Klein, Tony

Working Paper

Agree to Disagree? Predictions of U.S. Nonfarm Payroll Changes between 2008 and 2020 and the Impact of the COVID19 Labor Shock

QMS Research Paper, No. 2021/07

Provided in Cooperation with:
Queen's University Belfast, Queen's Business School

Suggested Citation: Klein, Tony (2021) : Agree to Disagree? Predictions of U.S. Nonfarm Payroll Changes between 2008 and 2020 and the Impact of the COVID19 Labor Shock, QMS Research Paper, No. 2021/07, Queen's University Belfast, Queen's Management School, Belfast, https://doi.org/10.2139/ssrn.3929635

This Version is available at:
https://hdl.handle.net/10419/271252

QUEEN'S UNIVERSITY BELFAST | MANAGEMENT SCHOOL

Working Paper Series - QMS Research Paper 2021/07

Agree to Disagree? Predictions of U.S. Nonfarm Payroll Changes between 2008 and 2020 and the Impact of the COVID19 Labor

Tony Klein

Queen's University Belfast

23 September 2021

Series edited by Philip T. Fliers and Louise Moss.
To check for updated versions of this paper here.
To subscribe click here.
To submit forward your paper to qms.rps@qub.ac.uk.

Agree to Disagree? Predictions of U.S. Nonfarm Payroll Changes between 2008 and 2020 and the Impact of the COVID19 Labor Shock

Tony Klein[a,*]

[a] *Queen's Management School, Queen's University Belfast, UK*

Abstract

We analyze an unbalanced panel monthly predictions of nonfarm payroll (NFP) changes between January 2008 and December 2020 sourced from Bloomberg. Unsurprisingly, we find that prediction quality varies across economists and we reject the hypothesis of equal predictive ability. In an error decomposition, we find evidence of significantly biased forecasts. Participation rate in the survey is affecting this bias. We find that survey participants under-predict job losses in times of market turmoil while also under-predicting the recovery thereafter, especially during the COVID19 labor shock. For prediction of NFP changes, autoregressive models are outperformed by a deep learning long short-term memory network. However, the consensus forecast yields better forecasts than model-based approaches and are further improved by combining the forecasts of the best performing economists. The COVID19 labor shock is shown to have adverse effects on the prediction performance of economists. However, not all economists are affected equally.

Keywords: COVID19, Employment, Forecasting, Machine Learning, Survey Data

JEL classification: G12, G17, J11

1. Introduction

Nonfarm payroll (NFP) figures and monthly changes thereof are important and immediate indicators of the development of the economy in the U.S., particularly the labor market itself. Published by the Bureau of Labor Statistics (BLS) on a monthly basis, nonfarm payroll represents the number of payroll jobs and its month-to-month changes.

*eMail: t.klein@qub.ac.uk.

The NFP covers most of the non-agricultural industry contributing roughly 80% of the GDP. As such, the monthly development in the labor market is an important precursor to the development and publication of other macroeconomic variables. Monthly NFP releases cause short- and medium term reactions to stock, bond, and FX markets which is documented in literature (Fleming & Remolona, 1999, Dungey et al., 2009, Dungey & Hvozdyk, 2012). The released numbers are perceived with a signaling effect, in particular when released numbers exceed or fall short of (market) expectations. Measuring and correctly quantifying these expectations—as for any micro- or macroeconomic variable—are of relevance in view of their impact and more importantly, their economic implications.

However, research and literature on NFP forecasts, their quality, sampling, and sample composition of forecasters are scarce in general. Forecasts of macroeconomic variables such as GDP growth and inflation—in particular those of the Survey of Professional Forecasters (SPF)—attract much more academic attention. For NFP forecasts, there exists no established nor agreed-on forecast format or expectation measure. Different data providers offer proprietary—and varying—data sets on forecasts and expectations derived from questionnaires of (academic) experts, economists, and other participants of financial markets.

The measure of these expected NFP changes—or a consensus thereof—is of utmost importance to determine surprises by over- or undershooting expectations. We make use of raw data of Bloomberg's qualified economists survey, which collects NFP change predictions in advance of their official publication from from 70 to 100 mainly U.S. and EU/UK based economists and academics. This paper is one of the first studies that utilizes this set of NFP forecasts by a heterogeneous set of economists, in particular in terms of forecaster bias and shocks within the Davies & Lahiri (1995) framework. The collated individual forecasts of NFP changes are dissected as unbalanced panel of forecasters. Focusing on the anatomy of forecast errors, we isolate temporal shocks which affect all forecasters equally. These shocks generally translate to the difference of expected nonfarm payroll figures to actual published ones. This quantification of over- and underestimation of expectations is analyzed in detail. In addition, the utilized framework also offers a

measure for a systematic idiosyncratic error of each forecaster which is used to address the question of forecasting quality of new-joiners and leavers in some relation to the findings of Clements (2021). Further, we directly address the question of equal predicting ability of this group of predictors with the relative measures of D'Agostino et al. (2012) and find that some individual forecasters outperform while others systematically underperform. As all forecasters are financial professionals, academics, and market participants from institutions worldwide, the obtained data offers a viable and informed cross section of expectations for the macroeconomic variable at question—nonfarm payroll changes.

NFP figures are published on a monthly basis on the first Friday of each month. These figures include the numbers for the current period as well as revisions on previously published NFP figures. NFP publications for the most recent month base on an incomplete survey as not all businesses have yet reported their employment numbers. Roughly 70% to 75% of responses are available for the first release, while the two months later, the collection rate is between 90% and 95%. Hence, these publication numbers are regularly revised in subsequent monthly publications to account for additional responses pertaining to an earlier period and to overcome nonsampling bias. This poses some challenge to analyzing the quality of forecasts as the target variable might change in subsequent months. We provide evidence that economists in the Bloomberg survey tend to systematically under-predict these more precise and updated NFP figures compared to its first release. Arguably, measures of surprise should be based on the first (but likely incomplete) release given that these numbers are new information anticipated by market participants, causing an immediate reaction due to a possible mismatch of expected and realized value. As such, the first release might be the most important one in terms of impact on financial markets. However, as almost all NFP publications are prone to changes based on updated surveys, we also consider the NFP figure based on the most complete survey which is usually the third release. All those figures are seasonally un-adjusted. Accounting for seasonality, we further include the *most recent* release which is seasonally adjusted as an additional benchmark for prediction quality.

As our observation sample spans the worldwide spread of COVID19, we also examine

its impact on the U.S. labor market with particular focus on the error decomposition and effect on prediction quality. This helps quantifying the uncertainty and inconclusiveness of economists around this labor event and its adverse effects on financial markets. We find that economists fail to predict the true dimension of these job losses and additionally, the rapid recovery. Ultimately, this yields shock or surprise measures several magnitudes larger than historic values, which aligns with the extreme market swings and volatility in equity and fixed income markets observed during this first COVID wave.

The remainder of this paper is structured as follows. Section 2 summarizes and systematizes existing literature on forecasts of macroeconomic variables and concepts relevant to this work. Section 3 introduces nonfarm payroll data in more detail, while we distinguish between the Establishment Survey Data in Subsection 3.1.1 and the Qualified Economist Survey outlined in Subsection 3.3. The applied methodology of error decomposition and prediction quality is detailed in Section 4. Findings are presented in discussed in Section 5 while the impact of COVID19 is analyzed in detail in its last subsection. Section 6 concludes this work.

2. Literature Review

Nonfarm payroll publications are an important indicator of the employment situation in the U.S. and affect equity, fixed-income, and FX markets not only locally but also in global financial markets. Edison (1997) provides early evidence on the effect of nonfarm payroll surprises on exchange rates where positive surprises yield an appreciation of the U.S. Dollar. These findings are further extended in Fleming & Remolona (1999) who describe the relationship between U.S. bond prices and employment data. Bond price shocks are linked to publications of employment data that are shown to be the strongest contributor to shocks. Ramchander et al. (2003) describe the significant relationship between surprises in several macroeconomic indicators, including NFP changes, on the volatility of money market instruments. The fact that these monthly changes in NFP numbers do not only affect the short end of the yield curve is further explored in Dungey et al. (2009). It is shown that movements across maturities of the U.S. term structure

can be traced directly to the difference in expected and published NFP numbers, which trigger a jump in bond prices. Dungey & Hvozdyk (2012) extend these findings to jumps in high frequency data of other asset classes which are triggered by surprises in the NFP releases. Gregory & Zhu (2014) address the predictive quality of the Bloomberg consensus forecast for *private sector* NFP predictions in comparison to the informational content of the Automatic Data Processing (ADP) report, as an additional data provider, but do not focus on individual contributions nor on overall NPF changes due to the nature of the private sector ADP data published two days prior to the official publication of NFP figures. It is found that the Bloomberg consensus forecast carries as much informational content as the ADP data and both are useful in predicting NFP changes.

Overall, there is clear evidence that NFP releases cause intraday reactions as well as short- to medium-term movements of financial markets. The magnitude and direction of these reactions is directly related with the *surprise* caused by the mismatch of expected and published numbers. The framework of Davies & Lahiri (1995) in which we dissect this forecast error distinguishes between temporal shock, bias, and idiosyncratic error. This offers an additional view on rationality of forecasters as shown in Isiklar et al. (2006), Lahiri & Sheng (2010), and Dovern & Weisser (2011).

When it comes to the analysis of macroeconomic forecasts with regard to prediction quality, it is usually the Survey of Professional Forecasters that is put in focus of academic discussion. Montgomery et al. (1998) analyze the quarterly SPF predictions for unemployment rate and identify an asymmetric behavior of unemployment rate itself which also has an effect on forecasters. Prediction quality and forecasting model performance differs in economic expansions and contractions. Similar results are found by Koop & Potter (1999), who document sudden negative shocks to U.S. unemployment rates that are followed by gradual increases featuring a strong asymmetry. This will be of particular interest to the discussion of forecaster performance during and after the COVID19 shock to the labor market.

Capistrán & Timmermann (2009a) address the rationality of forecasters and identify biased forecasters in inflation forecasts of the SPF. A positive serial correlation in forecast

errors is found. In addition to the presence of bias, Capistrán & Timmermann (2009b) further ascertain the frequent entry and exit of experts as a complicating factor while this fluctuation of participants requires attention in combination forecasts. In a recent study, Clements (2021) renews this evidence by showing that joiners of the SPF inflation seem to be less accurate, attributed to individual effects. However, there seems to be no difference for GDP predictions across joiners and leavers. We make use of these findings and focus on a possibly differing forecasting ability across participation rates in the Bloomberg survey. For SPF inflation data, Rich & Tracy (2010) find a positive association of disagreement in forecasts and the level of inflation. We confirm these findings for NFP numbers, in particular during the COVID19 labor shock.

D'Agostino et al. (2012) propose a normalized error statistic which accounts for the unbalanced nature of some panels that source from individual predictors and varying response numbers. Further, a bootstrapping approach to determine error percentiles is suggested to address the hypothesis of equal predicting ability across forecasters. We make use of this approach in a two-fold manner. Firstly, we determine if all forecasters have similar ability to forecasting NFP changes; if a certain percentile of forecasters shows lower (higher) prediction errors outside of the confidence interval of bootstrapped values, it is assumed that the hypothesis of equal forecasting ability is rejected if these percentiles pertain to the best (worst) performing forecasters. Secondly, we make use of the findings of Brown et al. (2008) that are of high relevance to this paper. Brown et al. (2008) find that prediction quality of economists in Bloomberg surveys are usually persistent and that some conditional consensus forecasts are better than the mean survey prediction. Hence, we construct NFP predictions conditional on previously *best performing* subsets of forecasters based on percentiles constructed with the methodology of D'Agostino et al. (2012). Clements (2020) applies D'Agostino et al. (2012) methodology on SPF histogram and point forecasts of GDP growth rates and the deflator and finds differences in forecasting ability of the SPF participants. However, Demetrescu et al. (2021) raise the issue of a time-varying effect in forecasting ability and by not accounting for this phenomenon, tests for forecasting ability might be biased.

A relatively recent strand of literature focuses on the prediction quality of exogenous factors, in particular those linked to an individuals' employment and economic situation. Vosen & Schmidt (2011) use different approaches to forecasting private consumption and finds that models including categorized Google search volume outperform survey forecasts. In D'Amuri & Marcucci (2017), it is shown that monthly U.S. employment rate forecasts are dominated by models based on Google search indices, which outperform conventional models. This is confirmed in Maas (2020) who presents evidence on the short-term usefulness of Google search data for job market growth, however, this usefulness decreases with longer forecasting horizons. Similar findings are presented in Borup & Schütte (2020) where Google search activity outperforms macroeconomic forecasts for future employment growth. In a more general setting, Kotchoni et al. (2019) show for employment growth among other macroeconomic variables, data-rich models help forecasting in the long-run but in the short run, simple univariate models perform reasonably well.

Reactions to macroeconomic shocks of survey participants is an important issue when examining forecasting quality and analyzing the dissection of error and bias. Coibion & Gorodnichenko (2012) show that mean forecasts of SPF data fail to completely adjust on impact to shocks. There is a significantly delayed response of economists to including shocks in their expectation formation process. This is caused by information rigidities, which for a broader setting is shown again in Coibion & Gorodnichenko (2015). In relation to this reaction to shocks and information rigidities, the revisions of macroeconomic variables might also play a role in the expectation formation process. Beckmann & Czudaj (2020) further demonstrate that the expectation formation process features spillovers across variables. Clements & Galvão (2021) show, based on SPF data, that data revisions affect and contaminate expectation shock estimations. This is relevant to this paper as NFP numbers are usually revised three times, including a seasonal adjustment. This certainly affects how expectation shocks are quantified and processed by participants in terms of the available and individual information set.

In addition to the error decomposition analysis, we compare the predictive quality of

the Bloomberg survey and its individual economists with time-series models as well as a deep-learning based network. For the network approach, we utilize a Long Short-Term Memory (LSTM) network of Hochreiter & Schmidhuber (1997) to reveal dependencies within the NFP data structure. Krauss et al. (2017), for example, show that deep learning applications are effective in times of market turmoil. Overall, we find the LSTM network to provide superior in-sample fit on the data and to outperform the autoregressive models and even the consensus forecasts. However, the out-of-sample prediction quality of the LSTM in ochanges does not benefit from this application.

3. Data

3.1. Nonfarm Payroll Data Sources and Revisions

3.1.1. The Establishment Survey Data

Nonfarm payroll data are obtained from the Employment Situation Summaries (ESS) of the U.S. Bureau of Labor Statistics (BLS). We focus on these monthly reports published between January 2008 and December 2020. The ESS consists of two separate survey parts. The first section bases on *Household Survey Data* which is not addressed in detail in this research. The second section of the ESS reports on the *Establishment Survey Data* which sources its data from private, local, state, and federal businesses. This monthly survey spans roughly 700 000 worksites from 145 000 businesses and government agencies across the USA.[1] The survey reports how many employees are on payroll based on responses from each work site for the pay period including the 12th of each month. The ESS is made public every first Friday of the month at 8:30 AM (EST).[2] The release is highly anticipated as the report serves as a regular indicator for the state of the U.S. economy, pooling information on employment and unemployment numbers. However, as with most surveys, the provided information is prone to several sources of bias (Bureau of Labor Statistics, 2020), most importantly the *nonsampling* bias. Following corrections of the

[1]Noteworthy, roughly 40% of businesses have fewer than 20 employees. More information on industries and areas included in the Current Employment Statistics—and their rotation—are found on `https://www.bls.gov/ces/`.

[2]If Friday is a bank holiday, the report is usually published on Thursday.

nonfarm payroll numbers in the subsequent months are revisions which correct preliminary values. In the next subsection, we address these systematic revisions in detail and highlight their importance to the economist surveys sourced from Bloomberg.

3.1.2. Nonfarm Payroll Revisions

As previously outlined, the deadline for reporting payroll information to the BLS is the 12th of each month and refers to the payroll period including the deadline date. After approximately three weeks, the BLS then reports the so-called *first* preliminary estimate for this month. This relatively short window to report is a contributing factor to the nonsampling bias. Respondents might fail to report numbers in time. Other causes for this bias are incorrect reporting by respondents, errors during collection and processing of the data, or sectoral clustering of non-responses. In order to overcome this bias, respondents are asked to report corrections to preliminary responses in the next survey. Each EES contains the first estimate or release for the current month, a revision for the previous month—the so-called *second* release—and a revision of the payroll information two months ago—the so-called *third* release. Hence, each month is revised twice resulting in three possibly different estimates in three consecutive reports. After these two revisions, the survey of this month is considered final.

For example, the December 2019 nonfarm payroll is published on January 10, 2020 and reports the first estimate for December 2019 (NFP +145 000). In the January 2020 ESS released on February 7, 2020, the December 2019 NFP is revised by +2 000 to +147 000, which is the second release for this month. The third revision is published in the February 2020 ESS on March 6, 2020, and revises the NFP by +37 000 to +184 000 which is now the third—and final—release for NFP changes in December 2019.

Lastly, all labor data is benchmarked annually and realigned with unemployment data as well as seasonally adjusted. This yields a final NFP change figure which is referred to as *most recent* in the available data set.[3] This further extends the available data structure as two preliminary and seasonally unadjusted figures (first and second release), a final yet

[3]For further details, we refer to the benchmark documentation of the BLS found at `https://www.bls.gov/web/empsit/cestn.htm#section7`.

seasonally unadjusted third release, and finally a seasonally adjusted and benchmarked most recent release.

3.2. Nonfarm Payroll Data

We obtain raw data on monthly vintages for nonfarm employee numbers from the Federal Reserve Bank Philadelphia from November 2007 published the following December to December 2020.[4] From this matrix, we extract the first, second, third, and *most recent* changes to NFP.[5] For our data set, the first observation refers to first release of NFP changes for December 2007 published on January 4, 2008. This starting date is chosen as a compromise between availability of individual predictions with a certain degree of coverage in the unbalanced panel of survey participants (see Section 3.3) and features of the series itself.

Figure 1 visualizes the response rate (or completion rate) averaged for the January to December survey each year for the first release and the subsequent two revisions. It is apparent why the first and second release of NFP data are considered preliminary by the BLS. For the first release, average completion rates range between 66% and 78%, while we observe the highest rates in the years 2013–2015. In recent years, completion rates are below 75%. This translates to the released NFP numbers being based on incomplete first surveys as only roughly 3 out of 4 employers reported numbers thus far. Hence, released numbers are prone to different biases as discussed before. The second release in the subsequent ESS appears to be much more reliable as completion rates range between 85% (2007) and 95% (2014). Recently, second release completion rates are around 90%. The third release—considered final before the regular updates such as seasonal adjustments—incorporates responses of 90-97% of businesses in the poll. These differences in response rates are the main contributing factor to the variation in NFP numbers across their

[4]This raw data is shaped as a (158×158) upper triangular matrix $\boldsymbol{A} = (a_{i,j})_{i,j=1}^{158}$ with the first releases on its main diagonal ($\{a_{i,i}\}_{i=1}^{158}$), the second releases on the minor diagonal to the right of the main diagonal ($\{a_{i,i+1}\}_{i=1}^{158}$) and so forth. This matrix is then used to calculate changes in the NFP resulting in a (157×157)-dimensional diagonal matrix of month-to-month changes $\tilde{\boldsymbol{A}}$.

[5]This data is publicly available at `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/employ`. We note that the data on the first, second, third, and most recent release of NFP changes is readily available from the same source.

releases.[6]

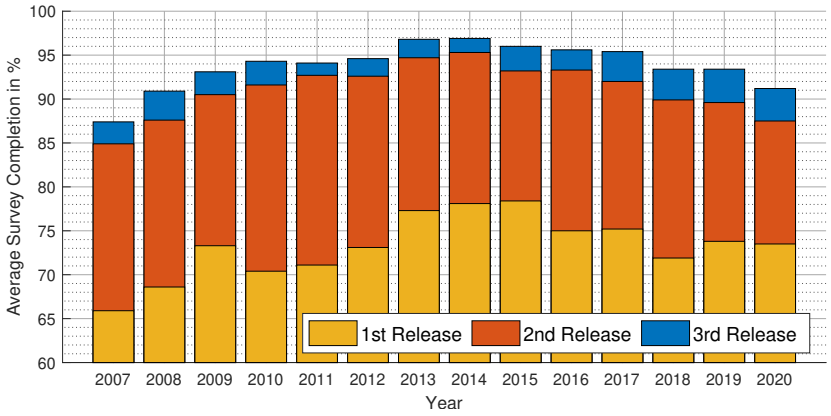


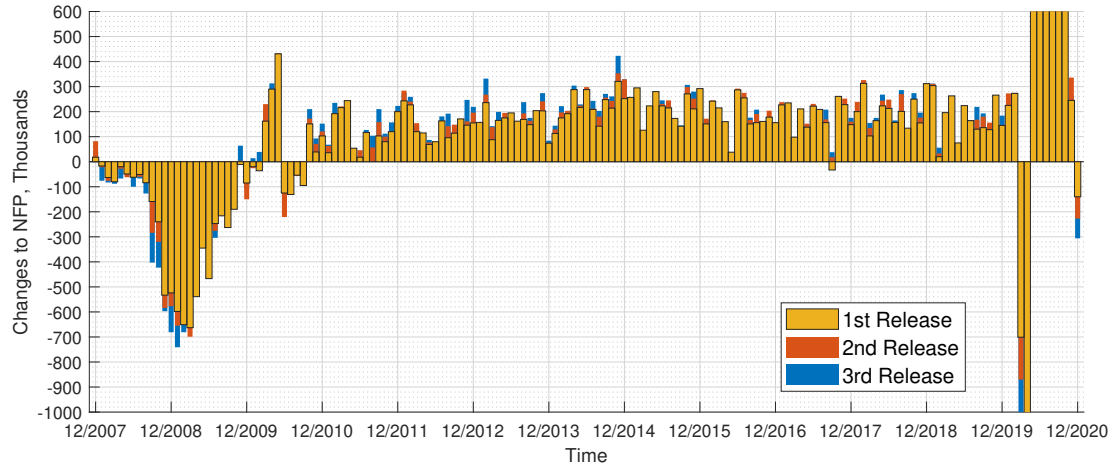
Figure 1: Averages over all months per sample year of the BLS collection rates across first, second, and third release.

Figure 2 visualizes the first, second, and third release in our data set.[7] Evidently, almost all first releases (yellow) are revised in the second (orange) and third (blue) release. In 61.8% of releases, the absolute value of changes is corrected upwards from first to second release. From second to third release, the correction is observed in 62.4% of releases. More important to our analysis of the surveys is the difference of the final estimate—before seasonal corrections—to the first preliminary release as we compare the performance of Bloomberg's qualified economists to the preliminary NFP as well as final numbers. This upward correction of absolute values between first and final release happens in 61.8% (97 out of 157) of all observed monthly NFP releases in our sample.

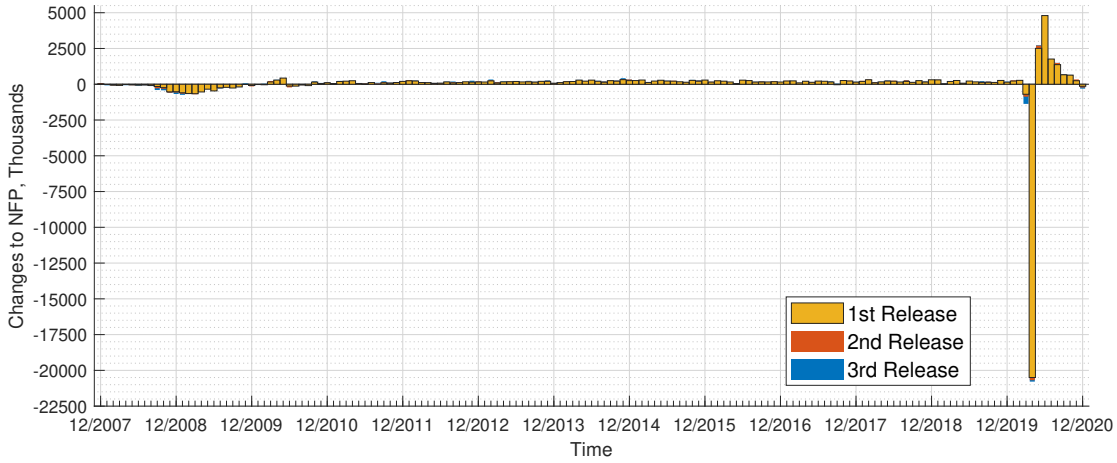
In five instances, the sign of the NFP change is revised from either increase to decrease or vice versa from the first to the third release. A recent example of such a significant revision is September 2017 with the first NFP change release of −33 000, the revised second release corrected to +18 000, and the third final release of +38 000, yielding a total correction from first to third release of +71 000.

[6]Data for response rates is available at `https://www.bls.gov/web/empsit/cesregrec.htm`.

[7]We note that a similar figure published by the BLS compares first to third to *newest* estimate here: `https://www.bls.gov/web/empsit/cesvininfo.htm`. However, given the nature of the predictive survey data described in the next section, we make use of the closest available data for each prediction survey.



(a) *limited ordinate for visibility of changes prior to COVID19*



(b) *unlimited ordinate for visibility of the COVID19 impact*

Figure 2: Nonfarm payroll changes in thousands as published by the BLS in the first release (yellow), second release (orange), and third release (blue) for December 2017 to December 2020 for differing scaling of the ordinate.

3.3. Nonfarm Payroll Survey Data (QES)

We source monthly survey data on month-to-month changes of the nonfarm payroll from Bloomberg—which we refer to as qualified economist survey (QES). We note that the SPF also offers predictions of nonfarm payroll employment on a quarterly resolution.[8] Due to the nature of the applied framework here, we opt for a larger set of predicting economists on a monthly publication schedule of the Bloomberg survey. In what follows, we address the one-period ahead forecasts in nonfarm payroll *changes* simply as *NFP* data predictions. We process the responses of $k^* = 239$ individual economists who predict these

[8]The SPF processes quarterly NPF predictions of around 30 professional forecasters and offers average monthly changes only.

changes roughly five to seven trading days ahead of the official first release of the BLS.[9] Usually, there is one submitting individual per institution per period. However, this submitter might change throughout the sampling period. Ultimately, the cleaned data set consists of $k = 181$ submitting accounts. Details on additional cleaning of the raw data set is found in Appendix A.1.

The cleaned QES data is visualized in Figure 3. From the left-hand side plot in this figure it becomes apparent that not all economists enter a prediction for all months or stop doing so completely while for others, responses begin later as they are rotated into the qualified economists group by Bloomberg. The plot on the right-hand side of Figure 3 visualizes the sorted number of responses. Evidently, we face a heavily unbalanced panel of response data prone to participation or non-response bias.

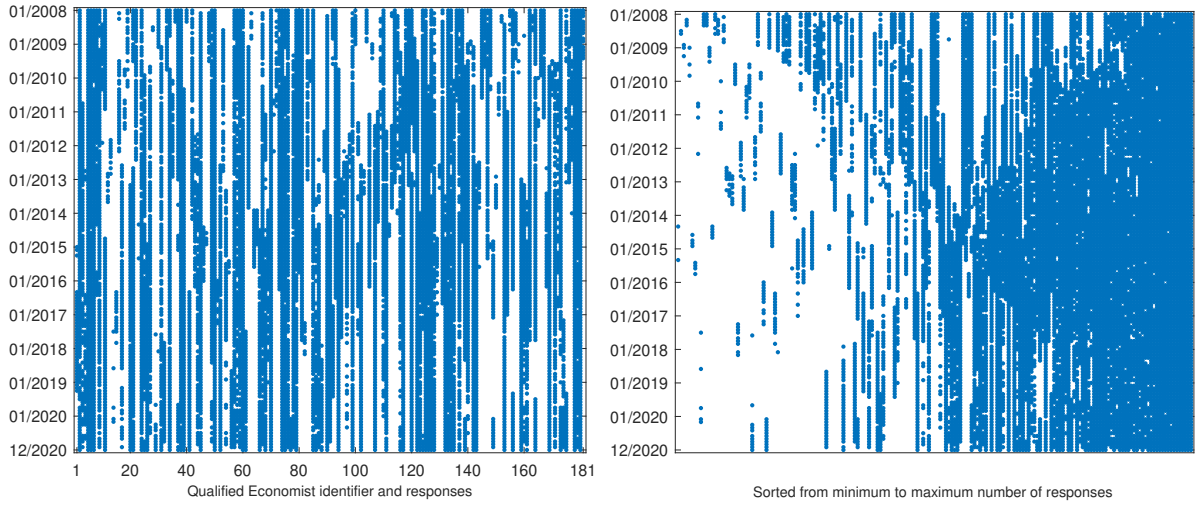


Figure 3: *Left:* Raw data of responses of Bloomberg's *Qualified Economists* ($k = 181$) with numerical identifiers to predict monthly changes ($n = 157$) in nonfarm payroll employment for the ESS releases of January 2008 to December 2020. Each blue square is a response for the respective survey month. *Right:* Economists sorted from minimum to maximum number of total responses.

Figure 4 visualizes a histogram of responses (NFP predictions) per economist/submitter (left-hand side plot) and the total number of responses for each NFP change survey month (right-hand side plot). Note that the maximum number of possible responses is 157. Fewer than 50 responses are recorded for 78 economists. Thirty-seven economists predict 50 to

[9]The raw data set consists of a unique identifier, the full name, the institution, the response date, and the response value for each economist.

100 surveys while 66 economists participate in more than 100 monthly surveys. In relative terms, 89 economists predict more than 50% of the sampled months.

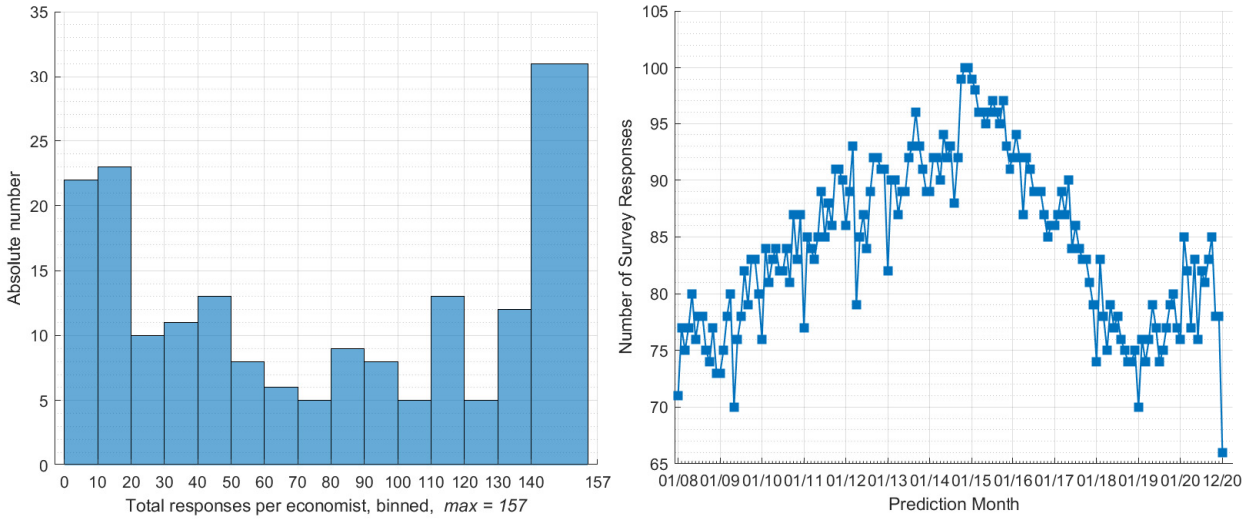


Figure 4: *Left:* Histogram of survey responses per economist, corresponding to the cardinality of T_i^*. *Right:* Number of responses from Bloomberg's *qualified economists* recorded for each NFP release from January 2008 to January 2021, which corresponds to the cardinality of the index sets I_t^*.

We further deepen the analysis by implementing an exclusion threshold of the number of predictions made by an economist and compare the prediction performance across subgroups of economists with more than 50 as well as more than 100 predictions. By comparing these groups, we address the question if regular participants perform differently from new joiners and leavers. We discuss the effect of this exclusion on the error decomposition outlined in Subsection 4.1 in detail.

3.4. COVID19 and NFP data

From Figure 2, it becomes apparent how extreme the job loss in March and April 2020 is. The dimension of job losses and subsequent recovery in in the following months is unprecedented in the history of the U.S. labor situation. Due to this *black swan*-like labor event, we split the sample and separate the analysis of the impact of COVID19 on NFP prediction error decomposition, the QES, and model- and network-based predictions. The split is introduced after the publication of the February 2020 figures, which coincides with the time the COVID19 pandemic started spreading and impacting economies world-wide. Hence, the non-COVID19 sample set ends with February 2020 NFP changes, effectively reducing the number of monthly NFP changes to 147. We later make use of the full sample

of 157 monthly NFP observations and discuss implications of such extreme events on the frameworks outlined in this work, in particular in view of calculation of idiosyncratic bias and shocks across predictions made by economists for these months.

4. Methodology

4.1. Average predictions and error decomposition

We adapt identical nomenclature as in Davies & Lahiri (1999) and denote the actual published nonfarm payroll figure in month t with A_t for $t = 1, \ldots, T$. The prediction made by economist i for time t, with $i = 1, \ldots, k$, is referred by $F_{i,t}$. In what follows, we adapt the framework of Davies & Lahiri (1995), subsequently extended in Davies & Lahiri (1999) and Davies (2006), among others. Note that our data structure implies $h = 1$ without adjunct or overlapping prediction horizons. Hence, the present model framework is a special and less complex case of the original framework of Davies & Lahiri (1995).

The number of predictions for each month varies as shown before. We define and index set $I_t^* := \{i = 1, \ldots, k | F_{i,t} \in \mathbb{R}\}$ which contains the indices (identifiers of the economist in the QES) of all numerical predictions that were made for month t. Hence, the average prediction $\overline{F}_t$ for month t is defined as

$$\overline{F}_t = \frac{\sum_{i \in I_t^*} F_{i,t}}{|I_t^*|}, \tag{1}$$

where the cardinality $|I_t^*|$ translates to the total number of predictions made for month t. Analogously, we define the set $T_i^* := \{t = 1, \ldots, T | F_{i,t} \in \mathbb{R}\}$, which contains the indices of all predictions made by economist i in the QES. Its cardinality then refers to the total number of predictions for forecaster i.

Davies & Lahiri (1995) formulate a decomposition of the individual forecasting error at time t, $A_t - F_{i,t}$, into three components. These components refer to a temporal shock affecting all forecasters with the same magnitude, a forecaster-specific bias which might vary across forecasters but remains constant for the observed period, and an idiosyncratic

error for a specific point in time for each forecaster. This decomposition is formalized as

$$A_t - F_{i,t} = \lambda_t - \phi_i - \varepsilon_{i,t} \tag{2}$$

for forecaster $i = 1, \ldots, k$ and time $t = 1, \ldots, T$. The idiosyncratic forecaster bias over all predictions is referred to by ϕ_i. If $\phi_i > 0$, then the predictor systematically over-predicts NFP changes and vice versa. This bias varies across predicting economists. Cumulative shocks are modeled by including λ_t which affect all predicting economists. Cumulative shocks describe the divergence of the information set at the prediction time and the actual NFP figures published at a later date, translating to an unforcastable component. If $\lambda_t > 0$, then a positive shock occurred in t translating to actual NFP figures being higher then predictions adjusted for idiosyncratic bias. In our data setting, this cumulative shock might be affected by nonsampling bias and revisions made across monthly publication. The idiosyncratic error of economist i at time t is denoted $\varepsilon_{i,t}$.

Following Davies & Lahiri (1995), the expected values of the prediction error components can be estimated by calculating

$$-\hat{\phi}_i = \frac{1}{|T_i^*|} \sum_{t \in T_i^*} A_t - F_{i,t}, \tag{3}$$

$$\hat{\lambda}_t = \frac{1}{|I_t^*|} \sum_{i \in I_t^*} A_t - F_{i,t} + \hat{\phi}_i, \text{ and} \tag{4}$$

$$\hat{\varepsilon}_{i,t} = -A_t + F_{i,t,h} - \hat{\lambda}_t + \hat{\phi}_i, \tag{5}$$

for all $i \in \{I_t^*\}$ and all $t \in T_i^*$, which translates to all prediction made by economists. Note that our notation slightly varies from Davies & Lahiri (1995) as we work on a strictly unbalanced panel, where not all economists make a prediction for every month, yielding $\sum_{i=1}^{k} |T_i^*| = \sum_{t=T}^{k} |I_t^*| =: M < kT$. In fact, we have $M = 13\,241$ predictions of $kT = 28\,417$ total possible economist-month predictions.

Similar to Davies & Lahiri (1999), we have to compress the error variance-covariance

matrix Σ which reads in its original form without compression induced by missing data

$$\Sigma = \begin{pmatrix} A_1 & B & B & \dots & B & B \\ B & A_2 & B & \dots & B & B \\ \vdots & & & & & \\ B & B & B & \dots & B & A_k \end{pmatrix}_{kT \times kT},$$

where $A_i = \sigma^2_{\varepsilon_i} I_T + \mathbf{B}$. The variance of the idiosyncratic error for economist i is denoted by ε_t and I_T refers to the identity matrix of dimension $T \times T$. The matrix $\mathbf{B}$ then comprises of submatrices $\mathbf{b}$ to $\mathbf{g}$ that describe covariances across targets and horizons.[10] Given the simple structure of our data without any overlapping prediction/publication windows and a forecast that only predicts one period ahead, $\mathbf{b}$ degenerates to a scalar while matrices $\mathbf{c}$ to $\mathbf{g}$ become zero. The compressed variance-covariance matrix Σ is then used in a generalized method of moments estimation to determine the standard errors of the error components, in particular the idiosyncratic bias ϕ_i.[11] We apply the framework of Davies & Lahiri (1995) of Eq. (2) with the error component estimators in Eq. (3)-(5) on the full sample with and without the inclusion of the COVID19 labor shock.

It is noteworthy that the prediction horizon of the economists is not exactly one-period ahead. The prediction is made at a time between the months t and $t+1$ for the publication in $t+1$, and the submission time varies across economists. This implies that the information sets vary across economists given the difference in time of the survey entry. It also implies that the economists have a larger information set than the following model approaches that strictly base their predictions on the information set at time t, Ω_t. It is not within the scope of this work to determine the value of this informational advantage of the survey, in particular in view of individual information rigidities.[12]

[10] An overview is found in Davies (2006), p. 385.

[11] For reasons of brevity, no further details on the methodology are included as Davies & Lahiri (1999), Section 6, describes all necessary steps for a similar missing-data and compression problem.

[12] I thank an anonymous reviewer for raising this.

4.2. Model-based predictions

Time series models, in particular of autoregressive structure are widely applied to model and predict macro-economic variables, if stationarity has been confirmed. For example, Ang et al. (2007) include Autoregressive Moving Average (ARMA) models in their prediction analysis of inflation rate surveys. In the vein of the usual notation in time series analysis, we set $y_t := A_t$.

The Autoregressive Moving Average model—in general ARMA(p, q) notation (Box & Jenkins, 1976)—is defined as

$$\phi(L)y_t = \mu + \theta(L)u_t, \tag{6}$$

where $\phi(L)$ describes the autoregressive lag polynomial and $\theta(L)$ the moving average lag polynomial. The disturbance u_t has zero mean. We test different combination of p and q following the Box-Jenkins approach and find $p = q = 1$ to feature the lowest BIC.[13] Hence, we subsequently set $p = q = 1$, such that Eq. (6) reads

$$y_t = \mu + \phi_1 y_{t-1} + \theta_1 u_{t-1} + u_t.$$

The ARMA model is compared to the mean forecasts in terms of in-sample fit and out-of-sample prediction performance given a loss function, outlined in subsequent sections. For the out-of-sample prediction, the model is trained on an extending training window used to produce one period-ahead forecasts, y_{t+1}, of NFP changes. That is, we make use of all available data of NFP changes to predict the next NFP change figure for every time period in the out-of-sample period.[14]

4.3. Predictions based on deep learning

We use the *long short-term memory* (LSTM) network as an example of applied deep learning. As RNNs suffer from some degree of memory loss across longer dependency

[13] Additionally, we compared ARIMA$(1, d, 1)$ and seasonal ARIMA (sARIMA) models, which all showed a lower quality of fit for the job market figures, be it First of Most Recent release.

[14] An alternative to this expanding window is a rolling window estimation of model parameters, in which the size of the training window is held constant across all prediction periods. All forecasts were additionally carried out with a rolling of 48 months for the ARMA. The loss functions only differed marginally. These results are available upon reasonable request.

structures (Bengio et al., 1994, Hochreiter, 1998), their application seems unfruitful for data sets in which long-memory or elevated persistence—either in levels, differences, or variances—is expected.

Long short-term memory networks, first introduced in Hochreiter & Schmidhuber (1997), solve common issues of RNNs by allowing for longer dependencies across sequences of information by incorporating long-term memory channels. LSTMs are commonly applied in speech and handwriting recognition where longer sequences are a common feature. LSTMs are also applied in Finance to uncover patterns stock performance, see for example Fischer & Krauss (2018). While portfolio selection based on other machine learning techniques, such as deep neural networks, is already shown to outperform the market portfolio for larger indices (Moritz & Zimmermann, 2016, Krauss et al., 2017), LSTMs might be capable of further improving prediction accuracy as suggested in Fischer & Krauss (2018).

Here, we apply a simple sequence-to-sequence regression LSTM; that is, we train the network to predict one-step ahead NFP changes. This is achieved by shifting the standardized NFP observation vector (sequence input, $(y_t)_{t=1}^{T-1}$) forward by one period to represent the response ($(y_t)_{i=2}^{T}$) onto which the model is trained. The LSTM then *learns* to predict the value at the next time step of the input sequence (regression output layer). The number of features and responses is one in this case. We compare the in-sample performance as well as the out-of-sample performance of one-period ahead predictions. For the latter, we update the network on an expanding observation window that aligns with the expanding information set available to the other model-based approaches.

4.4. Loss function and prediction quality

For each individual forecaster i, we calculate the root mean squared error (RMSE) as loss functions, which are defined as

$$RMSE_i = \sqrt{\frac{1}{|T_i^*|} \sum_{t \in T_i^*} (A_t - F_{i,t})^2}, \tag{7}$$

The loss function for the average or consensus prediction defined in Eq. (1) reads

$$RMSE^{\text{avg}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(A_t - \overline{F}_t\right)^2}.$$

Loss functions for times series models and the deep learning LSTM model are defined accordingly, where $F_{i,t}$ is replaced with the fitted value for the in-sample analysis. For the out-of-sample analysis, the one period-ahead prediction is used in these loss functions, in which the summation of RMSE summands is only carried out across the out-of-sample period.

As the model-individual and economist-individual RMSE are sensitive to outliers, e.g. very large prediction-realization differences, forecasting performance can be sufficiently evaluated with the model confidence set (MCS) of Hansen et al. (2011) incorporating bootstrapped re-sampling that reduces the impact of bias induced by loss function outliers. The MCS then yields a set of models or combinations of forecasters that significantly outperform—with respect to the chosen loss function—those that are not an element of this model confidence set.

However, since our data is an unbalanced panel as not all economists predict for each month, the number of observations the RMSE is calculated on varies, rendering a bootstrapped re-sampling of losses infeasible on an individual basis. We apply the MCS on the mean, the mean of the *best* economists, and model-based monthly predictions for the in-sample and out-of-sample exercises.

As an individual-level alternative, we sort the prediction performance of each economist into clusters or in our case—intervals—which are produced by a simple one-dimensional k-means algorithm which belongs to the category of unsupervised learning.[15] This unsupervised learning technique now allocates economists in groups of similar prediction performance. The above problem of bias caused by outliers still persists but we aim to gain better understanding of performance groups by compression and comparing centroid

[15]The k-means algorithm partitions data—here the individual loss function performance—into k clusters where each in-cluster variance across cluster elements is minimized. We use this clustering to induce a grouping of the prediction performance of economists.

values (cluster or interval means) across the produced intervals. The value of $k = 5$ is chosen to obtain a cluster separation between lowest loss values (cluster 1), medium (cluster 3), and highest loss values (cluster 5).

Given that the panel is unbalanced, we additionally make use of the test statistic outlined in D'Agostino et al. (2012) who define a *normalized squared error statistic* as

$$E_{i,t} = \frac{(A_t - F_{i,t})^2}{|I_t^*|^{-1} \sum_{i=1}^{I_t^*} (A_t - F_{i,t})^2}, \tag{8}$$

with the average defined as

$$S_i = \frac{1}{|T_i^*|} \sum_{t \in T_i^*} E_{i,t}. \tag{9}$$

Following the bootstrapping approach of D'Agostino et al. (2012), we then randomly reassign each individual normalized squared error $E_{i,t}$ to a set of $|I_t^*|$ simulated forecasters at time t. This step is repeated for all t. This way, we end up with a simulated panel of forecast errors, that are randomly allocated, with an identical number of forecasts per period and number of predicted periods per economist. Note that forecast errors are not reshuffled across periods. We also follow D'Agostino et al. (2012) by sampling with replacement. We generate $N = 10\,000$ simulations of this panel and for each simulation, we calculate the average normalized score S_i^j according to Eq. (9) for all forecasters for a random panel j, with $j = 1, \ldots, 10\,000$, yielding 10 000 distributions of these scores. These simulations are then used to calculate percentiles and their confidence intervals that allow to answer the question if some forecasters are truly better than others.

According to D'Agostino et al. (2012), the intuition behind this shuffling of errors is as follows. If there are forecasters that are truly superior to others—that is, we reject the null hypothesis of equal predicting ability—their historical performance measured by S_i, defined in Eq. (9), should be significantly different from those obtained by the random reshuffling. We calculate confidence intervals for the average normalized squared error for the bootstrapped 5%, 25%, 50%, 75%, and 95% percentile and additionally, for the single best and worst forecaster.[16] We reject the null hypothesis of equal predicting ability if

[16] While D'Agostino et al. (2012) describe a simple bootstrap percentile method that might be prone to

average realized error measures are found to be outside of the 10% confidence interval of the bootstrap percentiles.

5. Findings

5.1. Survey error decomposition of the QES

5.1.1. The role of shocks

Following the framework of Davies & Lahiri (1999), in which we allow for predictor-individual bias in addition to a general shock that affects all predictions made in $t-1$ for NFP changes in month t, we focus firstly on the temporal shocks λ_t. Estimates are calculated according to Eq. (4) for the sample that ends before the impact of COVID19, yielding a sample size of 147 monthly observations. In some contrast to the original framework of Davies & Lahiri (1995), extended in Davies & Lahiri (1999), shocks do not affect prediction as they do not overlay with prediction horizons. Predictions are made between $t-1$ and t for t, translating to a simple one-period ahead prediction. Hence, shocks do not affect predictions for future periods, such as for $t+1$, as the predicting individual incorporates the shock in the available information set. This only holds for our data set; for multi-period prediction, for example inflation rates (Davies, 2006, Ang et al., 2007, Boero et al., 2008), shocks affect several forecasts at differing horizons yielding an accumulative effect of shocks.

We focus on the predominant shock in our data structure that is observable to predicting economists, the shock calculated based on the first publication of NFP figures. We put focus on this measure as this shock affects not only predicting entities directly but is also observed as most recent news impacting equity and fixed income markets around the publication date. Shocks based on the second or third revised NFP figure would be available only with the next publication dates with a diminishing surprise or news affect as these figures would not refer to the most recent publication. Hence, we attribute the highest relevance to the figures of the first publication.

bias, this paper calculates the confidence interval based on the *Bias Corrected and accelerated percentile* method (BCa) of DiCiccio & Efron (1996).

Given its construction, the estimate for the temporal shock λ_t, affecting all predicting economists at time t, is the average forecasting error across all (participating) economists adjusted for the individual bias estimate $\hat{\phi}_t$. As such, it is directly related with the mean prediction made by economists. This is visualized in Figure 5. Shocks are presented as orange line in comparison to the first release of NFP figures (yellow bars) and the QES mean (black, dashed line). The sign of the shocks $\hat{\lambda}_t$ correspond directly to the direction of the *news* or surprise. If a job market figure is lower than expected, translating to *bad news* or a negative, unexpected effect, the sign of $\hat{\lambda}_t$ is negative. If published numbers are exceeding the expectations of forecasters and market participants, the sign is positive as we face *good news*. In Figure 5, we observe an alternating pattern of estimated shocks $\hat{\lambda}_t$. During the financial crisis in the U.S., it becomes evident that predictors have underestimated the effects on the labor situation in the beginning of a prolonged phase of negative developments, yielding negative shocks. However, during the recovery phase, predicting economists also underestimate the decline in job losses, which is picked up as *good news* as fewer than anticipated jobs were lost.

Additionally, Figure 5 shows that average predictions, which are a main contributor to the shock estimation, vary only little and show high autocorrelation and a lagged reaction to shocks. Several examples exist in the present sample. The survey mean does not predict larger deviations from a trend, which yields several positive and negative spikes in temporal shocks $\hat{\lambda}_t$, while individually, some economists have strong fluctuations in their predictions.

Most importantly, we find that shocks show some degree of seasonality in its pattern in which winter figures are usually over-estimated yielding negative shocks. For the majority of years, we also observe positive shocks in summer and fall. This might indicate that economists do not incorporate seasonal patterns in their predictions on average.[17] In order to to control for seasonality, we estimate temporal shocks with respect to the most recent, seasonally adjusted job market figures. We find differences in these estimates, in

[17]We find elevated autocorrelations for lag 11 and 12, albeit of no statistical significance. For reasons of brevity, these results are not reported in detail.

particular for peak values of shocks which seem to be higher for the seasonally adjusted figures in absolute terms. However, we also find that the alternating pattern remains mainly intact. This is visualized in the Appendix in Fig. B.9. We detect statistically significant autocorrelation in the first difference of the shock series' for the first lag, both for shocks based on the seasonally unadjusted first release and the most recent, seasonally adjusted release. This might be an indicator for the effect of large errors on the forecasters. However, as this effect is also present in the most recent releases,[18] it is more likely that this is a residual of the significant autocorrelation of the mean forecast across the panel.

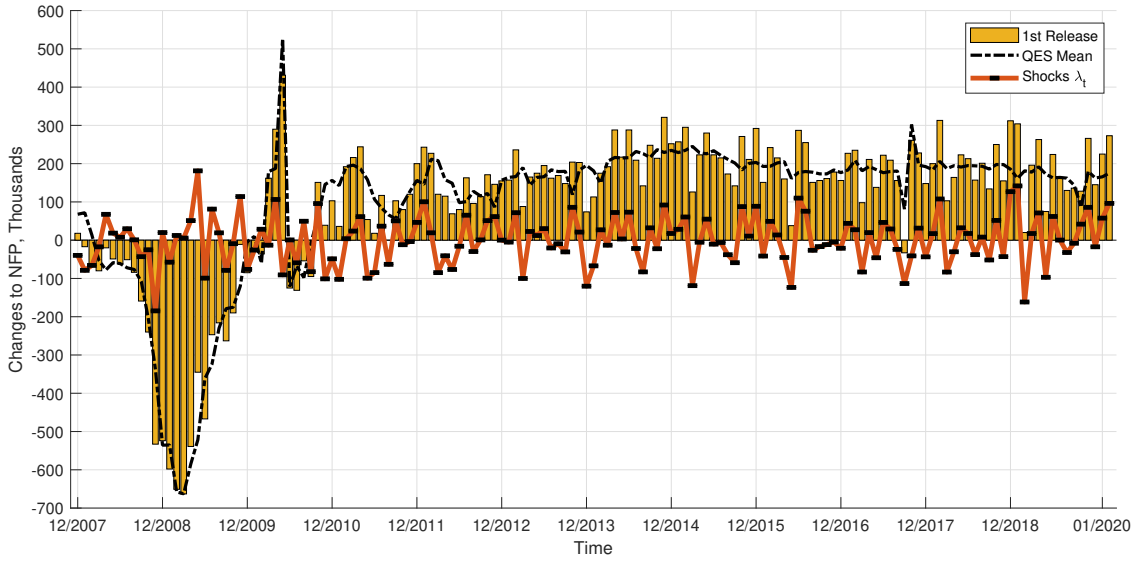


Figure 5: Temporal shocks $\hat{\lambda}_t$ (orange line) estimated as defined in Eq. (4), NFP changes of the first publication (yellow bars), and QES mean (black dashed line) for NFP changes published from December 2017 to February 2020.

5.1.2. Individual bias of economists

We estimate the individual bias of each economist according to Eq. (3). A forecaster systematically over-predicts with a significant $\hat{\phi}_i > 0$. In our sample, we find several economists who show a statistically significant bias, either positive or negative. Figure 6 visualizes the bias estimates $\hat{\phi}_i$ for all economists with more than 50 survey entries. Red squares are marking bias that is statistically significantly different from zero. Evidently,

[18]These most recent releases are not available to the economists at the lag of the detected autocorrelation.

most of the elevated biases in absolute terms are significant and indicate some dissociation from a rational forecast. Notably, the number of predictions made varies and as such, there is no universal threshold bias to determine significance jointly for all economists.

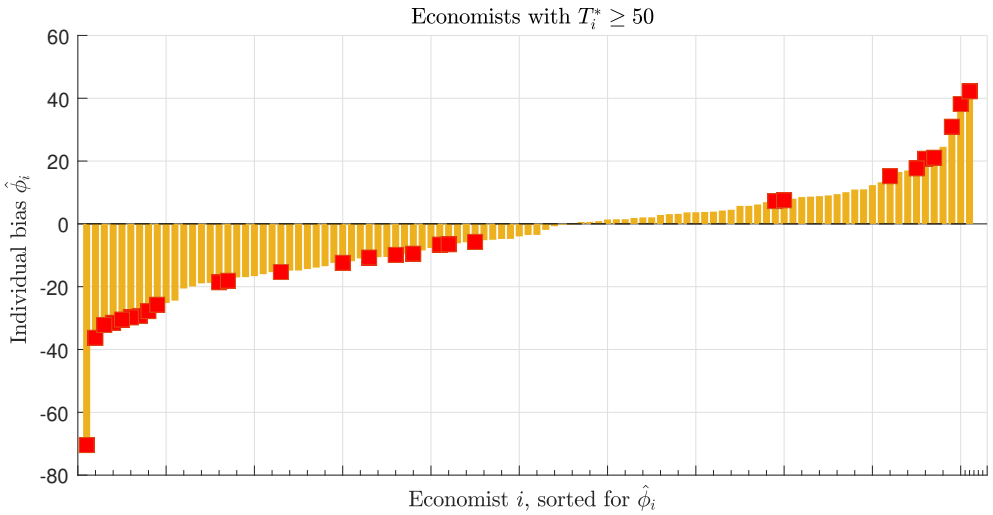


Figure 6: Idiosyncratic bias $\hat{\phi}_i$ across economists with at least 50 predictions based on the first publication of NFP numbers. Red squares indicate a statistically significant bias at the 10% significance level.

For the bias estimation, we find that the choice of publication figure plays a major role. This is in some contrast to the shock measure in the previous subsection. In what follows, we discuss this bias for all economists and for groups of economists with at least 50 or 100 predictions, respectively. We compare the bias across the first and the final figure of NFP changes.

The estimated bias is visualized in Figure 7. We sort the bias estimates from lowest to highest, as the order of economist identifiers is of no interest for this analysis. The three subplots refer to three nested groups of economists. The left-hand plot visualizes $\hat{\phi}_i$ for all economists. If the number of predictions is low, the estimate might be prone to small sample bias. The remaining two figures calculate the bias for economists with a reasonable number of predictions which ultimately reduces the number of economists to 103 and 66, respectively.

Several important observations are drawn from Figure 7. Firstly, we find that with a higher participation rate, that is with a higher number of predictions made, the individual bias seems to be decreasing. By removing economists with predictions below the threshold of 50 or 100, we remove some of the extreme values of positive and negative bias as

well. This is of interest as it shows that we find a reduction in systematic over- or under-prediction the more predictions an economist enters into the survey. These findings suggest that the more regularly an economist enters survey responses in the QES, the lower the systematic error. This could indicate differing information sets to other economists based on learning effects from past prediction differences or differing information rigidities.

As a second important observation, we find the individual bias to differ significantly if we base the calculation on the third and final publication of NFP figures. Bias calculated based on the third publication are plotted in blue in Figure 7. We observe a downward shift and a generally negative bias. The NFP figures are more precise as the response rates of businesses reported job numbers are much higher than for the first publication. This underlines a tendency to under-predict true or more precise values of NFP changes. The magnitude of this effect can only be partially explained by the observed upward correction in absolute terms from first to third revision in roughly 60% of the observations, which would causes the bias to decrease as A_t increases for some t. Of course, some of this increase would also be offset by a downward correction of releases. Hence, it is suggested that the role of the final figures for NFP changes play a less important role for the prediction of economists and the focus remains on the first release. Additionally, we find that the impact of these revisions affects the temporal shock $\hat{\lambda}$ to a lesser extent than the bias component $\hat{\phi}$.

5.2. Prediction performance

5.2.1. In-sample analysis

We now turn to the individual prediction quality and firstly, focus on an in-sample view covering all observations from December 2007 to February 2020 and in a second in-sample analysis, to December 2020 to analyze the impact of COVID19. Based on the individual predictions of economists, loss functions are calculated and results based on the first release are presented. As outlined previously, the number of predictions made by each forecaster varies which proves challenging for an evaluation of outperformance. To separate prediction quality implied from RMSE, a k-means clustering is applied on (1) all economists with at least 50 observations ($|T_i|^* \geq 50$), which yields 101 economists

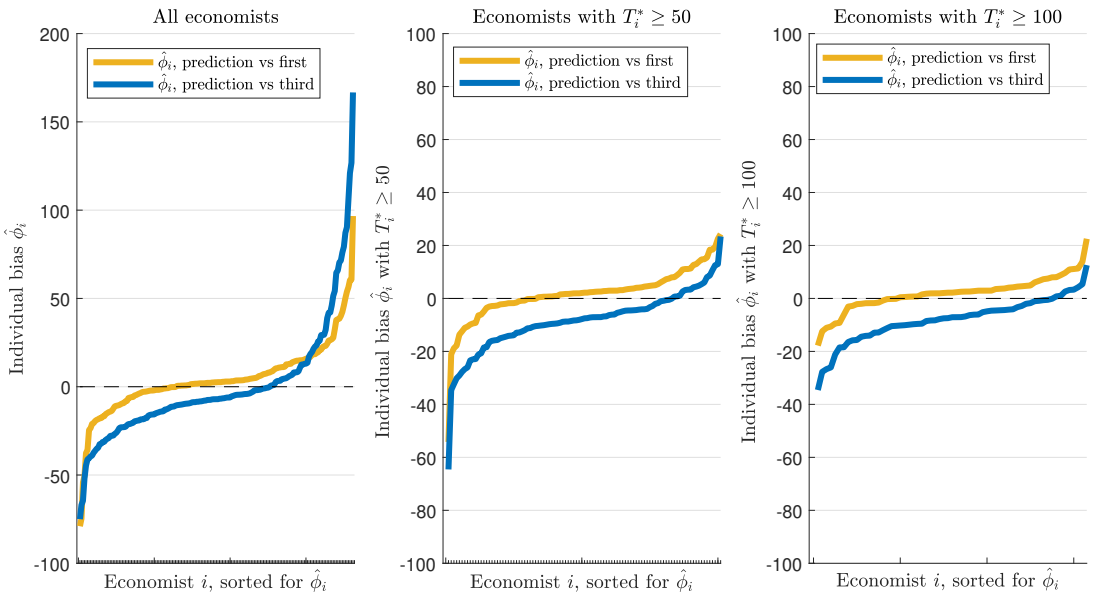


Figure 7: Idiosyncratic bias $\hat{\phi}_i$ across all economists *(left)*, economists with at least 50 predictions *(middle)*, and economists with at least 100 predictions *(right)*, sorted from low to high for each subplot. Yellow lines refer to the bias based on the first publication of NFP changes. Blue lines refer to bias based on the third and final publication of NFP figures, published in $t+2$ for period t.

remaining in the sample and (2) all economists with at least 100 observations ($|T_i|^* \geq 100$), which further reduces the sample to 63 economists. The number of clusters or in this specific data set—intervals—is chosen to be $k = 5$ as motivated in Subsection 4.4.

For RMSE figures for economists with at least 50 predictions, we find the majority of economists in the second and third cluster. The best performing economists are grouped in the first cluster, whose RMSE ranges from 57.5 to 64.3 while for the worst performing cluster, RMSE ranges from 91.7 to 99.2. Centroid values and cluster boundaries vary only marginally for the first three groups of economists with at least 100 observations. We cannot directly infer on outperformance of economists in the first cluster compared to all others but we can certainly say that they predict better, on average, than those in the fifth cluster. Interestingly, we find that if we only focus on economists with a higher number of predictions, this homogeneity of loss functions increases with the intercluster distance decreasing, which can be partly attributed to a smaller sample and the exclusion from economists with fewer predictions. This, in turn, indicates that economists with fewer predictions might produce higher losses. Detailed results are found in the appendix in Table C.10. These findings further motivate an analysis of equal predicting ability across economists and also justify filtering economists by participation.

The in-sample fits of the QES means, the QES means based on the selection of best performing economists following D'Agostino et al. (2012), the ARMA(p,q) model, and the deep learning network as introduced in Subsection 4.2 and 4.3 for both in-sample periods are presented in the in-sample columns of Table 1.

For both in-sample periods, the QES mean of all economists and their reduction to only include those with 50 or 100 predictions performs similar with respect to the RMSE. The no-change forecast, the prevailing mean over three, six, and twelve months, and the ARMA(1,1) perform much worse. We find the LSTM network to show superior data fit which is due to the nature of training of the model itself. The LSTM is the only model included in the MCS, and as such, significantly outperforms all others. If we exclude the LSTM from the model selection, all QES mean forecasts that are produced from the selection of the best, the best 5%, and the best 25%—based on the relative error and selection of D'Agostino et al. (2012)—are included in the MCS. This indicates that these forecasts provide a better prediction quality than standard QES means, which extends the findings of Brown et al. (2008) based on the measure of D'Agostino et al. (2012). In the subsequent section, we address this evident rejection of the equal predictive ability hypothesis. However, this outperformance only holds for the pre-COVID19 sample. For the sample which includes COVID19, we find no model to perform statistically better than others since errors are exceptionally high, with the exception of the LSTM network approach.

Loss	Model	In-sample 1	In-sample 2	Out-of-sample 1	Out-of-sample 2
RMSE	QES mean all	64.26	808.0	65.52	1 683
	QES mean 50	64.14	808.4	65.34	1 684
	QES mean 100	63.98	799.8	**65.08**	1 666
	QES best*	**62.12**/**61.87**/**61.23**	698.8/731.4/786.9	66.39/66.90/65.55	1 697/1 687/1 676
	QES best 50*	**61.34**/**61.60**/**61.48**	753.9/740.5/794.5	66.35/68.55/65.52	1 921/1 942/1 717
	QES best 100*	63.31/**61.79**/**61.45**	734.8/708.6/807.5	72.00/**64.80**/**64.44**	**1 506**/1 564/**1 510**
	No-change	108.7	2 446	119.89	5 105
	mean (3/6/12)	108.4/122.9/141.7	2 097/1 934/1 861	87.71/81.15/77.80	4 333/3 955/3 726
	LSTM	<u>33.99</u>	<u>354.8</u>	72.84	2 999
	ARMA(1,1)	100.7	1687	103.0	7 208

Note: In-sample 1 refers to the data set that ends in February 2020 (pre-COVID) with $n_1 = 147$ and In-Sample 2 refers to the full data set covering the COVID19 crisis ending in December 2020 with $n_2 = 157$. *For the QES best series', the best, the best 5%, and the best 25% of economists based on the squared loss over the respective sample are used as predictor for in-sample performance. For the out-of-sample exercise, the best series are updated monthly based on an expanding window, allowing for changes in the set of forecasters across the out-of-sample period.

Table 1: In-sample and out-of-sample RMSE for QES mean predictions and time-series models.

5.2.2. Equal predictive ability

We address the hypothesis of equal predictive ability of the participants of the Bloomberg survey and follow the approach of D'Agostino et al. (2012). We compare the findings obtained from using the first publication of NFP changes with the forecasting performance benchmarked against the most recent and seasonally-adjusted release. Table 2 shows the distribution of the average normalized squared error across the best, 5%, 25%, 50%, 75%, 95%, and worst percentile including bootstrap 10% confidence intervals, based on the first publication of NFP changes. We reject the hypothesis of equal predicting ability by finding statistically significant evidence of better-performing economists. For example, the best 25% of economists, even when controlling for survey participation, is shown to perform better than the remainder. In addition, by only including economists that participate regularly, we even find the group of the best 5% to be statistically significantly better than the remainder. Similar to D'Agostino et al. (2012), we further identify groups of economists that perform significantly worse.

Table 3 repeats this analysis but bases the error measure on the most recent, seasonally-adjusted publication. The results of a rejection of the equal predictive ability of economists are confirmed. We even observe an even higher spread between significantly better performing economists and the worst performing economists. By restricting the sample by the participation rate, we find the significance of over-performance to increase as well.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.1028	0.5994	**0.7068**	**0.8139**	0.9557	1.3813	3.5633
CI	(0.0421; 0.4215)	(0.5221; 0.6372)	(0.7404; 0.7829)	(0.8314; 0.8714)	(0.9262; 0.9834)	(1.1425; 1.3777)	(1.5819; 4.5791)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5141	0.6246	**0.7237**	**0.8139**	0.9510	1.3670	3.5633
CI	(0.3661; 0.5806)	(0.6078; 0.6833)	(0.7549; 0.7934)	(0.8351; 0.8733)	(0.9213; 0.9745)	(1.0953; 1.2701)	(1.3523; 2.3176)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5730	**0.6310**	**0.7227**	**0.8095**	**0.8905**	1.2655	3.5633
CI	(0.5612; 0.6756)	(0.6705; 0.7256)	(0.7651; 0.8023)	(0.8272; 0.8648)	(0.8929; 0.9403)	(0.9986; 1.0989)	(1.0879; 1.3897)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.6288	**0.6450**	**0.7265**	**0.8118**	0.8874	1.2643	1.3722
CI	(0.6221; 0.7139)	(0.6895; 0.7442)	(0.7645; 0.8042)	(0.8148; 0.8549)	(0.8665; 0.9141)	(0.9410; 1.0233)	(0.9840; 1.1578)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 2: Distribution of the forecasting performance relative to the first publication of NFP changes for the in-sample period from December 2007 to February 2020.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.3442	0.5942	0.7352	**0.8379**	0.9495	1.4565	3.2762
CI	(0.1564; 0.5162)	(0.5755; 0.6646)	(0.7587; 0.7932)	(0.8390; 0.8729)	(0.9256; 0.9763)	(1.1110; 1.2914)	(1.4462; 3.4634)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5002	0.6261	**0.7383**	**0.8376**	0.9446	1.4669	3.2762
CI	(0.4079; 0.5848)	(0.6245; 0.6934)	(0.7664; 0.7998)	(0.8396; 0.8725)	(0.9181; 0.9654)	(1.0722; 1.2159)	(1.2789; 2.1816)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5548**	**0.6539**	**0.7394**	**0.8290**	0.9057	1.1488	2.6415
CI	(0.5789; 0.6863)	(0.6831; 0.7345)	(0.7733; 0.8064)	(0.8312; 0.8645)	(0.8914; 0.9335)	(0.9894; 1.0833)	(1.0759; 1.3437)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.6283**	**0.6799**	**0.7550**	0.8206	0.8920	1.1061	1.2978
CI	(0.6472; 0.7285)	(0.7070; 0.7550)	(0.7737; 0.8083)	(0.8190; 0.8529)	(0.8645; 0.9044)	(0.9303; 1.0027)	(0.9692; 1.1207)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 3: Distribution of the forecasting performance relative to the seasonally-adjusted, most recent publication of NFP changes for the in-sample period from December 2007 to February 2020.

As those findings could be affected by exceptionally worse performing subgroups of economists, we drop the worst performing 20% of economists of each group and repeat the bootstrap. Results are shown in Table 4 for the first publication and in Table 5 for the most recent revision. By removing the worst performing 20%, we reduce the average error and the levels of confidence intervals as large errors are excluded from the bootstrap. We find our previous results to be robust. The best performing economists are significantly better than others, in particular for the groups that only include those economists who participate more regularly. Results with respect to the third release are found in Table C.11 and C.12.

Hence, we reject the hypothesis of equal predictive ability of economists of the QES, in line with what D'Agostino et al. (2012) found for the SPF. These findings also corroborate the impact of the significant bias of some forecasters, which are found within the Davies & Lahiri (1995) framework. We further show that these results hold across revisions of the NFP changes, while the for the most recent, seasonally-adjusted release we obtain the most compelling results.

5.2.3. Out-of-sample analysis

We now compare the one period-ahead prediction performance of the QES, individually and as mean, with the performance of time series model and deep learning. We run

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.1028	0.5884	**0.6830**	0.7864	0.8448	**0.9665**	**1.0065**
CI	(0.0424; 0.4446)	(0.5090; 0.6175)	(0.6900; 0.7240)	(0.7576; 0.7872)	(0.8235; 0.8659)	(0.9724; 1.1444)	(1.2091; 3.2170)
Panel A: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5141	0.6176	**0.6961**	0.7928	0.8446	0.9537	**0.9936**
CI	(0.3923; 0.5769)	(0.5860; 0.6507)	(0.6982; 0.7305)	(0.7574; 0.7868)	(0.8168; 0.8563)	(0.9334; 1.0581)	(1.0822; 1.6166)
Panel A: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5730	0.6267	**0.6957**	0.7781	0.8211	0.9052	0.9500
CI	(0.5364; 0.6345)	(0.6209; 0.6702)	(0.6975; 0.7289)	(0.7456; 0.7757)	(0.7936; 0.8306)	(0.8702; 0.9560)	(0.9321; 1.1645)
Panel A: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.6288	**0.6444**	**0.7115**	0.7864	0.8211	0.9152	0.9508
CI	(0.5995; 0.6797)	(0.6504; 0.7021)	(0.7134; 0.7487)	(0.7542; 0.7889)	(0.7946; 0.8342)	(0.8493; 0.9189)	(0.8779; 1.0064)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 4: Distribution of the forecasting performance relative to the first publication of NFP changes for the in-sample period from December 2007 to February 2020, restricted to the best 80% of forecasters.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.3442	0.5747	**0.7032**	0.7931	0.8690	0.9553	**0.9954**
CI	(0.1892; 0.5229)	(0.5566; 0.6408)	(0.7157; 0.7462)	(0.7806; 0.8084)	(0.8467; 0.8880)	(0.9949; 1.1414)	(1.1963; 2.6436)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5002	**0.5949**	**0.7133**	0.7944	0.8689	**0.9505**	**0.9942**
CI	(0.4339; 0.5780)	(0.5971; 0.6616)	(0.7211; 0.7507)	(0.7802; 0.8071)	(0.8397; 0.8768)	(0.9559; 1.0716)	(1.0879; 1.5848)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5548**	**0.6408**	**0.7185**	0.7931	0.8588	0.9082	**0.9444**
CI	(0.5636; 0.6581)	(0.6478; 0.6956)	(0.7258; 0.7546)	(0.7721; 0.7994)	(0.8182; 0.8521)	(0.8921; 0.9699)	(0.9497; 1.1624)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.6283**	**0.6700**	**0.7274**	0.7957	0.8564	0.8980	0.9105
CI	(0.6312; 0.7055)	(0.6788; 0.7256)	(0.7371; 0.7691)	(0.7752; 0.8062)	(0.8128; 0.8482)	(0.8629; 0.9274)	(0.8899; 1.0060)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 5: Distribution of the forecasting performance relative to the seasonally-adjusted, most recent publication of NFP changes for the in-sample period from December 2007 to February 2020, restricted to the best 80% of forecasters.

an out-of-sample analysis on two overlapping windows. The first window contains 36 predictions and runs from January 20217 to December 2019. The second out-of-sample window runs from January 2018 to December 2020 and includes the COVID19 shock. We separate these two prediction windows to disentangle the effect of the labor shock on the prediction quality.

Overall, we find the prediction performance of the ARMA, the LSTM, the no-change forecast, and the prevailing mean to be insufficient as none of these models is included in the MCS, while the LSTM shows the relative best performance of this group. Simple QES

mean predictions and a selection of best economists produce much lower losses. The MCS reveals that the mean forecasts built on the best performing 5% and 25% of economists, that very regularly participate, yield superior forecasts. The simple QES mean of the same subgroup of economists is also included in the MCS. This expands the findings of Brown et al. (2008) by showing that we are able to construct superior forecasts based on a selection of the best performing economists. It further shows that the framework outlined in D'Agostino et al. (2012) offers a viable choice as a survey combination method to improve prediction quality. Lastly, it complements Clements (2021) as we show that there is a difference between those economists that participate very regularly and those who drop in and out.

The out-of-sample performance during the period of COVID19 is discussed separately in Subsection 5.3.

5.3. The impact of COVID19

The impact of COVID19 on the U.S. job market is comparable with the characteristics of a *black swan* event; an unpredictable incident with negative consequences for a majority of a population. COVID19 started its rapid and lethal spread throughout every state of the U.S. in February, if not earlier. Fatalities and infection rates increased in the beginning of March[19] to which drastic counter measures were rolled out (Janiak et al., 2021). Closures of businesses, in particular in the hospitality sector, as well as close-contact industries, shut-downs of plants, and a recession followed shortly after. However, Albanesi & Kim (2021) highlight that this recession effect is unlike anything previously observed as close-contact businesses and in particular women are asymmetrically affected. In addition, this effect was nation-wide, causing a joint economic disruption across the U.S. (Rojas et al., 2020). The impact on the U.S. job market is unprecedented as absolute job losses of this magnitude have never been recorded before. In turn, the subsequent recovery due to re-opening of businesses (Bartik et al., 2020) and a generally different approach to dealing with the pandemic caused an increase in jobs of previously unknown and unseen

[19]See for example `https://covid.cdc.gov/covid-data-tracker/`.

dimensions as well.

Table 6 gives an overview of these numbers. In the first NFP publication for March 2020, a loss of 701 000 jobs is reported. Underlining the abnormality of the situation, this number is later revised to roughly double that loss, at 1 373 000 jobs lost published in the third release and further corrected in the seasonally-adjusted recent release to 1 683 000 jobs lost. The largest effect on the job market is recorded for April 2020, with a loss of 20 500 000 jobs which is corrected to 20 679 000 in the recent release. Job losses that large have never been recorded before. For comparison, job losses accumulated from January 2008 to December 2009, the peak of the financial crisis in terms of job market impact, equal 6 887 000 over a period of two years, see Figure 2, panel (a). In May to August, some of these lost jobs are reclaimed and a recovery of 2 833 000 in May and 4 846 000 in June are observed. This strong recovery continues until November, albeit still in a net loss due to April 2020. This job market recovery is of a previously unknown and abnormal magnitude as well, see Figure 2, panel (b).

		March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Release	first	−701	−20 500	2 509	4 800	1 763	1 371	661	638	245	−140
	second	−870	−20 687	2 699	4 791	1 734	1 489	672	610	336	−227
	third	−1 373	−20 787	2 725	4 781	1 761	1 493	711	654	264	−306
	recent	−1 683	−20 679	2 833	4 846	1 726	1 583	716	680	264	−306
Obs.		148	149	150	151	152	153	154	155	156	157

Table 6: Nonfarm payroll changes from March to December 2020.

Naturally, these extreme events also impact the forecasting error decomposition. Due to the scaling of the data during the COVID19 sub-sample, a visual representation of the complete sample becomes unfeasible. As such, Fig. 8 limits the view on 2020 and shows that the forecasting error follows suit in terms of magnitude. Table 7 shows the QES mean prediction for each month and the temporal shocks during these months following the error decomposition of Davies & Lahiri (1995). It is evident that the survey participants, on average, underestimated the immediate impact of COVID on the U.S. job market in March. The resulting temporal shock $\hat{\lambda}_{\text{March}}$ takes the highest negative value on record with $\hat{\lambda}_{\text{March}} = -593\,000$, translating to an extreme negative shock at that time of under-predicting job losses by half-a-million jobs. Surprisingly, the prediction for April roughly

aligns in terms of dimensions but this time, significantly over-predicts job losses. This in turn yields a positive shock as not as many jobs are lost as anticipated. The largest deviation from prediction to realized value is recorded for May 2020. The QES mean predicts a loss of approximately 7.3 million jobs while the actual number of jobs increased by 2.5 million. This causes a positive shock of 9.7 million jobs. In June, the QES again under-predicts NFP changes incurring a shock of 1.2 million jobs. With regard to individual forecaster bias ϕ_i, the estimates are now negatively biased due to the large deviations in these four months. Additionally, we observe some evidence for information rigidities as forecasters first underestimate the situation in March but then repeatedly overestimate job losses or underestimate the recovery, in particular in May, yielding a repeated occurrence of positive shocks. This slow reaction of economists is also found for inflation predictions during this time as demonstrated in Armantier et al. (2021). Forecasters, on average, produce more conservative predictions after April 2020. While the estimates for the shock variable λ in Table 7 are estimated across the full sample, we note that by construction—particularly in view of the forecaster bias estimated in Eq. (3) and the shock itself in Eq. (4)—the shock variable is now biased downwards due to the pivotal difference of expected labor development and realized numbers in March, April, May, and to some extend in June. These few but large outliers in differences are affecting the idiosyncratic bias with a negative shift. This also shifts the temporal shock variable downwards. When estimated across the whole sample including the COVID19 period, almost all temporal shocks are negative, underlining their biasedness due to these large outliers compared to those estimates obtained with the non-COVID19 sample.

	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
NFP Change (first)	−701	−20 500	2 509	4 800	1 763	1 371	661	638	245	−140
QES mean	−237	−21 944	−7 357	3 389	1 424	1 343	897	616	456	43
QES mean 50	−254	−22 132	−7 344	3 448	1 467	1 360	897	617	458	42
Shocks $\hat{\lambda}_t^{all}$	−593	1 314	9 725	1 289	209	−98	−361	−100	−335	−307
Shocks $\hat{\lambda}_t^{50}$	−538	1 540	9 753	1 261	205	−77	−322	−66	−299	−275

Table 7: Nonfarm payroll changes from March 2020 to December 2020, during the first wave of COVID19 with the QES mean of all economists as well as those with at least 50 predictions and temporal shock estimates $\hat{\lambda}_t$ of these economists.

Turning to the prediction quality of the QES, its means, the best selection, and the

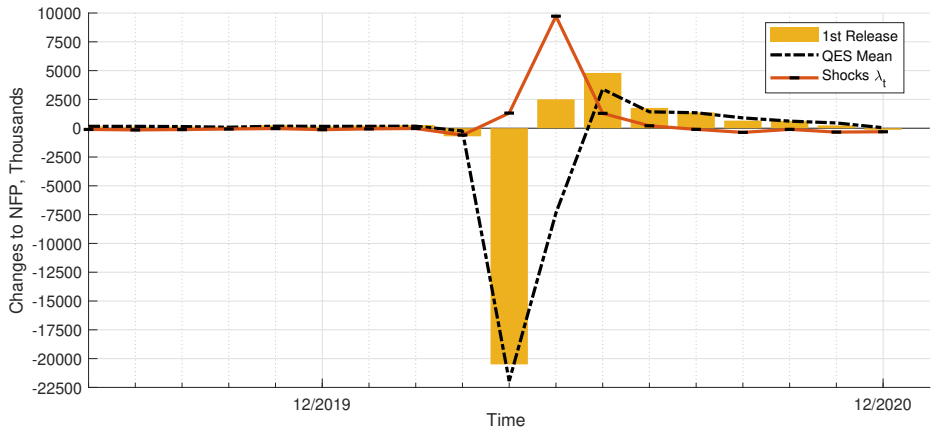


Figure 8: Temporal shocks $\hat{\lambda}_t$ (orange line) estimated as defined in Eq. (4), NFP changes of the first publication (yellow bars), and QES mean (black dashed line) for NFP changes published from July 2019 to December 2020.

model approaches, we revisit the fourth column of Table 1, and focus on the Out-of-sample 2 performance. We observe elevated RMSE figures and find distinctive differences between the consensus forecast and the model based approaches. The QES mean forecast produces similar prediction quality while only the forecasts produced by a selection of the best economists are elements of the MCS, translating to a statistically significant outperformance relative to all other predictions. In detail, predictions based on the best 25% of economists that regularly participate produce significantly lower losses and are superior to the other predictions. This again documents the existence of *smart predictions* as in Brown et al. (2008), even in times of extreme differences of expected and realized value. It further confirms that forecasters that regularly participate seem to have better predicting ability, in some relation to the findings of Capistrán & Timmermann (2009a) and Clements (2021).

The model-based predictions are unable to cope with the March loss and due to the stationarity restrictions, under-predict the April loss as they react too slow to these new values. These models also cannot predict the sudden change to recovery. The LSTM network shows a lagged reaction and predicts positive NFP changes for March and April while prediction are negative for May and June. All model-based approaches are producing very large residuals for these months that drive the RMSE far beyond what is observed from the QES consensus figures. This is, however, expected behavior of model-

based predictions as labor events surrounding COVID19 are unexpected shocks, in which the structure of data evolution changes completely. By its very nature, none of these models are included in the MCS. In light of the extreme magnitude of this event, the prediction behavior is in line with past observations. Economists tend to under-predict losses, while the turn to a recovering job market is predicted with lagged periods as observed after the financial crisis—at lower levels. During the months of the labor shock, very high standard deviations across individual predictions are observed and very high prediction errors both on individual and on consensus level follow. This phenomenon is known from inflation predictions (Rich & Tracy, 2010) and shown to be present in these NFP predictions as well.

In what follows, we briefly address the distribution of the forecasting performance across performance percentiles of forecasters with regard to the equal predictive ability hypothesis. Tables 8 and 9 show the percentiles and bootstrap confidence intervals, for all economists and a limited set by dropping the worst performing 20% as argued above. Including the COVID19 labor shock until December 2020 does not change the qualitative result of rejecting the hypothesis of equal predictive ability. We still find that the best 5% and 25% of economists with more than 50 or 100 survey entries predict better than the bootstrap distribution would suggest, even when controlling for negative outliers by removing the worst performing 20%. These findings are robust to the choice of release as shown in Tables C.13 - C.16 in the Appendix. This is evidence that not all economists are affected equally by the uncertainty surrounding the true impact of COVID19 on the labor market in March, April, and the following months.

6. Concluding Remarks

We analyze an unbalanced panel of nonfarm payroll predictions from January 2008 to December 2020 from 181 forecasters. Based on the framework of Davies & Lahiri (1995), we decompose the forecasting error of each forecaster into three components, of which two are further studied. Firstly, we focus on the temporal shock component that affects all forecasters equally per forecasting period. These shocks, a general over- or

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.1028	0.6020	**0.7197**	**0.8173**	0.9606	1.3750	3.5633
CI	(0.0440; 0.4357)	(0.5298; 0.6457)	(0.7449; 0.7869)	(0.8353; 0.8748)	(0.9308; 0.9883)	(1.1476; 1.3886)	(1.5907; 4.5369)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4794	0.6144	0.7293	0.8173	0.9514	1.3582	3.5633
CI	(0.3444; 0.5747)	(0.6092; 0.6862)	(0.7582; 0.7969)	(0.8385; 0.8768)	(0.9256; 0.9793)	(1.1002; 1.2803)	(1.3645; 2.5128)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5737	**0.6303**	**0.7295**	**0.8059**	0.8985	1.2953	3.5633
CI	(0.5601; 0.6765)	(0.6722; 0.7280)	(0.7672; 0.8046)	(0.8298; 0.8670)	(0.8952; 0.9441)	(1.0029; 1.1049)	(1.0943; 1.3933)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.6095**	**0.6307**	**0.7299**	**0.8065**	0.9123	1.2845	1.3590
CI	(0.6164; 0.7097)	(0.6865; 0.7408)	(0.7631; 0.8024)	(0.8146; 0.8537)	(0.8673; 0.9151)	(0.9441; 1.0318)	(0.9907; 1.1758)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 8: Distribution of the forecasting performance relative to the first publication of NFP changes for the in-sample period from December 2007 to December 2020.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.1028	0.5884	**0.6799**	0.7774	0.8502	**0.9737**	**1.0212**
CI	(0.0440; 0.4664)	(0.5201; 0.6261)	(0.6943; 0.7286)	(0.7607; 0.7916)	(0.8284; 0.8722)	(0.9778; 1.1536)	(1.2183; 2.9947)
Panel A: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4794	0.6106	**0.6965**	0.7901	0.8505	0.9554	**1.0096**
CI	(0.3546; 0.5713)	(0.5814; 0.6506)	(0.7000; 0.7321)	(0.7603; 0.7903)	(0.8222; 0.8638)	(0.9473; 1.0862)	(1.1044; 1.7648)
Panel A: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5737	0.6261	**0.6972**	0.7686	0.8390	0.9141	0.9419
CI	(0.5392; 0.6360)	(0.6246; 0.6730)	(0.7002; 0.7315)	(0.7478; 0.7778)	(0.7956; 0.8327)	(0.8730; 0.9546)	(0.9336; 1.1582)
Panel A: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.6095	**0.6294**	**0.7043**	0.7704	0.8296	0.9160	0.9419
CI	(0.5904; 0.6717)	(0.6445; 0.6947)	(0.7083; 0.7428)	(0.7490; 0.7837)	(0.7901; 0.8299)	(0.8468; 0.9178)	(0.8774; 1.0099)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table 9: Distribution of the forecasting performance relative to the first publication of NFP changes for the in-sample period from December 2007 to December 2020, restrcited to the best 80% of forecasters

under-prediction of *all* forecasters for a particular month represents a news effect where an under-prediction of job increases is considered a positive shock and vice versa. From these estimated shocks, we find that the sample of predicting economists under-estimate job losses in times of prolonged market turmoil. In addition, recovery phases are under-predicted as well, leading to positive shocks.

In general, we find that the mean predictions are rather stable, causing the shock estimate to alternate regularly. Secondly, we focus on the individual bias, which describes a systematic over- or under-prediction of a particular forecaster. We find the bias of

several forecasters to be statistically significant. More importantly, we find that with increasing participation rate, the individual bias is decreasing, yielding a lower prediction error. This indicates that economists that regularly make predictions are incorporating differing information sets than those with very few predictions. If we decompose the forecast errors based on a more precise measure for job market figures, the most recent publication, we observe a downward shift and a generally negative bias, underlining a tendency to under-predict true or more precise values of NFP changes. This suggests that forecasters make limited use of subsequent revisions of NFP changes and their focus remains on the initial and preliminary numbers. In view of the applied framework, we find that the impact of these revisions affects the temporal shock to a lesser extent than the individual bias.

We apply several model-based approaches and compare their in-sample fit and out-of-sample prediction quality with the predictions made in the survey. Additionally, we employ a deep learning LSTM network. The LSTM shows superior in-sample fit and outperforms the time series models in the out-of-sample forecasting. However, compared to the mean forecast of the qualified economists, the quality of these model-based predictions is lower. This implies that exogenous factors play a major role for nonfarm payroll forecasts. Additionally, we show that *smart consensus forecasts* that base on the selection of the best performing economists identified with the methodology of D'Agostino et al. (2012) outperform all other forecasts. This outperformance is statistically significant. These findings are in line with Brown et al. (2008). This rejects the hypothesis of equal predictive ability and we find stronger evidence against this hypothesis for economists with higher participation rates. This relates to Clements (2021).

We analyze the impact of COVID19 on the U.S. labor market and highlight the abnormal character of these job market figures, which is shown for temporal shocks in the error decomposition. It is evident that the survey participants, on average, underestimate the immediate impact of COVID but also under-predict the following recovery. However, this is a known prediction pattern; job losses are under-predicted at first while later on, the recovery is also predicted to be lower than realized, which had been observed for earlier

labor shocks already. Additionally, we find that extreme outliers in the difference between expected and realized NFP changes in April, May, and June have adverse effects on the estimation of idiosyncratic bias of economists and temporal shocks within the Davies & Lahiri (1995) framework. The shock estimates are negatively biased due to the impact of COVID19. Including COVID19 in addressing the hypothesis of equal predicting ability across different groups of economists further strengthens the evidence for rejecting this hypothesis, which is robust to the choice of release and seasonally-adjustment. To further study the effects of COVID19 on information rigidities and the behavior of survey participants, private nonfarm payroll changes could be focused on due to the asymmetrical impact on small-sized firms and the service industry. It is left for further research if the predictions made for the private NFP changes draw a better picture of the actual decline and recovery.

The findings presented herein are of relevance as it is suggested that NFP consensus or survey forecasts suffer in precision when forecasters are participating less frequently. We further show that autoregressive models, unlike for other macroeconomic variables, show insufficient prediction quality. A deep learning network yields superior in-sample fit but does not outperform the consensus forecast in an out-of-sample exercise. Future research could address how a combination approach of time series models and consensus forecasts, as carried out in Ang et al. (2007) for example, benefits the prediction of nonfarm payroll changes. Having shown that the LSTM network yields superior fit but lower prediction performance than the consensus forecast, the LSTM should be extended in the number of features aiming for an improvement of prediction quality. In view of the effects of nonfarm payroll announcements on financial markets, additional attention could be dedicated to the shocks identified in the Davies & Lahiri (1995) framework to further dissect nonfarm payroll news impact.

References

Albanesi, S., & Kim, J. (2021). Effects of the COVID-19 Recession on the US Labor Market: Occupation, Family, and Gender. *Journal of Economic Perspectives*, *35*, 3–24. URL: https://pubs.aeaweb.org/doi/10.1257/jep.35.3.3. doi:10.1257/jep.35.3.3.

Ang, A., Bekaert, G., & Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, *54*, 1163–1212. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304393206002303. doi:10.1016/j.jmoneco.2006.04.006.

Armantier, O., Koşar, G., Pomerantz, R., Skandalis, D., Smith, K., Topa, G., & van der Klaauw, W. (2021). How economic crises affect inflation beliefs: Evidence from the Covid-19 pandemic. *Journal of Economic Behavior & Organization*, *189*, 443–469. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167268121001839. doi:10.1016/j.jebo.2021.04.036.

Bartik, A., Bertrand, M., Lin, F., Rothstein, J., & Unrath, M. (2020). *Measuring the labor market at the onset of the COVID-19 crisis*. Technical Report National Bureau of Economic Research Cambridge, MA. URL: http://www.nber.org/papers/w27613.pdf. doi:10.3386/w27613.

Beckmann, J., & Czudaj, R. L. (2020). Professional forecasters' expectations, consistency, and international spillovers. *Journal of Forecasting*, *39*, 1001–1024. URL: https://onlinelibrary.wiley.com/doi/10.1002/for.2675. doi:10.1002/for.2675.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*, 157–166. URL: https://ieeexplore.ieee.org/document/279181/. doi:10.1109/72.279181.

Boero, G., Smith, J., & Wallis, K. F. (2008). Evaluating a three-dimensional panel of point forecasts: The Bank of England Survey of External Forecasters. *International Journal of Forecasting*, *24*, 354–367. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207008000526. doi:10.1016/j.ijforecast.2008.04.003.

Borup, D., & Schütte, E. C. M. (2020). In Search of a Job: Forecasting Employment Growth Using Google Trends. *Journal of Business & Economic Statistics*, (pp. 1–15). URL: https://www.tandfonline.com/doi/full/10.1080/07350015.2020.1791133. doi:10.1080/07350015.2020.1791133.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control*. (2nd ed.). San Francisco: Holden-Day.

Brown, L. D., Gay, G. D., & Turac, M. (2008). Creating a "smart" conditional consensus forecast. *Financial Analysts Journal*, *64*, 74–86.

Bureau of Labor Statistics (2020). The Employment Situation - March 2020. *News Releases Bureau of Labor Statistics*, *USDL-20-05*, 1–41.

Capistrán, C., & Timmermann, A. (2009a). Disagreement and Biases in Inflation Expectations. *Journal of Money, Credit and Banking*, *41*, 365–396. doi:10.1111/j.1538-4616.2009.00209.x.

Capistrán, C., & Timmermann, A. (2009b). Forecast Combination With Entry and Exit of Experts. *Journal of Business & Economic Statistics*, *27*, 428–440. doi:10.1198/jbes.2009.07211.

Clements, M. P. (2020). Are Some Forecasters' Probability Assessments of Macro Variables Better Than Those of Others? *Econometrics*, *8*, 16. doi:10.3390/econometrics8020016.

Clements, M. P. (2021). Do survey joiners and leavers differ from regular participants? The US SPF GDP growth and inflation forecasts. *International Journal of Forecasting*, *37*, 634–646. doi:10.1016/j.ijforecast.2020.08.003.

Clements, M. P., & Galvão, A. B. (2021). Measuring the effects of expectations shocks. *Journal of Economic Dynamics and Control*, *124*, 104075. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165188921000105. doi:10.1016/j.jedc.2021.104075.

Coibion, O., & Gorodnichenko, Y. (2012). What Can Survey Forecasts Tell Us about Information Rigidities? *Journal of Political Economy*, *120*, 116–159. doi:10.1086/665662.

Coibion, O., & Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*, 2644–2678. URL: https://pubs.aeaweb.org/doi/10.1257/aer.20110306. doi:10.1257/aer.20110306.

Davies, A. (2006). A framework for decomposing shocks and measuring volatilities derived from multi-dimensional panel data of survey forecasts. *International Journal of Forecasting*, *22*, 373–393. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207005001184. doi:10.1016/j.ijforecast.2005.09.007.

Davies, A., & Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, *68*, 205–227. URL: https://linkinghub.elsevier.com/retrieve/pii/030440769401649K. doi:10.1016/0304-4076(94)01649-K.

Davies, A., & Lahiri, K. (1999). Re-examining the Rational Expectations Hypothesis Using Panel Data on Multi-Period Forecasts. In *Analysis of Panels and Limited Dependent Variable Models* (pp. 226–254). Cambridge University Press.

Demetrescu, M., Hanck, C., & Kruse, R. (2021). *Robust Inference under Time-Varying Volatility: A Real-Time Evaluation of Professional Forecasters*. Universitätsbibliothek Dortmund. URL: https://wisostat.uni-koeln.de/sites/statistik/user_upload/Real-Time_Evaluation_of_Professional_Forecasters.pdf.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*. doi:10.1214/ss/1032280214.

Dovern, J., & Weisser, J. (2011). Accuracy, unbiasedness and efficiency of professional macroeconomic forecasts: An empirical comparison for the G7. *International Journal of Forecasting*, *27*, 452–465. URL: https://linkinghub.elsevier.com/retrieve/pii/S016920701000110X. doi:10.1016/j.ijforecast.2010.05.016.

Dungey, M., & Hvozdyk, L. (2012). Cojumping: Evidence from the US Treasury bond and futures markets. *Journal of Banking & Finance*, *36*, 1563–1575. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378426612000064. doi:10.1016/j.jbankfin.2012.01.005.

Dungey, M., McKenzie, M., & Smith, L. V. (2009). Empirical evidence on jumps in the term structure of the US Treasury Market. *Journal of Empirical Finance*, *16*, 430–445. URL: https://linkinghub.elsevier.com/retrieve/pii/S0927539808001011. doi:10.1016/j.jempfin.2008.12.002.

D'Agostino, A., McQuinn, K., & Whelan, K. (2012). Are Some Forecasters Really Better Than Others? *Journal of Money, Credit and Banking*, *44*, 715–732. doi:10.1111/j.1538-4616.2012.00507.x.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*, 801–816. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207017300389. doi:10.1016/j.ijforecast.2017.03.004.

Edison, H. J. (1997). The Reaction of Exchange Rates and Interest Rates to News Releases. *International Journal of Finance & Economics*, *2*, 87–100. URL: http://doi.wiley.com/10.1002/%28SICI%291099-1158%28199704%292%3A2%3C87%3A%3AAID-IJFE39%3E3.0.CO%3B2-8. doi:10.1002/(SICI)1099-1158(199704)2:2<87::AID-IJFE39>3.0.CO;2-8.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*, 654–669. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221717310652. doi:10.1016/j.ejor.2017.11.054.

Fleming, M. J., & Remolona, E. M. (1999). What Moves Bond Prices? *The Journal of Portfolio Management*, *25*, 28–38. URL: http://jpm.pm-research.com/lookup/doi/10.3905/jpm.1999.319756. doi:10.3905/jpm.1999.319756.

Gregory, A. W., & Zhu, H. (2014). Testing the value of lead information in forecasting monthly changes in employment from the Bureau of Labor Statistics. *Applied Financial Economics*, *24*, 505–514. URL: http://www.tandfonline.com/doi/abs/10.1080/09603107.2014.887190. doi:10.1080/09603107.2014.887190.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, *79*, 453–497. doi:10.3982/ECTA5771.

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *06*, 107–116. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218488598000094. doi:10.1142/S0218488598000094.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780. URL: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735.

Isiklar, G., Lahiri, K., & Loungani, P. (2006). How quickly do forecasters incorporate news? Evidence from cross-country surveys. *Journal of Applied Econometrics*, *21*, 703–725. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.886. doi:10.1002/jae.886.

Janiak, A., Machado, C., & Turén, J. (2021). Covid-19 contagion, economic activity and business reopening protocols. *Journal of Economic Behavior & Organization*, *182*, 264–284. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167268120304753. doi:10.1016/j.jebo.2020.12.016.

Koop, G., & Potter, S. M. (1999). Dynamic Asymmetries in U.S. Unemployment. *Journal of Business & Economic Statistics*, *17*, 298–312. doi:10.1080/07350015.1999.10524819.

Kotchoni, R., Leroux, M., & Stevanovic, D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, *34*, 1050–1072. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.2725. doi:10.1002/jae.2725.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*, 689–702. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221716308657. doi:10.1016/j.ejor.2016.10.031.

Lahiri, K., & Sheng, X. (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics*, *25*, 514–538. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.1167. doi:10.1002/jae.1167.

Maas, B. (2020). Short-term forecasting of the US unemployment rate. *Journal of Forecasting*, *39*, 394–411. doi:10.1002/for.2630.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., & Tiao, G. C. (1998). Forecasting the U.S. Unemployment Rate. *Journal of the American Statistical Association*, *93*, 478–493. doi:10.1080/01621459.1998.10473696.

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*, . URL: http://www.ssrn.com/abstract=2740751. doi:10.2139/ssrn.2740751.

Ramchander, S., Simpson, M. W., & Chaudhry, M. (2003). The impact of inflationary news on money market yields and volatilities. *Journal of Economics and Finance*, *27*, 85–101. URL: http://link.springer.com/10.1007/BF02751592. doi:10.1007/BF02751592.

Rich, R., & Tracy, J. (2010). The Relationships among Expected Inflation, Disagreement, and Uncertainty: Evidence from Matched Point and Density Forecasts. *Review of Economics and Statistics*, *92*, 200–207. URL: https://direct.mit.edu/rest/article/92/1/200-207/57788. doi:10.1162/rest.2009.11167.

Rojas, F. L., Jiang, X., Montenovo, L., Simon, K., Weinberg, B., & Wing, C. (2020). *Is the Cure Worse than the Problem Itself? Immediate Labor Market Effects of COVID-19 Case Rates and School Closures in the U.S.*. Technical Report National Bureau of Economic Research Cambridge, MA. URL: http://www.nber.org/papers/w27127.pdf. doi:10.3386/w27127.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*, 565–578. doi:10.1002/for.1213.

Appendix A.

Appendix A.1. Data Cleaning

As mentioned above, the raw panel of monthly predictions contains $k^* = 239$ submitting accounts with a large number of entries and exits of submitting economists. In some instances, these submitting economists change within a submitting entity such as a bank or financial firm or rotate out and back in at a later point.

For example, the QES entries of *Morgan Stanley* originate from four individuals who report their estimates. The raw data set contains these submission as four separate, non-overlapping series. Hence, the raw data is processed to account for changes in the submitting account of an institution and merges these responses, if they are non-overlapping. This reduces spurious non-sampling bias due to differing economist identifiers of one submitting institution. However, this might also affect how individual bias is calculated as merging these non-overlapping predictions assumes that the forecasters are prone to the same forecasting bias within the Davies & Lahiri (1995) framework. On the other hand, we argue that the information set available to the forecaster should be very similar, if not identical, as they stem from the same firm. Note that we do not merge predictions if the affiliation is identical but the location or branch differs. This, for example, is the case with some submission by economists of different branches of UBS or JPMorgan, among others.

We further restrict the sample to submitting economists with at least three consecutive submissions to the survey. The merging as outlined above and this threshold reduces the panel to $k = 181$ participants in the QES. During the analysis, we further restrict this sample to economists with at least 10, 50, and 100 survey submissions and compare the results across these different groups.

Appendix B.

Appendix B.1. Additional Figures

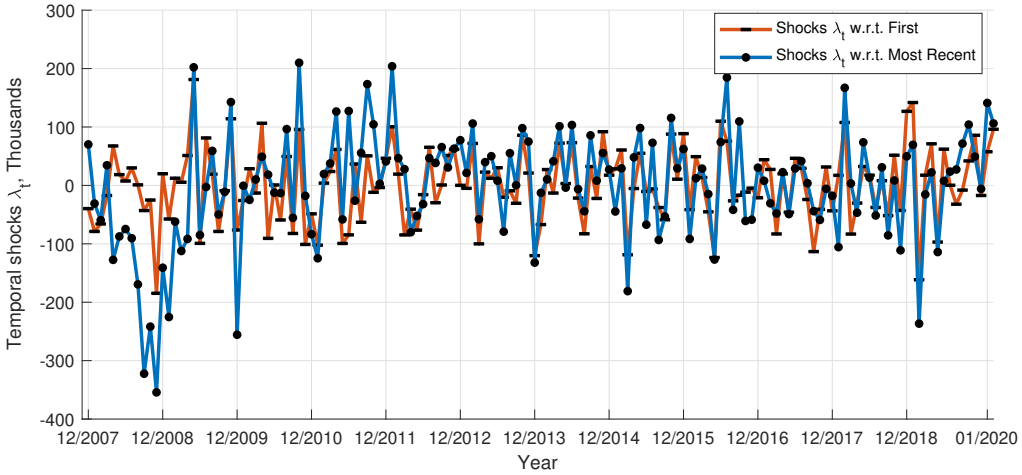


Figure B.9: Shock estimates λ_t based on the first release (orange line) and on the most recent, seasonally adjusted release (blue line).

Appendix C.

Appendix C.1. Additional Tables

			Cluster					
Loss	$\lvert T_i^* \rvert$		1	2	3	4	5	$\sum_i n_{C(i)}$
RMSE	≥ 50	$n_{C(i)}$	18	37	32	12	2	101
		$[C(i)]$	$[57.5, 64.3]$	$[64.6, 69.2]$	$[69.4, 74.4]$	$[74.8, 80.0]$	$[91.7, 99.2]$	
		$\overline{C}(i)$	62.1	67.1	71.7	77.4	95.4	
	≥ 100	$n_{C(i)}$	10	14	17	17	5	63
		$[C(i)]$	$[61.7, 64.6]$	$[64.9, 67.3]$	$[67.7, 70.4]$	$[71.1, 75.0]$	$[76.3, 79.2]$	
		$\overline{C}(i)$	63.3	66.5	68.7	72.7	77.9	

Table C.10: In-sample RMSE clusters of all economists with more than 50 or more than 100 predictions based on a k-means algorithm with $k = 5$ number of clusters. The number of economists per cluster is denoted by $n_{C(i)}$. The cluster boundaries are given as interval $[C(i)]$ with cluster centroid $\overline{C}(i)$.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5610	**0.6678**	**0.7747**	0.8942	1.3791	2.8301
CI	(0.0849; 0.4261)	(0.5043; 0.5947)	(0.6946; 0.7337)	(0.7865; 0.8224)	(0.8849; 0.9428)	(1.1162; 1.3475)	(1.5471; 4.3220)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5751	**0.6819**	**0.7861**	0.8930	1.3542	2.8301
CI	(0.3945; 0.5425)	(0.5681; 0.6299)	(0.7081; 0.7438)	(0.7895; 0.8252)	(0.8790; 0.9311)	(1.0684; 1.2423)	(1.3230; 2.3658)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5693	**0.5927**	**0.6779**	**0.7659**	0.8732	1.1353	2.4819
CI	(0.5061; 0.6128)	(0.6115; 0.6680)	(0.7163; 0.7523)	(0.7815; 0.8171)	(0.8495; 0.8957)	(0.9638; 1.0739)	(1.0661; 1.3973)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5693**	**0.6104**	**0.6926**	**0.7659**	0.8669	1.1217	1.3509
CI	(0.5919; 0.6755)	(0.6537; 0.7034)	(0.7243; 0.7615)	(0.7730; 0.8104)	(0.8237; 0.8699)	(0.9005; 0.9875)	(0.9463; 1.1397)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.11: Distribution of the forecasting performance relative to the third publication of NFP changes for the in-sample period from December 2007 to February 2020.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5505	0.6518	0.7433	0.8152	**0.9007**	**0.9265**
CI	(0.0957; 0.4336)	(0.4798; 0.5717)	(0.6500; 0.6843)	(0.7245; 0.7542)	(0.7994; 0.8440)	(0.9712; 1.1529)	(1.2333; 2.9347)
Panel A: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	**0.5724**	0.6650	0.7509	0.8232	**0.8990**	**0.9265**
CI	(0.4066; 0.5356)	(0.5456; 0.6020)	(0.6616; 0.6925)	(0.7276; 0.7558)	(0.7946; 0.8356)	(0.9356; 1.0666)	(1.0930; 1.6948)
Panel A: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5693	0.5833	**0.6628**	0.7494	0.8050	0.8787	**0.8929**
CI	(0.4912; 0.5843)	(0.5737; 0.6240)	(0.6636; 0.6948)	(0.7178; 0.7470)	(0.7699; 0.8062)	(0.8605; 0.9586)	(0.9401; 1.2036)
Panel A: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5693**	**0.6081**	**0.6810**	0.7510	0.7959	0.8723	0.8887
CI	(0.5728; 0.6488)	(0.6223; 0.6702)	(0.6835; 0.7163)	(0.7234; 0.7551)	(0.7621; 0.8001)	(0.8165; 0.8867)	(0.8467; 0.9836)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.12: Distribution of the forecasting performance relative to the third publication of NFP changes for the in-sample period from December 2007 to February 2020, restricted to the best 80% of forecasters.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5625	**0.6855**	**0.7785**	0.9049	1.4149	2.8301
CI	(0.0961; 0.4476)	(0.5237; 0.6142)	(0.7074; 0.7451)	(0.7953; 0.8308)	(0.8916; 0.9483)	(1.1193; 1.3570)	(1.5480; 4.1882)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.6034	**0.6933**	**0.7827**	0.9075	1.3934	2.8301
CI	(0.3910; 0.5600)	(0.5850; 0.6465)	(0.7183; 0.7541)	(0.7973; 0.8330)	(0.8856; 0.9395)	(1.0734; 1.2502)	(1.3376; 2.6176)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5626	**0.6097**	**0.6902**	**0.7693**	0.8681	1.1764	2.4561
CI	(0.5112; 0.6241)	(0.6218; 0.6786)	(0.7243; 0.7600)	(0.7882; 0.8232)	(0.8546; 0.9014)	(0.9684; 1.0781)	(1.0695; 1.3987)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5626**	**0.6080**	**0.6925**	**0.7693**	0.8674	1.1467	1.4240
CI	(0.5801; 0.6724)	(0.6515; 0.7038)	(0.7263; 0.7641)	(0.7767; 0.8139)	(0.8286; 0.8754)	(0.9070; 0.9969)	(0.9555; 1.1476)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.13: Distribution of the forecasting performance relative to the third publication of NFP changes for the in-sample period from December 2007 to December 2020.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.3442	0.6140	**0.7383**	**0.8366**	0.9647	1.4565	3.2762
CI	(0.2273; 0.5362)	(0.5952; 0.6804)	(0.7651; 0.8000)	(0.8439; 0.8776)	(0.9295; 0.9801)	(1.1162; 1.3024)	(1.4467; 3.4516)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4904	**0.6335**	0.7408	**0.8348**	0.9541	1.4700	3.2762
CI	(0.3961; 0.6040)	(0.6380; 0.7048)	(0.7723; 0.8059)	(0.8445; 0.8778)	(0.9232; 0.9707)	(1.0788; 1.2228)	(1.2993; 2.3497)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5938	**0.6485**	**0.7478**	**0.8281**	0.9137	1.1405	2.6088
CI	(0.5710; 0.6892)	(0.6874; 0.7390)	(0.7778; 0.8106)	(0.8347; 0.8682)	(0.8949; 0.9376)	(0.9913; 1.0851)	(1.0781; 1.3496)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5938**	**0.6532**	**0.7407**	0.8205	0.9054	1.0838	1.3816
CI	(0.6385; 0.7248)	(0.7037; 0.7531)	(0.7729; 0.8074)	(0.8188; 0.8536)	(0.8652; 0.9070)	(0.9344; 1.0126)	(0.9771; 1.1365)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.14: Distribution of the forecasting performance relative to the seasonally-adjusted, most recent publication of NFP changes for the in-sample period from December 2007 to December 2020.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.3442	0.5944	**0.7220**	0.7936	0.8654	**0.9731**	**1.0020**
CI	(0.2833; 0.5498)	(0.5813; 0.6613)	(0.7249; 0.7550)	(0.7869; 0.8153)	(0.8527; 0.8945)	(0.9993; 1.1488)	(1.2060; 2.6355)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4904	0.6265	**0.7252**	0.7945	0.8658	0.9728	**0.9954**
CI	(0.4284; 0.6000)	(0.6151; 0.6761)	(0.7278; 0.7564)	(0.7849; 0.8124)	(0.8450; 0.8829)	(0.9659; 1.0847)	(1.1070; 1.7066)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5938	**0.6455**	**0.7252**	0.7934	0.8475	0.9236	0.9572
CI	(0.5529; 0.6579)	(0.6479; 0.6971)	(0.7276; 0.7569)	(0.7741; 0.8015)	(0.8201; 0.8541)	(0.8931; 0.9689)	(0.9505; 1.1540)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	**0.5938**	**0.6387**	**0.7250**	0.7885	0.8447	0.9124	0.9340
CI	(0.6207; 0.6976)	(0.6723; 0.7188)	(0.7322; 0.7629)	(0.7696; 0.8001)	(0.8071; 0.8426)	(0.8594; 0.9241)	(0.8875; 1.0086)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.15: Distribution of the forecasting performance relative to the seasonally-adjusted, most recent publication of NFP changes for the in-sample period from December 2007 to December 2020, restricted to the best 80% of forecasters.

Panel A: All Economists							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5577	**0.6606**	0.7433	0.8283	**0.9164**	**0.9487**
CI	(0.1081; 0.4508)	(0.5001; 0.5904)	(0.6621; 0.6955)	(0.7329; 0.7626)	(0.8061; 0.8539)	(0.9833; 1.1746)	(1.2618; 2.9553)
Panel B: Economists with more than 10 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.4583	0.5803	0.6745	0.7572	0.8334	**0.9227**	**0.9570**
CI	(0.4004; 0.5520)	(0.5614; 0.6185)	(0.6731; 0.7036)	(0.7359; 0.7645)	(0.8028; 0.8444)	(0.9455; 1.0858)	(1.1116; 1.7774)
Panel C: Economists with more than 50 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5626	0.6080	0.6727	0.7408	0.8058	0.8735	**0.8966**
CI	(0.4942; 0.5933)	(0.5821; 0.6329)	(0.6702; 0.7010)	(0.7231; 0.7521)	(0.7747; 0.8114)	(0.8677; 0.9641)	(0.9473; 1.2023)
Panel D: Economists with more than 100 predictions							
	Best	5%	25%	50%	75%	95%	Worst
QES	0.5626	**0.6080**	**0.6711**	0.7426	0.7960	0.8678	0.8749
CI	(0.5597; 0.6418)	(0.6169; 0.6653)	(0.6811; 0.7137)	(0.7216; 0.7535)	(0.7610; 0.7990)	(0.8176; 0.8884)	(0.8486; 0.9884)

Distribution of the forecasting performance measured with the average normalized squared error proposed in D'Agostino et al. (2012). Numbers in parentheses are the bootstrap 10% confidence interval obtained from the BCa of DiCiccio & Efron (1996).

Table C.16: Distribution of the forecasting performance relative to the third publication of NFP changes for the in-sample period from December 2007 to December 2020, restricted to the best 80% of forecasters.