REVEAL: A Paradigm for Official Statistics[1]

Thomas A. Louis
U.S. Census Bureau, Suitland, MD, USA
Johns Hopkins University, Baltimore, MD USA – Thomas.Arthur.Louis@Census.Gov

Ronald Prevost*
U.S. Census Bureau, Suitland, MD, USA – Ronald.C.Prevost@Census.Gov

Big Data present particular challenges to National Statistical Organizations; there are opportunities and threats. Customers require statistics more rapidly and on more topics than ever before. Big data can provide answers, but with varying and many times unknown degrees of accuracy and focus. In this context, how can statistical agencies best provide policy–relevant information? Traditional approaches such as surveys and censuses are not always timely in their production of statistics. However, with their known quality characteristics they present the basis for benchmarking continuous near term projections. We propose a paradigm for official statistics, REVEAL, that leverages the strength statistical agencies possess – the ability to effectively measure ground truth. The process is: Reassemble information from known sources such as the internet and administrative records; Estimate from benchmarks such as prior surveys or censuses; Validate estimates with field operations; Evaluate and describe measurement differences; Analyze sources of error and determine if those data sources should be used in the future; and Link quality measures and inferences back to improve the projection methodology. Rather than using estimates to weight current surveys, surveys validate current estimates.

Keywords: Big Data; Official statistics; Method; Estimate.

1. Introduction

The current survey based infrastructure of national statistical agencies is not adequate to meet the timeliness and detailed coverage of domains (e.g., geography, industry, product etc.) and flexibility of informational products to meet emerging societal and policy demands. New approaches to the collection, processing, analysis, and dissemination of official statistics are required to develop relevant and cost-effective statistics.

We are surrounded by a sea of digital information know as Big Data that exists in both structured and unstructured forms including sensor, web, social media, cell phone data and direct feeds including sources referred to as the Internet of Things. Every day, we and our stakeholders are presented with new informational products created by new technologies and processes that ingest, store, query, and analyze all of this information. From the perspective of a national statistical agency, Big Data includes information that is difficult to collect, store or process within conventional systems; with volume, velocity, structure or variety that require new statistical software processing techniques and IT infrastructure. These enable relevant, transparent, cost-effective insights to be made while maintaining the confidentiality of the provider[i].

The challenge faced by statistical agencies is how to manage stakeholder expectations that, “Big Data will solve all the world’s problems,” and “why haven’t we already developed the solutions?” Big Data sources and techniques present promise for reducing the cost, improving the relevance and timeliness, and increasing the content and quality of future official statistics. As national statistical agencies we have a legal commitment to produce unbiased, “gold-standard” information. While many companies

[1] Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau

have touted success in using Big Data and its techniques to support decision-making, improve products and customer support, and generate revenue, does their use of the information meet the rigors of accuracy and transparency expected by the stakeholders of official statistics? The presence of several poorly estimated date of points is, from a methodological viewpoint, to be expected and tolerated. However, from an administrative viewpoint, even a few poor estimates can be extremely troublesome because official estimates are usually found in conjunction with resource allocation or service planning. The presence of only a few areas that are frequently estimated poorly can lead to administrative and even legal entitlements for the group responsible for producing estimates.

2. Further Defining the Challenges

Using Big Data as a source for official statistics has many commonalities to approaches developed decades ago when the Census Bureau began using administrative records from federal, state, and local agencies to create population statistics. At that time we were constrained by technology that was not able to effectively receive and process microdata records from providers. Instead we often relied on aggregate statistics provided by those sources. These aggregated statistics included registered events such as the number of births, deaths, marriages, automobiles, driver's licenses, voters' employment statistics, taxes collected, and school enrollments. The incentive for data providers to forward their statistics was driven by return on investment. In the United States, there are laws and regulations for the use of statistics to accurately and fairly distribute federal resources and measure the impact of public policy. For example, in FY2008, U.S. federal domestic assistance programs used Census data to guide the distribution of $446.7 billion, 31% of all federal assistance. Census guided grants accounted for $419.8 billion, 75% of all federal grant funding[ii].

As technological ability has expanded so has the amount of information produced each day. IBM has estimated that each day we create 2.5 quintillion bytes of information[iii] with much of this information created and stored in private sector repositories. Harnessing this information, the majority of which is in an unstructured format, is a challenge. Unlike administrative records submitted by governmental entities, the value proposition for providing data is different for private sources. Statistical agencies currently provide value to private for profit companies through:

- Developing independent historical baselines upon which the quality and trends in real-time information can be assessed;
- Providing statistics that indirectly reduce the costs of bringing goods to market – such as road improvements based on traffic statistics, or determining the location of distribution centers to better and more quickly deliver goods and services.
- Providing federal, state, and local governments with information to fuel economic development programs designed to attract and develop businesses.
- Direct purchase of data products.

There is no single governmental or non-governmental repository that can be accessed to capture all the data points and elements statistical agencies require to accomplish their missions. Individual companies may be concerned about providing information that they deem confidential and not normally shared outside of their firms. Data are valuable resources and sharing those resources with any external entity presents a risk of loss that must be calculated against value gained. Additionally, private companies share certain information with groups such as trade associations who are developing sophisticated measuring and monitoring system and returning statistics in near real time. The question remains: What can statistical agencies do to incentivize businesses to provide the critical source information needed?

3. Meeting the Challenges

Big Data sources and techniques are directly applicable to understanding and improving the collection and quality of traditional information. In response to stakeholder needs the Census Bureau is developing the Center for Statistical Modernization and Data Innovation (CSMDI). This center will

research the applicability of Big Data sources and techniques to support a 21st century data driven nation with the next generation of official economic and social statistics. The CSMDI will focus on several core elements required to redesign statistical operations:

- Methodological - to produce more timely and relevant, scientifically valid estimates and uncertainty measures of economic and social statistics from data collected from a wide variety of sources, most of which were not designed to produce inputs to the production of official statistics. Empirical and methodological research on linkage, bias, variance, new estimation procedures and total survey error measurement. Many new potential insights can be gained through the merger of Big Data sources such as new and more timely measures of business productivity and supply chains, intra and international trade, retail sales, housing trends, daytime populations, and social mobility, and more geographically detailed legacy products to name a few
- Computational - to acquire and develop the hardware, software, and human capital essential for collecting, computing and disseminating statistics constructed from a variety of sources including surveys, administrative sources, transaction data, social media, sensors, and satellites.
- Policy - to secure legal permissions and stakeholder buy-in to utilize non-traditional sources of data for the production of official statistics. Requirements include legal agreements with data providers, and engaging the complete set of stakeholders in the legal and privacy space in a transparent way to ensure all understand the cost, benefits and risks.
- Transparency – to provide the methodological understanding and access to statistics for external validation of results while maintaining the confidentiality of source information.
- Consistency – to provide a means to create accurate, consistent, and trusted longitudinal statistics from a wide variety of volatile data sources.
- Outreach and marketing – to market new products, manage customer expectations, and foster customer satisfaction by educating users on how to best draw inferences from estimates constructed in novel ways that may vary widely from traditional survey based statistics. Also, educate consumers on the quality and value of products created by statistical processes from the wide variety of information available to them through other sources.

In addition to developing the organizational and information technology infrastructure the Census Bureau has developed classes in Big Data techniques and is in the process of conducting several exploratory projects. Specifically, the use of Big Data is being examined to:

- Supplement existing monthly/annual retail surveys;
- Support the Commodity Flow Survey;
- Measure or improve housing and educational statistics in demographic surveys; and
- Create new measures of innovation and assessing its impact on economic growth by integrating university data on federally funded grants with Census Bureau data assets.

The larger goal of the Census Bureau is to incrementally modernize economic statistics by the year 2027. Accomplishing this requires Big Data and its techniques to i) re-engineer the Business Register to improve the integration of the Bureau's major administrative records sources on business activities both across source and over time, ii) move as much direct data collection from businesses from surveys/censuses to passive modes to reduce steady state reporting burden, improve timeliness, expand the measures produced and improve quality, and iii) utilize surveys more strategically to focus on items and populations not available from administrative or business records.

By increasing both the scope and timeliness of data obtained through automated passive collections, the Census Bureau will be in a position to address the two shortcomings of our economic statistics most often cited by our users. First, by collecting and processing data in near real-time, the Census Bureau will be in a position to improve the latency of both its current and periodic economic statistics.

Second, by moving to passive collection modes that can cost-effectively intake data from a much broader swath of the business universe, the Census Bureau can produce near real-time economic statistics for more detailed domains (e.g., geographic areas, industries, business characteristics, etc.). These timelier and more detailed data will be very useful, possibly key to decision makers in federal, state and especially local governments. Perhaps more importantly, the broad and timely coverage of the business sector will position the Bureau for the first time to be able to offer a range of data products (e.g., benchmarking reports, local labor marker analyses) for businesses.

4. Some lessons we are learning

Given the massive volume of information currently available, it is fair to say that no one organization possesses the computing power and human capital necessary to understand, acquire, ingest, and process the domains of information required to support the wide array of information produced by statistical agencies. Therefore, traditional methods used to acquire, process, and manage information must be changed. Further, there are numerous sources of information available on the internet and many of them contain restrictions on the use of web scraping to acquire their information. We must also develop new standard quantitative measures that depict the reliability and uncertainty not only of single data sources, but the combination of data sources and the methodologies used to create final products.

This process begins by developing a standardized scorecard that can rapidly assess whether a data source is both applicable and fit for use. This scorecard should including the following questions:

- What is the universe, time period referenced, lag time before availability, period of recurrence, and geographic detail of the information available?
- If this is a recurring product, have any universe or element definitions changed over time?
- Does the information represent an event, action, or a sentiment?
- Is metadata available that describes the data collection, edit, and imputation processes used to create the source information?
- How often is information missing? Are only certain data elements considered important or accurate to meet current user needs? What quality control procedures have been applied?
- Have the data been provided to others (this information is required to remove duplicate sources of information that initially appear to be independent).
- Are the data available as aggregated tallies or individual records, and do the records contain unique identifiers?
- Are there any restrictions to use, including legal, privacy, or ethical concerns?
- Is the information cost prohibitive to acquire?
- Are there constraints regarding the ownership, intellectual property rights, retention, reuse, and redistribution of the information?

It is time consuming to review all the potential sources of information available, however, there seems to be the early signs of a new business model statistical collection. Statistical agencies will need to increasingly partner with both upstream and downstream providers to produce economic statistics and to provide analysis and interpretation of their own products and those produced by others.

From the work the Census Bureau has been conducting on its exploratory projects it seems that the best approach to furthering the acquisition and use of Big Data is to develop partnerships with information clearinghouses and trusted sources such as universities, trade associations, and private data vendors who focus on understanding specific information domains. Their ability to understand and preprocess information is likely key to addressing the volume, velocity and volatility of Big Data.

As companies look to monetize their data assets, an independent scorecard or clearinghouse is needed to assist consumers in their understanding of the content, coverage, quality, and timing of numerous

specific data products. If there were such an honest broker, data consumers could acquire and use information more effectively, and market forces would increase the accuracy and coverage of the underlying information, because data would no longer be considered mere by-products.

5. Defining Emerging Best Practices

The traditional approach to creating statistics usually includes the use of administrative records or prior field data collections to develop survey frames, mail-out of forms; conduct field follow-up operations to collect respondent information, tallying of unweighted results, and the application of controls to produce estimates. Where do these controls come from? As an example, for decades demographic surveys have used population estimates with characteristics to weight survey results. These population estimates have been generated by the collection of aggregate and microdata administrative records from federal, state, and local agencies and the application of a wide variety of methodologies to develop the resultant estimates[iv]. They are benchmarked against prior data collections, reviewed for accuracy and employed to weight the final survey results.

Statistical agencies are facing increased requirements for the production of statistics within the realm of constrained budgets and falling response rates in both business and household surveys. The most costly components of any survey are its mail-out and field operations. If we have to reduce the cost of a survey by drastically reducing its sample size, variance measures significantly increase and accuracy decreases. Additionally, if a new survey is required, it takes years to develop the process, questionnaires, develop modes of delivering the questionnaires to respondents, conduct field operations, analyze and distribute the results. Consequently, we cannot continue to produce statistics as we have in the past. Is there another approach? Rather than using microdata administrative records to control survey results, perhaps survey delivery and field operations can be used to control, validate, correct for biases, and enhance the estimates produced from administrative records and Big Data sources.

The following approach (REVEAL) leverages the strength statistical agencies possess – the ability to measure ground truth:

R Reassemble information from known sources such as the internet and administrative records;
E Estimate from benchmarks such as prior surveys or censuses;
V Validate estimates with field operations;
E Evaluate and describe measurement differences;
A Analyze sources of error and determine if those data sources should be used in the future; and
L Link quality measures and inferences back to improve the estimation/projection methodology.

REVEAL raises a host of questions that must be addressed:

- How can big data techniques such as propensity modeling (or as the private sector calls it, recommendation systems) be used for adaptive survey deployment -- saving time and money by not sending people to houses when no one is home?
- How can one determine that sufficient “respondents” have been collected to satisfy statistical requirements so that collection can halt on that sub-population, again reducing costs?
- How can established sampling/survey methods be used to detect and/or correct for bias from big data sources? Should surveys include additional questions to provide a basis for measuring bias (such as mobile device use, do you have smart meters on your home etc.)?
- What are the best methods for ensuring sufficient transparency and replicability with big data sources? How can the reliability, authenticity, and provenance of official statistics be maintained when integrating big data from outside sources?
- How can distributed processing and continuous data management techniques be used to support continuous estimation? How can data review algorithms, be automated, informing daily field survey tactics, reducing cost and improving quality?

- How can new methods of estimation be communicated effectively to stakeholders and users? How do you communicate reliability?
- How can we effectively manage a whole host of data quality issues – including changes in collection processes and data definitions over which we have no control, missing information, imputed or estimated data whose methodology is protected as a trade secret, and the capacity for a provider to intentionally change information in order to achieve a desired outcome such as moving a market.
- How can we be transparent in our methodologies and data releases while controlling for the Mosaic effect – the effect of combining and potentially releasing seemingly innocuous information items that can be used to re-identify an individual person or business[v][vi]? The limitation of disclosure is particularly difficult when a data provider such as a private company possesses a subset of the components used to create the statistical product.

Many answers have been unfolding, and have generated numerous questions and opportunities. We are faced with exciting possibilities to improve statistical systems, better serve customer data needs, and focus the improvement of statistical methods beyond the current focus on survey coverage and variance to one of measuring bias and relevance.

6. References

[i] Adapted from ”How Big is Big Data, Exploring the Role of Big Data in Official Statistics Version 0.1”, United Nations Economic Commission for Europe, on behalf of the international statistical community, March 2014.
http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307

[ii] “Counting for Dollars: The Role of the Decennial Census in the Distribution of Federal Funds.” Andrew Reamer Metropolitan Policy Program, Brookings Institute, March 2010
http://www.brookings.edu/~/media/Research/Files/Reports/2010/3/09-census-dollars/0309_census_report.PDF

[iii]“Big Data are you ready for Blastoff.” Mathew Wall, Business Reporter, BBC News March 4, 2014
http://www.bbc.com/news/business-26383058

[iv] Evaluating the Utility of Population Estimates as the Basis for Distributing Federal Revenues. Bowling Green State University doctoral dissertation, Ronald Prevost, December 1991.

[v] Whitehouse Warns of Open Data Mosaic Effect, Molly Bernhardt Walker, Fierce Government IT, May 2013
http://www.fiercegovernmentit.com/story/white-house-warns-open-data-mosaic-effect/2013-05-14

[vi] The Mosaic Effect and Big Data, Adam Mazmanian, Federal Computer Week, May 2014
http://fcw.com/articles/2014/05/13/fose-mosaic.aspx