Using Polls to Forecast Popular Vote Share for US Presidential Elections 2016 and 2020: An Optimal Forecast Combination Based on Ensemble Empirical Model

Joshy Easaw[*a], Yongmei Fang[b] Saeed Heravi [a]
[a]*Cardiff University Business School,UK*
[b]*College of Mathematics and Informatics, South China Agricultural University, China*

July 2022

Contact*:* Joshy Easaw *EasawJ1@cardiff.ac.uk*

Using Polls to Forecast Popular Vote Share for US Presidential Elections 2016 and 2020: An Optimal Forecast Combination Based on Ensemble Empirical Model[+]

Abstract:

This study introduces the Ensemble Empirical Mode Decomposition (EEMD) technique to forecasting popular vote shares in general elections. The technique is useful when using polling data. Our main interest in this study is shorter- and longer-term forecasting and, thus, we consider from the shortest forecast horizon of 1-day to three months ahead. The EEMD technique is used to decompose the election data for the two most recent US presidential elections; 2016 and 2020. Three models, Support Vector Machine (SVM), Neural Network (NN) and ARIMA models are then used to predict the decomposition components. Subsequently, the final hybrid model is constructed by comparing the prediction performance of the decomposition components. The predicting performance of the combination model is compared with the benchmark individual models: SVM, NN, and ARIMA. Finally, this is also compared to the single prediction market *IOWA Electronic Markets*. The results indicate that the prediction performance of combined EEMD model is better than that of the individual models.

Keywords: Forecasting Popular Votes Shares; Electoral Poll; Forecast combination, Hybrid model; Support Vector Machine

[+] We gratefully acknowledge the comments and suggestions of the anonymous referee and Associate Editor. Nevertheless, any omissions and errors are entirely ours.

1: Introduction:

The purpose of this paper is to consider how well polling data can forecast the popular vote shares in general elections using the Ensemble Empirical Mode Decomposition (EEMD) approach. The application we consider are the two most recent US Presidential elections; 2016 and 2020. We are able compare the 2016 US presidential elections where neither candidate was the incumbent president and the 2020 elections where the incumbent was seeking a second term. In 2016, voters had to assess the candidates' election platform and campaign rather than their actual performance in office. While in 2020, the voter could also take into consideration the incumbent's competence in office. Importantly, too, we are not predicting the winner of the elections. This dependents on other outcomes, for example, the electoral college for the US and number of parliamentary seats won for the UK. Nevertheless, predicting candidates share of the votes is a leading indicator of the ultimate winner.

It is important to highlight the two distinct features of the last two presidential elections. Firstly, it included an unconventional candidate and president, Trump (see Panagopoulos et al (2018) and Panagopoulos (2020)). Secondly, the 2020 elections took place during a global pandemic and one that directly affected the US. This enables us to compare the performance of a single candidate (notable Trump) both as a non-incumbent and, subsequently, as an incumbent. Also, we can compare the relative ability to forecast US Presidential election outcomes during normal times (2016) and during a global pandemic (2020).

There are three dominant approaches to modelling and forecasting election

outcomes. Firstly, there is the Structuralists approach, which is firmly grounded in theoretical explanations of election outcomes (for example, Abramowitz, 2012, Holbrook, 2012 and Lewis-Beck and Tien, 2012). This approach uses a core political economy explanation. This includes economic variables such as economic growth, unemployment and inflation rate. It also accounts for political ones such as presidential popularity. Recent assessments of the Structuralists approach are found in Graefe et al (2015) and Lauderdale and Linzer (2015). In a recent study Nadeau et al (2020), applying a structural forecasting model, found that information on long-term factors is still valuable when making accurate predictions of electoral outcomes. Hence, they question the assumption that campaigns matter more now than they did in the past.

Another common approach in the literature is the Aggregators. This approach uses multiple public opinion polls to gauge voters' preferences (see, for example, Blumenthal, 2014 and Jackman, 2014, Reade and Vaughan Williams, 2019). Reade and Vaughan Williams (2019) also extend the existing literature by comparing opinion polls to prediction markets, which is another source of election forecasting. The Aggregators take an *atheoretical* approach relying on multiple polls using dynamic and repeated estimates throughout the campaign. Finally, the third approach is a synthesis of the preceding two approaches and appropriately termed the Synthesizers. (recent examples, see Erikson and Wlezien, 2014 and Linzer, 2013). This approach employs a political economy theory of voting, while using a number aggregated and contemporary opinion polls. The polled data used are either at a state or national level.

The approach of the present study falls firmly in the second category. We focus

on average opinion polls comparing the 2016 and 2020 US presidential elections outcomes. In addition, this compared to the single prediction market *IOWA Electronic Markets*. In 2016, voters can only judge candidates based on their platforms rather than achievements in office. But in the 2020 the incumbent candidate's competence and electoral platform can be compared to the challenger's electoral platform and campaign. Both elections are noteworthy for their respective novelty. As mentioned, in 2016 Trump was considered an unconventional or maverick candidate and 2020 he was judged as an unconventional president during a global pandemic.

Recent studies using EEMD have concluded that the method outperforms other statistical methods for forecasting high-frequency volatile components and shorter-term forecasts (see Fang et al, 2020 and references therein). Indeed, electoral polls tend to be high-frequency and volatile with, at least, daily polling. Furthermore, when forecasting using electoral polls, more accurate shorter-term forecast is enormously beneficial.

This paper makes a novel contribution to the existing literature as we use the EEMD techniques to forecast the share of votes for presidential elections using daily average polls compiled by *RealClearPolitics*. The EEMD technique is used to decompose the data into linear and non-linear characteristics. Subsequently, prediction models are applied on the decomposed components. The results of the combination model are then compared with the individual models of Support Vector Machine (SVM), Neural Network (NN) and ARIMA as the benchmark models. Our main interest in this study is on shorter- and longer-term forecasting, and thus we consider from the shortest forecast horizon of 1-

day to three months. The results indicated that the prediction performance of EEMD combined model is better than that of the individual models, especially for all forecasting horizons.

2. Methodology: Ensemble Empirical Mode Decomposition

In this section, we briefly describe the EEMD approach (a detailed outline can be found in Fang et al (2020)). EEMD is an adaptive method suitable for effectively capturing non-stationary and non-linear behavior in time series data. EEMD decomposes the time series into *n* Intrinsic Mode Functions (IMF) with different frequency and amplitude as follows:

1. Determine the maximum (minimum) values of the original time series.
2. Apply a cubic interpolation and connect all the maximum (minimum) to generate the upper(lower) envelope.
3. Obtain the local mean values of the two envelopes

$$m_1(t) = \left[x_{\max}(t) + x_{\min}(t) \right] / 2 \tag{1}$$

4. Subtract the means obtained in (1) from the original time series data

$$h_1(t) = x(t) - m_1(t) \tag{2}$$

5. If $h_1(t)$ satisfies the IMF conditions, then repeat step 1 to step 4 until the remainder becomes a monotonic function and no more IMF can be extracted, in which the series is decomposed into *n* independent IMFs and a remainder,

$$x(t) = \sum_{i=1}^{n} h_i(t) + r(t)$$

We use an improved EEMD that avoids misidentification, resulting from the empirical mode decomposition, by adding noise to the data set. The process of EEMD

is as following.

1. A white noise series confirming to normal distribution $\varepsilon_n(t)$ is added to the original time series, which generates a new time sequence as:

$$x_n(t) = x(t) + \varepsilon_n(t) \tag{3}$$

2. Decompose the time series data obtained in (3) into IMFs.

3. Repeat step 1 and step 2 *m*-times, with adding different white noise series.

4. As the result, compute the averages of the corresponding IMFs obtain in the decomposition, step 2.

$$h_n(t) = \frac{1}{m}\sum_{i=1}^{m} h_{in}(t) \tag{4}$$

The advantage of EEMD is that the added noise cancels each other in the end results and the chance of mode mixing is significantly reduced. The final decomposition result is given as:

$$x(t) = \sum_{i=1}^{n} h_i(t) + r(t) \tag{5}$$

where $h_i, i = 1,2,....n$ are the final IMFs, and r is the remainder. The intrinsic model functions and the remainder obtained by EEMD preserve the non-stationary and non-linear features of the original time series data while avoid the modal aliasing.

3. The Data and EEMD Decomposition

The data used in this study is taken from *Real Clear Politics*[1] which compiles the daily

[1] Please note that the 2020 presidential polling data is only available from the 1st of September 2019. Further details and breakdown of the data is found in: https://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton-5491.html and https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html.

average of all the major pollsters, and is outlined in Figure 1a and 1b below:

Figure 1a and 1b [about here]

We use polling data starting from the 1st of July 2015 and ending on polling day, the 8th of November 2016. For the 2020 US Presidential elections we employ the data from 1st September 2019 to 3rd November 2020. The summary statistics for the polling data relating for the 2016 and 2020 US presidential elections are outlined in Tables 1a and b respectively:

Table 1a and 1b [about here]

As can be seen from Table 1a, the standard deviation and range is slightly lower for Hilary Clinton. Trump's coefficient of variation is 6.2%, which indicates a slightly higher volatility and fluctuation. The average poll is 6% higher for Clinton in this period. Table 1b presents the summary statistics for 2020 election. The data leading up to the 2020 presidential elections are much less volatile for both Trump and Biden. The average poll is 7.3% higher for Biden in this period.

The EEMD method is used to decompose the transformed data. Figures 2a, 2b, 2c and 2d show the decompositions for the respective presidential candidates in both 2016 and 2020 (from high to low frequency) and a remainder.

Figures 2a, 2b,2c, 2d. [about here]

The fluctuation period reflects the time length, and the amplitude highlights the magnitude of the shock on the polling data. The remainder displaying a monotonous increasing trend and determines the long-term trend of polling data. This is consistent with the termination conditions of EEMD. In this paper, we combine the decomposed

components into one high frequency component and one low frequency component.

4. Forecast Evaluation

We now turn to the main issue of the present study, which is to evaluate the optimal combination forecasting performance using the EEMD approach. Our interest is in both shorter- and longer-term forecasting. Hence, we consider the 1-day, 7-days, 14-days, 30-days and 90- days ahead forecast horizons. To predict the k-days ahead forecasts, Neural Network, Support Vector Machine and ARIMA models were estimated using the first (498-k) observations. Additionally, this is compared to the single prediction market *IOWA Electronic Markets*. The daily market predictions for each candidate, starting from September 2015 and 2019 for the respective presidential elections, is outlined in Figures 3a and b respectively[2]:

Figures 3a and b [about here]

Post sample k-days ahead forecasts for these models and the optimal combination are then computed and the absolute errors (AE) given below is used to measure their respective performance:

$$\text{AE} = ABS(\hat{x}_{t+k} - x_{t+k}) \tag{6}$$

where $\hat{x}_{t+k}$ is the k-step ahead forecast computed either by the optimal combination or SVM, NN, ARIMA or IOWA predictions. x_{t+k} is the actual popular vote share for US presidential election in 2016; for Clinton (48.2%) and Trump (46.1%). For the 2020 election the popular votes for Biden and Trump are 51.3% and 46.9% respectively.

[2] Due completeness and availability, we use the *Last Price* daily quote. Further details and breakdown of the data is found in: https://iemweb.biz.uiowa.edu/markets/.

4.1 Forecast Results

We combine the forecast results from high frequency and low frequency components to obtain the final combination prediction results for each candidate. SVM, NN and ARIMA models are chosen as the benchmark for comparison. Table 2a and 2b presents the post-sample Absolute Error (AE) for the 1-day,7-days,14-days,30-days and 90- days ahead forecast horizons for the Clinton and Trump polling data respectively.

We can see from Table 2a and 2b that the prediction errors among the two individual models; SVM and ARIMA, are almost the same. While the prediction errors for the NN model is slightly lower, that is they are more accurate. Nevertheless, overall, the prediction error of the combined model is much smaller than that of SVM, NN and ARIMA models for all the forecasts. It suggests the superiority of the combined model utilizing the EEMD approach. In the case of Trump, the combined EEMD model continues to prevail as it consistently provides the least predication errors. Overall, in the case of Clinton, the combine model prediction errors are twice as accurate as the single models. In the Trump case, the prediction errors of the combined models are approximately a third or less than the single models.

Table 2a and 2b [about here]

Finally, we compare these with the single prediction market *IOWA Electronic Markets* (IOWA). In the case of Clinton, IOWA is more accurate than SVM and ARIMA but not the combined EEMD model for the 1-day ahead forecast. Regardless, considerably higher prediction errors for the longer forecasting periods. Overall, in the case of Trump, IOWA is more accurate than Clinton. Nevertheless, it is still less accurate than the

combined EEMD model. Interestingly, it is now more accurate that the single model predictors (SVM, NN and ARIMA) for the 7-day ahead forecast, compared with the 1-day ahead forecast for Clinton.

There are a couple of noteworthy points to consider. The prediction error for Clinton is lowest for the 90-day ahead forecast. The combined models are particularly accurate when it comes to longer range forecast of vote shares. The 1-day ahead prediction errors for the Trump popular vote share is about 16 and 10 times larger than the 7 and 90-day ahead forecasts respectively. It should also be noted that a similar outcome can be found for Clinton's 90-day ahead forecast. Nevertheless, the relatively larger prediction error for 1-day ahead forecast for Trump suggests the polls did not reflect Trump's actual support and potential votes just prior to polling day. This could be due to many Trump voters dissembling when responding to pollsters. Also, as a recent study Panagopoulos et al (2018) highlights there was a substantial number of late deciding voters and most of them voted for Trump.

Turning to the 2020 US Presidential elections, the corresponding results are outlined in Tables 2c and 2d for Biden and the now incumbent President Trump.

Table 2c and 2d [about here]

The challenger Joe Biden and eventually victor, as we saw from Figure 1b, consistently polled higher than the incumbent and with a comfortable gap of over 7% points. The combined EEMD predictor has lower prediction errors. Except for the 30-day ahead forecasts, the EEMD approach was considerably more accurate than its closest rival the SVM predictor. In the case of the 7-day ahead forecast its prediction error is a tenth that

of the SVM and very marginally less accurate for the 30-day ahead forecast.

The prediction errors for the incumbent President Trump appears to be higher than the 2016 ones. For instances, the combined EEMD model prediction errors are approximately four and fifteen times higher than its 2016 counterpart's 1- and 7-days ahead forecasts respectively. The EEMD model prediction errors are the lowest only for the 14-day ahead forecasts. In contrast to 2016 and the Biden 2020 predictions, IOWA has the least prediction errors for three of the five Trump forecasts. Especially, in the last week and month leading to election day.

An important overall point to make is that the prediction errors are considerably lower and, hence, the forecast more accurate for the Democratic candidate in 2020 than the 2016 elections. Notably, neither were the incumbent during these elections. In the case of Trump, the prediction errors of the opinion polls were considerably larger for the 2020 elections too. The 2020 prediction errors were at least three to four times larger. This contrasts with the descriptive statistics outlined in the preceding Tables 1a and b, where the prediction errors for Trump 2020 is considerably larger than 2016 even though the 2020 opinion poll is considerably less volatile than the 2016 poll. Conversely, IOWA has the least prediction errors when predicting Trump 2020.

The explanations for the lack of accuracy in predicting the vote shares for Trump in both 2016 and 2020 are varied. The inability to account for actual Trump votes in the polls could be due to, for instances, voters dissembling when responding to pollsters and pollster bias when sampling. Claassen and Ryan (2020) and Panagopoulos

(2021) both consider the notion that a large portion of 'shy Trump' voters did not respond honestly to pollsters. Others (for example, Mellman, 2020) argue that pollsters may not have accounted adequately for late deciders. As 2016, these late deciders voted disproportionately for Trump. Voters who decided in the final week of the 2020 election also favored Trump over Biden by a 54- 42 margin, according to the national exit poll.

Such biasness may be absent for IOWA, a single prediction market. Clearly, both the polls and IOWA were largely inaccurate for the 2016 elections, but IOWA was considerably more accurate in 2020 especially with respect to Trump. Nevertheless, by and large, the opinion polls prevailed.

5. Conclusions

The purpose of this note is to introduce the combined EEMD technique to forecasting popular vote shares in general elections. The combined technique is highly accurate for both shorter and longer horizon forecasting. Polling data is still an important indicator of election outcomes. Nevertheless, polling data had consistently large prediction errors for Trump in both 2016 and 2020. Indeed, in the 2020 elections the single prediction market *IOWA Electronic Markets* (IOWA) fared better. The polling data is probably better at predicting election outcomes for more conventional presidential candidates. While due to the various biasness in the surveyed data, it may be less reliable for unconventional candidates such as Donald Trump.

References:

Abramowitz, A. (2012), Forecasting in a polarized era: The time for change model and the 2012 presidential election. *PS: Political Science & Politics* 45(4), 618–619.

Blumenthal, M. (2014), Polls, forecasts, and aggregators. *PS: Political Science & Politics* 47(2), 297–300.

Claassen, R., & Ryan, J. B. (2020). Why Did the Polls Undercount Trump Voters? *Washington Post*, November 13.

Erikson, R.S., & Wlezien, C. (2014), Forecasting US presidential elections using economic and noneconomic fundamentals, *PS: Political Science & Politics* 47(2), 313–316.

Fang, Y., Guan. B., Wu, & Heravi, S. (2020), Optimal Forecast Combination Based on Ensemble Empirical Mode Decomposition for Agricultural Commodity Futures Prices, *Journal of Forecasting*, 39(6), 877-886.

Graefe, A., Küchenhoff, H., Stierle, V. & Riedl, B. (2015). Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems. *International Journal of Forecasting,* 31(3), 943-951.

Holbrook, T.M. (2012) Incumbency, national conditions, and the 2012 presidential election. *PS: Political Science & Politics* 45(4), 640–643.

Jackman, S. (2014) The predictive power of uniform swing. *PS: Political Science & Politics* 47(2), 317–321.

Lauderdale, B & Linzer, D. (2015) Under-performing, over-performing, or just performing? The limitations of fundamentals-based presidential election forecasting. *International Journal of Forecasting,* 31 (3), 965–979.

Lewis-Beck, M.S. & Tien, C. (2012). Election forecasting for turbulent times. *PS: Political Science & Politics* 45(4), 625–629.

Linzer, D.A. (2013). Dynamic Bayesian forecasting of presidential elections in the States. *Journal of the American Statistical Association*. 108 (201), 124-134.

Mellman, M. (2020). Polling Isn't Broken. But We Too Often Miss its Hidden Signals. *Washington Post*, November 16.

Nadeau, R., Dassonneville, R., Lewis-Beck, M., & Mongrain, R. (2020), Are election results more unpredictable? A forecasting test *Political Science Research and*

Methods, 8 (4), 764-771.

Panagopoulos, C., (2021), Accuracy and Bias in the 2020 U.S. General Election Polls. *Presidential Studies Quarterly*, 51 (1) 214–227

Panagopoulos, C., Endres, K., & Weinschenk, A.C. (2018), Pre-election poll accuracy and bias in the 2016 U.S. general elections. *Journal of Elections, Public Opinion and Parties*, 28 (2) 15-172.

Reade, J & Vaughan Williams, L., (2019), Polls to probabilities: Comparing prediction markets and opinion polls. *International Journal of Forecasting,* 35 (1) 336-350

Figure 1a: Average Daily Polling for Clinton and Trump (1st July 2015 to 8th Novemember 2016)

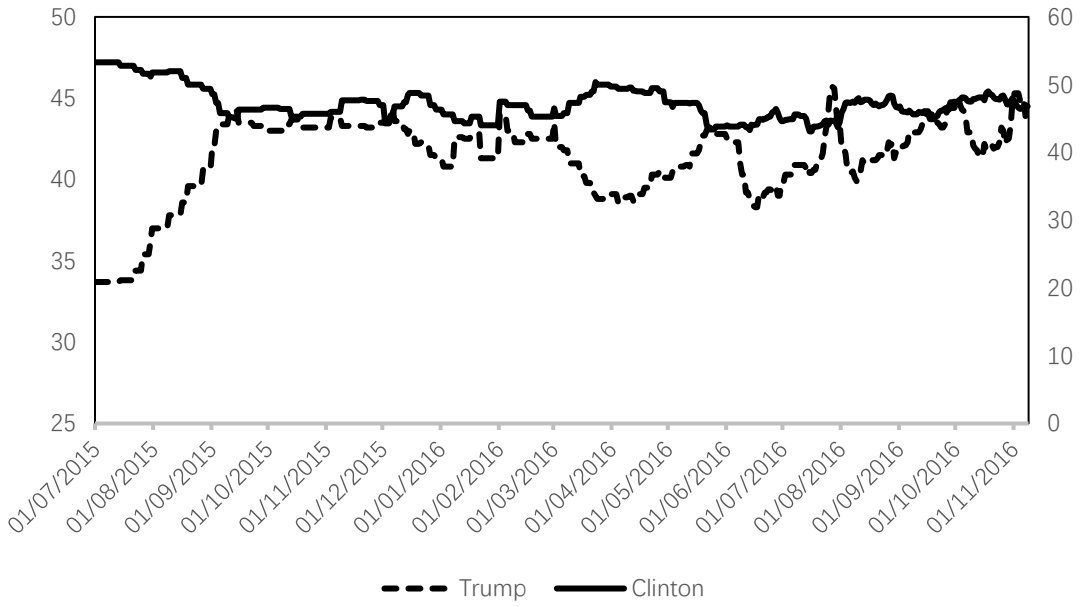

Figure 1b: Average Daily Polling for Biden and Trump (1st September 2019 to 3rd Novemember 2020

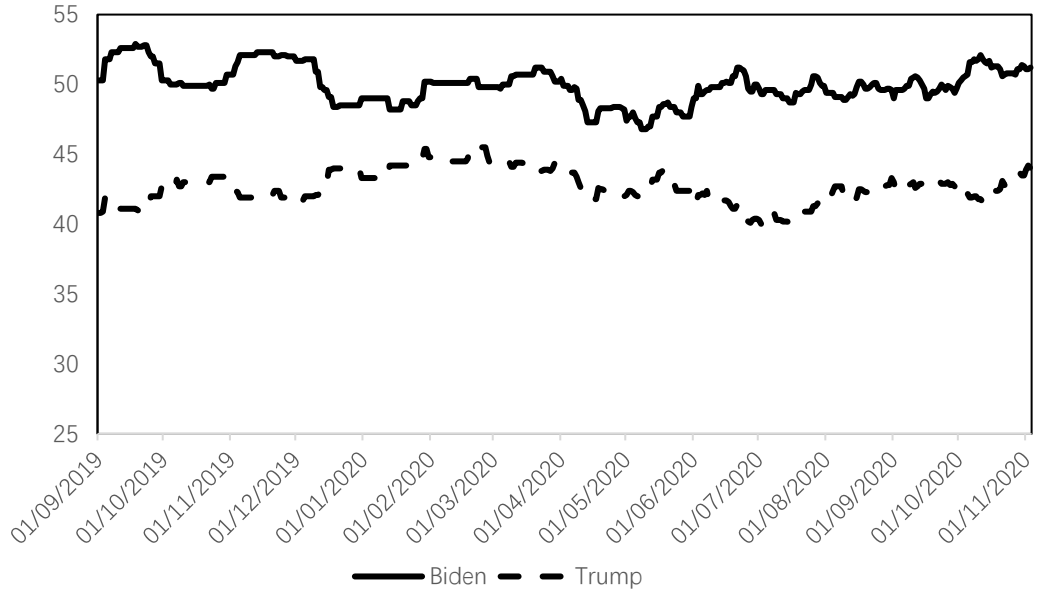

Table 1a: Descriptive Statistics for Daily Polling Data
Presidential Election 2016

Category	Mean	Standard deviation	Minimum	Maximum	Range
Clinton	47.2%	2.4%	43.1%	53.3%	10.2%
Trump	41.3%	2.6%	33.7%	46.1	12.4%

Table 1b: Descriptive Statistics for Daily Polling Data
Presidential Election 2020

Category	Mean	Standard deviation	Minimum	Maximum	Range
Biden	50.0%	1.3%	46.8%	52.9%	6.1%
Trump	42.7%	1.2%	40.0%	46.9%	6.9%

Figure 2a: Intrinsic Model Functions: Decomposed Polling Data 2016: Clinton

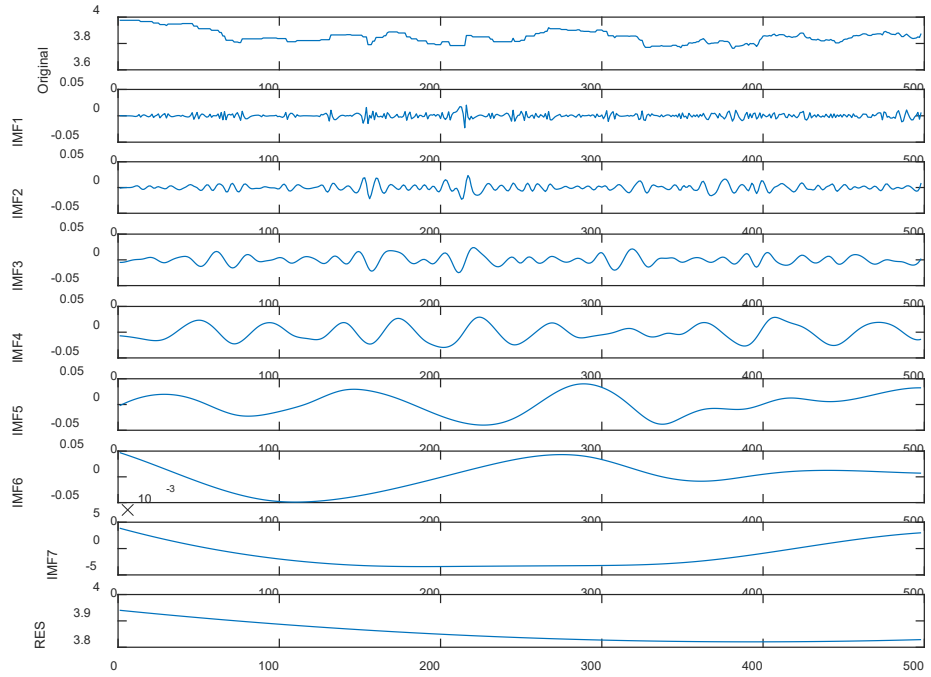

Figure 2b: Intrinsic Model Functions: Decomposed Polling Data 2016: Trump

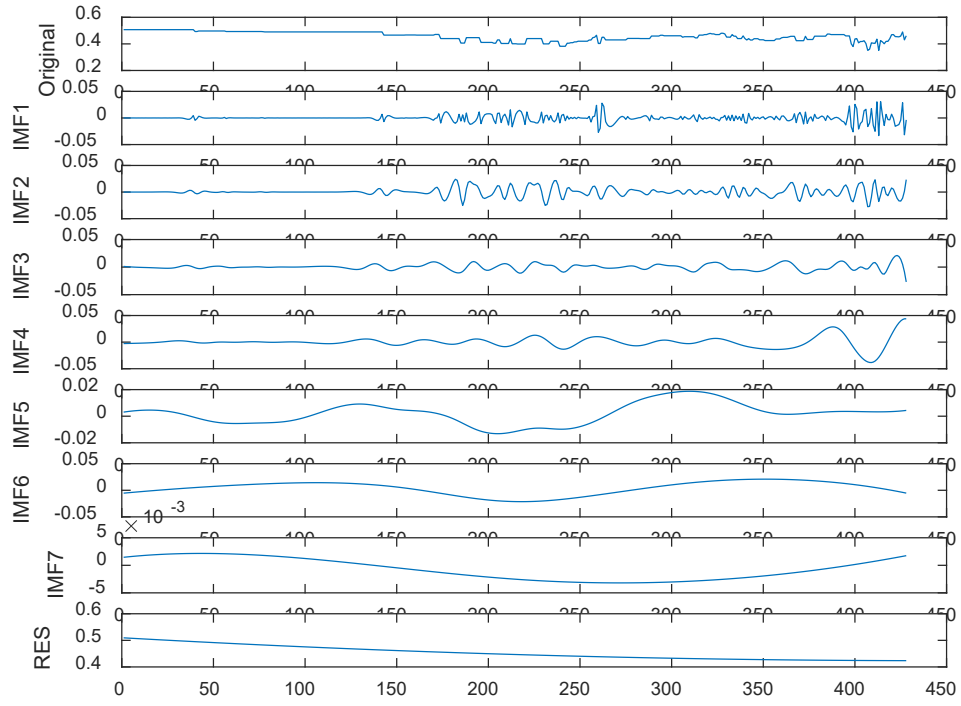

Figure 2c: Intrinsic Model Functions Decomposed Polling Data 2020: Trump

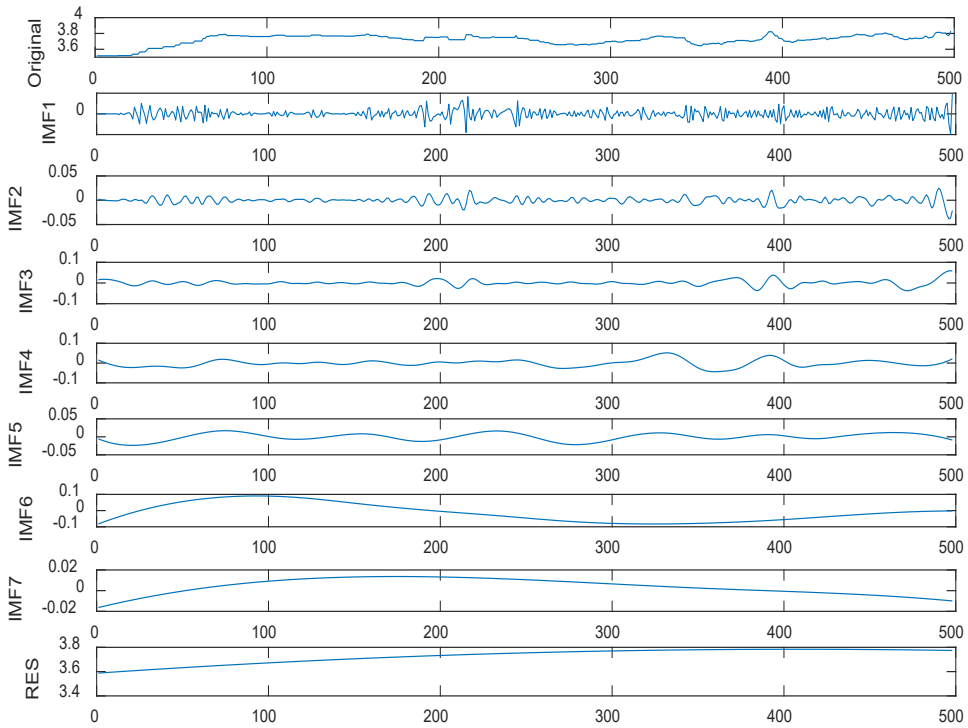

Figure 2d: Intrinsic Model Functions Decomposed Polling Data 2020: Biden

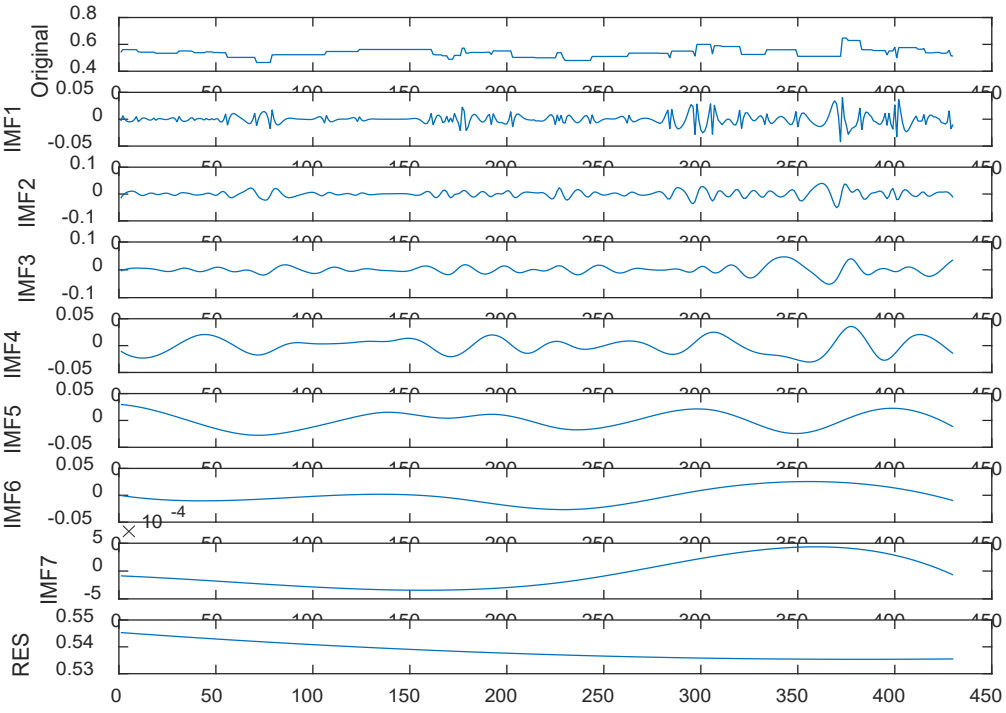

Figure 3a : IOWA Electronic Market Daily Predictions for Clinton and Trump (1st September 2015 to 8th Novemember 2016)

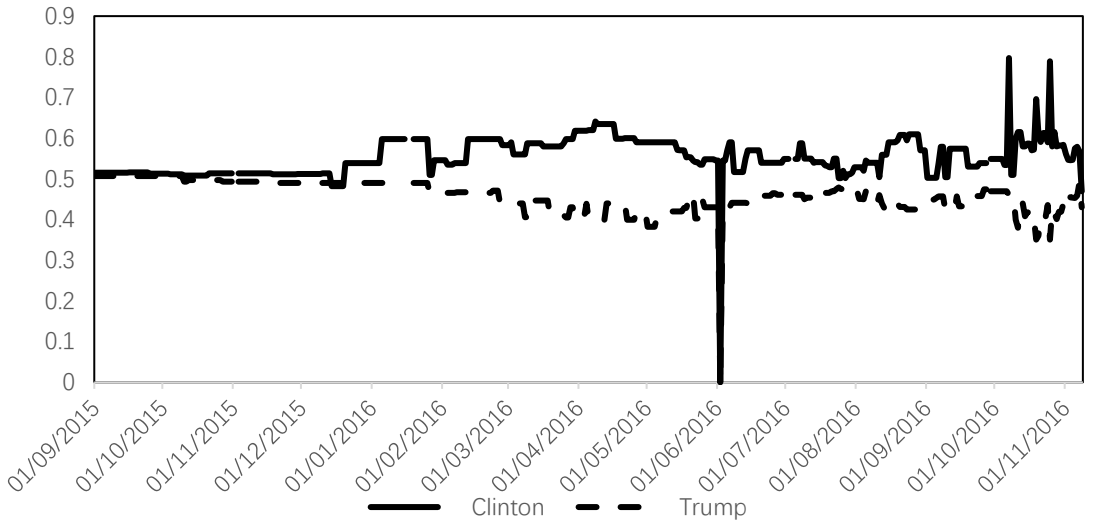

Figure 3b : IOWA Electronic Market Daily Predictions for Biden and Trump (1st September 2019 to 3rd Novemember 2020)

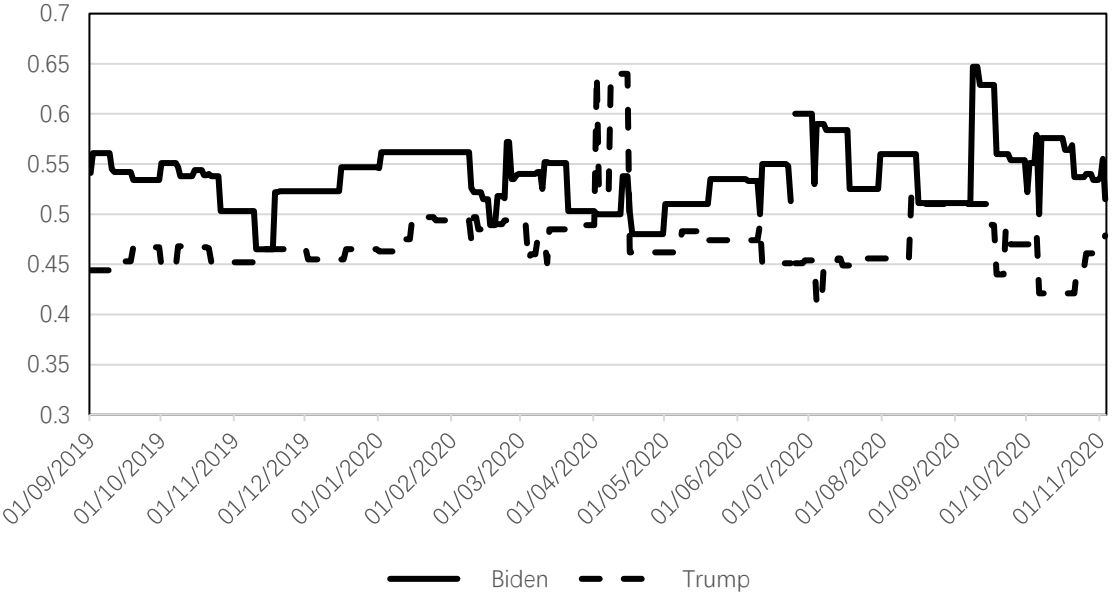

Table 2a: Absolute Error (AE) for Clinton (2016)

	1-day ahead	7-days ahead	14-days ahead	30-days ahead	90-days ahead
EEMD	**0.7410**	**0.7809**	**0.8172**	**0.8886**	**0.0482**
SVM	1.3488	1.4257	1.4433	1.4032	1.3917
NN	1.0676	1.3712	1.0207	1.4106	1.0169
ARIMA	1.4090	1.4631	1.5262	1.6701	2.2056
IOWA	1.2	6.5	9.8	11.7	5.8

Table 2b: Absolute Error (AE) for Trump (2016)

	1-day ahead	7-days ahead	14-days ahead	30-days ahead	90-days ahead
EEMD	**0.6682**	**0.0428**	**0.1075**	**0.0796**	**0.0675**
SVM	2.4691	2.5035	2.5016	2.5413	2.8062
NN	1.6926	2.7993	2.7321	2.7851	2.7705
ARIMA	2.4719	2.3028	2.1046	1.6484	0.1050
IOWA	3.1	0.9	3.6	6.1	2.1

Notes: The figures in bold denotes least prediction error for the respective forecasts

Table 2c: Absolute Error (AE) for Biden (2020)

	1-day ahead	7-days ahead	14-days ahead	30-days ahead	90-days ahead
EEMD	**0.01**	**0.01**	**0.04**	0.09	**0.09**
SVM	0.08	0.10	0.11	**0.08**	0.15
NN	0.19	0.92	0.25	0.26	0.15
ARIMA	0.10	0.08	0.07	0.03	0.11
IOWA	0.20	2.7	2.4	6.6	0.20

Table 2d: Absolute Error (AE) for Trump (2020)

	1-day ahead	7-days ahead	14-days ahead	30-days ahead	90-days ahead
EEMD	2.5	2.59	**2.16**	2.80	2.83
SVM	2.90	2.91	2.29	2.91	2.88
NN	2.45	3.42	2.89	2.89	2.91
ARIMA	2.89	2.80	2.70	2.48	**1.64**
IOWA	**1.0**	**0.80**	4.8	**1.30**	4.1

Notes: The figures in bold denotes least prediction error for the respective forecasts