Expanding the Frontier of Economic Statistics Using Big Data: A Case Study of Regional Employment

Abe Dunn, Eric English, Kyle Hood, Lowell Mason, and Brian Quistorff*†

July 10, 2024

Abstract

Big data offers potentially enormous benefits for improving economic measurement, but it also presents challenges (e.g., lack of representativeness and instability), implying that their value is not always clear. We propose a framework for quantifying the usefulness of these data sources for specific applications, relative to existing official sources. We specifically weigh the potential benefits of additional granularity and timeliness, while examining the accuracy associated with any new or improved estimates, relative to comparable accuracy

*Brian Quistorff, Abe Dunn and Kyle Hood: Bureau of Economic Analysis. Eric English: U.S. Census Bureau. Lowell Mason: Bureau of Labor Statistics. Brian Quistorff is the corresponding author (email: brian.quistorff@bea.gov). DISCLAIMER: Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Bureau of Economic Analysis, the Bureau of Labor Statistics, or the U.S. Census Bureau. This paper does not use any confidential Census Bureau data.

†We would like to thank participants at the Opportunity Insights seminar series, the Society of Economic Measurement (2023), and Federal Committee on Statistical Methodology (2023) and seminar participants at the Bureau of Economic Analysis. We would like to thank Ben Bridgman, Ryan Decker, Nathaniel Hendren, Justine Mallatt, Steven Mance, Mahsa Gholizadeh, Ludwig Straub, Henry Hyatt, and Santiago Pinto for comments. Jim Spletzer and numerous staff members from the Census and the Bureau of Labor Statistics provided invaluable early feedback on the analysis of the payroll data and official series. We would also like to thank John Friedman, Michael Stepner, and Raj Chetty for accelerating the analysis of these data and fostering the collaboration among the statistical agencies. In particular, the time and attention of John Friedman and Michael Stepner in facilitating research, data, and discussion on this topic has been invaluable. We would like to thank Sonya Porter of the U.S. Census Bureau for excellent leadership and coordination between the Opportunity Insights group and the statistical agencies. In addition, we would like to thank Julie Hatch and Joseph Piacentini for their involvement and leadership. We also thank the provider of the payroll data used to build the various series in this paper, which was crucial for this research effort.

produced in existing official statistics. We apply the methodology to employment estimates using data from a payroll processor, considering both the improvement of existing state-level estimates, but also the production of new, more timely, county-level estimates. We find that incorporating payroll data can improve existing state-level estimates by 11% based on out-of-sample mean absolute error, although the improvement is considerably higher for smaller state-industry cells. We also produce new county-level estimates that could provide more timely granular estimates than previously available. We develop a novel test to determine if these new county-level estimates have errors consistent with official series. Given the level of granularity, we cannot reject the hypothesis that the new county estimates have an accuracy in line with official measures, implying an expansion of the existing frontier. We demonstrate the practical importance of these experimental estimates by investigating a hypothetical application during the COVID-19 pandemic, a period in which more timely and granular information could have assisted in implementing effective policies. Relative to existing estimates, we find that the alternative payroll data series could help identify areas of the country where employment was lagging. Moreover, we also demonstrate the value of a more timely series.

1 Introduction

Economic statistics rely extensively on household and business surveys. While survey response rates have declined, threatening the reliability of traditional survey data sources, technological advances have led to a tremendous expansion in the availability of “big data” sources. Big data offers potentially enormous benefits, including improved accuracy and coverage, faster production, cost savings, and new insights, all of which may help provide more useful information to the public and policymakers (Abraham, 2022). The benefits and unique insights from these data have been realized in numerous research studies, especially during the pandemic as the demand for timely and granular information ballooned (e.g., Aladangady et al. (2019), Autor et al. (2022), Chetty et al. (2020), Cox et al. (2020), Cajner et al. (2020), and Dunn et al. (2021)). However, big data sources also raise measurement concerns, including issues related to quality—big data may be messy and unstructured—and coverage—big data are not typically constructed to be representative of the population (Abraham, 2022). These problems may bias or diminish the accuracy of potential statistics, casting doubt on the utility of these data for statistical agencies. This paper lays the groundwork for a streamlined architecture that integrates traditional and big data sources by developing a method to systematically measure the benefits and costs of that integration. We consider these ideas in the context of improving regional employment statistics using big data sources.

While "big data" is itself a relatively new concept, statistical agencies have long traded off accuracy for timeliness and granularity when comprehensive data is delayed. There are numerous examples. The Bureau of Economic Analysis (BEA) publishes timely "advanced" national Gross Domestic Product (GDP) statistics that are revised multiple times as more accurate information becomes available from the US Census Bureau (hereafter, Census) and from other sources.[1] In this paper, we consider the specific case of the Current Employment Statistics (CES) survey data, a principal indicator of employment for the US. The Bureau of Labor Statistics publishes these data at the state level with a delay of approximately five weeks, even though more accurate information becomes available later from the Quarterly Census of Employment and Wages (QCEW), a far more comprehensive census of the US workforce.

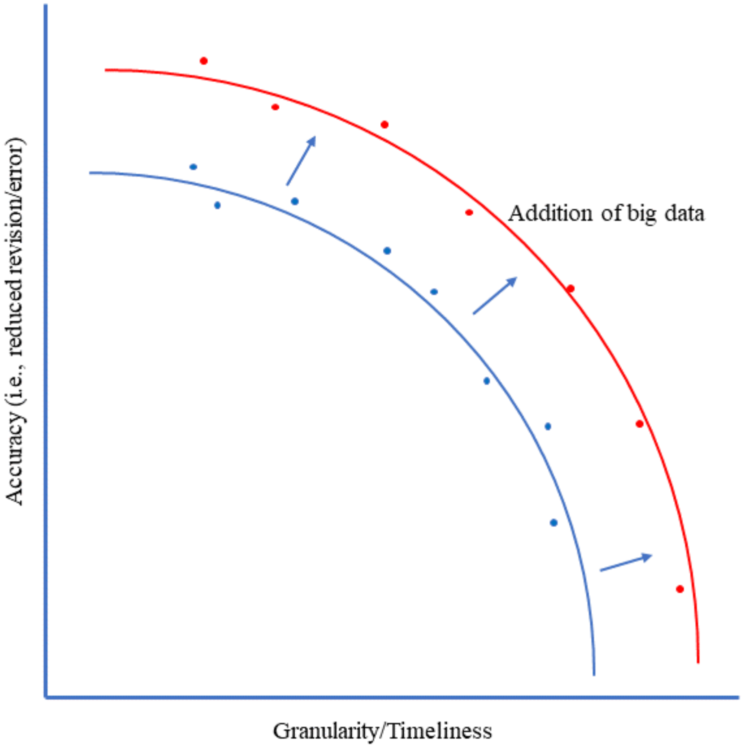
In each of these cases, the agencies have decided to publish data to provide more timely or granular information to the public, even though more accurate information is available in the future. This trade-off—whether made implicitly or explicitly—can in many cases be measured. Revisions observed between agencies' initial and final statistics at various increments of timeliness, aggregation, and granularity reveal the trade-offs that these agencies are presently accepting. Observed revisions offer a useful benchmark that may be applied to measure the value of improved estimates or new estimates.

Conceptually, the trade-offs among the various estimates produced by the statistical system may be viewed along a production possibility frontier, reflecting the competing goals to produce the most accurate estimates possible, but also produce the most timely, relevant and granular statistics. The main idea is illustrated in Figure 1. The key question is whether big data can be used to expand the existing frontier of economic measurement by increasing the accuracy (i.e., reducing revisions) of existing estimates or producing new, granular and more timely estimates than previously possible.

We focus on the potential benefit of big data to provide more granular and timely information about employment statistics than is available from traditional sources. We make two main contributions: First, we develop a framework for assessing economic statistics that weighs the

[1] The geographically most accurate information for regional GDP estimates is available only every five years with each Economic Census, yet BEA publishes state-level statistics at an annual and quarterly frequency.

Figure 1: Conceptual Production Frontier of Economic Measurement



Notes: The figure provides a conceptual production possibility frontier showing a trade off between the accuracy of a statistics and its granularity or timeliness. The blue line captures the existing frontier of the statistics. The red line represents the potential expansion of the frontier with the introduction of big data.

potential benefits of additional timeliness and granularity against potential bias and noise of these estimates. The foundation of the assessment that we propose is built from information on the errors currently observed in official measures to establish a benchmark, which may be viewed as the current frontier. Second, we apply the framework to assess the value of adding an alternative data source from a payroll processor to improve or expand on published statistics of employment. In other words, we measure whether the frontier can be expanded along some dimension, such as improved accuracy, timeliness or granularity relative to existing statistics.

Specifically, we apply these ideas to consider the value that data from a payroll processor may add to existing published employment statistics from the CES at the regional level. We use data from a major payroll processor that administers payroll services for hundreds of thousands of clients and pays over 5 percent of private sector employees in the United States, primarily covering small and medium-sized businesses. Certain limitations of CES suggest that additional payroll data may add some unique value. One limitation is that the CES data is a survey, so there is some loss of accuracy relative to the QCEW estimates, so it is possible that combining the CES data with payroll data could produce more accurate estimates of state or MSA employment than is produced currently by using CES alone. In addition, the CES estimates are available at the state level and for some select MSAs, but not available at the more geographically detailed county level. The payroll data, which we use at the county level, can be used in combination with CES to produce county-level estimates of employment. To potentially improve estimates at the state level or produce new county-level estimates, we use the established relationship between the CES initial estimates and final estimates based on QCEW as a benchmark.

In the background of this analysis, a critical detail lingers regarding how precisely to incorporate the information contained in an alternative data source into official statistics. We view this as comprising two steps: First, it is necessary to obtain the best signal possible from an alternative data source, the payroll data in this case. Next, it is necessary to determine how to use the alternative series, potentially combining the constructed series in a model with other available information to estimate a statistical series with desirable properties. In our application, we find that to add value, this first step is of critical importance. The input series from the payroll data is constructed to remove certain sources of noise, and in particular, focus on a rolling panel of employers, similar to the methodology that currently applies to the CES estimates. We demonstrate that this alternatively constructed constant-employer series has substantially more explanatory power than a raw or unadjusted payroll series.

Applying the refined payroll series, we use statistical models to construct estimated series in an attempt to improve on existing measures (that is, we apply the second step noted above). Specifically, we construct measures to improve on the existing CES state-level, 2-digit industry series, and we also produce a new county-level, 2-digit industry series. Next, with these series, we apply one last critical step, to construct cross-validation performance metrics to formally evaluate how well the new estimates perform relative to the existing economic measures. Using our best performing series, we find that these payroll data can provide increased accuracy (i.e., reduced revisions) when used in the process of estimating official statistics, thereby expanding the production frontier of statistical agencies. More precisely, we find that when CES is used in combination with payroll data, accuracy improves and errors are reduced in state-level estimates, reducing mean absolute error by 8-11 percent overall. We find around a 19 percent improvement in mean absolute error when we focusing on industries and mid-size employment cells where the payroll data adds significant value. These industry-employment categories where we observe substantial improvement account for around 50 percent of the state-industry cells and 39 percent of employment.

At the county level, there are no current official CES estimates, so determining whether estimates at this level of granularity meet or exceed existing standards poses a unique challenge. To address this issue, we develop a new methodology that uses existing official measures to establish the expected "frontier" given the level of granularity. An important contribution of this paper is that we develop a statistical test of whether the county level estimates are consistent with the existing frontier. This is the first statistical test that we are aware of that tests whether a new series meets or exceeds the standards implicit in official statistics. We find that our newly produced county estimates meet or exceed the level of accuracy needed for that level of granularity. In other words, we cannot reject the null hypothesis that the county estimates are consistent with the existing standards. In addition, we find that compared to using the associated MSA or State CES estimates applied to counties (e.g., using CES estimate for the corresponding counties within an MSA), the county level estimates improve mean absolute error by 9-13 percent. In summary, we produce county-by-industry level estimates with error rates that are reasonable based on the existing frontier, offering additional geographic granularity not previously possible.

We focus on the combined payroll and CES series for much of our analysis as the revised payroll data series alone produce errors that are higher than the CES estimates. However, we show that in times of greater economic volatility, the signal to noise ratio increases in the payroll data, providing a more meaningful signal of local employment that is more timely than the CES. In other words, the timeliness of the payroll data becomes more valuable during large economic fluctuations, precisely when economic measurement is most important.

We explore the practical importance of timely and granular estimates from alternative data series by examining an application, which is motivated by the economic stimulus initiated at the onset of the COVID-19 pandemic. Numerous programs were implemented at a national level to address the pandemic, including stimulus checks, expanded unemployment benefits, and the Paycheck Protection Program (PPP). During the onset of the pandemic, there was limited timely information available making precise targeting of the funds challenging, and prompting broadly targeted stimulus. The PPP program was intended to be more targeted to saving jobs for employers needing assistance, but subsequent research by Autor et al. (2022) and Chetty et al. (2023) demonstrate that the costs associated with the PPP were around $150k-$300k per job saved. This raises the question of whether more timely and granular information could help improve policy decisions and lower the cost of achieving specific policy goals.

Rather than examine a specific policy, we examine counterfactual scenarios around the pandemic to demonstrate how more timely and accurate information might help improve targeting and the efficient allocation of resources. The analysis we consider is how well the timely payroll data could help identify the industry-county pairs that were hardest hit by the pandemic. We find that the more timely payroll data alone performed substantially better than assuming areas were affected equally, providing useful information during this especially sharp downturn in the economy. In the months following the pandemic, we also show that the combined payroll-CES data can improve the accuracy of identifying the hardest hit counties by around 13%, relative to the CES estimates alone.

Our discussion and reporting of these errors and inherent trade-offs is in the spirit of Manski (2015) that advocates for statistical agencies to better communicate estimates of uncertainty. A contribution of our paper is to communicate that economic measures are produced with some error, and this is not a mistake of the statistical system. The errors are a natural part of the production possibility frontier of the agencies attempting to produce both the most accurate statistics, with the fewest revisions possible, but also more timely and granular statistics, which may have additional error.[2] The additional error of the detailed statistics has the added benefit of providing unique and actionable information to policymakers and the public.

Our paper also builds on comments in Abraham (2022), which notes that statistical agencies should be asking whether the incorporation of alternative data into official statistics results in statistics that are of similar or higher quality. We agree that this is an important idea, but it also raises challenging questions. When are new statistics considered to be of similar quality or an improvement? What if national statistical offices are producing a range of statistics of varying accuracies? Our paper provides a framework for answering these questions. Indeed, the goal of this paper is to lay out a methodology to clearly expand the frontier of economic measurement.

[2] For example, the mission statement of BEA is to promote: "... a better understanding of the U.S. economy by providing the most timely, relevant, and accurate economic accounts data in an objective and cost-effective manner."

Formalizing these ideas not only assists in measuring the value of new data at the agencies and whether the production of a new series is worth it, but it also helps in communicating to data users the value of new data sources.

In the next section we discuss some background literature leading up to this work. After this, we discuss the data sources, with particular focus on details such as how they are constructed and how this affects the accuracy and timeliness of statistics produced from these sources. Next, we dedicate a section to a study of the accuracy of existing official statistics, including the preliminary CES in predicting revised CES or in predicting QCEW, expanding on the concepts illustrated in Figure 1. The next section is dedicated to the statistical methods we use to model CES or payroll information (or to integrate information from both) to produce early estimates of employment by state or county and by industry. Finally, we provide an application motivated by policies during the pandemic, to demonstrate the usefulness of more granular and timely data.

2 Background

In recent years, several studies have used big data for novel measures of economic activity (Aladangady et al., 2019; Autor et al., 2022; Chetty et al., 2020; Cox et al., 2020; Cajner et al., 2020; Dunn et al., 2021). Each of these data sources benchmark off of official data sources and have also proven useful throughout the pandemic recovery. Our work most closely builds off of Chetty et al. (2023). Chetty et al. (2023) is particularly noteworthy as they develop multiple private sector data sources, including multiple payroll sources, to understand trends around the pandemic, and also make these series available to the public in real time. A major point of their paper is to demonstrate the usefulness of private sector data for public statistics, encouraging government statistical agencies to build on this work by generating complementary statistics. Our paper is intended to move this agenda forward by demonstrating objective improvement in official measures and expansion of the frontier currently offered by official statistics.[3] Perhaps as important as showing where the private sector alternative data adds value, we are also able to show where the value is more limited, ensuring that the focus is on the best available estimates.

To emphasize the points made earlier, statistical agencies are very familiar with the idea of working with various data sources of differing accuracy, timeliness and granularity. Currently,

[3] This goal is related to that of Chen et al. (2019), who focus on reducing revisions for the services component of GDP.

BEA publishes a whole collection of quarterly national accounts statistics one month, two months, and three months after the end of each quarter, with revisions between each publication. These statistics include component (sector) detail as well as industry detail for the business sector. BEA draws on a variety of data sources, including from statistical agencies at Census and BLS, but also numerous private sector data sources, as emphasized in Moyer and Dunn (2020). BEA also publishes a revision of these statistical series each year and benchmarks them every five years based on the most recent economic census. In addition, BEA publishes a similar (if somewhat less complete) set of statistics by state for each quarter and annually publishes GDP and other statistics by county.

Census has numerous examples of this idea as well. Census produces an advance estimate of monthly retail sales as well preliminary and revised estimates one and two months later, respectively.[4] The monthly data is then revised using an survey information from the Annual Retail Trade Survey (ARTS). These surveys are then followed by even more complete economic census measures. Similarly, BLS publishes a variety of employment statistics, including monthly information from the CES. CES-based statistics are later revised based on new information, especially when QCEW information becomes available. It is this last example that we focus on in-depth in this paper. Across all of the agencies there is a long history of using a variety of data sources and updating estimates as more complete information becomes available.

3 Data

In this section we detail the relevant features of the data that we consider, including the information source from which the statistics are constructed, timeliness (the elapsed time between the reference period and publication), and measures of accuracy. Our work focuses on three main data sources: CES, QCEW and payroll data, which we discuss in turn below. Because the CES and QCEW are well-established sources of employment information, we include only a general overview of these programs. In contrast, the payroll data source is discussed in more detail, with some space dedicated to a discussion of the construction of these data. Finally, because we are focused on potentially reducing the latency of employment information, we provide a detailed discussion of the timing of the availability of CES statistics and payroll

[4] The advance measure is produced using the Advance Monthly Retail Trade Survey whereas the preliminary and revised estimates are produced from the Monthly Retail Trade Survey responses.

information (the latter being somewhat customizable).

3.1 Current Employment Statistics (CES)

The CES survey is conducted by the BLS and provides timely and detailed information about employment in various industries and geographic areas across the United States. The CES program is known for its timeliness, as its employment statistics are published on a monthly basis, usually within the first week of the month following the reference month. The CES sample covers around 27 percent of employment, and samples larger firms at a higher rate than smaller firms.[5] For example, the employment statistics for March generally come out in the first week of April at the national level, although there is more latency in the regional statistics, which usually come out approximately two to three weeks after the national estimates. These statistics allow policymakers, businesses, and economists to closely monitor changes in employment and economic trends, although even more timely information could arguably be beneficial during economic shocks, such as during the COVID pandemic.

CES publishes data at various geographic resolutions, including national employment statistics and employment statistics by state and for metropolitan areas. This granularity allows users to analyze employment trends in specific areas, helping to inform local economic decisions and strategies. However, statistics by county and finer granularities are not available from CES.

The published CES numbers are revised as more information becomes available, especially after statistics from the QCEW are published for the reference quarter. For our analysis, we are working with monthly employment statistics from CES at the state and metro area from January 2013 to June 2021. Our data include all vintages of these statistics, and we evaluate the accuracy of the estimates using a real-time concept.

3.2 Quarterly Census of Employment and Wages (QCEW)

The Quarterly Census of Employment and Wages (QCEW) is a vital data source maintained by the BLS that offers comprehensive insights into employment, wages, and establishments across various industries in the United States. The data is collected through mandatory quarterly reporting by employers to state workforce agencies, which verify and compile the data before submitting it to the BLS. It includes information on job numbers, total wages, and average

[5] The use of sample weights in the estimation process prevents larger or smaller firm bias (See https://www.bls.gov/web/empsit/cestn.htm#tb1).

wages, broken down by industry, geography (from national to local levels), and establishment characteristics.

The QCEW is more comprehensive and accurate than CES at a granular level because it covers nearly all employers, provides detailed establishment-level data, and offers finer geographic and historical detail. Because the QCEW is more accurate, it is used in the process of revising the CES data. CES data is adjusted to align with QCEW figures, improving the accuracy of CES estimates and ensuring they better reflect actual employment trends. This revision process occurs annually and helps fine-tune the reliability of CES data.

However, the more detailed QCEW data is only available with a substantial latency of about six months after the reference quarter. This implies that the more accurate information cannot be used to inform the public in a more timely manner. The QCEW data that we use in our analysis provides monthly employment data and covers the time period from January 2017 to June 2021.

3.3 Private Sector Payroll Data

The payroll data we use is from a company that provides payroll services to businesses, particularly small and medium-sized enterprises (SMEs) in the U.S. As part of their regular business, the payroll processor collects detailed microdata on employment. The payroll data covers over 5 percent of private sector employment. Because the company processes the payroll microdata, the estimates generated from the data are customizable in regards to timeliness and granularity. All data used in this work is aggregated and anonymized (see the Appendix for more details).

Similar data was used by Chetty et al. (2023), who demonstrated how private sector data sources, like the one we use here, could be used to generate insights around the pandemic. In addition, they were able to provide the associated data source to the public. Our analysis builds off their work, and focuses specifically on how data from the payroll processor can add value, relative to existing official statistics.

The payroll data is a convenience sample, so the properties of the data and coverage match those of the payroll business, which specializes in helping small- and medium-sized businesses. In addition to being balanced toward SMEs, the sample reflects the payroll processing business environment which varies across industries and geographies.

3.3.1 Alternative Samples of the Payroll data

Alternative data sources, such as the payroll data set we are working with, have limitations because they are based on convenience samples that may not be representative of actual changes in the economy. For example, if the market share of the payroll processor were to increase substantially, using the raw data would suggest a large, fictitious increase in total employment. Numerous examples in the literature suggest that how these alternative data sources are constructed may have important implications for the data signal extracted, including for card transaction data (Aladangady et al., 2019), medical claims data (Dunn et al., 2014), and employment data (Cajner et al., 2020). In each case, the study sample varies with the business practices of the data contributor, which does not necessarily reflect the broader economic activity that researchers are attempting to capture. For employment, Cajner et al. (2020) hold the employer constant from period-to-period and examine changes in employment over time (which is also a similar approach to how the CES is constructed).[6] At the same time, it is not immediately obvious that holding the employers in the sample constant is preferable, as it will miss entry and exit from the sample. In particular, Chetty et al. (2023) examine unadjusted payroll data immediately around the pandemic, a period of time when the effects of entry and exit in the economy are a dominant feature.[7] In our analysis we construct both samples, full unadjusted payroll data and a continuing sample, and we evaluate the signal from each.

3.4 Timeline of Estimates—CES and Payroll Data

The CES program publishes statistics on fixed timelines, whereas it is feasible to compute payroll data at varying delays relative to the reference period. However, reductions in latency are associated with losses of information. Figure 2 shows the timeline of both data sources. For the CES, the reference period is the pay period that includes the 12th day of the month.[8] In our example, we focus on the reference period of March 12th, to maintain comparability with CES.

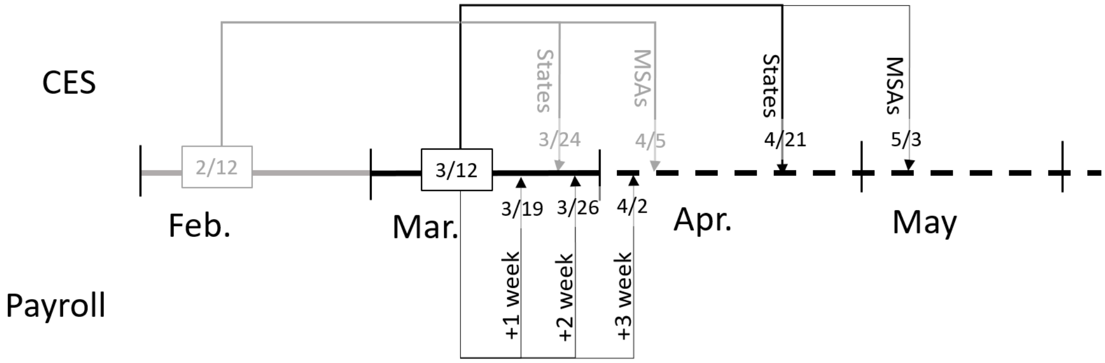
[6] Aladangady et al. (2019) examine card transactions based on a large card processing intermediary, Fiserv, and find they obtain a better signal by constructing data based on a rotating panel of merchants. Dunn et al. (2014) examine claims data from a data source that includes multiple contributors from large employers and insurance companies, and they find more stable estimates by holding the data contributors in their sample constant over time. In addition to tracking employers over time, the BLS also applies a firm birth-death model to account for entering and exiting employers.

[7] The methodology of Chetty et al. (2023) does make adjustments for extreme outliers, but does not rely on a continuing sample.

[8] For example, if a company pays biweekly and their payroll covers the period from March 5th to March 18th, they will report payroll for that period. Similarly, monthly payroll or weekly payroll that includes that reference data are similarly recorded. If the 12th day of the month falls on a weekday, the reference period includes that day. If it falls on a weekend or holiday, the reference period includes the nearest weekday.

For this reference date, the state and MSA level data is not available until around the 3rd week of the following month, April 21st in this example.[9]

Figure 2: Timeline of Current Employment Survey and Payroll Data



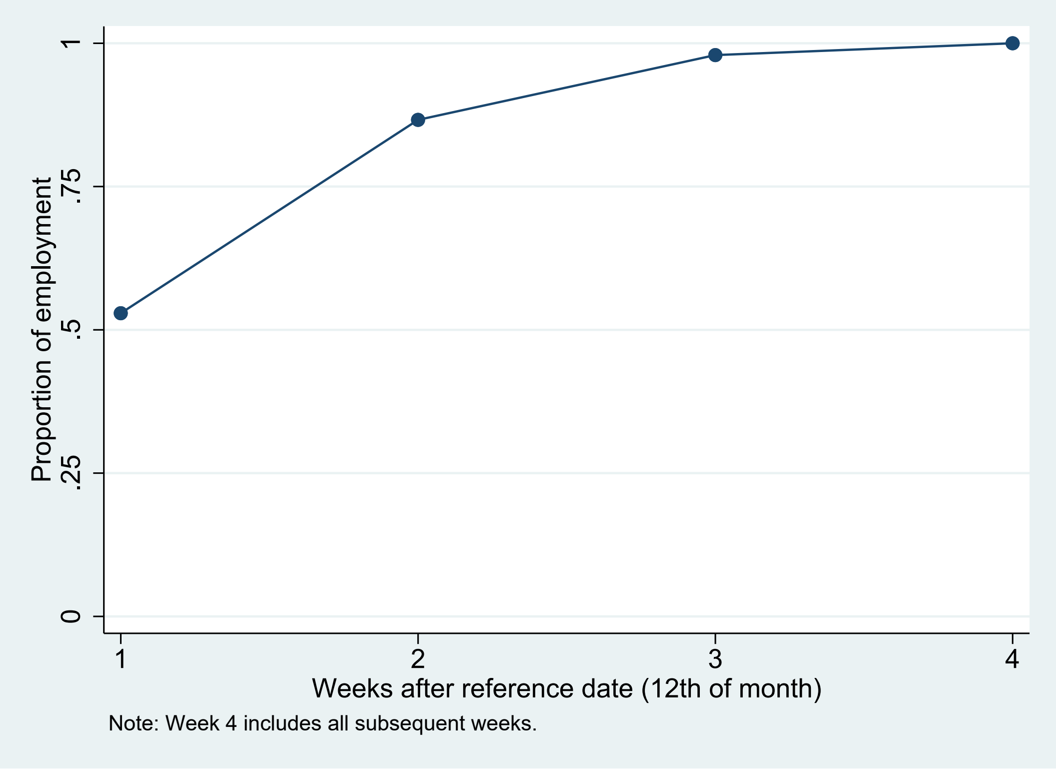
Notes: This figure shows the timeline of the CES and the payroll data. The line represents a timeline of dates. The top portion of the figure shows the key dates for CES. March 12th is the reference period we focus on in the figure. The state-level and MSA-level estimates for this reference date are not available until April 21st, about five weeks after the reference period. National estimates, not shown here, come out about three weeks after the reference date. The bottom portion of the figure shows the timing of the payroll data that we examine. Recall that the data is customizable, but we have chosen to analyze four cuts of the data. Employment estimates one week, two weeks, and three weeks after the reference period, as well as the "final" estimates, which includes all payments and adjustments at the time we extracted the data.

The information from the payroll processor is instantaneous. Therefore, the timeliness of the payroll data is customizable, so that we can examine the payroll any time after the reference date. As shown on the bottom of Figure 2, we analyze the data one week, two weeks, three weeks after the reference period, as well as the full data, which includes all subsequent weeks and any potential corrections to payroll.

While the payroll data has no lag in payroll information, the payroll frequency of employers affects the timing of payroll information. Returning to the example, after one week from the reference period, because some of the employers are on a bi-weekly pay schedule or monthly pay schedule, they may not yet have payroll that includes the reference date. After a second or third week, more of those employers are likely to enter the sample, providing more information on the payroll trends for that month. Figure 3 shows the proportion of employment reported at each period. After one week, we see 50 percent of employment reported, after two weeks we see over 80 percent reporting, and after three weeks, we have over 95 percent reporting.

[9] It is worth noting that the CES data at the national level is available with less of a lag, recall that it is generally the first week after the reference month, so the first week of April in this example.

Figure 3: Payroll Frequency and Payroll Timing



Notes: This figure shows the share of the employment total that comes in after the reference period. While Payroll information is instantaneous, the pay frequency affects when payroll information is recorded. For instance, employers that pay weekly will be recorded one week after the reference period, but those that pay bi-weekly may not be captured. The x-axis shows the timing after the reference period, 1-week, 2-weeks, 3-weeks, or the full sample (captured by the "4" along the x-axis). The y-axis is the share of employment, relative to the full employment based on the full sample. For example, one week after the reference period we see about 50 percent of employment information is available. By definition, the full sample, indicated by a 4 on the x-axis, captures 100 percent of employment.

Just as we analyze the different sampling strategies (i.e., full or continuing samples), we also evaluate the signals based on payroll data with varying latencies after the select reference period (i.e., 1 week, 2-weeks, 3-weeks, and the maximum period).

4 Observed Error in Employment Estimates

Before evaluating how payroll data may contribute to improving regional employment measures, we must first evaluate the current accuracy (in other words, "tolerance for error" or "tolerance for revisions") in official measures of employment. Table 1 shows a table of CES errors between the initial and the final CES estimates. We show this error at different levels of aggregation by geography and industry. Two left-hand columns show the level of geographic aggregation (i.e., MSA, state, or national) and industry aggregation (i.e., 3-digit, 2-digit or aggregate). The next three columns report alternative measures of error, including the mean absolute error (MAE), mean absolute scaled error (MASE)[10], and 1-R^2. Each measure captures alternative ways of measuring errors, and all three are measured so that lower values indicate greater accuracy.[11] The last column shows the number of observations.[12]

A clear pattern emerges from Table 1, which demonstrates the implicit trade-off between granularity and accuracy. Across all three measures, we generally find that error declines at higher levels of geographic or industry aggregation, or conversely, that error increases at finer levels of detail. This error is presumably acceptable, as the greater detail also provides additional information to the public regarding the industry and geography of changes in employment.

Table 1: CES Differences (CES initial vs CES final)

Geo	NAICS digits	Mean Abs. Error	MASE: MAE/MAD(Y)	1-R2	N
MSA	3d	0.0129	0.7726	0.2998	94,172
MSA	2d	0.0119	0.7814	0.3873	157,528
MSA	0d	0.0068	0.5828	0.3856	29,024
State	3d	0.0111	0.6605	0.2136	136,207
State	2d	0.0086	0.5580	0.1636	86,712
State	0d	0.0030	0.2615	0.0339	4,876
USA	3d	0.0048	0.3221	0.0951	6,900
USA	2d	0.0032	0.2347	0.0342	2,024
USA	0d	0.0014	0.1306	0.0052	92

Notes: The first column of the table indicates the geographic unit (i.e., MSA, State, or National), the second column indicates the industry detail using the hierarchical North American Industry Classification System (NAICS) (i.e., 3-digit NAICS, 2-digit NAICS, or aggregate), and columns (3), (4) and (5) indicate different measures of error. MAD is Mean Absolute Deviation. $R^2 = 1 - MSE/Var(Y)$ where MSE is the mean squared error. Observations are weighted by lagged QCEW employment. The figure shows that errors generally increase at finer levels of industry or geographic granularity.

We produce similar measures of error at the county level in Table 2. While there is no timely county-level estimate from CES, we assume that local information regarding employment may still be inferred from a higher level of CES. For instance, for understanding employment in

[10] This is, $\frac{MeanAboluteError(MAE)}{MeanAbsoluteDeviation(MAD)}$

[11] For example, MASE and 1-R^2 account for how much of the variation is explained by the prediction, while MAE only captures the absolute error. In addition, MASE and 1-R^2 are different as MASE and MAE are based on L1 errors, while 1-R^2 is an L2 error (i.e., a squared error).

[12] Due to suppression, we observe more observations at the 2-digit industry level for MSAs, relative to the 3-digit industry level.

Montgomery County, Maryland, which is in the Washington, D.C. metro area, policymakers in Montgomery County may look at the associated CES estimates for Washington, D.C. while they wait for more detailed information from QCEW. In this spirit, we compute the error at the county level between the initial CES estimate and the final QCEW estimate.[13] Similar to before, we find that the more disaggregate estimates show higher levels of error, across all three measures. We can also see this across the two tables. For example, in Table 2 the MAE at the 2-digit county-level is 0.018, while the error at the 2-digit MSA-level in Table 1 is 0.012.

Table 2: CES Differences with County (CES initial vs QCEW)

Geo	NAICS digits	Mean Abs. Error	MASE: MAE/MAD(Y)	1-R2	N
Counties	3d	0.0214	0.8907	0.6684	4,844,287
Counties	2d	0.0180	0.8779	0.6240	3,124,127
Counties	0d	0.0102	0.7511	0.5234	179,114

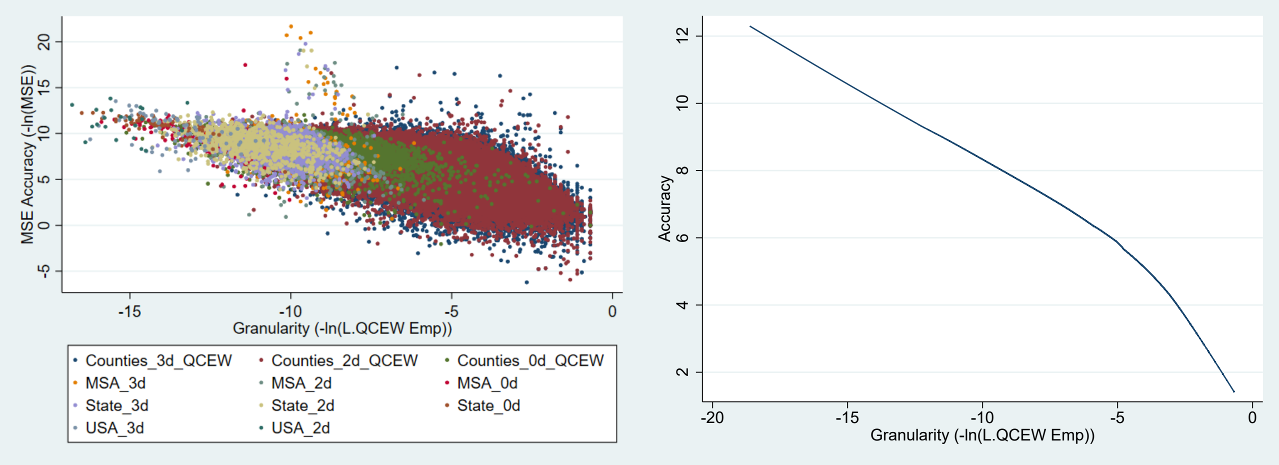
Notes: The first column of the table indicates the geographic unit (i.e., county in this table, in contrast to Table 1), the second column indicates the industry detail (i.e., 3-digit NAICS, 2-digit NAICS, or aggregate), and columns (3), (4) and (5) indicate different measures of error. CES is not defined for counties so we use the MSA growth rate for MSA counties and the state growth-rate otherwise. MAD is Mean Absolute Deviation, $N^{-1}\sum_i |y_i - \bar{y}|$. $1 - R^2 = MSE/Var(Y)$ where MSE is the mean squared error. Observations are weighted by lagged QCEW. The figure shows that errors generally increase at finer levels of industry or geographic granularity.

The main point is that there is an implicit trade-off between accuracy and granularity that may be gleaned more clearly from Figure 4. The left panel shows a scatter plot of errors at different levels of granularity. More precisely, the vertical access measures accuracy, so that accuracy increases moving up the y-axis. The y-axis is specifically measured as the negative of the log of the mean squared error (MSE). The x-axis is a measure of granularity, where granularity increases moving farther to the right. The x-axis is measured as the negative of the log employment level. The axis of this figure match those of the conceptual Figure 1. We generally see from this picture that the mass of points falls as estimates become more granular, demonstrating the trade-off between accuracy and granularity. The trade-off becomes more stark when we estimate a fitted line over the mass of these data points, as shown in the right panel of Figure 4.

This figure empirically describes the current production possibility frontier based on official statistics currently produced. While this trade-off is widely understood implicitly, this is the first empirical documentation of this trade-off that we are aware of. Next, we attempt to determine if

[13] For non-metro areas, we use the CES estimate for the state-level to compute the error.

Figure 4: Scatterplot of Error by Granularity and Associated Fitted Line



Notes: The figure in the left panel is a scatter plot of the errors based on different levels of granularity. The y-axis is actually the inverse of the errors, captured by the negative of the log of the mean-squared error (i.e., $-ln(MSE)$), so that farther out on the y-axis captures a reduction in error or improved accuracy. The x-axis is granularity, captured by the inverse of the employment size captured by the $-ln(lagQCEW)$. The units are for the corresponding geographic-industry categories from Tables 1 and 2. The figure shows as estimates are more granular, the accuracy falls. This is more clearly shown in the right-hand panel, which shows the lowess fitted values of the left-hand panel, indicating a clear downward slope.

the payroll data source may be used to expand this frontier.

5 Methodology

The goal of this paper is to investigate the potential for expanding the frontier of current employment statistics. The methodological steps for accomplishing this are:

1. Define the evaluation criteria and assess the accuracy of existing statistics across different levels of granularity or timeliness. This first step was accomplished in the previous section, as we demonstrate how L_1 and L_2 errors vary at different levels of granularity.

2. Produce statistics based on alternative data. In the case of our analysis here, we will focus on county- or state-level employment growth at the 2-digit industry level.

3. Apply cross-validation (described below) to determine the out-of-sample error of these estimates.

4. Finally, evaluate the performance of the new statistics by comparing errors to existing

accuracy levels. For improving existing estimates, we want the cross-validated errors to be lower than existing errors. For producing new estimates, we want the cross-validated errors to fall within a reasonable range, given existing accuracy.

As a concrete example, consider the case of producing new county-level estimates at the two-digit industry level. Based on Tables 1 and 2 we have an established accuracy level for this level of granularity. Current CES estimates applied to the county, 2-digit level produce a MAE of 0.018, so errors from a newly produced series would clearly need to be better than this. Another standard we could use is the error level for the closest series that is explicitly produced by CES – i.e. 0.012 for the MSA, 2-digit estimate. Since this series covers a higher level of aggregation, anything approaching this would have an acceptable level of error.

Now suppose we produce alternative estimates and find a MAE of around 0.20. In this case, the new estimate actually performs worse than CES alone, so the estimate provides no additional value and does not expand the frontier. However, suppose we produce an estimate that has a MAE of 0.13 for the 2-digit county level. In this case, the estimate is clearly better than the current county-level error of 0.18, so it likely provides some information relative to existing estimates. In addition, it is close to the MAE that we observe at the MSA level, suggesting this is near the existing accuracy level given the level of granularity selected. We could argue that this new estimate expands the frontier relative to existing statistics.

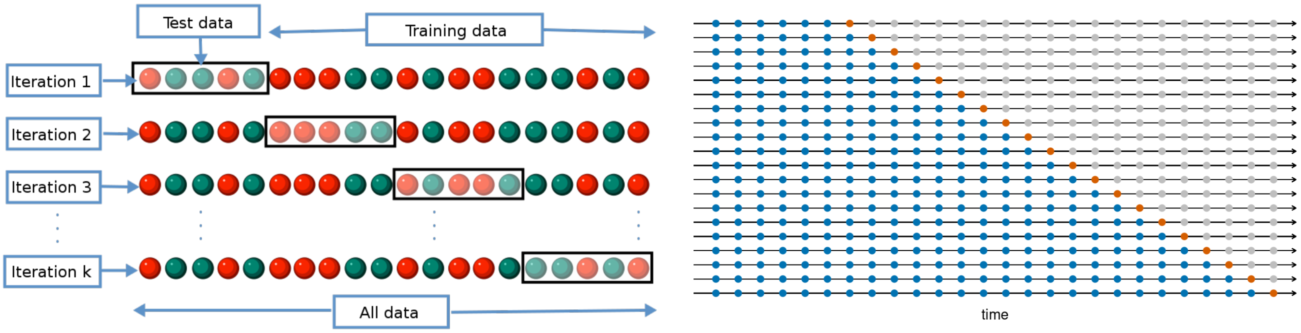
5.1 Model of Estimation

In this analysis we examine the additional value of payroll data, relative to, and in addition to, CES estimates at the regional level. We compare unmodeled estimates of CES with alternative models that include payroll and CES. Specifically, the prediction model takes the following form:

$$Y_{i,j,t} = f(CES_{ijt}, Payroll_{ijt}) + \epsilon_{i,j,t} \tag{1}$$

where $Y_{i,j,t}$ is monthly QCEW employment growth for geography i, industry j, and time t. The function $f()$ indicates linear functional forms for growth rates derived from CES and payroll data, where we also control for characteristics of these data sources (e.g., accounting for payroll coverage in an industry/geography). We also allow the payroll estimate to enter by county-industry, state-industry and aggregate county.

Figure 5: Cross-validation Methods: Cross-fold and Rolling One-step-ahead



Notes: The figure visually displays the two methods of cross-validation that we apply in our analysis. The left panel shows k-fold cross-validation that randomly selects a portion of the data as a hold-out sample, trains the model on the remaining data, then evaluates the model on the hold-out sample. This is repeated k-times, allowing all portions of the data to be part of the hold-out sample. The panel to the right shows the rolling one-step-ahead cross validation method. An initial time period is used as the training sample, while subsequent time period after the training time period, is the hold-out sample. The error is evaluated. Next, the training time period and the hold-out sample are advanced to the next period, process is repeated. This is done until the end of the sample is reached.

5.2 Cross-Validation

We evaluate the performance of the estimates by applying cross-validation. We apply two types of cross-validations, which are shown visually in Figure 5. The first type of cross-validation is called a cross-fold and the corresponding visualization is shown in the top panel of Figure 5. Cross-fold cross-validation involves dividing the dataset into k subsets or "folds," training the model k times on different combinations of training and validation sets, and aggregating the results to estimate the model's overall performance. This technique helps in evaluating a model's robustness, identifying potential issues like overfitting, and obtaining a more reliable estimate of its performance. The advantage of this approach is it uses all the time-series variation in the data to evaluate the error, but a major disadvantage is it does not match how estimation works in practice. In practice, all geographies are estimated at a single point in time, rather than estimating randomly selected locations across different points in time.

The second cross-validation is a rolling one-step ahead estimation, also known as time series cross-validation, and is shown in the bottom panel of Figure 5. Rolling one-step ahead cross-validation involves iteratively training a model on historical data up to a specific point in time, and then using the model to make one-step-ahead predictions for the next time step. This process is repeated for each data point in the time series, allowing for the assessment of the model's performance in a time-sequential manner. The main advantage of this approach is that it is closer to how estimation works in practice for the series we are producing, so it is the preferred methodology. The key limitation of the one-step ahead cross-validation is that it does not use the entire time series to validate the data.[14]

6 Results

In this section, we examine the value of adding payroll data series to estimating regional employment estimates. To do this, we first measure the value of alternative series produced using the payroll data set alone. Next, we use the payroll data series to examine whether we can improve or expand on existing official series.

6.1 Evaluating Payroll Data Samples

In evaluating the payroll data source, we examine the value of different constructions of the payroll data series.

6.1.1 Sample Size: Full vs Continuing

Recall that we have produced a variety of series, one that uses the full payroll raw data and one that uses a sample of continuing employers. One limitation of the raw data, mentioned earlier, is that it varies with the growth of the payroll business, which may vary in ways that are different than observed economic activity.[15]

The difference between the raw payroll data and the constant sample can be shown at the regional level using simple prediction models in Table 3. The table shows various estimates of state-by-two-digit industry on the final CES estimate. In the first column of Table 3 we show estimates based on the initial CES estimate, where CES is unmodeled, as shown by the coefficient held to a value of 1. The bottom of the table shows the key results of the various cross-validation strategies, including cross-fold and rolling, as well as MAE and $1 - R^2$. We will primarily discuss the rolling MAE cross-validation, as it is arguably the simplest measure that

[14] For example, one might think with the COVID pandemic in the data series, we would be much better at predicting future large swings in employment, so cross-fold may do a better job capturing future error rates, relative to the one-step ahead estimation that only uses the pre-pandemic period to predict variation during the pandemic.

[15] For example, a payroll processor may take a larger share of the market for payroll processing services, even as the employment in the economy falls.

more accurately captures the real-world production process, but flag the other measures when there are notable differences. The unmodeled CES estimate shows a MAE of 0.0088. Column (2) of Table 3 shows the prediction of final CES using the full payroll data series. The coefficient on the payroll employment growth rate is positive and significant, with a coefficient of 0.12, but the MAE is substantially higher than the initial CES estimate at 0.014, an error rate almost 60 percent above CES levels. Column (3) is the same as column (2), but uses the payroll growth rate from the continuing sample. The coefficient on the growth rate is 0.59, so it is substantially less attenuated than the full payroll sample estimate, and the prediction is also substantially better, with a MAE of 0.0124. Similar results are also found when we apply weights based on the size of the employment cells in columns (4) and (5).

Overall, this analysis shows that the economic signal coming from the continuing sample is substantially stronger than the raw full sample. However, it also shows that both CES and payroll are strong predictors of employment growth. This suggests that it may be beneficial to use payroll data in combination with CES data, which we explore later.

6.1.2 Timeliness

We also examine how the accuracy of the payroll data changes with the timeliness of the data extracted. Recall from Figure 3 that the amount of data available from payroll processors increases as the reference date increases, reflecting more employers submitting payroll around the reference date.

In Table 4 we examine how the payroll series of different levels of timeliness affect the predictive accuracy of the estimates. Column (1) of Table 4 shows the baseline CES results, identical to column (1) of Table 3. Column (2) shows estimates using the payroll growth rate constructed using a timely series of employment growth just one week after the reference period, as well as the various cross-validation estimates along the bottom of the table. In particular, it shows the rolling MAE to be 0.0130. Columns (3) and (4) are less timely and are constructed two or three weeks after the reference period, and we see the cross-validated error decreases as more data becomes available farther away from the reference date. By the third week, all the cross-validation measures are below those of the first week.

The fifth column shows estimates based on the fully adjusted payroll data (i.e., estimates

Table 3: Prediction Using Full Raw Payroll Data and Continuing Sample Payroll Data — State by Two-Digit Industry Estimates

	(1)	(2)	(3)	(4)	(5)
CES Emp Gr	1				
	(.)				
Payroll Emp Gr (all)		0.121^{***}			
		(23.21)			
Payroll Emp Gr (cont)			0.586^{***}		
			(42.94)		
Payroll Emp Gr (all; wgt)				0.0603^{***}	
				(24.09)	
Payroll Emp Gr (cont; wgt)					0.482^{***}
					(36.71)
Observations	41,924	41,924	41,924	41,924	41,924
$1 - R^2$ OOS (rolling)	0.177	0.823	0.371	0.430	0.430
$1 - R^2$ OOS (cf)	0.179	0.465	0.321	0.330	0.335
MAE OOS (rolling)	0.00876	0.0141	0.0120	0.0125	0.0124
MAE OOS (cf)	0.00853	0.0120	0.0108	0.0112	0.0111

Notes: This table shows how the full and continuing sample payroll data compares with the CES data. The first column shows the predictions based on the unmodeled CES estimates, where we have fixed the coefficient on CES estimates to 1. The second column shows the raw payroll data and the third column shows the continuing sample payroll data. We find that the continuing sample payroll data performs substantially better than the full sample data, which is not surprising given the pattern in Figure 2. However, we find that the unmodeled CES estimates perform substantially better than the two series constructed using payroll data. Columns (4) and (5) repeat estimates from columns (2) and (3), but are weighted based on lagged employment that incorporates QCEW.

Table 4: Prediction Using Payroll Data By Timeliness of Series

	(1)	(2)	(3)	(4)	(5)
CES Emp Gr	1				
	(.)				
Payroll Emp Gr (cont; t=1)		0.468***			
		(37.92)			
Payroll Emp Gr (cont; t=2)			0.603***		
			(43.81)		
Payroll Emp Gr (cont; t=3)				0.612***	
				(44.32)	
Payroll Emp Gr (cont)					0.586***
					(42.94)
Constant		-0.000478	-0.000672	-0.000837	-0.000827
		(-0.49)	(-0.74)	(-0.94)	(-0.93)
Observations	41,924	41,924	41,924	41,924	41,924
$1 - R^2$	0.179	0.379	0.330	0.319	0.318
$1 - R^2$ unweighted	0.309	0.719	0.622	0.585	0.585
$1 - R^2$ (excl. 2020-03 to -05)	0.347	0.772	0.695	0.677	0.677
$1 - R^2$ OOS (rolling)	0.177	0.509	0.389	0.374	0.371
$1 - R^2$ OOS (cf)	0.179	0.383	0.333	0.322	0.321
MAE	0.00853	0.0114	0.0109	0.0108	0.0108
MAE OOS (rolling)	0.00876	0.0130	0.0121	0.0120	0.0120
MAE OOS (cf)	0.00853	0.0114	0.0109	0.0108	0.0108

t statistics in parentheses

Outcome is QCEW Emp Growth (Mean Abs Dev=0.014). Observations at the State-NAICS2-Month level.

Dates=[2017,2021-06]. Variables are NSA. [aw=L_QCEWEmp2dST].

Restricted to observations with valid CES and Payroll Growth. Non-CES Equal models include state FEs.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table demonstrates how the signal form the payroll data changes with additional weeks from the reference period. Columns (1), (2) and (3) show the results from the first, second and third week after there reference period, respectively. Column (4) shows the estimate based on the complete payroll data, at the time of extraction accounting for all additional payments and adjustments. The estimates improve with each additional week after the reference period, as well as with the complete payroll data. Column (5) shows the CES estimates, which performs better than all the alternative payroll estimates.

that include all available changes or corrections to the payroll at the time the data is accessed).[16] This last estimate shows no improvement relative to column four. This finding is not surprising given that from Figure 3 over 95 percent of employment is accounted for aver the third week, so relatively little improvement is seen using data from subsequent weeks that may have more complete Payroll and recorded.

The unmodeled CES estimates in column (1) substantially outperform all the payroll series. The difference is particularly large, with the L2 error of the most timely payroll series, measured using the rolling $1 - R^2$, being four times the size of the corresponding CES estimate.

6.1.3 Summary

There are two important points to highlight in this section. First, the CES data performs remarkably well relative to payroll alone, which immediately suggests that the value added of payroll data may be to combine payroll data with CES to examine how the combined information may improve estimates or potentially lead to new and more granular estimates. That is, given the current level of timeliness of the CES data, can we expand the frontier by improving either the accuracy of existing statistics or producing more granular estimates? We will focus on this question in the next section.

The second point is that while the payroll data is noisier than the CES data, it can also be customized to be more timely. In other words, the payroll data can expand the production frontier along the timeliness dimension, relative to official statistics, but if the noise in the timely payroll data is sufficiently large, then it may not add value.

Determining the value of more timely data is challenging, as noted in Dunn et al. (2021), the value of the more timely data may depend on the magnitude of the signal relative to the noise, so the value of the estimates may change over time. Indeed, Chetty et al. (2023) demonstrated the value of payroll data sources during the COVID pandemic, a period in which the value of timely data grew substantially. We will revisit the value of payroll data as a more timely signal in a later section.

[16] Corrections or adjustments to payroll may be run well into the future, so these data may not reflect all future changes.

6.2 Improving Existing Estimates — State Estimates

In this section we examine whether the payroll data can be used to improve on existing state-level 2-digit estimates, applying simple linear models that combine CES and payroll information. When considering the combined data, we use payroll data three weeks after the reference period, as this information is available prior to the CES data and captures nearly all the employment estimates from the payroll processor.

The results of the analysis are reported in Table 5. Column (1) reports the baseline estimates using the unmodeled initial CES data, as shown in prior tables. Column (2) applies a linear regression to the CES data alone, and shows no improvement relative to column (1) on many of the metrics. Column (3) introduces a simple linear growth rate of the payroll data. Specifically, it includes both the CES initial estimate and the payroll estimate for the corresponding 2-digit industry category and state geography. Incorporating the payroll data in this way improves the estimates. Finally, in column (4) we incorporate additional information from the payroll data, including information on payroll coverage, as well as an aggregate measure of payroll data that combines information across industries. Incorporating this additional information improves all of the cross-validation measures, with the MAE rolling estimate declining by about 11%, and the cross-fold estimate declining by about 14%.

Figure 6 shows the improvement graphically and demonstrates how this combined information from the payroll data and CES expands the frontier of existing estimates. The axis in Figure 6 include measures of accuracy and granularity that match the axis of Figure 4. Similar to before, we plot the Lowess estimate of the errors along these two dimensions, comparing the accuracy of the modeled estimates and the unmodeled CES estimates along different dimensions of granularity. We find that the modeled estimates improve accuracy across a range of granularity levels. However, we do not observe improvement at the highest levels of aggregation (i.e., the left of the figure) likely because the large numbers of observations in the CES data already produce highly accurate information, so the marginal benefit of payroll data is small. As a reference, the value of -14 in the x-axis corresponds to an employment-cell size of around 1 million employees. We also observe little improvement in the estimates at the most granular level. When the employment-cell size is around 1,000 (corresponding to about -7 on the x-axis) we see no improvement in the estimates.

To highlight the potential magnitude of improvement, we can focus on areas where payroll

Table 5: Improving State-Level Estimates using CES + Payroll Data

	(1)	(2)	(3)	(4)
CES Emp Gr	1	0.853***	0.680***	0.607***
	(.)	(438.49)	(258.44)	(217.93)
Payroll Emp Gr (cont; t=3)			0.213***	0.140***
			(89.42)	(15.06)
Payroll (cont; t3) Coverage x Emp Gr				2.008***
				(27.72)
Payroll Emp Gr (st-Agg; t3)				0.0369***
				(10.52)
Constant		0.000138	-0.000601	-0.000173
		(0.21)	(-0.99)	(-0.30)
Observations	41,924	41,924	41,924	41,924
$1 - R^2$	0.179	0.179	0.150	0.135
$1 - R^2$ unweighted	0.309	0.309	0.281	0.249
$1 - R^2$ (excl. 2020-03 to -05)	0.347	0.347	0.317	0.306
$1 - R^2$ OOS (rolling)	0.177	0.216	0.176	0.154
$1 - R^2$ OOS (cf)	0.179	0.179	0.151	0.136
MAE	0.00853	0.00801	0.00758	0.00731
MAE OOS (rolling)	0.00876	0.00865	0.00810	0.00777
MAE OOS (cf)	0.00853	0.00802	0.00759	0.00733

t statistics in parentheses

Outcome is QCEW Emp Gr (Mean Abs Dev=0.014). Observations at the State-NAICS2-Month level.

Dates=[2017,2021-06]. Variables are NSA. Analytic weights are lagged QCEW 2-digit NAICS empl.

Payroll Coverage is calculated as lagged Payroll employment divided by lagged QCEW employment.

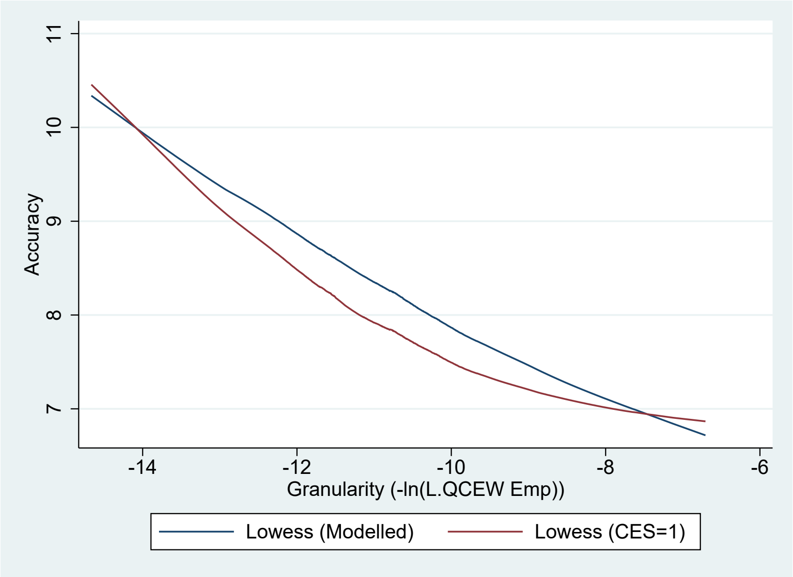
Coverage included when coverage is interacted with Gr.

Restricted to observations with valid CES and Pay GR. Non-CES Equal models include state FEs.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table examines the performance of the estimates combining CES and payroll data at the state, two-digit industry level. Column (1) shows the baseline, unmodeled CES estimates, column (2) allows for CES to be a fitted value, column (3) includes payroll and CES data, and column (4) includes additional covariates based on the payroll data. Column (4) demonstrates that the combined estimates perform better than the baseline CES estimates.

Figure 6: Production Frontier for State-level Estimates: Fitted Lowess Values of Accuracy and Granularity



Notes: The figure shows a lowess estimate that is parallel to the right panel of Figure 4. The red line shows the lowess fitted value based on CES alone, and the blue line shows the lowess value based on the combined modeled payroll and CES estimates, using the out-of-sample errors. The accuracy is measured in the $-ln(MSE)$ as in the prior figure, and the x-axis is the inverse of the log employment of the bin. The figure shows that the estimates may be improved along much of the range, except the most disaggregate estimates (i.e., the right portion of the figure) and the most aggregate estimates (i.e., the left portion of the figure).

data appears to add the most value. To do this, we run a simple regression of the accuracy of the blended payroll/CES estimate minus the accuracy of the CES estimate, where accuracy is measured as the negative of the absolute error. A larger positive value indicates a bigger improvement from the blended payroll/CES estimate. On the right hand side of the regression, we include different covariates, including dummy variables of the cell size of the employment estimate range (e.g., 10K-50K or 50K-100K employees), a measure of the payroll coverage for the cell, along with 2-digit industry fixed effects. Estimates are shown in Table A1. Consistent with the Figure 6, where a positive sign indicates that the accuracy of the blended CES payroll data is higher. We find the largest gains for mid-sized cells. We also find that the error difference to be partly explained by industry. Some NAICS industries like educational services (61) and real estate (53) show relatively large improvements.

We select the categories from the regression that indicate the largest improvement, including employment cell size range from 1K to 500K employees and 10 industries with the largest coefficients out of 18 total. These select categories account for 53 percent of the state-industry cells, and 39 percent of total employment. Using these select categories, we then evaluate the improvement in mean absolute error. For this subsample, we find the mean absolute error based on the CES is 0.0098 and the mean absolute error based on payroll and CES combined is 0.0079, which is about 19 percent improvement, nearly double the aggregate improvement. The magnitude of the improvement is important for two reasons: (1) it shows that there are areas where the improvement is substantial, which is obscured in the aggregate number; and (2) it highlights what the possible improvements might be if a data source larger than our payroll data were available.

Overall, the estimates in this section demonstrate the potential for payroll data to improve the accuracy of existing CES estimates, potentially reducing revisions and providing more accurate information sooner.

6.3 Producing New and More Granular Estimates — County Estimates

We next explore the potential use of payroll data to be used in combination with CES data to generate new county-level estimates. Table 6 shows the county level estimates. The first column reports the baseline estimates using the unmodeled CES data. When applying the CES data at the county-level, we use the corresponding CES estimate for the associated geography. If the county resides in a MSA, then we use the CES estimate for that MSA, but if the county resides outside an MSA, we use the corresponding CES estimate for the state. Column (2) fits a simple regression only including CES in the model, allowing the coefficient to be different than 1. Column (3) uses only a simple linear function of payroll county-level information and we find estimates to be worse relative to using solely CES data (i.e., columns (1) or (2)). Similarly, even when more complex functional forms of payroll data are added in column (4) (e.g., interactions with coverage, state-aggregates by 2-digit industry, and county-aggregates across industry), we still find the fit to be a bit worse than unmodeled CES data. The last two columns combine CES and payroll data. In column (5) payroll data enters in a simple linear specification, and we find most of the cross-validation metrics are better with the combined information, compared to the unmodeled CES data. Finally, column (6) includes both the CES data and additional payroll estimates. Here we see improvements across all the metrics, with about 9% improvement in the rolling MAE relative to unmodeled CES estimates. The improvement is 13% based on the cross-fold metrics.[17]

Overall, the MAE is considerably smaller than the baseline in column (1). Moreover, it is just a bit larger than the MAE we observe for existing MSA level estimates at the 2-digit industry level, reported in Table 1. We think these estimates suggest great potential for using these alternative series for producing more granular-level county estimates.

Similar to the state-level Figure 6, we can produce Figure 7 at the county-level showing accuracy on the y-axis and granularity on the x-axis. For counties, we see that all of the gain in prediction is for the medium and large industry-county cell sizes ranging in size from around 1,000 (i.e., approximately -7 on the figure) to around 500,000 (i.e., approximately -13 on the figure). For the smallest cell sizes of around 200 employees or less (i.e., around -5.5 on the figure), there is no improvement from adding the payroll data, likely due to considerable noise in both the CES and payroll data sources for this cell size.

The evidence from Table 1 shows how the payroll data can improve the prediction of county level estimates relative to CES alone, and Figure 6 shows that most of the improvement is for larger county-level cells. However, the BLS does not release CES data with the intention of representing MSA or state-level estimates as county-level estimates.

[17] The improvement based on the $1 - R^2$ metric is smaller, but we think much of that may be attributable to the pandemic, a period of time where there were special adjustments to the CES to account for large exits in the market. This is discussed in greater detail in the next section.

Table 6: Improving County-level Estimates using CES + Payroll Data

	(1)	(2)	(3)	(4)	(5)	(6)
CES Emp Gr (MSA or State)	1	0.843***			0.769***	0.667***
	(.)	(713.96)			(600.29)	(485.81)
CES State Emp Gr * 1(non-MSAs)		-0.210***			-0.301***	-0.244***
		(-137.28)			(-178.90)	(-145.91)
Payroll Emp Gr (cnty-naics2; t3)			0.0986***	0.0256***	0.0274***	0.00749***
			(46.98)	(10.91)	(13.56)	(3.77)
Payroll Emp Gr (t3) * 1(non-MSAs)				0.0117**	0.0492***	0.0318***
				(2.91)	(14.12)	(9.31)
Payroll Emp Gr (Cnty-naics2) x Payroll Coverage (naics2) (t3)				0.463***		0.293***
				(65.41)		(48.83)
Payroll Coverage vs QCEW (Cnty-naics2, 2020-01, t3)				-0.00162		-0.00248*
				(-1.21)		(-2.17)
Payroll Emp Gr (cnty-Agg; t3)				0.144***		0.0897***
				(143.58)		(104.94)
Payroll Emp Gr (state-naics2; t3)				0.307***		0.111***
				(342.04)		(133.09)
Constant		0.00110	-0.00166	-0.00171	0.000356	0.000109
		(0.17)	(-0.21)	(-0.24)	(0.06)	(0.02)
Observations	898,139	898,139	898,139	898,139	898,139	898,139
$1 - R^2$	0.501	0.490	0.717	0.590	0.441	0.423
$1 - R^2$ (excl. 2020-03 to -05)	0.236	0.253	0.0270	0.104	0.267	0.287
$1 - R^2$ (rolling)	0.497	0.562	0.934	0.732	0.524	0.484
$1 - R^2$ (CrossFold)	0.501	0.494	0.721	0.617	0.445	0.431
MAE	0.0157	0.0144	0.0171	0.0157	0.0139	0.0136
MAE (rolling)	0.0159	0.0151	0.0185	0.0170	0.0149	0.0144
MAE (CrossFold)	0.0157	0.0144	0.0172	0.0158	0.0139	0.0136

t statistics in parentheses

Outcome is QCEW Emp Growth (Mean Abs Dev=0.018). Observations at the County-NAICS2-Month level. Full dates=[2017,2021-06].
Variables are NSA. Analytic weights are lagged QCEW 2-digit NAICS empl. Payroll Coverage is calculated as
lagged Payroll employment divided by lagged QCEW employment. County aggregate growth is employment growth across all industries within the county. State-NAICS2 is state-wide growth for the given industry.
Restricted to observations with valid CES employment growth.

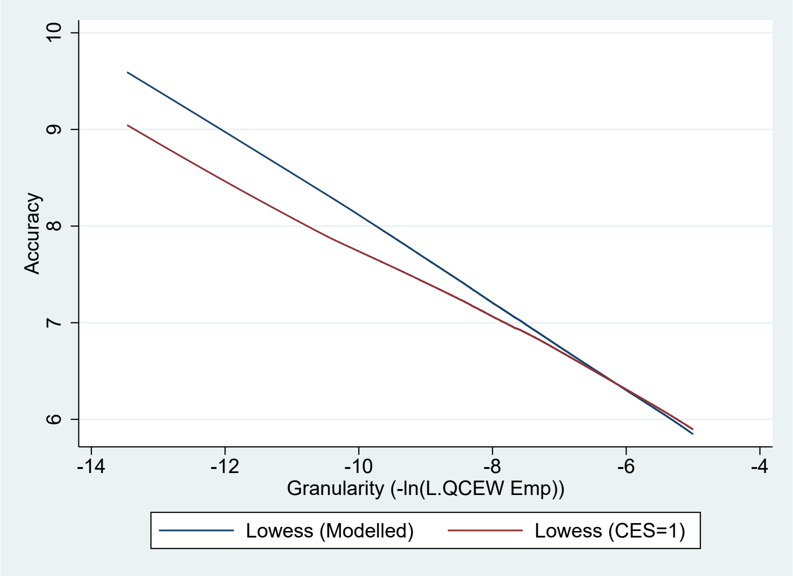
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table examines the performance of the estimates combining CES and payroll data at the county, 2-digit industry level. Column (1) shows the baseline, unmodeled CES estimates, column (2) allows for CES to be a fitted value, column (3) includes payroll data alone and without CES, and column (4) includes functional forms of payroll data, without CES. Columns (5) and (6) include both CES and payroll data. Overall, we find that performance, as measured by MAE, improves when CES and payroll data are combined, especially in column (6), where functional forms of the payroll data are included.

6.3.1 Statistical Test of Frontier Expansion at the County Level

Because the above bases of comparison for county-level accuracy are State and MSA statistics that are not necessarily intended to be used to convey county information, we propose an alternative approach that is based on tolerances inferred exclusively from official series. To

Figure 7: Production Frontier for County-level Estimates: Fitted Lowess Values of Accuracy and Granularity



Notes: The figure shows a lowess estimate that is similar to the right panel of Figure 4 showing the accuracy on the y-axis and granularity on the x-axis. The red line shows the lowess fitted value based on CES alone, and the blue line shows the lowess value based on the combined modeled payroll and CES estimates, using the out-of-sample errors. The accuracy is measured in the $-ln(MSE)$ as in the prior figure, and the x-axis is the inverse of the log employment of the bin. The figure shows that the estimates may be improved along much of the range, except the most disaggregate estimates (i.e., the right portion of the figure).

determine whether the estimated county-level series from Table 1 column (6) meets current standards, we estimate the current production possibility frontier for different levels of granularity. We then extrapolate the accuracy of county-level statistics based on the estimated relationship between accuracy and granularity, and test whether the newly minted statistics lie on or beyond the estimated frontier. We first include a figure that furnishes the intuition for this approach before we present the formal statistical test.

Just as Figure 4 above, 8 depicts MSE rates from CES estimates against varying levels of granularity, where the y-axis is the negative of the log of the MSE and the x-axis is the negative of the log of the level of granularity. The green points are revisions between official CES initial and final estimates for different geographic areas and industries (i.e., National, State, and MSA; and NAICS 3-digit, 2-digit, and all industries). The red line is fitted to the green points, capturing the relationship between accuracy and granularity. The blue points represent hypothetical revisions between the county QCEW and proposed county-level estimates that use the combined CES and payroll data. Intuitively, we consider the hypothesis that the blue dots lie, on average, along the red line. A rejection of this hypothesis for a collection of blue dots that lie on average above the blue line provides evidence that the proposed estimates are an improvement relative to the accuracy extrapolated from published official statistics, while a failure to reject would indicate estimates that are consistent with current standards[18]. Rejecting this hypothesis when the blue dots lie below the line would indicate that the new estimates perform worse than expected based on this standard.

This test is formalized in a relatively simple algorithm: We first regress accuracy (the negative log MSE) on granularity (the negative of log cell size), for only the official series. This establishes the baseline relationship, and from this, we extrapolate the anticipated accuracy levels of the new statistics (based on the main specification Table 1 column (6)) as a function of their granularities—that is, we compute the expected accuracy of the proposed county-level statistics based on this regression line. We then compute the residuals between the expected and actual accuracy for each of these points, performing a t-test on the difference between the groups of points above and below this line. As a comparison, in addition to the main specification that combines CES and payroll data, we also test this same relationship using CES for the corresponding county estimate (i.e., looking at the MSE if the corresponding CES is applied at the county level).

The estimated relationship between granularity and accuracy depends on the collection of statistics on which the relationship is estimated. Thus, we consider two approaches for establishing this expected relationship. In the first, we estimate this relationship making use of all official CES levels of aggregation. In the second, we use only MSA and state geographies and 2-digit and 3-digit NAICS industries. One may argue that the latter, more granular state and MSA statistics are more comparable to the county-level estimates.[19]

[18] Such a statistical test incorporates an estimate of the variance in accuracy conditioned on granularity, and we recognize that failure to reject a hypothesis can depend on this quantity in a relevant manner: Precisely estimated zeroes are distinct from failures to reject due to a large residual variance. In this case, however, it should be noted that a large residual variance does not imply that the newly constructed estimates are noisy, but rather, that there is a greater range of noisiness in the estimated statistics (the noisiness of the estimates, conversely, is reported on the vertical axis). Nevertheless, the fact of a relatively wide confidence interval (especially for a negative coefficient) may necessitate some judgment or further exploration from the statistical agency.

[19] Each collection of statistics that is published provides its own distinct estimate of the relationship between accuracy and granularity, as demonstrated in Figure A1 of the appendix. Because of this, the statistician might consider estimating the relationship from the set of statistics that is in some sense closest to the proposed; but in using a smaller set of observations, there is a risk of estimation error. We leave further exploration of this trade-off to future work.

The results of these t-tests are shown in Table 7. The rows show t-tests from two separate models. The first row represents the test associated with county estimates based exclusively on state and MSA CES, while the second row shows t-tests associated with the main specification that combines payroll and CES information. The two columns show t-tests from the two ways that we estimate the expected relationship between granularity and accuracy: The first column is associated with a relationship based on all available CES statistics, while the second column is associated with a relationship estimated from only the more disaggregated series.

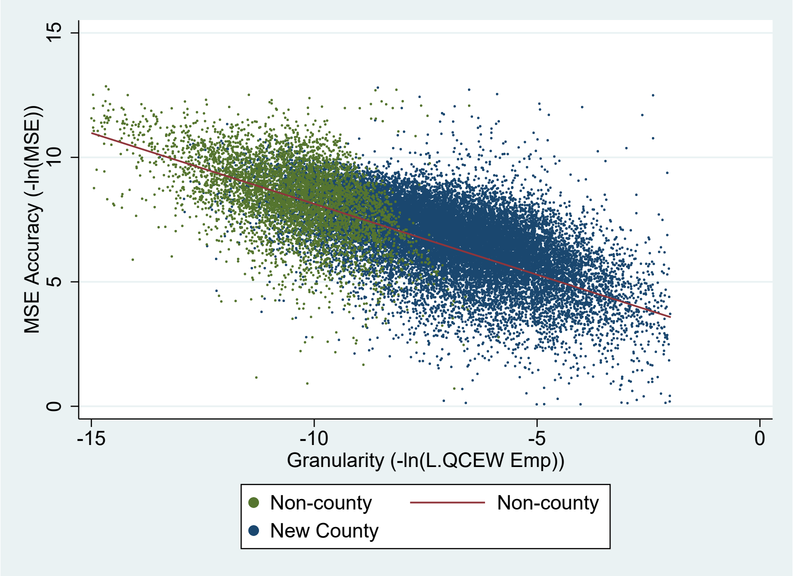
The t-test in the second row and first column of Table 7 reveals that our main specification has accuracy levels that are higher than the current frontier when compared to all official statistics, while the t-test in the second column suggests that the accuracy of our main statistics does not significantly differ from the frontier when compared to the more disaggregated MSA and state official statistics. We interpret row 2 as providing evidence that the level of accuracy we observe from our main specification is at least as good as current practice, conditional on granularity. In contrast, while the first column of the second row of the table suggests that CES does as well as should be expected, based on the relationship that is estimated from all levels of aggregation, the second column indicates that CES does significantly worse when compared to the accuracy estimated from the more granular official series. That is, based on the second column, CES at the state and MSA level predicts county-level growth with a lower accuracy than what is expected given the level of granularity.

Table 7: Comparing Existing Trends to Various County Specifications

	(1)	(2)
CES const. counties	-0.0358	-0.112***
	(0.102)	(0.000)
Main specification	0.0455**	-0.0303
	(0.033)	(0.166)

p-values in parentheses
The coefficient is the average residual for the row sample from a linear trend prediction that was estimated over existing statistics. The p-value is from a 2-sample t-test of means of the row sample's residuals against the residuals from the existing statistics. The first specification compares each row against the trend for all (non-county) existing estimates. The second specification limits existing estimates to just State or MSA estimates at either the 2- or 3-digit level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 8: Comparing Statistical Accuracy of Existing Statistics and New County Estimates



Notes: This figure shows a scatter plot of Mean Squared Error (MSE) estimates from CES initial and final estimates from different levels of granularity, similar to Figure 4. The y-axis is the negative of the log of the MSE and the x-axis is the negative of the log of the level of granularity. The green points are error rates based on official data sources for different geographic and industry granularities (i.e., National, State, MSA, and NAICS 3-digit, 2-digit, and aggregate). The red line is a fitted line based on the green dots, capturing the expected value of the MSE based on the level of granularity. The blue dots are the new county-level estimates using the combined CES and payroll data.

6.4 Exploring the Application of Machine Learning

The focus of the paper is primarily on linear models, as this somewhat reduces the scope of the analysis. In addition, we believe our supplementary payroll data source is likely to have a linear relationship with the underlying employment statistics, as they are similar measures. Linear models also have the advantage of being more interpretable than nonlinear models. At the same time, we recognize the potential of machine learning (ML) models to have tremendous predictive power. To highlight both the potential and distinct issues involved in ML analysis, we repeat our previous analysis at the state level, but apply ML tools.

The main results are reported in Appendix Table A2. Overall, we find that our main specification performs very similarly to a Lasso model, which is not surprising given that both Lasso and our main specification are linear.

We find very interesting and distinct results using a random forest model. A random forest prediction model works by combining the predictions of multiple decision trees, each trained on a random subset of the data, to make more accurate and robust predictions. One limitation of random forest predictions, is that the predictions are based on discrete values of historical growth rates. This shows up in the performance of the algorithm. Using the cross-fold cross validation which uses the entire sample, including the pandemic period, the random forest performs considerably better than the other algorithms. However, when our preferred rolling cross-validation is applied, we find the errors to be substantially higher, likely because the algorithm poorly fits the pandemic period, as no prior changes match the observed changes over the pandemic. This is less problematic for linear models, which implicitly assume linear relationships between the variables.[20]

These results highlight potential avenues for future research and demonstrate the predictive potential of machine learning, but they also provide a cautionary lesson of the limitations of nonlinear models.

7 Improving Timeliness with Payroll Data

One potential advantage of the payroll data is that it is more timely than the CES data. As discussed earlier, due to the larger sample size of CES, which covers 27 percent of employment, and the fact that it is based on a nationally representative survey, we find that CES contains less noise relative to the payroll data.[21] However, if the magnitude of the noise in the payroll data is relatively fixed, then the signal-to-noise ratio may rise during large economic fluctuations.[22] Using timely and high frequency spending data, Dunn et al. (2021), show that alternative and timely spending data increased in correlation with official statistics around the pandemic, consistent with the idea that the signal may increase relative to the noise during large shocks to the economy.

In this section, we use rich regional data from payroll data and CES by 2-digit industry

[20] We apply ML algorithms to county-level data in Table A3. We obtain results that are similar, in some ways, to when ML methods are applied at the state-level. For instance, the Lasso algorithm performs similar to our main specification. However, when the random forest algorithm is applied, we have a different result. The MAE rolling estimates show an improvement, perhaps because there is more county variation to build the model, but the MAE crossfold actually shows a substantial increase in error. As the $1 - R^2$ metric declines, it may be that the algorithm is reducing the L2 error more than the L1 error.

[21] https://www.bls.gov/web/empsit/cesfaq.htm

[22] For clarity, suppose the noise in the growth rate of employment is normally distributed mean 0 and S.D. 0.02, then when the actual growth rate is 0.02, the error will be proportionally larger than when the actual growth rate is 0.10.

at the state level, and examine how the ratio of the error to the actual change in employment changes with the magnitude of the aggregate economic fluctuation. We show this pattern using Figure 9. The y-axis is the error to signal ratio for each state-2-digit industry.[23] The x-axis is the national absolute change in the employment, which is shown on a log-scale. The red and blue dots capture the values based on CES and payroll data, respectively. The solid blue line is the lowess fitted value for the payroll data, while the red line is the lowess fitted value of the CES data. In general, the CES errors are lower than the payroll errors, as shown by the red line appearing below the blue line, consistent with our previous discussion. The figure shows that for large aggregate fluctuations, the payroll data comes close in value to the CES data as measured by the relative error rate. While payroll estimates perform slightly worse than the CES estimates, even for these large fluctuations, the payroll data is substantially more timely, offering an informative signal that is not currently provided by official statistics.

For large shocks in the economy, this trade-off is arguably reasonable given the greater timeliness of the payroll data, and the value of economic information during large shocks to the economy. The trade-off of timeliness and accuracy for large economic shocks, is similar to the trade-off from geographic granularity discussed earlier, where CES regularly reports MSA estimates with errors that are substantially larger than corresponding national- or state-level estimates.

This section shows that the more timely payroll data is arguably more useful when there are large fluctuations in the economy. We demonstrate this through an application in the following section.

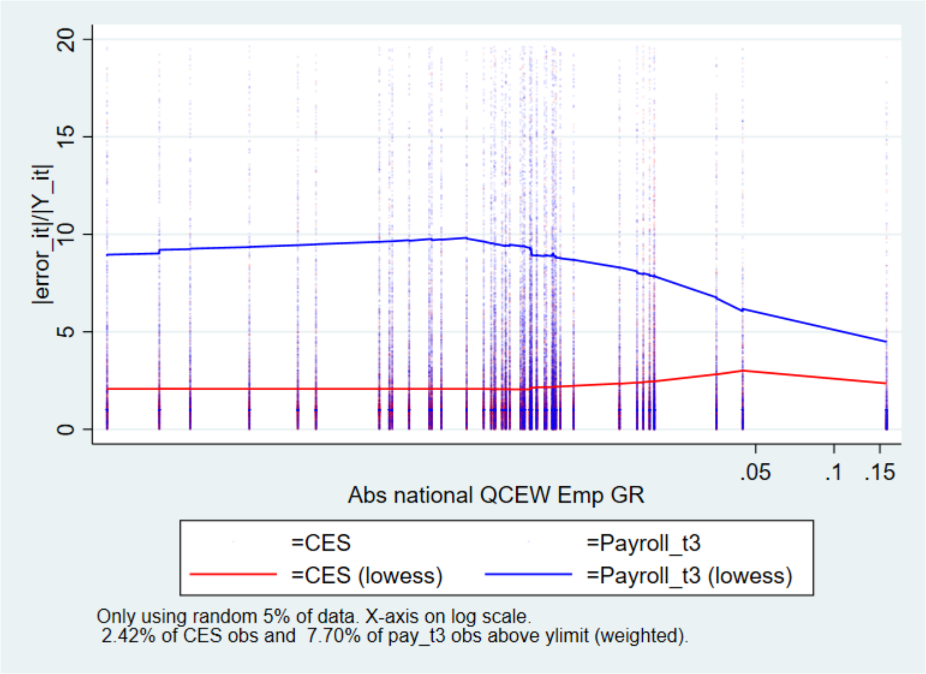
8 Application of Alternative Data to the COVID-19 Pandemic

In this section, we describe a plausible and practical application of the alternative payroll data. In this application, we highlight the potential value of the timely series that is based on the payroll continuing sample alone, which can be generated just one week after the reference period, as well as of a series that combines CES and payroll data.

A compelling use case for timely economic statistics is in fashioning rapid policy responses to

[23] The measure is similar to the MASE error concept, but the denominator is the actual economic change in the cell, rather than the mean absolute deviation.

Figure 9: Signal-to-Noise Ratio and Economic Shocks



Notes: The y-axis shows the ratio of the error, relative to the observed change in the growth rate. This is essentially a ratio of the noise to the observed signal. The x-axis shows the size of the aggregate economic shock. The blue line shows the lowess fitted values for the payroll data and the red line shows the lowess fitted value for CES. The figure demonstrates that for large economic shocks, the payroll data performs similar to the CES data.

large and unexpected (or unprecedented) economic events. A salient example of such an event is the recent COVID-19 pandemic and the economic fallout accompanying local policy responses such as shutdowns and stay-at-home orders. In this case, there may be significant benefit to low-latency, high resolution economic statistics, to rapidly pinpoint the areas that sustained the the most intense economic repercussions.

While policymakers were undeniably punctual in assembling a package of policy responses to the COVID-19 pandemic, these were generally not targeted to specific geographical areas where hard-hit industries were struggling to cope with the aftermath of the event. In many cases, it has been argued that the policies suffered from higher than expected costs relative to the measurable benefits.[24] In the example below, we show that using payroll data could have yielded timely information on the hardest hit counties. Yet it should be noted that our goal here is not to comment on any past policies (which were made based on information available at the time) nor to suggest how statistics that we introduce should be used. Rather, our goal is to provide a look at the way such timely statistics might be used as a starting point for crafting a more economically efficient response to unexpected events in the future.

The two key dates for this application are the March 13th initial declaration of a state of emergency, and the March 12th reference date for CES, which falls one day prior to the declaration. Importantly, the downturn in employment did not occur until mid to late March. As such, more than two months had elapsed by the time state-level and MSA-level CES data were available that covered the immediate aftermath of the emergency declaration. Thus, to have possessed actionable labor market statistics immediately following March 13 would potentially have been helpful throughout the pandemic response. Another key point is that numerous policies were made for several months after the pandemic, so not only would more timely information have been helpful, but also more accurate information as the CES estimates became available. An abbreviated timeline of the COVID-19 Pandemic declaration, its economic effects, and subsequent policy responses is recorded in the Appendix.

To better gauge the potential value of the payroll data series, we examine how various

[24] One policy response to the pandemic was Paycheck Protection Program (PPP), an initiative that provided forgivable loans to small businesses and nonprofits to help them retain employees and cover essential expenses. The program aimed to mitigate economic hardship and maintain business operations during the pandemic. However, some researchers have argued that the program was poorly targeted, costing between about $150k to over $300k per job saved according to Autor et al. (2022) and Chetty et al. (2023). At the same time, the funds for PPP were dispersed very quickly, as highlighted in the timeline in the Appendix, which lays out both the timing of the COVID-19 pandemic, the PPP policy as well as the timeline for employment statistics and when policymakers would receive critical employment information.

estimates can help in identifying the hardest hit counties and industries around the pandemic. To do this, we conduct a couple of distinct exercises. First, for the month of April 2020, when 98.2 percent of county-industries experienced job loss, we identify the worst hit counties and 2-digit industries. More precisely, we rank the county-industry pairs by employment loss, and see how many county-industries would need to be selected to account for 50% of the actual loss nationally, where each county is weighted by QCEW employment in March. If the county-industries were selected randomly, we would expect that 50% of the counties would be needed to capture 50% of the employment loss.

The results of this first exercise are shown in Table 8. The first line shows the baseline of random allocation, which requires identifying 50% of the county-industry cells to identify 50% of the employment loss. The last row "QCEW (Ideal Allocation)" uses the QCEW data as the "truth" to define the employment loss. Using the QCEW, we find that 18.5% of the county-industry combinations must be identified to capture 50% of the employment loss. However, as mentioned previously, the QCEW data is based on administrative data that is only available with a lag of about 5 months, which is far too slow for policy purposes. Between these two extremes the rows of the estimates are ordered by level of timeliness, from payroll t1, which is the most timely estimates based solely on payroll data one week after the reference period, to the CES and Main estimates that have the same latency.

As mentioned before, the payroll timeliness is customizable and it performs relatively better during large shocks to the economy. Below our baseline random allocation benchmark, the next three rows of Table 8 show performance of the payroll data one-to-three weeks after the reference period (i.e., one week - payroll t1, two weeks - payroll t2, and three weeks - payroll t3). These latencies would have been sufficiently low for policymakers to consider this information prior to the second round of PPP funding, for example. These estimates perform far better than random assignment for each series, only needing to identify around 26-29% of the county-industry cells to capture 50% of the employment loss. Improved allocation in where resources are spent could lead to substantial improvement in jobs saved per dollar spent. However, it is also important to note that the timely payroll series does not perform as well as CES and the main estimate (which combines CES and payroll), as shown in the next two rows.[25]

The CES performs only slightly worse than the QCEW and substantially better than random

[25] Recall that this difference is likely due to the especially large sample size of CES, which covers 27% of employment.

allocation, needing to select around 20.1% of county-employment cells to identify 50% of the employment loss. Part of the exceptional performance of the CES data over this period was related to the careful analysis by staff, who made adjustments to the standard firm birth-death model around the pandemic.[26] In particular, staff took into consideration and empirically modeled the fact that many of the business exits and entry during the pandemic may be important to incorporate. Indeed, this type of adjustment is not feasible with the payroll data, and may be reflected in the slightly worse performance of the combined payroll and CES data relative to the CES estimate. The combined payroll and CES modeled estimate needs to select 20.6% of the county-industry cells to identify 50% of the loss (Table 6 and main estimate row (4)).

Table 8: Select County and 2-Digit Industry to Capture 50% of the Employment Loss

Specifications	2020-04
Baseline Random Allocation	0.500
=Payroll t1	0.286
=Payroll t2	0.265
=Payroll t3	0.261
=CES	0.201
Main	0.206
QCEW (Ideal Allocation)	0.185
total losses (000s)	-17,101
Share of counties w/ losses	0.982

Notes: The table shows how many county-2-digit industry categories would need to be identified to capture 50% of employment loss as of April of 2020, where county-2-digit industries are weighted by lagged employment from QCEW. The first row shows the result based on random guesses, which is 50% by construction. The following remaining rows are shown in order of timeliness, showing the payroll series, followed by CES and main, and then QCEW, which is the ideal allocation. The ideal allocation based on the QCEW, which emplies 18.5% of counties would need to be identified to capture 50% of the employment loss. The fifth and sixth rows show CES and the main estimates, where the main estimates combine the payroll and CES data (column (6) of Table 6). The performance is only slightly worse than the ideal for both estimates. The next three rows show estimates based on payroll data, with different timing (1, 2 or 3 weeks after the reference period). For example, one week after the reference period the payroll data would need to identify 28.6% of counties to capture 50% of employment loss.

After the initial downturn in April 2020, it was important for policymakers to monitor the recovery and determine what areas of the country may be lagging in employment gains. Again, accurate information may affect where resources are allocated. To capture this notion, we first use the QCEW data to perfectly identify the 25 percent of counties where the employment

[26] https://www..gov/covid19/effects-of-covid-19-pandemic-and-response-on-the-employment-situation-news-release.htm

increase was the lowest. We then identify what share of the worst performing counties are captured using the different estimates for the first 6 months of the pandemic.

The results are reported in Table 9. As a baseline, it is important to note that if county-industries are identified randomly, then we would only capture 25 percent of the areas with the lowest employment growth, as shown on the top row of Table 9 for all months. At the other extreme, using QCEW as the benchmark, we see that it perfectly identifies the 25% of counties with the lowest employment growth, as indicated by the 1 along the bottom of the table for all months.

As in the prior table, the rows are listed in order of timeliness. The fifth and sixth rows track the performance of the CES and the Main specification in identifying the lowest 25 percent of counties as measured by employment growth. For the first month of the pandemic in April, the CES performs the best, identifying 83.2% of the counties, although the main specification performs nearly as well, identifying 82.7% of the counties. After the initial month, the main specification consistently performs better than the CES estimate, by 2-3 percentage points each month. While the percentage point difference is arguably small in a particular month, the improvement in the allocation over the entire time period adds up to around 13 percentage points, a sizeable difference. This suggests that use of the combined CES and payroll data may lead to measurably improved outcomes in the long run.

9 Conclusion

Our research has explored the potential benefits and challenges of incorporating big data sources into economic statistics, with an application targeting regional employment statistics. We have designed a structure that weighs the advantages of increased timeliness and granularity against potential bias and noise in these estimates, using revisions observed in official measures as a benchmark. Through our analysis, we conclude that the integration of data from a major payroll processor can significantly enhance the accuracy of state-level employment by industry statistics and can enable the production of county-level estimates. We also provide evidence that the timely payroll data provides useful information on its own, especially in a setting of significant economic variation.

We illustrate the value of these alternative data sources and estimates through an application motivated by the policies implemented during the pandemic. In this counterfactual analysis, we

Table 9: Identify 25% of County and 2-digit Industries with the Lowest Increase In Employment

Specifications	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09
Baseline Random Allocation	0.25	0.25	0.25	0.25	0.25	0.25
=Payroll t1	0.587	0.437	0.379	0.325	0.320	0.332
=Payroll t2	0.614	0.430	0.397	0.366	0.345	0.338
=Payroll t3	0.619	0.430	0.407	0.358	0.336	0.351
=CES	0.832	0.510	0.552	0.412	0.389	0.397
Main	0.827	0.541	0.581	0.438	0.424	0.415
QCEW	1	1	1	1	1	1

Notes: This table preforms a counterfactual exercise of determining how well the weakest performing counties may be identified across the different series. Specifically, we see how well the algorithms correctly identify the 25% of county-2-digit industry pairs. The first row shows the random baseline allocation. If county-industries are selected randomly, then on average only 25% of the worst performing counties will be identified. The second row shows the ideal based on QCEW, where 100% of the counties will be identified. The next three columns identify the worst performing areas using the timely payroll series alone. The fourth and fifth rows show the estimates using CES and the main estimates combining CES and payroll estimates (column (6) of Table 6). The main estimates perform better than CES alone, except for the first month, where CES performs slighly better. The last row is the "ideal" estimate using QCEW, which by definition perfectly identifies the worst hit counties.

find that the use of alternative data could assist in both timelier and more accurate targeting of employers that were hit hardest by the pandemic, providing a prototype for the deployment of alternative data sources and how they may contribute to effective policy decisions.

We illustrate the feasibility of using alternative data to expand the production frontier of statistical agencies, offering valuable insights for policymakers and the public. However, the effective incorporation of such data sources requires careful consideration along each part of the statistical assembly line: From constructing and adjusting the input series, to selecting and evaluating models (which potentially combines alternative and traditional survey data sources), to testing and evaluating results. Our study provides a demonstration of one way to thoughtfully and rigorously execute such a program, with a particular focus on providing appropriate context for statistical evaluation.

As the stream of available data sources has swollen to a torrent, navigating among the range of potential applications has become an increasingly daunting task. We journey beyond the basic question of whether such data appears to offer some value, instead exploring and more formally testing whether the data is valuable, where and when it is valuable, and how much value the data adds to improving official statistics. This systematic approach simultaneously looks back at the accuracy of historical series to provide grounding for new or improved estimates while it looks forward in extrapolating this information to assess the incremental value of potential improvements. These efforts help chart a way toward improved statistics, and thus, toward more informed decision making.

References

Abraham, K. G. (2022). Big data and official statistics. *Review of Income and Wealth*.

Aladangady, A., S. Aron-Dine, W. Dunn, L. Feiveson, P. Lengermann, and C. Sahm (2019). From transactions data to economic statistics: Constructing real-time, high-frequency, geographic measures of consumer spending. Technical report, National Bureau of Economic Research.

Autor, D., D. Cho, L. D. Crane, M. Goldar, B. Lutz, J. Montes, W. B. Peterman, D. Ratner, D. Villar, and A. Yildirmaz (2022). An evaluation of the Paycheck Protection Program using administrative payroll microdata. *Journal of Public Economics 211*, 104664.

Cajner, T., L. D. Crane, R. A. Decker, J. Grigsby, A. Hamins-Puertolas, E. Hurst, C. Kurz, and A. Yildirmaz (2020). The US labor market during the beginning of the pandemic recession. Technical report, National Bureau of Economic Research.

Chen, J. C., A. Dunn, K. Hood, A. Driessen, and A. Batch (2019, July). *Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators*. University of Chicago Press.

Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, et al. (2020). The economic impacts of COVID-19: Evidence from a new public database built using private sector data. Technical report, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, M. Stepner, and the Opportunity Insights Team (2023, October). The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. *The Quarterly Journal of Economics*.

Cox, N., P. Ganong, P. Noel, J. Vavra, A. Wong, D. Farrell, F. Greig, and E. Deadman (2020). Initial impacts of the pandemic on consumer behavior: Evidence from linked income, spending, and savings data. *Brookings Papers on Economic Activity 2020*(2), 35–82.

Dunn, A., K. Hood, A. Batch, and A. Driessen (2021). Measuring consumer spending using card transaction data: Lessons from the COVID-19 pandemic. In *AEA Papers and Proceedings*, Volume 111, pp. 321–25.

Dunn, A., E. Liebman, and A. H. Shapiro (2014, September). *Developing a Framework for Decomposing Medical-Care Expenditure Growth: Exploring Issues of Representativeness*, pp. 545–574. University of Chicago Press.

Manski, C. F. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after morgenstern. *Journal of Economic Literature 53*(3), 631–653.

Moyer, B. C. and A. Dunn (2020). Measuring the gross domestic product (gdp): The ultimate data science project. *Harvard Data Science Review 2*(1), 1–9.

Appendix

.1 Description of the Payroll Data and Series Construction

The analytics group of the private payroll company maintains several analytical databases through nightly queries of their production databases (i.e., the databases that support the various payroll systems). The analytical databases thus provide near real-time data (i.e., there is at most a 24-hour lag between when a client (i.e., employer) schedules a payment and when the payment is saved to the analytical databases). The analytical databases also provide historic data via monthly archiving.

The analytical databases contain attributes of the payroll's clients and their employees, as well as summary and detailed information on the payments employers make to their employees. Payment data are temporal, and thus are organized by the year and month the payments were processed.

These data are proprietary and are hence siloed inside the payroll company's computing environment. Only after extensive processing to aggregate and anonymize the data so that employer and employee personal information are protected are the transformed "cellular" data transferred out of the companies computing environment. At this point, the cellular data are comingled with public statistics as described in this paper.

In general, processing is similar to some CES methods: `https://www.BLS.gov/opub/hom/ces/`.

- First, only a subset of payments that employers make to their employees are considered. These CES-compatible payments are: regular pay, both salaried and hourly; overtime; holiday and vacation pay; sick leave; tips; piecework and commissions; and supplemental pay such as bonuses and severance, so long as the payments are regular in both amount and schedule.

- Secondly, following CES, processing is monthly and only payments made for a pay period that includes the 12^{th} of each month are considered.

- Finally, processing follows CES matched sample methodology. A employer (and all of its current employees) is only included in processing if the employer existed in both the current and prior months.

Because only payments centered around the reference date of the 12^{th} of the month are considered, an important consequence is that multiple months before and after the reference month must be accessed. This is due to the fact that payments can be made outside of the reference month (e.g., prepayments, corrections, and payments made by employers that pay by calendar month).

Specifically, employment and wage processing are based on payments stored in five months of analytical wage databases: the reference month as well as the two prior and subsequent (if these exist) months. For example, for the payments made in the five months centered on the payroll reference date of 6/12/2022, practically all payments a (97.76 + 1.88 = 99.64%) are made during the reference month (m) (97.76%) and in the month following the reference month (m+1) (1.88%). However, a non-trivial percentage of payments (0.36%) are made in months m-2, m-1, and m+2. (Not shown are payments made outside of this 5-month window, which are undoubtedly a very small but non-trivial percent).

More detailed processing steps are given below. Note that while this paper only discusses employment, future endeavors may also examine payments and wages. In addition, the concept of employment and payment for work are closely tied, as we define employment as a period of time over which an employee received payment. The primary benchmark for wages would also be QCEW, but the wage data in the QCEW is recorded quarterly, rather than monthly, so the process of validating the estimates would be different. Both employment and wage processing are discussed in this section.:

1. Filtering the payment data:

    - Filter payments such that only employers and their employees that are active as of the payroll reference date for the particular year and month (i.e., as of the 12^{th} of the month) are included.
    - Further filter these employers such that only continuing units (i.e., those employers that were also active as of the prior monthly payroll reference date) are included.
    - Filter for employers whose payment frequency is weekly, bi-weekly, semi-monthly, or monthly only.

2. Simple corrections and/or transformations:

- If any two of payment amount, hours, or rate are present, but the third is missing, impute using the formula: $rate = amount/hours$.
- If either the payment amount or hours is negative but the other is not, set both to be negative
- If payment frequency is missing, flag the payment as needing imputation. Otherwise, use payment frequency to convert amount and hours to weekly units by scaling amount and hours by the following factors:
  - Monthly: $\frac{12}{52}$
  - Semi-monthly: $\frac{24}{52}$
  - Bi-weekly: $\frac{26}{52}$

3. Added variables: Determine the week relative to the payroll reference date they payment was processed. This variable takes on the values:

   - 1: if scheduled in the first week (i.e., from the 12th - 18th of the month),
   - 2: if scheduled in the second week following the 12th,
   - 3: if scheduled in the third week following the 12th, and
   - 4+: if the payment was scheduled 4 or more weeks after the 12th of the month.

4. Payment corrections: Corrections occur when errors in payments are subsequently corrected. Corrections are legitimate payments and should be included in processing assuming both the initial error and correction(s) can be paired (otherwise they may need to be imputed). Unfortunately, being able to identify the error and subsequent correction(s) is likely a rare occurrence. This is due to many factors, including:

   - The timing of the correction (does it occur during the same pay period or in a different one?)
   - The type of correction (was the initial error an underpayment or overpayment?)
   - How the corrections are made (is it a single correction or multiple repeated partial corrections?)

   Corrections that occur within the same pay period as the initial error are the easiest to deal with. For example, a employer may initially enter an incorrect amount for a particular payment and zero this out by entering the negative of the amount. “Zero-sum” corrections such as these are flagged but otherwise left alone.

Corrections that occur in a different pay period are far more difficult to deal with. This is due in part to the way payments are stored in the analytical databases (i.e., by the year and month the payments are scheduled). If we only looked at the payments made in the reference period month, there is a good chance either the initial error or the correction would be missing. To alleviate this, we look for payments that occur during the reference period by looking in the reference period month as well as in the 2 prior months and the 2 subsequent months (if these occur).

Despite looking at multiple months, it is not unusual that the full payment history (i.e., both the error and the correction) is incomplete. This will occur if the corrections (or original payments) occur outside of the 5-month window considered during processing. Even if the error and correction are made within this window, the employer must enter the same pay period start and end dates for both the error and correction in order for both to be extracted.

Errors can be due either to an initial underpayment or overpayment. The latter case is more readily identified if the resulting correction is a single negative payment. In this case, the presence of negative correction is flagged during processing but otherwise is not treated.

If rather than a single negative payment, the correction of an initial overpayment consists of garnished future payments (i.e., payments that are reduced by a partial repeated amount), identifying the initial error and subsequent corrections is an altogether more complicated problem. This is also the case if the initial error is an underpayment, regardless of whether the correction consists of single payment or multiple repeated partial payments. In these situations, corrections are better considered as contributing to anomalous wages, and treated in the same manner as anomalous wages are.

5. Imputation: Negative payments are flagged. If both the initial error and correction can be identified, and the resulting amount is not negative, these are included when determining wages. Otherwise, these payments are imputed.

6. The final processing step is to aggregate the processed employment and wage microdata to

safeguard payroll client (i.e., employer) privacy. The standardized weekly payments are summed for each employer/employee, giving the weekly wage. Employment is a binary variable equal to 1 if any payments were made to the employee for a payroll that includes the monthly payroll reference date. Finally, the employer/employee wages and employment are summed to the cellular level. The cells are defined by the following employer and employee characteristics:

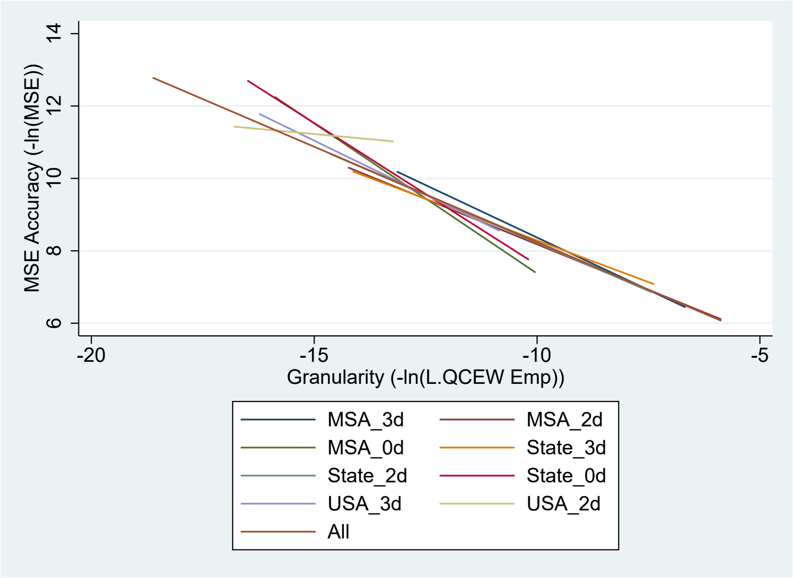
- 2-digit NAICS
- current and prior month client size class
- county
- current and prior month payment week

.2 COVID Timeline

- March 12, 2020 - CES reference date for March employment
- March 13, 2020 - The President of the United States declares a national emergency.
- March 15, 2020 - Many states begin to implement shut-downs and stay-at-home orders.
- March 27, 2020 - CARES Act signed into law (individual payments and PPP loans).
- April 24, 2020 - Some states begin to reopen/roll back pandemic policies. A second round of funding was signed into on law and disbursements made from April 27-August 8, 2020.
- May 9, 2020 - Official national unemployment rate in the U.S. hits 14.7%.
- May 18, 2020 - HHS awards states, territories and local jurisdictions funding from CARES Act for COVID-19 response (mostly testing).
- May 22, 2020 - CES state and metro estimates for April revealing the depth of pandemic loss for regions.
- July-August 2020 - Beginning of deals between the U.S. Government and vaccine providers for production and distribution.
- September 23, 2020 - HHS and CDC awards states, territories and local jurisdictions funding from CARES Act to support vaccine distribution.
- December 27, 2020 - A third round of funding for CARES act was signed and disbursements made from January - May 2021.
- January 6, 2021 - HHS and CDC announces $22 billion in funding for states, territories and local jurisdictions to support vaccine distribution.

.3 Additional Analysis

Figure A1: Granularity Trends in Accuracy Across Different Aggregation Levels for Existing Estimates



Notes: This figure shows fitted lines of Mean Squared Error (MSE) estimates from CES initial and final estimates from different levels of granularity. The y-axis is the negative of the log of the MSE and the x-axis is the negative of the log of the level of granularity. Each line is a fitted value based on the different series. For example, MSA 3-digit NAICS, State 2-digit NAICS or USA 3-digit. The estimates show the different relationships between accuracy and granularity across the different series.

Table A1: Regression of State-Level Accuracy on Employment Cell Characteristics (Blended Payroll/CES Accuracy Minus CES Accuracy)

	(1)	(2)	(3)
Payroll Coverage	0.00816***	-0.00506	0.00181
	(2.65)	(-1.31)	(0.49)
>500k Emp	-0.000449*		0.000828
	(-1.92)		(1.55)
>250k Emp & $\leq$ 500k	0.0000178		0.00133***
	(0.10)		(2.51)
>100k Emp & $\leq$ 250k	0.00113***		0.00224***
	(5.95)		(4.36)
>50k Emp & $\leq$ 100k	0.00214***		0.00325***
	(8.68)		(5.91)
>10k Emp & $\leq$ 50k	0.00254***		0.00368***
	(9.48)		(6.92)
>5k Emp & $\leq$ 10k	0.00392***		0.00540***
	(5.65)		(6.66)
>1k Emp & $\leq$ 5k	0.000581		0.00228***
	(0.68)		(2.05)
$\leq$ 1k	-0.00339***		-0.00137***
	(-442.16)		(-2.60)
Constant		0.00288***	
		(6.67)	
Observations	38,767	38,767	38,767
R^2	0.017	0.013	0.026
Adjusted R^2	0.017	0.013	0.025
NAICS 2d FEs		X	X

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The accuracy is measured in the negative of the absolute error. The dependant variable is the accuracy of the blended estimate (Payroll/CES) minus the CES accuracy. On the right hand side of the regression, we include different covariates, including dummy variables of the cell size of the employment estimate range (e.g., 10K-50K or 50K-100K employees), a measure of the payroll coverage for the cell, along with 2-digit industry fixed effects. Column 1 includes coverage and industry cell size; column 2 includes coverage and industry fixed effects; and column 3 includes cell sizes and industry fixed effects.

Table A2: Exploring State-Level Estimates using CES + Payroll and Machine Learning Estimation

	(1)	(2)	(3)	(4)
	cnsreg	reg	Lasso	RF
CES Emp Gr	1	0.607***	0.608	23.79
	(.)	(217.93)		
Payroll Emp Gr (cont; t=3)		0.140***	0.0790	12.54
		(15.06)		
Payroll (cont; t3) Coverage x Emp Gr		2.008***	1.955	7.677
		(27.72)		
Payroll Emp Gr (st-Agg; t3)		0.0369***	0.0363	8.512
		(10.52)		
Constant		-0.000173	0.0000269	
		(-0.30)		
Observations	41,924	41,924	41,924	41,924
$1 - R^2$	0.179	0.135	0.135	
$1 - R^2$ unweighted	0.309	0.249	0.253	
$1 - R^2$ (excl. 2020-03 to -05)	0.347	0.306	0.308	
$1 - R^2$ OOS (rolling)	0.177	0.154	0.153	0.393
$1 - R^2$ OOS (cf)	0.179	0.136	0.137	0.105
MAE	0.00853	0.00731	0.00731	
MAE OOS (cf)	0.00853	0.00733	0.00734	0.00641
MAE OOS (rolling)	0.00876	0.00777	0.00776	0.0122

t statistics in parentheses

Outcome is QCEW Emp Gr (Mean Abs Dev=0.014). Observations at the State-NAICS2-Month level.

Dates=[2017,2021-06]. Variables are NSA. Analytic weights are lagged QCEW 2-digit empl.

Payroll Coverage is calculated as lagged Payroll employment divided by lagged QCEW employment.

Coverage included when coverage is interacted with growth.

Restricted to observations with valid CES and Payroll Gr. Coverage included in RF.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports state-level by two-digit industry county estimates, but applying standard ML tools to the analysis. The first two columns show our baseline models, where column (1) shows predictions based on the unmodeled CES estimates. Column (2) shows the main estimates that combine payroll and CES data. Column (3) applies a Lasso estimation method. Column (4) applies the random forest ML algorithm and includes “feature importance” metrics rather than coefficients.

Table A3: Exploring County-Level Estimates using CES + Payroll and Machine Learning Estimation

	(1) cnsreg	(2) reg	(3) Lasso	(4) RF
CES Emp GR (MSA or State)	1	0.667^{***}	0.664	521.0
	(.)	(485.81)		
CES State Emp GR * 1(non-MSAs)		-0.244^{***}	-0.239	170.6
		(-145.91)		
Payroll Emp Gr (cnty-naics2; t3)		0.00749^{***}	0.0110	155.2
		(3.77)		
Payroll Emp Gr (t3) * 1(non-MSAs)		0.0318^{***}	0.0120	89.42
		(9.31)		
Payroll Emp Gr (Cnty-naics2) x Payroll Coverage (naics2) (t3)		0.293^{***}	0.294	
		(48.83)		
Payroll coverage vs QCEW (Cnty-naics2, 2020-01, t3)		-0.00248^{*}	0.00604	88.00
		(-2.17)		
Payroll Emp Gr (cnty-Agg; t3)		0.0897^{***}	0.0899	212.0
		(104.94)		
Payroll Emp Gr (state-naics2; t3)		0.111^{***}	0.111	386.4
		(133.09)		
Constant		0.000109	-0.000825	
		(0.02)		
Observations	898,139	898,139	898,139	898,139
$1 - R^2$	0.501	0.423	0.426	0.129
$1 - R^2$ (rolling)	0.497	0.484	0.489	0.475
$1 - R^2$ (CrossFold)	0.501	0.431	0.429	0.429
MAE	0.0157	0.0136	0.0137	0.00384
MAE (rolling)	0.0159	0.0144	0.0146	0.0141
MAE (CrossFold)	0.0157	0.0136	0.0136	0.0160

t statistics in parentheses

Outcome is QCEW Emp Gr (Mean Abs Dev=0.018). Observations at the County-NAICS2-Month level.

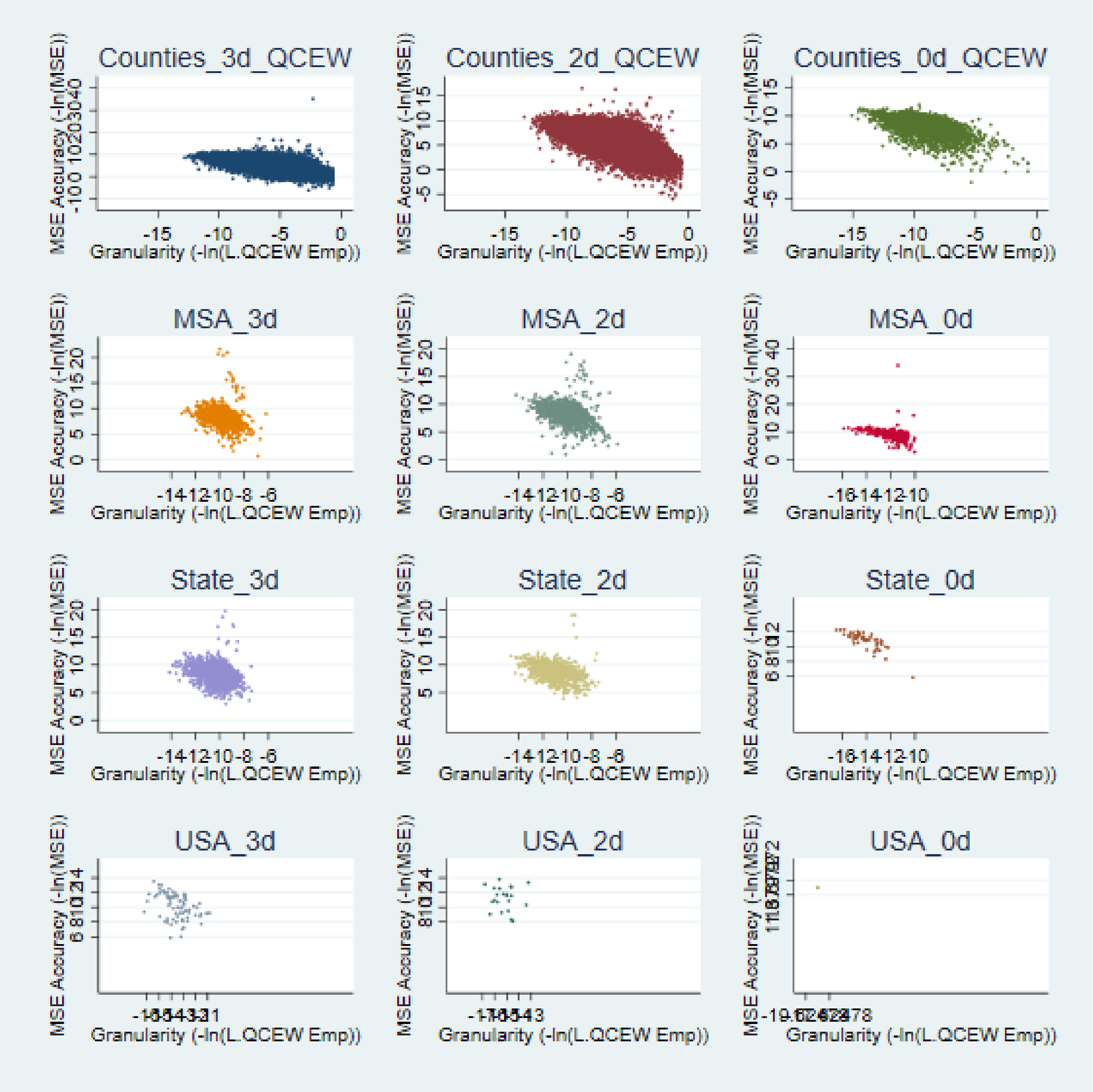
Full dates=[2017,2021-06]. Variables are NSA. Analytic weights are lagged QCEW 2-digit employment

Restricted to observations with valid CES employment growth.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports county-level by two-digit industry county estimates, but applying standard ML tools to the analysis. The first two columns show our baseline models, where column (1) shows predictions based on the unmodeled CES estimates. Column (2) shows the main estimates that combine payroll and CES data. Column (3) applies a Lasso estimation method. Column (4) applies the random forest ML algorithm and includes "feature importance" metrics rather than coefficients..

Figure A2: PPF
CES initial vs CES final (USA, State, MSA) and QCEW (County)



Notes: The table shows the scatterplot shown previously in Figure 4, but one distribution at a time.