Opportunities and Challenges for using Big Administrative Data November 2016

David M. Talan

U.S. Bureau of Labor Statistics

2 Massachusetts Ave., N.E., Washington D.C., 20212

Abstract

The demand for information, insights and data is increasing in every segment of society. The ability to respond to ever increasing demand forces government statistical agencies to search for new opportunities and new methods.

In these days of instantaneous demands, waiting a few years to build new data sources, investing valuable staff to the new development, then waiting for a sufficient time series to build for analysis, is usually an insufficient response to the original demand. If there is any other way to gain sufficient insight into a problem without waiting for several years to get a few answers to basic items, data users will gravitate to other, immediately available sources.

The Bureau of Labor Statistics' Quarterly Census of Employment and Wages program produces a dataset that some consider Big Data. There is no standard agreed upon definition of Big Data, but it is often defined as a data set with the following dimensions: volume, velocity, variety, and veracity. While the QCEW does not have the velocity of some other datasets, it certainly does have the volume, variety, and veracity that characterize Big Data.

This paper describes the opportunities and challenges for using Big Data at BLS through its Business Register. We present some opportunities for developing new products such as recently released hurricane zone data, and employment and wage estimates for non-profits; and also prospects for employment estimates from foreign direct investment and other opportunities. We also present some challenges of using big data that include statistical, legal, and technical infrastructure issues. We also present the evolving nature of Big Data and describe how it may yield promising areas of future development.

Key Words: Bureau of Labor Statistics, big data, administrative data, matching

1. Motivation

The demand for information, insights and data is increasing in every segment of society. The ability to respond to this ever increasing demand has forced government statistical agencies to search for new opportunities and methods.

It is well recognized that the ability to quickly change existing surveys or mount new surveys to collect new data is limited. Adding questions to existing surveys not only takes time and resources, but also adds to respondent burden. Building new surveys is costly, and can take several years. For example, the Bureau of Labor Statistics (BLS) monthly Job Openings and Labor Turnover Survey (JOLTS) was funded in late 1998 and the first results for February 2004 were published in April 2004, nearly five years later. The BLS Green Goods and Service (GGS) survey was funded in FY 2010 and the first annual results were published in March 2012, two and a half years later.

In a world of instant gratification, taking years to develop new data sources, and then waiting for a sufficient time series for analysis, is an insufficient response. Modern data users gravitate toward immediately available, albeit sometimes incomplete, data sources.

Big Data concepts are based, in part, on using available data in new ways to address existing and evolving information needs.

The Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) program produces a dataset that some consider Big Data. There is no single agreed upon definition of Big Data, but it is often defined as a data set with the following dimensions: volume, velocity, variety, and veracity. While the QCEW does not have the velocity of some other datasets, it certainly does satisfy the other dimensions that characterize Big Data.

This paper describes the opportunities and challenges for using Big Data at BLS through its Business Register. The second section of this paper presents various definitions of big data and provides background on the BLS Business Register. The third section describes opportunities for using big data within BLS and incorporating it into the Business Register. Potential opportunities include developing new products such as hurricane zone data; employment and wage estimates for non-profits; and employment estimates from foreign direct investment. Section four describes challenges of using big data which include statistical, legal, and technical infrastructure issues. The last section describes the evolving nature of Big Data and how experimenting with it may yield promising sources of future development.

2. Definitions of Big Data

There is no commonly agreed upon definition of Big Data. Wikipedia describes Big Data as "a term for data sets that are so large or complex that traditional data processing applications are inadequate." (Wikipedia, 2016)

Another often cited definition of big data is the 3Vs model: volume, variety, and velocity. This definition was first developed in a 2001 research paper by META Group (now Gartner) analyst Douglas Laney where he described the challenges and opportunities of data growth as being three-dimensional: increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). The Gartner group and much of the technology industry use this characteristic-focused definition to describe big data.

Others have since added a fourth "V" to the 3Vs model, veracity. Veracity is an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions. (Villanova, 2014).

Some statisticians have referred to Big Data as non-sampled data, characterized by the creation of databases from electronic sources, whose primary purpose is something other than statistical inference. Others refer to this as "unstructured data," where data is created for another purpose, but can be used to create something else (examples include online job vacancies).

Big Data can also include text in addition to traditional data. For example, the text on the Internet grows every day. This text can be mined as Big Data. The data generated from sources such as weather related devices, financial transaction logs, biomedical logs, radio-frequency identification (RFID) readers, and wireless sensor networks, etc. may be considered to be examples of Big Data.

Another way to use Big Data is to leverage existing data by using it in new ways. For example, matching existing data to other sources may allow new data products and economic information to be produced at low cost or with no respondent burden. Several of these approaches will be discussed below.

The discussion of Big Data also includes the use of paradata. Paradata is data about the data that can be used to streamline operations, optimize resource use and burden reduction. Paradata has been used for many years without the term. For survey operations, and in a simple example, mailings might be organized by postal area to maximize postal discounts. Telephone surveys target businesses only during business hours and call scheduling accounts for time zones. Workload planning may also include call times or the number of call attempts before declaring a refusal or shifting strategies to "reluctance" approaches.

Another way of adding value to existing data is to improve visualization techniques. Graphing and mapping are longstanding visualization tools. More sophisticated applications are increasingly used to make information understandable in new ways. The challenge is in developing staff with the vision and skills to create suitable outputs for existing and new data.

One of the motivations to direct resources to the Big Data concept is the high cost and relative rigidity of surveys. Surveys are often developed to focus on measuring a single, targeted concept.

Monthly surveys are, by definition, short and focused due to the overall goal of gathering and publishing timely data. The BLS Current Employment Statistics (CES) survey has been finely honed to provide the earliest measure of business activity determined by employment and hours worked. These data are released 2 to 3 weeks after the reference period. Any added variables, regardless of usefulness, may challenge this eagerly anticipated, closely scrutinized, and timely publication.

Similarly, the BLS Job Openings and Labor Turnover Survey (JOLTS) collects job openings, hires, quits and layoffs with near-CES timeliness. Economists have suggested the usefulness of adding more questions to the JOLTS collection. For example, adding questions on the duration of vacancies may support insights into the tightness of the labor market. Allowing employers to differentiate between replacement hires and new positions would provide evidence of employers' growth intentions. Adding occupation of hires would be a useful indicator of occupational demand and wage pressure. Despite usefulness, it must also be considered whether current respondents would have easy access to these new data items.

In both the CES and JOLTS surveys, however, adding questions to gather new information would necessitate changing forms, systems, and possible changes to sample design. While these changes are not insurmountable, they would require new resources that are currently not available.

Business Registers offer advantages over surveys as a platform for Big Data initiatives. For example, the wide coverage of the Business Register does not require sample plan changes. Business Registers usually include a number of identifiers that can be used to match to other datasets. Tax systems require unique identifiers that then become widely used for various purposes. In the U.S., the Internal Revenue Service (IRS) provides Employer Identification Numbers (EIN) to every business for federal tax purposes. Other entities also include the EIN in data collections, making the EIN an extremely valuable tool for matching administrative datasets.

The BLS business register is the Quarterly Census of Employment and Wages (QCEW). It provides monthly employment and quarterly wages for all businesses with employees covered by Unemployment Insurance (UI), about 98% of total employment. The UI administrative data is augmented by two additional collections. First, the Annual Refiling Survey (ARS) updates industry, geography, and respondent information on roughly a three-year cycle. The Multiple Worksite Report (MWR) captures employment and wage data for individual establishments under multi-unit businesses, allowing the QCEW to locate establishments in the correct industry and local area. These three sources, combined with intense review and editing of the data, provide the best and most current universe for sampling and employment benchmarks. The QCEW is published within 6 months of each reference quarter.

Four “V’s”:

For the purposes of this paper related to Business Registers, the Big Data concept will focus on the four “V’s”.

Volume: Business Registers are by definition voluminous covering all eligible business establishments and/or firms. The QCEW holds 9.6 million establishment records per quarter. The entirety of the longitudinally linked dataset contains 850 million records from first quarter 1990 through fourth quarter 2015. This number grows with the economy and time.

Variety: Business Registers can cover a large number of data items providing a rich source of analysis on many dimensions.

Velocity: Velocity refers to the frequency and timeliness of the data. The QCEW is produced quarterly and is published 6 months after the reference period. Few Business Registers can match these characteristics.

Veracity: Veracity is also a strength of Business Registers. Mandatory reporting through tax agencies helps to provide complete reporting, but does not ensure accuracy. Post collection editing and review can improve accuracy. In the case of the QCEW, a range of edits and validating measures, including direct respondent re-contact and supplemental collections, are applied to transform these administrative data into highly accurate and reliable economic statistics.

3. BLS Uses of Big Data

While Big Data is a relatively new term, the concepts behind it are not. The QCEW, for example, is the sample universe for all BLS business surveys. Probability sampling insures that survey responses represent the universe. In addition, it is the employment benchmark for the Current Employment Statistics (CES) survey, the Occupations Employment Statistics (OES) survey, and others. As the employment benchmark, the vast coverage, mandatory reporting and intense editing of the QCEW allows surveys to leverage the strength and robustness of the QCEW universe.

Business Demography: One of the first uses of Big Data by BLS was the development of the Business Employment Dynamics (BED) data from the QCEW Business Register. BED is the BLS term for business demography. The BED data are created by linking individual business register establishment records longitudinally. These records are then tabulated to create aggregate time series for business establishment openings, closings, expansions and contractions. Additional dimensions of these data have been created to include establishment age and survival data, along with firm size data. These data have gained a wide user base where economists and data users examine the role of employment dynamics in the U.S. economy.

The BED series were first published in 2003. Because it was built on the QCEW Business Register, the BED series were constructed to start from 1993, providing an "instant" decade of data for analysis. Any new series built on the QCEW history may be relatively inexpensively developed to show multiple years worth of analysis and classification.

The creation of the BED data elements – age, size, birth, death, industry, geography, etc. – essentially add new variables to the business register – adding additional variety.

The easiest way to add Variety is to merge the existing Business Register with other administrative datasets. Assuming one or more common identifiers, matching records can add new data elements. Identifiers include, but are not limited to, Employer Identification Number (EIN), business name, and address.

Non-Profit Economy: The QCEW program has been matched to a publicly available list of non-profit businesses. These new research data covering 2007-2012 were released in September 2014 meeting a longstanding need for current, detailed industry and geographic detail on this sector of the economy. The non-profit sector covers about 10% of employment and has higher than average wages, making it an important to understand segment of the economy.

Geospatial Hurricane Zones: The QCEW microdata contain physical location addresses that have been geocoded to create precise points on the earth. Recently, QCEW staff completed a project to overlay existing hurricane evacuation zones (polygons) over the geocoded business locations. Maps and tables have been published on the BLS website (http://www.bls.gov/cew/hurricane_zones/home.htm) showing the number of establishments, and the accompanying employment and wages that are exposed to potential damage under hurricane conditions of varying strengths.

Foreign Direct Investment: Other existing datasets include mandatory reports from the Bureau of Economic Analysis's Survey on Foreign Direct Investment (FDI). A similar effort by BLS to match and review records will yield a research data series tabulated by industry and state. FDI data is in demand for many uses, including local economic development efforts, where interest lies in targeting established foreign investments from specific countries.

With a pilot test underway, it may be possible to create, even with limited resources, a short historical series, again building on the linked longitudinal features of the QCEW.

Comment Codes: The local state QCEW staff reviews thousands of records each quarter and re-contacts respondents to clarify data, resulting in a large proportion of those records with significant changes. The results of these contacts are recorded as "comment codes" explaining these movements. Examples of such comments are:

- Strike, lockout, or other labor dispute
- More business/less business
- Layoffs
- New business establishment
- Fire disruption
- Natural disaster disruption
- Non-natural disaster disruption
- Energy shortage

Further analysis of the patterns of existing codes (by size, age, expanding or contracting) may point to business expectations, future hiring or layoff events.

Also, as a part of the QCEW data stream, analysts learn about whether a business was bought or sold, or a part was involved is such a transaction and are looking at ways to study patterns in these events to learn more about the dynamics of businesses, employment and wages.

The US federal government has a number of administrative data sets that might, if merged or matched with other datasets yield new information at low costs. However, many of these have restricted access regulations. The Office of Management and Budget (OMB) is beginning a process of documenting the range of potential uses and benefits of linking the existing array of administrative datasets with an apparent eye toward attempting to change laws and other barriers, paving the way for fully leveraging these opportunities in the future

The remaining barriers include staff resources, linking methods and perhaps new techniques.

Local Hires: The National Directory of New Hires (NDNH) was constructed to allow the tracking of parents owing a variety of payments to their children. Businesses are required to report information on new hires within a few weeks to state agencies, which is then forwarded to the Department of Health and Human Services to allow cross state tracking and reduce avoidance of child support payments. Current laws severely restrict access to these new hire data records. If laws were changed to allow data sharing with statistical agencies, there is potential to link these business records to the QCEW, for example, by common identifiers; this would create data on hires and wages at low levels of geography and industry, far more detailed than would be possible with a sample expansion to JOLTS.

Entrepreneurship: The Internal Revenue Service (IRS) collects vast amounts of information on businesses. These data are also severely restricted to a small set of agencies for specific uses. The Census Bureau is allowed to obtain these data, but not BLS. Decades of attempts to change the laws have failed. If BLS were to have the same access as Census, research and related analysis may improve the consistency of U.S. statistics. Furthermore, access to IRS data on the self-employed, currently not covered in the BLS Business Register or business surveys, may allow for new studies on entrepreneurship and the business creation process.

4. Challenges for using Big Data

The challenges for statistical agencies using Big Data include data capture, storage, sharing, analysis and visualization (Wikipedia, 2016).

One critical challenge is building the staff who can manipulate, merge and link large datasets and bring the needed curiosity to continuously seek new insights, tool, techniques and statistical tools to new and evolving situations (Royster, 2013).

Can the Business Register community learn how to quantify the characteristics of other datasets in ways that give us the confidence to cite results of analysis? If a dataset has the same distributions as another proven source, can we use it? Or is it enough to provide the dataset characteristics and cite results, letting the user make decisions about suitability?

The community of official government statistics gains it credibility based on professionalism, transparency and sound statistics. We have to maintain those features as we move into Big Data. The challenge is to apply good, sound statistics to new data sets. It will take some years of development and research to learn how to describe and justify other methods than those we have relied on for 50 years.

5. Future Developments using Big Data using the Business Register

The scope of Big Data applications is being evaluated at BLS and other agencies. Specifically for the QCEW Business Register, a number of initiatives have been completed, are in process, or are being evaluated.

First, BLS is linking the QCEW to available datasets as resources allow. As these research projects are attempted and completed, lessons are learned, techniques refined and analyses improved. Resources used on these pilot projects are useful in training staff, developing new techniques and products, working with users and growing the number and types of products on the U.S. economy. For each project listed above, there are known users; however, it is possible that some new products do not find enough customers to justify continued production. Resources are then moved to other opportunities.

The insights, linking approaches, and other aspects of learning by doing with Big Data will inform BLS on opportunities, train staff, and build infrastructure for future work.

The new era of Big Data is creating interest in Business Registers and their specific characteristics and potential to contribute. It is incumbent on the Business Register community to profile and market Business Register capabilities during this opportunity. The success in building Big Data capabilities, tools, curious staff and new products will determine the value of the community's efforts.

Any opinions expressed in this paper are those of the author and do not constitute policy of the U.S. Bureau of Labor Statistics.

References

Business Employment Dynamics website: http://www.bls.gov/bdm/

Clayton, Richard; James Spletzer, John Wohlford. "Conference Report: JOLTS Symposium," Monthly Labor Review, February 2011.

Horrigan, Michael. Big Data: A Perspective from the BLS, Amstat News, January, 2013.

Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety." Gartner. Retrieved 6 February 2001.

OECD, EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH Mapping the Policy Issues Raised by "Big Data," June 2013.

Quarterly Census of Employment and Wages website: http://www.bls.gov/cew/home.htm

Royster, Sara. "Working with Big Data", Occupational Outlook Quarterly, Fall 2013.

What Does "Big Data" Mean for Official Statistics? United Nations Economic Commission for Europe, Conference of European Statisticians.

“What is Big Data?" Villanova University: http://www.villanovau.com/resources/bi/what-is-big-data/#.U85_uPbD-mc

Wikipedia: http://en.wikipedia.org/wiki/Big_data (viewed July 12, 2016)