Federal Reserve Bank of Chicago

Predicting Benchmarked US State Employment Data in Realtime

Scott A. Brave, Charles Gascon, William Kluender, and Thomas Walstrum

REVISED
June 3, 2020

WP 2019-11

https://doi.org/10.21033/wp-2019-11

Predicting Benchmarked US State Employment Data in Realtime

Scott A. Brave,[a] Charles Gascon,[b] William Kluender[c], and Thomas Walstrum[d]

June 3, 2020

Abstract: US payroll employment data come from a survey and are subject to revisions. While revisions are generally small at the national level, they can be large enough at the state level to alter assessments of current economic conditions. Users must therefore exercise caution in interpreting state employment data until they are "benchmarked" against administrative data 5–16 months after the reference period. This paper develops a state-space model that predicts benchmarked state employment data in realtime. The model has two distinct features: 1) an explicit model of the data revision process and 2) a dynamic factor model that incorporates realtime information from other state-level labor market indicators. We find that the model reduces the average size of benchmark revisions by about 11 percent. When we optimally average the model's predictions with those of existing models, the model reduces the average size of the revisions by about 14 percent.

Keywords: Benchmarking methods; Real-time data; Revisions; Forecasting accuracy; Time series; Nowcasting; US employment

Acknowledgements: The authors wish to thank Keith Phillips and seminar participants at the Federal Reserve Bank of Chicago for their comments, along with organizers and participants at the 2nd Conference on Forecasting at Central Banks and the day-ahead meetings of the Federal Reserve Regional System Committee for their comments on an earlier draft.

---

[a] Federal Reserve Bank of Chicago; 230 S. LaSalle St., Chicago, IL, USA 60604; sbrave@frbchi.org.

[b] Federal Reserve Bank of St. Louis; 1 Federal Reserve Bank Plaza, St. Louis, MO, USA 63102; charles.s.gascon@stls.frb.org.

[c] Compass Lexecon; 332 S Michigan Ave #1300, Chicago, IL, USA 60604; bkluender14@gmail.com.

[d] Federal Reserve Bank of Chicago; 230 S. LaSalle St., Chicago, IL, USA 60604; twalstrum@frbchi.org; corresponding author.

1 Introduction

State nonfarm payroll employment is one of the most important indicators of regional economic activity in the United States. Produced as part of the Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) program, it is a strong predictor of other coincident measures of economic activity (Crone and Clayton-Matthews (2005)). It is also a rare example of a state economic indicator with a reporting lag of less than one month. Many other state level indicators take nearly two quarters to be released. For example, state gross domestic product, arguably the broadest measure of state economic activity, is released five months after the end of the reference quarter.

To achieve such a quick turnaround, the BLS relies on a survey of nonfarm businesses establishments. As a result, the data are subject to revision. The first revision occurs one month after the initial release and includes any missing or corrected responses. The second revision occurs every March when the CES survey results are "benchmarked" against administrative data sources that cover the universe of nonfarm workers. The state revision process is similar to that of the national CES employment data, but revisions to the national data generally receive little attention because they tend to be small: The average national 12-month growth rate revision is 0.2 percent. In contrast, revisions tend to be much larger at the state level, with an average 12-month growth rate revision of 0.6 percent across the 50 states since 2005. Revision sizes are larger than the national average for every state except Pennsylvania, and the largest average revision size is for North Dakota, at 1.0 percent.

In some instances, benchmark revisions are large enough to swing an apparent net job loss in a state for the year to a sizeable net gain. To illustrate the potential magnitude of the state level

benchmark revisions, Figure 1 compares two data vintages of CES employment in Illinois from January 2014 to December 2015. The blue line shows data from the January 2016 vintage, where the dashed portion of the line represents data that have not been benchmarked. The black line shows data from the March 2016 vintage, which included newly benchmarked data through September 2015. The newly benchmarked data indicate that rather than losing 3,000 workers from December 2014 to December 2015, Illinois actually gained 51,000 workers.

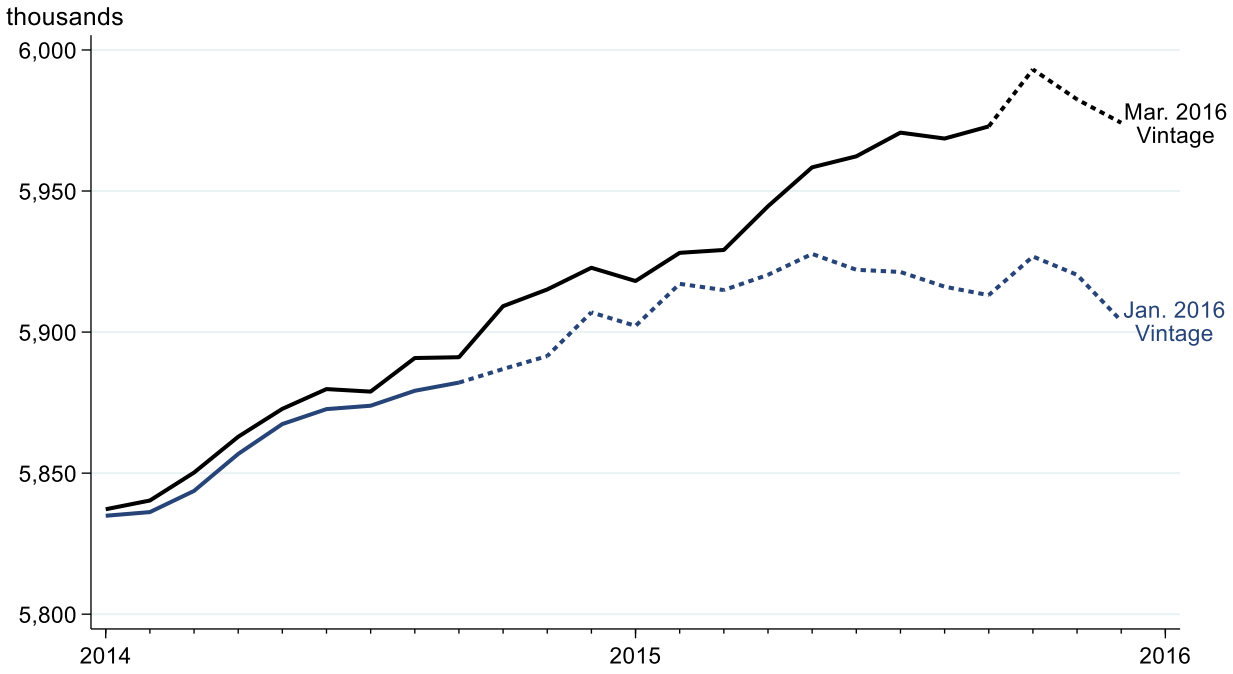

Figure 1. Current Employment Statistics (CES) data for Illinois before and after the March 2016 benchmark. The short-dashed lines represent survey-based data that are not yet benchmarked against administrative data. Source: St. Louis Fed archive of data from Haver Analytics.

This paper develops and tests a state space model that predicts benchmarked state employment data in realtime. The model has two distinct features: 1) an explicit model of the data revision process and 2) a dynamic factor model that incorporates realtime information from other state level labor market indicators. We find that across all 50 US states, the model reduces the average size of the benchmark revisions by about 11 percent on a mean absolute error basis. That said, the model's performance varies by state, and for some states does not improve on the initial

release of the official CES data. For this reason, we employ a model averaging technique that allows us to optimally average our model's predictions with those of the official CES data and an existing model developed by Berger and Philips (1993). We find that by averaging models we can reduce the average size of the revisions across states by about 14 percent.

The literature has long recognized the challenge posed by data revisions for analyzing current economic conditions at the national and local levels. For example, Croushore and Stark (2001) created a realtime dataset of macroeconomic indicators for the US that allowed macroeconomists to examine how data revisions affect forecasts. Their research started an extensive literature that explores the reliability of initial releases of macroeconomic data. An example from this literature is Orphanides and van Norden (2002), which uses the Croushore and Stark (2001) dataset to examine the unreliability of output gap measures. While national data like the output gap are a major focus of the data revision literature because of their relevance for national fiscal and monetary policymakers, state and local economic indicators are also important because state and local policymakers rely on them to make decisions. For example, policymakers use the data to project tax revenues and outlays and to estimate the impact of economic development spending.

In spite of the importance of state and local data and their much larger revision sizes, research on revisions to state and local data is limited. That said, this paper builds on two earlier papers that address this problem and take approaches related to ours. Coomes (1992) uses a state space framework similar to the one we develop to improve upon estimates of employment in Virginia and the Louisville metropolitan area. And Berger and Philips (1993) develop a model for predicting CES benchmark revisions for Texas. We estimate the Berger and Philips (1993) model for all 50 states to serve as a comparison for the model in this paper, and like our model, it performs well for some but not all states. Because our models are different in some important ways, they play

complementary roles in this paper's model averaging exercise. We explain how the Berger and Philips (1993) "early benchmark model" is estimated in this paper's appendix.

The state space framework we develop builds on the work of Coomes (1992) by explicitly modeling the BLS's succession of data revisions, which take between 5 and 16 months to complete (depending on the calendar month) and involve four data releases. We outline this process in detail in section 2. Statistical agencies tend to treat the data revisions we model in this paper as an analytical problem that involves reconciling information across multiple sources. This is the approach of Berger and Philips (1993), who reconcile realtime administrative data with the official data to produce an "early benchmark" of CES employment. In section 3, we instead pose the revision process as a signal extraction problem where the "true" or "final" value is unobserved (see, for example, Aruoba (2008)) and must be filtered in realtime from multiple noisy observations of CES state employment.

Another unique feature of our state space framework is that it includes a dynamic factor model that incorporates other state level labor market indicators that are unrelated to the revision process but contain relevant information for state employment. This is a new approach to modeling the CES revision process, but is similar in spirit to approaches used for the realtime tracking of business conditions (see, for example, Aruoba, Diebold, and Scotti (2009)). It is also related to the literature on "nowcasting" (that is, forecasting the current period), as in Giannone, Reichlin, and Small (2008), who nowcast real gross domestic product, Knotek II and Zaman (2014), who nowcast CPI inflation, and Monteforte and Moretti (2013), who nowcast Euro area inflation.

While the models of Coomes (1992) and Berger and Philips (1993) made progress in predicting benchmarked state employment data, they were difficult to evaluate because of the limited availability of realtime data. In section 4, we test our model and the Berger and Philips (1993)

early benchmark model using a realtime dataset of state CES employment with vintages that go back to 2005. We find that our model has smaller errors compared to the initial CES release for 36 of 50 states and smaller errors compared to the early benchmark model for 31 of 50 states. Because our model is not always the best performer, we draw on the literature on forecast combinations (see, for example, Timmermann (2006)). In section 4, we show that when we optimally combine our model's predictions with those from the early benchmark model and the initial official CES release, we can achieve better performance than that of any single model on its own.

2 The CES Data Revision Process

In order to achieve a quick turnaround in its release of payroll employment data, the BLS conducts a survey of business establishments and then revises its estimates as more comprehensive data become available. According to the BLS (2016), the CES program covers all establishments with employment over 1,000 and surveys a representative sample of smaller establishments. Roughly 689,000 establishments and 40 percent of nonfarm workers are covered by the program.

While the survey is very reliable for producing national statistics, estimates for state and local areas necessarily rely on smaller samples, which range from around 80,000 establishments in California to 2,100 establishments in Rhode Island. State level industry subsamples are even smaller. In instances where samples are too small to publish reliable employment counts, the BLS utilizes a small domain model that produces a weighted least squares estimate based on the CES survey, an ARIMA projection, and employment for that industry in adjacent states. The reported value is calculated as a weighted average of these three estimates, where the weights are based on forecast

accuracy. The BLS reports that 43 percent of state and local CES series are calculated in this way.[1] All told, the BLS's methods for producing state and local CES data means that they are subject to revisions due to sampling error as well as non-sampling error such as response bias, survey nonresponse, model misspecification, methodological changes.

State and local CES data go through two primary revisions. One month after the first "preliminary" release, the BLS produces a second "final" estimate, which includes additional data from establishments that did not submit survey responses in time for the preliminary release. Then, each March, the BLS releases "benchmarked" estimates for October of two years prior through September of the prior year.

The benchmarked estimates are calculated from administrative data that comprise nearly the entire universe of nonfarm workers. The administrative data come from a number of sources, but the primary source is the unemployment insurance program, which covers about 97 percent of nonfarm workers. The remaining 3 percent come from a variety of other sources, but primarily from the Railroad Retirement Board and County Business Patterns (Bureau of Labor Statistics, 2017). Once the data are benchmarked, subsequent revisions are typically quite small.

The BLS also releases data on the universe of workers who are covered by the unemployment insurance program (including some farm workers) through its Quarterly Census of Employment and Wages (QCEW) program. The data are monthly, but released on a quarterly basis 5 to 7 months after the end of the reference period. Because the QCEW data are released quarterly and are the primary sources for the CES benchmark revisions, it is possible to use the QCEW data to

[1] See https://www.bls.gov/sae/additional-resources/guaranteed-publication-levels-and-the-ces-small-domain-model-sdm.htm

predict the future benchmark revisions for the earliest portion of the non-benchmarked CES data. This is the approach of the Berger and Phillips (1993) early benchmark model.

As an example of the CES data revision process, Figure 2 shows a timeline of CES releases associated with the data for April 2016. The first (preliminary) CES release of April 2016 CES employment was on May 20, 2016. On June 17, 2016, the BLS released its second (final) estimate of April employment (concurrent with the preliminary release of data for May 2016). In December 2016, seven months after the first CES release of April employment, the QCEW data for April 2016 were released. Finally, on March 13, 2017, the BLS released the benchmarked CES data for April 2016.

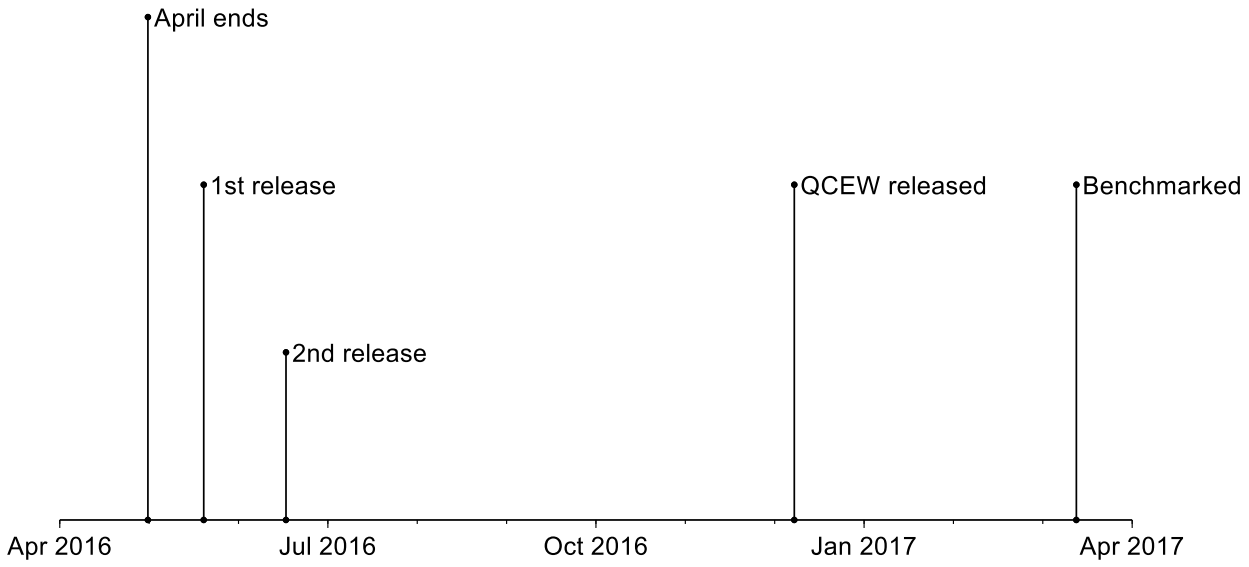

Figure 2. Revision timeline for state CES data for April 2016.

3 A State Space Model of Benchmarked CES Data

In this section, we outline our realtime data construction and our state space model for benchmarking state employment data in realtime, which builds on the models of Coomes (1992) and Berger and Phillips (1993). We estimate our model separately for each state because differences in state size, industry composition, and data collection methods (see section 2 for details) can lead to

large differences in model parameters. Our state space model contains two sub-models. First, we model the CES data revision process and we call this sub-model the "revision model." Second, we develop a dynamic factor model that incorporates additional realtime state level labor market indicators that are separate from the CES data program. We call this sub-model the "factor model."

3.1 Realtime Data Construction

Our realtime data come from a weekly archive of the Haver Analytics database maintained by the St. Louis Fed.

To capture the realtime properties of the CES benchmarking process, we construct a number of vintage data series from the seasonally adjusted CES releases and the QCEW. First, we construct a vintage series of the first CES releases for each month, $CES1$. The series includes, for example, the June 2014 observation released in July 2014, the July 2014 observation released in August 2014, and so on. We next construct a vintage series of second releases, $CES2$, which are the CES releases after additional responses have been incorporated. The series includes, for example, the May 2014 observation released in July 2014, the June 2014 observation released in August 2014, and so on. Finally, we construct a series of benchmarked values, $CESBench$, which contains only the portion of the data from a given vintage that are benchmarked. The series includes, for example, observations through September 2013 from the July 2014 vintage.

For the QCEW series, $QCEW$, we match as closely as possible CES nonfarm payroll employment, which means we remove all employees in the agriculture, forestry, fishing, and hunting sector (NAICS 11) with the exception of those in the logging sector (NAICS 1133). QCEW data are subject to revision until the BLS finalizes them with the release of data for the first quarter

of the following year. While we do not have realtime data for the QCEW, the BLS (2019) indicates that the revisions are typically minor.[2]

A key component of our model is that we incorporate other state employment indicators, Y, in addition to those directly involved in the revision process. We use only indicators for which we have realtime data for all states. In practice, it may be desirable to include indicators in the model that are available only for some states or for which realtime data are not available. For consistency in testing our specification across states, we exclude such data and develop a model that can apply to all states and can be tested on realtime data.

The additional series we include are a state's initial claims for unemployment insurance (seasonally adjusted by us), national employment from the BLS's Current Population Survey, national employment from the CES, and two measures we calculate ourselves that require some explanation. The first measure we calculate aims to capture any spatial correlation in employment changes across neighboring states. For example, employment changes in Indiana and Wisconsin may be correlated with employment changes in Illinois. For a given state, our spatial correlation measure is the weighted average of the change in CES employment across the other 49 states, where the weights are a neighboring state's population in 2010 divided by the square of the Euclidian distance from a neighboring state's population centroid to the given state's population centroid. The second measure we calculate is an estimate of total state employment based on the combination of national CES industry-level data and state industry employment shares. For a given state and 3-digit NAICS industry, we multiply total national CES employment in that industry by the state's share of

[2] The BLS does not release seasonally adjusted QCEW data, so we first convert finalized QCEW data into pseudo-real time vintages and then seasonally adjust the series using the Census Bureau's X12 procedure.

national employment in that industry according to the QCEW. We sum across all industries to get our estimate of total state employment.

3.2 Forecasting CES State Employment

We model revisions to CES state employment data on a state-by-state basis using the unobserved components framework described in this section. All series shown are in logs, with first differences denoted by the operator Δ.

$$CESBench_t = E_t$$

$$CES2_t = E_t + B_t \qquad B_t = \kappa + \rho B_{t-1} + \eta_t$$

$$CES1_t = E_t + B_t + R_t \qquad R_t = \omega_t$$

$$QCEW_t = E_t + W_t \qquad W_t = \delta + \sum_i \lambda_i W_{t-i} + \nu_t$$

Each latent state in our framework captures an element of the revision process or CES employment dynamics. Our target variable, E_t, represents benchmarked CES data, $CESBench_t$, which are observed 5 to 16 months after the reference month. The difference between $CESBench_t$ and the once-revised second release $CES2_t$ is the benchmark revision B_t.[3] We assume that B_t follows a first-order autoregressive process around a potentially nonzero conditional mean, κ, imposing $|\rho| < 1$ to ensure stationarity. This specification for B_t is common for all states and allows us to flexibly model revisions to the average level and slope of $CES2_t$. The difference between $CES2_t$ and the initial release $CES1_t$ is captured by R_t. We assume R_t is an iid mean zero random variable, which is based on an examination of sample averages and autocorrelation functions for all 50 states. Finally, the

[3] Note that for historical data beyond the most recent benchmark, this difference will reflect the initial and subsequent benchmark revisions. Because the size of subsequent benchmarks tend to be small (except when methodological changes occur), we do not model them as a separate source of error for these observations.

difference between $CESBench_t$ and $QCEW_t$ is a time-varying "wedge", W_t, that we assume follows an AR(p) process around a conditional mean δ and state-specific lag order p.[4]

To estimate E_t using the relevant information contained in the additional state labor market indicators, Y_t, we assume that a dynamic factor structure exists between them.

$$\Delta E_t = \alpha + f_t + \zeta_t$$

$$\Delta Y_t = \gamma + \Gamma f_t + \upsilon_t$$

$$f_t = \theta f_{t-1} + \varepsilon_t$$

$$\upsilon_t = \psi \upsilon_{t-1} + \vartheta_t$$

The scale and sign of the factor, f_t, are set by constraining the factor loading for ΔE_t to be 1 and restricting the sign of the element of the loading vector Γ on unemployment insurance claims in ΔY_t to be negative. We assume that the factor follows a stationary AR(1) process with $|\theta| < 1$ and that the idiosyncratic errors of the factor model, υ_t, do as well ($|\psi| < 1$). Together with the estimated dynamics of the revision process and the QCEW wedge, these dynamic processes allow us to forecast the benchmarked values of the state CES data beyond the last available $CESBench_t$ observation.

3.3 State Space Model and Estimation

With the further assumptions of iid normally distributed errors for the latent variables and observables, the model can be estimated by maximum likelihood methods with the Kalman filter as described in Durbin and Koopman (2012). The state space representation of our model that we use

[4] We choose p according to the Bayesian Information Criterion (BIC) for each individual state. Similar tests were also used to determine the lag order for other dynamic specifications. Note that the structure of the latent variables makes it possible to construct BIC statistics simply by taking differences of observed data series, i.e. $B_t = CES2_t - CESPost_t$, and estimating autoregressive specifications by OLS.

for estimation is shown below. Note that the autoregressive specification for W_t is presented in companion form, as the exact number of lags used for estimation varies by state.[5]

$$\begin{bmatrix} CESBench_t \\ CES2_t \\ CES1_t \\ QCEW_t \\ \Delta Y_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \gamma \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & \beta & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} E_t \\ f_{t+1} \\ B_t \\ R_t \\ W_t \\ \upsilon_{t+1} \end{bmatrix}$$

$$\begin{bmatrix} E_t \\ f_{t+1} \\ B_t \\ R_t \\ W_t \\ \upsilon_{t+1} \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \kappa \\ 0 \\ \delta \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & \theta & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi \end{bmatrix} \begin{bmatrix} E_{t-1} \\ f_t \\ B_{t-1} \\ R_{t-1} \\ W_{t-1} \\ \upsilon_t \end{bmatrix} + \begin{bmatrix} \zeta_t \\ \varepsilon_{t+1} \\ \eta_t \\ \omega_t \\ \nu_t \\ \vartheta_{t+1} \end{bmatrix}$$

To provide some intuition for how the state space model incorporates the observable data at our disposal into an estimate of E_t, we turn again to our example of the January 2016 CES employment vintage for Illinois. We first examine our model's estimate of the Kalman smoothed E_t, which is shown in Figure 3 in red. Our primary target is the value for December 2015, which is first benchmarked in the BLS's March 2017 vintage (black). We also show the BLS's January 2016 vintage (blue). In this case, the state space model provides an estimate that is much closer to the benchmarked data from the March 2017 vintage than the BLS's January 2016 vintage is.

How does the state space model arrive at its estimate for Illinois's December 2015 employment value? We provide some intuition for this in Figure 4, where we decompose the state space model's estimate for Illinois shown in Figure 3 into contributions from the observable data. The decomposition emphasizes the fact that our model is estimated in both levels and growth rates—the revision portion in levels and the factor model portion in growth rates.

[5] This information is available from the authors upon request. We do not present results for lag orders beyond one for the factor and idiosyncratic errors because they tended to produce inferior forecasts.

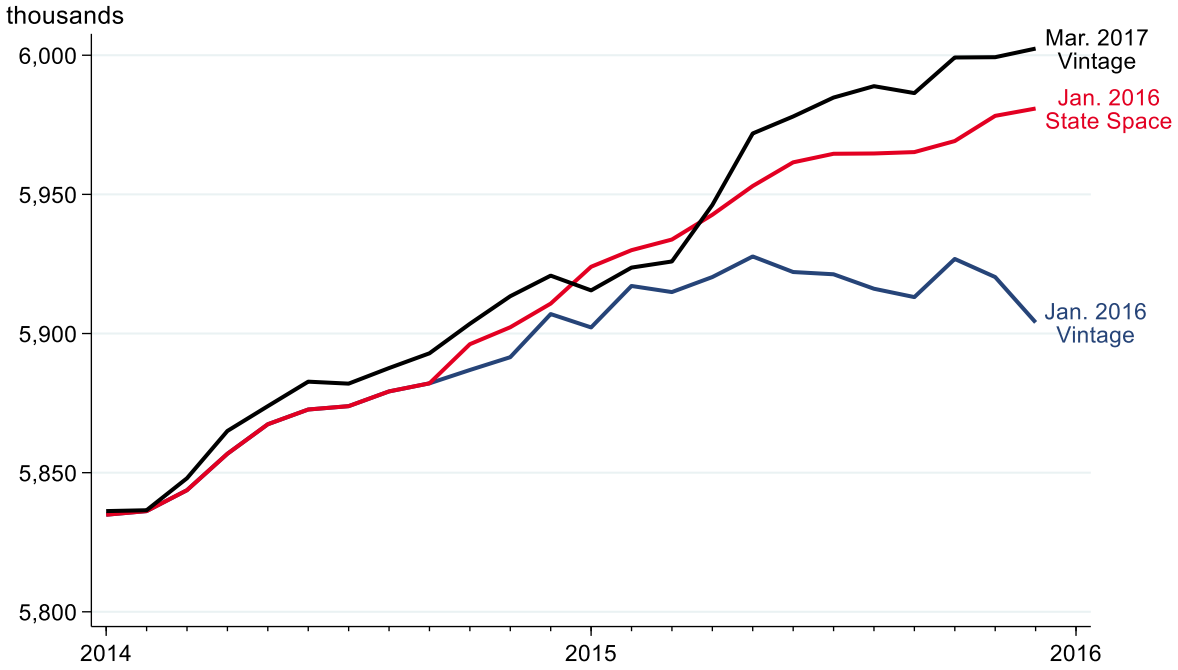

Figure 3. Comparison of Illinois CES Employment by vintage or Model. The January 2016 vintage was the first release to include data for December 2015 and the March 2017 vintage was the first to include benchmarked data for December 2015. The state space and early benchmark model estimates are based on data available as of January 2016. Sources: St. Louis Fed archive of data from Haver Analytics and authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

We first consider panel A, which highlights how contributions from the observable data that go into the revision model evolve over time. The black bar on the far left is the value for September 2014, which is the month where the benchmarked data end in the January 2016 vintage. The model initially puts substantial weight on the benchmarked data. However, because the benchmarked data are not available after September 2014, the model progressively puts more weight on the QCEW data. When the QCEW data are no longer available (July 2015), the model starts to rely more on the CES second release data and the factor model indicators.

Note that the CES second release series ends in November 2015 (the second to last bar), so it is perhaps surprising that the CES first release does not contribute to the December 2015 estimate. This is because the model relies on the factor model data for its December 2015 estimate of E_t and

relies on the CES first release solely for estimating R_t, the white noise error term that is the difference between the first and second CES releases. While the difference between the first and second releases is typically small, it is apparently large enough that the model chooses to rely solely on the information in the factor model for its estimate of the change in employment from November to December. This result is true across all states and vintages. When we estimate the revision model by itself (that is, without the factor model), the CES first release does contribute to the estimate of E_t. In this case, it receives an especially large weight for vintages' final observations, since it is the only available data series.

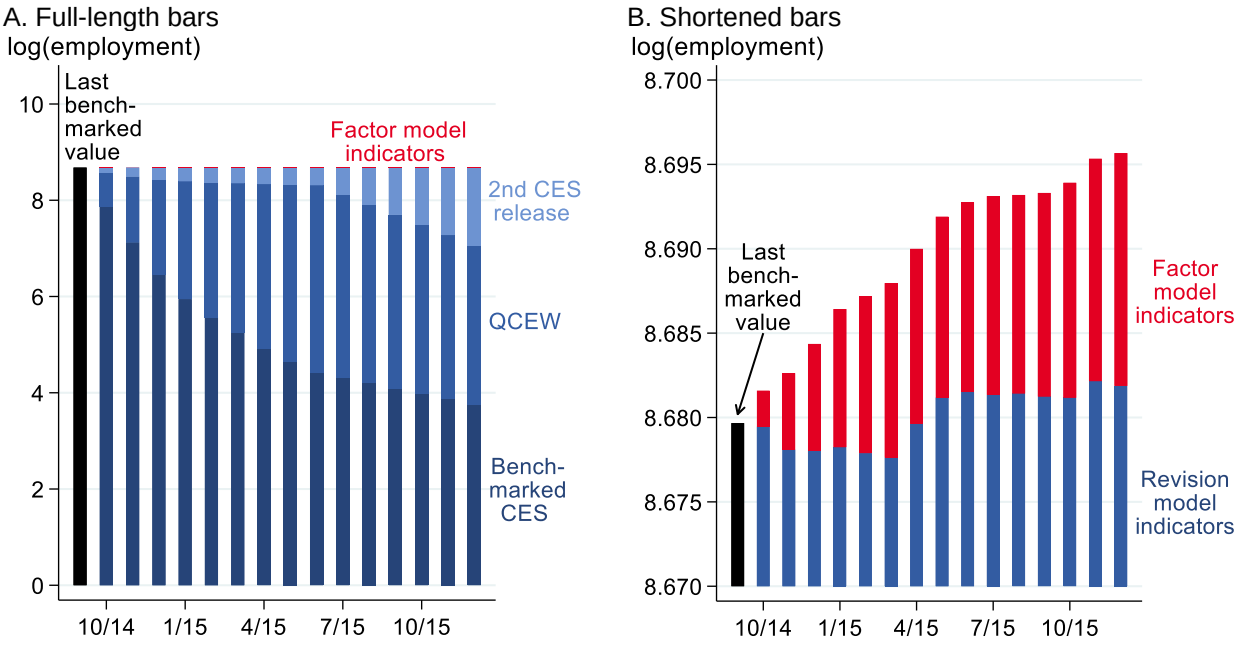

Figure 4. Contributions to E_t for the Illinois January 2016 vintage. The figure depicts an estimate of the natural logarithm of Illinois CES employment using data available as of January 2016. Benchmarked data are in black and were available through September 2014. The state space model estimates of E_t begin in October 2014. The revision model bars in panel B represent the consolidated contributions of the benchmarked CES, QCEW, and CES second release data shown in panel A. Sources: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

Panel A of Figure 4 shows that the factor model makes contributions starting in October 2014, but the contributions are barely visible because it enters the model in growth rates. That is, the

factor model is estimating the month-to-month change in bar height, but it is impossible to see month-to-month changes in panel A. In panel B, we cut off the bars at 8.67 and consolidate the contributions of the revision model to highlight how the factor model contributes to the estimate of E_t. Panel B shows that the factor model plays an important role in estimating the month-to-month change in E_t even in October 2014, and that its contribution grows over time.

Another way to describe how the state space model works is to examine the estimated unobserved components of the model. Figure 5 shows our Kalman smoothed estimate of E_t for the Illinois January 2016 vintage along with B_t (the benchmark revision), R_t (the revision to the first CES release), W_t (the wedge between the CES and the QCEW), f_t (the factor) and ζ_t (the idiosyncratic error term in the formula relating f_t to ΔE_t). The earlier, blue portions of the E_t, B_t, R_t, and W_t lines are not estimated because benchmarked CES data are available and the model equates E_t to the benchmarked data. This means that, for example, B_t is simply $CESBench_t - CES2_t$ (note that by this formulation, upward benchmark revisions to $CES2_t$ result in a negative B_t). The red portion of the lines, then, is what the model estimates in the absence of $CESBench_t$ based on the observable data and the model's parameters.

It is easy to see the persistence in B_t and W_t and why R_t is treated as white noise. Part of the persistence in B_t is the fact that revisions tend to be in the same direction for a given benchmark period. That is, when a new set of benchmarked data are released each March, the revisions in that set are usually all in the same direction. The benchmark revisions also tend to be in the same direction from one benchmark period to the next, but this is not always the case. In contrast, the revisions between the first and second releases of the CES data are much less persistent and frequently are of different signs from month-to-month, although this not always the case. For

example, while the time series of R_t as a whole displays a very low autocorrelation coefficient for Illinois, during the Great Recession, consecutive revisions tended to be negative for many months.

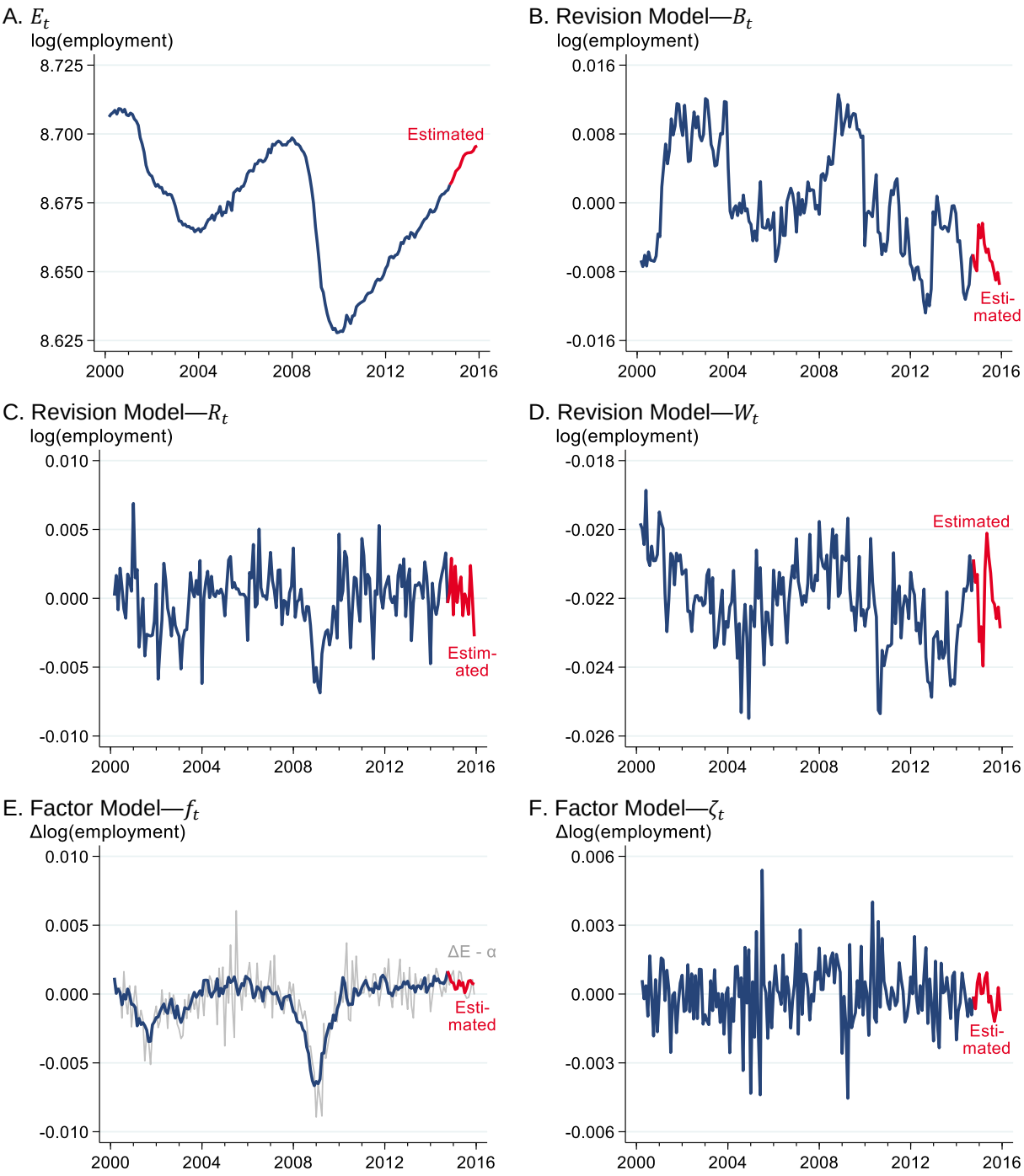

Figure 5. Unobserved components for the Illinois January 2016 vintage. Data for B_t, R_t, and W_t do reflect the opposite direction of revision. For example, a negative value for B_t means that the CES data for that month were revised up. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

It is also important to note the different scales for each of the unobserved components. The scale of E_t is of course much larger than that of B_t, R_t, W_t, and f_t. The scale of W_t indicates that the wedge between the CES and QCEW is about 2 percent for Illinois. And while R_t is usually smaller than B_t, it is often similar in size to ΔE_t (shown in panel E with f_t), which helps to explain why the model doesn't use $CES1_t$ to estimate E_t. Panel E also shows that f_t is much smoother than ΔE_t, as it emphasizes the persistent common variation in state employment indicators at the expense of the idiosyncratic volatility in $\Delta CESBench_t$, ζ_t. Unlike the idiosyncratic components of the factor model indicators, we treat ζ_t as white noise, and its lack of persistence is confirmed for Illinois in panel F.

4 Out-of-Sample Analysis

To test the predictive ability of the state space model, we perform an out-of-sample analysis of its end-of-sample prediction, or nowcast, for benchmarked CES employment for all 50 states spanning realtime data vintages from March 2005 through October 2018, with a sample period extending back to January 1990.[6] Our target for judging model performance is the 12-month log percent change in employment as reported in the first vintage in which a given month's employment is benchmarked for the first time. We label this target $\Delta_{12} CESFin_t$, where Δ_{12} represents the 12-month change in log levels. For example, in the case of the December 2015 observation for Illinois that we discuss above, we compare the log percent change in employment from December

[6] Maximum likelihood estimation of our state space model requires initial parameter values for each state, which we obtain from an autoregressive model of the expected interactions between the observed data and our model's latent state variables for the first data vintage. Beginning with the second vintage, we use the parameter estimates from the previous vintage to initialize each of the subsequent 11 vintages in a benchmark period and then update these values every 12 vintages to account for the impact of annual revisions to the Current Population Survey data included in the factor model before starting the process over again for the subsequent benchmark.

2014 to December 2015 as reported in the January 2016 vintage to that reported in the March 2017 vintage, as March 2017 is when the December 2015 observation is first benchmarked.[7]

The 12-month growth rate measure allows us to focus on revisions made to the portion of the data that were not previously benchmarked. The 1-month growth rate is a common target for testing model performance, but we prefer the 12-month growth rate because it approximately covers the whole portion of the time series that are not benchmarked. In the end, as shown in appendix section 7.3, our results are qualitatively similar whether we assess the model's performance in levels, 1-month growth rates, or 12-month growth rates, but we report the 12-month growth rate results below because we believe they provide the cleanest performance assessment.

4.1 Model Performance across States and Over Time

We assess the performance of the state space model by comparing its predictions to 1) a naïve model that predicts no difference between the first CES release of employment and its benchmarked value and 2) the early benchmark model of Berger and Phillips (1993).

The Berger and Phillips (1993) approach is to follow the BLS benchmark procedure, but rather than doing it annually as the BLS does, doing it quarterly when the QCEW data are released. This method begins with the historical benchmarked portion of the CES series. When the benchmarked portion runs out, the series is extended by applying the month-to-month growth rates from a QCEW-based series that is constructed to reflect as closely as possible the nonfarm worker population.[8] Because the QCEW data typically provide a very good prediction of the benchmark

[7] Alternatively, one could focus on the level of employment by comparing the December 2015 value reported in the January 2016 vintage to the one reported in the March 2017 vintage. However, such comparisons can include level shifts in the data that are the result of methodological changes, not sampling error, and affect all of the previous values in the series.

[8] See the appendix for further discussion of Berger and Phillips (1993) method.

revisions, there is little difference between the state space model and the early benchmark model until after the QCEW data are no longer available. Thus, the comparison between our model and the early benchmark model is a test of which model does best once the QCEW data are no longer available.

Our performance metric is based on a comparison of mean absolute errors (MAEs). For each model and for each state, we calculate the MAE across the $T = 164$ vintages in our sample. We then take the difference between the MAE of the state space model and that of the other comparison models. Formally, our performance metric is:

$$Difference\ in\ MAE_s = \frac{\sum_{t=1}^{T}|\Delta_{12}CESFin_t - \Delta_{12}E_t|}{T} - \frac{\sum_{t=1}^{T}|\Delta_{12}CESFin_t - \Delta_{12}ComparisonModel_t|}{T}.$$

We show the distribution of the MAE metric across all 50 states in Figure 6, where panel A is the distribution for the comparison with the first CES release and panel B is the distribution for the comparison with the early benchmark model. By design, bars with negative x-axis values correspond to an improvement in MAE for the state space model over the comparison models. With a mean of −0.048 percentage points and median of −0.062 percentage points, panel A indicates that the state space model succeeds on average in nowcasting state employment better than the first CES release for most states. The average MAE across states is 0.56 percentage points, which indicates that, on average, the state space model reduces the MAE of the first CES release by about 9 percent. However, in 14 states the difference in MAE is positive. In section 4.2, we estimate a regression model of the difference in MAE that helps to explain why the state space model does better for some states than for others.

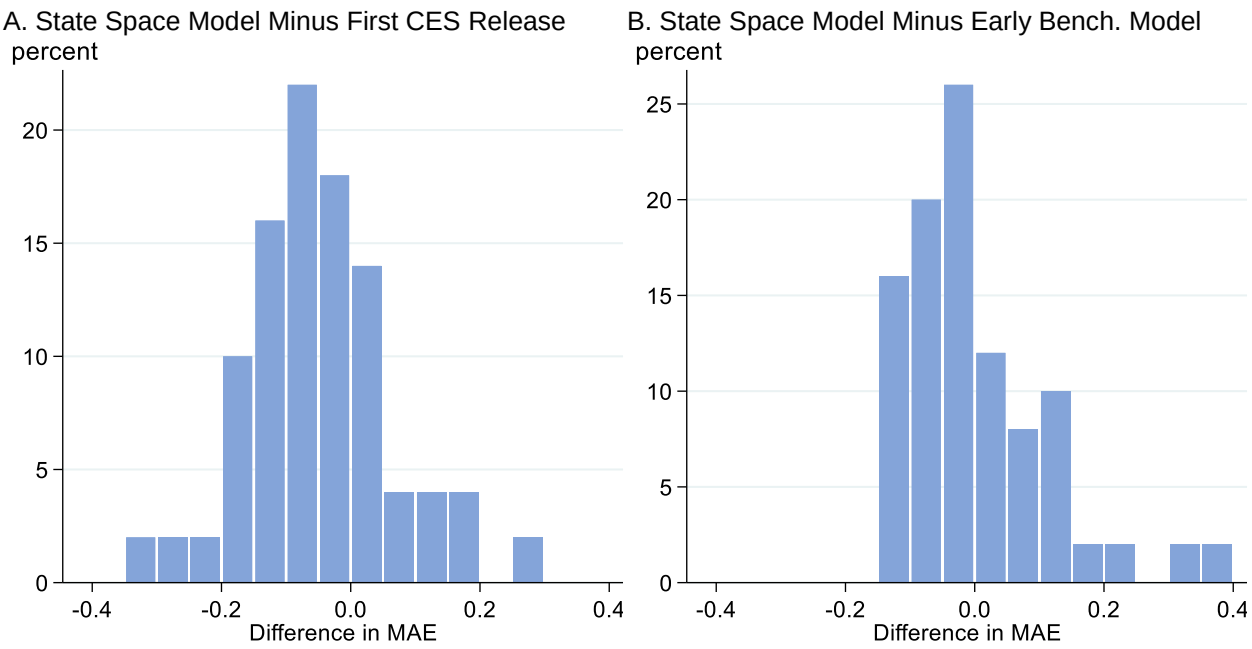

Figure 6. Distribution across states of the difference in mean absolute error (MAE) between the state space model and comparison models. Mean absolute error (MAE) calculations are for all 50 states and are based on end-of-sample 12-month log percent changes for data vintages spanning from March 2005 to October 2018. Bars with negative x-axis values represent an improvement in MAE for the state space model over other predictors. To provide further context for the x-axis scales, the average MAE across states is 0.56 percent for the first CES release and 0.51 percent for the early benchmark model. The mean of the distribution in panel A is −0.048 percentage points and the median is −0.062 percentage points. The mean of the distribution in panel B is 0.00008 percentage points and the median is −0.024 percentage points. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

Panel B of Figure 6 shows how the state space model performs compared to the early benchmark model. The distribution looks very similar to the one in panel A but is shifted to the right, reflecting the fact that the early benchmark model improves on average over the first CES release—though there are 10 states for which it doesn't. The mean of this distribution is 0.00008 percentage points and the median is −0.024 percentage points. The state space model performs better than the early benchmark model for 31 states, but worse for 19. It is important to note here that for 6 of the 10 states where the early benchmark model does worse than the first CES release, the state

space model does better. This fact makes a strong argument for combining the predictions of the state space and early benchmark models, an approach we explore in section 4.3.

4.2 Explaining Model Performance across States and Over Time

The results shown in Figure 6 indicate our model can be valuable in predicting revisions of state employment data. However, its performance varies considerably across states. What can explain these differences? Using regression analysis, we find that four explanations can explain more than half of the variation. They are a state's: 1) size, 2) employment volatility, 3) average benchmark revision size, and 4) tightness of link to national employment.

We construct variables to capture these explanations as follows. State size is average total employment over the sample period of March 2005 to October 2018. Employment volatility is the standard deviation of the 12-month log employment change over the sample period. Average benchmark revision size is the MAE of the first CES release. And the tightness of link to national employment is the natural logarithm of the share of months where the sign of 12-month growth in a state differs from that of the US. That is, it is the share of months where state growth is positive and US growth is negative or vice versa.

To make it easy to compare coefficient magnitudes across our four variables, we standardize each variable to be mean zero, standard deviation one. We then estimate the following linear regression model on our cross-section of 50 U.S. states:

$$Difference\ in\ MAE_s = \alpha + \beta_1 Size_s + \beta_2 Volatility_s + \beta_3 RevisionSize_s + \beta_4 NatStLin_s + \varepsilon_s.$$

We estimate the regression using a robust regression algorithm that reduces the large weight outliers receive in standard OLS regression.[9]

[9] The robust regression algorithm is implemented by the *Stata* (2019) command *rreg*.

Figure 7 shows the estimation results in the form of partial regression plots. The x-axis for a given panel is the value of the respective independent variable conditional on all the other independent variables in the model, and the y-axis is the value of the difference in MAE conditional on all the other independent variables in the model.[10] All four variables make an important contribution to explaining the performance of the state space model relative to the CES first release and together explain more than 60 percent of the variation in the difference in MAE as measured by adjusted r-squared.

The signs of the coefficients in the regression are telling: the state space model does well for bigger states, states with low employment volatility, states subject to large revisions, and states whose employment changes typically follow those of the US as a whole. These results generally make sense in light of the state space model's design. The national employment data typically receive a large loading in the factor model, so the state space model expects CES data that conflict with the national data to be revised away. Figure 5 shows that the state space model also smooths through the unrevised data, which means that it generally treats volatility in the CES data that is not persistent or common to other state level labor market indicators as noise instead of signal. It is no surprise then that the state space model does better for states that typically have lower employment volatility. There is nothing in the design of the model to suggest that it should do better for larger states or states subject to large revisions, but we view the result for states subject to large revisions as a positive one.

[10] Formally, the x-axis in each panel is the residual from a robust regression of the respective independent variable on all the other independent variables and the y-axis is the residual from a robust regression of the difference in MAE on all the other independent variables.

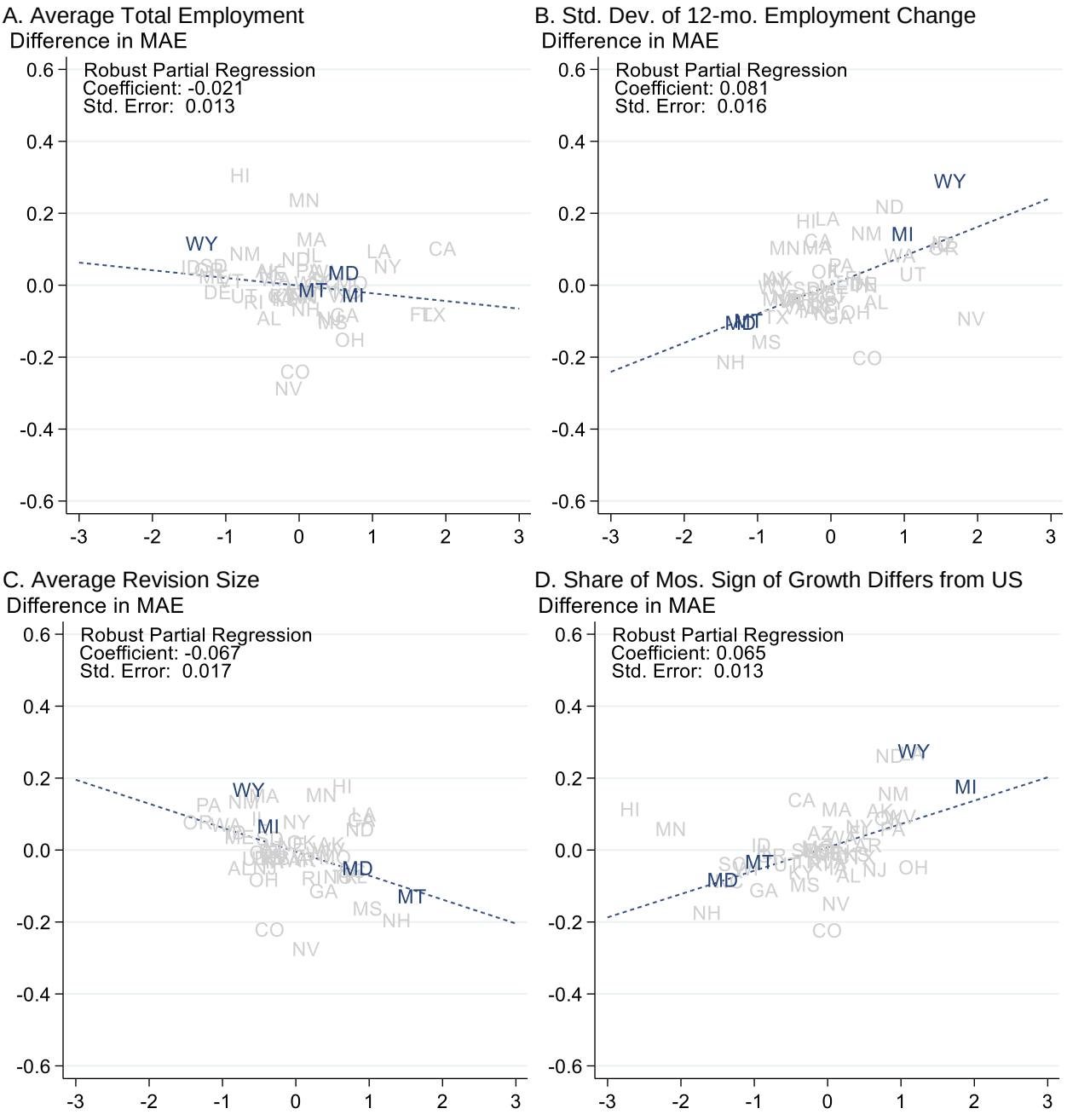

Figure 7. Robust partial regression plots from model of state space model performance. This figure shows partial regression results for four explanatory variables included in a regression model of the difference in MAE between the state space model and the first CES release. The distribution of the difference in MAE across states is displayed in panel A of Figure 6. To account for outliers (particularly Louisiana), we estimate a robust regression. All explanatory variables are standardized to be mean zero, standard deviation one in order to easily compare the coefficients' magnitudes. See the text for details on how the four explanatory variables are calculated. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

It is also important to note that the four variables we use to explain the performance of the state space model are not uncorrelated: larger and less volatile states have smaller revisions, as do states that closely follow the national economy. While this correlation structure is such that for many states, disadvantages for the state space model along one dimension are balanced out by advantages along another, this is not the case for all states. For example, Figure 7 shows that the state space model performs especially poorly for Michigan and Wyoming because their (conditional) employment is volatile, their (conditional) revisions tend to be small, and their (conditional) employment trends tend to differ from that of the US as a whole. For Maryland and Montana, the case is exactly the opposite.

Table 1. Lowest Mean Absolute Error by Model

	Share of states
State Space Model	31
Full model	22
No CPS data	7
Revision model only	2
Early Benchmark Model	15
First CES Release	4

Notes: This table shows the number of states with the lowest mean absolute error for each model of the 12-month log percent change in benchmarked state CES employment. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

Because of the mixed performance of the state space model across states, we estimate two alternate versions of the model that seek to address some of the full model's shortcomings. One alternative acknowledges that a factor structure may not exist for some states and estimates only the

revision portion of the model, with univariate dynamics for ΔE_t.[11] The other alternative omits the CPS data from the factor model because they underwent a noticeable methodological change in 2010. For some states, this change significantly alters the correlation structure with other state level employment indicators and causes the factor structure to deteriorate in a way that ultimately negatively affects nowcast performance.

Taking the performance of our alternative state space models into account, Table 1 summarizes the best performing model across states. The alternative state space models perform best for only a handful of states. If we sum across the three variants of the state space model, jointly it performs the best for 31 states; while the EB model performs best for 15, and there are 4 states where simply sticking with the first CES release works best.

To this point, we have focused on results across states where we aggregate over time. We now turn to how our results evolve over time when we aggregate over states. To do this, we focus on the median state's MAE from the full state space model and the early benchmark model relative to that of the first CES release, a measure which is robust to outlier concerns. Specifically, for the early benchmark model and ours, we calculate the ratio of the median state's MAE in each vintage to that of the first CES release. Formally, the measure is given by:

$$Median\ State\ Relative\ MAE_t = \frac{Median\ State\ MAE\ of\ Comparison\ Model_t}{Median\ State\ MAE\ of\ First\ CES\ Release_t}.$$

By this formulation, a value of less than 1 implies better performance for the state space or early benchmark models relative to the first CES release, and a value of greater than 1 implies worse performance.

[11] See the appendix for further details on this specification.

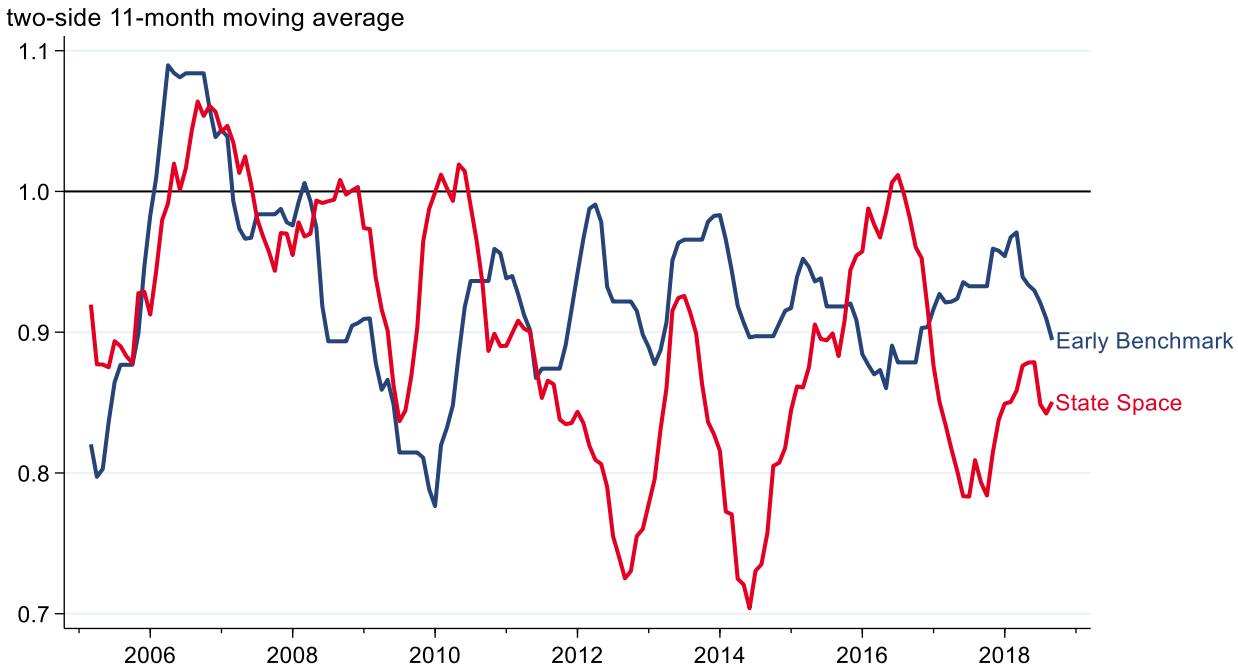

Figure 8. Median state's relative mean absolute error by data vintage. This figure displays, for the state space and early benchmark models, the median state's mean absolute error relative to that of the first CES release. The data are for vintages from March 2005 to September 2018. We use medians rather than means to remove the influence of outliers. To highlight the general trends, we smooth the series using a two-sided 11-month moving average. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

Figure 8 plots the median state's relative mean absolute error for each vintage in our sample period. To highlight broad patterns over time, we smooth the series by taking a two-sided 11-month moving average of it. Figure 8 shows that at times, like in the cross section of states, the state space model does not perform better than the early benchmark model or first CES release. There is, however, a long period (2011–2016) where the state space model either matches or outperforms both.

All told, the results in this section make clear that the state space model works well in a majority of states and much of the time, but not in all states or all of the time. The reasons for the variability in performance are to a large extent inherent to the model's design. For some states, the

state space model's design is a benefit, but for others it's a cost. This is a strong argument for combining predictions of the state space and early benchmark models, which we do next.

4.3 Combining Models to Improve Performance

Because no single model works best across states or time, we now show how averaging across the models at our disposal results in better nowcast performance than any individual model can provide. To do this, we estimate optimal model weights in realtime using a least squares regression that restricts the coefficients on the models at our disposal to be positive and sum to one. We include the first CES release, the early benchmark model, and all three versions of the state space model in the regression. Formally, the regression used to calculate the averaging weights for each state is:

$$\Delta_{12}CESFin_t = \beta_1\widehat{CES1}_t + \beta_2\widehat{EB}_t + \beta_3\widehat{SSFull}_{nt} + \beta_4\widehat{SSNoHH}_{nt} + \beta_5\widehat{SSRevOnly}_t + \varepsilon_t$$

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$$

$$\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\} \geq 0.$$ [12]

To arrive at a single nowcast for each state, we estimate this regression on past model nowcasts and then take its predicted value for each state using current model nowcasts. In this way, our realtime model combination strategy assigns greater weight to models with superior past performance.

We employ two approaches for calculating these weights in relation to a model's historical performance. The *recursive* estimation method uses information on model performance from the entire prior history of each model's performance. It assumes that we are learning about the true, constant best weights across models with each new benchmark revision. The *rolling window* estimation method uses information from the previous three benchmarks. This approach assumes

[12] This framework is similar to one found in Brave and Fisher (2004).

that the best model weights can vary over time. For example, if one model does better during recessions and another does better during expansions, using only the recent data might better capture this.

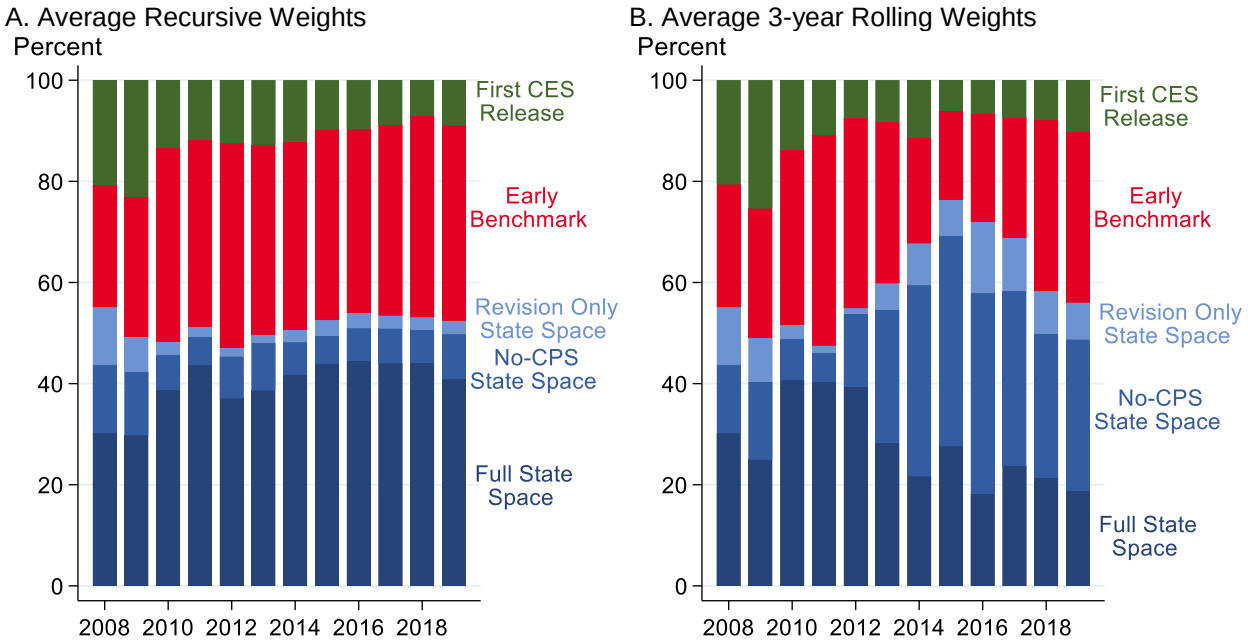

Figure 9. Average across states of recursive and rolling weights by benchmark year. This figure shows the weights obtained from a model averaging regression exercise. Recursive weights are based on the entire prior history of each model's performance and rolling weights are based on each model's performance over the previous three benchmarks. The x-axis dates reflect data available with the release of new benchmarked data in March of a given year. For example, the 2019 bar is for benchmark data available as of March 2019. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

Figure 9 shows the time series of realtime weights in each vintage averaged across the 50 U.S. states for both the recursive and rolling estimation methods. Because the model averaging methods require an initial sample of nowcasts to compute model weights, the sample period runs from March 2008 through December 2018. The recursive weights shown in figure 9 are stable throughout time, reflecting the small month-to-month differences in estimation samples, while the rolling weights vary much more over time. For both methods, the three state space models in total receive at least 50 percent of the weight, with the early benchmark model accounting for the vast

majority of the remaining weight. Interestingly, the rolling weights seem to suggest that the methodological change that occurred in 2010 for the CPS data negatively affected the performance of the state space models. By 2013, when this change has been fully incorporated into the rolling weights, the state space model that omits the CPS data clearly dominates the model that instead includes it. That said, it is difficult to disentangle the impact of the change to the CPS data from the possibility that some models may perform better during recessions and others during expansions.

To summarize the performance of the model averaging and individual model results, we construct a metric based on their mean absolute error (MAE) relative to a naïve benchmark model. Our naive model (referred to below as *Naïve)* is the first CES release.

$$Average\ Percent\ Gain\ in\ MAE = \frac{1}{50} * \sum_{i=1}^{50} 100 * \left(1 - \frac{MAE(Model)}{MAE(Naïve)}\right)$$

Error! Reference source not found. shows values of this metric, with stars indicating a statistically significant average percent gain based on a Diebold and Mariano (1995) test of equal forecast accuracy. To highlight differences in the performance of the model averaging methods over time, we consider the full sample period and a separate sample period that excludes the period of the Great Recession and runs from March 2013 through September 2018.

A few results from the full sample period (column 1) stand out. First, the rolling average and recursive methods perform the best across the 50 US states, reducing MAEs by a statistically significant 14 percent on average over the naïve benchmark. Second, both averaging methods are also a statistically significant improvement over any of the individual nowcast models that we consider. Third, the state space models play an important role in the superior performance of the averaging methods. If we only average across the first CES release and the early benchmark model,

we see a noticeable drop in performance—and this is the case even though the revision-only state space model performs worse than the first CES release.

Table 2. Average Percent Gain in Mean Absolute Error by Sample Period

	2008–18	2013–18
Model Average—3-year rolling window	13.5*	17.3*
Model Average—Recursive	14.0*	16.0*
State Space—Full model	10.7*	9.7*
State Space—No CPS data	8.8*	9.2*
State Space—Revision model only	-4.2*	8.2*
Model Average—Recursive, EB Only	8.6*	7.7*
Early Benchmark Method	9.9*	7.6*
Model Average—3-year rolling window, EB Only	8.4*	7.4*
CES First Release	0.0	0.0

Notes: EB means early benchmark. This table presents the average percent gain in mean absolute error of each nowcasting model of the 12-month log percent change in post-benchmark revision CES state employment relative to the CES initial release for various sample periods. We judge statistical significance from zero, *, at the 95 percent confidence level using the Diebold and Mariano (1995) test of equal forecast accuracy. We apply the test to our panel of 50 states with heteroskedastic and autocorrelation consistent standard errors calculated by clustering on state identifiers and Student's t critical values. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

When we examine the results for the post-Great Recession sample period (column 2), we find that model averaging helps even more during the 2013–18 period, with the recursive and rolling average methods reducing mean absolute errors by around 16 percent and 17 percent on average. This improvement is statistically significant over the first CES release and also when measured relative to any other individual model, including the model average that does not incorporate the state space models. The improvement in performance of the rolling averaging method in the post-Great Recession sample is particularly pronounced. This result likely combines two elements: 1) its

robustness to performance differences across models during recessions and expansions, and 2) its robustness to methodological changes in the data underlying the dynamic factor model like the change in the CPS data structure that occurred in 2010. Both elements make it an example of the robustness of forecast combination strategies to structural changes of the kind described in Clark and McCracken (2009).

5 Conclusion

This paper develops and tests in realtime a state space model of benchmarked US state nonfarm payroll employment that successfully reduces the uncertainty surrounding the initial estimates of employment growth in 36 of 50 states. When compared with previous models of benchmarked state employment data, the model outperforms in 31 of the 50 states. The model explicitly tracks the revision process of the CES data and estimates a single common factor for state employment growth that makes use of additional state level labor market indicators that are not directly related to the CES data or revision process. Much of the model's strength comes from incorporating these additional indicators, suggesting that models of benchmarked state employment data have the potential for further improvement as higher quality realtime data become available.

While much of the literature has focused on forecasting revisions to US national data, this paper contributes to the limited research on revisions to US state and local data. Models of US sub-national data provide two key benefits. First, state and local data are generally subject to larger revisions than national data because of smaller sample sizes, so such models have the potential to be very useful. For the data we examine, the average 12-month growth rate revision is 0.6 percent, while it is only 0.2 percent for the national data. Second, while the data we use are collected following a consistent methodology, they contain a cross sectional dimension that is absent from the

national data. This allows us to test the robustness of other models' specifications and ours across a much larger and more varied data sample than what is available at the national level.

We find that there is indeed sizeable variation in how our model performs across states and that four characteristics can explain more than half of the variation in performance. They are a state's size, employment volatility, average benchmark revision size, and strength of the relationship with national employment. That our model's performance depends on a state's employment volatility and the strength of its relationship with national employment are inherent to the model's design. For some states, the design is a benefit and for others it is a cost. For this reason, researchers interested in producing benchmarked US state employment nowcasts for a particular state may consider adjusting our general model to the specific characteristics of the state. In addition, the nowcast combination method we develop and test in this paper may provide further improvements, as it did in the general case we explored.

6 References

Aruoba, S. B. (2008). Data Revisions Are Not Well Behaved. *Journal of Money, Credit and Banking, 40*(2-3), 319–340.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics, 27*(4), 417–427.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow. In G. Elliot, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vols. 2, Part A, pp. 195–237). Amsterdam, Netherlands: Elsevier. doi:10.1016/B978-0-444-53683-9.00004-9

Berger, F. D., & Phillips, K. R. (1993). Reassessing Texas Employment Growth. *The Southwest Economy*, 1-3.

Berger, F. D., & Phillips, K. R. (1994). Solving the Mystery of the Disappearing January Blip in State Employment Data. *Economic Review*(Second Quarter), 53-62.

Brave, S. A., & Fisher, J. D. (2004). In Search of a Robust Inflation Forecast. *Economic Perspectives, 28*(4), 12–31.

Bureau of Labor Statistics. (2016). *CES Stata and Metro Area Frequently Asked Questions*. Retrieved January 24, 2017, from Bureau of Labor Statistics: https://www.bls.gov/sae/790faq2.htm

Bureau of Labor Statistics. (2017). Employment, Hours, and Earnings from the Establishment Survey. In *Handbook of Methods.* Retrieved January 24, 2017, from https://www.bls.gov/opub/hom/pdf/homch2.pdf

Clark, T. E., & McCracken, M. W. (2009). Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts. *International Economic Review, 50*(2), 363–395.

Coomes, P. A. (1992). A Kalman filter formulation for noisy regional job data. *International Journal of Forecasting*, *7*(4), 473-481.

Crone, T. M., & Clayton-Matthews, A. (2005). Consistent Economic Indexes for the 50 States. *The Review of Economics and Statistics*, *87*(4), 593–603.

Croushore, D., & Stark, T. (2001). A Real-Time Data Set for Macroeconomists. *Journal of Econometrics*, *105*(1), 111–130.

Delgado, M., Porter, M. E., & Stern, S. (2016). Defining Clusters of Related Industries. *Journal of Economic Geography*, *16*(1), 1–38. doi:10.1093/jeg/lbv017

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–265.

Durbin, J., & Koopman, S. J. (2012). *Times Series Analysis by State Space Methods* (2nd ed.). Oxford, UK: Oxford University Press.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, *55*(4), 665–676.

Knotek II, E. S., & Zaman, S. (2014). On the Relationships between Wages, Prices, and Economic Activity. *Economic Commentary*, *2014*(14), 1–6.

Monteforte, L., & Moretti, G. (2013). Real-Time Forecasts of Inflation: The Role of Financial Variables. *Journal of Forecasting*, *32*(1), 51–61.

Orphanides, A., & van Norden, S. (2002). The Unreliability of Output-Gap Estimates in Real Time. *The Review of Economics and Statistics*, *84*(4), 569–583.

QCEW Questions and Answers. (2019). Retrieved from Bureau of Labor Statistics website: https://www.bls.gov/cew/questions-and-answers.htm

StataCorp. (2019). Stata Statistical Software: Release 16. College Station, TX.

Timmermann, A. (2006). Forecast Combinations. In G. Elliott, C. W. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Amsterdam, Netherlands: Elsevier.

7 Appendix

7.1 Berger and Phillips (1993) Early Benchmark Model

Formally, the model is:

$$EB_t \begin{cases} CES_t & if\ 0 \le t < i \\ CES_i \cdot \sum_{t=i}^{t} \left(\frac{QCEW_t}{QCEW_{t-1}} \right) & if\ i \le t < j \\ CES_i \cdot \prod_{t=i}^{t} \left(\frac{QCEW_t}{QCEW_{t-1}} \right) \cdot \prod_{t=j}^{t} \left(\frac{CES_t}{CES_{t-1}} \right) & if\ j \le t < T \end{cases}$$

where i is the first month after the benchmarked portion of the CES data run out and j is the first month after the QCEW data run out.

The QCEW includes some farm workers, which are removed from a state's total employment count. Specifically, Berger and Phillips (1993) remove all employees in the agriculture, forestry, fishing, and hunting sector (NAICS 11) with the exception of those in the logging sector (NAICS 1133). This is the early benchmarked portion of the series. For months when the QCEW data are not yet available, the series is extended by applying the month-to-month growth rates from the non-benchmarked portion of the CES. It is important to note that the non-benchmarked data have different seasonal patterns than the benchmarked data (Berger and Phillips (1994)).

Figure 10 shows an example of the early benchmark model for Illinois's CES data through December 2015, which were initially released in January 2016. The blue series is the initial release, and the dashed portion of the series represents data that were not benchmarked using the QCEW as of January 2016. The early benchmarked series is in red, and is benchmarked using growth rates from the QCEW starting in October 2015. As of January 2016, QCEW data were available through June 2015, so starting in July 2015, the early benchmark model relies on growth rates from the initial

release (the blue line). For this reason, the blue and red series run parallel starting in July 2015. The black series is the March 2016 vintage of the Illinois CES data through December 2015, which is benchmarked using the QCEW through September 2015. It is clear that the early benchmarked series is closer to the March 2016 benchmarked series than the initial release is. It is also important to note that data for the already-benchmarked portion of the data (January through September 2014) changed from as part of the March 2016 benchmark. The BLS revises already benchmarked portions of the data, but (as is clear in Figure 10) the revisions are typically much smaller than the revisions from non-benchmarked to benchmarked data.

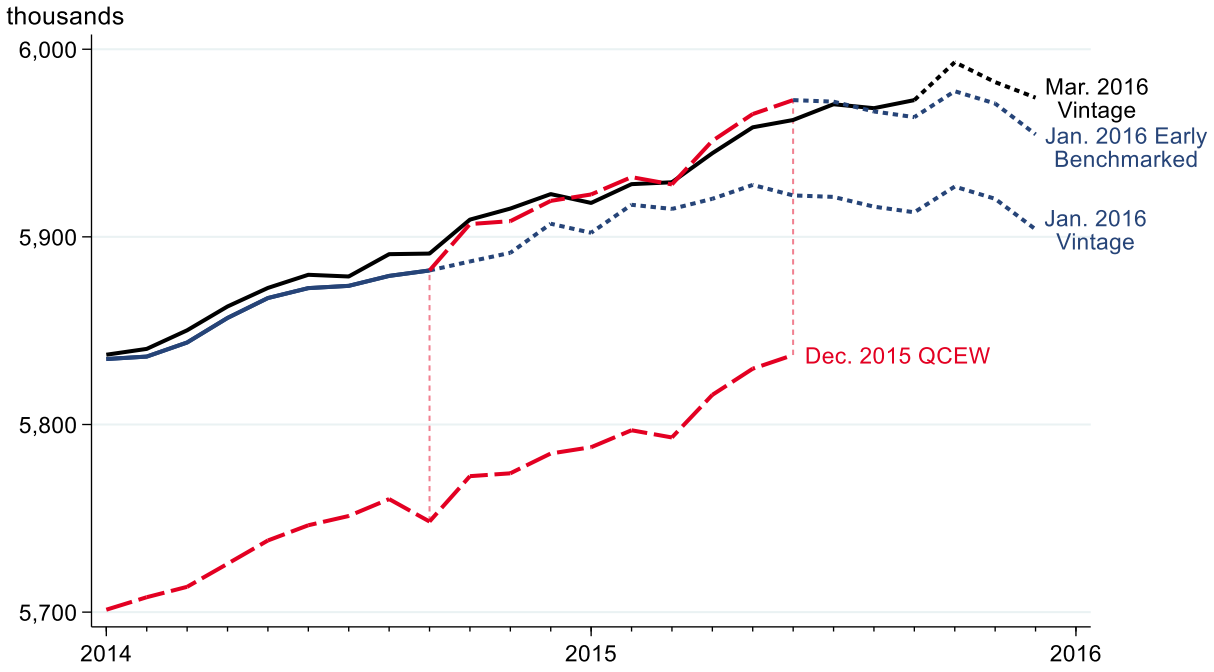

Figure 10. Example of the Berger-Phillips early benchmarking method for Illinois. The short-dash lines represent survey-based data that are not yet benchmarked against administrative data. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.

7.2 Revision-only State Space Model

In section 4, we consider an alternative univariate dynamic specification for the growth rate of E_t with state-specific AR(p) dynamics around a conditional mean α.

$$\Delta E_t = \alpha + \sum_i \beta_i \, \Delta E_{t-1} + \zeta_t$$

The state space representation of this *revision only* model is shown below in companion form.

$$\begin{bmatrix} CESBench_t \\ CES2_t \\ CES1_t \\ QCEW_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} E_t \\ E_{t-1} \\ B_t \\ R_t \\ W_t \end{bmatrix}$$

$$\begin{bmatrix} E_t \\ E_{t-1} \\ B_t \\ R_t \\ W_t \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \kappa \\ 0 \\ \delta \end{bmatrix} + \begin{bmatrix} 1+\beta & -\beta & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau \end{bmatrix} \begin{bmatrix} E_{t-1} \\ E_{t-2} \\ B_{t-1} \\ R_{t-1} \\ W_{t-1} \end{bmatrix} + \begin{bmatrix} \zeta_t \\ \varepsilon_{t+1} \\ \eta_t \\ \omega_t \\ \nu_t \end{bmatrix}$$

The state-specific lag orders for both ΔE_t and W_t are chosen based on minimum BIC values.

7.3 Results by Alternative Performance Measures

Table A1. Average Percent Gain in Mean Absolute Error or Mean Absolute Log Percent Error by Performance Measure

	12-Month Log Pct. Chg.	1-Month Log Pct. Chg.	Endpoint Level
Model Average—Recursive	14.0	30.4	13.9
Model Average—3-year rolling window	13.5	30.8	12.7
State Space—Full model	10.7	26.0	11.9
Early Benchmark Method	9.9	†	10.3
State Space—No household data	8.8	25.5	10.4
Model Average—Recursive, EB Only	8.6	†	8.6
Model Average—3-year rolling window, EB Only	8.5	†	8.3
CES First Release	0.0	0.0	0.0
State Space—Revision model only	-4.2	26.9	-2.0

Notes: EB means early benchmark. This table presents model performance results for the sample period March 2008 to December 2018 according to three measures: average percent gain in mean absolute error for a model's estimate of the 12-month log percent change relative to the CES initial release; average percent gain in mean absolute error for a model's estimate of the 1-month log percent change relative to the CES initial release; and average percent gain in mean absolute log percent error for a model's estimate of the endpoint level relative to the CES initial release. †By construction, the EB model makes the same 1-month log percent change prediction as the initial release. See section 7.1 for a full description of the EB model. Source: Authors' calculations based on St. Louis Fed archive of data from Haver Analytics.