

Pros Vs. Joes: Comparing Professional and Crowdsourced Forecasting of NFP

Amar M. Patel | Peter N. Murphy | Advised by Professor Michael D. Aguilar

Herbert Brown Mayo Research Fellowship in Financial Economics, Summer 2017

Acknowledgements

We deeply appreciate the opportunity to work on this project backed by the Herbert Brown Mayo Summer Research Fund and made possible with the generous support from Leigh Drogan, founder and CEO of Estimote Inc., in making Estimote data available to us and providing us with guidance on our paper.

In addition, we would also like to thank Prof. Mike Aguilar for his early and continued advice during the Mayo project. From brainstorming topics of interest to advising us on the analysis of the paper, this endeavor would not have been made possible without his help.

“There is this to be said for the Many: each of them by himself may not be of a good quality; but when they all come together it is possible that they may surpass – collectively...although not individually – the quality of the few best...”

Aristotle, Politics, 3: Ch. 11

Introduction

Financial markets are a perfect example of this wisdom of crowds: together, multitudes of participants collectively process an equally vast amount of information to assign prices to assets. Economic indicators are a key type of this information and are a major driver of financial markets. With their regular release dates, they represent important “known unknowns” (Rumsfeld 2002) of which any investor should be conscious. Miao, Ramchander, and Zumwalt (2013) find "a strong association between macro news and price jumps. Over three-fourths of the price jumps between 8:30 am and 8:35 am and over three-fifths of the jumps between 10:00 am and 10:05 am are related to news released at 8:30 am and 10:30 am, respectively."

Change in Nonfarm Payrolls (NFP) is arguably the most important economic indicator. Compiled and released by the Bureau of Labor Statistics on the first Friday of each month, NFP is equally important on Main Street and Wall Street. It serves as both a barometer of wellbeing for working American families and a measure of the overall growth of the US economy. Miao, Ramchander, and Zumwalt (2013) find that "Among several types of news releases considered, Non-farm Payroll and Consumer Confidence are found to be significantly related to price jumps."

Along with significantly affecting financial markets, NFP is also a key influence on US monetary policy. Taylor (2010) discovered that “Non-farm Payrolls and Civilian Unemployment surprises are significantly (and consistently) priced in the federal funds futures market.” Specifically, “there is a significant relationship between abnormal federal funds futures rate changes and the unexpected component of these announcements.” Surprises in NFP data affect longer-dated fed funds futures because people assume that the Fed will incorporate these surprises

into their decision-making about US monetary policy: essentially, NFP is a significant piece of the Fed's monetary policy decisions. NFP clearly has wide-ranging effects, so having accurate forecasts is important to helping markets improve their information.

Currently, a variety of platforms exist for aggregating NFP forecasts, with differing degrees of accuracy. Most consist of the opinions of professional economists and researchers, often affiliated with large financial institutions and universities. In this paper, we explore the usefulness of a unique crowdsourced platform: Estimize. Unlike the various aggregations of forecasts made solely by professional economists, Estimize allows anyone with an internet connection to contribute their expectations of future NFP reports. In this paper, we ask the following:

1. What are Estimize economic forecasts?
2. How do Estimize NFP forecasts differ from traditional Consensus forecasts?
3. Are there certain characteristics that make Estimize and/or Consensus forecasts more accurate?

Estimize Background and Literature Review

Estimize was launched in 2011 as an “open financial estimates platform designed to collect forward looking financial estimates from independent, buy-side, and sell-side analysts, along with those of private investors and academics” (Estimize, Inc. [2017]). The platform began sourcing Earnings Per Share (EPS) estimates on equities and today has more than 50,000 contributors and 650,000 estimates across 2,200 stocks. On its EPS platform, Estimize’s user base is evenly split between investment professionals, independent researchers, individual traders, and students (Drogen and Jha [2013]). This differs from traditional aggregating platforms like the Institutional Brokers’ Estimate System (IBES), generally considered to be the consensus for EPS forecasting, which consists entirely of sell-side equity research analyst forecasts.

Several sources have found forecasts made on Estimize's EPS platform to be more accurate than traditional sell-side forecasts. Jame (2014) compares Estimize forecasts to those of IBES, and finds Estimize forecasts to be "equally accurate at shorter horizons" and "less biased and bolder (further from the combined IBES–Estimize consensus)" than IBES forecasts. Jame (2017) identifies several explanations of why IBES estimates may not incorporate the most up-to-date information: sell-side analysts are "dependent on managers for information and subsidized by investment banking revenues," which causes "analysts [to] have incentives to bias their research to please managers and facilitate investment banking activities."

Jame finds that approximately half of Estimize forecasts are issued in the two days prior to the earnings announcement date, while less than 2% of IBES forecasts are issued in the same period. This gives evidence to the argument that sell-side analysts are incentivized to not revise earnings forecasts that would "rock the boat," even when additional information is available to incorporate into forecasts.

Drogen and Jha (2013) find that Estimize is consistently more accurate (51%-65% of the time) in forecasting EPS, and its advantages increase as the number of analysts that cover a stock increases.

Estimize launched its Economics platform several years later in the first quarter of 2014. The platform provides users with the ability to forecast over 80 economic indicators across developed and major emerging markets. Major US indicators consistently receive close to 50 estimates per release, whereas international indicators typically receive less than 20.

Fig. 1a: Number of Forecasts Submitted per Month on Estimize, by Major Indicator

Month	ADP	Change in Nonfarm Payrolls	Change in Nonfarm Payrolls	Consumer Price Index	Core Consumer Price Index	Durable Goods New Orders	Existing Home Sales	Housing Starts
	Change in Nonfarm Payrolls							
4/2014	6	66	22	0	2	8	17	
5/2014	19	45	5	4	4	7	7	
6/2014	9	41	12	8	9	14	8	
7/2014	11	42	12	9	7	12	14	
8/2014	12	35	17	11	15	12	14	
9/2014	10	50	20	15	18	18	19	
10/2014	18	63	21	17	15	17	14	
11/2014	16	59	20	17	13	17	20	
12/2014	8	60	18	17	14	20	20	
1/2015	13	46	13	7	9	10	14	
2/2015	11	47	12	10	7	14	14	
3/2015	19	55	27	14	18	20	19	
4/2015	14	55	19	11	12	16	15	
5/2015	20	56	21	15	15	12	19	
6/2015	17	42	23	15	8	20	20	
7/2015	21	45	9	6	5	4	7	
8/2015	11	47	15	9	8	3	11	
9/2015	17	61	22	11	20	26	20	
10/2015	21	91	38	3	17	24	21	
11/2015	19	81	25	20	14	25	21	
12/2015	16	72	22	17	16	22	17	
1/2016	23	89	26	14	13	17	26	
2/2016	18	78	21	16	12	13	23	
3/2016	15	65	26	18	60	19	21	
4/2016	25	141	53	24	42	22	23	
5/2016	32	119	43	23	30	20	23	
6/2016	22	75	51	41	30	57	40	
7/2016	48	124	35	27	13	44	26	
8/2016	25	176	35	24	8	27	23	
9/2016	43	125	48	43	15	47	47	
10/2016	42	106	36	27	11	44	28	
11/2016	33	53	39	33	18	37	35	
12/2016	26	43	51	51	14	57	52	
1/2017	28	56	52	47	10	42	45	
2/2017	30	78	33	30	18	41	23	
3/2017	5	4	1	2	0	2	2	
Total	723	2,491	943	656	540	810	768	



 >50 Estimates

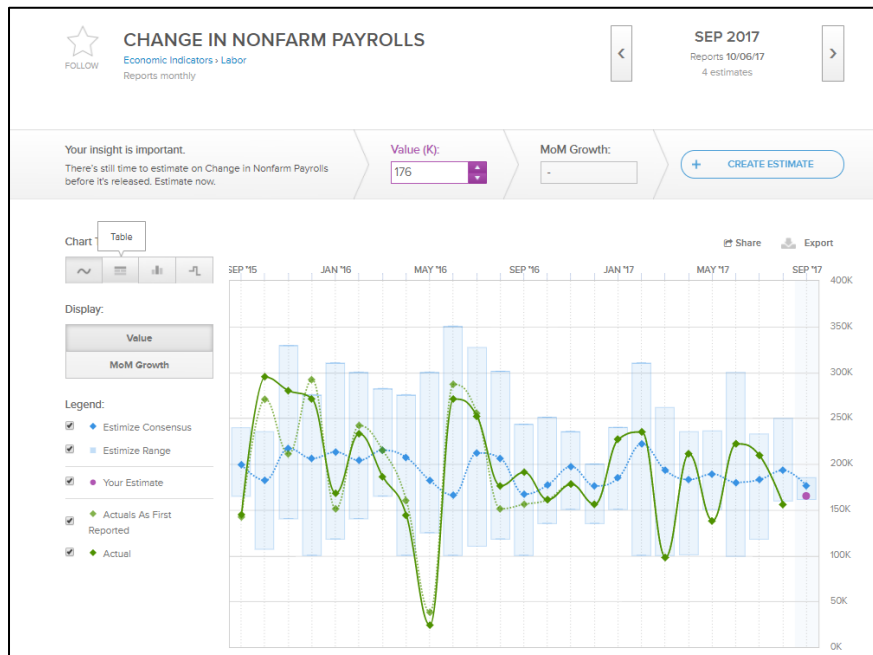
Fig. 1b: Number of Forecasts Submitted per Month on Estimize, by Major Indicator

Month	Industrial Production	ISM Non-Manufacturing Index	Manufacturing New Orders	New Single Family Houses Sold	Purchasing Managers Index	Umich Consumer Sentiment	Unemployment Rate
4/2014	8	12	3	7	5	0	19
5/2014	4	5	2	7	5	0	14
6/2014	4	3	3	10	3	0	13
7/2014	12	9	0	7	4	0	21
8/2014	12	10	6	19	2	0	17
9/2014	18	14	13	24	4	0	24
10/2014	19	19	10	18	13	0	26
11/2014	17	22	6	16	12	0	27
12/2014	13	9	10	16	5	0	34
1/2015	11	14	4	10	6	0	30
2/2015	12	6	14	8	9	0	26
3/2015	14	18	9	13	6	0	39
4/2015	16	14	9	15	14	0	30
5/2015	7	14	14	18	6	0	32
6/2015	8	7	21	15	11	0	29
7/2015	8	16	4	12	7	0	30
8/2015	6	17	11	10	12	0	31
9/2015	12	21	13	33	16	25	27
10/2015	15	20	10	14	21	20	45
11/2015	7	11	11	20	13	30	39
12/2015	25	17	11	18	17	14	31
1/2016	22	18	14	17	15	18	44
2/2016	22	18	16	13	18	13	37
3/2016	20	26	20	19	20	19	39
4/2016	25	26	24	26	57	43	85
5/2016	28	24	20	18	48	40	65
6/2016	23	26	20	43	38	27	51
7/2016	15	27	14	24	27	45	62
8/2016	12	19	14	20	17	27	57
9/2016	16	16	19	52	11	29	44
10/2016	18	24	10	33	21	49	82
11/2016	11	8	11	32	17	33	53
12/2016	41	10	11	45	16	31	56
1/2017	32	11	14	32	11	59	60
2/2017	6	10	1	43	13	28	59
3/2017	0	3	0	2	5	33	2
Total	539	544	392	729	525	583	1,380

 >50 Estimates

Nonfarm Payrolls are the most frequently forecasted indicator on Estimize, with 2,491 forecasts created from 4/2014 through 3/2017, compared with 1,380 for the next-highest indicator. This lends support to the argument that NFP is the most important and high-profile economic indicator.

Fig. 2: A Screenshot of Estimize’s Forecasting Interface



Estimize provides several measures for ease of use and quality control. A non-complicated user experience design allows for a broad range of people to use the platform, ensuring no one is disqualified from participating in the forums. The platform offers a visual history of Estimize’s accuracy as well as the actual values for the indicator’s release and subsequent revisions. After at least one forecast has been submitted for a monthly release of an indicator, an average will appear in the “Value” box, allowing users to observe what the community believes. Additionally, forecasts that are deemed unreasonably unrealistic or are entered in the wrong units are “flagged,”

making them easy to filter out of the data set.¹ In a system that is anonymous and open to the public, there are bound to be outliers. For the analysis in this paper, we remove all estimates that were flagged.

Data

For Estimize, we use an archive of every Estimize US economic forecast recorded since the Economic forecasting platform was launched in April 2014. The dataset provides user ID's, forecasts, actual release data, and timestamps of when each estimate was created, allowing for a variety of analyses.

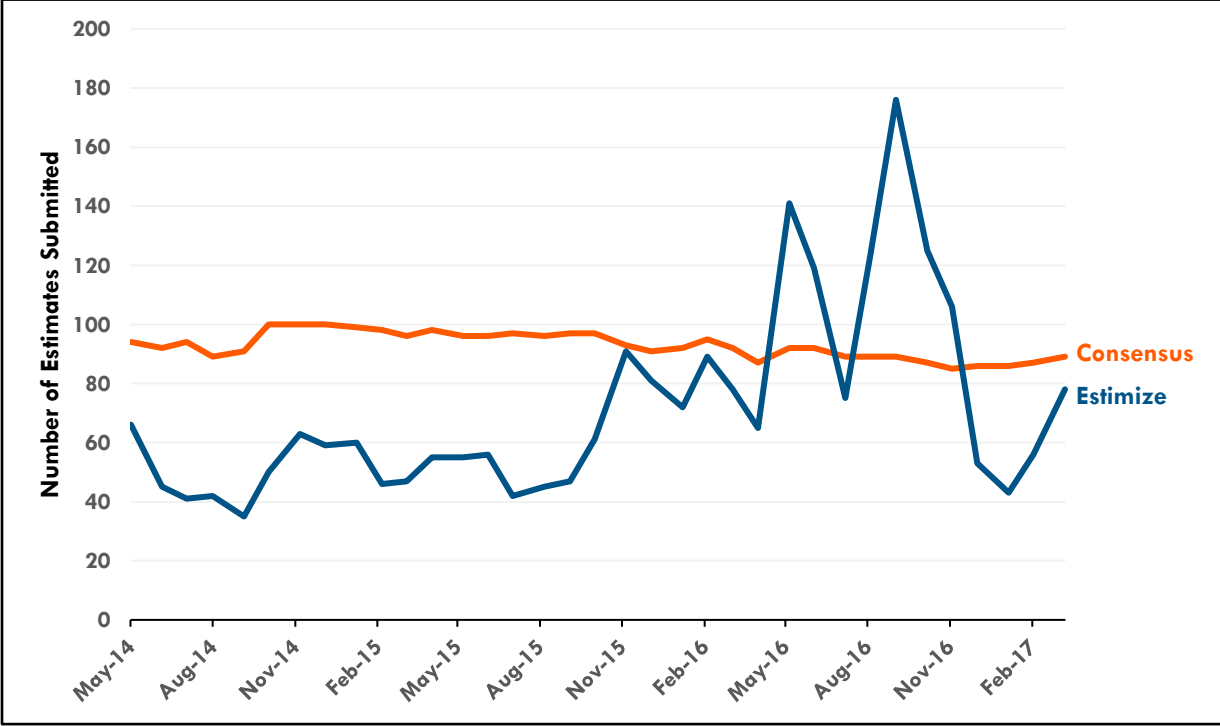
To compare the Estimize community's forecasts to more well-known professional forecasters, we use the Bloomberg Consensus as our professional data set. While there are many professional services that forecast consensus (e.g. Wall Street Journal, Consensus Economics, the Philadelphia Federal Reserve Branch's Survey of Professional Forecasters (SPF)), many of these sources are incompatible with our research due to differing time horizons (monthly forecasting on Estimize vs. quarterly for the SPF), pricing on data (Consensus Economics), or difficulty retrieving historical data (WSJ). Bloomberg is decidedly the best fit, providing user identity, date of forecast, and actual release data similar to Estimize. The forecast polls 80-100 economists, from both academia and major financial institutions each month, making it a suitable proxy for the consensus of professional economists.

The data covers the time period from the second quarter of 2014 through the first quarter of 2017. The time horizon is limited by the availability of Estimize data: the Estimize Economics platform was only launched in April 2014, and the data given to us runs through March 2017. While small sample size was a concern at first, we believe this 3-year horizon, which contains

¹ We do not know the exact methodology that Estimize uses to flag forecasts.

more than 30,000 estimates across all indicators and more than 2,400 estimates on NFP, should be a large enough sample to draw meaningful conclusions from.

Fig. 3: Number of NFP Estimates Submitted on Each Platform, per Month



As shown in Fig. 3, the consensus forecast routinely includes about 100 economists per month, while Estimize’s number of forecasts varies widely. While the number of Estimize forecasts per month generally increases over the three-year horizon as the forecasting platform gains popularity, the number of forecasters still varies more widely month-to-month than Consensus. Because March 2017 only contains 4 estimates on Estimize², we have omitted this month from our dataset.

² This is likely because the data was retrieved before all forecasts for March were recorded.

Methodology

Because Estimote's Economics platform is newer than its EPS platform, no academic literature focusing on the Estimote Economic Indicators yet exists, to the best of our knowledge.

We compare Estimote to Consensus (Bloomberg) in several dimensions. Following Jame (2014), we compute accuracy, bias, boldness, and dispersion for both Estimote and Consensus forecasts.

Accuracy is a directionally-agnostic measure of how close a forecast is to the actual indicator release for a given month. Accuracy is measured using the following metrics:

Absolute Forecast Error (AFE)	$ F - A $
Mean-Squared Forecast Error (MSFE)	$(F - A)^2$
Proportional Mean Absolute Forecast Error (PMAFE)	$(AFE - AFE_{avg}) / AFE_{avg}$

Where F is the predicted forecast and A is the value announced at the time of initial release (not a revised point at a later date). The value announced at the time of release tends to move markets more than subsequent revisions, so we use this first announced actual NFP value for creating our metrics.

Bias adds a directional element to accuracy. Bias is measured using the following metrics:

Directional Forecast Error (FE)	$(F - A)$
Percent Forecast Error	$(F - A) / A$
Standardized Score	$(F - A) / \sigma$ or FE / σ

Where F is the predicted forecast, A is the actual indicator as released, and σ is the standard deviation of all forecast errors in that period, for that grouping (Estimize or Bloomberg).

Boldness examines how far each forecast is from the average forecast, across both Estimize and Consensus. Boldness is defined as:

$$|F - F_{\text{avg, all}}| / F_{\text{avg, group}}$$

Where $F_{\text{avg, all}}$ is the average forecast from the combined groups of Estimize and Consensus made prior to Forecast F , and $F_{\text{avg, group}}$ is the average forecast within the group (Estimize or Consensus) made prior to Forecast F .

Dispersion measures how varied the forecasts are within a given month. Higher dispersion implies more uncertainty among forecasters. Dispersion is measured using the following metrics:

Variance	σ_F^2
Standard Error	σ_F / \sqrt{n}
Percentiles	Minimum 5% Quantile 25% Quartile Median 75% Quartile 95% Quantile Maximum
Ranges	Maximum – Minimum 95% - 5% Interquartile Range (IQR)

Where σ_F^2 is the variance in forecasts within one group, σ_F is the standard deviation of forecasts within one group, and n is the number of forecasts within one group.

Comparing Estimize and Consensus' forecasting power

Estimize and Consensus forecasts show what two different groups of forecasters are expecting from each month's NFP release. Which group of forecasters performs better on the metrics outlined above?

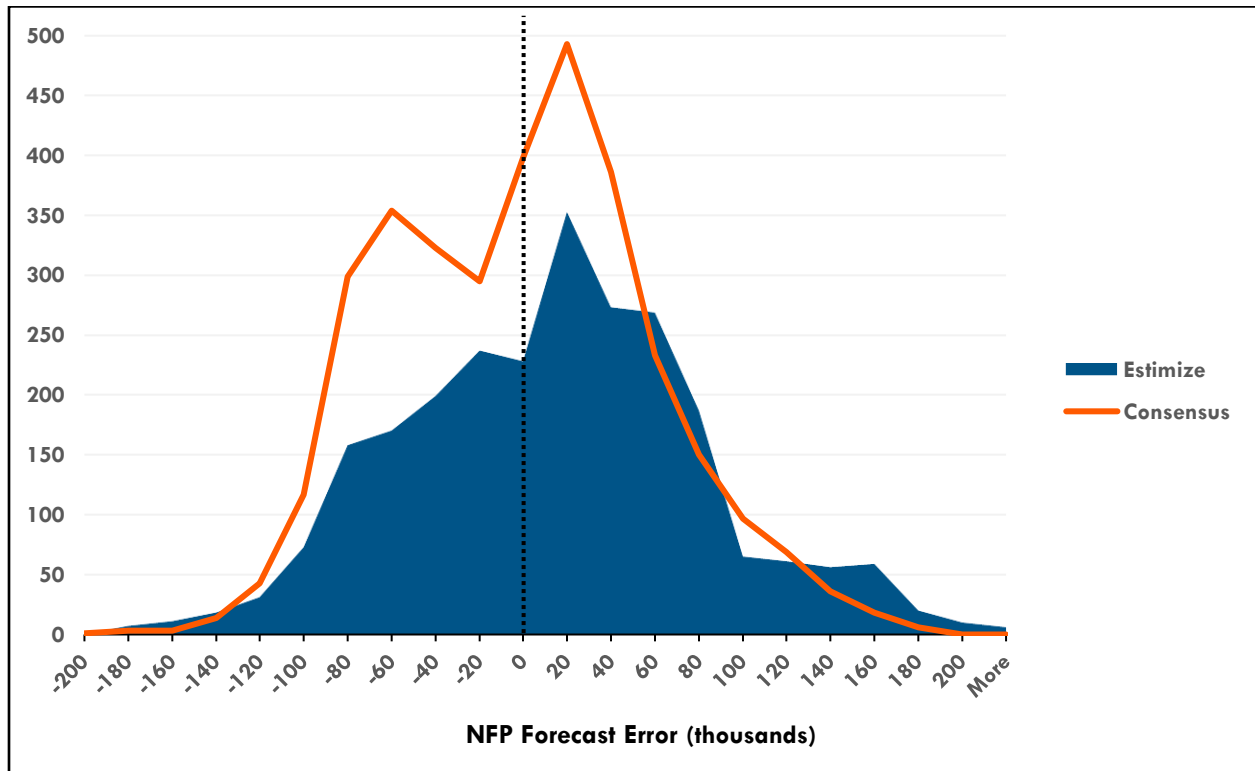
Fig. 4 compares Estimize and Consensus on the measures of accuracy, bias, and boldness described above. The p-values associated with two sampled t-tests for each metric are effectively 0, indicating that the errors between Estimize and the Bloomberg Consensus are not equal to a high degree of confidence. Consensus tends to be significantly more accurate on an absolute basis by roughly 5,000 jobs per month. Estimize tends to be less biased, with the average forecast error coming in only 5,000 jobs over the actual indicator, while Consensus estimates about 11,000 jobs below the actual value. Estimize tends to be bolder, meaning that its users make estimates that are further from the combined average of the two groups.

Fig. 4: Average Accuracy, Bias, and Boldness Metrics for Estimize and Consensus

	n	AFE	MSFE	PMAFE	Forecast Error	% FE	Standardized Score	Boldness
Estimize	2,487	54.62	4,720.24	0.01	4.69	-0.77	0.19	0.11
Consensus	3,251	48.22	3,578.42	-0.08	-13.53	-1.00	-0.58	0.09
<i>p-value</i>		<i>0.00</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>	<i>0.00</i>

In the distribution shown in Fig. 5, both groups' forecast errors display a central tendency near zero and appear positively skewed. Estimize's distribution shows a larger positive tail than Consensus, meaning that a number of Estimize users predict significantly higher than the actual result. The Consensus group is somewhat bimodal, with forecast errors concentrated around -60,000 and 20,000 jobs.

Fig. 5: Histogram of NFP Forecast Errors for Estimize and Consensus



Figs. 6a and 6b show metrics of dispersion for Estimize and Consensus. The tables reveal greater dispersion among Estimize forecasts. While Consensus has a greater overall range, this is most likely an outlier³. Estimize has greater standard deviation, IQR, and 95-5 percentile ranges than Consensus. 90% of Consensus forecasts tend to be within 100,000 jobs of each other—roughly 3.3 standard deviations. This is a very tight grouping among Consensus forecasts, showing evidence of herding among professional economists.

Fig. 6a: Measures of Dispersion for Estimize and Bloomberg

	<i>n</i>	Std. Dev	Std. Err	Range	IQR	95 - 5
Estimize	2487	38.12	0.76	250	47.88	125
Consensus	3251	31.19	0.55	320	42	98

³ We did not create any cut-off for outliers in the Consensus dataset, assuming that all participants polled were professional economists and did not contain the same issues that causes Estimize to “flag” certain results.

Fig 6b: Selected Quantiles, Estimize and Bloomberg

	Min	5%	25%	Median	75%	95%	Max
Estimize	100	150	184	210	232	275	350
Consensus	50	160	188	210	230	258	370

In summary, Estimize outperforms consensus on a pure measure of directional forecast error, statistically significant to a very high level. Consensus forecasts tend to exhibit lower dispersion and boldness, demonstrating herding behavior among professional economists. Estimize's forecasts are more dispersed, with a large positive skew.

Which other factors affect the accuracy, bias, boldness, and dispersion of forecasts?

From the above results, we can tell that Estimize tends to be more accurate and less biased, yet also tends to have higher dispersion between forecasts.

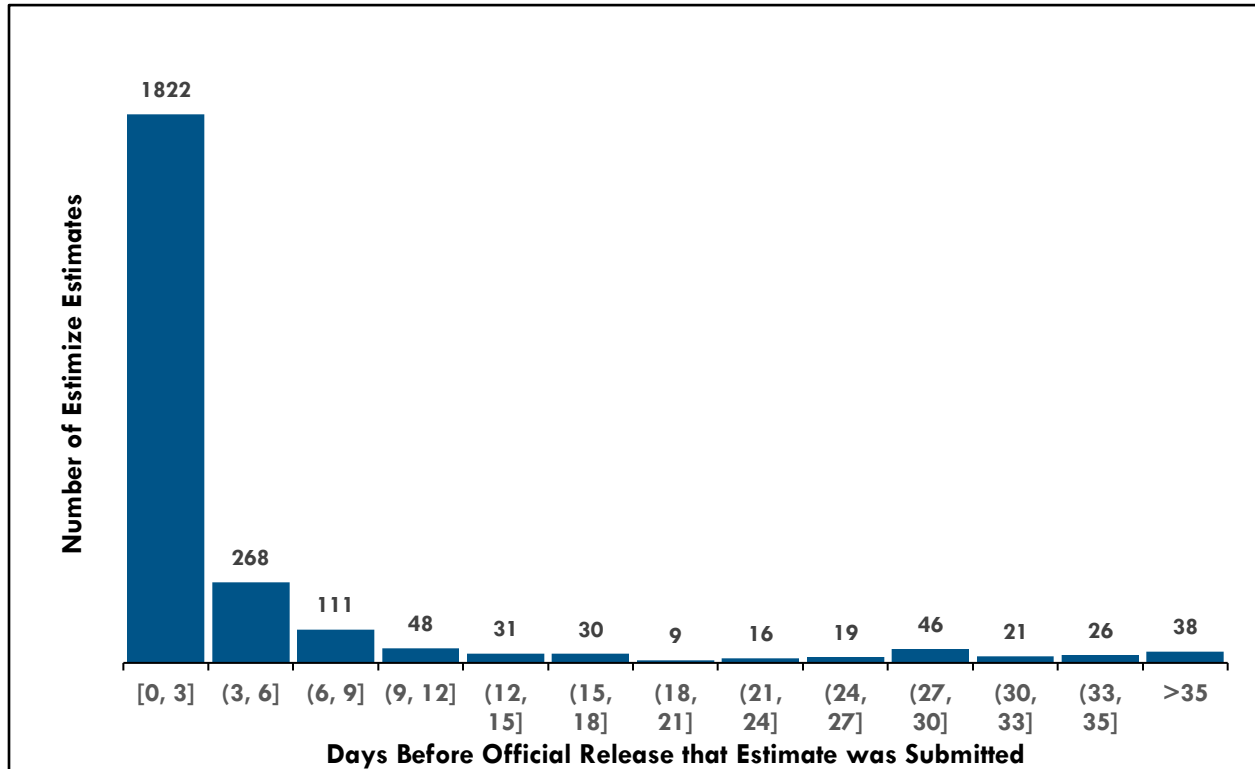
Next, we examine several other factors that may affect the quality of forecasts. One factor is *term structure*: does a forecast made closer to the release date incorporate new information, and therefore is it more accurate? Another factor is *seasonality*: are forecasters better in certain months of the year? A third factor is *dispersion*: in months with higher dispersion among forecasts, are these forecasts less accurate? We determine whether each of these factors affects forecasting quality, and whether they affect Estimize and Consensus differently.

Term Structure: Examining the time horizon of forecasts

The efficient market hypothesis assumes that forecasts incorporate all available information when they are made. This would imply that forecasts made closer to the release date would incorporate more information, and would therefore be more accurate.

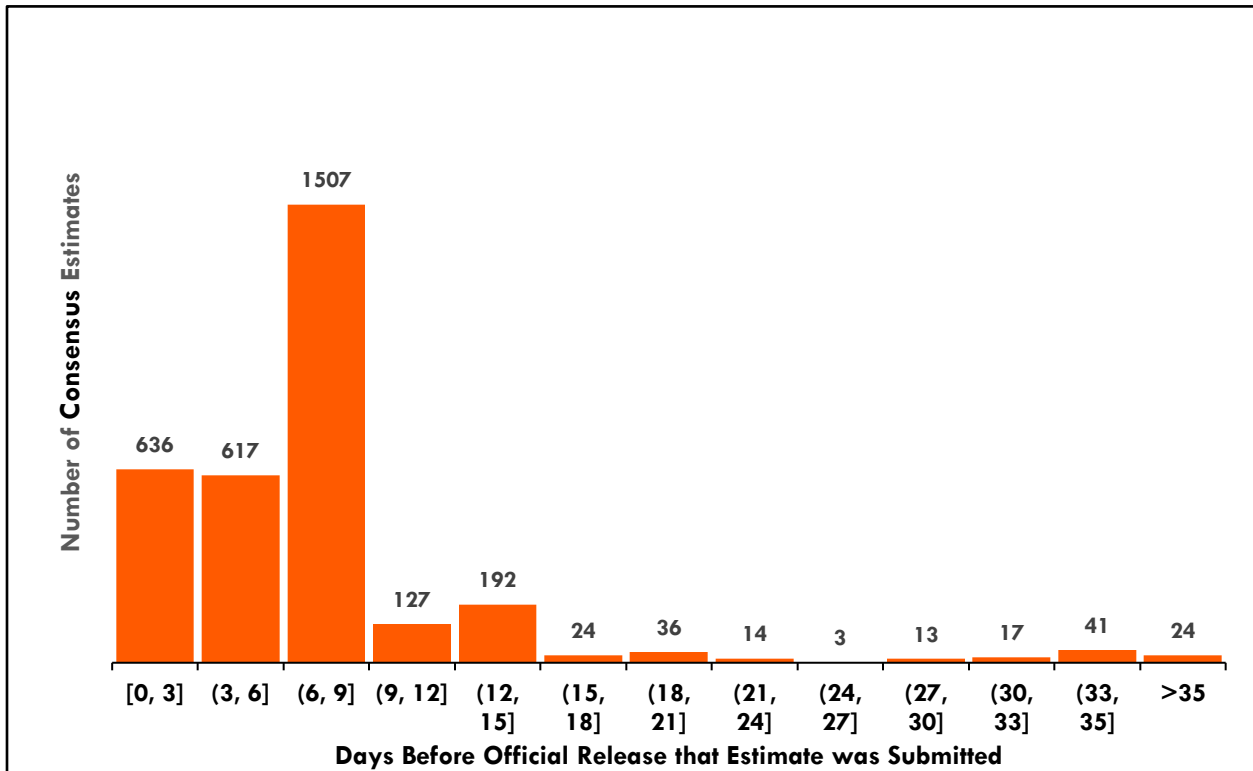
In fixed income, “term structure” refers to how interest rates vary with bond maturity. We adopt this term to refer to the number of forecasts and the variation in forecasts that occurs as the time horizons of these forecasts change.

Fig. 7a: Term Structure of Estimize Forecasts



Estimize’s term structure features a smooth drop-off over the first 30 days. Looking at the consensus, there is a large pick-up in forecasting activity about 10 days before the monthly indicator is released, with a less consistent drop-off over both this 10-day period and over time horizons greater than 10 days.

Fig. 7b: Term Structure of Consensus Forecasts



The term structure of Consensus forecasts does not exhibit the same smooth decrease present in the Estimize term structure. Instead, the majority of forecasts are made 7-9 days before an indicator is released. Consensus has fewer forecasts made beyond 15 days than does Estimize. Overall, Estimize forecasters make their forecasts closer to the release date.

Fig. 8a: Absolute Forecast Error by Days Before Release, Estimimize and Consensus

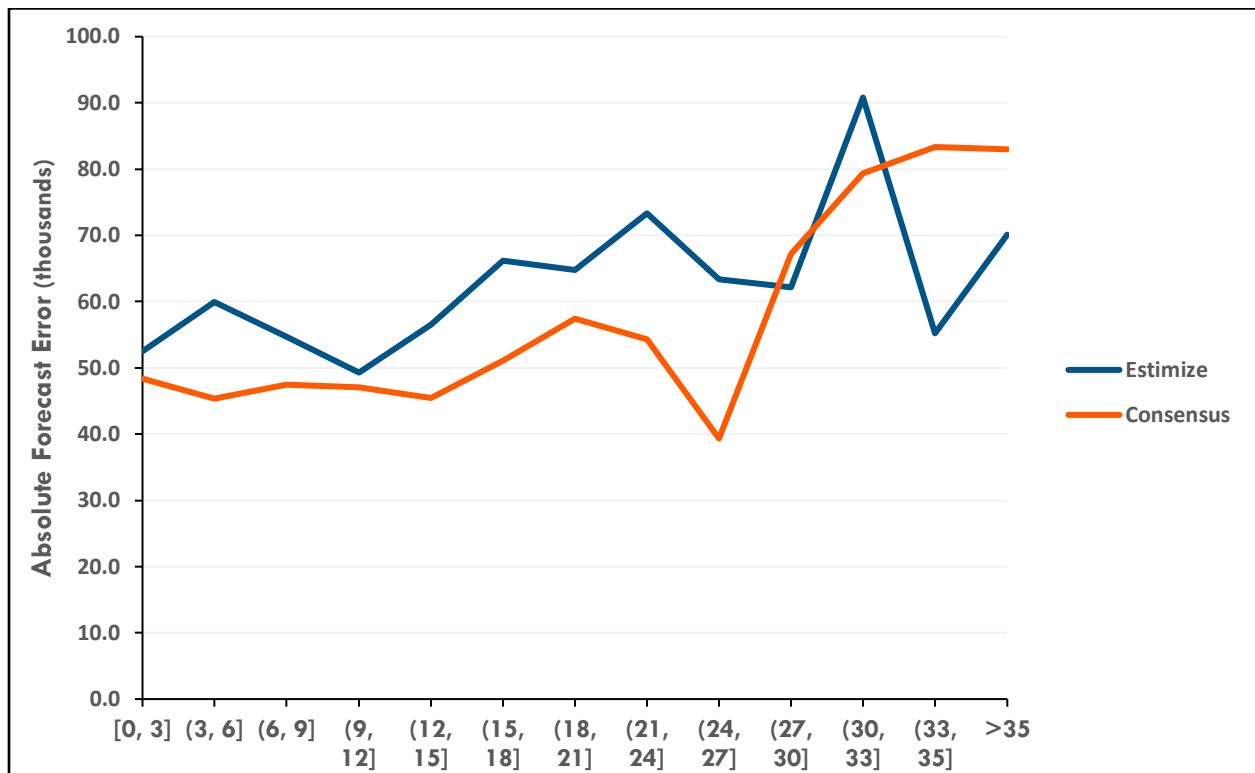
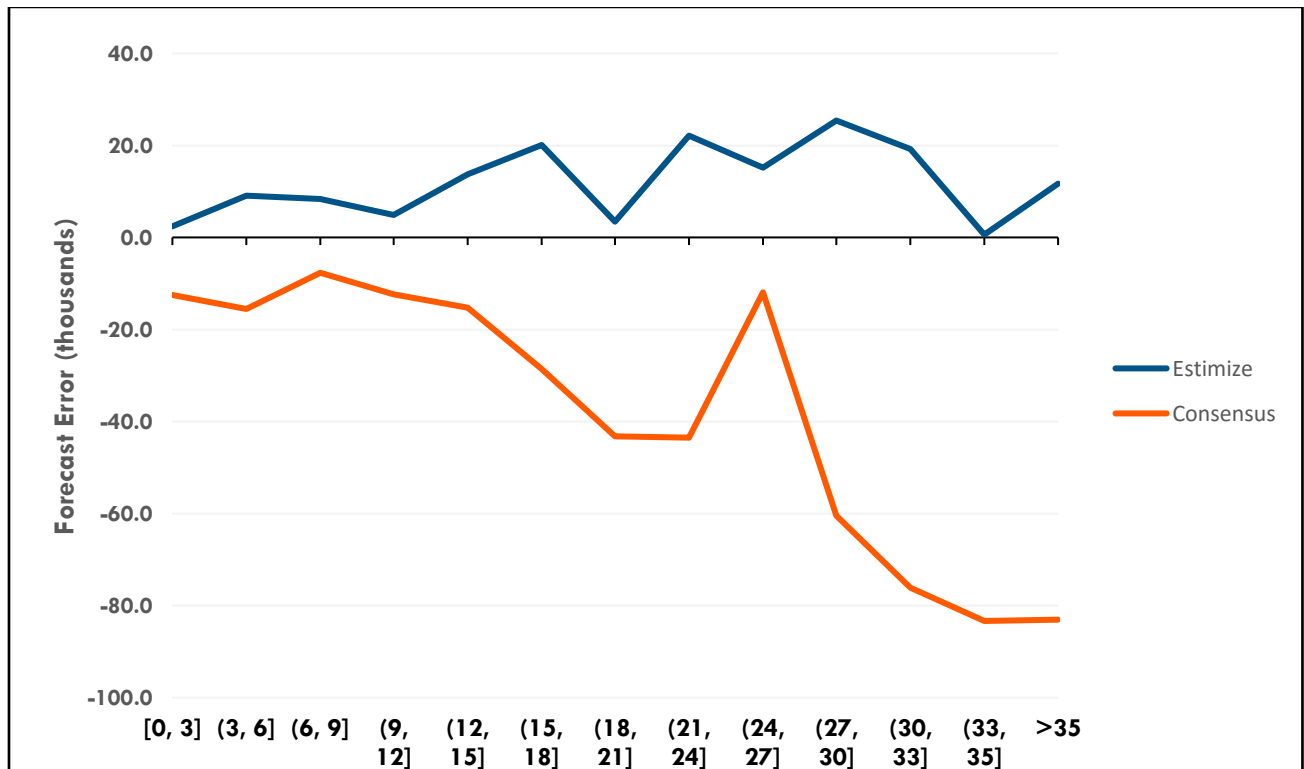


Fig. 8b: Forecast Error by Days Before Release, Estimimize and Consensus



The graphs in Figs. 8a and 8b casually show how the accuracy and bias change as the time horizon increases. Although the term structure of forecasts is different for Estimize and Consensus, the accuracy of both datasets tends to be similar. As time horizons increase, the groups exhibit different biases. Estimize bias remains positive and relatively flat. Consensus drops off sharply, with forecasts that were made 15 or more days out severely underestimating the number of jobs.

Fig. 9a: Summary Table by Time Horizon of Forecast, Estimize

Time Horizon	n	AFE	MSFE	PMAFE	Forecast Error	% FE	Standardized Score	Boldness
[0, 3]	1822	52.49	4,421	-0.02	2.38	-0.79	0.10	0.10
(3, 6]	268	59.91	5,256	0.09	9.06	-0.73	0.35	0.10
(6, 9]	111	54.74	4,542	0.08	8.34	-0.83	0.36	0.11
(9, 12]	48	49.29	4,015	-0.10	4.81	-0.79	0.17	0.10
(12, 15]	31	56.52	4,748	0.09	13.67	-0.76	0.66	0.12
(15, 18]	30	66.23	6,436	0.12	20.05	-0.60	0.75	0.13
(18, 21]	9	64.82	6,016	0.10	3.49	-0.58	-0.03	0.12
(21, 24]	16	73.36	8,417	0.14	22.13	-0.60	0.74	0.15
(24, 27]	19	63.36	5,471	0.23	15.17	-0.57	0.56	0.09
(27, 30]	46	62.17	5,933	0.16	25.40	-0.55	0.85	0.16
(30, 33]	21	90.81	12,026	0.22	19.28	-0.11	1.04	0.11
(33, 35]	26	55.18	4,951	0.10	0.61	-0.95	0.17	0.15
>35	38	70.08	7,481	0.26	11.62	-0.78	0.44	0.15

Fig. 9b: Summary Table by Time Horizon of Forecast, Consensus

Time Horizon	n	AFE	MSFE	PMAFE	Forecast Error	% FE	Standardized Score	Boldness
[0, 3]	636	48.33	3,819	-0.07	-12.52	-1.00	-0.51	0.11
(3, 6]	617	45.32	3,135	-0.06	-15.51	-1.02	-0.75	0.08
(6, 9]	1507	47.47	3,443	-0.07	-7.68	-0.97	-0.30	0.08
(9, 12]	127	47.02	3,393	-0.06	-12.32	-1.00	-0.53	0.11
(12, 15]	192	45.50	3,315	-0.11	-15.28	-1.01	-0.63	0.11
(15, 18]	24	51.08	3,956	0.00	-28.58	-1.08	-1.25	0.13
(18, 21]	36	57.42	5,122	0.01	-43.19	-1.14	-2.10	0.12
(21, 24]	14	54.29	4,398	-0.05	-43.57	-1.15	-2.19	0.14
(24, 27]	3	39.33	2,485	-0.12	-12.00	-1.02	-0.64	0.08
(27, 30]	13	67.23	5,920	-0.30	-60.46	-1.20	-2.56	0.13
(30, 33]	17	79.35	6,831	-0.33	-76.06	-1.27	-3.27	0.10
(33, 35]	41	83.34	7,249	-0.30	-83.34	-1.29	-3.58	0.07
>35	24	83.00	7,349	-0.30	-83.00	-1.29	-3.56	0.09

Figs. 9a and 9b show the aforementioned metrics over the time horizon for both Estimize and Consensus. Based on these tables, forecasts made closer to date seem to be more accurate on average. For Estimize, absolute forecast errors remain below 60k until 15 days out, when they pick up and remain above 60k for the rest of the time horizons. Estimize directional forecast errors also tend to increase slightly after 15 days as well. Estimize forecasts made within 3 days of release are amazingly accurate, missing the actual value by an average of 2,380 jobs over a sample of 1,822 forecasts. Consensus forecasts show a gradual decrease in accuracy as time horizon is increased. Forecasts made further in advance also tend to be more negatively biased, meaning that Consensus forecasters tend to be more pessimistic about job growth at longer time horizons. Standardized scores follow a similar trend as forecast error: forecasts made 30 or more days in advance are more than three standard deviations below the average forecast error. From these tables, forecast error has a greater effect on Consensus forecasts than Estimize.

Perhaps there are additional nuances that are not being captured by the histogram buckets. To provide more rigor, we decided to utilize several OLS regressions to examine the effects of time horizon on accuracy and bias. These regressions use the individual forecasts, not the time horizon buckets used in the histograms and data tables above, to create a more continuous rather than categorical model. With a larger sample size (2500-3000 forecasts vs. 5-10 buckets), this model should have more statistical power.

To analyze accuracy, time horizon (variable name `days_early`) is used as an explanatory variable on absolute forecast error. To analyze bias, directional forecast error is regressed on the same explanatory variable of time horizon.

Fig. 10a: OLS Regression – Forecast timing as explanatory variable on AFE, Estimize

<i>Regression Statistics</i>								
Multiple R	0.08							
R Square	0.01							
Adjusted R Square	0.01							
Standard Error	41.57							
Observations	2,487.							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	25,988	25,988	15.04	0.00			
Residual	2,485	4,294,654	1,728					
Total	2,486	4,320,642						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	53.58	0.88	61.23	0.00	51.87	55.30	51.87	55.30
days_early	0.18	0.05	3.88	0.00	0.09	0.28	0.09	0.28

Fig. 10b: OLS Regression, Forecast timing as explanatory variable on AFE, Consensus

<i>Regression Statistics</i>								
Multiple R	0.13							
R Square	0.02							
Adjusted R Square	0.02							
Standard Error	35.09							
Observations	3,251							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	73,915	73,915	60.02	0.00			
Residual	3,249	4,001,379	1,232					
Total	3,250	4,075,294						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	42.52	0.96	44.37	0.00	40.65	44.40	40.65	44.40
days_early	0.78	0.10	7.75	0.00	0.58	0.98	0.58	0.98

Fig. 10c: OLS Regression, Forecast timing as explanatory variable on FE, Estimize

<i>Regression Statistics</i>								
Multiple R	0.06							
R Square	0.00							
Adjusted R Square	0.00							
Standard Error	68.47							
Observations	2,487.00							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	35,514	35,514	7.58	0.01			
Residual	2,485	11,648,940	4,688					
Total	2,486	11,684,453						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	3.49	1.44	2.42	0.02	0.66	6.31	0.66	6.31
days_early	0.22	0.08	2.75	0.01	0.06	0.37	0.06	0.37

Fig. 10d: OLS Regression, Forecast timing as explanatory variable on FE, Consensus

<i>Regression Statistics</i>								
Multiple R	0.18							
R Square	0.03							
Adjusted R Square	0.03							
Standard Error	57.37							
Observations	3,251.0							
	0							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	344,425	344,425	104.64	0.00			
Residual	3,249	10,694,012	3,291					
Total	3,250	11,038,436						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-1.24	1.57	-0.79	0.43	-4.31	1.83	-4.31	1.83
days_early	-1.69	0.16	-10.23	0.00	-2.01	-1.36	-2.01	-1.36

Figs. 10a and 10b show that the trend observed from visual representations of the data is statistically significant at a very high level. This holds true for both Estimize and Consensus, with t statistics of 3.88 and 7.75 respectively. Consensus forecasts are more sensitive to time horizons than Estimize: for every additional day out, Estimize is less accurate by 180 jobs, while Consensus misses the actual value by 780 jobs.

Figs. 10c and 10d show that the relationship between time horizon and forecast error is also significant for both Estimize and Consensus. In the Estimize regression, the two factors exhibit a positive relationship; for every additional day out, Estimize forecasts are an additional 220 jobs too high. Consensus exhibits the opposite relationship: increasing the time horizon by a day tends to make forecasts underestimate the actual release by 1,690 jobs.

Overall, the timing of forecast plays a role in how accurate forecasts are. For both Estimize and Consensus, the relationship between forecast error—both absolute and directional—and time horizon of forecast is statistically significant.

Seasonality of forecasts: do forecast errors follow a seasonal pattern?

Many economic indicators show some degree of seasonality. As such, many are seasonally adjusted, including NFP. Despite this adjustment, seasonality models can be flawed, as many have argued with GDP over the last few years (Lunsford [2017]). As accounting for seasonality is difficult even for professional statisticians and economists at the BEA, Federal Reserve, and other organizations, it is likely that both professional economists and Estimize users struggle with it as well. To test for seasonality, we compute aforementioned metrics for Estimize and Consensus on a monthly basis.

Fig. 11a: Absolute Forecast Error by Month

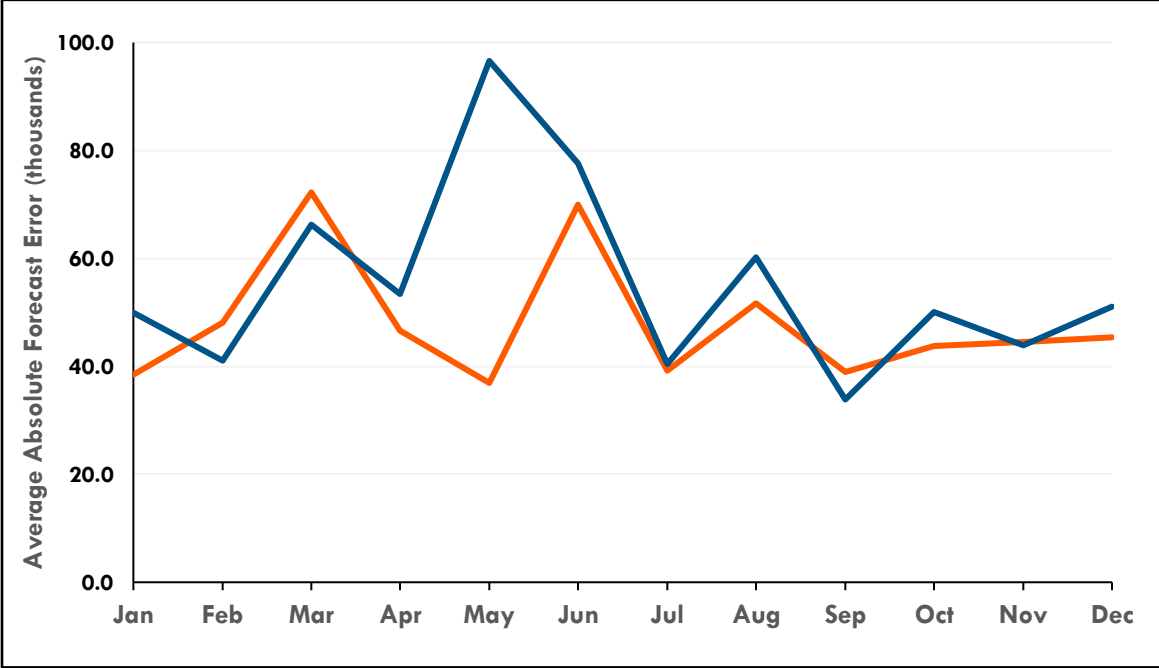
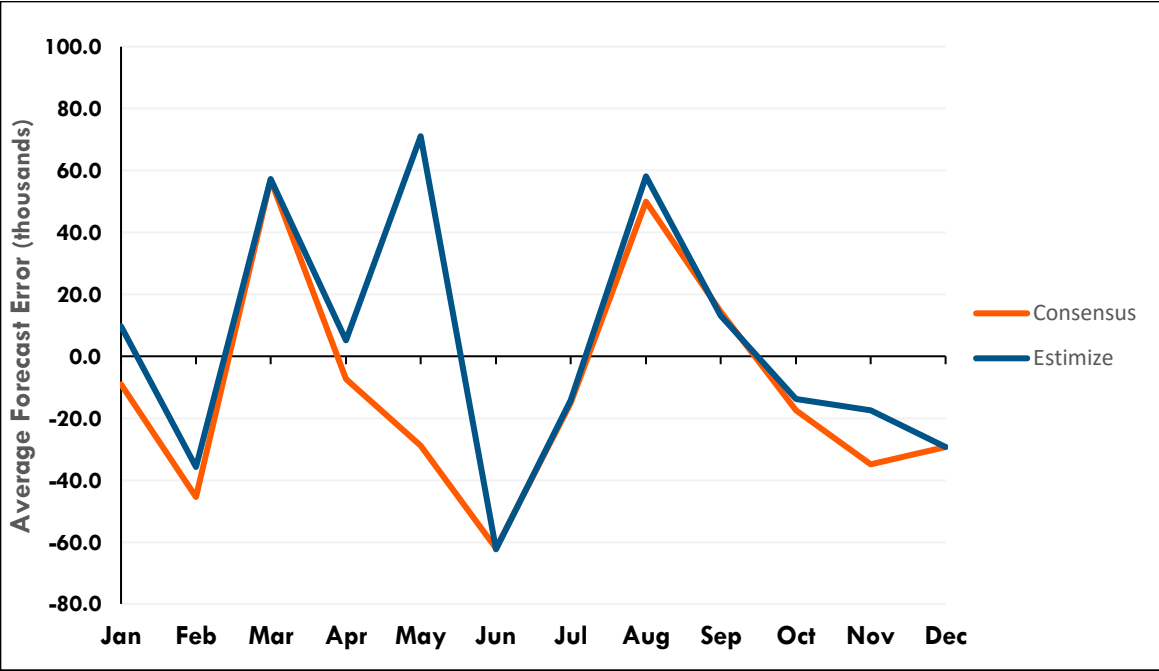


Fig. 11b: Forecast Error by Month



The data does not follow a smooth annual cycle, but the Estimize and Consensus datasets do seem to follow very similar patterns. Errors in AFE and FE both tend to be lower in winter months and increase during the summer.

Examining the average number of forecasts made by month, Consensus tends to receive the most estimates in June, at 372 over the 3-year period compared to about 260 on average for the other months. Estimize receives a lower-than-average number of estimates in June, at 158 compared to an average of 211 per month during the rest of the year⁴.

Fig. 12a: Seasonality of forecasts, Estimize

Month	<i>n</i>	AFE	PMAFE	FE	% FE	Standardized Score	Boldness
1	191	49.97	0.06	9.77	-0.89	0.22	0.11
2	203	41.11	-0.09	-35.73	-1.14	-1.38	0.10
3	120	66.25	-0.06	57.17	-0.55	1.66	0.10
4	262	53.39	0.05	5.23	-0.91	0.18	0.11
5	220	96.58	0.12	71.09	1.05	2.39	0.12
6	158	77.61	-0.03	-62.36	-1.21	-1.55	0.15
7	211	40.41	-0.08	-14.22	-1.05	-0.32	0.12
8	258	60.24	0.13	58.17	-0.62	1.87	0.12
9	236	33.85	-0.06	13.12	-0.89	0.96	0.08
10	260	50.11	0.05	-13.72	-1.02	-0.52	0.09
11	193	43.88	-0.01	-17.39	-1.03	-0.80	0.11
12	175	51.08	-0.03	-29.35	-1.08	-1.00	0.08

⁴ This variation in number of Estimize forecasts per month seems more attributable to the gradual ramp-up in activity over the 3-year window, which might give certain months more data than others. However, the Consensus forecast service on Bloomberg has been active for at least 10 years, making the June increase in forecasts less attributable to variation.

Fig. 12b: Seasonality of forecasts, Consensus

Month	<i>n</i>	AFE	PMAFE	FE	% FE	Standardized Score	Boldness
1	280	38.45	-0.07	-9.11	-1.00	-0.50	0.09
2	277	48.11	0.03	-45.55	-1.17	-1.80	0.10
3	185	72.22	-0.06	57.00	-0.53	2.48	0.09
4	282	46.60	-0.11	-7.38	-0.98	-0.51	0.10
5	188	36.92	-0.10	-28.81	-1.10	-1.16	0.09
6	372	69.98	-0.10	-62.02	-1.21	-2.51	0.09
7	274	39.15	-0.03	-15.14	-1.05	-1.04	0.10
8	277	51.65	-0.14	50.01	-0.67	2.38	0.09
9	284	38.94	-0.02	14.48	-0.87	0.68	0.08
10	278	43.81	-0.13	-17.41	-1.05	-0.72	0.09
11	277	44.52	-0.12	-34.83	-1.11	-1.50	0.09
12	277	45.42	-0.05	-29.29	-1.08	-1.33	0.09

Both Estimize and Consensus seem to follow a similar seasonal pattern of forecasting errors, with low AFE scores in fall and winter months. Bias has less of a seasonal trend than absolute accuracy. While there visually appears to be a seasonal pattern of accuracy for both groups, we have not conducted tests to examine the statistical significance of this casual observation.

Dispersion: do months with wider dispersion of forecasts exhibit larger errors?

Intra-month dispersion, defined by a larger standard deviation of forecasts made within a given month (σ_F), could likely indicate greater uncertainty among forecasters about the month's NFP release. In a month where forecasters are more uncertain, it is likely that their forecasts are more inaccurate and therefore that month exhibits larger errors.

Fig. 13a: Percentile Ranges for Estimize

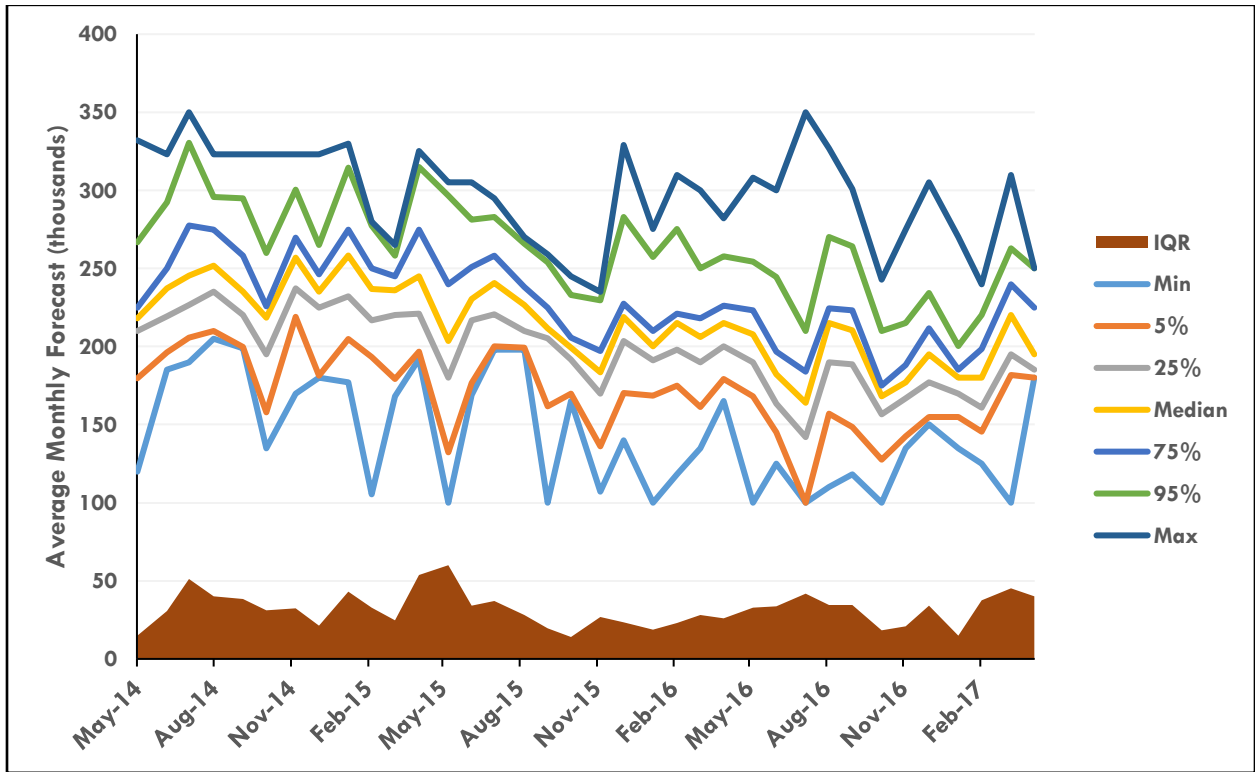
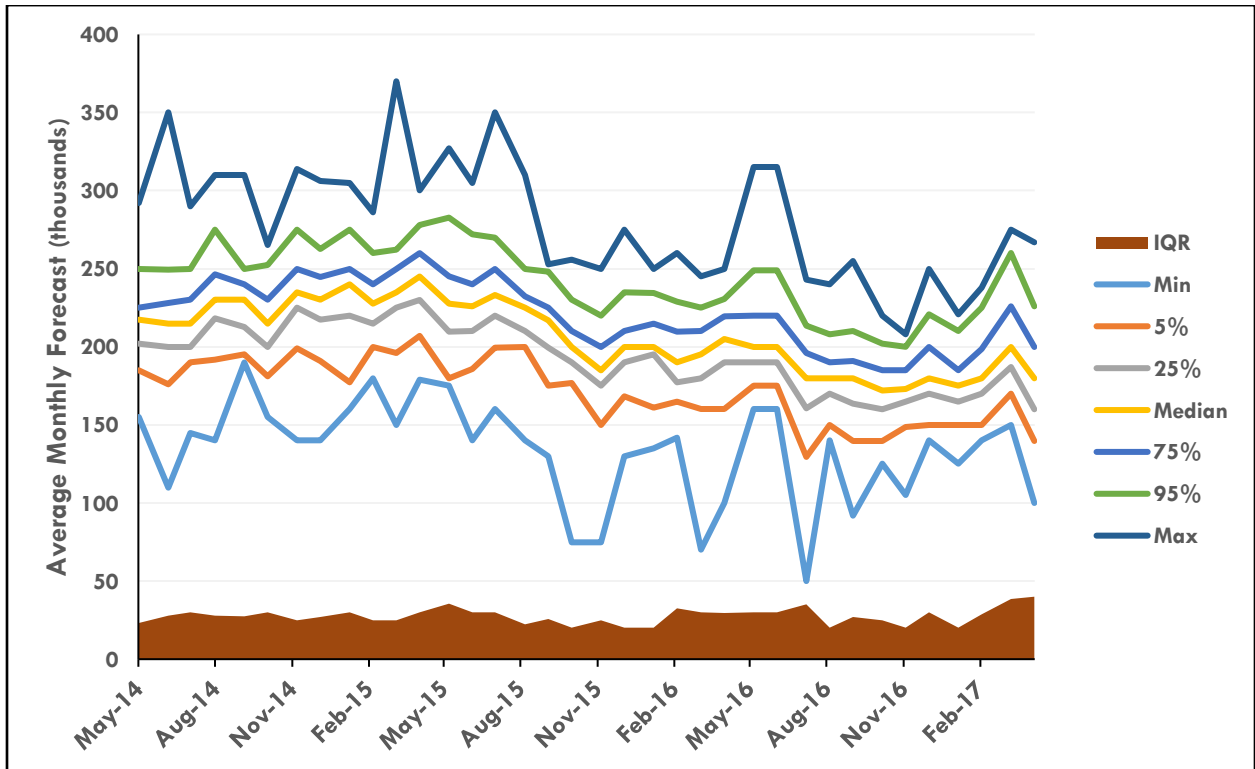


Fig. 13b: Percentile Ranges for Consensus



Figs. 13a and 13b examine dispersion within the forecasts of each month. Consensus exhibits a consistent IQR of about 25,000, while the IQR for Estimize is much more volatile. At the top of the distribution, the groups differ sharply: The Max and the 95th percentile of the Consensus tend to move together, while the Max and 95th percentile of Estimize move in opposite directions at times, such as mid-2016, where most forecasts decreased while the top 5% increased. Even when most forecasters are predicting a decrease in NFP, a small subset predict high amounts of growth. The Consensus distribution from 5% to 95% is remarkably close together, which gives evidence to our theory that professional forecasters tend to cluster. Estimize tends to exhibit wider dispersion compared to Consensus, especially in the 75-95% range.

To test whether this observed dispersion affects forecasting ability, we examine the relationship between forecast error (both absolute and directional) and dispersion, defined as the standard deviation of forecasts within each month.

Fig. 14a: Scatterplot, Monthly Standard Deviation vs. Absolute Forecast Error, Average by Month, Estimize and Consensus

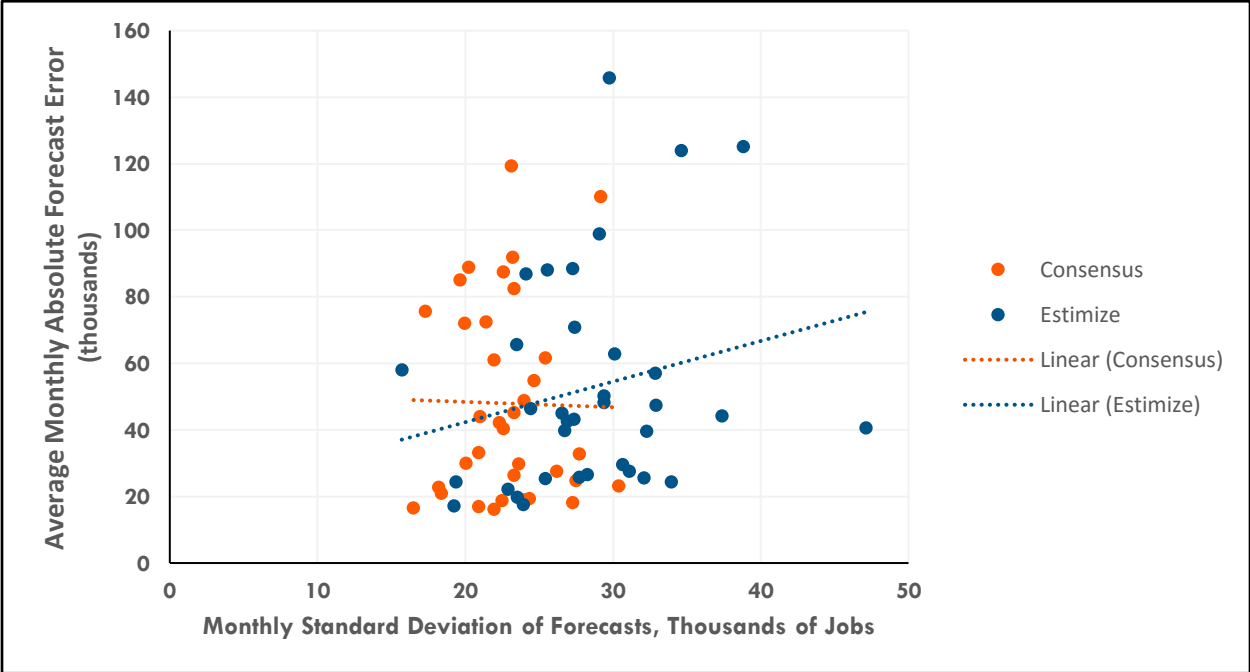


Fig. 14b: Scatterplot, Monthly Standard Deviation vs. Forecast Error, Average by Month, Estimize and Consensus

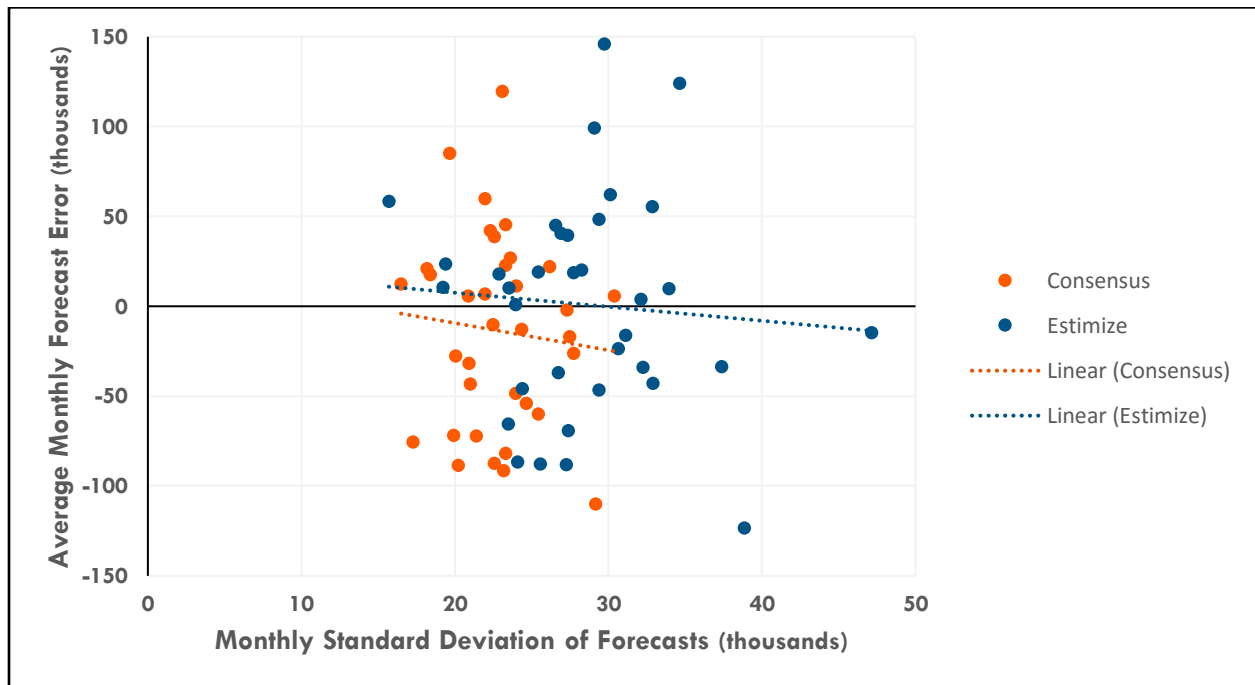


Fig. 14a displays a positive relationship between dispersion and forecast error for Estimize, and a slightly negative relationship for Consensus. Fig. 14b shows a negative relationship for both groups. On both scatterplots, Estimize tends to have higher dispersion, and especially high forecast errors in months with high dispersion. Note the narrow range of standard deviations for Consensus, showing that this group exhibits less month-to-month dispersion.

This analysis of visual data patterns is casual evidence of a relationship between dispersion and accuracy, and serves as motivation for more rigorous statistical analysis to determine a relationship. We create OLS regressions to determine whether the visual trends observed in the data hold under higher levels of confidence. The aforementioned measure of dispersion, intra-month standard deviation (stdev in regressions), is used as an explanatory variable on the average monthly AFE and FE.

Fig. 15a: OLS Regression using Monthly Standard Deviation as Explanatory Variable on AFE, Estimze

<i>Regression Statistics</i>								
Multiple R	0.22							
R Square	0.05							
Adjusted R Square	0.02							
Standard Error	32.46							
Observations	35							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	1,767	1,767	1.68	0.20			
Residual	33	34,779	1,054					
Total	34	36,545						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	17.90	27.42	0.65	0.52	-37.87	73.68	-37.87	73.68
stdev	1.22	0.94	1.29	0.20	-0.70	3.14	-0.70	3.14

Fig. 15b: OLS Regression using Monthly Standard Deviation as Explanatory Variable on AFE, Consensus

<i>Regression Statistics</i>								
Multiple R	0.02							
R Square	0.00							
Adjusted R Square	-0.03							
Standard Error	30.15							
Observations	35							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	9	9	0.01	0.92			
Residual	33	30,000	909					
Total	34	30,008						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	51.60	37.09	1.39	0.17	-23.86	127.07	-23.86	127.07
stdev	-0.16	1.61	0.10	0.92	-3.44	3.12	-3.44	3.12

Fig. 15c: OLS Regression using Monthly Standard Deviation as Explanatory Variable on FE, Estimze

<i>Regression Statistics</i>								
Multiple R	0.08							
R Square	0.01							
Adjusted R Square	-0.02							
Standard Error	61.30							
Observations	35							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	723	722	0.19	0.66			
Residual	33	124,021	3,758					
Total	34	124,744						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	23.15	51.77	0.45	0.66	-82.18	128.48	-82.18	128.48
Std. Dev	-0.78	1.78	-0.44	0.66	-4.40	2.84	-4.40	2.84

Fig. 15d: OLS Regression using Monthly Standard Deviation as Explanatory Variable on FE, Consensus

<i>Regression Statistics</i>								
Multiple R	0.09							
R Square	0.01							
Adjusted R Square	-0.02							
Standard Error	54.67							
Observations	35							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	799	799	0.27	0.61			
Residual	33	98,644	2,989					
Total	34	99,442						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	20.81	67.26	0.31	0.76	116.03	157.65	-116.03	157.65
Std. Dev	-1.51	2.92	-0.52	0.61	-7.45	4.43	-7.45	4.43

Examining the relationship between forecast errors and dispersion, these four regressions show that the relationships observed in the scatterplots are not significantly relationship. Because this sample is only 35 months, more data could reveal a more significant relationship.

Conclusion and Takeaways

Our research revealed several key takeaways about the differences in NFP forecasting between Estimize forecasts and more traditional Consensus forecasts created by professional economists. Estimize forecasts tend to be more accurate and less biased, yet exhibit larger measures of dispersion, kurtosis, and boldness than forecasts by professionals.

Examining the dispersion of forecasts, Consensus forecasters tended to group more closely to each other, with the 5th - 95th percentiles remaining very close. This herding behavior may be explained by the fact that these professional forecasters' estimates are public and are tracked by their peers, so they fear being wildly inaccurate; it could also be because they are basing their forecasts off the same information.

We developed one theory for why Estimize forecasts may be more accurate but also more dispersed. Professional economists base their forecasts off other economic data which has been recently released. This data is inherently lagging, with releases coming months or even quarters after the period on which it reports has passed. On the other hand, Estimize forecasters may develop their estimates from their personal experiences: for example, if the forecaster or their close relatives got a job or a raise, they may incorporate this information into their forecasts more so than economists do. This makes their forecasts more reflective of real-time information that is not as lagged as official economic data.

However, the sample size of a forecaster's personal network is inherently small. With each forecaster using their own smaller "sample" of personal experiences when forecasting, these

smaller samples are going to be more dispersed than a large national economic survey. This may explain the larger amounts of dispersion among the Estimote community.

Looking at the term structure of forecasts reveals a lot about the how the data is collected by each aggregation platform. Estimote is available to its users all the time, so forecasts can be made as far in advance as the forecaster wishes. This explains the more gradual decrease in number of forecasts made as the time horizon increases, and the larger number of forecasts made far in advance (>15 days). We suspect⁵ that the Consensus forecast aggregator, Bloomberg, contacts economists at a regular monthly time, about a week before the indicator is released. Because Bloomberg does not allow its forecasters to submit forecasts before they call, there are far fewer forecasts further than 9 days out from the release of the indicator.

On average, forecasts in winter months tend to be more accurate, while forecasts in summer months tend to vary more widely. However, we did not prove this statistically but merely analyzed the data qualitatively. The number of forecasts per month tends to increase in the summer for Consensus, and decrease for Estimote. This could be because during the summer months of “sell in May and go away” on Wall Street and summer breaks at colleges, economists employed by financial institutions and universities have more time to forecast economic indicators because they are less focused on other professional responsibilities.

Months in which forecasts were more dispersed tended to exhibit greater errors, both absolutely and directionally. Directionally, forecasts made in months with more dispersion underestimated the number of jobs. This implies that in general, forecasters estimate too low when there is more uncertainty and dispersion. Forecasters seem to be “playing it safe” when they are

⁵ We were unable to find out more details into the methods by which Bloomberg aggregates its forecasts.

unsure of the true NFP value, a logical conclusion. However, this relationship was not found to be statistically significant in any meaningful way.

These findings have critical implications for financial markets. As proven in previous literature, NFP is a major factor on US equity and bond markets, and the ability to develop a more accurate forecast can give market participants a huge information asymmetry. Because consensus are developed days or weeks in advance of the release of monthly NFP, markets tend to price in this consensus days or weeks ahead of release. When the actual report is released, markets quickly move to adjust to the actual number of nonfarm payrolls added that month. This is where the biggest market movements occur.

To predict the most accurate values of NFP, market participants should look to Estimize users who forecast very shortly before the release date. If these forecasts are higher than the overall Consensus forecast, it implies a long position in US stocks and a short position in bonds (as interest rates tend to positively correlate with economic activity). Conversely, if this subset of forecasters is predicting lower values than the average Consensus forecast, it suggests a short position in equities and a long position in bonds, *ceteris paribus*.

Our research on dispersion also implies the possibility of a strategy to take advantage of volatility. If dispersion is high, the absolute forecast error will be high on average, and markets will react to this surprise with higher levels of volatility. Using a long straddle options strategy to take advantage of volatility could be another way to capitalize on the additional information that Estimize provides to markets. However, because this strategy was not proven with the same degree of statistical confidence as the long/short strategy, market participants may be cautious about using this strategy until a longer time horizon is available to prove this relationship with more confidence.

With a dataset that, to the best of our knowledge, has never before been analyzed, this research represents a first step upon which others can build additional research. Several ideas interest us as topics for future investigation:

Most importantly, Estimize tracks over 50 indicators, of which 5-10 receive substantial forecasts each month (though none as many as NFP). Performing the analysis in this paper on other indicators would be useful to examine whether Estimize consistently outperforms Consensus, or whether it varies by indicator.

Estimize tracks the occupation and industry of its forecasters, and Bloomberg assigns names and employers to each of its survey participants, which makes it easy to segment the data into several groupings of academia, students, investment professionals, etc. It would be interesting to see whether one status has better forecasts, either across both platforms or within Estimize or consensus.

Even more detailed than forecaster groupings, both sources track individual forecasters using names and ID numbers for Consensus and Estimize, respectively. Tracking individuals based on their number of forecasts would be a fascinating study on whether forecasting ability can be learned over time, and at what degree of experience do forecasters begin to improve.

Estimize gives forecasters the opportunity to revise their forecasts, theoretically an unlimited number of times, before the indicator is released. One would assume that a forecaster would only revise if significant new information became available, but it would be interesting to prove this hypothesis and discover whether new revisions are more accurate than the original forecasts. The interaction between user identity and revisions—whether academics revise more than professionals or students, for example—would be another interesting topic for further research.

Seasonalized data was one area where we lacked the technical knowledge to perform the times series analyses required to determine whether the seasonality was statistically significant. After we complete more courses on time series analysis, we hope to re-examine the seasonality and provide a rigorous model to test the significance of the seasonal variation in forecasting.

As Aristotle remarked, crowds often exhibit a collective knowledge, even when individuals lack the expertise to compete with a small number of technocrats. Financial markets are one example of this phenomenon, where asset valuations are the result of a very large, heterogeneous crowd. In a field like economic forecasting, which is traditionally dominated by the expertise of elites, our research demonstrates that crowds, using the proxy of Estimote, indeed add value to the field. Estimote's platform is an exciting step towards greater democratization and crowd participation in economic forecasting.

Bibliography

Adebambo, B., & Bliss, B. (2015). The Value of Crowdsourcing: Evidence from Earnings

Forecasts. Unpublished. Retrieved from

[http://com.estimize.public.s3.amazonaws.com/papers/Barbara%20Bliss%20-](http://com.estimize.public.s3.amazonaws.com/papers/Barbara%20Bliss%20-%20University%20of%20San%20Diego%20-%20The%20Value%20of%20Crowdsourcing%20-%20Estimize%20-%20July.pdf)

[%20University%20of%20San%20Diego%20-](http://com.estimize.public.s3.amazonaws.com/papers/Barbara%20Bliss%20-%20University%20of%20San%20Diego%20-%20The%20Value%20of%20Crowdsourcing%20-%20Estimize%20-%20July.pdf)

[%20The%20Value%20of%20Crowdsourcing%20-%20Estimize%20-%20July.pdf](http://com.estimize.public.s3.amazonaws.com/papers/Barbara%20Bliss%20-%20University%20of%20San%20Diego%20-%20The%20Value%20of%20Crowdsourcing%20-%20Estimize%20-%20July.pdf)

Da, Z., & Huang, X. (2015). Harnessing the Wisdom of Crowds. Unpublished. Retrieved from

<https://ssrn.com/abstract=2731884>

Drogen, L., & Jha, V. (2013). Generating Abnormal Returns Using Crowdsourced Earnings

Forecasts from Estimize. Unpublished. Retrieved from <https://ssrn.com/abstract=2337709>

Estimize, Inc. (2017). About Estimize. Retrieved September 17, 2017, from

<https://www.estimize.com/about>

Jame, R., Markov, S., and Wolfe, M. (2017). Does Crowdsourced Research Discipline Sell-Side

Analysts? Unpublished. Retrieved from <https://ssrn.com/abstract=2915817>

Jame, R., Johnston, R., Markov, S., & Wolfe, M. (2016). The Value of Crowdsourced Earnings

Forecasts. *Journal of Accounting Research*, 54(4). <https://doi.org/10.1111/1475-679X.12121>

Lunsford, K. (2017). Lingering Residual Seasonality in GDP Growth. *Economic Commentary*

2017-06. Retrieved from [https://www.clevelandfed.org/newsroom-and-](https://www.clevelandfed.org/newsroom-and-events/publications/economic-commentary/2017-economic-commentaries/ec-201706-lingering-residual-seasonality-in-gdp-growth.aspx)

[events/publications/economic-commentary/2017-economic-commentaries/ec-201706-lingering-residual-seasonality-in-gdp-growth.aspx](https://www.clevelandfed.org/newsroom-and-events/publications/economic-commentary/2017-economic-commentaries/ec-201706-lingering-residual-seasonality-in-gdp-growth.aspx)

Miao, H., Ramchander, S., & Zumwalt, J.K. (2014). S&P 500 Index-Futures Price Jumps and

Macroeconomic News. *Journal of Futures Markets*, 34(10). <https://doi.org/10.1002/fut.21627>

Rumsfeld, D. (2002, February 12). DoD News Briefing - Secretary Rumsfeld and Gen. Myers.

US Department of Defense. Retrieved from

<http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636>

Taylor, N. (2010). The Determinants of Future U.S. Monetary Policy: High-Frequency Evidence.

Journal of Money, Credit and Banking, 42(2/3). Retrieved from

<http://www.jstor.org/stable/20685105>