Editor's Introduction (August 2, 2020)

Big Data for 21st Century Economic Statistics: The Future Is Now

Katharine G. Abraham, Ron S. Jarmin, Brian C. Moyer and Matthew D. Shapiro

The infrastructure and methods for official U.S. economic statistics arose in large part from the federal government's need to respond to the Great Depression and Second World War. As it emerged from the war, the U.S. economy was heavily goods based, with more than a third of employment in manufacturing. Although censuses of manufacturing activity had been undertaken as early as 1810, the first comprehensive quinquennial economic census was conducted in 1954. Economic census data provide the backbone for the measurement of nominal economic activity in the national income and product accounts. Surveys based on probability samples that were developed in the period after World War II collect accurate statistics at lower cost than complete enumerations and make central contributions to high-frequency measurements. Administrative data, especially data on income from tax records, play an important role in the construction of the income side of the accounts and in imputing missing data on the product side.

The deflators used to construct estimates of real product were developed separately from the measurement system for nominal income and product. The earliest Consumer Price Index (CPI) was introduced in 1919 as a cost of living index for deflating wages. The CPI and Producer Price Index (PPI) programs provide the price measurements that are used to convert nominal measures into estimates of real product.[1]

This measurement infrastructure, established mostly in the middle part of the 20th century, proved durable as well as valuable not only to the federal government but also to a range of other decision makers and the research community. Spread across multiple agencies with separate areas of responsibility, however, it is less than ideal for providing consistent and comprehensive measurements of prices and quantities. Moreover, as has been noted by a number of commentators, the data landscape has changed in fundamental ways since the existing infrastructure was developed. Obtaining survey responses has become increasingly difficult and response rates have fallen markedly, raising concerns about the quality of the resulting data (see e.g. Baruch and Holtom 2008, Groves 2011, Meyer, Mok and Sullivan 2015). At the same time, the economy has become more complex and users are demanding ever more timely and more granular data.

In this new environment, there is increasing interest in alternative sources of data that might allow the economic statistics agencies to better address users' demands for information. As discussed by Bostic, Jarmin and Moyer (2016), Bean (2016), Groves and Harris-Kojetin (2017),

[1] See Carson (1975) and Goldbert and Moye (1985) for discussions of the development of the existing infrastructure for the production of economic statistics.

and Jarmin (2019), among others, recent years have seen a proliferation of natively digital data that have enormous potential for improving economic statistics. These include item-level transactional data on price and quantity from retail scanners or companies' internal systems, credit card records, bank account records, payroll records and insurance records compiled for private business purposes; data automatically recorded by sensors or mobile devices; and a growing variety of data that can be obtained from websites and social media platforms. Incorporating these non-designed Big Data sources into the economic measurement infrastructure holds the promise of allowing the statistical agencies to produce more accurate, more timely and more disaggregated statistics, with lower burden for data providers and perhaps even at lower cost for the statistical agencies. The agencies already have begun to make use of novel data to augment traditional data sources. More fundamentally the availability of new sources of data offers the opportunity to redesign the underlying architecture of official statistics.

In March 2019, with support from the Alfred P. Sloan Foundation, the Conference on Research in Income and Wealth convened a meeting held in Bethesda, Maryland to explore the latest research on the deployment of "Big Data" to solve both existing and novel challenges in economic measurement. The papers presented at the conference demonstrate that Big Data together with modern data science tools can contribute significantly and systematically to our understanding of the economy.

Some of the same themes were explored at a CRIW conference on *Scanner Data and Price Indexes* organized by Robert Feenstra and Matthew Shapiro held in the fall of 2000 in Arlington, Virginia. Authors at the earlier conference examined use of retail transaction data for price measurement. While there was considerable interest at that time in this new source of data, many of the papers pointed to problems in implementation and with the performance of the resulting measures. Research continued, but for a variety of reasons, innovations in official statistics to make use of the new data were slow to follow.

Twenty years on, the papers in this volume highlight applications of alternative data and new methods to a range of economic measurement topics. An important contribution to the conference was the keynote address given by then Statistics Netherlands Director General Dr. Tjark Tjin-A-Tsoi. He reported on that agency's impressive progress in supplementing and replacing traditional surveys with alternative Big Data sources for their statistical programs. Notwithstanding the issues and challenges that remain to be tackled to realize the full potential of Big Data for economic measurement at scale, there was much enthusiasm among the conference participants regarding their promise.

The message of the papers in this volume is that Big Data are ripe for incorporation into the production of official statistics. In contrast to the situation two decades ago, modern data science methods for using Big Data have advanced sufficiently to make the more systematic incorporation of these data into official statistics feasible. Indeed, considering the threats to the current measurement model arising from falling survey response rates, increased survey costs and the growing difficulties of keeping pace with a rapidly changing economy, fundamental changes in the architecture of the statistical system will be necessary to maintain the quality and utility of official statistics. Statistical agencies have little choice but to engage in the hard work and significant investments necessary to incorporate the types of data and measurement approaches studied in this volume into their routine production of official economic statistics.

While the challenges are significant, the COVID-19 crisis has driven home the importance of modernizing the federal data infrastructure by incorporating these new sources of data. In a crisis, accurate and timely data are of critical importance. The high frequency information by location and type of activity that Big Data make possible already have proven to be highly valuable. For example, location data from smartphones have provided timely and detailed information on the response of aggregate activity to the unfolding health crisis (Google 2020, University of Maryland 2020). Based on data incorporated from variety of private sources, Opportunity Insight's Economic Tracker is providing decision makers with weekly indices of employment, earnings and consumer spending (Chetty et al. 2020). Statistical agencies also responded nimbly to the crisis. For example, in addition to introducing two new Pulse Surveys providing important information on the response of households (Fields et. al. 2020) and businesses (Buffington et. al., 2020) to the crisis, the Census Bureau released a new measure of weekly business creation based on administrative data. Unfortunately, the use of Big Data for real time granular economic measurement is still in a nascent state, with the infrastructure for systematically constructing key economic statistics based on robust and representative Big Data sources not yet developed. Rather, as the proliferation of new working papers using novel data sources attests, there has been a blossoming of uncoordinated measurement efforts that capture particular aspects of the pandemic's economic impact. Our hope would be that, at the point when the American economy experiences any future crisis, the statistical agencies, making use of near-real-time Big Data, will be prepared to provide systematic and comprehensive information that policy makers can use to guide essential decisions

The Promise of Big Data for Economic Measurement

As already noted, the current infrastructure for economic measurement has been largely in place since the mid-20th century. While organized in various ways, with some countries adopting a centralized model (e.g., Canada) and others a decentralized one (e.g., the United States), official economic measurement typically uses a mix of data sourced from sample surveys, government administrative records and periodic censuses to support key statistics on output, prices, employment, productivity and so on. For decades, as the primary collectors, processors, and curators of the raw information underlying economic statistics, government statistical offices were near monopoly providers of this information. Organizations such as the Census Bureau and the Bureau of Labor Statistics (BLS) collected information from household interviews or paper questionnaires completed by business survey respondents based on company records. In many cases, the information was digitized only when it was entered in the statistical agencies' computers. Today, in contrast, staggering volumes of digital information relevant to measuring and understanding the economy are generated each second by an increasing array of devices that monitor transactions and business processes as well as track the activities of workers and consumers.

The private sector is now the primary collector, processor and curator of the vast majority of the raw information that potentially could be utilized to produce official economic statistics. For example, the information systems of most retailers permit tracking sales by detailed product and location in near real time. The private sector also is increasingly a disseminator of economic statistics to the public as in the case of ADP's monthly employment report, the Conference Board's Help Wanted Online publications, and the statistical information produced by the JPMorgan Chase Institute. Though often extraordinarily rich, many of the new sources of data lack representativeness, covering only particular subpopulations such as the businesses that use a particular payroll service or customers of a particular bank. In addition, exactly what is collected is outside of the agencies' control. These considerations point in the direction of a blended survey-Big Data model for incorporating new sources of information into official statistics. Finding ways to do this effectively holds the promise of vastly more timely and detailed economic statistics. To be clear, we do not suggest that official statisticians should want to produce estimates of Cheerios sold in Topeka last week. Rather, we believe it is possible to do much better than producing only monthly estimates of retail sales at the national level as is currently done.

Even for official economic indicators that continue to be published at a monthly or quarterly frequency, access to timely information across wide swaths of economic activity can help reduce revisions in the statistics. The estimates of Gross Domestic Product produced by the Bureau of Economic Analysis go through multiple rounds of sometimes substantial revisions, largely because the information that undergirds the initial estimates is sparse and better information arrives only with a substantial delay. These revisions can cause significant problems for users of the data. Recent research, including papers in this volume, shows that even incomplete information from private sources available on a timely basis can help with producing better initial estimates that are less subject to later revision.

In addition, Big Data can support the production of more granular statistics than are possible using survey data. Users of economic statistics commonly express a particular interest in geographically disaggregated information. State and local agency representatives who met with members of a recent Committee on National Statistics panel reviewing the Census Bureau's annual economic surveys, for example, made clear that even state level data are of limited use and that ideally they would like data that could be aggregated into custom local geographies, such as a user-specified collection of counties (Abraham et al. 2018). Survey sample sizes, however, often limit what can be produced with any degree of reliability to national or perhaps state estimates. Big Data hold the promise of supporting statistics at considerably finer levels of geography than are currently published that may be more meaningful to many users of the data.[2]

Finally, new tools should make it possible to automate much of the production of economic statistics. To the extent that processes can be re-engineered so that natively digital information flows directly from the source to the agency or organization responsible for producing the relevant economic statistics, the need for survey data can be reduced and scarce survey resources can be directed to measurement domains it is not possible to address with Big Data. In the longer run, this use of Big Data has the potential for reducing the cost and respondent burden entailed with surveys and with enumerations such as the manual collection of prices in the CPI program.

The future is now, or so we say in the subtitle to this essay. Given the successes documented in the papers in the volume, we believe the time is ripe for Big Data to be incorporated systematically into the production of official statistics.

Using Big Data for Economy-Wide Economic Statistics

Major innovations in official statistics often have followed improvement in source data. The first five papers in the volume feature research using data sources that are new to economic measurement. The authors of these papers all are interested in using these new data sources to improve the timeliness and granularity of economic statistics. While the findings are encouraging, the authors are quick to point out that incorporating these new sources into routine production of economic statistics is not trivial and will require substantial investments.

In their paper, Gabriel Ehrlich, John Haltiwanger, Ron Jarmin, David Johnson and Matthew Shapiro offer a vision of what integrated price and quantity measurement using retail transaction level data might look like. Currently, retail prices and quantities are measured separately (prices by the BLS and quantities by the Census Bureau), using separate surveys drawn from different frames of retail businesses. Collecting prices and quantities separately limits how the resulting data can be used. Furthermore, the survey-based methodologies employed to collect the data limit the timeliness as well as the geographic and product specificity of the resulting estimates. Computing estimates of prices, quantities and total retail sales directly from point of sale transactions data—which record both the price and the quantity of an item sold at a particular location—can overcome all these issues. The trick is first to secure access to transactions level data and second to develop the computational and analytic infrastructure to produce reliable estimates from them. Ehrlich et al. use a subset of transactions-level data from Nielson and NPD to demonstrate feasible methods for accomplishing this. They describe many of the practical challenges involved in using

[2] Producing more granular statistics does raise challenges related to the preservation of privacy and confidentiality, challenges we discuss further below.

transactions-level data for economic measurement, especially for measuring price changes. A key feature of transactions-level data is the large amount of product turnover. While the methods proposed by Ehrlich et. al. show promise, the authors stress that much more work on methodological and data access issues is needed before the agencies can use transactions level data for measuring retail prices and quantities at scale.

The paper by Crystal Konny, Brendan Williams and David Friedman examines several alternative data sources the BLS has studied for use in the Consumer Price Index (CPI). First, they describe efforts to use transaction summaries from two corporate retailers, one of which is unwilling to participate in traditional BLS data collections, as a replacement for directly collected prices. One important issue encountered in the data for one of these firms was the presence of large product life cycle price declines. Absent sufficiently rich descriptions of the products being priced, there was not a good way to deal with this. Second, Konny, Williams and Friedman discuss how the BLS has used data obtained from several secondary sources, prioritizing product areas with reporting issues. In the case of data on new vehicle sales from JD Power, BLS has been able to field a successful experimental series and intends to introduce these data into regular CPI production. This is expected to be more cost effective than existing collection methods. Finally, the authors report on efforts to scrape data on fuel prices from a crowd sourced website (GasBuddy) and using APIs to obtain data on airline fares. Overall, the authors describe excellent progress at the BLS on introducing new data sources into the CPI. The work to date, however, relies on idiosyncratic methods related to the specific data sources and product or services involved. This may limit the ability of the BLS to scale these approaches across additional items in the CPI basket or to expand the basket to include a larger subset of the potential universe of items.

Rebecca Hutchinson's paper describes ongoing work at the Census Bureau to obtain alternative source data for retail sales. The Census Bureau's monthly retail trade survey has experienced significant declines in response rates and thus has been prioritized for modernization (Jarmin 2019). Like Ehrlich et al., Hutchinson uses data from NPD's database, but rolled up to observations on the dollar value of sales at the product by store level. She examines how well the NPD numbers map to the retail sales data collected for the same companies and also how closely movements in the aggregated NPD numbers align with national level Census estimates. Work is underway to examine how the product codes in the NPD data map to those used for the 2017 Economic Census. The results are very encouraging. Indeed, the Census Bureau has replaced monthly survey data with NPD sourced retail sales for over 20 companies and is working with NPD to increase that number. Hutchinson provides an excellent summary of the Census Bureau's process for negotiating access to and testing of the NPD data. It is instructive to see how much effort was required in what was, compared to other alternative data efforts, a relatively straight forward process. In addition to the explicit cash costs for third-party data acquisition, these implicit costs will need to come down through increased experience if the agencies are to scale these efforts under realistic budget assumptions.

The paper by Aditya Aladangady, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengerman and Claudia Sahm uses anonymized credit card transactions data from First Data, a large payments processor, for retail stores and restaurants. The data permit the authors to look at daily spending within tightly defined geographic regions with a lag of only a few days. They show that national monthly growth rates in the data track fairly well with the Census Bureau's monthly retail trade estimates. Then they use the daily feature of the data to track the impact of shocks, such as the 2018-2019 government shutdown and natural disasters, on consumer spending. Before the data can be used for analysis, a number of filters must be applied. A key filter controls for the entry and exit of particular merchants from the database. The necessity of accounting for features of an alternative data source that complicates its application to economic measurement is a feature of many of the papers in this volume. Aladangady et al., demonstrate that the careful application of filters to raw Big Data sources can result in data that are fit for various measurement tasks.

The final paper in the section, by Tomaz Cajner, Leland Crane, Ryan Decker, Adrian Hamins-Puertolas and Christopher Kurz, aims to improve real time measurement of the labor market by combining timely private data with official statistics. Many efforts to use alternative data for economic measurement attempt to mimic some official series. Cajner et al. depart from this by bringing multiple noisy sources together to better measure the true latent phenomenon, in their case payroll employment. Thus, they model payroll employment using private data from the payroll processing firm Automatic Data Processing (ADP) together with data from the BLS Current Employment Statistics (CES) survey. Importantly for policymakers, forecasts using the authors' smooth state space estimates outperform estimates from either source separately. An important feature of the ADP data, which are available weekly, are their timeliness. This featured critically when the authors, in collaboration with additional coauthors from ADP and academia, recently used these data and methods to produce valuable information on employment dynamics during the COVID-19 crisis (Cajner et al. 2020).

Uses of Big Data for Classification

Many data users care not only or even primarily about aggregate measurements but also about information by type of firm, product or worker. Published official statistics are based on standardized classification systems developed with the goal of allowing agencies to produce disaggregated statistics that are categorized on a comparable basis. In a "designed data" world, information about industry, product category, occupation and so on is collected from the firm or worker and used to assign each observation to an appropriate category. In some cases, expense precludes the collecting of the information needed to produce statistics broken out in accord with a particular classification. Even when it is collected, the responses to the relevant questions may be missing or unreliable. Responses from businesses about organizational form or industry, for example, frequently are missing on surveys, and when provided, the information can be unreliable since the question asks about a construct created by the agency rather than a variable that has a natural counterpart in businesses' operations. The next three papers provide examples of how non-designed data can be used to produce statistics broken out along dimensions relevant to users of the data or to better categorize the information already being collected by the statistical agencies.

In their paper, Arthur Turrell, Bradley Speigner, Jyldyz Djumalieva, David Copple and James Thurgood begin by noting that the statistics on job openings available for the United Kingdom are reported by industry but are not broken out by occupation. Turrell et al. use machine learning methods in conjunction with information on job advertisements posted to a widely used recruiting website to learn about occupational vacancies. Using matching algorithms applied to term frequency vectors, the authors match the job descriptions in the recruitment advertisements to the existing Standard Occupational Classification (SOC) documentation, assigning a 3-digit SOC code to each advertisement. Turrell et al. then reweight the vacancy counts so that total vacancies by industry match the numbers in published official statistics. The result is estimates that integrate official job openings statistics designed to be fully representative with supplementary Big Data that provide a basis for further disaggregation along occupational lines.

Joseph Staudt, Yifang Wei, Lisa Singh, Shawn Klimek, Brad Jensen, and Andrew Baer address the difficult measurement question of whether or not an establishment is franchise-affiliated. Franchise affiliation was hand-recoded in the 2007 Census, but due to resource constraints, this was not done for the 2012 Census. While commercial sources showed an increase in the rate of franchise affiliation between 2007 and 2012, the Economic Census data showed a significant decline, suggesting a problem with the Economic Census data. The paper automates the recoding process by making use of web-scraped information collected directly from franchises websites as well as data from the Yelp API. The authors then use a machine-learning algorithm to probabilistically match franchise establishments identified in these data sources to the Census Business Register (BR), allowing them to code BR establishments as franchise-affiliated. This approach leads to a substantial increase in the number of establishments coded as franchise-affiliated in the 2017 Economic Census.

Similar to the Staudt et al. paper, John Cuffe, Sudip Bhattacharjee, Ugochukwu Etudo, Justin Smith, Nevada Basdeo, Nathaniel Burbank, and Shawn Roberts use web-scraped data to classify establishments into an industrial taxonomy. The web-scraped information is based on text; it includes variables routinely used by statistical agencies (establishment name, address, and type) and novel information including user reviewers that bring a novel dimension—customer assessment—to informing the classification of businesses. As with the previous paper, establishments identified via web scraping are matched to the BR and coded with a characteristic—in this case, a North American Industry Classification System (NAICS) industry classification. This approach yields a fairly low misclassification rate at the 2-digit NAICS level. Further work is needed to evaluate whether the general approach can be successful at providing the more granular classifications required by agencies.

Uses of Big Data for Sectoral Measurement

In addition to their uses for producing broad measures of economic activity and assisting in the assignment of activity to categories, Big Data also can be applied to address measurement issues relevant to a specific sector. Several of the papers in the volume leverage data sources that are generated owing to how activity in a particular context is organized, taxed, or regulated to produce new measures in the areas of international trade, health care and housing. Because of the way in which foreign trade is taxed and regulated, there is much more detailed and frequent administrative data on the prices and quantities associated with international transactions than of those associated with domestic transactions. Owing to the fact that medical care typically is accompanied by insurance claims, there are rich data on health care diagnoses, treatment costs, and outcomes. State and local property taxation means that there are rich data on the valuations and sales of residential real estate. Other regulated or previously-regulated sectors (e.g., transportation, energy utilities) also have rich and often publicly-available sources of data that are a byproduct of the regulatory regime. Industrial organization economists long have used these data for market analyses. The analyses in several of the volume's papers show how unique data available for several sectors beyond those covered by retail scanners can be used to produce meaningful measures.

New types of data generated by social media and search applications provide novel opportunities for measurement based on the wisdom of crowds. The paper by Edward Glaeser, Hyunjin Kim and Michael Luca addresses the fact that official County Business Patterns (CBP) statistics on the number of business establishments at the Zip Code level do not become available until roughly a year-and-a-half, or in some cases even longer, after the end of the year to which they apply. There would be considerable value in more timely information. Glaeser et al. ask whether information gleaned from Yelp postings can help with estimating the number of new business startups for Zip Code geographies in closer to real time. Yelp was founded in 2004 to provide people with information on local businesses and the website's coverage grew substantially over the following several years. The data used by Glaeser, Kim and Luca span a limited period (2012 through 2015) but have broad geographic coverage with more than 30,000 Zip Code tabulation areas. They use both regression and machine learning methods to develop forecasts of growth in the Zip-Code-level CBP establishment counts. Adding current Yelp data to models that already include lagged CBP information substantially improves the forecasts. Perhaps not surprisingly, these improvements are greatest for Zip Codes that are more densely populated and have higher income and education levels, all characteristics that one would expect to be associated with better Yelp coverage.

Several papers in this volume examine the use of unit values for retail scanner data for price measurement. The paper by Don Fast and Susan Fleck looks at the feasibility of using administrative data on the unit values of traded items to calculate price indexes for imports and exports. Although the paper uses a fairly granular definition for what constitutes a product, making use of information on each transaction’s 10-digit harmonized system (HS) code, the items in these categories are considerably more heterogeneous than, for example, the products used to construct traditional matched model price indexes or the products identified by retail UPC codes. This creates a risk that changes in average prices in a given 10-digit HS category could reflect changes in product mix rather than changes in the prices of individual items. Although they do not have information that allows them to track specific products, Fast and Fleck do have other information that they argue lets them get closer to that goal, including the company involved in the transaction and other transaction descriptors. Fast and Fleck report that there is considerable heterogeneity in transaction prices within 10-digit HS code but that this heterogeneity is reduced substantially when they use additional keys, i.e., the other transaction descriptors available to them. Their work suggests that, by using the additional descriptors to construct sets of transactions that are more homogeneous, it may be feasible to construct import and export price indexes using the administrative data.

There have been substantial advances in using large-scale datasets on medical treatments for the measurement of health care. The BEA (Dunn, Rittmueller, and Whitmire, 2015) uses insurance claims data to implement the disease-based approach to valuing health care advocated by Cutler, McClellan, Newhouse, and Remler (1998) and Shapiro, Shapiro, and Wilcox (2001). Health insurance claims data can provide comprehensive measurements of inputs and outputs for the treatment of disease. This volume’s paper by John Romley, Abe Dunn, Dana Goldman, and Neeraj Sood uses data for Medicare beneficiaries to measure multifactor productivity in caring for acute diseases that require hospitalization. Output is measured by health outcomes, which, in the absence of market valuations, provide a proxy for the value of healthcare (Abraham and Mackie, 2004). Utilizing a rich claims dataset, the authors make comprehensive adjustments for factors that affect health outcomes such as comorbidities and social, economic, and demographic factors thereby allowing them to isolate the effect of treatments on outcomes. While they find evidence for improvements in the quality of many health treatments, which would lead price indexes that do not adjust for quality change to overstate healthcare price inflation, their results imply that quality improvement is not universal. For heart failure, one of the eight diseases studied, there is evidence that, over the years study, the productivity of treatment declined.

Case and Shiller (1989) introduced the idea of using repeat sales of houses to construct a constant-quality measure of changes in price. Building on these ideas, the increasing availability of data on transactions prices from local property assessments and other sources has revolutionized the residential real estate industry. Zillow provides house price estimates based on repeat sales at the house level. Marina Gindelsky, Jeremy Moulton, and Scott Wentland explore how the Zillow data might be used in the national income and product accounts. The U.S. national accounts use a rental-equivalence approach to measuring the services of owner-occupied housing. Implementing the rental equivalence approach requires imputation since, by definition, owner-occupied housing does not have a market rent. An important difficulty with this approach is that it relies on there being good data on market rents for units that are comparable to owner-occupied units. The chapter discusses the challenges to the implementation of the rental equivalence approach and the steps the BLS and BEA take to address them.

The chapter then asks whether a user cost approach facilitated by Big Data house prices is a useful alternative to the rental-equivalence approach. As explained in detail in the paper, the real user cost of housing depends on the price of housing, the general price level, the real interest rate, the depreciation rate and the real expected capital gain on housing. In the chapter's analysis, the empirical variation in user cost comes almost exclusively from variation in the price of housing. During the period under study, the U.S. experienced a housing boom and bust. Likely because the paper's user costs calculations do not explicitly account for systematic variation in expected capital gains, which economic theory suggests should be related to the level of prices, the user cost estimates reported in the paper mirror this boom and bust cycle in housing prices. The observed fluctuation in house prices seems highly unlikely to reflect a corresponding fluctuation in the value of real housing services. Hence, while the paper contains a useful exploration of housing prices derived from transactions-based data, it is difficult to imagine the method outlined in the paper being used for the National Income and Product Accounts.

Methodological Challenges and Advances

As already mentioned, one significant impediment to realizing the potential of Big Data for economic measurement is the lack of well-developed methodologies for incorporating them into the measurement infrastructure. Big Data applications in many contexts make use of supervised machine learning methods. In a typical application, the analyst possesses observations consisting of a gold-standard measure of some outcome of interest (e.g., an estimate based on survey or census data) together with Big Data she believes can be used to predict that outcome in other samples. A common approach is to divide the available observations into a training data set for estimating the Big Data models, a validation data set for model selection and a test data set used to assess the model's out-of-sample performance. Validation and testing are important because overfitting can produce a model that works well in the training data but performs poorly when applied to other data.

The fact that Big Data suitable for the production of economic statistics have only relatively recently become available, however, means the standard machine learning approaches often cannot simply be imported and applied. That is the challenge confronted in the paper by Jeffrey Chen, Abe Dunn, Kyle Hood, Alexander Driessen and Andrea Batch. Chen et al. seek to develop reliable forecasts of the Quarterly Services Survey (QSS) series used in constructing Personal Consumption Expenditures (PCE). Complete QSS data do not become available until about two-and-a-half months after the end of the quarter and their arrival often leads to significant PCE revisions. Chen et al. consider several types of information including credit card and Google trends data as potential predictors of QSS series for detailed industries to be incorporated into the early PCE estimates. They also consider multiple modeling approaches, including not only moving average forecasts and regression models but also various machine learning approaches. Because the 2010Q2 through 2018Q1 time period for which they have data spans just 31 quarters, splitting the available information into training, validation and test data sets is not a feasible option. Instead, Chen et al. use 19 quarters of data to fit a large number of models using different combinations of source data, variable selection rule and algorithm. They assess model performance by looking at predicted versus actual outcomes for all of the QSS series over the following 12 quarters. The intuition behind their approach is the idea that modeling approaches that consistently perform well are least likely to suffer from overfitting problems. Chen et al. conclude that, compared to current practice, ensemble methods such as random forests are most likely to reduce the size of PCE revisions and incorporating nontraditional data into these models can be helpful.

Rishab Guha and Serena Ng tackle a somewhat different methodological problem. Use of scanner data to measure consumer spending has been proposed as a means of providing more timely and richer information than available from surveys. One barrier to fully exploiting the potential of the scanner data, however, is the lack of well-developed methods for the seasonal adjustment of weekly observations. In essence, the challenge is that events that can have an important effect on consumer spending may occur in different weeks in different years. As examples, Easter may fall any time between the end of March and the end of April; the 4th of July may occur during either the 26th or the 27th week of the year; and both Thanksgiving and Christmas similarly may fall during a different numbered week depending on the year. Unless the data are seasonally adjusted, movements in spending measures based on scanner data cannot be easily interpreted. Customized seasonal adjustment models could be developed for a series or two, but that is not feasible when the time period covered is short and the number of series to be adjusted is large.

Guha and Ng work with weekly observations for 2006-2014 for each of roughly 100 expenditure categories by county. Their modeling first removes deterministic seasonal movements in the data on a series-by-series basis and then exploits the cross-section dependence across the observations to remove common residual seasonal effects. The second of these steps allows for explanatory variables such as day of the year, day of the month, and county demographic variables to affect spending in each of the various categories. As an example, Cinco de Mayo always occurs on the same day of the year and its effects on spending may be greater in counties that are more heavily Hispanic. Applying machine learning methods, Guha and Ng remove both deterministic and common residual seasonality from the category by county spending series, leaving estimates that can be used to track the trend and cycle in consumer spending at a geographically disaggregated level.

Erwin Diewert and Robert Feenstra address another important issue regarding the use of scanner data for economic measurement, namely, how to construct price indexes that account appropriately for the effects on consumer welfare when commodities appear and disappear. Using data for orange juice, the paper provides an illustrative comparison of several empirical methods that have been proposed in the literature for addressing this problem. On theoretical grounds, they say, it is attractive to make use of the utility function that has been shown to be consistent with the Fisher price index. On practical grounds, however, it is much simpler to produce estimates that assume a constant elasticity of substitution (CES) utility function as proposed by Feenstra (1994) and implemented in recent work by Redding and Weinstein (2020) and Ehrlich et al. in this volume. The illustrative calculations reported by Diewert and Feenstra suggest that results based on the latter approach may dramatically overstate the gains in consumer welfare associated with the introduction of new products. A possible resolution, currently being explored by one of the authors, may be to assume a more flexible translog expenditure function that better balances accuracy with tractability.

Increasing the Use of Big Data for Economic Statistics: Challenges and Solutions

The papers in the volume document important examples of the progress thus far in incorporating Big Data into the production of official statistics. They also highlight some of the challenges that will need to be overcome to fully realize the potential of these new sources of data.

One of the lessons learned from successful current partnerships between federal agencies and private data providers is the necessity of accepting Big Data as they exist rather than requiring data providers to structure them in some pre-defined fashion. What that means, however, is that the agencies need to be nimble in working with data that were not originally designed for statistical analysis. As illustrated by the papers in the volume there are a number of respects in which information generated for commercial or administrative purposes may not readily map into measurements that are immediately useful for statistical purposes:

- Data created for business purposes may not be coded into the categories required for the production of official statistics. As an example, scanner data contain product level price information, but to meet the operational needs of the CPI program, the individual items must be mapped into the CPI publication categories (Konny, Williams and Friedmen, this volume).
- More generally, the variables generated by business and household data frequently do not correspond to the economic and statistical concept embodied in official statistics. Agencies, of course, run risks by mechanically accepting a survey response (see Staudt et al. and Cuffe et al. earlier in this volume). Inhaling Big Data, however, would force agencies to create approaches to map the imported data into desired measurement constructs, problems that many of the papers in the volume confronted.
- There are many complications related to the time-intervals of observations. Weekly data on sales do not map readily to months or quarters (see Guha and Ng, this volume). Payroll data refer to pay period, which may not well-align with calendar periods, especially for bi-weekly payrolls (see Cajner et al., this volume). The BLS household and establishment surveys deal with this problem by requiring responses for a reference

period, which shifts the onus to respondents for mapping their reality into an official survey, but moving to the use of Big Data puts the onus for dealing with the issue back onto the statistical agency.

- Data generated as a result of internal processes may lack longitudinal consistency, meaning there may be discontinuities in data feeds that then require further processing by the statistical agencies. Even if the classification of observations is consistent over time, turnover of units or of products may create significant challenges for the use of Big Data (see e.g. Ehrlich et al. and Aladangady et al., this volume).

Producing nominal sales or consumption totals is conceptually simpler than producing the price indexes needed to transform those nominal figures into the real quantities of more fundamental interest. Product turnover causes particular difficulties for price index construction. The BLS has developed methods for dealing with product replacement when specific products selected for inclusion in price index samples cease to be available, but these methods are not feasible when indexes are being constructed from scanner data that may cover many thousands of unique items. As pioneered by Feenstra (1994) and advanced by Ehrlich et al (this volume), Diewert and Feenstra (this volume), and Redding and Weinstein (2020), dealing with ongoing product turnover requires new methods that take advantage of changes in spending patterns to infer consumers' willingness to substitute across products.

Another set of issues concerns the arrangements under which data are provided to the statistical agencies. Much of the work that has been done to date on the use of Big Data to improve economic statistics has been done on a pilot basis, to assess the feasibility of using the data, or to fill specific gaps in data (see Hutchinson and Konny, Williams, and Friedman, both this volume). In several instances, the use of Big Data has been initiated when companies preferred to provide a larger data file rather than be burdened by enumeration (Konny, Williams, and Friedman, this volume). Even when data are more comprehensive, they may be provided under term-limited agreements that might not have the stability and continuity required for use in official statistics. This volume's papers by Federal Reserve Board authors using credit card and payroll data are examples of cases in which this appears to be the case. Several of the papers in this volume make use of retail scanner data made available through the Kilts Center at the University of Chicago under agreements that specifically exclude their use by government agencies.

At least given the statistical agencies' current budgets, unfortunately, scaling the existing contracts at a similar unit cost would be cost-prohibitive. Some data providers may find it attractive to be able to say that their information is being used in the production of official statistics, perhaps making it easier for the agencies to negotiate a mutually agreeable contract for the continuing provision of larger amounts of data. In general, however, new models are likely to be needed. As an example, Jarmin (2019) suggests that existing laws and regulations could be changed to better encourage secure access to private sector data for statistical purposes. One promising path would be to allow third-party data providers to report to the federal statistical agencies on behalf of their clients, making that a marketable service they could sell. For example, as part of the services small businesses receive from using something like QuickBooks, Inuit could automatically and securely transmit data items needed for economic statistics to the appropriate agency.

In some cases, publicly facing websites contain information that could be used to improve existing economic statistics. This volume's papers by Konny, Williams, and Friedman; Staudt et al.; Cuffe et al.; and Glaeser, Kim, and Luca all make use of such information. Even where data are posted publicly, however, the entities that own the data may place restrictions on how they can be used. As an example, the terms of use on one large retailer's website state "(Retailer) grants you a limited, non-exclusive, non-transferable license to access and make non-commercial use of this website. This license does not include… (e) any use of data mining, robots or similar data gathering and extraction tools." What this typical provision would appear to mean, however, is that any statistical agency wanting to use information from this retailer's website will need to negotiate an agreement allowing that to happen. Multiplied across all the websites containing potentially useful information, obtaining these agreements could be a daunting task. In some cases, it may be possible to obtain desired information using an Application Programming Interface (API) provided by an organization, but this by no means is guaranteed.

One concern often cited with regard to the use of Big Data in the production of economic statistics is that the data could cease to be available or be provided in an inconsistent fashion over time, jeopardizing continuity in the production of statistical estimates. To be sure, in the face of sharply declining survey response rates, the sustainability of the statistical agencies' current business model is itself very much an open question. These recent trends suggest strongly that business as usual is simply not an option. Further, unexpected events such as the recent COVID-19 crisis can disrupt planned survey data collections and the timing of deliveries of key administrative data. Perhaps in such circumstances the flow of Big Data might actually be less vulnerable to interruption than the flow of data from traditional sources. Although Big Data are not produced primarily with the federal statistical agencies in mind, there often are other data users who are paying customers and rely on continuity of data provision. This may provide some assurance that the data will continue to be available. Contractual agreements also can help to ensure that a data source does not disappear without warning. One might worry that a sole-source contract with a data provider could lead to a hold-up problem. In the context of a continuing relationship, however, this seems unlikely to be in the data provider's interest. All of this said, care should be taken in vetting potential private sector suppliers of data essential to the production of key economic statistics. In addition, increasing reliance on Big Data also may require the development of Plan B's that could be implemented in the event incoming data are disrupted.

Another concern with the use of commercial Big Data, especially from data aggregators, is that the data provider may have advanced insight into official statistics by virtue of providing major inputs into them. This can be addressed by legal and data security arrangements, but is likely to remain a fundamental concern unless aggregator data are blended with sufficiently diverse sources of other data to reduce their leverage as a data input for any particular statistic. A related concern already experienced by the statistical agencies is the reluctance of publicly traded companies to provide data prior to releasing quarterly earnings reports.

Beyond these issues related to accessing new sources of Big Data, the capabilities of the existing statistical agency staff is another factor that could impede the incorporation of these data into ongoing statistical production. Reflecting the needs of existing production processes, most of these staff have backgrounds in statistics or economics rather than in data science. This is a problem that surely can be corrected over time as staff receive training in the use of the relevant data science methods. The Census Bureau has collaborated with academia to develop a rigorous training curriculum for agency staff (see Jarmin et. al. 2014). This evolved into the Coleridge Initiative, a collaboration among researchers at New York University, the University of Maryland, and the University of Chicago that is providing growing numbers of agency staff with hands-on training on data linkage and data science applications. Further, new hires increasingly will arrive with data science skills acquired as part of their college educations. That said, the statistical agencies will need to make concerted investments to build the skills required to acquire, process and curate data sets that are larger and less structured than the surveys and administrative records on which the agencies have relied historically.

In the meantime, partnerships with those at academic and other research institutions with relevant expertise will be especially important for the agencies. The NSF-Census Research Network (NCRN) is a successful example of such collaboration across a number of universities and the Census Bureau (see Weinberg et al. 2019). The CRIW and NBER have also long been a nexus of collaboration between agencies and academics on measurement issues. This volume is a good example, with several of the papers including both agency and academic coauthors. A more recent nexus of collaboration that is directly relevant to data and methods discussed in this volume are the Tech Economics Conferences held by the National Association of Business Economists. These have featured economists and data scientists from academia, the public sector but especially tech and other companies that have pioneered using data in new and innovative ways.

One of the most attractive features of the economic statistics that could be generated from Big Data also poses one of the biggest challenges associated with their production. Big Data offer the opportunity to produce very granular statistics. Protecting the privacy of the individuals or businesses underlying such detailed statistics, however, is inherently difficult. The fundamental challenge is that as more and more accurate statistics are computed from a private dataset, the more privacy is lost (Dinur and Nissim, 2003). Formal methods, such as differential privacy, allow data publishers to make precise choices between privacy protection and data utility. A balance must be struck, however, between these competing objectives (see Abowd and Schmutte, 2019). As the controversy around the Census Bureau’s adoption of differential privacy as the protection methodology for products from the 2020 Census demonstrates, coming to an agreement about what is appropriate can be a difficult process. That said, the Census Bureau has also used these methods to protect privacy in products such as the Post-Secondary Employment Outcomes without much controversy (Foote, Machanavajjhala and McKinney, 2019). A key distinction is that there are well established and politically sensitive use cases for decennial census data whereas products like the PSEO are new and would be impossible to produce with sufficient accuracy and privacy protection without using modern disclosure avoidance techniques. This gives us hope that new economic statistics computed from detailed transactions, geolocation and other sensitive sources can be released with an acceptable tradeoff between utility and privacy and be broadly accepted by data users.

While not the focus of this volume, the computing infrastructure of the agencies will be need to be improved for the agencies to benefits from these new data sources and tools. This is especially the case if the agencies intend to access data in new automated ways such as through APIs or taking advantages of approaches like secure multi-party computing. There has been recent progress on moving some agency computing infrastructure to the cloud. Continued progress and investments in modern computing capabilities is a necessary condition for success in the Big Data era.

A final set of challenges for the statistical agencies relates to their organizational structure and ability to collaborate across organizational boundaries. Historically, each of the three main economic statistics agencies—the BLS, Census Bureau and Bureau of Economic Analysis—has had a well-defined set of largely distinct responsibilities. Although there always has been considerable collaboration among the agencies, the largest part of the core work of each agency was carried out independently. To illustrate, the Census Bureau carries out surveys to measure nominal sales by industry; the BLS carries out surveys to produce the price indexes needed to convert nominal sales into real quantity measures; and the Bureau of Economic Analysis uses these separately-produced measures in the production of the National Income and Product Accounts. In the new Big Data world, however, as discussed by Ehrlich et al. (this volume) the same underlying data could be used to produce both sales and price statistics in a much more integrated fashion. To take another example, both BLS and Census produce employment statistics based on employer surveys. Big Data in the form of payroll records of the type analyzed by Cajner et al. (this volume) could potentially strengthen all of these estimates.

All of this implies that, in a Big Data world, the agencies' production processes will need to be more integrated than in the past. Rather than each agency negotiating separately with the providers of Big Data related to sales, prices or employment, a single agreement ideally would be negotiated on behalf of all of the interested agencies. In a world in which statistics are based on built-for-purpose surveys, there is a rationale for separating the production of sales statistics and price statistics, though even in this world this separation has caused problems. In a world in which these statistics increasingly are based on Big Data, however, there is a compelling rationale for the integration of these data programs. This may not, however, be something the agencies can accomplish on their own.

At present, the Census Bureau and BEA are located together in the Department of Commerce, whereas the BLS is a part of the Department of Labor. Having official statistics spread over the measurement and estimation programs of multiple agencies creates barriers to realizing the full potential of Big Data. Over the decades, there have been multiple proposals for consolidating the agencies or, absent such reorganization, for reducing legal and institution barriers to coordinating their measurement programs. In the new Big Data world, the potential benefits of coordination or reorganization loom much larger than in the past. Absent reorganization, legal changes that will allow the agencies to coordinate their activities more effectively would advance the agenda for using Big Data to improve official statistics.

Despite the challenges and the significant agenda for research and development they imply, the papers in the volume point strongly toward more systematic and comprehensive incorporation of Big Data to improve official economic statistics in the coming years. Indeed, the future is now.

References

Abowd, John and Ian Schmutte. 2019. “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices,” *American Economic Review*, 109(1): 171-202.

Abraham, Katharine G., Constance F. Citro, Glenn D. White, Jr., and Nancy K. Kirkendall, eds. 2018. *Reengineering the Census Bureau's Annual Economic Surveys*. Washington, DC: National Academies Press.

Abraham, Katharine G. and Christopher Mackie, eds. 2004. *Beyond the Market: Designing Nonmarket Accounts for the United States*. Washington, D.C.: National Academies Press.

Baruch, Yehuda and Brooks C. Holtom. 2008. “Survey response rate levels and trends in organizational research,” *Human Relations, 61*(8): 1139-1160.

Bean, Charles. 2016. *Independent Review of UK Economic Statistics.* London, UK: Cabinet Office and H.M. Treasury.

Bostic, William G., Ron S. Jarmin and Brian Moyer. 2016. “Modernizing Federal Economic Statistics,” *American Economic Review: Papers and Proceedings,* 106(5): 161-164.

Buffington, Catherine, Carrie Dennis, Emin Dinlersoz, Lucia Foster and Shawn Klimek, 2020. "Measuring the Effect of COVID-19 on U.S. Small Businesses: The Small Business Pulse Survey," Working Paper 20-16, Center for Economic Studies, U.S. Census Bureau.

Cajner, Tomaz, Leland D. Crane, Ryan A. Decker, John Grigsby, Adrian Hamins-Puertolas, Erik Hurst, Christopher Kurz, and Ahu Yildirmaz. 2020. “The U.S. Labor Market During the Beginning of the Pandemic Recession,” *Brookings Papers on Economic Activity* [June 25, 2020 conference draft.]

Carson, Carol. 1975. “The History of the National Income and Product Accounts: The Development of an Analytical Tool,” *Review of Income and Wealth*, 21(2); 153-181.

Case, Karl E., and Robert J. Shiller. 1989, "The Efficiency of the Market for Single-Family Homes," *American Economic Review* 79(1): 125-37.

Chetty, Raj, John Friedman, Nathaniel Hendren and Michael Stepner. 2020. “How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data,” NBER Working Paper 27431.

Cutler, David M., Mark McClellan, Joseph P. Newhouse, and Dahlia Remler. 1998. “Are Medical Prices Declining? Evidence from Heart Attack Treatments,” *Quarterly Journal of Economics,* 113(4): 991–1024.

Dinur, Irit and Kobbi Nissim. 2003. “Revealing Information while Preserving Privacy,” *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202-210. https://dl.acm.org/doi/proceedings/10.1145/773153.

Dunn, Abe, Lindsey Rittmueller, and Bryn Whitmire. 2015. “Introducing the New BEA Health Care Satellite Account.” *Survey of Current Business,* 95(1).

Feenstra, Robert C. 1994. “New Product Varieties and the Measurement of International Prices,” *American Economic Review,* 84(1): 157-177.

Feenstra, Robert C. and Matthew D. Shapiro, eds. 2003. *Scanner Data and Price Indexes.* Studies in Income and Wealth No. 64, Chicago: University of Chicago Press.

Fields, Jason, Jennifer Hunter-Childs, Anthony Tersine, Jeffrey Sisson, Eloise Parker, Victoria Velkoff, Cassandra Logan, and Hyon Shin. 2020. “Design and Operation of the 2020 Household Pulse Survey.” Mimeo, U.S. Census Bureau.

Foote, Andrew, Ashwin Machanavajjhala and Kevin McKinney. 2019. “Releasing Earnings Distributions using Differential Privacy: Disclosure Avoidance System for Post-Secondary Employment Outcomes (PSEO). *Journal of Privacy and Confidentiality,* 9(2).

Goldberg, Joseph P. and William T. Moye. 1985. *The First Hundred Years of the Bureau of Labor Statistics*. Washington, DC: U.S. Department of Labor.

Google. 2020. “COVID-19 Community Mobility Reports” [Website] https://www.google.com/covid19/mobility/. Accessed July 24, 2020.

Groves, Robert M. 2011. “Three Eras of Survey Research,” *Public Opinion Quarterly*, 75(5): 861-871.

Groves, Robert M. and Brian A. Harris-Kojetin, eds. 2017. *Innovations in Federal Statistics*. Washington, D.C.: National Academies Press.

Jarmin, Ron S. 2019. “Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics.” *Journal of Economic Perspectives,* 33 (1): 165–84.

Jarmin, Ron S., Julia Lane, Alan Marco and Ian Foster. 2014. “Using the Classroom to Bring Big Data to Statistical Agencies,” *AMSTAT News*. November, 12-13.

Meyer, Bruce D., Wallace K.C. Mok and James X. Sullivan. 2015. “Household Surveys in Crisis,” *Journal of Economic Perspectives*, 29(4): 199-226.

Redding, Stephen J. and David E. Weinstein. 2020. “Measuring Aggregate Price Indices with Taste Shocks: Theory and Evidence for CES Preferences.” *Quarterly Journal of Economics*, 135(1): 503–560.

Shapiro, Irving, Matthew D. Shapiro, and David W. Wilcox. 2001. “Measuring the Value of Cataract Surgery.” In David M. Cutler and Ernst R. Berndt, eds., *Medical Care Output and Productivity*,Studies in Income and Wealth No. 62, Chicago: University of Chicago Press, 411-437.

University of Maryland. 2020. “COVID-19 Impact Analysis Platform” [Website]. Accessed July 30, 2020.

Weinberg, D.H., J.M. Abowd, R.F. Belli, N. Cressie, D.C. Folch, S.H. Holan, M.C. Levenstein, K.M. Olson, J.P. Reiter, M.D. Shapiro, and J. Smyth. 2019. "Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Secure the Future of the Federal Statistical System?" *Journal of Survey Statistics and Methodology* 7(4):589-619. https://doi.org/10.1093/jssam/smy023