part of the Economic Census. It provides a modal picture of national freight flows, and represents the only publicly available source of commodity flow data for the highway mode. The CFS was conducted in 1993, 1997, 2002, 2007 and most recently in 2012.[3]

Workshop participants. Seventeen experts participated in the workshop (16 in person). These experts were drawn from senior leadership in the transportation industry, industry associations, federal government, and national experts in academia.

Discussion was conducted under Chatham-house rules, which restricts attribution of individuals, and individual statement without prior explicit approval. A number of participants volunteered to identify themselves for the purposes of this public summary:

- Micah Altman, Massachusetts Institute of Technology (Facilitator)
- William Bostic, Assoc. Dir. Economic Programs, U.S. Census Bureau
- Cavan Capps, Big Data Lead, U.S. Census Bureau
- Alberto Cavallo, Massachusetts Institute of Technology
- Ronald Duych, Commodity Flow Survey, Bureau of Transportation Statistics, Dept. of Trans.
- James Hinckley, Chief Commodity Flow Survey Branch, U.S. Census Bureau
- Ron Jarmin, Deputy Assoc. Dir. for Research & Methodology, U.S. Census Bureau
- Kimberly Moore, Division Chief for Economic Reimbursable Surveys, U.S. Census Bureau
- Ronald Prevost, Senior Statistician, Research & Methodology, U.S. Census Bureau

[3] Kriger, David S. *Freight Transportation Surveys*. Vol. 410. Transportation Research Board, 2011.

Commodity Flow Survey Challenges

The CFS relies on traditional survey instruments for its data collection, and thus is subject to nonresponse and measurement errors (due to inaccurate reporting by respondents, etc.). These non-sampling errors are controlled through survey administration, but their effects on estimates are challenging to estimate. Further, the CFS collects data only from domestic shipping establishments. As a result, it provides an incomplete measurement of the use of domestic transportation infrastructure, since it does not include all components of intermodal shipping, and excludes shipments to domestic destinations that originate from outside the 50 states, and those that traverse the U.S. from one foreign location to another.

The CFS is conducted every five-years, and requires substantial time to generate estimates.: It takes approximately a year to collect the data, a year to process the data, and another year to weight, review, and release the data.

Patterns of shipping are changing due to globalism; increased use of multiple modes of transportation; changes in warehousing and shipping requirements driven by reductions in brick and mortar stores, centralization of distribution networks driven by internet sales; and the increasingly rapid lifecycle of business. Shipping statistics are further complicated by rapid changes in modern commercial transportation. This includes accelerating Internet sales and associated same day delivery, changing industrial/business categories; changing classification systems; growing use of multi-modal carriers; and increasingly complex relationships among shippers, carriers, and third-party logistic providers.

Discussion Framework

Discussion among workshop participants ranged over a wide variety of issues; including key government decisions supported by the CFS; the most important ways in which businesses use transportation and commodity flow information; perceived gaps in the CFS and in other related estimates; sources of information for understanding business transportation patterns, loads, uses, and flows; related technologies and protocols; and barriers to data collection, access, sharing, and use. Across these discussions three main themes emerged: decision support (inference), data sources, and incentives.

Decision support.

The CFS data increasingly supports different sets of decisions for different user communities. In particular, government and business use the data in different ways.

Data from the CFS is used primarily by government to form inferences regarding the flow of shipping, and the inferred use and needs of the transportation infrastructure. Specifically, it supports a number of government decisions for highway infrastructure planning into and between regional metropolitan centers; the identification of freight transportation bottlenecks,

and potential costs associated with those freight bottlenecks; and the identification of needs associated with using multiple modes of transportation.This approach ensures that associated infrastructure needs are documented and trade-offs illuminated for the growth of transportation in a region.

Estimates of commodity flows and transportation infrastructure load can be used for other purposes. In theory, the data could identify the effect of different transportation and shipping configurations on the larger supply chain. In practice, some businesses use the CFS data as a baseline when they forecast transportation costs and times (including traffic) in order to make strategic decisions -- such as market entry, and when they make other logistical decisions -- such as warehouse siting and sizing and locating distribution points.

Businesses need more timely data to support their forecasts -- near real-time data is needed for operations, while a lag of a few months may be acceptable for service planning. Transportation infrastructure is planned and constructed over five-ten years, business strategy is conducted over timescales of several years, service planning over months, and operations over days. Businesses also need this data to be collected more frequently as time-series in order to make more reliable forecasts. Both businesses and local governments require data that can be used to make estimates for more detailed geographies in order to assist in planning public and private infrastructure including roads and warehousing Finally, businesses seek data that directly measures the cost of shipping, and that allows them to forecast traffic disruptions and the associated expense.

Data Sources

There are a variety of sources for data that have the potential to inform the key decisions above. A few of these sources could be used to increase the accuracy and completeness of the CFS baseline data collection, or reduce the costs and burdens associated with establishing this baseline -- these are useful for long-term strategic planning of major infrastructure. Other sources, although incomplete in coverage, and potentially inaccurate, could increase the detail, timeliness, and granularity of estimates. These could be of tremendous value for local infrastructure and service planning, and help to incentivize the same businesses that contribute to baseline data collection.

The data sources identified by participants in the workshop can be grouped into several broad categories:

- *Enterprise systems data* collected from parcel tracking systems (sometimes uses in conjunction with RFID tags), ERP (enterprise resource planning) and DRP (distribution resource planning) systems, and EDI (electronic data interchange) transactions across systems and companies could provide information on content types, modes, loads, times, origin/destination, complete shipping paths, and costs across both shippers and carriers. Automated collection of this data could increase the reliability and completeness of coverage while reducing respondent burden.

- *Traffic and transportation systems* data, collected from sources such as toll transponders, license-plate readers, in-roadway sensors, weigh station, traffic cameras, and traffic monitoring apps (such as Waze), can provide information about traffic patterns, flow, and, in some cases, load.
- *Personal and vehicle sensor* data, collected from sources such as embedded vehicle tracking systems and the GPS sensors in cellphones, can provide information about traffic patterns, and flow.
- *Imagery data*, including satellite and aerial imagery may also provide information on traffic patterns, disruptions, construction, and road damage.

If data from these varied sources could be reliably and confidentially linked the inferential power would be significantly improved over the collection from any single source or from the varied unconnected sources.

Incentives

Each of these sources of data are associated with a different set of stakeholder. For some sources, participants identified key stakeholders who could provide access to substantial data collections or streams. Enterprises like Walmart, UPS, Amazon, Fedex, and Target are potential large sources of EDI or ERP/DRP data; logistics companies and industry verticals (such as Pittney Bowes) and EDI software vendors such as Intuit that focus on small businesses, are sources of EDI data for other companies; telecom providers such as AT&T and Verizon, and mobile phone operating system vendors such as Google and Apple, were identified are potential sources of sensor and imagery data; FastLane and EzPass are potential source of traffic systems data.

Working with these sources to obtain data will require careful analysis of the incentives of these stakeholders and management of the relationships with them. A number of specific issues will need to be addressed with each stakeholder:

- *Value.* Agencies collecting data from stakeholders must be able to communicate and provide value to the participating data provider. Added value could include:

    - providing access to more timely detailed and granular estimates;
    - create new useful public data products, by linking this new data internally with other data sources to meet the forecasting and other needs of business, local and the federal government(on infrastructure needs for example)
    - reducing burden, by automating data collection and reducing the need for manual responses.

    For example, Foreign Trade data could be linked to domestic shipping data. Linkage of Trucking GPS, RFID and shipping survey and/or electronic ordering and tracking data would also improve how the data could be used.

- *Trust.* Stakeholder value their data, and many will expect to establish relationships before sharing. Participants identified a number of strategies for developing trust:
    - Engage stakeholders in workshops, in advisory boards, or in collaborations to develop more useful estimations.
    - Stakeholders will need increased assurance of complete data confidentiality as more data is collected from them.
    - IT security concerns will also increase as data sharing increases. Identify these concerns and transparently and actively address them.

- *Technology.* Many of the most promising primary on-ramps to big data are through existing technical systems. But creating new or modifying current technical systems businesses can be very expensive; and custom reports developed on an establishment-by-establishment basis can be very costly Moreover, it is challenging to deploy a single technology approach that can be that handles both small independent businesses and large enterprises. Participants identified a number of approaches to technology in this domain:

    - Identify opportunities to use existing technologies to reduce cost and burden from data collection while improving quality.
    - Leverage existing software systems -- for example the *Electronic Data Interchange(EDI), Electronic Resource Planning(ERP), Distribution Resource Planning(DRP) systems* listed above.
    - Engage with vendors and standards-drivers to develop data collection modules and data interchange standards and practices.

Next Steps for Big Data and Commodity Flow Estimation

The workshop participants reached a general consensus that the next step towards incorporating big data into commodity flow estimation would be to experiment with the sources above, and to engage those stakeholders that are gateways to those sources. Specific panel recommendations comprised:

1. Evaluate the feasibility of linking existing data such as Foreign Trade data to shipping data in order to provide a more accurate baseline estimate and estimates of port delays.
2. Initiate discussions with vendors of ERP and DRP software, such as SAP and ORACLE to evaluate the feasibility of developing an official “compliance” plugin software module that would automatically and securely send aggregate information to agencies. Such a product could reduce respondent burden and increase the completeness, accuracy, detail and timeliness of information -- while creating a sustainable product for the vendor.
3. Initiate contacts with shippers, carriers, and EDI vendors to explore the possibility of automatic statistical collection of EDI shipping purchase orders, receipts and shipping progress orders.

4. Deploy small teams of data scientists and subject matter experts to conduct 4-6 week projects to identify businesses and other entities that have the potential to provide access to (respectively) substantial collections of information collected from smartphone sensor data; embedded vehicle tracking systems; parcel tracking technologies; traffic cameras, toll transponders, license plate readers, and traffic monitoring applications such as Waze; highway imagery data -- including satellites; and state-level transportation data -- such as weigh station measures, and shipping port information.
5. Convene a workshop or series of meetings to engage with key stakeholders in order to build trust relationships and to explore data sharing approaches. Such a workshop might identify:
    a. Options to access data for experimentation -- e.g. by sharing data snapshots or data sandboxes.
    b. Identification of key individuals and small groups who can partner on experiments and initiatives
    c. Identification of prototypes and proof-of-concepts that could be generated quickly and which would demonstrate value and feasibility to business source/stakeholders.

General Observations On Agency Use of Big Data

Broad new sources of information have the potential to bring increased granularity, detail and timeliness to the next generation of official statistics -- while reducing survey burden.The next generation of official statistics will utilize broad sources of information, potentially linked together, to provide increasing granularity, detail, and timeliness, while reducing cost and burden.

Incorporating big data into statistical agencies will require a number of adjustments. Agencies will need to adapt their strategies, and broaden the current narrow focus on data collection, to include a broader focus on information provisioning.

Agencies will require different sources of data to support different types of decisions. Traditional *designed* survey data contributes to the construction of baselines, and continues to provide the broad context needed for long-term decisions on infrastructure and strategic planning. Designed data collection may be augmented by big data sources that increase coverage, accuracy, or decrease the costs and burden of survey implementation -- but carefully designed data collection will remain essential to establish reliable baselines for population-parameters.

Increasingly however, businesses and governments need more than baselines -- they also need timely, frequent, and granular estimates to inform strategy, service, and operations. Big data that comes from transactional and other *found* sources could enable such estimates.

Utilizing big data requires creating new relationships with businesses. Many of the primary sources of big data are businesses that use data intensively to guide decisions: Big data

sources are also critical stakeholders. In order to obtain access to data from businesses that create and use it, it is critical to both provide a value proposition to the business and to develop a trust relationship.