Measuring the Gross Domestic Product (GDP): The Ultimate Data Science Project

Brian C. Moyer¹ Abe Dunn²

¹National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland, United States of America,

²United States Bureau of Economic Analysis, United States Department of Commerce, Suitland, Maryland, United states of America

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.414caadb>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editor's Note: *For the inaugural article of Effective Policy Learning, Brian Moyer, Director of the U.S. Bureau of Economic Analysis (BEA), and Abe Dunn, Assistant Chief Economist at BEA, explain all that goes into capturing economic activity in one single number: the Gross Domestic Product (GDP). Data science can help the next generation of economic statistics to be even more relevant, timely, accurate, and detailed.*

Keywords: GDP, big data, data science, economic statistics, national accounting

---

How fast is the economy growing? Is it slowing down or speeding up? What sectors are driving changes in economic growth? To answer these questions, businesses, policymakers, economists, and many others turn to gross domestic product (GDP) and associated statistics. Answers to these questions often bring in their wake significant policy changes, such as the monetary authorities' call to lower or raise interest rates. GDP statistics may offer a magnifying glass on important sectors of the economy, such as housing and finance, of particular importance in the United States during the 2008 financial crisis. The GDP statistics can signal to economists that a recession may be looming or give confidence to businesses that consumers are freely spending. Around the world, the public pays attention to these figures during tumultuous economic events, such as the 2014 Brazilian recession and the debt-driven Greek recession of 2008. In the United States, these measures are produced by the Bureau of Economic Analysis (BEA), but similar estimates are produced by various statistical agencies around the world.

GDP is considered the broadest measure of a nation's economic performance. Technically, GDP is the sum of a few key components:

$$GDP = Consumption + Investment + Government Spending + (Exports - Imports)$$

)

Despite the theoretical simplicity, measuring the \$21 trillion U.S. economy is a complex effort. To accurately achieve this remarkable feat on a quarterly basis, BEA draws from more than 300 data sources each month—mostly from government surveys, private companies, and administrative records—and analyzes thousands of data points.

While this article focuses on U.S. estimates, statistical agencies around the world share the same challenges BEA faces when measuring the economy. Namely, how do we leverage hundreds of data sources to produce GDP every quarter? How do we produce accurate and timely estimates in the face of lagged data? And what role does data science play in economic statistics? This column provides an overview of the methods currently applied to estimate GDP in the United States and how data science plays an increasingly important role.

Current U.S. Methods

The U.S. GDP measures all goods and services produced in the United States, including associated statistics on production, profits, and income (refining information by industry, economic sector [for example, business or government], and location). These estimates are produced with a couple of key goals. First, BEA must release estimates in a timely manner that allows users, including businesses, policymakers, researchers, and households, to understand current trends as they are occurring. Second, BEA must release estimates at a fine enough level of detail to identify contributors to major trends and shifts in the economy.

To the first goal of timeliness, given that BEA receives different data at different points in time, GDP and associated statistics undergo regular revisions once new information becomes available. Some data flow in fairly quickly (e.g., monthly retail sales) while some take months to arrive (e.g., quarterly hospital revenues). The data with the most complete coverage, from the nation's Economic Census, come only once every 5 years. Yet BEA's economists make their first estimate of a quarter's GDP only about a month after the quarter ends. BEA economists turn these estimates around so quickly by working with the best data available and, when necessary, filling in the gaps with projections based on alternative data sources and, in some cases, historical trends. Once more data become available, BEA updates GDP and associated estimates for a given quarter. Each subsequent estimate improves the accuracy of GDP estimates and provides a clearer picture of the U.S. economy.

While the backbone of many U.S. GDP estimates is based on traditional sources from government surveys produced by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), there are many components of the GDP estimates that have, for decades, relied on nontraditional private sector and administrative data sources. Those include: *Wards' Automotive Reports* for autos; IQVIA for pharmaceutical sales; AM Best for insurance revenues and profits; and the Federal Deposit Insurance Corporation (FDIC) Call Reports for banking statistics, among many [others](#). For income information, BEA relies on Internal Revenue System (IRS) tax returns data as well as the Quarterly Census of Employment and Wages based on BLS, state

employment, and security agencies data. BEA's reliance on nontraditional data sources to inform certain components of the U.S. economy is exemplified in the field of digital and electronic products. As the range of digital and electronic products entering the U.S. market expanded, data users became interested in additional product detail. For example, BEA efforts to produce finer levels of detail include using point-of-sale scanner data to track consumer spending on electronics gear, data from HarrisX to estimate consumption on video streaming, cable, satellite, and other live TV services, and data from Nielsen to measure consumption on new products like vaping devices.

Measuring the “Real” Economy in the Digital Age

GDP statistics measure the total U.S. dollar amount of production in the nation's economy. However, for a more ‘apples to apples’ comparison, economists prefer ‘real’ output, which removes changes in spending due to price changes. This real output allows data users, for instance, to know whether an additional \$10 spent on apples reflects purchases of an additional 10 apples priced at \$1 each, or a \$1 increase in the price of the same 10 apples consumers purchased last quarter. To calculate real output, BEA starts with a measure of growth in current expenditures and then subtracts growth due to price changes (i.e., inflation). BEA uses detailed price inflation series computed by BLS, along with standard economic formulas, to ultimately obtain estimates of real growth for GDP and its components.

Despite the simplicity of adjusting for inflation, the dynamic nature of the U.S. economy throws a wrench into this seemingly straightforward calculation. Increasingly, goods that are purchased throughout the economy are from tech industries (e.g., computers and cell phones) that are continually innovating to provide more value per dollar spent. To account for these value changes, BLS adjusts prices for the changes in the quality of the products sold, and BEA incorporates these adjusted series into its real GDP and its components. For instance, the official estimates for the quality-adjusted price of computers show that a dollar spent on computers today buys 20 times the computing power of a dollar spent on computers two decades ago.

Quality-adjusted prices are often formed with the use of standard data science estimation tools, such as regression analysis, alongside alternative source data. In one recent example, BEA collaborated with other government and academic researchers to measure the quality changes for cell phones (Aizcorbe, Byrne, & Sichel, 2019). The researchers used multivariate regression analysis combined with private sector data on cell phone sales from the International Data Corporation (IDC) to measure quality-adjusted price changes. Using a variety of statistical models, BEA found that prices for cell phones fell quite rapidly after accounting for gains in their quality. Another example comes from the field of advanced imagery and medical equipment, which has undergone steady advancement in technology. For this industry, BEA found, using private sector data, strong evidence of improving quality and decreasing prices of such technologies once they had been adjusted for quality.

The Next Generation of GDP Statistics

While BEA has used nontraditional data sources for decades, there has been an explosion in recent years in the number of available alternative data sources as well as in the power of computing and artificial intelligence tools to fully leverage these new data resources. These new data sources and methods have spurred research into improving estimates in a variety of areas.

Getting More Accurate Information Faster

The first, early look at U.S. GDP has proven reliable, and, thus, is in high demand from businesses, investors, policymakers, and others who want timely information on the U.S. economy. While revisions to the statistics are expected as new information arrives, BEA endeavors to reduce revisions whenever possible. Doing so amounts to providing more accurate information to users more quickly. To achieve more timely and accurate production, BEA has made great strides researching the combined use of machine learning tools with alternative data sources.

One of the key inputs missing from the early U.S. GDP estimates is the Quarterly Services Survey (QSS), which measures revenues and expenses of businesses that provide services in industries such as health care (e.g., doctor's office visits or hospital care), legal services (e.g., drafting a will or preparing a contract), and spectator sports (e.g., putting on basketball or baseball games). Unfortunately, results from the QSS do not become available until nearly two months after the initial estimate of GDP. To get a more accurate, early read of these estimates, BEA researchers have developed models using an array of alternative data from credit card transactions and Google Trends, as well as labor force, wage, and price statistics. Under traditional estimation methods, such a large amount of data cannot be incorporated into economic prediction models because the number of variables entering the model exceeds the number of data points available for estimation. In contrast, off-the-shelf machine learning algorithms such as random forests, LASSO, and XGBoost are well-suited to this environment. BEA researchers have tested a model using these tools and they have found that it greatly reduces revisions for key areas of the service sector (Chen, Dunn, Hood, Driessen, & Batch, 2019). This model is currently applied to help inform the initial estimates of GDP, reducing revisions and providing more accurate and timely information to users of our data. In general, using alternative data sources with unknown properties can be tricky, but for BEA's application, the machine learning models are carefully trained to match the QSS data and are only considered if the forecasts meet stringent quality standards for accuracy and consistency. Moreover, the QSS data enter the official estimates when they become available.

Diagnosing the Health Care Sector

As health care spending in the United States has grown to over 17% of GDP and has dominated discussions of the U.S. economy, there has been increasing interest in better understanding this sector by viewing it from multiple perspectives. Currently, the official estimates for the health care sector track the inputs into treatment, such as doctor's office visits and prescription drug purchases, but do not track the actual output of the sector,

which is the treatment of a health condition. Health economic experts have advocated for alternative measures that focus more directly on the treatment of a condition. One such alternative measure is provided in BEA's [Health Care Satellite Account](#) (HCSA). The HCSA departs from traditional measures of health care by reporting at the disease level (i.e., the cost of diabetes or hypertension) rather than at the place of service (i.e., the cost of a hospitalization) (Dunn, Rittmueller, & Whitmire, 2015).

The estimates that the HCSA provides for over 200 diseases are only possible with the use of large claims databases that contain information on millions of enrollees and billions of claims from both public and private insurance claims data sources. This type of detailed estimate of spending by medical condition would not be possible using more traditional survey data sources because, for many conditions, traditional survey data include only small numbers of patients, and would yield highly volatile estimates. These estimates move us a step closer to understanding the value of dollars spent in this critical sector.

Putting Our Housing Numbers in Order

After health expenditures, housing is the next largest expenditure category in the U.S. economy, accounting for around 10% of GDP historically. The interest in the sector has grown considerably, particularly after the 2008 financial crisis. To improve our understanding and accuracy of measurement in this sector, BEA is working with [microdata from Zillow](#) which is a technology company focused on real estate and associated services. As part of Zillow's business, they maintain a large data set containing detailed information on millions of U.S. housing sales. BEA researchers have already investigated how these data can be used to form alternative measures for the housing sector, but this research is just the tip of the iceberg (Gindelsky, Moulton, & Wentland, 2019). Ultimately, the millions of transactions observed in the data mean that the housing sector can be examined at a much finer level of geographic granularity and frequency than would be possible using traditional survey data sources alone. At the same time, in order to continuously improve measurements of the housing sector, BEA is collaborating with the Census Bureau to combine Zillow data with more traditional data from the American Community Survey to augment and better understand the quality of the traditional data as well as the Zillow data. Together, this potent combination should lead to a steady pipeline of improved estimates for this critical sector of our economy.

What About the Economy Where I Live?

While BEA's national estimates of GDP garner a lot of media attention, there is also much interest in understanding local economies. The broad national trends may be quite different from what is happening in local geographic markets. BEA has produced GDP at the state level and for select cities for several years, but only recently did BEA release [GDP estimates for thousands of counties](#). These new county GDP statistics were made possible due to a substantial effort by BEA researchers to combine a vast collection of various public and private sector data sources. As BEA continues to produce these statistics on a regular basis, it is also diligently searching for new ways to enhance these statistics by exploring new data sources (e.g., credit card data) and

new methods (e.g., data science tools) to deliver more frequent and timely statistics or provide new insights at the local level.

What Else?

BEA is responsible for measuring a growing and changing U.S. economy, and it must continue to innovate to keep pace with these changes, while maintaining its reputation for producing accurate estimates of GDP. Cutting-edge work isn't limited to GDP statistics. BEA is engaged in a variety of other data projects, including: measuring 'free' digital goods (Nakamura, Samuels, & Soloveichik, 2017); measuring ride-sharing; and measuring the distribution of personal income (Fixler, Gindelsky, & Johnson, 2019). BEA has also conducted research into measuring household production, which includes nonmarket activities, such as preparing meals at home or home child care (Bridgman, 2016; and Kanal, & Kornegay, 2019). These items are traditionally excluded from GDP in large part because of data limitations.

Where Do Statistical Agencies Go From Here?

With the opportunities offered by the growth in alternative data sources and novel data science tools, also come challenges. How do statistical agencies, such as BEA, select the most promising projects that will lead to the greatest overall improvement in its statistics? How can agencies turn research projects into viable long-term data products that businesses, policymakers, researchers, and the public value? How can agencies produce existing statistics more efficiently using the new data science tools that have become available? Finally, and perhaps most importantly, how can agencies identify, hire, or train employees internally with the latest statistical and machine learning tools necessary to realize the full potential of these new opportunities?

Graduates in the data science field are often lured by private sector jobs or academia, both areas well-known for applying the latest data science tools. However, data science research at BEA and other statistical agencies lies somewhere at the intersection of academia and the private sector. Much like academia, statistical agencies tackle challenging problems that do not have off-the-shelf solutions, often vetting and communicating these new methods through research papers and publications in academic journals. And, much like the private sector, many agencies, such as BEA, embrace alternative data sources to inform practical solutions for data users (businesses, households, policymakers, and researchers, to name a few).

Measuring GDP is an undertaking that for decades has drawn upon many of the methods in the data science toolbox. However, recent exponential increases in data availability and advances in data science tools have expanded both the possibilities and the challenges for statistical agencies around the world. BEA is excited about the prospect of further engaging with the data science field, while prevailing over such challenges—guided by the goals of providing ever more relevant, timely, accurate, and detailed economic statistics.

Acknowledgments

The writing of this paper benefited substantially from the contributions of BEA staff, including Lasanthi Fernando, Kyle Hood, and Jeannine Aversa.

Disclosure Statement

Brian Moyer and Abe Dunn have no financial or non-financial disclosures to share for this article.

References

- Aizcorbe, A., Byrne, D. M., & Sichel, D. E. (2019). *Getting smart about phones: New price indexes and the allocation of spending between devices and services plans in personal consumption expenditures* (No. w25645). Cambridge, MA: National Bureau of Economic Research.
- Bridgman, B. (2016). Accounting for Household Production in the National Accounts, 1965-2014. *Survey of Current Business*, 96(2), 1–5. Retrieved from <https://apps.bea.gov/scb/2019/06-june/0619-household-production.htm>
- Chen, J. C., Dunn, A., Hood, K. K., Driessen, A., & Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In K. G. Abraham, R. S. Jarmin, B. Moyer, & M. D. Shapiro (Eds.), *Big Data for 21st Century Economic Statistics*. Chicago, IL: University of Chicago Press.
- Dunn, A., Rittmueller, L., & Whitmire, B. (2015). Introducing the new BEA health care satellite account. *Survey of Current Business*, 95(1), 1–21. https://apps.bea.gov/scb/pdf/2015/01%20January/0115_bea_health_care_satellite_account.pdf
- Fixler, D., Gindelsky, M., & Johnson, D. (2019) *Improving the measure of the distribution of personal income*. BEA Working Papers, Bureau of Economic Analysis.
- Gindelsky, M., Moulton, J., & Wentland, S. A. (2019). Valuing housing services in the era of big data: A user cost approach leveraging Zillow microdata. In K. G. Abraham, R. S. Jarmin, B. Moyer, & M. D. Shapiro (Eds.), *Big Data for 21st Century Economic Statistics*. Chicago, IL: University of Chicago Press.
- Kanal, D., & Kornegay, J.T. (2019). Accounting for household production in the national accounts: An update, 1965–2017. *Survey of Current Business*, 99(6), 1–9. <https://apps.bea.gov/scb/2019/06-june/0619-household-production.htm>
- Nakamura, L., Samuels, J., & Soloveichik, R. (2017). *Measuring the free digital economy within the GDP and productivity accounts*. BEA Working Papers, Bureau of Economic Analysis.
-

©2020 Brian Moyer and Abe Dunn. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.