Graph-based Model Election Forecasting

Matthew Goodman
University of Pennsylvania
goodmanm@seas.upenn.edu

Abstract

We developed a model that uses historical election results and other datasets to forecast future election results. This model builds an election network with counties represented as nodes on fully-connected graph. We used sampling theory specific to graph signal processing (GSP) to determine which counties to poll in order to best reconstruct the election signal. We tested this model using the 2012 Presidential election and matched the actual result with approximately five percent mean error.

1. Introduction

The fundamental characteristic of Democratic society is the use of voting to elevate citizens to public office. In the United States, elections are held to select representatives for the legislative branch, as well as to select the head of the executive branch - the President of the United States. Elections are cyclical; elections for the presidency are quadrennial, while elections for the legislature occur every two years on a rotating cycle. The importance of the election process naturally invites widespread attention from the general public. A natural extension of this importance is the popularity of predicting election results through polling. Traditional election polling consists of querying a subset of the eligible voting population for their preferred choice. This subset is selected by drawing randomly from predetermined categories divided along socioeconomic, geographic, racial, and age lines [11] with the expectation that the preferences of a few individuals from each group will reflect the behavior of the whole in the election. Election polling is inherently difficult; polling is a resource intensive process because it is both difficult to acquire data and difficult to account for polling bias across categories. The need for bias correction cannot be overlooked because polling data is not verifiable; there is a definite probability that polling responses are inaccurate. Beyond this, it is uncertain who will even show up on election day, a phenomenon compared to reconstructing a signal from the sample of a sample [7]. The desired behavior is to develop a forecasting model that only requires accurate polling from a minimal set of counties that still forecasts the entire election.

While traditional poll forecasting is widely used and effective in some cases [7], results are consistently inaccurate and easily subject to human error. In one example, a simple linear regression based on racial composition and a custom 'geographic factor' was used to predict Hilary Clinton's vote share in recent primary elections [1]. The flaw in this logic is that elections are inherently non-linear phenomena - they play out in a geographic map where every individual county is in some way influenced by the rest of the country. We are better able to model non-linear phenomena using non-linear data. One such useful model for non-linear data is the use of a network, which is mathematically modeled as a graph. By describing data in a network we can capture the relations between all counties providing us more information to build an election model. Furthermore, our non-linear data set is composed of real election returns and census data. That this data is verifiable means that incorrect predictions are purely the result of a flaw in the model and not the result of faulty data.

1.1. Addressing the Problem

The goal of this research is to provide a graph-based model for predicting elections based on limited data. We aim to model the counties as nodes in a graph and to design the edges between these counties based on available election and census data. We set counties as nodes because national elections are physically organized on the county level; each county is responsible for maintaining voting locations and sending votes to counting facilities. First, we consider the relationship between voting patterns from previous Presidential elections and predicted election results. Second, we consider the relationship between voting patterns from previous Senatorial elections and predicted election results. Senate election results are powerful because they are useful to improve our model at the state level. Senate elections are both infrequent and polarizing which results in higher turnout than elections for the House of Representatives. Overall this yields a better model of how counties in the same state influence each other. Counties are

related to one another by historical voting patterns. Two physically adjacent counties will almost certainly have the same historical voting pattern meaning they are strongly correlated; an accurate poll of county X should predict how county Y will vote. Using historical election data we can build a graph shift operator that relates the connections between counties as some factor of 'connectedness'. We can also add other census data such as median income or poverty index to further establish the connections that exist between counties.

By describing the underlying electoral structure as a network we can exploit the emerging theory of Graph Signal Processing (GSP) [12, 13]. We define the election result on each county as a graph signal and we analyze the spectral properties of this signal on the graph. Then, we proceed to exploit the sparse frequency representation of the election signal and use GSP sampling theory [5, 10] to determine which counties are the most representative of the voting signal and therefore where to allocate the bulk of polling resources. In essence, this research will focus on building a graph signal model that predicts election outcomes with similar accuracy to traditional polling techniques [8].

1.2. Validation

Measurements to validate these proposed methods are self-evident because we are attempting to reconstruct real results. Put another way, we can use any of the commonly used metrics from discrete signal analysis. The simplest validation measurement would be to count how many incorrect predictions our model makes and compare to more established models. Another validation would be to determine the mean squared error at the county-by-county level; a partial success would be the model calling the election at the state by state level yet generating a high mean squared error (error-rate) at the county level. At a broader level, this research advances the field if it can achieve similar accuracy to traditional polling using less resources, but this could only be validated using historical results. If successful on historical data, we plan to test the model on the 2016 Presidential election in real-time.

2. Methods

2.1. Equipment and Facilities

The entirety of this research is based in theory and MATLAB software. Therefore, the only equipment needed to duplicate this research is any computer with MATLAB software. The code we have written falls into two groups. The first group is used to clean and access the data needed to build the graph shift operator. The second group of scripts is used to analyze the efficacy of any constructed model with various plots and metrics. All code for this research is written in MATLAB. All datasets were accessed through existing University of Pennsylvania library subscriptions, resulting in negligible cost.

2.2. Setup and Procedures

The data architecture for this project is separated into two partitions. The first data partition consists solely of data used to build the graph shift operator (county model); this consists of the county-level quadrennial Presidential election data and county-level biennial Senatorial election data. The data should be organized in folders according to a shared naming convention so that it is easy to programmatically compile and use. Data accessor functions should be placed in the shared folder for easy data access. Election data is accessible through a variety of sources. For this research we used county-level data from the U.S. Political Stats dataset provided by CQ Press. The second partition should consist of the physical MATLAB files the research team produces to analyze the data.

The goal of this research is to reconstruct election results based on a model of the connection between counties. This model is directly produced as a graph shift operator describing the connections between counties based on historical election and income data. The computational procedures necessary to realize the core functionality of this research consist of those that clean the raw data and the mathematical methods described in the following section.

Raw election data comes as a table containing the results for each county in several metrics. From this, we calculate the percentage of citizens that voted Republican for each county with the whole consisting of Republicans and Democrats. Votes for Independent candidates are ignored as an assumption of this research. This data is filtered to remove newly created counties and the entire state of Alaska due to odd election district structuring, and any other anomalous data. Senatorial election data is cleaned in the same manner with the additional caveat that elections where the winning candidate either ran unopposed or was an Independent candidate (e.g., Bernie Sanders in Vermont) are removed. The raw election data is reduced to a vector of numbers that shows the Republican/Democratic share of the vote for each county. A vector is produced for each election cycle. The concatenated election vectors form a matrix. This matrix of election vectors is hereafter referred to simply as the election data.

2.3. Mathematical Basis

A brief overview of the core theory and requisite terminology used for this research is presented. This is derived mainly from an introduction by Marques, Segarra, et. al.

A graph $\mathcal{G}$ is defined as $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$. The set of nodes $\mathcal{N}$ has size N, the set of edges $\mathcal{E}$ is such that edge $(i, j) \in \mathcal{E}$ if node i is connected to node j, and every edge in the set $\mathcal{E}$ has a corresponding weight in the set $\mathcal{W}$. A

signal $\mathbf{x} = \begin{bmatrix}[x_1 & x_2 & \cdots & x_n\end{bmatrix}^T$ is defined on the graph $\mathcal{G}$ where x_i is the value of the signal corresponding to node i. Formally, $\mathbf{x} \in \mathbb{C}^N$. We have defined the graph structure, as well as an arbitrary signal defined on the graph.

The graph $\mathcal{G}$ has a graph-shift-operator $\mathbf{S}$ defined as an $S \in R^{N \times N}$ matrix satisfying $S_{ij} = 0$ for $i \neq j$ and $(i, j) \notin \mathcal{E}$. We assume that $\mathbf{S}$ is diagonalizable, so there exists an $N \times N$ eigenvector matrix $\mathbf{V}$ and an $N \times N$ eigenvalue matrix $\mathbf{\Lambda}$ that can be used to decompose $\mathbf{S}$ into $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. If $\mathbf{S}$ is a normal matrix $\mathbf{S}\mathbf{S}^H = \mathbf{S}^H\mathbf{S}$. This means that $\mathbf{V}$ is unitary and gives $\mathbf{V}^{-1} = \mathbf{V}^H$. This yields the decomposition of the graph shift operator $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$.

The graph shift operator allows for the representation of the signal in the frequency domain, understanding as such, as the use of a basis that is invariant to linear filtering. The Graph Fourier Transform (GFT) is defined as $\tilde{\mathbf{x}} = \mathbf{V}^{-1}\mathbf{x}$ where $\tilde{\mathbf{x}}$ are the frequency components of the signal. The inverse Graph Fourier Transform (iGFT) is therefore $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}$. We say that the signal $\mathbf{x}$ is K-bandlimited on the graph shift operator $\mathbf{S}$ if its GFT has at most K nonzero components. That is, $\mathbf{V}^{-1}\mathbf{x} = \tilde{\mathbf{x}} = \begin{bmatrix}\tilde{\mathbf{x}}_K^T & \mathbf{0}_{N-K}^T\end{bmatrix}^T$ where $\mathbf{0}_{N-K}$ is a vector of $N - K$ zeros. If this is the case then the original signal can be reconstructed from a sampled version of the complete signal. Define a $K \times N$ selection matrix $\mathbf{C}$ that samples K of N of the graph nodes. Observe that $\mathbf{C}$ is a selection matrix if it is binary and has exactly one nonzero entry per row and at most one nonzero entry per column. This sampled signal is defined as $\bar{\mathbf{x}} = \mathbf{C}\mathbf{x}$. The reconstruction can be carried out by

$$\mathbf{x} = \mathbf{V_k}\tilde{\mathbf{x}}_\mathbf{K} = \mathbf{V_K}(\mathbf{C}\mathbf{V_K})^{-\mathbf{1}}\tilde{\mathbf{x}} \tag{1}$$

whenever a selection matrix $\mathbf{C}$ that satisfies $rank\{\mathbf{C}\mathbf{V}_K\} = K$ is used.

The difficulties specific to this research include generating a selection sampling matrix $\mathbf{C}$ that satisfies the above equation, working with a graph signal that is not cleanly bandlimited, and working with a graph shift operator that does not decompose to a full rank eigenvalue matrix.

2.4. Results

We attempted to forecast the 2012 Presidential election using only historical data. This includes Presidential elections from 1984-2008 and Senatorial election results for cycles between 1984-2008. Several graph shift operators were tested in forecasting the 2012 election. The methods section describes the graph shift operator we use as a simple covariance matrix. There are actually many different methods of relating the county 'connectedness' that are more complex than covariance. Results will be reported in terms of each of these methods.

The table in Fig. 1 gives several metrics to analyze the frequency spectrum of the graph Fourier transform applied to the election signal. As described in the mathematical methods section, each graph shift operator is diagonalizable. We define the Hermitian of the eigenvector matrix as the graph Fourier transform. We then apply this transform to this election signal to get the frequency output.

The number of non-negligible frequency coefficients in Fig. 1 determined the number of counties we sampled to reconstruct the original signal. The table in Fig. 2 shows the reconstruction errors achieved with each of the different graph shift operators.

The results are available to my research mentor as the script that runs the experiment. This research implements a deterministic procedure, therefore anyone who runs the forecasting script with the same inputs we used will obtain the same results.

3. Discussion

Looking at the results in Fig. 2 we see that different variants of the graph shift operator S give different numbers of non-negligible frequency components, but that most give approximately 23. This is the result of having a limited data set of approximately 23 election cycles that we use to produce S. For the moment, the bandlimitedness is loosely contained if at all; the rank of the eigenvalue matrix is approximately 25 giving 25 degrees of freedom in the eigenvector basis and the election signal is bandlimited to approximately the same number of coefficients. This is a critical area that needs to be addressed to ensure the validity of the proposed method. The state-of-the-art concept we employ is to reconstruct the election signal by defining it on a network structure. This whole concept relies on the assumption that the election signal is bandlimited to K frequency coefficients or as having K degrees of freedom. To determine what K is we must first have a graph shift operator that satisfies $rank(S) > K$. If $rank(S) \leq K$ we can't measure whether the signal is actually K bandlimited, or whether we hit the boundary on the degrees of freedom set by $rank(S)$. By incorporating more data we aim to increase the rank of the graph shift operator's eigenvalue decomposition matrix, which will solidify that the bandlimited signal has less degrees of freedom than the rank of the eigenvalue basis (see Section 2.3).

It is important to note that we used a heuristic to determine which frequency components were considered non-negligible. The fluid definition of "non-negligible" was adequate for this early stage research but will not be satisfactory for more robust efforts. We would need to use a principled statistical basis to determine what is negligible.

In Fig. 2 we see a different picture emerge. The 'vanilla' covariance matrix was clearly the best graph shift operator for signal reconstruction. As we incorporate more data to

Graph Shift Operator Method	Non-Negligible Freq. Components	Max Value	Min Value	Range
dy Covariance	K=23	3.122	0.043	71.786
Pearson	K=22	3.102	0.056	55.325
Kendall	K=253	3.017	0.0001	25705.401
Spearman	K=22	3.162	0.028	112.409
Cosine	K=24	2.892	0.028	102.884
Correlation pdist	K=23	2.947	0.003	989.297
Spearman pdist	K=23	3.342	0.020	163.902

Figure 1. Frequency components by variant of graph shift operator

Reconstruction Errors		
Graph Shift Operator Method	Mean Absolute Error	Mean Squared Error
Covariance	0.0499	0.0637
Pearson	0.0618	0.0766
Kendall	0.0834	0.1054
Spearman	0.0812	0.1054
Cosine	0.0592	0.0765
Correlation pdist	0.0793	0.0972
Spearman pdist	0.0838	0.1057

Figure 2. Reconstruction Errors by variant of graph shift operator

improve the rank of the graph shift operator decomposition, it will be interesting to see if the covariance matrix remains the best choice for reconstruction. Using more datasets may require non-linear methods to combine datasets that are not similar (e.g. income data and election data). Non-linear datasets may be more effectively represented in a graph shift operator that is more expressive than the covariance.

Our model actually supplies a list of counties to sample. The minimal set of counties includes: Geneva County (AL), Ashley County (AR), Searcy County (AR), Mono County (CA), Lafayette County (FL), Clayton County (GA), Worth County (GA), Shelby County (IL), Dallas County (IA), Rush County (KS), Jefferson Davis Parish (LA), Barnstable County (MA), Benzie County (MI), Monroe County (MO), Meagher County (MT), Bernalillo County (NM), Burleigh County (ND), Franklin County (OH), Union County (OR), Gregory County (SD), Roane County (TN), Albemarle County (VA), Wood County (WI). It is interesting to observe the geographic diversity of the county set. There are counties representing each major geographic region of the United States. Further analysis is needed to know if there is any significance in the specifics of this set.

4. Conclusion

There are several steps that will be taken in the immediate future. We will implement algorithms to combine multiple datasets into one comprehensive graph shift operator. We will then use these algorithms to merge our cleaned election data with Census median income data. As stated in Section 4 this will improve the validity of this research and should improve the accuracy of our forecasts. There are several future directions this research can take. The first is that we will apply this research to the 2016 election and improve our model based on the results. The second is exploring how noise effects the model. Currently we reconstruct the signal election signal based on implied perfect sampling. That is, in reconstructing the 2012 election we restrict ourselves to sampling the actual signal at K nodes. We then use these K nodes to predict the election signal. In reality we cannot sample the signal before it is predicted, so we would need to use polling in the subset of counties we identify and then "reconstruct" to predict the election.

References

[1] Abramowitz, A. (2016). A Simple Model for Predicting Hillary Clinton's Vote in the March 15 Democratic Primaries, *Larry J. Sabato's Crystal Ball*, University of Virginia Center for Politics

[2] Anis, A., Gadde, A., & Ortega, A. (2014). Towards a Sampling Theorem for Signals on Arbitrary Graphs, *Proceedings of ICASSP 2014*, 10.1109/ICASSP.2014.6854325

[3] Apte, C., Damerau, F., & Weiss, S. M. (1994). Automated Learning of Decision Rules for Text Categorization, *ACM Transactions on Information Systems (TOIS)*, 12(3), 233-251. 10.1145/183422.183423

[4] Brown, L. B., & Chappell, H. W. (1999). Forecasting Presidential Elections using History and Polls, *International Journal of Forecasting*, 15(2), 127-135. 10.1016/S0169-2070(98)00059-4

[5] Chen, S., Varma, R., Sandryhaila, A., & Kovacevic, J. (2015). Discrete Signal Processing on Graphs: Sampling Theory, *IEEE Transactions on Signal Processing*, 63(24), 6510-6523. 10.1109/TSP.2015.2469645

[6] Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Van Wedeen, J., & Sporns, O. (2008). Mapping the Structural Core of Human Cerebral Cortex, *PLoS Biology*, 6(7), 1479-1493. 10.1371/journal.pbio.0060159

[7] Hillygus, D. S. (2011). The Evolution of Election Polling in the United States, *Public Opinion Quarterly*, 75(5), 962-981. 10.1093/poq/nfr054

[8] Lewis-Beck, M. S., & Rice, T. W. (1984). Forecasting Presidential Elections: A comparison of naive models, *Political Behavior*, 6(1), 9-21. 10.1007/BF00988226

[9] Lowe, D. G. (1999). Object Recognition from Local Scale-invariant Features, *Proceedings of ICCV 1999*, 10.1109/ICCV.1999.790410

[10] Marques, A. G., Segarra, S., Leus, G., & Ribeiro, A. (2015). Sampling of Graph Signals with Successive Local Aggregations, *IEEE Transactions on Signal Processing*, 64(7), 1832-1843, 10.1109/TSP.2015.2507546

[11] Mokrzycki, M., Keeter, S., & Kennedy, C. (2009). Cell-phone-only Voters in the 2008 Exit Poll and Implications for Future Non-coverage Bias. *Public Opinion Quarterly*, 73(5), 845-865. 10.1093/poq/nfp081

[12] Sandryhaila, A., & Moura, J. M. F. (2013). Discrete Signal Processing on Graphs, *IEEE Transactions on Signal Processing*, 61(7), 1644-1656.

[13] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., & Vandergheynst, P. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Data Domains. *IEEE Signal Processing Magazine*, 30(3), 10.1109/MSP.2012.2235192

[14] Zhang, C., Member, S., Flor, D., Member, S., & Fellow, P. A. C. (2015). Graph Signal Processing - A Probabilistic Framework, *Microsoft Research Technical Report*, MSR-TR-2015-31, 1-10.