the review of income and wealth

Review of Income and Wealth
Series 68, Number 4, December 2022
DOI: 10.1111/roiw.12617

BIG DATA AND OFFICIAL STATISTICS†

BY KATHARINE G. ABRAHAM*

University of Maryland, College Park, MD, USA

The infrastructure and methods for developed countries' economic statistics, largely established in the mid-20th century, rest almost entirely on survey and administrative data. The increasing difficulty of obtaining survey responses threatens the sustainability of this model. Meanwhile, users of economic data are demanding ever more timely and granular information. "Big data" originally created for other purposes offer the promise of new approaches to the compilation of economic data. Drawing primarily on the U.S. experience, the paper considers the challenges to incorporating big data into the ongoing production of official economic statistics and provides examples of progress towards that goal to date. Beyond their value for the routine production of a standard set of official statistics, new sources of data create opportunities to respond more nimbly to emerging needs for information. The concluding section of the paper argues that national statistical offices should expand their mission to seize these opportunities.

JEL Codes: C82, E01

Keywords: big data, official statistics, survey nonresponse, small area estimates, real time data

The core infrastructure underlying the production of U.S. economic statistics dates in large part to the 1930s and 1940s (see, for example, Carson, 1975; Goldberg and Moye, 1985 and U.S. Census Bureau, 2022a). Official statistics on employment, unemployment, earnings, sales, and prices rest on surveys of households and businesses created specifically for the purpose. The national income and product accounts also rely heavily on data from these surveys. Samples for the surveys are selected using probability principles to represent the population of interest and the survey questions solicit information that conforms to economic concepts. Administrative data and periodic censuses are important parts of the statistical infrastructure, providing sampling frames, benchmarks for survey-based statistics

†This paper is based on the Ruggles Lecture I delivered at the 36th Annual Conference of the International Association for Research on Income and Wealth on August 23, 2021. I would like to thank Connie Citro, Abe Dunn, Erica Groshen, Marshall Reinsdorf, Matthew Shapiro, Erich Strassner and two anonymous reviewers for thoughtful comments on earlier drafts of the paper. Shelley Karlsson at the U.S. Census Bureau kindly provided me with data on response rates for selected U.S. Census Bureau business surveys. Brendan Williams and Rob Cage at the Bureau of Labor Statistics, Erich Strassner at the Bureau of Economic Analysis, and William Abriatis, Rebecca Hutchinson, Ron Jarmin and Aidan Smith at the U.S. Census Bureau shared valuable information regarding ongoing initiatives at their agencies involving the use of alternative sources of data for economic statistics. I also am grateful to Arthur Turrell at the United Kingdom's Office of National Statistics for sharing information about their work using data from nontraditional sources to produce real-time economic statistics. Any errors in the paper are of course my own.

*Correspondence to: Katharine G. Abraham,1218 Lefrak Hall, University of Maryland, College Park, MD 20742, USA (kabraham@umd.edu).

and, in some cases, source data used directly in the production of estimates, but survey data play a central role in the existing economic measurement system.

This model for the production of economic statistics has served the country well. The development of probability sample surveys that could support unbiased statistics at a fraction of the cost of a household or business census was a major innovation. Survey methodology has developed as a field, leading to improvements in sampling techniques, methods of adjusting for nonresponse, questionnaire design and other aspects of survey design and administration. In recent years, however, the traditional model's reliance on survey data for the production of economic statistics has come under increasing pressure, raising questions about its continued viability. In some cases, agencies have been able to expand their use of administrative data, especially tax data, to supplement or replace data collected through surveys. The use of administrative data for statistical purposes is an important topic in its own right. My focus today, however, will be on the potential uses of new sources of data generated as a byproduct of private sector economic transactions. Statistical system leaders face major questions about how to employ these new sources of data and the implications of their availability for the role of an official statistics agency in the modern world. I discuss these issues from the perspective of the U.S. statistical system, but hope that at least some of what I have to say will be relevant to the work of official statisticians in other countries.

Section 1 of the paper elaborates on emerging challenges to the existing model for the production of official economic statistics, including the significant declines in survey response rates experienced by national statistical offices and data users' growing demands for more timely and more disaggregated statistics. Section 2 considers the potential that previously untapped private sector data hold for addressing the challenges to the existing model, as well as the new challenges the adoption of these data sources will pose. Section 3 describes some of the interesting work underway at the U.S. statistical agencies to incorporate new types of private sector data into the production of key economic indicators. The agencies' work with these new data sources primarily has focused on improving the standard set of economic statistics for which they are responsible. Not infrequently, however, economic shocks raise important questions that the standard data series are not well designed to answer. As discussed in Section 4, this was certainly the case at the onset of the COVID-19 pandemic. I argue in Section 5 for an expanded view of the mission of a national statistical office that encompasses being prepared to use alternative data sources to produce information for addressing emerging questions, even when the resulting statistics may not be of the same high quality as the agencies' standard measures.

1. Challenges to the Existing Model for the Production of Official Economic Statistics

Data users long have viewed the official economic statistics produced by the Bureau of Labor Statistics (BLS), the Bureau of Economic Analysis (BEA) and the Census Bureau as gold standard measurements. This remains the case today, but cracks have begun to appear in the foundations underlying these measurements. Declining survey response rates have raised concerns about the continued viability

of the existing model for the production of economic statistics. At the same time, users are demanding more timely and more disaggregated information. Further, all of this is occurring in an environment in which, at least in the United States, the budgets of the statistical agencies have been stagnant or declining.

Figure 1 shows the trend in unit response rates for the monthly Current Population Survey (CPS), the source of official U.S. statistics on employment and unemployment; the Annual Social and Economic Supplement to the CPS (CPS-ASEC), the source of data used to produce official poverty statistics; and the quarterly Consumer Expenditure Survey (CEX), which produces important data on consumer spending. The U.S. Census Bureau fields all three of these household surveys. The onset of the COVID-19 pandemic created new problems for household data collection during 2020 and response rates did not return to pre-pandemic levels in 2021. Even before the pandemic, however, response rates for these surveys had fallen sharply. The response rate for the CPS, one of the U.S. statistical system's most important household surveys, fell from an average of 92.4 percent in 2009 to an average of 82.7 percent in 2019, a dismaying decline for a survey that for decades had consistently maintained a response rate in excess of 90 percent. The CPS-ASEC response rate, which reflects nonresponse to the supplement in addition to nonresponse to the monthly survey, fell from 85.5 percent in 2009 to 67.6 percent in 2019 and the CEX response rate fell from 74.5 percent to 53.7 percent. Although concerns about falling household survey response rates had been voiced as early as the 1990s (for a discussion, see Brick and Williams, 2013), the declines since 2009 have been far steeper than the earlier declines. The surveys for which response rates are shown in Figure 1 were selected because of their importance for economic measurement, but other household surveys fielded by the Census Bureau have experienced a similar pattern of response rate decline.[1]

Figure 2 displays unweighted unit response rates for three monthly business surveys fielded by the BLS—the Job Openings and Labor Turnover Survey (JOLTS); the Current Employment Statistics (CES) survey, the source of the monthly payroll employment figures; and the Employment Cost Index (ECI) survey. In contrast to the household survey response rates displayed in Figure 1, the response rates to these surveys do not exhibit consistent downward trends prior to the mid-2010s. By the late 2010s, however, the response rates for all three of these surveys were falling and all dropped sharply with the pandemic's arrival in 2020. Over a period of just 5 years, from 2016 to 2021, the JOLTS response rate fell from 65.8 percent to 45.7 percent; the CES response rate from 60.8 percent to 48.3 percent; and the ECI response rate from 68.6 percent to 54.3 percent.

The Census Bureau also has experienced notable declines in response rates for its business surveys. To illustrate, Figure 3 displays unweighted response rates for four selected Census Bureau annual business surveys—the Annual Survey of Manufactures (ASM), the Annual Wholesale Trade Survey (AWTS), the Annual Retail Trade Survey (ARTS) and the Services Annual Survey (SAS). In addition to being of interest in their own right, the data from these surveys are important inputs to

[1]As examples, the National Crime Victimization Survey, the Medical Expenditure Survey Household Component, and the American Time Use Survey all experienced a notably faster pace of response rate decline during the decade beginning in 2009 than during the prior decade.

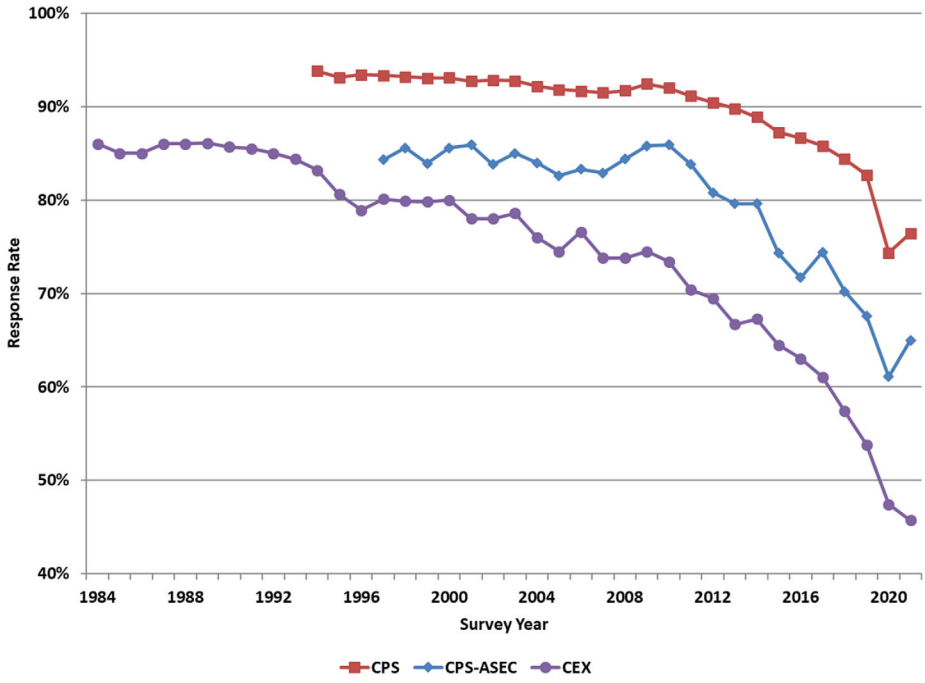

Figure 1. Response Rates for Selected Household Surveys Conducted by the U.S. Census Bureau, 1984–2021.
Source: CPS response rates downloaded using BLS Series Report tool (series ID LNU09300000). CPS-ASEC and CEX rates through 2013 from Meyer *et al.* (2015); later years' response rates for CPS-ASEC from annual supplement documentation and for CEX from BLS website at https://www.bls.gov/osmr/response-rates/. CPS = Current Population Survey. CPS-ASEC=CPS Annual Social and Economic Supplement. CEX = Consumer Expenditure Interview Survey. Annual or annual average response rates reported. [Colour figure can be viewed at wileyonlinelibrary.com].

the construction of the national income and product accounts. These annual surveys are fielded in the year following the year to which the data refer, so that the 2018 data, obtained during 2019, were the last collected before the start of the COVID-19 pandemic. Comparing the 2008 and 2018 surveys, the ASM response rate fell from 80.7 percent to 67.7 percent. Over the same period, the AWTS response rate fell from 81.3 percent to 71.3 percent; the ARTS response rate from 82.4 percent to 64.3 percent; and the SAS response rate from 80.4 percent to 69.6 percent. Perhaps because of their more extended fielding period, these surveys' response rates were less affected by the pandemic than those for the monthly and quarterly surveys for which response rates are shown in Figures 1 and 2.

The response rate for a business survey is of course an imperfect indicator of the effect of nonresponse on data quality. The response rates plotted in Figures 2 and 3 give equal weight to all sample units, but some business units are more important for the estimates than others. In addition, item nonresponse may mean that the information provided by a responding business is incomplete. The Census Bureau is able to use administrative or other data to fill in some of the missing data elements for its annual business surveys. For assessing the quality of the data from these surveys, the Census Bureau calculates a Total Quantity Response Rate (TQRR), defined as

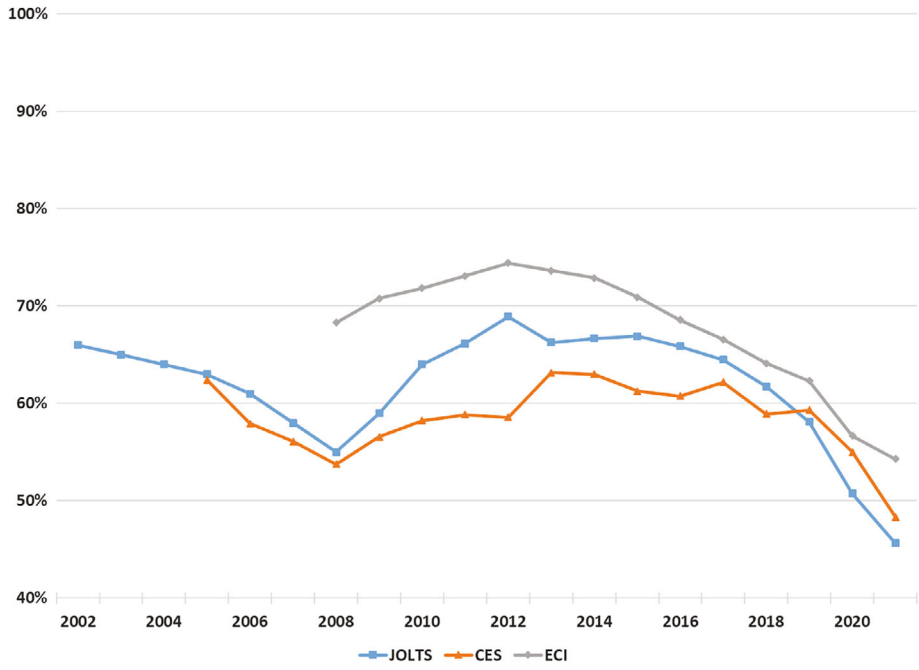

Figure 2. Response Rates for Selected Business Surveys Conducted by the Bureau of Labor Statistics, 2002–2021.

Source: Pre-2013 response rates from Bureau of Labor Statistics (2009, 2016). Response rates for later years from BLS website at https://www.bls.gov/osmr/response-rates/. JOLTS = Job Openings and Labor Turnover Survey. CES=Current Employment Statistics. ECI = Employment Cost Index. JOLTS, second closing unit response rates; CES final private sector unit response rates; ECI total unit response rates. All are annual averages. [Colour figure can be viewed at wileyonlinelibrary.com].

"the percentage of the estimated (weighted) total of a given data item reported by the active tabulation units in the statistical period or from sources determined to be equivalent-quality-to-reported data" (U.S. Census Bureau, 2022b). The TQRR for key data elements collected on the four annual economic surveys for which data are displayed in Figure 3 generally is higher and has fallen less than the unweighted unit response rate. Between 2008 and 2018, the TQRR for sales or revenue in the AWTS, ARTS, and SAS fell only slightly (from 86.1 percent to 85.9 percent, 92.6 percent to 91.7 percent, and 87.0 percent to 85.2 percent, respectively). The decline in the TQRR for revenue in the ASM, the only one of the four surveys for which the unit of observation is the establishment rather than the firm, was much larger; it fell from 83.0 percent to 59.8 percent. Because the other sources of data the Census Bureau deems to be of equivalent quality often contain only a subset of the survey data elements, however, declining survey response rates are a concern even if the TQRRs for revenue and other topside measures have fallen less.

At the same time that it has become more difficult to obtain survey responses, the demands on survey organizations are growing. Data users increasingly seek data that are more timely and more disaggregated than those the economic statistics agencies publish. Several years ago, I chaired a National Academies of Sciences panel charged with reviewing the Census Bureau's annual business survey program.

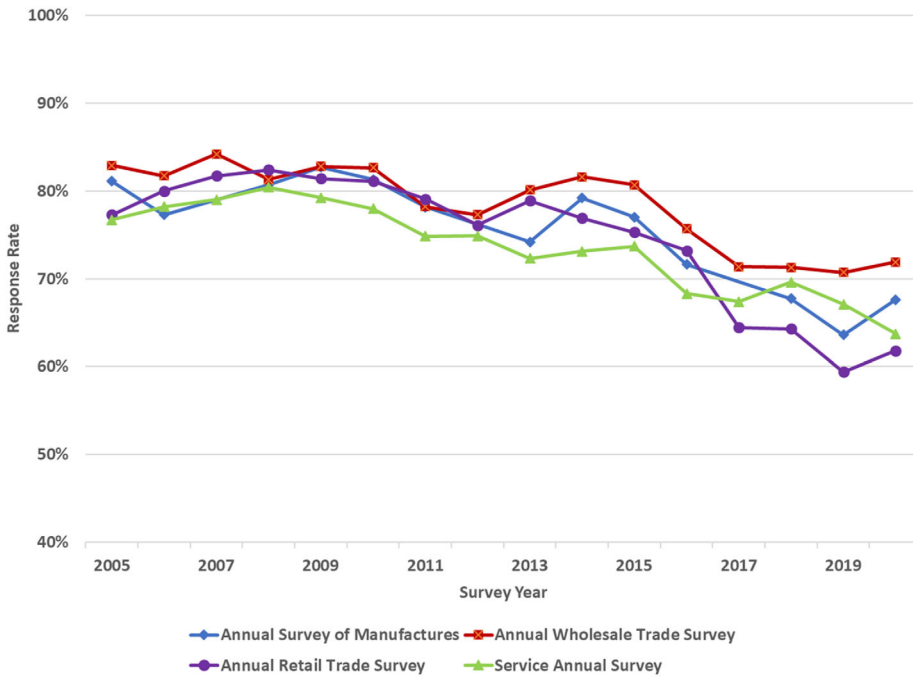

Figure 3. Response Rates for Selected Business Surveys Conducted by the U.S. Census Bureau, 2005–2020.

Source: Personal correspondence with Shelley Karlsson, Assistant Division Chief, Collection Instruments and Preparation, Economic Management Division, U.S. Census Bureau. Annual Survey of Manufactures (ASM) not conducted in years ending in 2 or 7. ASM response rate the percent of establishments mailed a survey for which a survey form submitted. Response rate for other surveys the percent of in-scope businesses for which a valid response provided. [Colour figure can be viewed at wileyonlinelibrary.com].

As part of its work, the panel hosted a workshop featuring presentations from several groups of data users, including one group comprising representatives from state and local business and government organizations, about their information needs. The state and local representatives said they appreciated the high quality of the Census Bureau business data, but wanted more timely data for more disaggregated geographies. The Census Bureau's annual business surveys provide only national- and selected state-level data; the agency's County Business Patterns program produces county-level statistics based on tax data, but these are published with a lag (preliminary data for 2020, for example, were not published until February 2022) and confidentiality considerations require a fair amount of data suppression. Because they could not obtain current-year data from the Census Bureau at the desired level of disaggregation, the state and local representatives who spoke with our panel often turned to private data providers who make use of Census Bureau data as an input to produce modeled estimates (Abraham *et al.*, 2018).

In the United States, the growing demands on the producers of official economic statistics have arisen in an environment of constrained or shrinking statistical agency budgets. Figure 4 shows the trends in real funding levels for the BLS, the BEA and the Census Bureau. In order to focus on the production of economic data

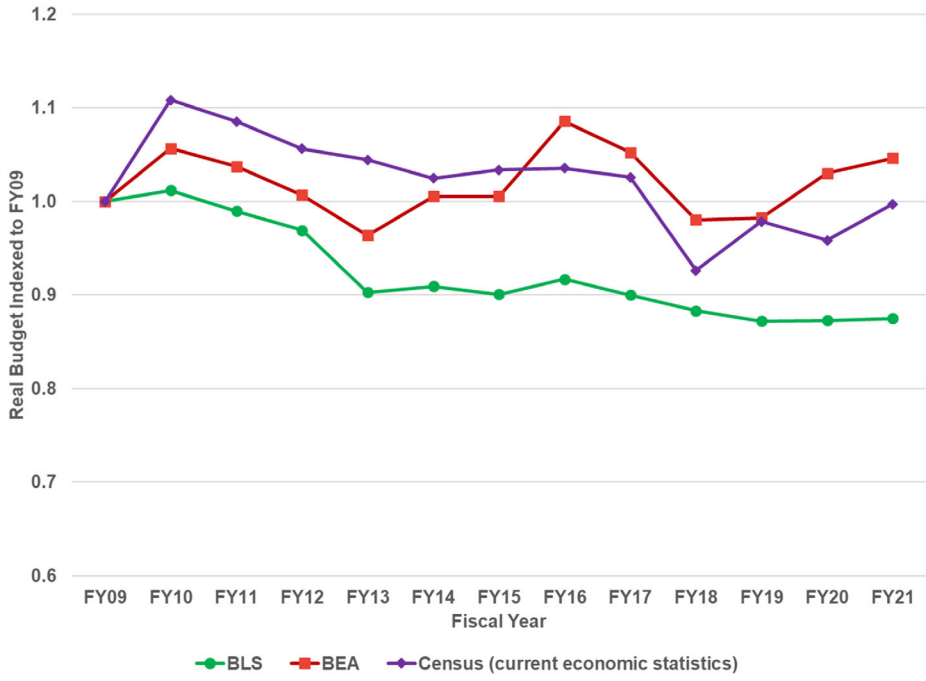

Figure 4. Index of Real Current Program Budgets for the Bureau of Labor Statistics, Bureau of Economic Analysis and U.S. Census Bureau, FY2009-FY2021 (FY2009 = 1.0).
Source: Statistical Programs of the United States Government, Office of Management and Budget, various years; Census Bureau budget documents, various years. Census figures refer to budget for current economic statistics program. Spending converted to constant dollars using GDP deflator. [Colour figure can be viewed at wileyonlinelibrary.com].

and abstract from the fluctuations in funding associated with the every-five-year economic censuses, the reported Census Bureau figures refer only to the budget for current economic statistics. Between Fiscal Year 2009 and Fiscal Year 2021, the Census Bureau's current economic statistics budget has only just kept pace with inflation and the overall BLS budget fell by 12.5 percent in real terms. Only the budget for the BEA—an agency whose activities primarily involve data integration as opposed to field data collections and whose budget is correspondingly smaller than that for the BLS or the Census Bureau's current economic statistics program—has grown, albeit modestly (less than 5 percent in real terms over the same period).

The numbers plotted in Figure 4 show how total real spending on current U.S. economic statistics has changed. Over the years for which the figure tracks overall spending, the number of economic actors and the overall size of the economy grew substantially. Between 2009 and 2021, U.S. employment grew by nearly 10 percent; the number of U.S. business establishments grew by more than 20 percent; and real U.S. Gross Domestic Product (GDP) grew by more than 25 percent.[2] Scaled relative

[2]Employment is annual average CPS employment and the business establishment numbers are third quarter Quarterly Census of Employment and Wages (QCEW) counts, both downloaded from the BLS website. GDP figures were downloaded from the BEA website.

to these trends, the real budgets for current economic statistics at all three agencies have fallen.

2. Are Big Data the Answer?

While collecting the survey data needed to feed the traditional process for producing official economic statistics has become more challenging, the availability of other types of data has grown. As discussed by Bostic *et al.* (2016), Bean (2016), Groves and Harris-Kojetin (2017), Jarmin (2019) and Abraham *et al.* (2022), among many others, recent years have seen a proliferation of natively digital data that have enormous potential for improving economic statistics. Some of these data come from administrative records generated by federal, state and local governments that could be more widely incorporated into the production of economic statistics. The private sector big data that are my focus include scanner data from retail outlets; price, product characteristic and other information posted to the web; credit card transactions data; bank account data; payroll processing and scheduling data; sensor data captured by satellite images, traffic cameras and mobile devices; medical insurance claims data; and many other types of novel data.

Even before the recent surge of interest in the potential statistical uses of natively digital data, the economic statistics agencies had relied in some cases on data from third-party sources. For example, for decades, the BEA has used Wards' Automotive Reports data (to estimate auto sales); IQVIA data (to estimate pharmaceutical sales); and AM Best data (to estimate insurance revenues and profits) (Moyer and Dunn, 2020). Similarly, the BLS has long used data from third-party sources to measure changes in used car prices and information supplied by the Post Office to measure changes in postage rates (Bureau of Labor Statistics, 2020a). The use of such data, however, has been the exception rather than the rule.

Today, the economic statistics agencies are actively exploring whether and how they might expand their use of alternative private data sources. To the extent that such natively digital data can substitute for survey responses, it may be possible to reduce respondent burden and free agency survey resources for the collection of information not obtainable in other ways. In some cases, the use of natively digital data may allow the agencies to improve the timeliness of official statistics or reduce the revisions in published numbers. Access to data sources with a larger number of observations than the typical survey data set may allow the agencies to produce more disaggregated statistics. There is even the possibility that taking advantage of new sources of private data ultimately could lower the statistical agencies' costs, though at least in the short run this is unlikely to be the case.

Although the potential benefits of naturally occurring or big data for statistical purposes are great, there is reason to proceed carefully with incorporating them into the statistical agencies' estimation processes. Table 1 summarizes some of the differences between survey data and naturally occurring private

TABLE 1
CONTRASTING SURVEY AND NATURALLY OCCURRING PRIVATE DATA

Characteristic	Survey Data	Naturally Occurring Private Data
Sample size and representativeness	Small but representative samples of target population	Large but not necessarily representative samples
Data elements	Data elements selected to meet statistical needs	Data elements reflect needs and constraints of business processes
Data quality	Quality control central to survey process, though errors in measurement may arise	Data elements relevant to business processes most likely to be accurate
Comparability of data over time	Comparability of data over time controlled by survey statistician	Comparability of data over time may be disrupted by changes in business requirements or the broader economic environment
Data structure	Data records designed for statistical analysis; typically well documented	Data records reflect business purposes; may or may not be well documented
Data ownership	Data "owned" by statistical agency, typically collected from respondents under a pledge of confidentiality	Data "owned" by business where it was generated; obtaining data may be expensive or raise legal, business or other concerns (including concerns about relying on a monopoly provider)
Fit with statistical agency capabilities	Agencies' human and physical infrastructure developed for collection and processing of survey data	Naturally occurring data sets require new staff skills and enhancements to computing capacity

data that are relevant when contemplating their use in the production of official statistics.[3]

2.1. *Sample Size and Representativeness*

One of the reasons natively digital data are so appealing is the very large size of many of these data sets. As an illustration, in one BLS study that evaluated the potential use of Nielsen scanner data for the Consumer Price Index (CPI), researchers had access to market-level data on the dollar value of sales and the number of units sold for nearly 1.5 million separate products as identified by their UPC codes (Fitzgerald and Shoemaker, 2013). To take a somewhat different example, a University of Maryland research team that used anonymized cell phone data to study patterns of geographic mobility in the United States early in the COVID-19 pandemic was able to exploit information from some 100 million devices (Pan *et al.*, 2020).

[3]Many of the same contrasts between survey data and naturally occurring private data could be drawn between survey data and administrative data.

The very large size of many of these naturally occurring data sets means they may support statistics that are much more disaggregated than statistics based on survey data. A potential drawback is that all of these data sets are non-designed samples, meaning they may or may not fully represent the population of interest. The Nielsen scanner data analyzed in the study by Fitzgerald and Shoemaker, for example, excluded drug stores with less than one million dollars in sales, grocery stores with less than two million dollars in sales, and one major national retailer (Fitzgerald and Shoemaker, 2013). Individuals who use cell phone apps that allow their movements to be observed may differ in important ways from the full population. While large naturally occurring data sets have the potential to be extraordinarily valuable, these sorts of coverage limitations imply that they will be most useful if data known to be representative are available for benchmarking. As a rule, big data will complement rather than fully substitute for survey data.

2.2. *Data Elements*

The nature of the data elements collected also differs between designed data, such as data collected through a survey, and naturally occurring private data. Survey questionnaires are structured to collect the information needed to produce desired statistical estimates, but naturally occurring private data sets contain only the information that is relevant to the business processes from which they emerge. For example, the Consumer Expenditure Survey collects information that supports statistics on income and spending patterns for different types of households. Banking and credit card records provide valuable information about consumer incomes and spending, but contain little information about household demographics or exactly what a household has purchased. To take another example, the Department of Transportation's trip behavior surveys collect information on the demographic characteristics of the survey respondent and the reasons for their trips; mobile phone data allow movements to be observed more directly and with much greater granularity, but provide no direct information about who is traveling or why.

In other cases, a naturally occurring data set may contain data elements that are close to what a survey designer would seek to measure but aligned imperfectly in some respect. As an example, the prices tracked for the CPI exclude temporary discounts available to certain customers at an outlet unless more than half of sales of an item occur at the discounted price. Transactions data, however, will reflect these discounts. Whether such differences are important is ultimately an empirical question.

2.3. *Data Quality*

Designed data and naturally occurring private data sets also may differ in the quality of the information they contain. In survey data, measurement problems may arise because the measurement construct is poorly operationalized; because survey respondents are unable or unwilling to provide an accurate response; or because post-processing of the answers introduces errors (Groves *et al.*, 2009). Survey methodologists are well aware of these potential issues and work to minimize their effects on survey estimates. In a naturally occurring data set, data elements

that are central to a business process are likely to be very accurate, but other information may be less reliable or less complete. Aladangady *et al.* (2022) report that a significant share of merchants in the credit card transactions data they analyze were assigned to line of business classifications that did not correspond to their actual activities. In addition, some merchants batched the processing of their credit card transactions so that the processing date did not necessarily correspond to the transaction date. Similarly, the Census Bureau has found inaccuracies in the classification of construction building permits obtained from third party sources (Aidan D. Smith, video interview with author, April 4, 2022). To take another example, working with data from Homebase, a company that provides scheduling and time clock services to small businesses, Kurmann, Lale and Ta found that the industry code was missing for about a third of the observations (Kurmann *et al.*, 2021).

2.4. *Comparability of Data Over Time*

Because so many of the users of economic statistics are primarily interested in assessing economic trends, it is especially important that these measurements be consistent over time. The agencies that produce economic statistics give high priority to maintaining the comparability of economic time series and, when a break in series is necessary, giving users of the data a way to bridge the break. I should acknowledge that, even if a survey's sample design and questionnaire remain the same, changes in response rates could mean that the data for different years are not fully comparable. On the whole, however, discontinuities in naturally occurring data are likely to be a greater threat to the comparability of economic time series than changes in the meaning of survey responses. Breaks in series related to changes in tax or benefit program rules are a not-infrequent problem with administrative data. Similarly, changes in business processes or in the broader economic environment can lead to discontinuities in the information captured in natively digital data sets. As an example of the latter, to the extent that the prevalence of cash transactions is changing over time, the trend in the volume of credit card spending could diverge from the trend in overall spending.

The turnover of units included in a data set also can create significant challenges for the use of naturally occurring data to measure economic trends. As noted by Aladangady *et al.* (2022), for example, the volume of transactions handled by a payments processor may change either because aggregate spending is in fact changing or because the processor's client base is changing. In their analysis of payment processor data, they address this problem by estimating changes in spending based on a rolling set of fixed-composition panels comprised of merchants present in the data for at least 14 months. Cajner *et al.* (2022) deal with a similar problem in their study using ADP data to track employment growth. To avoid distortions related to changes in the client base represented in the data, their measure of weekly growth in employment makes use of information for payroll accounts present in the ADP data in successive weeks. While these procedures remove spurious changes in the outcome of interest due to changes in the data provider's customer base, they also remove changes resulting from true firm births and deaths.

2.5. *Data Structure and Documentation*

Another frequent concern with naturally occurring private data sets is that they are neither structured nor documented with statistical analysis in mind. Understanding the nuances of a new data set and extracting useful information from it can be a challenging endeavor even in the best-case scenario. Lack of clear documentation makes things that much harder. In a survey of federal statistical agencies discussed by Reamer (2021a), 14 out of 18 agencies that had acquired private data for statistical purposes cited the lack of clear documentation regarding the methodology used to produce the data as a significant data analysis challenge.

Ingesting natively digital data from multiple companies providing information on their individual operations, rather than from a single data aggregator, introduces additional complications. Each of the companies may store different data elements and code them differently. The promulgation of voluntary standards for corporate data could make importing company data easier (Groshen, 2021). The Jobs and Employment DataExchange (JEDx) project, spearheaded by the U.S. Chamber of Commerce Foundation, illustrates how this might work. This initiative has the objective of designing voluntary standards for employers' records on jobs and employment, including information on worker demographics, hours, earnings, occupation and so on.[4] Although the project has broader motivations, were such a standard to be widely adopted, the BLS and other statistical agencies could more easily make use of raw company-provided data in the construction of statistics on employment and earnings.

A final point regarding the complications that may arise in ingesting natively digital data is that the way the data are structured may change over time, requiring the statistical agencies to change their ingestion procedures. As just one example, if a firm redesigns a website that an agency has been scraping, the agency must rewrite the scraping script.

2.6. *Data Ownership*

The fact that a private entity owns data a statistical agency might find useful also can affect the feasibility or advisability of incorporating the data into the production of official statistics. For individual firms, a big question can be what the company gains from agreeing to share its data. Some businesses may view their cooperation in data collection as a civic duty or decide that sharing data is worthwhile because it saves the effort of interacting with survey interviewers or filling out survey forms. Even when a firm is receptive to the idea of sharing its data, however, the process of negotiating a data provision agreement can be lengthy. One inducement for companies to share their data may be a promise from the statistical agency to provide customized reports that compare the firm to others in the same industry or area.

Private data aggregators typically view the data they have assembled as assets to be monetized rather than something to be freely shared with the statistical agencies. Nielsen has made retail scanner data available for research purposes through

[4]See https://www.uschamberfoundation.org/JEDx for additional information about the initiative.

the Kilts Center at the University of Chicago, but the agreements under which that sharing occurs specifically exclude their use by government agencies (Abraham *et al.*, 2022). Even where the use of such proprietary data otherwise might be attractive, purchasing the data may be more expensive than can be justified. The BLS, for example, has evaluated the use of third party scanner data to replace the collection of prices for selected Food at Home items, but concluded it was more cost effective to rely on field economists to collect the needed prices (Konny *et al.*, 2022).

Another concern with incorporating naturally occurring private data into the production of economic statistics is whether the agency can rely on their continued availability. To be sure, in the face of sharply declining survey response rates and sporadic response to non-mandatory surveys, the sustainability of the statistical agencies' current business model is itself very much an open question. Still, if proprietary data are to be incorporated into the production of official statistics, the statistical agency needs to be able to count on having access to them. The use of rolling multi-year contracts, such that an agreement for data provision is always in place for several years ahead, may reduce the risk that a data source simply disappears. In the case of private data aggregators, the existence of other paying customers who are purchasing the same data may provide some additional assurance of continued availability. Even with these sorts of protections, increasing reliance on natively digital private sector data may require the development of backup plans that the statistical agency can implement in the event of a disruption to incoming data flows.

Another issue related to privately sourced data is the risk that a firm supplying a substantial amount of raw input data might be able to anticipate changes in published statistics prior to their release. In the worst-case scenario, a data provider might even be able to manipulate the data it provided so as to affect the published numbers. Were that to happen, it would be a serious blow to public confidence in the integrity of the official statistics. Appropriately structured contract provisions or even legislation similar to existing laws that govern insider trading may be needed to protect the integrity of data production processes that rely heavily on private data sources.

2.7. *Fit With Statistical Agency Capabilities*

A final consideration regarding the use of naturally occurring private sector data is that the current statistical agency staff skill set and computing infrastructure were not built with this sort of data in mind. Most importantly, relatively few current agency staff members have experience working with very large unstructured data sets, though growing numbers are acquiring those skills. The Office of Personnel Management's recent announcement of a new job series for data scientists, introduced as of the end of 2021 (Heckman, 2021) is a positive development. Statistical agency leaders have said this should make it easier for them to hire people with the data science skills they need.

This discussion suggests a possible checklist for agencies deciding whether to incorporate data from naturally occurring sources into the ongoing production of official statistics. Questions to ask about incorporating alternative data include the following:

- Has the collection of data using current methods become difficult or proven to be inadequate to meet users' demands?
- Are the alternative data a good fit for the intended purpose?
- If the agency were to incorporate the alternative data into the estimation process, would the quality of the resulting statistics be of similar or higher quality?
- Would using the alternative data lower the costs of producing required statistics? If not, can any added costs be justified based on improvements to the estimates?
- Would using alternative data create risk due to reliance on 3rd-party data suppliers, such as the risk that the data might not continue to be available in the future? If so, can that risk be mitigated?

These questions are very similar to those the U.S. statistical agencies themselves report they are using to evaluate potential uses of big data in the production of official statistics (see, for example, Konny *et al.*, 2022).

3. Using Naturally Occurring Private Data in the Production of U.S. Official Statistics

While adopting alternative data sources may not be the right answer in every case, U.S. statistical agencies are increasingly interested in the opportunities for replacing survey data with various types of natively digital private data or using natively digital private data to improve or enrich published statistics. To illustrate both the promise of alternative data sources and some of the challenges their use can pose, I briefly describe a few recent examples of the adoption or potential future adoption of new types of data into the production of official statistics. These are drawn from recent work at the three major U.S. economic statistics agencies—the BLS, the BEA and the Census Bureau.

3.1. *New Sources of Data for the CPI*

How scanner and other alternative sources of data might be used for price measurement has been an active subject of research for more than 20 years (National Academies of Sciences, Engineering and Medicine, 2022). Historically, however, almost all of the price information used to produce the CPI has come from surveys carried out by BLS field economists (Bureau of Labor Statistics, 2020a). Obtaining price information in this way has become increasingly difficult. Building on both internal and external research, the BLS CPI program recently embarked on an ambitious program to incorporate nontraditional data collection methods and alternative data sources into the production of official statistics.

The BLS plan for obtaining data in new ways and from new sources envisions the use of several types of data—company data submitted from corporate headquarters as an alternative to in-store price collection, data from secondary sources and data scraped from the Web (Konny *et al.*, 2022). Changes already incorporated into production include the substitution of prices provided directly by two large companies for prices collected by BLS field economists. In June 2021, crowdsourced gasoline prices obtained from an online website replaced directly collected gasoline prices in the index. Plans to use data on new car prices purchased from

J.D. Power and to incorporate information on airfares from a national airline have been approved for implementation. BLS is actively exploring the possible use of a number of other alternative data sources. If all goes well, within a few years, up to 22 percent of the index could be constructed using data from alternative sources (Paplomatas, 2021).

The BLS is not alone among national statistical offices in moving from studying alternative sources of data on consumer prices to incorporating them into official price measures. Statistics Netherlands, Statistics Canada, the Australian Bureau of Statistics, and the Office for National Statistics (ONS) in the United Kingdom, among others, are moving or have already moved in this direction. Statistics Netherlands began to investigate the use of large retailers' scanner data for price measurement in the 1990s and introduced the first company scanner data into its CPI some 5 years later. By 2020, 35 percent of the CPI market basket in the Netherlands was priced using transactions data obtained from companies and 6 percent using data scraped from the web (Chessa, 2021). Statistics Canada first introduced scanner data into CPI production for the May 2018 reference month. By March 2020, prices for about 20 percent of the basket weight were collected from some alternative data source. The goal is to increase that to 55 percent by March of 2023 (Ertl *et al.*, 2020). As of 2020, nontraditional data sources accounted for 43 percent of the Australian CPI market basket (Merrington and Smyth 2020). The ONS is not as far along but expects to begin introducing data obtained from alternative sources into its CPI in 2023 (Office for National Statistics, 2021).

In addition to various technical issues related to index construction, one practical challenge in incorporating alternative data sources into these agencies' CPIs has been the difficulty of convincing firms to share their transactions-level data. The process of negotiating an agreement with a company to do so can take months if not years. Once companies begin to provide their data, the statistical agency staff must deal with company records submitted in multiple formats. Another practical challenge is the very large number of products represented in the transactions-level data. Manually assigning every product on every company's transactions file to an item category would be an unmanageable task. The development of natural language processing (NLP) approaches to item coding is key to being able to ingest transactions-level data at scale. Classifying products whose prices are scraped from the web raises similar issues. Although getting to this point has been a significant undertaking, data from alternative sources are beginning to play a major role in the production of official price statistics.

To this point, national statistical offices' use of scanner and web scraped data in price index production mostly has involved simply substituting prices from these sources for directly collected prices. Scanner data, however, contain information not only on prices but also on quantities. Recent research has investigated how item-level price and quantity data could be used in practice to produce price statistics that better account for consumer substitution and changes in the quality of the items that are purchased, together with internally consistent measures of nominal and real output (Ehrlich *et al.,* 2019, 2022). The statistical infrastructure for using scanner data in this way does not currently exist, but ongoing research has begun to sketch the outlines of what that infrastructure might look like.

3.2. *Improving the "Advance" Estimates of GDP*

Roughly a month following the end of each quarter, the BEA releases the initial or "advance" estimate of GDP. At that point, much of the data from the BLS, the Census Bureau and other federal agencies that will inform later estimates is not yet available. This includes data from the Census Bureau's Quarterly Services Survey (QSS), meaning that the Personal Consumption Expenditure (PCE) Services component of GDP is potentially subject to substantial revisions, making the data less useful to BEA's customers.

To address this problem, as an alternative to relying on existing methods for extrapolating PCE Services, BEA staff have investigated methods for "nowcasting" the QSS estimates with the goal of using the forecasts to reduce the size of the revisions between the advance estimates of PCE Services and those released 2 months later after the QSS data become available. One innovation in this research was the use of credit card transactions data and Google search queries in addition to official BLS statistics on employment and prices as predictors. The researchers tested a variety of machine learning models for making the predictions. The available time series were too short to divide the sample into training, test and validation data sets, as is typical in machine learning applications. Instead, the researchers evaluated the predictions based on the consistency of the improvement achieved across the models for different series (Chen *et al.*, 2022).

Predictions from nowcasting models are now prepared each quarter and predictions that differ from those based on the normal extrapolators are reviewed to determine whether the initial estimates should be adjusted. This has been done most regularly for the advance estimates of health care services and software investment (Erich Strassner, email to author, March 25, 2021). Similar nowcasting models could potentially be applied in other contexts at the BEA and elsewhere.

3.3. *Using Big Data to Produce State-level Retail Trade Estimates*

The Census Bureau's Monthly Retail Trade Survey (MRTS) collects information from approximately 13,000 retail and food services businesses each month. The survey collects data at the company level and there is no geographic component to the survey design. Together with the survey's relatively modest sample size, this means that reliable state-level estimates cannot be produced using the MRTS data alone. Historically, state-level retail sales estimates have been available only once every 5 years, at the time of the Economic Census. Point-of-sale data from the NPD Group, a third-party aggregator, have helped the Census Bureau meet the demand for more current geographically disaggregated data on retail sales. The Census Bureau released the first estimates from its initiative to produce monthly state-level retail sales data in September 2020. The new estimates are year-over-year rates of growth in sales for the retail sector as a whole (exclusive of non-store retailers) and each of 11 three-digit NAICS sub-sectors. They are available for the period from January 2019 forward.

The experimental estimates are based on a composite of top-down estimates that allocate total industry sales from the MRTS in line with annual state industry payrolls and bottom-up estimates that sum the sales of pre-selected multi-unit businesses covered by the NPD data, survey reporters operating in a single state, and

imputed values for other retailers. The top-down estimates assume that sales are proportional to payroll and that the month-to-month percentage changes in sales are the same in every state. The bottom-up estimates do not require these assumptions, but because the store-level data are incomplete and imputations are necessary, they can have a high variance. Incorporating more third-party data and publishing estimates of sales levels in addition to sales growth rates are goals for future iterations of this initiative (Hutchinson, 2021).

3.4. *Using Big Data to Produce Monthly Construction Statistics*

The Census Bureau produces statistics for residential construction based on the Building Permits Survey (BPS) and Survey of Construction (SOC). The BPS provides monthly information on the number and valuation of new privately-owned housing units authorized by building permits. The current BPS program design, introduced in January 2022, calls for monthly data to be collected from all local building permit offices that issue permits for an average of more than five units per year. This effectively makes the BPS a census and allows the production of monthly estimates not only for states but also for Metropolitan Statistical Areas and counties (U.S. Census Bureau, 2022c).[5] The SOC provides monthly information on housing starts, sales, and completions for the nation and for Census regions.

In principle, much of the information collected for the BPS from local building permit offices could be obtained from third party sources that already compile it for their own purposes. The Census Bureau is exploring the use of third party data to fill in for missing survey responses and perhaps even ultimately to replace the BPS. This has proven to be more challenging than originally anticipated. One significant complication is that the third party data include many different types of building permits. The data vendor codes the permits by type, but it can be difficult to distinguish permits for new housing units from other types of permits, such as those for commercial construction or remodeling projects, and the quality of the codes assigned by the vendor varies by jurisdiction. Another complication is that the Census Bureau has identified cases in which the address on the permit appears to lie outside the boundaries of the jurisdiction issuing the permit (Smith interview). Research on resolving these issues as well as on the possibility of using the third party data to support weekly estimates and estimates for sub-county geographies is ongoing (Studds and Abriatis, 2021).

The Census Bureau also is studying whether it might be possible to replace data on construction starts and completions currently collected through the SOC with information obtained from satellite images. Doing this successfully will require automating the categorization of images for identifying when construction starts and is concluded at residential construction project locations. An automated process might make it possible to collect information for a larger sample of projects and thus support more disaggregated residential construction activity estimates (Smith and Ferronato, 2021).

[5]Less active building permit offices account for only about 1 percent of all residential building permits. They are surveyed annually.

4. The Pandemic and the Demand for Real-Time Estimates

The work just described covers a spectrum of use cases for incorporating data from nontraditional sources into the routine production of official statistics. Agencies are substituting natively digital data for survey data (e.g., the BLS work to identify new ways of obtaining data for the CPI and the Census Bureau work to reengineer its collection of building permit data). They are using new sources of data to improve the preliminary estimates they publish (e.g., the BEA work to forecast late-arriving source data in order to improve the early GDP estimates). And they are using big data to support the production of more disaggregated estimates (e.g., the Census Bureau work on using credit card data to produce monthly state-level retail trade statistics and, potentially, its work on incorporating nontraditional sources of data into the construction statistics program). All of these examples, however, essentially represent improvements to existing data programs.

The pandemic changed the data landscape in some important ways. Especially during its early phases, policymakers were asking new questions that existing data programs had not been designed to answer. Even when data from existing programs could answer a question, they were arriving too slowly. In normal times, monthly data are more than adequate to guide fiscal and monetary policy decisions. Events moved so quickly at the onset of the pandemic, however, that monthly statistics often felt hopelessly out of date. Because the pandemic affected different communities in different ways, the demand for disaggregated data also grew.

4.1. *Real-Time Survey Data Collection*

The U.S. economic statistics agencies responded rapidly to the pandemic with new surveys designed to provide more timely information. In April 2020, in collaboration with several other federal agencies, the Census Bureau launched the Household Pulse Survey. The survey has collected information on topics including childcare, education, employment, energy use, food security, health, housing, household spending, Child Tax Credit payments, and COVID-19 vaccination. The survey went into the field on April 23, 2020 and the first data were released on May 20, 2020. Through July of 2020, the survey produced weekly state-level estimates; in later waves, it produced bi-weekly estimates and then, at the end of 2021, moved to a two-weeks-on, two-weeks-off estimation cycle. The sample for the Household Pulse Survey was drawn from households on the Master Address File (MAF) who could be matched to a phone number (available for 88 percent of addresses) and/or email address (available for 80 percent of addresses) from the Census Bureau Contact Frame. Data were collected using the Qualtrics online platform (Fields *et al.*, 2020).

Early in the pandemic, the Census Bureau also launched the Small Business Pulse Survey. This survey collected a variety of information about the pandemic's effect on small businesses, defined as single-location employer businesses with fewer than 500 employees. It included questions about business operations and the policies that businesses have adopted in response to the pandemic. The Census Bureau used the set of small businesses for which it had a valid email address as the sampling frame and data were collected online. The first wave of the Small Business

Pulse Survey went into the field on April 26, 2020 and the first estimates were published on May 14, 2020. With some gaps, weekly state-level estimates were produced through mid-April 2022, when the survey was discontinued.

The staff of the Census Bureau and other agencies involved with these surveys deserve enormous credit for their early recognition of the significant impact the pandemic was likely to have and for their work to design and implement these data collections. The decision to use sampling frames that allowed potential respondents to be contacted by email or text and to collect data online was crucial to producing data quickly. As anticipated, however, this came at the cost of far lower response rates than is typical for surveys conducted by the Census Bureau. The overall national weighted response rate for the first phase of the Household Pulse Survey, conducted in 12 waves between April 21 and July 23, 2020, averaged just 2.9 percent; the response rates in the second and third survey phases were higher, but still averaged just 9.3 percent and 6.8 percent, respectively (U.S. Census Bureau, 2022d). Response rates to Phases 1, 2 and 3 of the Small Business Pulse Survey, carried out between April 2020 and January 2021, were 26 percent, 23 percent, and 21 percent, respectively (Reamer, 2021b). Both surveys' estimates rely heavily on the answers of respondents with given observable characteristics being similar to the answers that would have been given by nonrespondents with the same observable characteristics. Still, the new survey data have filled major gaps in the information otherwise available and have been used by researchers and policy officials alike. In evaluating whether data from the Household Pulse Survey and Small Business Pulse Survey were "fit for use" (Groves and Lyberg, 2010), many people concluded that, despite their limitations, the relevance, timeliness and geographical disaggregation of these statistics made them valuable.

4.2. *Real-Time Big Data Estimates*

The response to policymakers' demand for data to help with navigating the pandemic was not restricted to the collection of new survey data. The pandemic prompted a tsunami of research using a wide variety of nontraditional data sources intended to shed light on its economic effects. I will describe just a few of these creative efforts.

As the pandemic began to spread, governments adopted policies intended to encourage social distancing and slow the spread of the virus. Policy officials needed to understand how the extent of in-person interaction was changing. Hoping to fill the gap in information about this, by late March of 2020, researchers at the University of Maryland had begun working to produce estimates based on mobile device location data of changes in travel outside the home compared to pre-pandemic levels. An April 13, 2020 press release announced their launch of an interactive dashboard that provided daily information down to the county level on changes in mobility, the extent of social distancing and other COVID-relevant metrics. Among the measures included on the dashboard were estimates of changes in travel-to-work behavior (Zhang *et al.*, 2021). The dashboard was updated regularly throughout the first year of the pandemic.

Policymakers were especially concerned about the pandemic's impacts on employment. The first official employment estimates to capture the pandemic's

effects, which provided a snapshot for the payroll period including April 12, 2020, were not released until May 8, 2020, almost 2 months after a national emergency had been declared.[6] Policymakers were desperate for data that were both more timely and more temporally granular. On April 15, 2020, based on their analysis of data from the payroll processing company ADP, researchers at the Federal Reserve Board published a paper containing week-by-week estimates of employment through the week ending April 4, 2020 (Cajner *et al.*, 2020). It was possible for the research team to move so quickly because, when the pandemic hit, they already had developed a methodology for using the data to produce weekly employment numbers that they had shown closely tracked the official payroll employment estimates. Their ongoing work with ADP data was part of a broader Federal Reserve Board research program that also has used scanner data from the NPD Group, job postings data from Indeed, and data on employee hours from Homebase, among other novel data sources, to produce high frequency estimates of consumer spending, job openings and small business activity. These estimates appear to have played an ongoing role in informing Federal Reserve Board thinking, but except for occasional research papers, are not readily available for public consumption (Stevens, 2021).

Another set of concerns for policymakers during the early part of the pandemic related to its impact on household spending. The Census Bureau released preliminary retail sales estimates for April 2020 on May 15, 2020, but because these statistics captured spending across the entire month, they were not well-suited to addressing questions such as whether the pandemic stimulus payments issued in mid-April of 2020 were bolstering overall consumption spending. There are no monthly or even quarterly official statistics on spending by households of different types.[7]

Here again, researchers were able to move quickly to provide useful information. On April 17, 2020, one academic research team released the working paper version of an article using data from a non-profit Fintech company to examine week-to-week changes in spending by the company's clients. Later research broadly confirmed the overall pattern of changes observed in this very early work (Baker *et al.*, 2020). On May 14, 2020, the JPMorgan Chase Institute reported estimates of changes in credit card spending through April 11, 2020 for a sample of 8 million households that were regular Chase credit card users. The JPMorgan Chase researchers had been working with the credit card data for several years prior to the pandemic. By linking the spending data to bank account information, they were able to segment the analysis by household income, finding that the pandemic had a modestly larger initial impact on spending by high-income households (Farrell *et al.*, 2020). On June 12, 2020, building on a collaboration with researchers at the Federal Reserve Board that had begun prior to the pandemic, BEA began to publish experimental estimates of weekly consumer spending in selected industries relative

[6]The BLS published employment statistics for the payroll period including March 12, 2020 on April 3, 2020, but little of the pandemic's effect on employment levels would have been visible in those numbers.

[7]The BLS Consumer Expenditure Survey (CEX) provides data on spending by different types of households, but the earliest CEX estimates covering any part of the pandemic period, for the 12-month period from July 2019 through June 2020, were not released until April 29, 2021.

to what it would have been absent the effects of the pandemic (Bureau of Economic Analysis, 2022). Then, on June 17, 2020, researchers at Opportunity Insights published a paper containing estimates based on private data sources of the pandemic's impact on a variety of outcomes. This included estimates of pandemic-induced changes in consumer spending based on data from Affinity Solutions, a data analytics company that gathers information through a daily feed of individual-level debit and credit card transactions. They found that there had been a markedly steeper decline in consumer spending in higher-income as compared to lower-income zip codes, due mainly to a dramatic shift away from in-person services and activities in the higher-income zip codes (Chetty *et al*,. 2020). Similar to the benchmarking of the Cajner *et al.* (2020) employment estimates to the official payroll employment numbers, Chetty *et al.* (2020) were able to confirm that, prior to the pandemic, changes in spending captured in the Affinity Solutions data had done a good job of tracking changes in spending in the Census Bureau's Monthly Retail Trade Survey.

The U.S. national statistical offices have of course been a key source of information about the pandemic's influence on the economy. Much of the early public understanding of the pandemic's impact, however, was informed by work with non-traditional sources of data done by academic and think tank researchers rather than by BLS, BEA or Census Bureau data.

5. Implications for the Role of National Statistical Offices in the Big Data World

The traditional role of the U.S. economic statistics agencies is to produce official statistics with well-documented properties and of the highest possible statistical quality that appear on a predictable monthly, quarterly or annual schedule. Historically, these mostly have been estimates based, either directly or indirectly, on survey data. Increasingly, in cases where survey nonresponse is a problem, the agencies are using administrative data to augment the survey responses. Naturally-occurring and third party private data, too, are beginning to play a larger role in the production of official U.S. economic statistics, as the BLS, BEA and Census Bureau explore the use of such data to replace hard-to-collect survey data, improve their early estimates or support more disaggregated statistics.

Based on recent experience, however, I would advocate for an expanded vision for these agencies that goes beyond producing a static array of regularly published official statistics. In this vision, the economic statistics agencies would be the go-to sources for information on emerging issues and concerns, whether based on regularly published statistics or produced using alternative sources of data as the need arose.

During the pandemic, academic and think tank researchers were important suppliers of information not routinely available from the statistical agencies. Relying on private actors to fill this role is not ideal. Data users typically have no easy way to evaluate the quality of estimates produced by an academic or think tank research team. Where detailed methodological information is available, some data users will be able to read the papers describing the construction of such estimates and assess their technical soundness, but most will not. In addition, the producers

of private economic indicators often provide limited documentation of their sources and methods. If the members of a research team have identifiable policy views, that may create uncertainty about the objectivity of the information provided. Further, researchers typically do not have either the motivation or the capacity to update estimates on a regular schedule or make updated estimates available in a form that is easy for potential data users to access. Researchers at a handful of organizations created regularly updated pandemic dashboards where anyone interested in the estimates they were producing could download them, but these efforts are the exception rather than the rule.

One possible reservation about involving national statistical offices in the production of quick-turnaround indicator information is that the quality of the estimates almost certainly will fall short of that normally associated with official statistics. The new sources of naturally occurring private data on which many of these indicators would be based typically do not represent the full population and the mapping of the data elements contained in these sources to the statistical constructs of interest can be far from perfect. It is entirely possible, however, for a statistical office to draw a distinction between its "gold standard" official statistics and more experimental measures that it produces in response to a specific need for information.

Indeed, there is a history of U.S. statistical agencies doing exactly that. Since 1999, the BLS has produced what until recently it termed the CPI-U-RS (Consumer Price Index for All Urban Consumers, Research Series). This series is an admittedly imperfect reconstruction of how the CPI-U, the headline CPI measure, would have behaved had it been constructed historically using current CPI methods (Stewart and Reed, 1999). When the BLS first began to discuss the idea of producing the CPI-U-RS, some BLS staff members expressed concern that it would fall short of the high standards for quality associated with other BLS statistics. I was the BLS Commissioner at the time; my view was that this series would provide information demanded by users and that the BLS staff were better positioned than anyone else to produce such a measure. Today, the BLS produces a number of CPI research series (Bureau of Labor Statistics, 2020b). The Census Bureau similarly has introduced a number of experimental series, including the monthly state retail sales estimates and the series based on the Household and Small Business Pulse Surveys mentioned earlier. Census describes these experimental series as "innovative statistical products created using new data sources or methodologies that benefit data users in the absence of other relevant products" and cautions that they "may not meet some of the Census Bureau's statistical quality standards" (U.S. Census Bureau, 2022e). Despite these cautions, users of the new data series seem very happy to have them. The BEA describes its new weekly retail sales measures in similar terms, noting that "(t)hey provide timely data but are a complement to, not a substitute for, the government's official data series" (Bureau of Economic Analysis, 2022).

New statistics developed for a specific purpose should not necessarily become a permanent part of a national statistical office's repertoire. An obvious consideration in making decisions about continuation of an experimental series is the extent to which interest in the information persists. At the onset of the pandemic, for example, weekly data from the Census Bureau's Household Pulse Survey met an urgent need for real-time information; as the situation stabilized, however, that

need waned. The survey fielding frequency was reduced to bi-weekly and then to once over a four-week period; at some point, it likely will be appropriate to consider discontinuing the survey altogether. The Census Bureau has already discontinued the Small Business Pulse Survey. Similarly, weekly statistics on employment based on alternative data sources were of enormous interest early in the pandemic, but as the economy recovered, there was less interest in weekly information. Another consideration that relates specifically to statistical series based on alternative data sources is whether they are likely to be sufficiently informative during normal times as to warrant their continued production. Such series are unavoidably noisy measures of the construct of interest. During a period when the economy has suffered a major shock, the signal-to-noise ratio in a measure may be high. In more normal times, however, the noise in a series may swamp the signal, making it less useful (Dunn *et al.*, 2021a).

What seems very much worth investing in is the capacity to respond quickly to demands for information to address new questions as they arise. Among other things, this will mean having agency staff who are actively engaged on an ongoing basis in working with nontraditional data from various sources. Experience has shown that learning how to extract useful information from novel data sources can be a time-consuming endeavor. Creating many of the new data series produced early in the pandemic was feasible only because researchers already had been working with the underlying data.

With the creation of its Center for Big Data Statistics in 2016, Statistics Netherlands became a pioneer in using novel sources of data to inform public understanding (Tjin-a-Tsoi, 2019). In the United Kingdom, ONS has created a Faster Indicators program that seeks to use real-time big data to provide more timely and more granular economic insights. Projects undertaken during the pandemic included using Barclaycard data to produce near-real-time information on consumer spending; data from Google Community Mobility reports to produce information on mobility patterns (e.g., travel to work, travel to retail establishments); and text extracted from business websites to learn about how they were responding to the pandemic (Arthur Turrell, email to author, April 12, 2021). In the United States, the Census Bureau has begun to highlight its production of experimental data series as an important agency activity (U.S. Census Bureau, 2022e). As part of this effort, working together with the BLS and BEA, the Census Bureau has embarked on a research collaboration with Opportunity Insights to "(e)xplore ways in which … alternative data sources may be used to complement and improve the data produced at the statistical agencies" (Dunn *et al.*, 2021b). The United Nations recently established a Network of Economic Statisticians charged with identifying possible steps towards a more agile and responsive system of economic statistics. One of the Network's initial focus areas has been access to privately held data for the production of economic statistics (Erich Strassner, email to author, May 3, 2022).

In my view, going down this path is essential. Statistical offices no longer have a monopoly on data provision and, if they do not respond with answers to important questions that key data users are asking, perceptions of their relevance will suffer. An agency that can be counted on to produce both the official statistics that data users rely on for consistent measurements of economic activity and the information needed for addressing new questions as they arise is likely to be perceived

more favorably. This might even put these agencies in a better position to obtain the funding needed to fully realize their mission.

The technical challenges to using new sources of data are often significant. A robust research program of the sort I envision for working with such data will require adequate resources, consistent leadership and buy-in from high-level government decision makers. Even recognizing the value of being able to respond agilely to emerging needs for information, the statistical agencies have finite capacity and will need to be selective about what they take on. The preceding discussion suggests a possible checklist for decisions about using naturally occurring data to produce experimental data series. Questions to ask include the following:

- Is there a significant demand for information that the normal suite of official economic statistics is not meeting?
- Are there sources of alternative data that could be used to create relevant experimental statistics? In the case of information that would be provided to meet an immediate need, could this be done sufficiently quickly to be helpful?
- Given the strengths and weaknesses of the alternative data, are the answers that the experimental series would provide sufficiently precise to be of value to data users?
- Can the experimental series be produced at reasonable cost?
- In the case of an experimental series created to meet an immediate need, what criteria will be applied, once that immediate need has passed, to decide whether the series should be made permanent or phased out?

One last point to make about the use of naturally occurring private data to improve official statistics or generate new statistics is that, in the United States, the decentralized structure of the statistical system may be an impediment to success. There are certainly arguments one can make for a decentralized system. Some have suggested that a statistical system with multiple smaller agencies may be more nimble and innovative. Others have argued that a decentralized system may be less susceptible to political pressure, though there also are reasons to think the opposite could be the case. On the other side of the argument, a centralized system could benefit from economies of scale and facilitate compatibility across different data series. Whatever one's view about the relative merits of centralized versus decentralized statistical systems, the growing potential of big data as a source of valuable information seems to me to strengthen the case for a reorganization that brings the BLS, the BEA and the Census Bureau's economic directorate together under one roof. Rather than having each of these agencies interacting separately with potential data providers and learning independently how to work with their data, an integrated approach could be more effective and efficient.

Indeed, the point about the value of collaboration can be made more broadly. Statistical offices around the world are wrestling with many of the same challenges regarding the use of new types of data to improve their economic statistics and working together would have many advantages. International bodies such as the Organisation for Economic Cooperation and Development and the United Nations can play an important role in coordinating these efforts.

Although there are considerable challenges to realizing the expanded vision I have sketched for the agencies responsible for the production of economic statistics, viewed in a positive light, it's an exciting time to be an economic statistician!

References

Abraham, K. G., C. F. Citro, G. D. White, Jr. and N. K. Kirkendall (eds), *Reengineering the Census Bureau's Annual Economic Surveys*, National Academies Press, Washington, DC, 2018.

Abraham, K. G., R. S. Jarmin, B. C. Moyer and M. D. Shapiro, "Big Data for Twenty-First Century Economic Statistics: The Future is Now," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 1–22, 2022.

Aladangady, A., S. Aron-Dine, W. Dunn, L. Feiveson, P. Lengermann and C. Sahm, "From Transaction Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 115–45, 2022.

Baker, S. R., R. A. Farrokhnia, S. Meyer, M. Pagel, C. Yannelis and J. Pontiff, "How Does Household Spending Respond to an Epidemic? Consumption during the 2020 COVID-19 Pandemic," *Review of Asset Pricing Studies*, 10(4), 834–62, 2020.

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

Bostic, W. G., R. S. Jarmin and B.C. Moyer, "Modernizing Federal Economic Statistics," *American Economic Review: Papers and Proceedings*, 106(5), 161–4, 2016.

Brick, M. and D. Williams, "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys," *Annals of the American Academy of Political and Social Science*, 645, 36–59, 2013.

Bureau of Economic Analysis. 2022. COVID-19 and Recovery: Estimates from Payment Card Transactions. Last modified April 7, 2022. https://www.bea.gov/recovery/estimates-from-payment-card-transactions. Accessed April 8, 2022.

Bureau of Labor Statistics. 2009. *Response Rate Status Report*, Survey Response Measurement Team. September.

Bureau of Labor Statistics. 2016. *Response Rate Status Report*, Survey Response Measurement Team. June.

Bureau of Labor Statistics. 2020a. *Handbook of Methods: Consumer Price Index*. https://www.bls.gov/opub/hom/cpi/pdf/cpi.pdf. Accessed March 22, 2022.

Bureau of Labor Statistics. 2020b. CPI Research Series. Last modified August 5, 2020. https://www.bls.gov/cpi/research-series/. Accessed April 8, 2022.

Cajner, T., L. D. Crane, R. A. Decker, A. Hamins-Puertolas, and C. Kurz. 2020. "Tracking Labor Market Developments during the COVID-19 Pandemic: A Preliminary Assessment," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2020–030. https://www.federalreserve.gov/econres/feds/files/2020030pap.pdf. Accessed April 10, 2022.

Cajner, T., L. D. Crane, R. A. Decker, A. Hamins-Puertolas and C. Kurz, "Improving the Accuracy of Economic Measurement with Multiple Data Sources: The Case of Payroll Employment Data," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 147–70, 2022.

Carson, C., "The History of the United States National Income and Product Accounts: The Development of an Analytical Tool," *Review of Income and Wealth*, 21(2), 153–81, 1975.

Chen, J. C., A. Dunn, K. Hood, A. Driessen and A. Batch, "Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 373–402, 2022.

Chessa, A. 2021. "Using Transaction Data in Consumer Price Index: Experiences at Statistics Netherlands," presentation to Panel on New Vision for Federal Data Infrastructure, Committee on National Statistics, National Academies of Sciences, Engineering and Medicine. December 9, 2021. https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a. Accessed April 5, 2022.

Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, and the Opportunity Insights Team. 2020. "How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data," unpublished working paper. https://iepecdg.com.br/wp-content/uploads/2020/06/tracker_paper.pdf. Accessed April 10, 2022.

Dunn, A., K. Hood, A. Batch and A. Driessen, "Measuring Consumer Spending Using Card Transaction Data: Lessons from the COVID-19 Pandemic," *American Economic Association Papers and Proceedings*, 111, 321–5, 2021a.

Dunn, A., J. Piacentini and S. Porter. 2021b. "Developing Experimental Statistics to Measure Economic Activity in Real Time," presentation to the Federal Economic Statistics Advisory Committee. June 11. https://apps.bea.gov/fesac/meetings/2021-06-11/Dunn-DESMEA-FESAC-2021-final.pdf. Accessed April 10, 2022.

Ehrlich, G., J. C. Haltiwanger, R. S. Jarmin, D. Johnson and M. D. Shapiro, "Minding Your Ps and Qs: Going from Micro to Macro in Measuring Prices and Quantities," *American Economic Association Papers and Proceedings*, 109, 438–43, 2019.

Ehrlich, G., J. C. Haltiwanger, R. S. Jarmin, D. Johnson and M. D. Shapiro, "Reengineering Key National Economic Indicators," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 25–68, 2022.

Ertl, H., S. Goussev, C. Li and R. Keshishbanoosy. 2020. "Statistics Canada CPI Modernization," presentation to Panel on Improving Cost-of-Living Indexes and Consumer Inflation Statistics in the Digital Age, National Academies of Sciences, Engineering and Medicine. December 7.

Farrell, D., F. Greig, N. Cox, P. Ganong and P. Noel. 2020. "The Initial Household Spending Response to COVID-19: Evidence from Credit Card Transactions," JPMorgan Chase Institute research report. https://www.jpmorganchase.com/institute/research/household-income-spending/initial-household-spending-response-to-covid-19. Accessed April 5, 2022.

Fields, J. F., J. Hunter-Childs, A. Tersine, J. Sisson, E. Parker, V. Velkoff, C. Logan and H. Shin. 2020. "Design and Operation of the 2020 Household Pulse Survey," U.S. Census Bureau working paper. https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_Background.pdf. Accessed March 28, 2022.

Fitzgerald, J. and O. Shoemaker. 2013. "Evaluating the Consumer Price Index Using Nielsen's Scanner Data," Bureau of Labor Statistics Office of Survey Methods Research working paper. https://www.bls.gov/osmr/research-papers/2013/st130070.htm. Accessed April 5, 2022.

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

Groshen, E., "The Future of Official Statistics," *Harvard Data Science Review*, 3(4), 2021.

Groves, R. M. and B. A. Harris-Kojetin (eds), *Innovations in Federal Statistics*, National Academies Press, Washington, D.C., 2017.

Groves, R. M. and L. Lyberg, "Total Survey Error: Past, Present and Future," *Public Opinion Quarterly*, 74(5), 849–79, 2010.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, *Survey Methodology*, Second ed., Wiley, Hoboken, NJ, 2009.

Heckman, J. 2021. "OPM Sets Bar for Agencies Hiring Data Scientists with New Job Qualifications," Federal News Network, December 29. https://federalnewsnetwork.com/hiring-retention/2021/12/opm-sets-bar-for-agencies-hiring-data-scientists-with-new-job-qualifications/. Accessed April 10, 2022.

Hutchinson, R. 2021. "New Census Bureau Experimental Data Product: Monthly State Retail Sales," webinar presentation. https://www.census.gov/data/academy/webinars/2021/explaining-monthly-state-retail-sales.html. Accessed March 23, 2022.

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

Konny, C. G., B. K. Williams and D. M. Friedman, "Big Data in the US Consumer Price Index: Experiences and Plans," in K. G. Abraham, R. S. Jarmin, B. C. Moyer, and M. D. Shapiro (eds), *Big Data for Twenty-First Century Economic Statistics*, University of Chicago Press, Chicago, 69–98, 2022.

Kurmann, A., E. Lale and L. Ta. 2021. "The Impact of COVID-19 on Small Business Dynamics and Employment: Real-Time Estimates with Homebase Data," Drexel University, unpublished working paper. July 30. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896299. Accessed April 10, 2022.

Merrington, L. and C. Smyth. 2020. "Use of Scanner and Web-scraped Data in Australia's CPI," presentation to Panel on Improving Cost-of-Living Indexes and Consumer Inflation Statistics in the Digital Age, National Academies of Sciences, Engineering and Medicine. December 7.

Meyer, B. D., W. K. C. Mok and J. X. Sullivan, "Household Surveys in Crisis," *Journal of Economic Perspectives*, 29(4), 199–226, 2015.

Moyer, B. C. and A. Dunn, "Measuring the Gross Domestic Product (GDP): The Ultimate Data Science Project," *Harvard Data Science Review*, 2(1).
National Academies of Sciences, Engineering, and Medicine, *Modernizing the Consumer Price Index for the 21st Century*, The National Academies Press, Washington, DC, 2022.
Office for National Statistics. 2021. "Transformation of Consumer Price Statistics." https://www.ons.gov.uk/economy/inflationandpriceindices/articles/introducingalternativedatasourcesintoconsumerpricestatistics/november2021. Accessed April 5, 2022.
Pan, Y., A. Darzi, A. Kabiri, G. Zhao, W. Luo, C. Xiong and L. Zhang, "Quantifying Human Mobility Behavior Changes during the COVID-19 Outbreak in the United States," *Nature Scientific Reports*, 10, 20742, 2020.
Paplomatas, A. 2021. "Alternative Collection Methods for Consumer Prices," presentation to the Bureau of Labor Statistics Data User Conference. https://www.bls.gov/cpi/additional-resources/alternative-collection-methods-price-programs-presentation.pdf. Accessed April 5, 2022.
Reamer, A. 2021a. "Federal Statistical Agency Uses of Private Sector Data: Study Findings to Date," presentation to the Panel on New Vision for Federal Data Infrastructure, Committee on National Statistics, National Academies of Sciences, Engineering and Medicine. https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a. Accessed March 29, 2022.
Reamer, A. 2021b. "Small Business Pulse Survey Phase 5 (5/17 to 7/18)-- Census invites comments (by 5/3)," EconSpark blogpost. April 2. https://www.aeaweb.org/forum/1914/small-business-pulse-survey-phase-census-invites-comments. Accessed April 10, 2022.
Smith, A. D. and H. Ferronato. 2021. "Modernizing Construction Indicators Through Machine Learning and Satellite Imagery," Proceedings of Statistics Canada Symposium on Adopting Data Science in Official Statistics to Meet Society's Emerging Needs.
Stevens, J. 2021. "Federal Reserve Board Experience Using Transactions Data," presentation to the Panel on New Vision for Federal Data Infrastructure, Committee on National Statistics, National Academies of Sciences, Engineering and Medicine. https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a. Accessed April 8, 2022.
Stewart, K. J. and S. B. Reed, "Consumer Price Index Research Series Using Current Methods, 1978-98," *Monthly Labor Review*, 122(6), 29–38, 1999.
Studds, S. L. and W. Abriatis. 2021. "Construction Re-engineering," presentation to Census Scientific Advisory Committee. https://www.census.gov/about/cac/sac/meetings/2021-03-meeting.html. Accessed March 23, 2022.
Tjin-a-Tsoi, T. 2019. "How CBS Intends to Use Big Data to Improve Official Statistics While Reducing Costs and Lowering Burden," keynote address, Conference on Big Data For Twenty-First Century Economic Statistics, Conference on Research in Income and Wealth, Washington, DC. March 15.
U.S. Census Bureau. 2022a. History: Economic Programs. https://www.census.gov/history/www/programs/economic/ . Accessed July 24, 2022.
U.S. Census Bureau. 2022b. ASM Methodology. Last modified October 8, 2021. https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html. Accessed April 9, 2022.
U.S. Census Bureau. 2022c. "Monthly Building Permits Survey Sample Methodology Change," U.S. Census Bureau working paper. https://www.census.gov/construction/pdf/bps_2022_faqs.pdf. Accessed March 23, 2022.
U.S. Census Bureau. 2022d. Household Pulse Survey Technical Documentation. Last modified March 14, 2022. https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html. Accessed April 7, 2022.
U.S. Census Bureau. 2022e. Experimental Data Series. Last modified March 4, 2022. https://www.census.gov/data/experimental-data-products.html. Accessed March 28, 2022.
Zhang, L., A. Darzi, S. Ghader, M. L. Pack, C. Xiong, M. Yang, Q. Sun, A. Kabiri and H. Songhua, "Interactive COVID-19 Mobility Impact and Social Distancing Analysis Platform," *Transportation Research Record.*, 2021. https://doi.org/10.1177/03611981211043813.