

Federal Reserve Bank of Chicago

The perils of working with Big Data and a SMALL framework you can use to avoid them

Scott A. Brave, R. Andrew Butters, and Michael Fogarty

REVISED
March 2, 2021

WP 2020-35

https://doi.org/10.21033/wp-2020-35

Working papers are not edited, and all opinions and errors are the responsibility of the author(s). The views expressed do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

The perils of working with Big Data and a SMALL framework you can use to avoid them

Scott A. Brave[1]

R. Andrew Butters[2]

Michael Fogarty[3]

March 2, 2021

❖ The authors wish to thank Daniel Aaronson, Gadi Barlevy, Ezra Karger, Spencer Krane, and Diane Alexander for their insights on many of the big and traditional data sources referenced in this article.

[1] Federal Reserve Bank of Chicago, Economic Research. 230 S. LaSalle St. Chicago, IL 60604: sbrave@frbchi.org.
[2] *Corresponding author*. Kelley School of Business, Indiana University. 1309 E. Tenth St. Bloomington, IN 47405, 812-855-5768: rabutter@indiana.edu.
[3] Federal Reserve Bank of Chicago, Economic Research. 230 S. LaSalle St. Chicago, IL 60604: mfogarty@frbchi.org.

The perils of working with Big Data and a SMALL framework you can use to avoid them

Abstract: The use of "Big Data" to explain fluctuations in the broader economy or guide the business decisions of a firm is now so commonplace that in some instances it has even begun to rival more traditional government statistics and business analytics. Big data sources can very often provide advantages when compared to these more traditional data sources, but with these advantages also comes the potential for pitfalls. We lay out a framework called SMALL that we have developed in order to help interested parties as they navigate the big data minefield. Based on a set of five questions, the SMALL framework should help users of big data spot concerns in their own work and that of others who rely on such data to draw conclusions with actionable public policy or business implications. To demonstrate, we provide several case studies that show a healthy dose of skepticism can be warranted when working with and interpreting these new big data sources.

Keywords: big data, economic statistics, business analytics, forecasting

JEL Codes: C53, C55, C80, C81

Introduction

The appeal to the use of "Big Data" to justify changes in public policy, explain fluctuations in economic activity, or drive business decisions is becoming nearly universal in policy circles and executive suites. There is perhaps no better example of this than current events, where researchers, journalists, and firms in greater numbers than ever before have used the large and novel datasets of private and public companies (see, for example, The Economist (2020)) to measure the health and economic consequences of the novel coronavirus and the global recession that transpired as a result of it.[4] In this article, we take a closer look at the general efficacy of "Big Data" in this context and provide a framework to guide its use for both consumers and producers of these new datasets.

The primary advantages offered by many big datasets are their granularity and high-frequency, which allows decision makers to track developments among disparate populations in real time.[5] These advantages can be beneficial when compared to more traditional statistics that are usually reported with a substantial lag or offer only coarse levels of both spatial and demographic disaggregation (e.g., see Abraham et. al, 2020). But with these advantages also comes the potential for pitfalls. Most traditional data sources have the benefit of being time-tested, with their vagaries well understood and publicly documented. This is rarely the case for many big data sources—at least in their infancy—and often it can lead big data users to draw inferences that they would otherwise not make.

Here, an example is helpful in providing context: The *Missing* 2013 Flu According to Google Trends (Lazer et. al, 2014). Google Flu Trends (GFT), was a predictive "Big Data" model designed to use Google search activity to predict the proportion of doctor visits for influenza-like illness. GFT made headlines in 2013 for predicting twice as many visits than the Centers for Disease Control and Prevention (CDC), a forecast that turned out to be a very large mistake.[6] To make its predictions, GFT matched the histories of 50 million search terms to a single time series of data. Lazer et. al cite "big data hubris", or the assumption that the size of a dataset alone can overcome traditional issues of measurement and statistical inference, as one of the primary causes of the large forecast error made by GFT. In essence, the GFT model drastically overfit the universe of Google search terms to just 1,000 data points. What appeared big on the surface in the end was actually rather small.

---

[4] See, for example, Chetty et. al (2020) who use a wide variety of private sector data sources to construct a database to track the real-time impact of the COVID-19 pandemic at https://tracktherecovery.org/.

[5] An example includes Alexander and Karger (2020) who use county-level data on consumer spending, small business revenues, and cellphone location to evaluate the impact of stay-at-home orders on consumer behavior during the early months of the pandemic.

[6] The large and arguably predictable mistake came despite the GFT model receiving considerable praise for capturing the power of big data resources (e.g., see Goel et. al. (2010), McAfee & Brynjolfsson (2012)).

In what follows, we lay out a framework that we have developed in order to help interested parties as they navigate the big data minefield and hopefully spot concerns in their own work and that of others who use big data to draw conclusions with actionable public policy or business implications. We refer to this framework as SMALL in an obvious play on words meant to draw a distinction between what big data sources purport to be and what they often can be instead if not used carefully. Based on our own experiences and other well-known examples that we describe below, we argue that a healthy dose of skepticism can be warranted when working with and interpreting these new big data sources.

We are by no means the first to go down this path. In fact, several of the concerns that we raise in our framework can be found in the analyses of others from the firm's perspective of the generation and use of big data (e.g., Sivarajah et. al, 2017).[7] The focus of this article—and the SMALL framework more generally—is instead on the *bigger picture* of the critical dimensions that users of big data should be aware of in considering the insights that are drawn from these data sources. For this reason, the case studies of this article share a significant time series component among them in using big data sources to track economic activity and evaluate public policy outcomes.[8] Even here, we are not the first to raise many of these concerns, but we hope that by providing a novel framework to guide the interested reader through our collection of "cautionary tales" they will be better equipped to engage with their next big dataset and the public policy debates that often accompany their use.

Perhaps more than anyone, the lessons we uncover with the SMALL framework should be particularly useful for those involved with the creation and dissemination of economic statistics and business analytics. As statistical agencies are beginning to make inroads into the use of big data and data journalists are starting to regularly include them in their work, taking the time to step back and re-examine what it is that is being captured and not captured by these big data sources has never been more important. We show below that this introspection can be made easier by asking a simple set of five questions, one for each element of the SMALL acronym. Not only should asking these five questions lead to more accurate interpretations of big data sources, but it should also help to avoid making the kinds of mistakes that the GFT model made in areas where public policy is concerned.

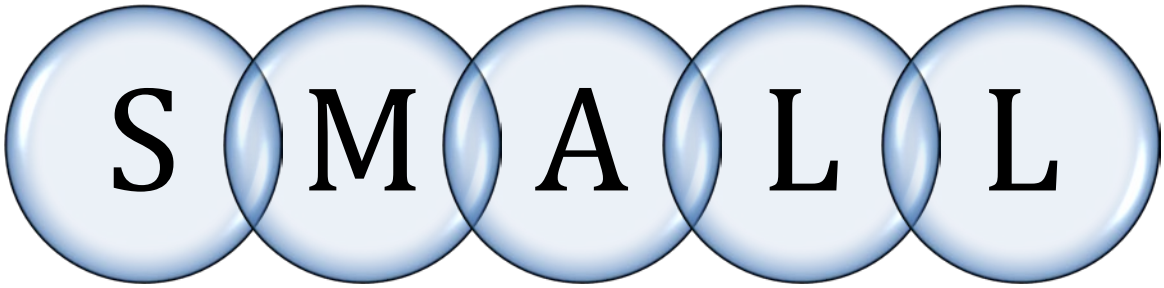
---

[7] Previous work has also shown that with the new opportunities big data provides (e.g. Gupta, et. al., 2020; Early, 2015; Njuguna and McSharry 2017; Pérez-Martín, A., et. al., 2018; Pham and Stack, 2018; Wang and Hajli, 2017), come a new set of challenges (Lee, 2017). Some of the challenges outlined range from quality issues, privacy considerations, and the required infrastructure typically associated with managing big data (Lee, 2017; Nagle, Redman, and Sammon, 2020) to the ability of a firm to monetize the data it collects and produce meaningful analytics from it (Walker, 2015).

[8] The use of Big Data has also become commonplace in settings that have less of a time sensitive nature to them and are instead more specific to firm performance. An example includes Netflix's predictive modelling exercise about the individual movie preferences of its customer base. While the framework we propose has implications and relevance for these applications as well, we leave this discussion to future work.

The SMALL Framework

The SMALL framework poses *five* relevant questions that all interested parties should ask themselves to guide their use of big data, each of which we explain in more detail in subsequent sections:



1) How are the data **S**ampled? (e.g., are the data ***representative*** in all dimensions?)
2) How are the data **M**easured? (e.g., is it ***consistently constructed over time***?)
3) How is it **A**ssembled? (e.g. are ***statistical methods*** used to alter the data?)
4) Do the data exhibit reporting **L**ags? (e.g. are there ***reporting delays***?)
5) Is it a **L**eading indicator? (e.g., is it ***predictive of traditional data***?)

At this point, readers may be asking themselves: "Ok, why do I care as long as I go in knowing the data has limitations?" And in many ways, those sentiments are exactly correct. The problem, however, often comes in defining what those limitations are in the first place. The SMALL framework should make this process more transparent, and in doing so, make the impact of these limitations on the subsequent predictions and conclusions drawn from big data sources easier to convey. For instance, the use of increasingly available retail scanner datasets to understand the pricing strategies of firms provides the perfect archetype for our first question on data **sampling**. More specifically, the use of these big datasets illustrates the tradeoff between the representativeness and granularity (or *scale and scope*) of a big data source captured in the first question of SMALL.

On the one hand, there is the work of Chevalier et. al. (2003) who use very detailed scanner data on product-level weekly sales, transaction prices, and wholesale prices from the grocery retailer Dominicks, a now defunct Chicagoland grocery chain, to investigate why prices seemed to *fall* during peak periods of demand (e.g., beer in the summer). An advantage of this big data source—albeit from a *single* retailer—was that it provided direct measures of *many* dimensions of the Dominicks' retail strategy including, pricing, promotional activities, and profit margins. This granularity afforded the authors an unprecedented opportunity to provide direct evidence indicating that the pricing patterns

observed were most consistent with a "loss-leader" type pricing strategy by the retailer.[9] It then took more than a decade for DellaVigna & Gentzkow (2019) to show that some of Dominicks' pricing strategies were not representative of the typical retailer in this industry.

On the other hand, there are the *multiple*-retailer scanner datasets (e.g., provided by IRI and Nielsen) that also provide price and quantity data at the product-store-week level. These datasets have allowed researchers to answer a number of questions related to both retailer decisions, such as pricing to market (Gopinath et. al., 2011), in-store advertising (Hitsch et. al., 2020), the pass-through of local cost shocks (Butters et. al., 2020), and broader macroeconomic phenomena, such as price rigidity over the business cycle (Coibin et. al., 2015). However, a typical drawback of these much broader—more representative—datasets is that they provide consistently measured information on *fewer* dimensions, or a lower level of granularity. For instance, in both the IRI and Nielsen datasets retail prices are not consistently measured across retailers, and wholesale price information is not systematically available. Furthermore, even the most comprehensive assessments of pricing practices with scanner data resources of this type typically do not include key players like Walmart and Whole Foods – and so are likely to miss a significant component of the price dynamics that move the grocery market as a whole.

The question then is: what can we learn from such big data sources? In the case of the Dominicks dataset, they might be incredibly revealing of a *particular firm's* business strategy or its customer base. But depending on the firm in question—not all that relevant for understanding the *industry* on a broader scale. Our view is that this is typically still more of a "feature" (as opposed to a "bug") of the big data source, in so much that the methods required to arrive at those insights in the first place become much more transparent. They come *from the data*, not some "eyewitness account" or a "black box" modelling approach which often becomes noted as anecdotal only after the *fact* reveals to be *fiction*. For the Nielsen and IRI datasets, generalizability becomes less the issue and the transparency or clarity of the data available becomes more the burden.

Beyond the obvious issues of scale and scope with big data sources, there still remain a number of other potential concerns, evidenced by the fact that our SMALL framework does not simply end with one question. Big data has been applied to a number of contexts, from macroeconomic forecasting to public policy evaluation, and each new context brings with it a new set of potential concerns. The remainder of the SMALL framework categorizes these concerns into four areas: the consistency of **measurement** over time, the statistical method of **assembly**, the presence of reporting **lags**, and predictive, or **leading,** ability.

To help introduce these other areas, we return to the use of Google Trends. We already discussed the GFT model and its shortcomings, but setbacks such as what occurred in 2013 have not prevented Google search activity from quickly becoming one of the most commonly used big data sources across a wide variety of disciplines. This is perhaps not

[9] Another Dominicks' pricing strategy uncovered by the scanner data was their use of highly localized (within Chicagoland) "neighborhood" pricing zones. That is to say, the price of a can of tuna during Lent in a Dominicks in Hyde Park was typically not the same as that same can of tuna in a Dominicks' in Rogers Park.

surprising given how easy Google Trends is to use and how far the Google search engine has penetrated into our daily lives. For a big data user, to say that a conclusion is derived from patterns found in Google searches for common words of phrases is to invite a level of familiarity that everyone can understand. Here, the level of transparency could not be more obvious, or so it would seem on the surface. To understand what one is really looking at with Google Trends requires going under the hood; and those who have done so have pointed out various features that might give us pause before jumping to conclusions.

Consider the use of Google Trends as a predictor. Choi and Varian (2010) showed that search activity related to unemployment correlated strongly with initial claims made for unemployment insurance (UI) during the 2007-09 U.S. recession. Ever since, Google Trends search activity has been used to predict everything from home prices (Beracha and Wintoki, 2013) to public health outbreaks (Towers, et. al., 2015). There are several features of the correlation underlying any such analysis of economic activity with Google Trends that readers should be aware of before drawing conclusions on the nature of this relationship. The exact specifics of the underlying construction of Google Trends are the proprietary trade secrets of Google. So, while the free and publicly available nature of Google Trends has led to its use in a wide array of applications and analysis—many aspects of the data are still subject to concerns our SMALL framework highlights.

First and foremost, Google Trends is itself a sample. While still ***representative*** of overall search activity, this means that at any given moment the data that we might pull from Google's website could be slightly different than the one you might pull. Second, the particular search term(s) used in Google Trends can matter a lot. There is often a tension between designing a search inclusive enough to be representative of everything desired but yet still exclusive enough to avoid capturing activities beyond the scope of study. For example, with the use of "jobs" as a search term one must consider the potential role of news-related search. As an underperforming labor market grabs the attention of the news media, the search activity of readers interested in the most recent analysis of the "jobs report" will necessarily confound the activity of individuals using similar search terms but in an effort to find a job. This "reverse causality" indicates that there is value in being able to distill which variation in the Google Trends data is best to use for each particular application.[10] At least in this example, however, there is still some underlying economic relationship of interest. This is not always the case. For instance, the dramatic uptick in individuals searching for "jobs" in October of 2011 was unrelated to any labor market considerations, but instead driven by the passing of Apple co-founder Steve Jobs.

These are examples of what we refer to as a problem with ***consistency of measurement***. While it might not necessarily matter all that much depending on how Google Trends is being used, for some applications it can matter a lot. One aspect for which it certainly matters is that the data made publicly available are an index and not a raw count of

[10] Brave et. al (2020a,b) show that the correlation between the time series of Google search activity for the unemployment topic even at the individual county level was reflective of changes in initial unemployment insurance (UI) claims during the Covid-19 pandemic.

searches. More specifically, the data are normalized such that the period that experienced the highest amount of search activity for a term (as a share of all searches) is scaled to be 100. Thus, all the remaining periods reflect the level of search activity relative to that period.[11] In effect, this means that two individuals using the same search term for different but overlapping points in time could receive different answers for the points in time that their queries share in common. Here, a problem with consistency bleeds into the issue of ***statistical methodology***. It is not uncommon for big data sources to receive some form of pre-treatment of this nature that users must take into account when analyzing the data.[12]

All that being said, we wish to avoid from this discussion giving the impression that Google Trends is necessarily a problematic big data source. On the contrary, in many applications it can be extremely informative once the considerations above are taken into account. In fact, on the remaining dimensions of our SMALL framework (***reporting delays*** and ***predictive ability***), it has been shown to be an extremely valuable high-frequency measure of economic activity (e.g., Aaronson, et. al., 2020a). Rather, we wish to leave the reader with the impression that even the best of the best big data sources still fall subject to the concerns raised by our SMALL framework. Next, we take each component of our framework and illustrate through several additional case studies how using it helps one avoid the potential pitfalls of big data.

1. How are the data [S]ampled?

Unlike the methods used to construct official economic and population statistics (e.g. the Census Bureau's host of surveys on activities like Retail Sales and Construction Spending), big data sources are commonly not collected and structured in a way that would allow them to easily address a broad economic question (e.g. "Is economic activity for the U.S. declining?"). Economists tend to call this an issue of *external validity*. While a dataset may be perfectly fine for addressing the characteristics of a firm's customer base, that customer base need not be representative of a firm's industry or the broader economy. This is not a critique of big data collection itself so much as it is just a reflection of the different aims of the private sector firm and the public sector data agency.

As an example, when the Census Bureau goes to collect data on retail sales, it starts by designing survey methods which aim to capture the broad population of retailers. Even though they can't possibly survey all of them (thereby giving up to a degree on the "big" aspect of big data), they still want to be as representative of the universe of retailers as possible. This requires that some structure be placed on the data collection process itself. Private sector big data tends to evolve along a different dimension, focusing instead on

[11] Where one is examining multiple geographies, the normalization is such that the period and geography with the most search activity for the search term is set to 100. For more discussion surrounding the construction of Google Trends data, see Stephens-Davidowitz (2014).

[12] For a survey of the empirical methods developed to handle big datasets, see Varian (2014).

broadly capturing the universe of a smaller population, i.e. the universe of a firm's customers or a website's users. In statistical terms, these private big data sources are often *convenience samples* of a firm's customers or a specific type of business activity that is more easily recorded or tracked, often as the means for some other business function.[13]

To further illustrate, consider the construction of the Consumer Price Index (CPI) by the U.S. Bureau of Labor Statistics (BLS). As Friedman et. al (2019) discuss, the BLS recently began experimenting with using big data in the CPI. In other words, rather than send people out to price particular items in stores, as they have traditionally done, they reached out to large retailers for this information in an effort to reduce the costs of constructing the CPI. On the surface, this seems like a very straightforward application of big data to an important problem, i.e. measuring changes in the general level of prices. However, in practice, it was not as successful as one might first imagine. When large retailers record price data, they are primarily interested in keeping track of whatever will best help them to maximize profits. This is not the goal of the BLS data collector who is instead primarily interested in collecting the data in a way that best facilitates the computation of the CPI (e.g. taking into account hedonics and related quality-adjustment considerations).

For this reason, we prefer to think of big data as a sample drawn from a broader universe of interest, the key feature of which is that it is highly unlikely to be drawn at random. The data are often generated as the outcome of a conscious decision-making process by an individual or group to engage in some sort of behavior (e.g. purchase a product, spend time searching a website, etc.) The aim of the data collector in this case is to be as unstructured as possible in order to cast a wide net for these types of actions. Therefore, using big data in other contexts often requires that a bridge of sorts be built to map the more narrow activity that is actually measured to the more general concept of interest.

In some cases, big data providers build this bridge for the researcher by attempting to replicate sampling strategies which reduce the scale of the data in order to better match the scope of a related traditional dataset.[14] How this is done can often bring about other concerns in our SMALL framework, but at least in terms of representativeness it is a step forward. Of course, transparency in any case is key. These are often private firms selling their data and not government statistical agencies charged with making all of their methods publicly available. Even when the data itself is not for sale, there is no guarantee that the proprietary methods used to create it are fully explained and documented; and if

---

[13] As an ironic example, in its infancy Netflix began "collecting" the preference data of its customers via the "queue" of movies a subscriber would populate as a means to ensure that its supply chain was appropriately equipped to handle the rising demand for its rental DVDs. As of 2020 Netflix's DVD business is virtually non-existent, but the data regarding its customer's preferences remain one of its most highly valued assets.

[14] An example that builds off of our earlier discussion of the CPI is the PriceStats measure of online prices that tracks pretty closely with the CPI on a nonseasonally adjusted basis. More information on this big data source can be found in Cavallo and Rigobon (2016) and Cavallo (2017). We thank Gadi Barlevy for pointing this data out to us.

they are, they can still be subject to vigorous debate and interpretation, particularly when the data is used to guide public policy decisions.

The use of cell phone mobility data during the pandemic is a good example. One of the premier sources of information used by public health officials to measure the potential for the spread of Covid-19, this data has been highly documented and studied in recent months.[15] Still, its representativeness has been called into question by Coston et. al (2020), who by linking the mobility data with voter rolls find that older and nonwhite voters are less likely to be captured by this data. Given the higher mortality rates among these populations during the pandemic, such a result, if it holds more broadly, is potentially of concern to public health officials. Below, we provide additional examples of representative issues based on big data sources of labor market information during the pandemic.

Application: The Covid-19 pandemic labor market

The Covid-19 pandemic here in the U.S. has caused some of the fastest and largest changes in economic activity on record. Although statistics from the Bureau of Labor Statistics' (BLS) Employment Situation report are among the timeliest of the traditional economic statistics used to track these changes, there has still been considerable demand from policy makers and business professionals for even more up-to-date labor market data. To illustrate a few of these newly popular big datasets, figure 1 shows a big data measure of weekly average hours worked and employment relative to similar measures from the BLS. In each case, we have been careful to align the timing and frequency of the BLS survey reference week to the big data source to facilitate comparison, but it is still worth noting that both big data sources are available at even higher frequencies. To highlight changes over time, all of the measures in the figures have been indexed against a similar week in January of 2020 and show deviations from this level through November 2020.

The measure of hours worked comes from Homebase, a firm that provides time scheduling software. While both measures capture a steep decline in hours worked in March and April followed by a recovery and stabilization to a lower level by June and July, the magnitudes of the changes are quite different – the decline in the Homebase data is about three times as large as the decline in the BLS data. This discrepancy is largely driven by the nature of the Homebase dataset itself. The firms in the Homebase sample differ from the universe of firms tracked in the BLS survey along at least two important dimensions. First, the Homebase firms tend to be concentrated in industries such as restaurants, retail, and personal services that were most affected by the virus and the public health measures that were imposed to mitigate the spread of the disease. Additionally, the Homebase sample skews towards smaller businesses, because the firms that use the Homebase product are

[15] We provide several examples of this data in later sections of the paper. Besides these examples, there is also the well-known data made publicly available by SafeGraph at https://www.safegraph.com/. It forms the basis of the Mobility and Engagement Index described in Atkinson et. al (2020) that is published weekly by the Federal Reserve Bank of Dallas.

"largely individually owned/operator-managed businesses."[16] These underlying differences illustrate why the Homebase data are not necessarily representative of the larger universe of firms, and this lack of representativeness needs to be taken into account when interpreting data from private big data sources like Homebase.

There are many other examples like the Homebase data that have crept into the popular press as a result of the pandemic.[17] While each has value in its own right, it is important to keep in mind their inherent limitations. While it is difficult to look at figure 1 and come away with anything but a sense of awe at the impact of the pandemic on labor market activity, it is also quite apparent that not everyone was affected equally by it. This is an important aspect of the SMALL framework's first question. Representativeness can have stark implications for public policy and welfare when the impacts of a policy or event are experienced unevenly across different populations. Big data can help to identify those effects, but we must first know what it is we are and are not capturing in the data. For instance, those workers who were able to successfully transition to remote work arrangements during the pandemic likely had a very different experience. For some of them, hours worked may have even increased with reductions in commuting time and lower productivity stemming from an increase in work and home life demands.[18]

The bottom panel of figure 1 provides a second example. Here, we compare changes in the level of employment from the BLS survey with a measure constructed by the Opportunity Insights (OI) lab from a variety of payroll and financial management service providers.[19] The OI series is constructed in a sophisticated way that attempts to account for both firm size and industry in order to better match with the overall composition of employers. This matching process can be viewed as an example of the type of bridge we discussed above linking the big data source to the economic activity of interest. The declines and subsequent increases in employment during the pandemic in the OI data are much more in-line with the BLS survey. In this case, the bridge that was built is a good first step toward addressing the SMALL framework's first question, but it is likely one of only several steps toward satisfying the remainder of the framework.

---

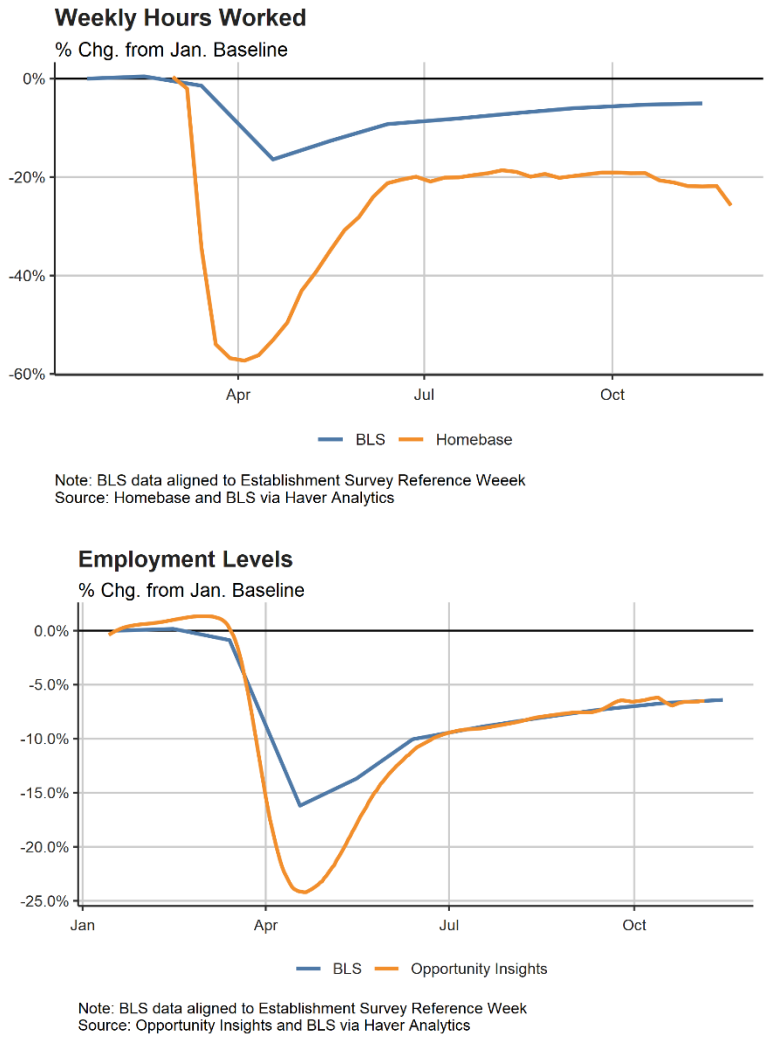
[16] The Homebase Coronavirus impact data are described here: https://joinhomebase.com/data/.

[17] Other examples include data from Opentable on restaurant reservations (available at https://www.opentable.com/state-of-industry), airline passenger counts reported by the Transportation Security Administration (https://www.tsa.gov/coronavirus/passenger-throughput), data on hotel occupancy and revenue from STR (https://str.com/data-insights/news/press-releases), and movie box office data from Box Office Mojo (https://www.boxofficemojo.com/date/?ref_=bo_nb_hm_secondarytab).

[18] For evidence of this, see the survey results discussed in Barrero, Bloom, and Davis (2020).

[19] These firms are Paychex, Earnin, Intuit, and Kronos.

Figure 1



2. How are the data [M]easured?

While representativeness issues most often take the form of concerns about the cross-section of data points (i.e. observations within a specified time period), a whole host of additional problems arise when the researcher is also interested in making comparisons across time periods. Here, problems often take the form of *internal validity* and concern the ability of a dataset to accurately capture what it purports to describe. For instance, a dataset that shows foot traffic patterns at geographic places of interest may be used to judge the state of business activity. But what happens when the way this data is collected differs across two points in time? If the researcher's aim is to draw comparisons across time in order to identify changes in patterns of business activity, concerns of this nature cloud all aspects of the ability of big data to be useful even for its intended purpose.

As an example, consider what would happen if the underlying number of "sensors" in such a dataset were to double unbeknownst to the researcher.[20] This would likely cause the number of "hits" to increase quite substantially, but does it represent a true increase in activity? The answer is likely to be *NO*, and if the researcher were to become aware of this they would be forced to throw up their hands in frustration and abandon the use of at least portions of the data. More often than not, issues of this nature prevent building the kind of bridge that is necessary to connect big data and official statistics.

Official statistics, however, are also not immune to these considerations, but they typically have developed time-tested methods to address them which make comparisons over time feasible. One example is the impact of firm entry and exit which affects everything from the measurement of employment, hours, and wages to firm productivity. Two separate—but often related—considerations are important in regards to handling the sometimes prevalent entry and exit of firms or individuals in a dataset. The first consideration involves how the entry and exit might influence interpretations of the *static* relationships amongst variables in the data in a given time period. The second consideration involves how the entry and exit requires accommodations when trying to recover *dynamic* implications of the data across time periods. In both cases the underlying complication hinges on the fact that the very act of a firm (or, individual) entering/exiting a data panel is typically for reasons that are far from random.

Most firms fail (exit) within the first year, and for reasons typically tied to profitability and other aspects of the firm (e.g., Abbring and Campbell, 2005). Alternatively, a firm's (de-novo) entry into a particular market is usually construed as the culmination of a series of particularly good events. If these deviations result (or, came about) from systematic differences in how these firms handle employment, pricing, or investment decisions—then a consumer of such datasets will have to think carefully about how to handle the classic *simultaneity bias* in even the static relationships in the dataset that one might be trying to recover.[21] This form of bias often manifests itself when there are aspects of demand or supply that are observable to the firms, individuals, or consumers *but not to the big data producer or consumer*. Perhaps, the classic example of this issue is if *firms* know how productive they are, but the *data analyst* does not and is trying to infer it from information on output and labor use. Consequently, any successful attempt in estimating production functions has had to overcome the fact that firms who are more productive typically use more labor (e.g., Olley and Pakes (1996) and Griliches and Mairesse (1999)).

Because the entry and exit of firms or individuals within a dataset is typically driven by specific factors that might not always be perfectly revealed in the dataset (e.g., last year's

---

[20] A real world example of this is the use of the Opentable online reservation system being used by colleges and universities during the pandemic to reserve dining hall space. If one were not careful to filter out these additional reservations, it would potentially result in a big spike in the data in the fall of 2020 once the academic calendar resumed that had little connection to how this data has been used to track restaurant activity during the pandemic. We thank Rick Mattoon for pointing this out to us.

[21] For a seminal reference on some of the issues associated with this econometric problem, see Griliches and Hausman (1986), and Arellano and Bond (1991).

profits and/or productivity), it also creates complications in efforts to recover the dynamic evolution of key features of the data of interest. A well-studied example of this is how the entry and exit of firms with varying productivity levels translates into the evolution of aggregate productivity in an industry or the economy (e.g., Olley and Pakes, (1996), Collard-Wexler and De Loecker (2015)). Here, the critical issue at the heart of the analysis rests on jointly accounting for how the entry/exit of firms in the dataset *reveal* information about their productivity levels—which then are appropriately attributed to how the differences in productivity levels of firms entering and exiting the market contribute to the evolution in aggregate productivity.

An influential insight that has manifested itself from this work is the disparate roles that firm entry and exit versus firm expansion and contraction can have in driving the evolution of key variables like aggregative productivity. For instance, aggregate productivity in an industry—say telecommunications—could increase not only because the firms in the industry individually become more productive, but also because the firms entering the industry are more productive than the firms exiting the industry. Finally, even without any entry or exit or improvement in firm-level productivity, aggregate productivity might increase if production is allocated towards the more productive firms and away from the less productive firms. That is to say, the more productive firms expand and the less productive firms contract. As we show in the next application, these measurement issues are just as—if not more—relevant for interpreting big data sources.

Application: Measuring exit and entry of firms and workers in real time

As we saw earlier, realtime data from firms such as Homebase have been crucial for obtaining up-to-date information about the labor market during the Covid-19 pandemic.[22] However, in addition to the representativeness issues discussed above, consumers of this data must also consider how the data are measured over time. Comparing the Homebase data with that of a competing firm called Kronos that is among the firms used by the OI lab in their measure of employment provides an illustrative example. Homebase and Kronos take different approaches to constructing the panel of firms that they use to measure changes in employment and hours worked. Homebase uses a fixed panel of firms. This means that if a business starts using their service, they are *not* added to the panel, while if a business stops using the software (distinct from having zero recorded working hours), they are dropped from the panel. In contrast, Kronos measures the raw number of shift punches and payrolls without accounting for changes in the sample of firms that use their software.

These two methods of panel construction provide different insights into changes in employment and hours worked. By using a fixed panel, Homebase captures changes along what economists refer to as the *intensive margin*, or the amount of labor that existing firms demand. The Kronos data instead captures both changes along the intensive margin as well

[22] Bartik et. al (2020) use data from Homebase and the BLS to measure the labor market collapse and recovery through July. Electricity production (Lewis et. al., 2020) and consumption (Cicala, 2020) data are other examples of high frequency and timely data that has been used to track the evolution of economic activity during the Covid-19 pandemic.

as changes along the *extensive margin*, or the number of firms that are demanding labor. The extensive margin accounts for changes due to the lifecycle of firms, measuring increases due to new firms opening up (firm births) and decreases from existing firms closing down (firm deaths). Both the intensive and extensive margins are of interest in their own right. The changes along the intensive margin may be especially interesting in the context of the Covid-19 pandemic for policy makers and managers who wish to understand how well businesses are (or are not) weathering the pandemic. At the same time, the developments along the extensive margin are an important component of the aggregate changes in labor market conditions.

Figure 2 shows two data series that together illustrate the intensive and extensive margins of business operations during the pandemic. The top panel of figure 2 shows that after a steep decline in March and April, new business applications that are identified by the Census Bureau as having a high propensity to become a business with a payroll are now up by 25% year-over-year. This measure of new business formation captures the extensive margin of firm births. The bottom panel of the figure instead contains a measure of small business openings from Womply, a credit card payment aggregator, comparing the number of businesses open to a January reference period. Because Womply is comparing the same set of firms at two different points in time, they capture the extensive margin of business closures.[23] The Womply data, however, cannot tell us whether or not these closures are temporary are likely to be permanent.

Together, these two big data sources suggest that through the summer months of 2020 many firms were still shut down due to the pandemic and public health restrictions, but that since the end of June there has also been a wave of newly formed businesses to fill some of this gap.[24] Reporting from the *Wall Street Journal* suggests that these new business are overwhelmingly small operations started by individuals whose prior employment arrangements had been disrupted by the pandemic. This fits in nicely with additional big data evidence which points to the remaining small business closures as less likely to be temporary than permanent. For example, Yelp, the business review website, published a report in September finding that, of businesses that were open on March 1 and had marked their business as "closed" on their website, 60% reported closures that were permanent.[25]

Given what we see in figure 2, drawing conclusions about the overall labor demand of small employers during the pandemic solely on the basis of the Homebase data is not a good idea. Doing so, one would likely miss a considerable amount of interesting variation along the

---

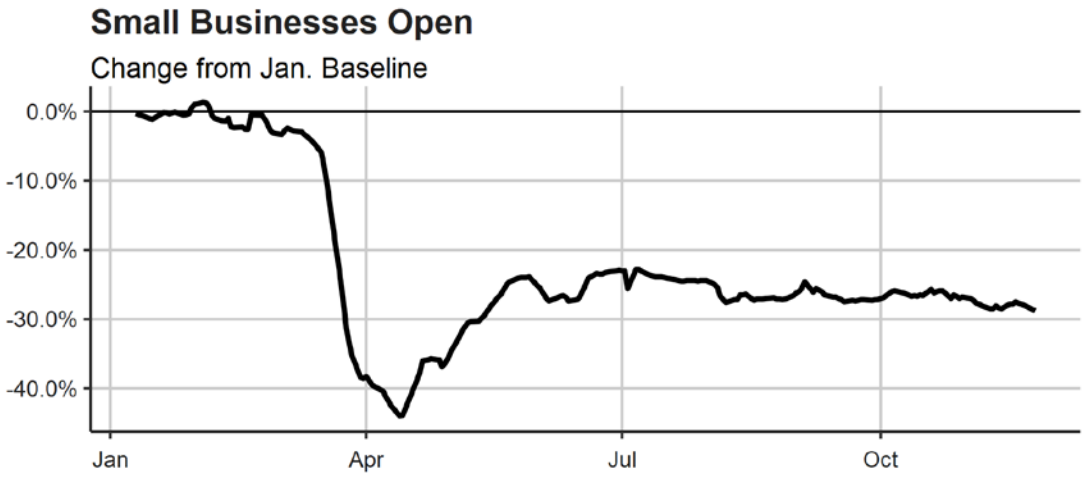
[23] Like the Homebase data, Womply also uses a fixed panel of firms. Womply defines these businesses as those that are "regularly open" from January through February.

[24] See https://www.wsj.com/articles/is-it-insane-to-start-a-business-during-coronavirus-millions-of-americans-dont-think-so-11601092841 and https://www.economist.com/united-states/2020/10/10/the-number-of-new-businesses-in-america-is-booming. Prior to the recent boom, there had been a well-documented decades-long secular trend of lower rates of new business formation, often referred to as reduced business dynamism. For more details, see Decker et. al (2016, 2020).

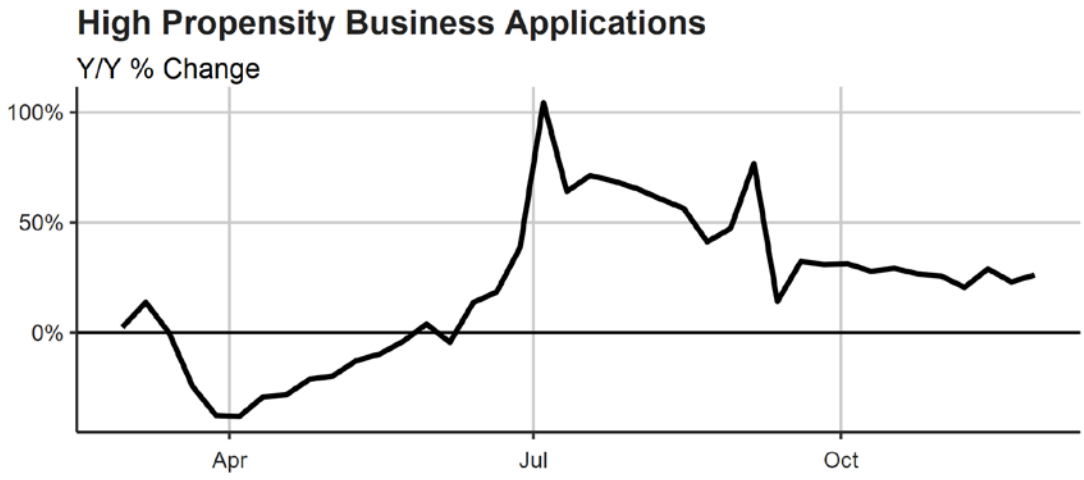
[25] https://www.yelpeconomicaverage.com/business-closures-update-sep-2020

extensive margin. The same can be said, however, for the other data sources that better capture the extensive margin. For instance, there is a well-known lag between new business applications and actual employment, so putting too much emphasis on that data alone is also probably not a good idea if the goal is to understand current labor market trends. Instead, to get the whole picture of small business labor market conditions it is better to consider all of these big data sources together in conjunction with traditional sources of information on the labor market.

Figure 2



Source: Womply via Opportunity Insights and Haver Analytics



Source: Census Bureau via Haver Analytics

One of those data traditional data sources that is commonly used for this purpose is initial unemployment insurance (UI) claims. Take-up rates for UI are viewed as a reliable indicator of turning points in the business cycle (Gordon, 2009). As firms let go of workers at the beginning of a recession, the demand for UI increases, typically peaking just prior to the trough of a recession. Unlike most economic statistics, new claims for UI are reported weekly and with only a week's lag, making them an exceptionally timely measure of labor market conditions. That said, there are still some instances where big data sources can improve upon the signal initial UI claims provides. Like the new business applications data, claims are just an application for insurance. Some will be accepted and some will be rejected as they make their way through the approval process. That administrative process can add lag times to the data, and often claims for the prior week are revised as a result.

While the processing delays and revisions are typically small, during the pandemic, when unemployment offices were suddenly flooded with applications and eligibility guidelines were rapidly changing as new laws were passed, they were sizeable. This has led many users of the data to question their value, and ultimately to a full scale audit by the Government Accountability Office (GAO).[26] As these others have noted, the official weekly UI claims statistics reported during the pandemic have been heavily influenced in particular by state-by-state measurement and reporting issues surrounding the federal Pandemic Unemployment Assistance (PUA) program, as well as substantial processing delays in states such as Florida and cases of fraud artificially boosting claims numbers in other states, including Washington, California, and Colorado. These issues have spurred interest in alternative big data measures capable of cutting through the noise in this data.

Based on the earlier work of Choi and Varian (2012), we have found in our own work that Google Trends search interest for the "unemployment" topic in some instances may actually be a *better* measure of the demand for UI than the official claims statistics for a number of reasons. To demonstrate, figure 3 plots Google Trends search interest for the unemployment topic and weekly initial UI claims (both regular state programs and PUA claims), both indexed to their value at the beginning of March. The figure makes clear that Google Trends was quick to pick up the increase in demand for UI even as processing delays slowed its appearances in the initial UI claims data.[27] Once most states had caught up, however, Google Trends continued to track very closely with new UI claims through the summer months. As new issues have arisen in the fall and states like California have temporarily stopped processing PUA claims in order to work out fraudulent submissions, a gap has emerged with Google Trends indicating greater continued demand for UI.

Going forward, tracking UI claims data will have to overcome a culmination of factors including the sunsetting of some unemployment programs (e.g., the extended benefits program), and any additional extensions/stimulus programs that might be implemented.

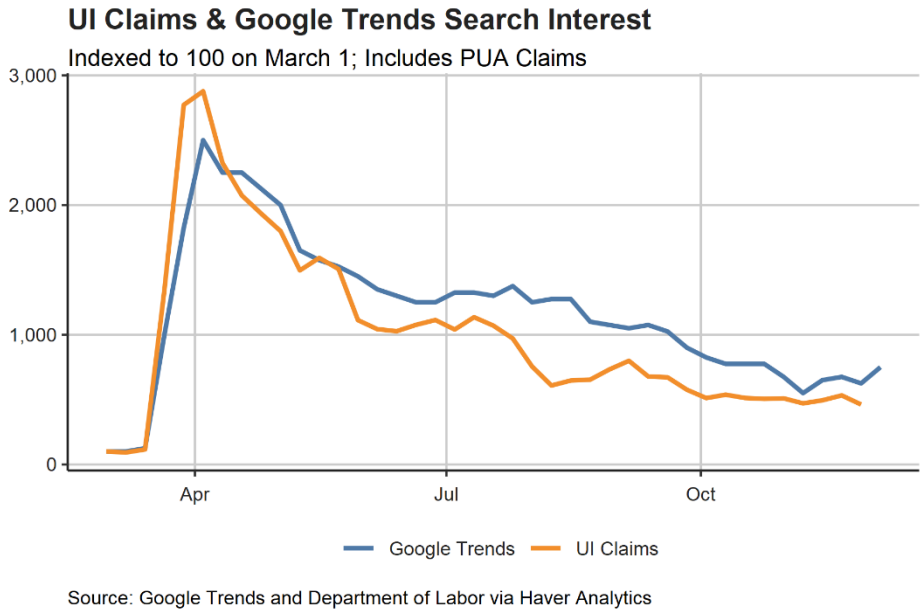
---

[26] For a discussion of the GAO's audit and report, see https://www.wsj.com/articles/labor-department-published-flawed-estimates-of-weekly-jobless-claims-watchdog-says-11606752477

[27] This timing can be seen even more clearly if you disaggregate across U.S. states and examine changes around the time that stay-at-home orders were implemented. See, for example, Aaronson et. al (2020b).

Each of these will have important implications on how past UI data can be used as a basis for comparison on what the incoming data say about the health of the labor market—as measures of the extensive margin will likely be of particular interest. This, of course, also highlights that even some of the most traditional of data sources are themselves not impervious to issues the SMALL framework attempts to highlight. As further evidence, while we did not highlight it here, even the BLS survey of payroll employment in the previous example must deal with the firm entry/exit issues noted above.[28]

Figure 3



3. How is it [A]ssembled?

Big data users must also take note of how the data are assembled—in particular, what, if any, transformations were used to construct them and whether or not any effort has been made to adjust them for normal seasonal fluctuations. Often, the data are made available in only one form, leaving little choice to the user about how to present them. But when it is possible to transform the data, it is important to note that many transformations used with traditional data sources may not always be as reliable to use on big data sources. To

[28] See https://www.bls.gov/web/empsit/cesbd.htm for an exposition of the net birth-death model that the BLS uses to adjust estimates from the CES survey for firm entry and exit.

facilitate further discussion, we break this section down into the type of transformation applied and its connection with any necessary seasonal adjustment.[29]

Common transformations used when presenting big data include levels, growth rates, or as a comparison to a baseline period (such as year-over-year percent changes or a comparison to a fixed baseline period, such as indexing a series to a value from a specific point in time). All of these choices have advantages and drawbacks, and the user must consider what they are trying to communicate with the data when determining which data transformation to apply. Transformations can often have a drastic impact on the conclusions drawn from data. Presenting a series in levels, for instance, provides an easily interpretable manner of expressing the activity that the series is designed to capture, but it may or may not also have direct meaning to the desired application to be studied.

Big data sources often represent only a small subsample of the entire universe we wish to track or make inferences about, such that the level of a big data series is probably not directly comparable to that of the broader object of interest in most cases. Growth rates, which can be directly compared to broader concepts of interest, are often more useful. However, growth rates, especially at the weekly or daily frequency, tend to be noisier (smaller number of observations in time leads to higher sample variance) and typically require some scaling (for example, annualization) in order to make direct comparisons with the overall concept of interest, or else some amount of temporal aggregation in order to match more standard frequencies of observation (e.g. monthly or quarterly).

Annualizing high-frequency growth rates, in particular, can create issues with communication. In other words, annualizing high-frequency data risks over-emphasizing large swings that occur in the data repeatedly over short time spans. Often, the better way of capturing slow-moving changes in a big data time series is to focus on year-over-year comparisons instead. In fact, it is very common to see such time series shown in this manner or other methods that highlight a comparison to a particular reference period.[30] This is a very attractive and simple way to answer the question "how different are these data from normal?" Using comparisons to a reference period, however, can create issues with holidays and other semi-regular but atypical events with seasonal elements to them.

Seasonality is a topic that will be familiar to many readers and close followers of official economic statistics. At the heart of the issue is the notion that there are many factors that are likely to contribute to the variation in a time series. A broad set of these typically revolve around seasonality, either through effects related to seasonal changes in weather, a schedule of events, or agriculture. For example, people buy more groceries during the final quarter of the year in conjunction with Thanksgiving and Christmas (Chevalier et. al., 2003,

[29] Other considerations related to the sorts of transformations that are often used include the use of fixed- vs. variable-weights in constructing indexes, annualizing and/or differencing the data, and any imputations, interpolations, or extrapolations used to handle missing data.

[30] Many private companies that have produced data products related to COVID-19 and economic activity have used a January 2020 baseline, typically taking the median value for each day of the week as the reference point. Some examples include the Homebase small business impact data and the Google mobility reports.

Butters et. al., 2020b). The seasonal fluctuations in demand for hotels and air travel varies widely within the year across different metro areas (Butters, 2020), and construction activity is significantly affected by seasonal weather patterns (Geremex and Gourio, 2018). Even growth in real Gross Domestic Product (GDP) experiences a seasonal "trough" in the first quarter each year that dwarfs the size of most recessions (Barsky and Miron, 1989).

The breadth and depth of these seasonal effects often requires that some adjustment be made in order to distill the variation driven by other considerations of interest instead, such as the business cycle for GDP.[31] Of course, to the extent that the transformation or method of assembly only partly addresses seasonality considerations it can lead to improper judgements by users of such data. This is why analysts often obsess over seasonal considerations in the daily flow of information available to track economic activity in real time. One need not look far whenever an anomalous reading of an economic statistic emerges for an explanation that cites an inappropriate seasonal adjustment. In point of fact, though, seasonality is something that statistical agencies address in painstaking detail, as anyone who has ever studied the Census Bureau's methods, for instance, can attest.[32]

Seasonal adjustment is a potentially even more dire problem for big data sources. Because they are typically measured at a very high frequency, established methods of correcting for seasonality that are regularly applied to traditional data sources are often not feasible, and methods that do exist for daily and weekly time series are many times considerably more difficult for the average user to implement.[33] Instead, what is often done is to appeal to particular transformations of the data which can be thought of as simple seasonal adjustments, or "filters." The simplest example is the use of year-over-year comparisons where days or weeks are matched across years. However, this is not always sufficient to do the trick, as we note in the application below. Many times this simple procedure still leaves behind traces of *residual seasonality*.[34] This can occur for a multitude of reasons; for example, through day-of-week effects or modest fluctuations in the timing of particular holidays within the year. Below, we highlight a few salient examples.

---

[31] It is also important to note that there are many instances in which abstracting from the heterogeneity in the seasonal fluctuations could in fact lead to improper inferences, say when investigating inventories or productivity when adjustment costs are present (e.g., see Krane and Braun (1991); Butters (2020)). An example with a commonly used big data source is the impact of within-month incentives and inventory management practices on auto dealer sales. We thank Paul Traub for pointing out this example to us.

[32] See U.S. Census Bureau (2017).

[33] For examples of such methods, see Cleveland and Scott (2007), McElory (2017), and McElory (2018) and the references within.

[34] For example, despite the Bureau of Economic Analysis' revisions to the seasonal adjustment procedure used for real Gross Domestic Product (the third and final phase was completed in July 2018), evidence still exists that residual seasonality remains (Consolvo and Lunsford, 2019).

Application: Transformations and seasonality in cell phone data

Despite their obvious advantages, applying year-over-year percent change and fixed reference period comparisons to big data time series is not a panacea. For instance, both can face issues with effects that are tied to particular days of the week and/or particular holidays which move around from year to year (e.g. Easter, Thanksgiving), although some of these issues can be mitigated to a degree by using a 7-day moving average of the data.

Figure 4 plots cell phone mobility data from a company called Unacast that captures the number of visits to nonessential businesses in the U.S. during the pandemic period. The figure nicely illustrates both points. The large negative spikes in the data correspond to the U.S. summer holidays – Memorial Day, the 4th of July, and Labor Day– and Thanksgiving. Without knowing their cause, it would be easy to misinterpret what is going on for those days, but it is also clear that even without that knowledge focusing on the 7-day moving average instead is a little more robust to drawing the wrong conclusion.

Figure 4

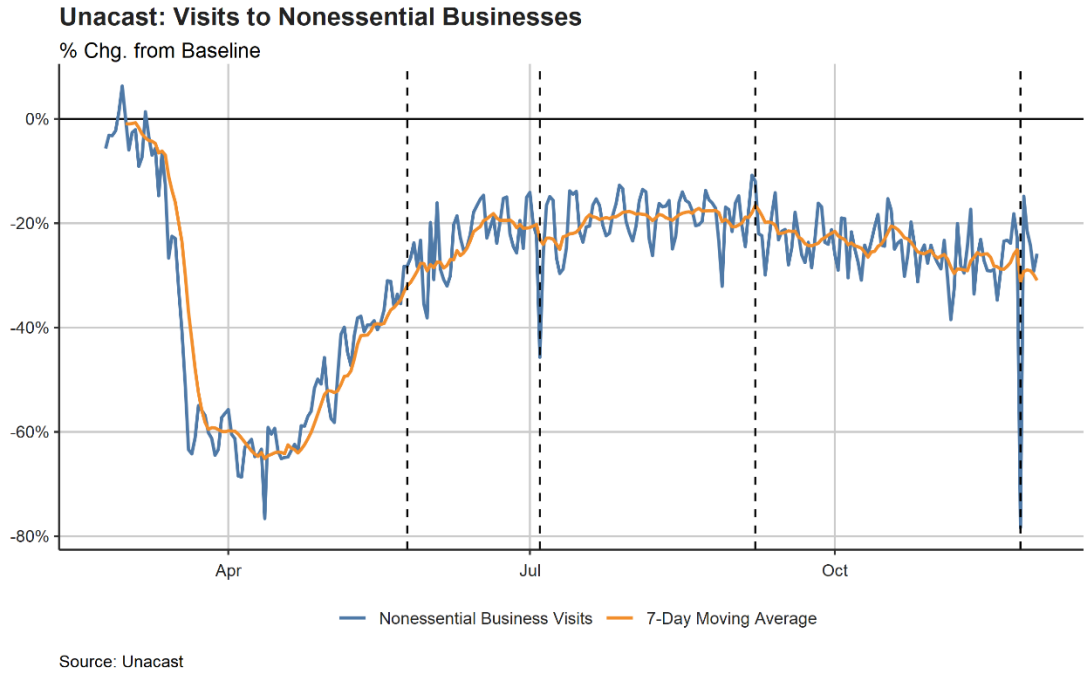
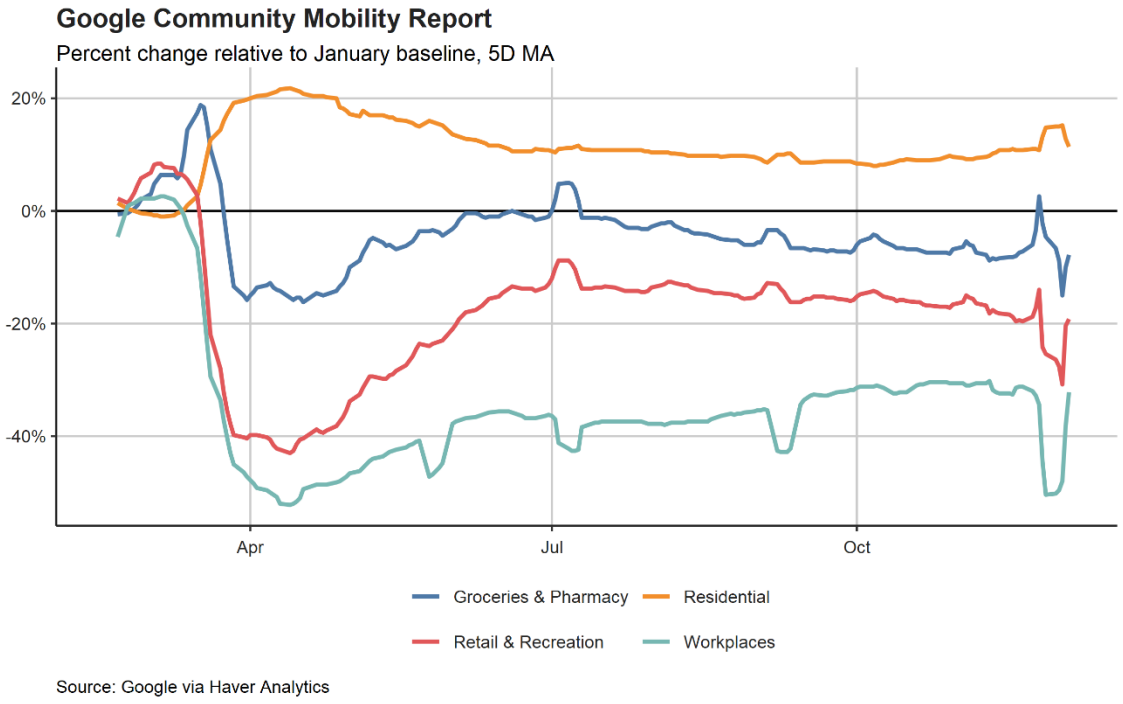


Figure 5 depicts another big data source from the Google Community Mobility Report that is connected with the Google Maps feature present on many cellphones. Here, visits are constructed as a percent change from a January baseline for several categories of places. In addition to the holiday effects noted for the Unacast data, there are very likely seasonal patterns in cell phone mobility data of this nature as well (e.g. fewer visits to workplaces during the summer, less recreation trips to beaches and parks in the winter) that cannot be disentangled from the underlying changes in mobility when constructed in this manner. Specifically, the time series in figure 5 is likely to understate the decrease in the Retail and

Recreation category while simultaneously overestimating the increase in the Residential category because the use of a fixed base period does not allow for the fact that the "normal" amount of mobility almost certainly varies over the course of the year within different categories.[35]

During the pandemic where social distancing restrictions have been prevalent throughout, this may not matter too much in terms of capturing the overall decline in mobility with data of this nature, but it will affect the precision of the estimates made over time as the seasons change from spring/summer to fall/winter. Furthermore, this is not something that a 7-day moving average can correct. To establish a seasonal "normal" requires at least one additional year of data in this case to use for comparison.

Figure 5



Source: Google via Haver Analytics

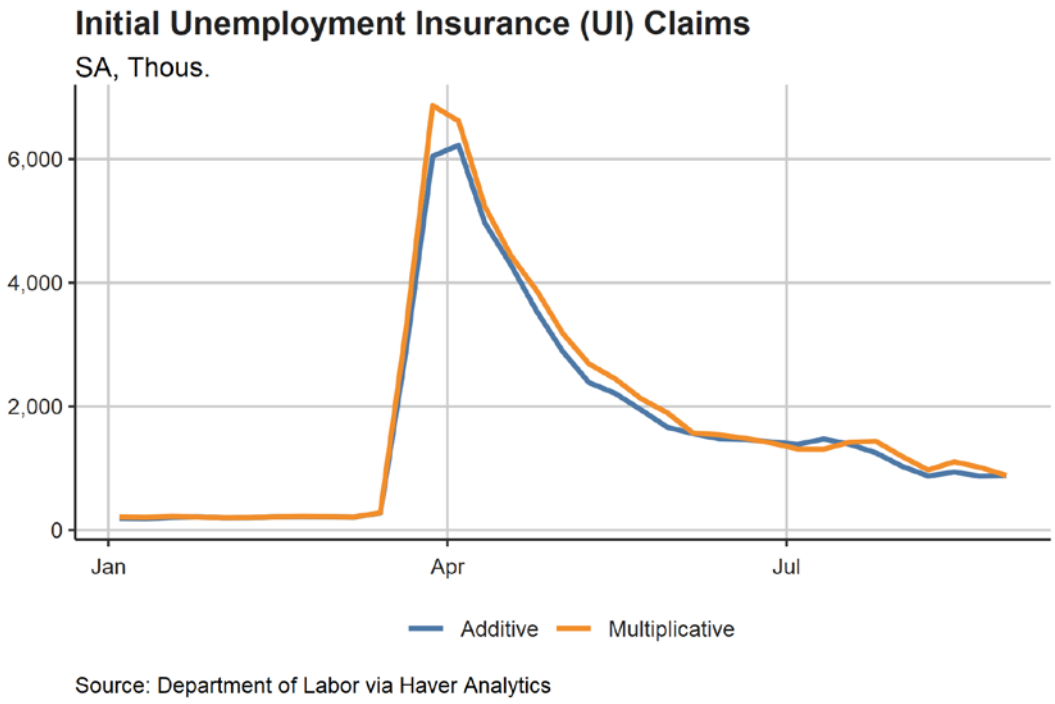
Even when longer time series are available, proper seasonal adjustment is still a major hurdle facing big data users. A good illustration of this is the Department of Labor's (DoL) seasonal adjustment of initial unemployment insurance (UI) claims during the pandemic. During the first weeks of the pandemic in the U.S., initial UI claims spiked to unprecedented levels, peaking at over 6 million claims filed in one week. Historically, the DoL has used multiplicative seasonal factors to seasonally adjust the UI claims data. However, economists and commentators expressed concerns that the multiplicative seasonal adjustment procedure was inappropriate given the dramatic change in the level of the

[35] This is an issue that also plagues the use of electricity demand data, since one must be careful to account for seasonal fluctuations that result from the transition to and from mostly electric-driven air conditioning to primarily natural gas-driven heating in the spring and fall months.

series during the pandemic. Using multiplicative seasonal factors to adjust initial UI claims when the baseline level shifted an order of magnitude larger than any prior baseline level in the time series led to seasonal adjustments that dominated pre-existing long-run trends.

Starting with the release on September 3,, 2020 for the week ending August 29th, the DoL switched from multiplicative to additive seasonal factors. Figure 6 plots the multiplicatively-adjusted series against an additively-adjusted series produced by Haver Analytics. The cumulative difference between the two time series suggests that the multiplicative adjustment overstated the number of initial UI claims on a seasonally adjusted basis by over four million claims. This regime change in the DoL seasonal adjustment process illustrates just one of the issues facing those who wish to seasonally adjust series from non-traditional big data sources. In addition to choosing between multiplicative and additive seasonal factors, big data users must also consider how the length of the time series factors into the method used, since most big data time series are too short for standard seasonal adjustment methods to be reliably applied, as well as the optimal frequency (daily or weekly) to apply the seasonal adjustment procedure.

Figure 6



Source: Department of Labor via Haver Analytics

4. Do the data exhibit reporting [L]ags?

One of the most attractive features of many of the new big data sources that have been used to track the economy during the pandemic is that these data are available much more rapidly than official statistics published by government statistical agencies. While the lag time on indicators such as the employment report or GDP is on the order of weeks to months, many private sector big data sources are available in near real time. That said, the key word in that sentence is *near*. Often, there are still structural hurdles that prevent the data from being truly useful in real time. Indeed, some of the big data sources we referenced earlier fall subject to this critique.[36] Understanding these publication delays should always be a first order concern.

The primary component of these publication delays is typically the lag time between when the activity of interest occurs and when the data are first available – this could be anywhere between one to two days to one to two weeks depending on the activity and the data source. However, the second and potentially more important delay tends to occur between the first "release" of the data for a given time period and when it can be considered "final." For some data sources, especially those based on cell phone location such as the Unacast and Google mobility reports, the first "release" of the data can be considered definitive.[37] However, for an important subset of big data sources, most notably debit and credit card spending data, there is an appreciable delay before the data can be considered final.

For big data sources that demonstrate such a delay, it can be necessary to allow several extra days for observations to filter into the dataset; otherwise analysts are at risk of detecting a spurious decline in the last few days of the time series. The obvious need to avoid this potential mistake demonstrates just how critical it is for big data consumers to understand if there is any type of structural processing delay in a given data source, how long it is, and how this type of reporting delay could impact the conclusions they draw from the data, especially near the end of the sample.

A good example that we can draw on to make this point is the use of citation counts in the study of research and development and innovation (e.g., see Azoulay et. al., 2015; Bryan and Ozcan, 2020). Broadly speaking, these studies use citation counts as a proxy for the dissemination of information (e.g., research innovations) and/or the value of such activity.

---

[36] Chetty et. al (2020) note the end-of-sample delays in the Womply data, for instance, and up to four-week lags in the payroll processing data that they use to track employment. Using Google Trends to track demand for unemployment insurance is another example. Much of this search activity occurs on weekends. In order to be certain to capture the full extent of search for the previous week, one often must wait until at least a few days into the current week in order to obtain accurate results.

[37] Even with the Google mobility data, one must still be careful about timing. For instance, if the object of interest is to characterize how the daily number of visits changes from week-to-week the patterns that exist for weekdays are often very different than those for weekends. In fact, Google presents its mobility reports separately for weekdays and weekends. We show the weekday data in figure 5. This is true as well for Google search activity, with it often displaying a noticeable spike on weekends compared to weekdays.

Abstracting from other inherent measurement issues, one substantive hurdle all investigations of this sort have to overcome is the censored nature of citation counts for recently produced research. In other words, the relevant (even relative) citation count of an innovative patent filed in 2001, for instance, will be much different than a comparable patent filed in 2020 for reasons unrelated to either patents' usefulness. While several creative research designs are still available to overcome this shortcoming, this setting underscores the other channels through which *lags* in the data might present themselves.

Application: Processing delays in credit and debit card transactions

For debit and credit card transaction data, reporting lags are a consequence of the processing delays in payment networks (i.e. the time it takes from a transaction clearing at the point of sale to its settlement between the financial institutions in the payment network). The settlement lag is typically 1-3 business days, making it important to take this lag into account when analyzing the data.[38] That said lags can differ greatly across different parts of the payment network and much work is already underway to reduce them, such that they are likely shrinking even as we write this.[39] The bigger picture that we want to emphasize with this example is that knowledge of the institutional arrangements that govern big data collection is absolutely essential to understanding and properly using it.

Figure 7 plots the month-over-month percent change in retail and food service sales from the Census Retail Trade Survey compared to a measure developed at the Federal Reserve Board and published by the U.S. Bureau of Economic Analysis (BEA) that uses credit and debit card spending data from Fiserv First Data, a large payment intermediary.[40] Recalling the earlier applications and lessons on representativeness, a fundamental limitation of the card data compared to the Census survey is that it, by construction, misses an important type of spending – cash transactions. That said, the card spending data have tracked the Census measure relatively closely since the beginning of the Covid-19 pandemic in the U.S.

The Fiserv data and their use by the BEA as a supplementary high-frequency data source for consumer spending is a good example of when structural reporting delays are manageable when properly understood. Census retail sales data is used by the BEA as source information for various elements of U.S. GDP. Being a quarterly statistic (or even monthly in the case of some of its major components like personal consumption expenditures), a delay of a few days is not of major concern when it comes to GDP calculations. Even with these delays, this information is still available and generally reliable well before the monthly releases of Census retail sales or personal consumption

---

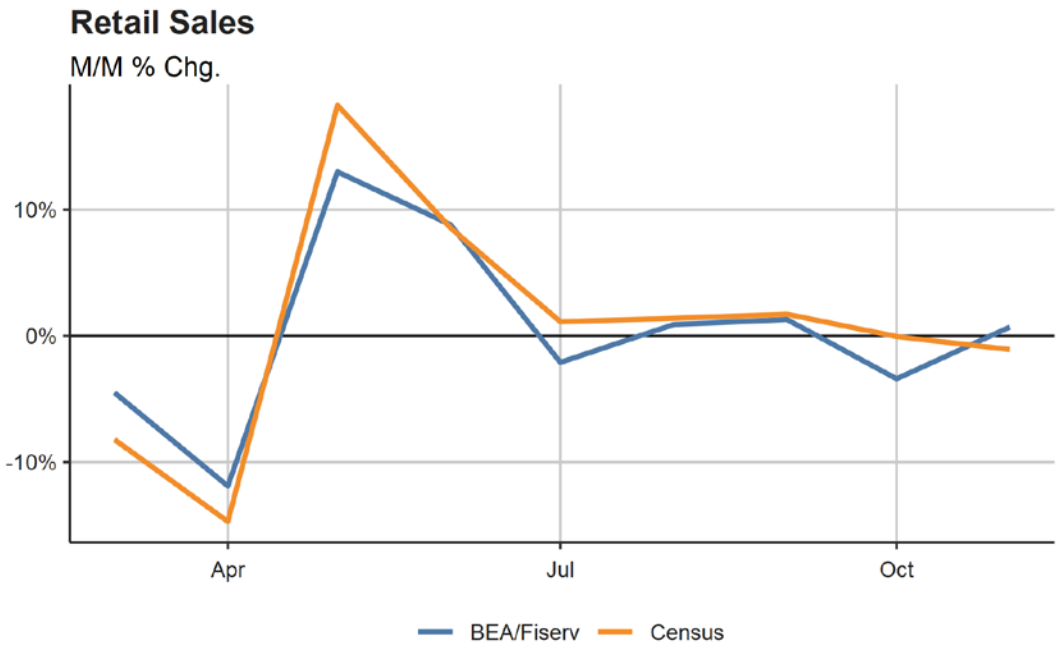
[38] See Herbst-Murphy (2013) for a detailed description of the clearance and settlement process for credit card payments.

[39] Real-time payments networks are an ongoing area of development in the payments industry with the Federal Reserve working to implement its own system, referred to as FEDNow, and several private sector competitors also set to roll out systems in the near future. See https://www.frbservices.org/financial-services/fednow/what-is-fednow.html for a description of the FEDNow realtime payments service.

[40] See Aladangady et. al (2019) for details on the construction of the Fiserv spending series

expenditures and quarterly GDP. Like many big data sources, determining its value comes down instead to how useful it is as a leading indicator, as we detail in the next section.

Figure 7



Source: Bureau of Economic Analysis and Census Bureau via Haver Analytics

5. Is it a [L]eading indicator?

If the concerns above can be addressed, the next question becomes are these big data sources still useful? In general, we believe the answer here is most often *YES*, but with an important caveat. For many big data users, the utility of a data source is likely to lie in part in how *predictive* it is of official statistics or other key variables of interest. One version of this is the process known as *nowcasting* which has increasingly been adapted to include big data sources.[41] Here, however, the lead time between the big data source and the key variable of interest is critical. A good forecast, or nowcast, that is available for weeks or a month ahead of time is much more valuable than one of the same accuracy that only comes out a day or two before. Thus, the *lead time*, or gap between the availability of the big data source and the official statistics, is an important determinant of its value. Additionally, the degree of leading information within a big data source—as measured by how predictive

[41] Some examples include Aaronson, et. al. (2016) which uses construction payments data to improve upon near-term forecasts of Census construction spending and Borup, et. al. (2020) and Coble and Pablo (2017) who use Google Trends to forecast near-term growth in employment and building permits, respectively.

and/or the ease with which such predictions can be generated—is also an important factor in its ultimate value.

Similarly, given their relative novelty, big data sources generally do not have long histories. This complicates truly testing their ability to generate reliable *out-of-sample* predictions in real time, as the power of any such test will be naturally limited. The lack of long histories is a hurdle that is not easy to overcome for even the most reliable big data sources. With not enough data points to likely truly establish the statistical significance of any predictive relationship, it falls to the economic significance of the argument to carry the day. Here, we come full circle, because economic significance is closely tied to the way the data is collected. In order to establish it, the researcher must have a firm understanding of what the data is measuring and how it was generated. In other words, the data is likely to be discounted if by design it is a "black box." Transparency is, therefore, almost itself a prerequisite for the use of big data in nowcasting.

Application: Leading signals from payroll and card processers

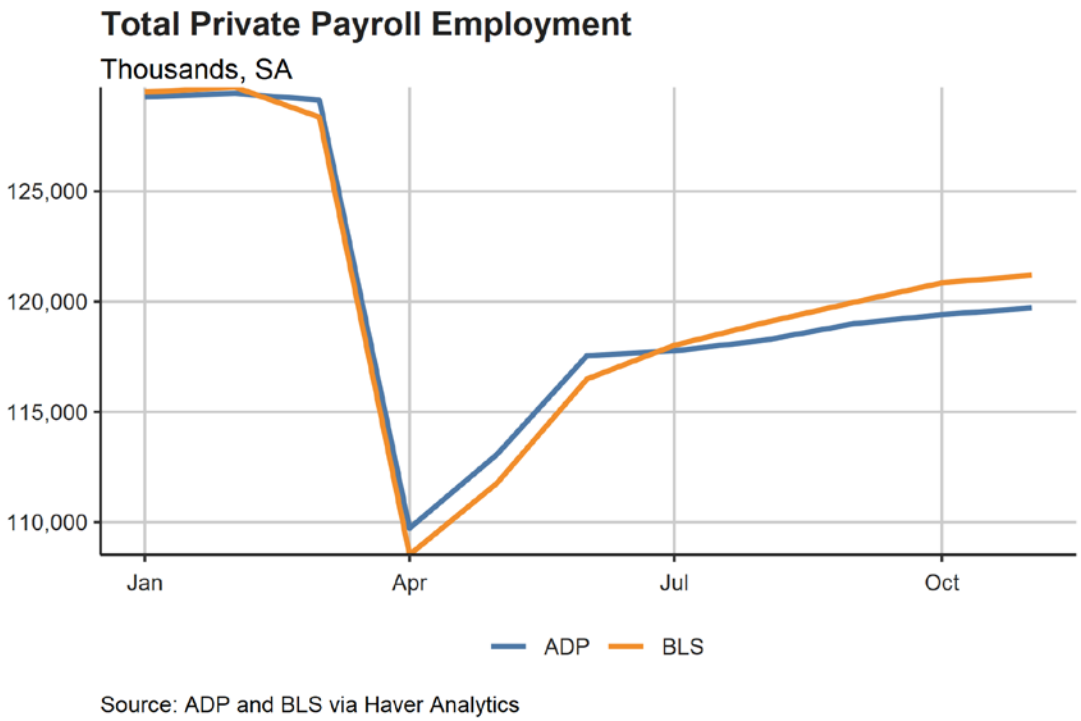
Several big data sources would seem to already fit the bill when it comes to being accurate predictors of official statistics; and, undoubtedly, more will be developed and tested as the need arises, but the process is likely to be slow and time intensive. One of the earliest and best documented examples is the data collected by the payroll processing firm ADP and published each month ahead of the BLS Employment Situation report. As figure 8 shows, the ADP data has tracked quite closely with the BLS estimates of total private payroll employment throughout the crisis and nascent recovery – of that there is little doubt. But to return to the point raised above, how much value is there in knowing the ADP number usually just a mere few days before the BLS number?

It depends on the audience, perhaps, but even with a high degree of accuracy the ADP information is likely to not be as valuable in real time for BLS payroll employment as the card spending data is for BEA GDP given that it leads the release of GDP by a much longer time period. That said, there could be other factors at play that still make the ADP data highly informative for other reasons. The BLS data, for instance, are from a survey of establishments and are subject to revision due to lagged reporting and other miscellaneous features (including the birth/death measurement of firms issue that we alluded to earlier). If the ADP data are predictive of these revisions, then the lead time is actually longer because the BLS revises its survey results only twice: once when the following month's results are released and another time in March of each year when the previous year's results are updated to reflect incoming administrative source data on employment. This is indeed the case with the ADP data.[42] More recent work has even tried to extend the lead time of the data for BLS payroll employment by looking at a weekly frequency instead.[43]

[42] Cajner et. al (2018).

[43] See Cajner et. al (2020).

Figure 8

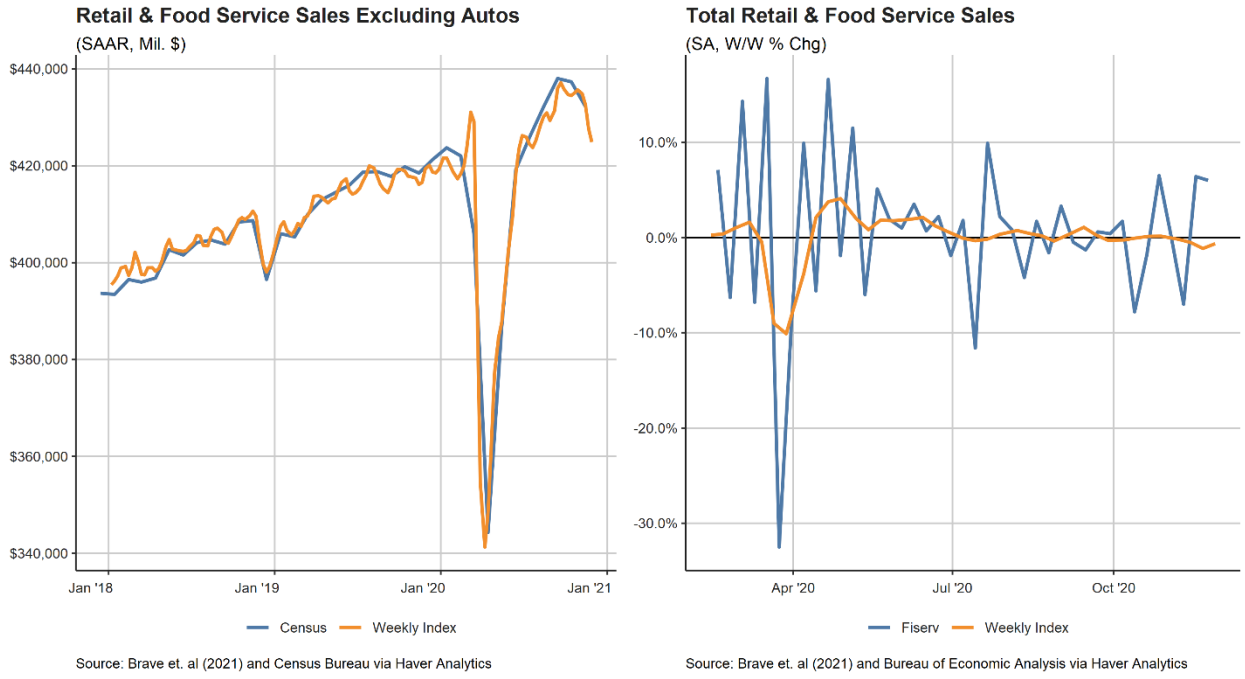


Source: ADP and BLS via Haver Analytics

The Covid-19 pandemic has also brought forth a great deal of attention on the credit and debit card transaction data. In the previous section, we noted the use of the Fiserv First Data series by the BEA as a high-frequency measure of consumer spending. However, Fiserv First Data is but one of many card processers that collects data of this nature. Like Fiserv First Data, many of these other big data sources also provide more detailed breakdowns of spending which allow the user to refine the universe of interest in order to better match official statistics. For example, auto sales are included in the Census Monthly Retail Trade survey, but are not likely to be represented in the card data. Therefore, many analysts focus on a subset of the data instead that excludes auto-related sales. Similarly, during the pandemic, food service sales have been disproportionately affected by government efforts to enforce social distancing. Thus, many analysts will include food service sales with retail sales in these comparisons as well.

Brave et. al (2021) combine the information from a number of these other credit and debit card data processers with retail foot traffic data and a survey of consumer sentiment to produce a weekly index of retail and food service spending excluding auto-related sales during the pandemic. Figure 9 plots their index as both a weekly level and growth rate in comparison to the Fiserv First Data series for total retail and food service sales. Both the weekly level and growth rate measures are indexed to the monthly data from the Census Monthly Retail Trade survey on retail and food service sales excluding autos (also shown in the figure). By combining the related information in all of these big data sources, the authors build the kind of bridge we referred to earlier which allows for a way of filtering through some of the high-frequency "noise" in the data to reveal the smoother "signal" that is often more relevant for official statistics.

Figure 9



Conclusion

The growing availability and acceptance of big data has produced an excellent opportunity for researchers, business executives, and policy makers to inform business strategies and policy decisions on highly granular and timely information. These opportunities, however, come with the need to be mindful of the obstacles those same data sources present. Fortunately, many of these obstacles have been faced before with more traditional data sources. Through a series of case studies, the distinct aspects surrounding sampling design, measurement, and data assembly were shown to have important implications on the types of conclusions and the ultimate value of big data. By collecting these "lessons learned" of past statisticians, econometricians, and other data practitioners, our SMALL framework should help aid consumers of big data in avoiding the potential pitfalls that can arise.

We encourage future work to continue to expand on outlining and addressing the issues that the more granular and often higher frequency aspects of many big data sources will have on their use in diverse fields. The continued dissemination of these "cautionary tales" of the use of big data should facilitate more informed business and policy decisions, as well as in turn shape the development of future big data sources themselves. While the focus of this article, and of the SMALL framework more specifically, was on the big data itself, we would be remiss not to acknowledge the coincident development of new statistical methods that facilitate the use of these data as well. It is perhaps here even more so that big data has the potential for serious misuse. We leave for future work a guide for how to navigate the forest of empirical methods now available to analyze big datasets.

References

Aaronson, Daniel et. al. (2016). "Using private sector `big data' as an economic indicator: The case of construction spending," *Chicago Fed Letter*, No. 366.

Aaronson, Daniel et. al. (2020a). "Using the eye of the storm to predict the wave of Covid-19 UI claims," *Covid Economics: Vetted and Real-Time Papers*, No. 9, April 24, 59–76.

Aaronson, Daniel et. al. (2020b). "The stay-at-home labor market: Google searches, unemployment insurance, and public health orders," *Chicago Fed Letter*, No. 436.

Abbring, J. and J. Campbell (2005). "A firm's first year," *Federal Reserve Bank of Chicago Working Paper*, No. 2003-11.

Abraham, Katherine G. and Ron S. Jarmin and Brian Moyer, and Matthew D. Shapiro (2020). Big Data for 21st Century Economic Statistics. *National Bureau of Economic Research: University of Chicago Press.*

Aladangady, Aditya, et al. (2019). "From Transactions Data to Economic Statistics: Constructing Real-time, High-frequency, Geographic Measures of Consumer Spending." *National Bureau of Economic Research Working Paper* 26253.

Alexander, Diane, and Ezra Karger (2020). "Do stay-at-home orders cause people to stay at home? Effects of stay-at-home orders on consumer behavior," *Federal Reserve Bank of Chicago Working Paper,* 2020-12, revised April 18, 2020.

Alharthi, Abdulkhaliq et. al. (2017). "Addressing barriers to big data," *Business Horizons*, 60 (3), 285-92.

Arellano, M. and S. Bond (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58(2), 277-97.

Atkinson, Tyler et. al (2020). "Mobility and Engagement Following the SARS-Cov-2 Outbreak," *Federal Reserve Bank of Dallas Research Department Workign Paper,* No. 2014, May 2020.

Azoulay, Pierre, et. al., (2019). "Public R&D investment and private sector patenting: Evidence from NIH funding rules," *Review of Economic Studies*, 86(1), 1-36.

Barsky, Robert. and Jeffrey Miron (1989). "The seasonal cycle and the business cycle," *Journal of Political Economy,* 97(3), 503–534.

Barrero, Jose Maria, Nick Bloom, and Steven J. Davis (2020). "60 million fewer commuting hours per day: How Americans use time saved by working from home." *Becker Friedman Institute for Economics Working Paper No. 2020-132.*

Bartik, Alexander W., et al. (2020). "Measuring the labor market at the onset of the COVID-19 crisis." *National Bureau of Economic Research Working Paper* 27613.

Beracha, Eli and M. Babajide Wintoki (2013). "Forecasting residential real estate price changes from online search activity," *The Journal of Real Estate Research*, 35(3), 283-312.

Borup, Daniel et. al. (2020). "In Search of a Job: Forecasting Employment Growth Using Google Trends," *Journal of Business & Economic Statistics*, Forthcoming.

Brave, S., et. al. (2020a). "A closer look at the correlation between Google Trends and Initial Unemployment Insurance Claims," *Chicago Fed Insights*, June 17, 2020.

Brave, S. et. al. (2020b). "Another look at the correlation between Google Trends and Initial Unemployment Insurance Claims," *Chicago Fed Insights*, July 7, 2020.

Brave, S et. al. (2021). "Tracking U.S. consumers in real time with a new weekly index of retail trade," *Federal Reserve Bank of Chicago Working Paper*, 2021-05.

Bryan, Kevin and Yascin Ozcan (2020). "The impact of open access mandates on invention," *Review of Economics and Statistics*, Forthcoming.

Butters, R. Andrew (2020). "Demand volatility, adjustment costs, and productivity: An examination of capacity utilization in hotels and airlines," *American Economic Journal: Microeconomics*, 12(4), 1-44.

Butters, R. Andrew et. al. (2020a). "How Do National Firms Respond to Local Shocks? Evidence from Excise Taxes" *Kelley School of Business Working Paper.*

Butters, R. Andrew et. al. (2020b). "Why Do Prices Fall During Seasonal Demand Peaks?" *Kelley School of Business Working Paper, 19-21.*

Cajner, Tomaz, et al. (2020). "Tracking labor market developments during the covid-19 pandemic: A preliminary assessment." *Federal Reserve Board Finance and Economics Discussion Series 2020-030.*

Cajner, Tomaz, et. al (2018), "Using Payroll Processor Microdata to Measure Aggregate Labor Market Activity". *Federal Reserve Board Finance and Economics Discussion Series 2019-005.*

Coston, Amanda et. al (2020), "Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for Covid-19 Policy," available at https://arxiv.org/pdf/2011.07194.pdf

Cavallo, Alberto (2017). "Are Online and Offline Prices Similar? Evident From Large Multi-Channel Retailers," *American Economic Review,* 107 (1), 283-303.

Cavallo, Alberto and Roberto Rigobon (2016). "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 30 (2), 151-178.

Chetty, et. al. (2020). “How did COVID-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data,” *NBER Working Paper*, June 2020 (27431). https://www.nber.org/papers/w27431

Chevalier, Judith, A. et. al. (2003). "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data ." American Economic Review, 93 (1), 15-37.

Choi, Hyunyoung, and Hal Varian (2012). “Predicting the present with Google Trends,” *Economic Record*, Vol. 88 (s1), 2-9.

Cicala, Steve, (2020). “Powering work from home,” *Working Paper*, http://www.stevecicala.com/papers/powering_wfh/powering_wfh.pdf.

Cleveland, W. and Stuart Scott (2007). “Seasonal adjustment of weekly time series with application to unemployment insurance claims and steel production,” *Journal of Official Statistics*, 23(2), 209-221.

Coble, David and Pablo Pincheira (2017). “Nowcasting building permits with Google Trends,” *MPRA Working Paper*, 76514.

Coibion, Olivier, et. al. (2015). "The cyclicality of sales, regular and effective prices: Business cycle and policy implications," *American Economic Review*, 105 (3), 993-1029.

Collard-Wexler, Allan, and Jan De Loecker (2015). "Reallocation and Technology: Evidence from the US Steel Industry." *American Economic Review*, 105 (1), 131-71.

Consolvo, Victoria and Kurt Lunsford (2019). “Residual seasonality in GDP growth remains after latest BEA improvements,” *Economic Commentary*, No. 2019-05, https://www.clevelandfed.org/~/media/content/newsroom%20and%20events/publications/economic%20commentary/2019/ec%20201905/ec%20201905%20pdf.pdf

Decker, R., et. al. (2016). “Where has all the skewness gone? The decline in high-growth (young) firms in the U.S.” *European Economic Review*, 86, 4-23.

Decker, R. et. al. (2020). “Changing business dynamism and productivity: Shocks vs. responsiveness,” *American Economic Review*, Forthcoming.

DellaVigna, Stefano and Matthew Gentzkow (2019). "Uniform Pricing in U.S. Retail Chains," *The Quarterly Journal of Economics*, 134(4), 2011-84.

Earley, Christine E. (2015). “Data analytics in auditing: Opportunities and challenges,” *Business Horizons*, 58 (5), 493-500.

The Economist (2020). “Why real-time economic data need to be treated with caution,” *Free Exchange*, July 25, 2020. https://www.economist.com/finance-and-economics/2020/07/23/why-real-time-economic-data-need-to-be-treated-with-caution

Friedman, David M. and Crystal G. Konny and Brendan K. Williams (2019). “Big Data in the U.S. Consumer Price Index: Experiences and Plans,” in Big Data for 21st Century Economic Statistics. *National Bureau of Economic Research: University of Chicago Press.*

Geremex, Menelik and François Gourio (2018). “Seasonal and business cycles of U.S. employment,” *Economic Perspectives*, 42(3).

Goel, S. et. al. (2010). “Predicting consumer behavior with web search,” *Proceedings of the National Academy of Sciences*, 107(41), 17486–90.

Gopinath, Gita, et. al. (2011). "International Prices, Costs, and Markup Differences." *American Economic Review*, 101 (6), 2450-86.

Gordon, Robert (2009). “Green shoot or dead twig: Can unemployment claims predict the end of the American recession?” *Vox*. Date accessed: 3/23/2020. https://voxeu.org/article/us-recovery-may-2009-new-evidence-based-surprisingly-robust-linkage

Griliches, Z. and J. Hausman (1986). “Errors in variables in panel data,” *Journal of Econometrics*, 31(1), 93-118.

Griliches, Z. and J. Mairesse (1999). “Production Functions: The search for identification,” *Econometrics and economic theory in the twentieth century: The Ragnar Frisch Centennial Symposium,* Cambridge University Press.

Gupta, Sumedha, et. al., (2020). “Tracking public and private responses to the COVID-19 epidemic: Evidence from state and local government actions,” *NBER working paper*, 27027.

Herbst-Murphy, Susan, (2013). “Clearing and Settlement of Interbank Card Transactions: A MasterCard Tutorial for Federal Reserve Payments Analysts,” *Federal Reserve Bank of Philadelphia Payment Cards Center Discussion Paper.*

Hitsch, G. J., et. al. (2019). “Prices and promotions in U.S. retail markets: Evidence from big data,” *Chicago Booth Research Paper*, 17(18).

Krane, S. and S. Braun (1991). “Production smoothing evidence from physical-product data,” *Journal of Political Economy*, 99(3), 558-81.

Lazer, et. al. (2014). “The parable of Google Flu: Traps in big data analysis,” *Science*, 14 March 2014, 1203-05.

In Lee (2017). “Big data: Dimensions, evolution, impacts, and challenges,” *Business Horizons*, 60(3), 293-303.

Lewis, Daniel et. al. (2020). “US economic activity during the early weeks of the SARS-Cov-2 outbreak,” *Covid Economics: Vetted and Real-Time Papers*, No. 6, April 17, 1–21.

McAfee, A., & Brynjolfsson, E. (2012). “Big Data: The management revolution,” *Harvard Business Review*, 90(10), 60–68.

McElroy, T. (2017). “Multivariate seasonal adjustment, economic identities, and seasonal taxonomy,” *Journal of Business & Economic Statistics*, 35(4), 611-25.

McElroy, T., et. al. (2018). “Modeling of holiday effects and seasonality in daily time series,” *Census Bureau Research Report Series*, 2018-01.

Nagle, Tadhg et. al. (2020). “Assessing data quality: A managerial call to action,” *Business Horizons*, 63(3), 325-37.

Njuguna, C., and P. McSharry (2017). "Constructing spatiotemporal poverty indices from big data." *Journal of Business Research*, 70, 318-27.

Olley, S. and A. Pakes (1996). “The dynamics of productivity in the telecommunications equipment industry,” *Econometrica*, 64(6), 1263-97.

Pérez-Martín, A., et. al. (2018). "Big data techniques to measure credit banking risk in home equity loans." *Journal of Business Research*, 89, 448-54.

Pham, Xuan and Martin Stack (2018). “How data analytics is transforming agriculture,” *Business Horizons*, 61(1), 125-33.

Sivarajah U., M.M Kamal, Z. Irani, and V. Weerakkody (2017). “Critical analysis of Big Data challenges and analytical methods.” *Journal of Business Research,* 70, 263-286.

Stephens-Davidowitz, Seth (2014). “The cost of racial animus on a black candidate: Evidence using Google search data,” *Journal of Public Economics*, 118, 26-40.

Towers S, et al. (2015). “Mass media and the contagion of fear: The case of Ebola in America,” *PLoS ONE*, 10(6).

U.S. Census Bureau (2017). “X-13ARIMA-Seats Reference Manual.” Washington DC: U.S. Census Bureau.

Walker, R. (2015). “From Big Data to Big Profits: Success with Data and Analytics,” Oxford University Press.

Wang, Y., and N. Hajli (2017). "Exploring the path to big data analytics success in healthcare." *Journal of Business Research*, 70, 287-99.Varian, Hal R. (2014). "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, 28 (2), 3-28.