Improving the Accuracy of Economic Measurement with Multiple Data Sources: The Case of Payroll Employment Data*

Tomaz Cajner
Federal Reserve Board
tomaz.cajner@frb.gov

Leland Crane
Federal Reserve Board
leland.d.crane@frb.gov

Ryan A. Decker
Federal Reserve Board
ryan.a.decker@frb.gov

Adrian Hamins-Puertolas
Federal Reserve Board
adrian.b.hamins-puertolas@frb.gov

Christopher Kurz
Federal Reserve Board
christopher.j.kurz@frb.gov

March 19, 2018
Extended Abstract for NBER/CRIW
Big Data for 21st Century Economic Statistics

Abstract

This paper seeks to efficiently combine two data sources on U.S. payroll employment in order to obtain the most accurate real-time measure of employment: the CES payroll employment series and a payroll employment series produced by Cajner, Crane, Decker, Hamins-Puertolas, Kurz, and Radler (2018) based on microdata provided by the payroll processor ADP. The two time series are derived from roughly equally-sized and roughly non-overlapping samples. This paper briefly overviews the ADP payroll microdata, reviews issues related to weighting for representativeness, and presents the results of an exercise that benchmarks the ADP data to an employment census (the QCEW). We argue that by combining CES and ADP data we can reduce the measurement error inherent in both data sources. In particular, we infer "true" unobserved payroll employment growth using a state-space model and find that the optimal predictor of the unobserved state puts approximately equal weight on the CES and ADP series.

Keywords: economic measurement, big data, state-space models.
JEL Classification: J2, J11, C53, C55, C81.

*The views in this paper are those of the authors and do not necessarily represent the views or policies of the Board of Governors of the Federal Reserve System or its staff.

1 Introduction

The ongoing data revolution is dramatically increasing the availability of information that can be used for economic research and measurement. Economists are still adapting to the new era of data abundance by exploring how newly accessible information, typically produced by private companies, can be most fruitfully used in empirical work and by improving econometric and statistical tools to analyze "big data" (Einav and Levin, 2014). One particularly promising area, where "big data" are expected to have a major positive impact in the future, is measurement of economic activity and creation of new aggregate economic statistics. While still in infancy, this line of inquiry has already produced some encouraging results, both from academic economists—e.g., the Billion Prices Project at MIT (Cavallo and Rigobon, 2016)—and from private companies—e.g., the payroll service provider ADP publishes monthly its National Employment Report (NER), MasterCard produces retail sales estimates, and Zillow calculates house price indices at the county level.

Cajner, Crane, Decker, Hamins-Puertolas, Kurz, and Radler (2018) describe the construction of weekly and monthly aggregate employment series based on ADP's weekly payroll microdata. Their aggregate series are designed to be an independent signal about labor market conditions (rather than an attempt to forecast monthly BLS employment figures). In the present paper we describe additional steps to improve our weighting methodology and benchmark the ADP-based series to the QCEW on an ***annual*** basis. For weighting, we experiment with trimming large weights, collapsing cells, and using Bayesian approaches. The benchmarking procedure is broadly similar to CES benchmarking and ensures that year-to-year changes are governed by the QCEW, while higher-frequency changes, and the period after the most recent benchmark, are mostly a function of the ADP data. Our belief is that while CES and ADP can be profitably combined to better measure high frequency and real-time employment changes, the QCEW is the "final word" for annual growth. This further illustrates the essential role that government statistics play even when there is significant value in nontraditional data sources.

More fundamentally, existing work using nontraditional data sources for economic measurement typically takes official government data as a benchmark for all frequencies. For example, the monthly NER series published by ADP are constructed with the goal of predicting the (fully-revised) BLS Current Employment Statistics (CES) data. In this paper we take a different approach by recognizing that CES payroll employment is subject to non-negligible measurement error (as are the estimates derived from the ADP data). As a result, our baseline model assumes that true U.S. employment growth follows an AR(1), and that both the CES and ADP employment estimates are

noisy signals of this underlying process. Then standard state-space tools allow us to estimate the latent process and the observation error associated with each series. It turns out that the optimal predictor of the unobserved state puts approximately equal weight on the CES and ADP series. This is not necessarily surprising, as the ADP sample covers roughly the same fraction of private nonfarm U.S. employment as the CES sample (about 20 percent).

In what follows, we first briefly describe the characteristics of ADP microdata and how we construct aggregate ADP employment indexes. We then discuss the uses and limits of ADP data for forecasting official employment numbers before turning to our preferred use of the ADP data as an independent signal that, coupled with official sources, can inform us about the "true" state of the economy. Throughout, we focus on the role of these privately generated data as a complement to existing official statistics. While there is no substitute for official statistics in terms of consistency, transparency, and scientific collection methods, official numbers do have limitations that alternative data sources can address.

2 Data

The main sources of timely employment data in the U.S. are CES and QCEW, both published by the BLS. CES, which is published a few days after each reference month, is based on a stratified-sample survey of about 500,000 private establishments covering 23 percent of all U.S. private employees.[1] QCEW, which is published quarterly with a lag of two quarters, is based on administrative data for unemployment insurance tax records and covers nearly the universe of (UI-covered) employment; while QCEW is not subject to traditional sampling error, it is subject to various sources of non-sampling error.[2] We now turn to a discussion of the ADP data.

2.1 Structure and Representativeness of the ADP Microdata

The raw data we use are weekly, establishment-level employment counts derived from ADP payroll records. ADP provides human capital management services, including payroll processing. Processing payroll for a client firm involves many tasks including maintaining worker records, calculating taxes, and issuing paychecks. The structure

[1] See https://www.bls.gov/web/empsit/cestn.htm. Note that the CES contains data for total nonfarm payroll employment, but here we focus only on private payroll employment, excluding government employment to be consistent with the reliable scope of ADP.

[2] For a detailed analysis of measurement challenges in CES and QCEW, see Groen (2012).

of the microdata is determined by the business needs of ADP. ADP maintains records at the level of Payroll Account Controls (PAC), which often correspond to business establishments (but may sometimes correspond to firms) as defined by the Census Bureau and the BLS. Each PAC generates a new record at the end of each pay period. The record consists of the date payroll was processed, employment information for the pay period, and many time-invariant PAC characteristics (such as an anonymized PAC identifier, NAICS industry code, zip code, etc.). PAC records include both the number of individuals employed (active employees) and the number of individuals issued a paycheck in a given pay period (paid employees). Active employees include wage earners with no hours in the pay period, workers on unpaid leave, and the like. Paid employees include any wage or salary workers issued regular paychecks during the pay period as well as those issued bonus checks and payroll corrections. Data available to us begin in July 1999.[3]

There are several important differences between ADP payroll microdata and the QCEW and CES data in terms of pay frequency, region, establishment size, and industry composition that should be noted.[4]

- **Pay Frequency:** the composition of businesses by pay frequency in the ADP data is roughly comparable to the population estimates available. In the ADP microdata, PACs reporting biweekly payroll frequency are somewhat more common, and those reporting weekly pay are somewhat less common, than in the official BLS data.

- **Region:** ADP data provide reasonable geographic representation of the country as a whole.

- **Size:** While the ADP data include businesses of all sizes (in terms of employment), ADP PACs generally tend to be larger than U.S. establishments. The ADP PAC size distribution marks a middle ground between the CES sample and the QCEW universe, with relatively more employment in small units compared to the CES sample. Notably, however, ADP has significantly more employment in mid-sized units than does CES, with a distribution that looks reasonably similar to QCEW.

- **Industry Composition:** Compared to both the CES sample and the QCEW universe, the ADP sample modestly overweights manufacturing employment and

[3]When accessing the microdata, we follow a number of procedures to ensure confidentiality. Business names are not present in the data we access.

[4]These composition differences are covered in detail in Cajner, Crane, Decker, Hamins-Puertolas, Kurz, and Radler (2018), which also provides additional detail on the construction of the indexes covered in section 2.2.

slightly overweights employment in services. Trade, transportation, and utilities employment is underweighted in the ADP data, while the weight of construction employment in ADP is similar to the CES sample, but both ADP and the CES sample underweight construction employment relative to QCEW.

2.2 Constructing Indexes of Employment using Private Payroll Microdata

We create seasonally adjusted indexes of employment at both the weekly and monthly frequencies. Importantly, the first and second moments of the weighted ADP indexes for active employment are similar to those found in the BLS data. Similarly, the correlations are high and significant, particularly during periods of sharp employment upswings or downswings.

The raw ADP payroll microdata we access are weekly data. Weekly frequency data are useful not only for providing timely updates but also for identifying high-frequency shocks (e.g., natural disasters) and the response of labor markets. We show the growth rate (at the weekly frequency) of our seasonally adjusted and weighted employment series in Figure 1. The weekly series is highly volatile, even after seasonal adjustment. The inset panel in Figure 1 transforms the growth rates into first differences. The within-month variation in employment changes is rather large, with different weeks contributing substantially to the overall monthly employment change.

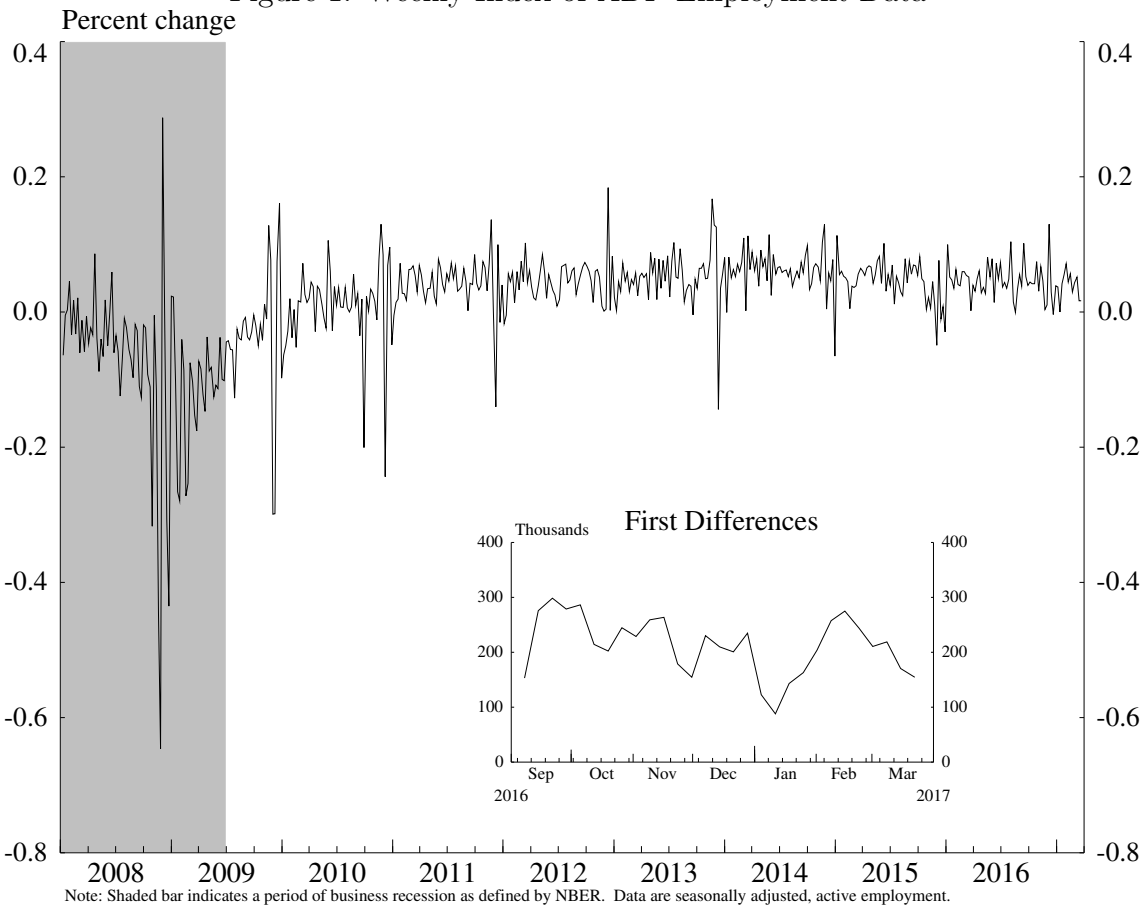
We now turn to the monthly series. The monthly data are converted from weekly data—by averaging the weekly levels—and then seasonally adjusted. We plot the ADP employment index, both weighted and unweighted, and the CES private payroll series employment in Figure 2. The top panel plots the growth rates and the bottom panel plots the levels, indexed to 100 in 2010.

Importantly, the weighted ADP series posts growth rates very similar to the CES. Moreover, the cyclical patterns of the ADP series and the CES data are very much aligned. We also observe that the ADP series are volatile, and many of the outsized monthly swings do not correspond with the monthly swings in the BLS data.

2.3 Strengths and Weaknesses of Different Types of Payroll Employment Data

Perhaps the most important issue when analyzing the quality of a dataset is its representativeness. Obviously, the QCEW data have a clear advantage here since these data

Figure 1: Weekly Index of ADP Employment Data

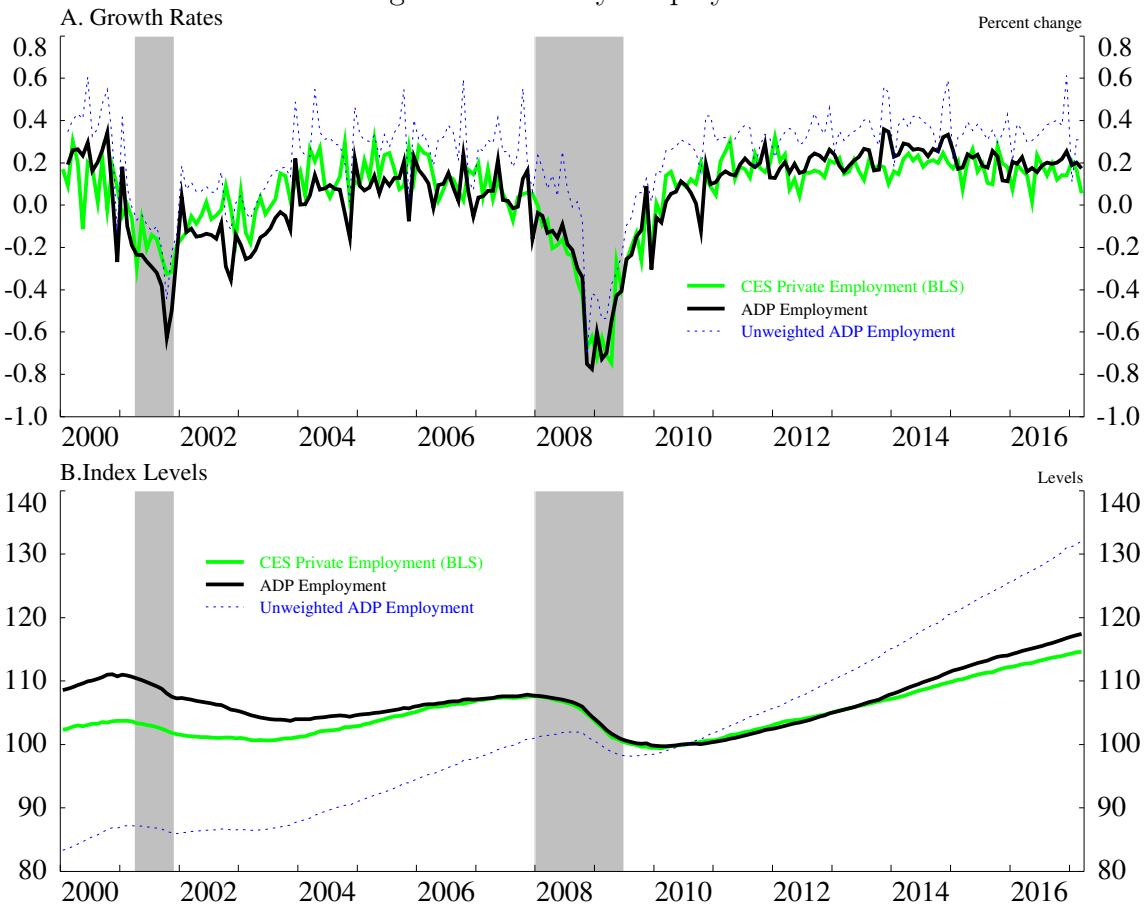


Note: Shaded bar indicates a period of business recession as defined by NBER. Data are seasonally adjusted, active employment.

represent population counts.[5] In contrast, CES and ADP estimates are sample based. As with CES, our ADP samples are adjusted with sampling weights that are meant to make the estimates representative of the U.S., but this does not solve all issues. In the case of ADP, an important sample selection issue exists since only the firms that hire ADP to manage their payrolls show up in the ADP data. In the case of CES, the data are based on a probability sample of establishments, but since the response rates are only about 60 percent (Kratzke, 2013), this introduces a sample selection issue as well. Additionally, both the ADP and the CES data are subject to dynamic selection issues related to establishment entry and exit. The BLS attempts to address this problem with the CES Net Birth/Death Model to predict net entry, while we make no such

[5]Note, though, that there is a small scope discrepancy between QCEW on the one hand and CES/ADP on the other hand: about 3 percent of jobs that are within scope for CES/ADP estimates are exempt from UI tax law. For more detail, see https://www.bls.gov/news.release/cewqtr.tn.htm.

Figure 2: Monthly Employment



Note: Shaded bars indicate periods of business recession as defined by NBER. Data are seasonally adjusted.
Current Employment Statistics (Survey: U.S.).

adjustment to the ADP data.

In terms of periodicity of data availability, the QCEW and the CES data are available monthly, while the ADP data are in principle available weekly. While for the purposes of analysis in this paper we use monthly data for all data sources, the ADP data nevertheless have a notable advantage. In particular, the QCEW and the CES data contain only information for the pay period that includes the 12th day of the month, while the ADP data incorporate all pay periods since July 2009 (and biweekly pay frequency between May 2006 and June 2009). As a result, the CES/QCEW data cannot measure employment activity over the entire month, which can be especially problematic in the case of temporary distorting events during the reference period. For example, an unusually large weather event (e.g., a hurricane or a snow storm) that reduced employment numbers during the reference period but left the rest of the month unaffected would result in a CES employment report that understates the strength of

the labor market throughout the month. In addition, higher frequency data in principle make it easier to identify outliers (for example, due to unusual hiring patterns) and thus seasonally adjust the data.

Regarding timeliness, the ADP data become available the week after the end of the pay period, the CES data usually three weeks after the end of the reference week, and the QCEW data six months after the end of each quarter.

Finally, the QCEW and the ADP data are both essentially administrative data and thus arguably somewhat less prone to reporting errors and nonresponse, which typically plague survey data such as the CES.

2.4 Weights

Estimating U.S. employment growth from the ADP sample requires reweighting the data for representativeness. We group businesses into employment size-industry cells and apply weights from the QCEW to make the estimates representative. The choice of cell size is critical here, as narrower cells help ensure the resulting estimates are unbiased, but at the cost of a high variance (due to sparse cells). To deal with these issues we use tools from the literature on nonresponse and post-stratification in surveys (reviewed in, e.g. Valliant, Dever, and Kreuter (2013)). We experiment with trimming large weights, collapsing cells, and using Bayesian approaches (as in Little (1993)) to find an acceptable compromise between variance and bias in our estimates.

2.5 Benchmarking the FRB Employment Indexes

We annually benchmark our monthly ADP series to the QCEW. Our procedure closely follows that of the BLS (March levels benchmarking and a linear wedge for the intervening months). The data are benchmarked from 2001 through March of 2017.

Our preliminary results indicate that, for the period 2001-2017, the root mean squared benchmark revision for ADP is 0.81 (in percentage points), while the root mean squared benchmark revision for CES from 2002 to 2017 is 0.36. Looking at the benchmark revisions since 2008 onwards, the root mean squared benchmark revision is 0.51 for the ADP data and 0.35 for the BLS CES data.

In the 10 years starting from 2008, the pre-benchmark ADP estimates were closer to the eventually published population counts in 4 years, while the pre-benchmark CES estimates were more accurate in 6 years. Interestingly, the ADP estimates outperformed the CES estimates during the business cycle swing associated with the Great Recession (2008-2010). In contrast, over the past 5 years the pre-benchmark ADP estimates consistently overpredicted population counts.

3 The Limits of Predicting Private Payroll Employment

Cajner, Crane, Decker, Hamins-Puertolas, Kurz, and Radler (2018) evaluates the ability of the FRB ADP employment indexes to predict labor market movements with both real-time and non-real-time indexes. We find that our indexes help to predict BLS payroll employment gains in real-time, even after accounting for market expectations of these gains. Indeed, the inclusion of our ADP employment indexes in the forecasting regression significantly reduces in- and out-of-sample root-mean-square errors (RMSE). We also find that ADP employment indexes have information that predicts revisions to BLS CES gains, even after factoring in market expectations.

To be sure, the success of the indexes derived from the ADP payroll microdata should not be overstated. The improvement to predictive ability may be statistically significant, but we should not expect too much improvement as the sample variance in the BLS CES data is large. For example from 2013 until 2017 the out-of-sample RMSE for predicting monthly payroll employment using the ADP data (along with other predictors) is 70,700 jobs, whereas the standard error of the CES data is 67,000. To the extent that sampling error is i.i.d., the CES standard error provides a lower bound on the forecastability of CES. This suggests that the main value of the ADP data may not be in forecasting applications. ADP data can be employed to introduce new statistics that are more granular, more up-to-date, and available at a higher frequency than the traditional sources of employment data currently being produced. In addition, the ADP data may be combined with the CES to reduce total measurement error. We now turn to a methodology that combines multiple data sources to estimate the "true" state of the labor market.

4 Inferring the "True" Underlying Private Payroll Employment Growth with a State-Space Model

Payroll employment growth is one of the most reliable business cycle indicators. Each postwar recession in the U.S. has been characterized by a year-on-year drop in payroll employment as measured by the BLS in CES, and outside of these recessionary declines, the year-on-year payroll employment growth has always been positive. Thus, if one knew the true underlying payroll employment growth, this would help enormously in assessing the state of the economy in real time. In this section, we present results from

a state space model to infer the true underlying payroll employment growth.[6]

Let ΔEMP_t^U denote the unobserved, true change in private payroll employment, which evolves according to the following state transition equation:

$$\Delta EMP_t^U = \mu(1-\rho) + \rho \Delta EMP_{t-1}^U + \epsilon_t^U.$$

The observed values (latest prints) of CES and ADP employment gains are a function of the underlying state according to the following measurement equations:

$$\begin{bmatrix} \Delta EMP_t^{ADP} \\ \Delta EMP_t^{CES} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Delta EMP_t^U + \begin{bmatrix} \epsilon_t^{ADP} \\ \epsilon_t^{CES} \end{bmatrix}.$$

We assume all shocks are Gaussian and uncorrelated[7]:

$$\Sigma = \begin{bmatrix} \sigma_U^2 & 0 & 0 \\ 0 & \sigma_{ADP}^2 & 0 \\ 0 & 0 & \sigma_{CES}^2 \end{bmatrix}.$$

The above model can be easily estimated with standard statistical packages. With these estimates in hand, we then proceed to calculating the optimal convex combination of ADP and CES employment growth measures.[8] In particular, let:

$$\Delta EMP_t^C = \lambda \Delta EMP_t^{ADP} + (1-\lambda)\Delta EMP_t^{CES}.$$

Both right-hand side variables are latest prints, the same as in the estimated state-space model. We would like to minimize the error:

$$\epsilon_t^C \equiv \Delta EMP_t^U - \Delta EMP_t^C.$$

Then, minimizing ϵ_t^C gives:

$$\lambda^* = \frac{\hat{\sigma}^2_{\epsilon_t^{CES}}}{\hat{\sigma}^2_{\epsilon_t^{CES}} + \hat{\sigma}^2_{\epsilon_t^{ADP}}}.$$

where $\hat{\sigma}^2_{\epsilon_t^{CES}}$ is the estimated variance of the observation error ϵ_t^{CES}, and similarly for $\hat{\sigma}^2_{\epsilon_t^{ADP}}$. Our main results are summarized in Figure 3. The thick blue line represents the underlying payroll gains estimated with the Kalman filter. It is clear from the figure that both the ADP and CES measures of employment gains deviate over time

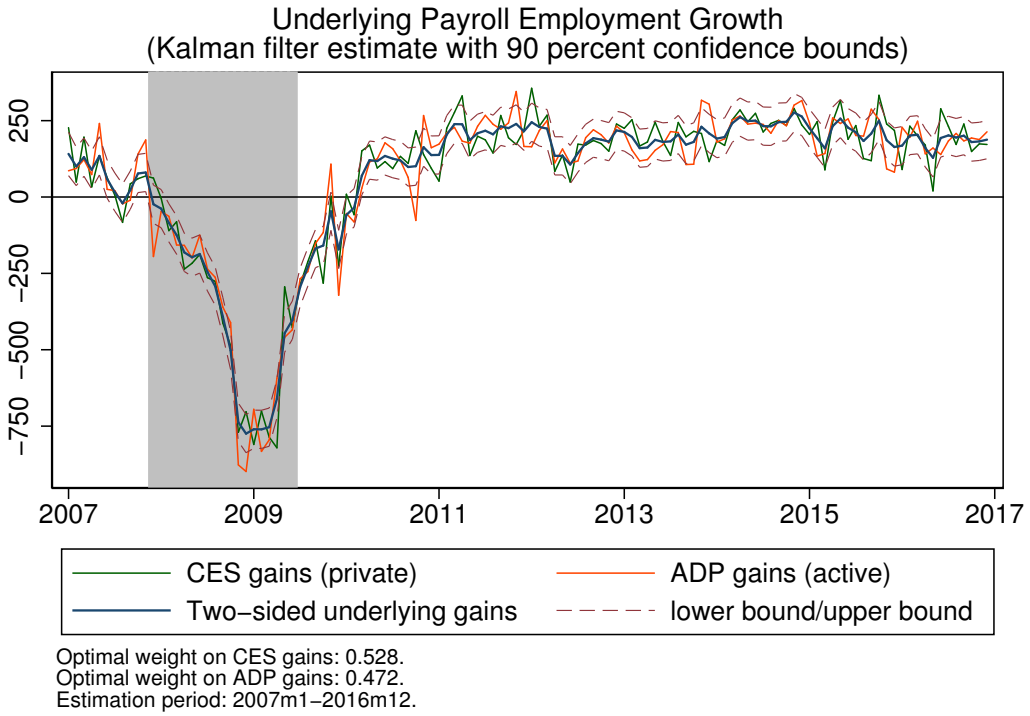
[6] Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2016) use a similar approach to provide a better measure of output.

[7] In ongoing work, we explore a model that allows for correlated shocks.

[8] For related approaches, see, for example, Mankiw, Runkle, and Shapiro (1984) and Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2013).

from this measure of underlying payroll gains. After calculating the standard deviation of these errors for both observed measures of payroll gains, we find that one should optimally put the weight of 53 percent on CES gains and 47 percent on ADP gains.[9] For comparison, we repeat the exercise and replace the employment measures with total nonfarm CES employment gains and adjusted CPS gains. In the latter case, one would put an optimal weight of only 4 percent on CPS employment gains and 96 percent on total nonfarm CES employment gains.

Figure 3: Underlying Payroll Employment Growth



At first blush, the result that one should optimally put a roughly equal weight on CES and ADP employment gains might seem surprising. However, both datasets cover roughly a similar share of private U.S. payroll employment (23 percent for CES, 20 percent for ADP) and thus the sampling error could plausibly be of similar magnitude. Indeed, the BLS calculates that the 90-percent confidence interval for monthly change in private CES payroll employment is plus/minus 110,000. Although the BLS eventually benchmarks CES payroll employment to the QCEW (which has nearly universal

[9]Note that the linear combination of the ADP and CES series is nearly identical to the smoothed two-sided state estimate from the Kalman filter.

coverage), the benchmarking is done only for March each year, while the month-to-month changes are largely unaffected due to the procedure of linearly wedging-back the discrepancy between March estimates of employment in CES and QCEW. As a result, if in a particular month the CES sample estimate of payroll employment gain is distorted due to the sampling error, it is likely that the error will survive even the subsequent revisions. Since the ADP data rely on a (mostly) different sample, it should be unsurprising that taking a Kalman filter estimate of underlying gains based on both observed measures should give a more precise estimate of the current pace of employment growth, with weights being roughly similar due to the similar sample size.

We also intend to provide several robustness checks and alternative specifications to the standard Kalman Filter approach outlined above. We then intend to evaluate the performance of the combined ADP-CES employment measure in some economic applications, such as nowcasting and recession probability models.

References

Aruoba, S. B., F. X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song (2013): "Improving U.S. GDP Measurement: A Forecast Combination Perspective," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, ed. by X. Chen, and N. R. Swanson, pp. 1–25. Springer New York, New York, NY.

Aruoba, S. B., F. X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song (2016): "Improving GDP measurement: A measurement-error perspective," *Journal of Econometrics*, 191(2), 384–397.

Cajner, T., L. Crane, R. A. Decker, A. Hamins-Puertolas, C. Kurz, and T. Radler (2018): "Using Payroll Processor Microdata to Measure Aggregate Labor Market Activity," Finance and Economics Discussion Series 2018-005, Board of Governors of the Federal Reserve System (U.S.).

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

Einav, L., and J. Levin (2014): "Economics in the age of big data," *Science*, 346(6210).

Groen, J. (2012): "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures," *Journal of official statistics*, 28, 173–198.

Kratzke, D.-T. (2013): "Nonresponse Bias Analysis of Average Weekly Earnings in the Current Employment Statistics Survey," Discussion paper, Bureau of Labor Statistics.

Little, R. J. A. (1993): "Post-Stratification: A Modeler's Perspective," *Journal of the American Statistical Association*, 88(423), 1001–1012.

Mankiw, N., D. E. Runkle, and M. D. Shapiro (1984): "Are preliminary announcements of the money stock rational forecasts?," *Journal of Monetary Economics*, 14(1), 15 – 27.

Valliant, R., J. Dever, and F. Kreuter (2013): *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.