 doi:10.20944/preprints202209.0092.v1

Type of the Paper (Article)

Asessing the accuracy of Google Trends for predicting presidential elections: The case of Chile 2006-2021

Francisco Vergara-Perucich [1]*

[1] Universidad de Las Américas, Chile;jvergara@udla.cl
* Correspondence: jvergara@udla.cl

Abstract: This article presents the results of reviewing the predictive capacity of Google trends for national elections in Chile. The electoral results of the elections between Michelle Bachelet and Sebastián Piñera in 2006, Sebastián Piñera and Eduardo Frei in 2010, Michelle Bachelet and Evelyn Matthei in 2013, Sebastián Piñera and Alejandro Guillier in 2017, and Gabriel Boric and José Antonio Kast in 2021 were reviewed. The time series analysed were organised on the basis of relative searches between the candidacies, assisted by R software, mainly with the gtrendsR and forecast libraries. With the series constructed, forecasts were made using the ARIMA technique to check the weight of one presidential option over the other. The ARIMA analyses were performed on 3 ways of organising the data: the linear series, the series transformed by moving average and the series transformed by Hodrick-Prescott. The result indicates that the method offers optimal predictive ability.

Keywords: Arima; elections; time series; forecasting; Chile

1. Introduction

Chile has a presidential system where the president acts as head of state and head of government. The nation has three branches: executive, legislative and judicial. Based on the principles of the political system defined in the constitution, only the executive and legislative branches are elected by popular, open and voluntary vote. However, voting has been voluntary only since 2009. The Chilean presidential system establishes that laws and regulations that require fiscal budgetary expenditure depend exclusively on the president of the republic, while other types of initiatives that arise from the legislative branch (chamber of deputies and senate) need presidential sponsorship or should not require fiscal expenditure. This reality means that presidential elections take on special relevance in defining the destiny of the nation, given that it will be the government project headed by the president in office that will determine the guidelines along which the country will advance during the four-year presidential term in Chile. In simple words, the nation defines its roadmap every four years by deciding who will be the next president of the republic. This condition makes presidential campaigns highly intensive in terms of media demand, information flow and civic and political interaction. The campaigns of each conglomerate usually last about 90 days, while the electoral period is close to two months. The main conglomerates are composed of two major right-wing parties: Union Democrata Independiente (UDI) and Renovación Nacional (RN), recently joined by Evopoli, Partido de la Gente and Republicanos. There is another centre-left conglomerate composed of the Christian Democratic Party (DC), the Party for Democracy (PPD), the Radical Party (PR) and the Socialist Party (PS). Since 2017, a conglomerate of progressive left-wing parties called Frente Amplio (FA) has been strongly established, which together with the Communist Party (PC) and other progressive factions of the Socialist Party now make up a relevant political force (the current president Gabriel Boric comes from these forces). On the left, there are other groups with less electoral weight in terms of presidential campaigns, which is what is reviewed in this article. The Chilean political landscape is currently undergoing significant structural redefinitions.

In October 2019, Chile entered a process of social revolt triggered by various reasons that stem from the degradation of democratic institutions in their ability to represent the needs of the population and the structural inequality in constant reproduction. (Arias-Loyola, 2021; Mayol, 2019). At a critical moment in Chile's political history, the political class decided to initiate a constituent process to replace the constitution implemented during the dictatorship of Augusto Pinochet with a new one drafted in democracy, which would redefine a large part of the electoral map. In addition, Pinochet's constitution had already been in force for 39 years and retained a set of locks that prevented changes in line with the social needs of the 21st century. (Salazar, 2020). By means of a transversal political agreement, a constituent process was initiated to draft a new constitutional text. With this process, the political scene is becoming more and more heated, and the 2021 presidential elections will become the most widely contested voluntary elections in the country's history. In this context of heightened civic activity, the results of this study would allow us to understand in part the advance of the public space where civic interaction usually takes place, with the virtual political space.

Google has become the main connector between questions and answers in the world, achieving significant penetration among its users. According to TRUELIST, 1.2 trillion global searches are performed annually. (TURELIST, 2022). The relationship between user interest and access to information is increasingly used in different studies to identify patterns, preferences, business opportunities and effectiveness of business campaigns, among multiple applications. It is also beginning to be used in scientific research to access data that facilitates the process of analysing certain concepts, trends and information flows, including the possibility of using this information to predict potential electoral results, thus awakening the interest of the public in the use of information. (Prado-Román et al., 2021)This has awakened the political world's interest in incorporating these monitoring and diagnostic elements into the design of campaigns. The greater the penetration of internet use in a population, the greater the accuracy of electoral forecasting tools based on these data sources should be. According to Trevisan, as early as 2014, 80% of web searches were conducted from Google worldwide, making it feasible for electoral forecasting tools based on these data sources to be more accurate. (Trevisan, 2014)thus making it feasible to explore the relationships between such searches and voters' electoral choices. Even more determinedly, Ma-Kellams et al. argue that Google searches are the main predictor of electoral choice over other alternatives. (Ma-Kellams et al., 2018)even discussing the accuracy options with probabilistic polls.

This research reviews the predictive value of data obtained from Google Trends for general elections in Chile from 2006 to 2021. The basis for this review is that currently in this nation the internet has a penetration of 82.3% and there are more than 15 million active users. (*Digital in Chile*, 2022). From the series of data collected in that period of time from Google Trends, a time series model is applied to make forecasts based on the autoregressive integrated moving average (ARIMA) technique for each of the elections. The aim is to test the efficacy of the methodology, reviewing scopes and analysing the results. A high predictive capacity of this instrument to identify the winners of each election is observed, in addition to a high accuracy of the result for 66% of the cases studied. In presenting and discussing the results, we conclude positively on the methodological value of the findings that emerge from this research, confirming that this modelling technique is adequate for the Chilean case.

2. Data and Methods

Predictive analytics using time series has different approaches, but they are based on the principle of searching for causality by using past values to predict future values, in serial time ordering from oldest to newest to generate results suitable for causal inference between observations. (Angrist & Pischke, 2008; Gujarati & Porter, 2009). In the case of electoral studies, the use of this methodological field for forecasting is becoming more and more common. Cantini et al. manage to demonstrate the analytical value for electoral climates from social network data in an effective way, as long as they manage to clean from the interactions agents that muddy the discussion or manage to remove concepts that confuse the object of search: voting intention and information about electoral options. (Cantini et al., 2022). Skoric et al. review predictive studies using social network data to predict elections and indicate that the highest accuracy is achieved with machine-learning methods using time series data. (Skoric et al., 2020). This finding is consistent with Schoen et al. who indicate that the best mechanism for predicting futures from social networks is through advanced statistical methods. (Schoen et al., 2013). Using data from Twitter and Facebook, Chauhan et al. indicate that the analysis of sentiment in social networks can generate accurate predictions about political scenarios, given that they allow us to understand the general climate of opinion in the face of elections. (Chauhan et al., 2021). Bilal et al. achieve significant accuracy of Pakistan's 2018 election results from Twitter data that after extensive cleaning can be used as valid factors to identify electoral intentions and potential outcomes at the ballot box. (Bilal et al., 2018). Schmidbauer et al. describe how tracking hashtags on Instagram presented valuable results for predicting that Trump would triumph over Clinton in the 2016 US election. (Schmidbauer et al., 2018). Chin and Wang apply predictive time series techniques to review the predictive value of social networks against the 2018 Taiwan election, indicating that incorporating Facebook into the analysis matrices used considerably increases the predictive value. (Chin & Wang, 2021). Unlike the aforementioned cases, this article contributes using a statistical method little explored for these cases, which is the prediction model using an ARIMA model from serial data collected in Google Trends for different social networks.

The use of Google data has proven to be effective for election forecasting, for which there is already scientific evidence. Trevisan et al. manage to demonstrate the importance of using Google trends to achieve a successful programmatic design of a candidate, being a useful tool to capture undecided voters, while allowing monitoring the progress of the campaign over time. (Trevisan et al., 2018). While some studies have indicated some problems in developing forecasts based on Google Trends, the errors can be corrected in the future. (Yasseri & Bright, 2014)The errors can be corrected based on the development of add-ons to the core sample that can be obtained from Google Trends searches. (Lui et al., 2011). Some studies based on the Google Trends study for the 2015 Greek referendum indicate that this tool has an important predictive capacity in short time intervals, despite the high volatility that can be seen in the political scenarios of such cases. (Askitas, 2015; Mavragani & Tsagarakis, 2019). Similarly, Graefe and Armstrong analysed presidential elections using Google Insights for Search, discovering significant productive power in the data used. (Graefe & Armstrong, 2012). Prado-Román et al. confirm the findings of previous studies that take Google trends to predict election outcomes and conduct a study for every presidential election in the United States and Canada from 2004 to 2019. (Prado-Román et al., 2021). This research is inspired in part by this cited work, to which they add as a predictive tool a time series modelling, taking binary choices that are synthesised in the rate of dominance of one over the other, to study the predictability of the sample.

This is an exploratory quantitative research approach based on an ARIMA model to develop univariate predictive analyses. These models do not assume exogenous structural conditions, since they work on the basis of the internal variations of each group of observations. It is based on the assumption that previous values and their standard errors contain the necessary information to predict future values. In that sense, the advantage of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

ARIMA models is that their consistency depends mainly on the data to be used rather than on other factors as in multivariate models. However, this can also be a limitation, since it does not consider other variables to place the analyses in broader theoretical contexts that seek to explain social phenomena. To achieve accuracy, ARIMA models require that the data for the time series be meticulously constructed, applying as many filters as possible to ensure that what is being asked of the predictive model is being measured. It can be said, then, that the ARIMA models are essentially exploratory (Lui et al., 2011) and thus fulfil the purpose of this research: to provide a methodologically valid, repeatable and reliable mechanism to assess whether elections in Chile could be predicted from data obtained from Google Trends. In addition, ARIMA models have proven to be tremendously useful for predicting scenarios in the short term, as is done in Google Trends. (Litterman, 1986; Stockton & Glassman, 1987)The aim of this research is to see how people's interest in an electoral option in a period of around 90 days achieves predictive capacity of the expected outcome.

The notation for the models to be used is expressed as ARIMA (p,d,q), where p is the number of autoregressive terms, q is the number of terms to consider for calculating the moving averages and d indicates the number of differences that must be incorporated into the model to ensure the stationarity of the sample. The process of calculating the ARIMA model starts by identifying the structural order of the model to be used, defining the integer values (p,d,q), estimating the coefficients for the formulation, checking the fit of the residuals based on a Ljung test and forecasting the future results for a certain number of observations. Before running the ARIMA models, it is essential that the data series is appropriate for evaluation, which is defined on the basis of an Augmented Dickey-Fuller (ADF) test, which allows checking for autocorrelation problems. In this research, the analysis is performed in R software, using the tseries (Trapletti & Hornik, 2022) and forecast (Hyndman & Khandakar, 2008) packages for the calculation of forecasts. The notation of the model can be explained as follows:

$$\left(1-\sum_{i=1}^{p\prime} \propto i\, Li\right) Xt = \left(1+\sum_{i=1}^{q} \theta i\, Li\right) \varepsilon t$$

In the function, ∝i corresponds to the autoregressive parameters of the model, θi corresponds to the moving averages, Li are the lags, Xt is an integrated index, p and q are the components of the series and εt represents the standard error. For this study, in order to reduce the computational error and to order results, we worked in R software.

In the R environment, the data is obtained from the gtrendsR (Massicotte & Eddelbuettel, 2022)package, which allows to extract trend information from Google, identifying variations in a set of periodic variables that assess the interest over time of some concepts searched from the R interface, in this case. The general search model applied followed the following first-order function:

```
Dataset <- gtrendsR:: gtrends(keywords=c('candidate 1 -candidate 2', 'candidate 2 -candidate 1'), geo = 'CL', time = 'YYYYY-mm-dd YYYYY-mm-dd')
```

The above-mentioned code allows to collect the data compared between one option and the other. After this search, data is extracted from the variable 'Hits' within the extracted data subset called 'Interest Over Time'. With the hits data, a single time series is composed based on the following criteria:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

Time series = Hits Candidate 1 / (Hits Candidate 1 + Hits Candidate 2)

This time series is then smoothed by two strategies: 7-day moving average and Hodrick - Prescott smoothing. This smoothing seeks to create a uniform criterion for all the studies developed, with the aim of reducing the problems associated with missing data for some days. Finally, an ARIMA forecast is applied for the three series: (i) series without transformation, (ii) series smoothed by moving average and (iii) series smoothed by Hodrick-Prescott. The forecasts are calculated using the R forecast library, developed by Rob Hyndman.

Table 1. Descriptive statistics for time series data 2005-2021

Elections	Min	1st quartile	Median	Mean	3rd quartile	Max	NA
Bachelet - Piñera 2006	0.2196	0.3630	0.4419	0.4877	0.6248	0.9286	6
Piñera - Frei 2010	0.3549	0.5485	0.6229	0.6055	0.6763	0.8175	0
Bachelet - Matthei 2013	0.4662	0.6177	0.6657	0.6549	0.7191	0.7825	3
Piñera - Guillier 2017	0.6400	0.7816	0.8220	0.8107	0.8622	0.9332	6
Boric - Kast 2021	0.2857	0.3572	0.4121	0.4238	0.4867	0.5952	6

The data series have a daily frequency and in order to unify the criteria, information is collected from 126 days before election day and forecast from day 5 before the election. In other words, 121 observations are used for the modelling.

3. Results

The results described below are favourable for the use of this data analysis technique. Each election is reviewed in detail and the models that best fit the final result are compared. In the first modelling (Table 2), we work with the 2006 presidential campaign, between Michelle Bachelet and Sebastián Piñera. Three ARIMA models were applied: (0,1,1), (2,1,3) and (2,1,2), with sigma2 values suitable for the modelling process. One of the differences between the three models applied can be seen in the standard error which is highly variable. However, the ARIMA modelling for the Hodrick-Prescott smoothed series, which has a very low standard error, gave an excellent forecast, differing by only 0.78% from the final election result of 53.5% for Michelle Bachelet. On the other hand, the moving average forecast only had an error of 0.36% in relation to the final election result, but with a standard error of 10.53%, so the most reliable and effective modelling series in this case was Hodrick-Prescott.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

Table 2. Forecasting results election between Michelle Bachelet vs Sebastián Piñera 2006.

Variables	Time Series	ARIMA	Sigma ARIMA model	p-value Box Test by Ljung-Box	Average forecasting result	Election result	Difference between Forecast and Election result	Standard Error
Bachelet – Piñera 2006 (Relative)	Normal	(0,1,1)	0. 1632037	0.5138308	0.5084575	0.535	0.02761574	0.405063
	Moving Average	(2,1,3)	0.004500995	0.9093858	0.5389227	0.535	0.003642637	0.1053066
	Hodrick-Prescott	(2,1,2)	4.774217e-06	0.9439797	0.527154	0.535	0.007846023	0.01271612

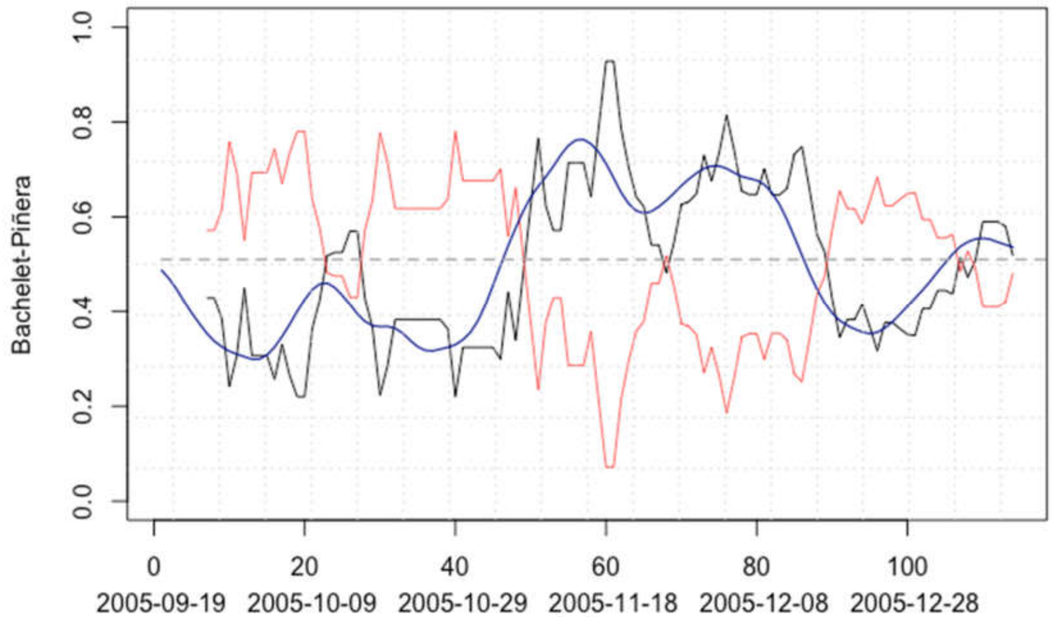

Figure 2. Forecasting results Bachelet – Piñera 2006.

In the second modelling (Table 3), we work with the 2009-2010 presidential campaign between Sebastián Piñera and Eduardo Frei. Three ARIMA models were applied: (0,0,1), (1,0,0) and (4,1,0), with sigma2 values suitable for the modelling process. One of the differences between the three models applied can be seen in the standard error which is highly variable although not as divergent as in the previous case. However, the ARIMA modelling for the series smoothed by Hodrick-Prescott is again the one with a very low standard error and offers the best forecast, differing only by 0.092% from the final election result of 51.5% for Sebastián Piñera. On the other hand, the moving average forecast in this case had an error of 5.32% in relation to the final election result, but with a standard error of 6.43%, so the most reliable and effective modelling series in this case was Hodrick - Prescott. If in this case and in the previous one the modelling had been done only by moving average, the standard error does not allow detecting the definitive winner, since the variance may fall below 50% of preferences on who won the election, which is problematic, beyond the fact that in all the averages of the forecasts the winner is given as the one who finally won the election.

Table 3. Forecasting results election 2010 between Sebastián Piñera and Eduardo Frei.

Variables	Time Series	ARIMA	Sigma ARIMA model	p-value Box Test by Ljung-Box	Average forecasting result	Election result	Difference between Forecast and Election result	Standard Error
Piñera - Frei 2010 (Relative)	Normal	(0,0,1)	0.04846787	0.9696941	0.5992357	0.515	0.08423567	0.223711
	Moving Average	(1,0,0)	0.002123367	0.19944	0.5682475	0.515	0.05324748	0.06432831
	Hodrick-Prescott	(4,1,0)	9.841646e-07	0.8237633	0.5140846	0.515	0.0009154432	0.006677285

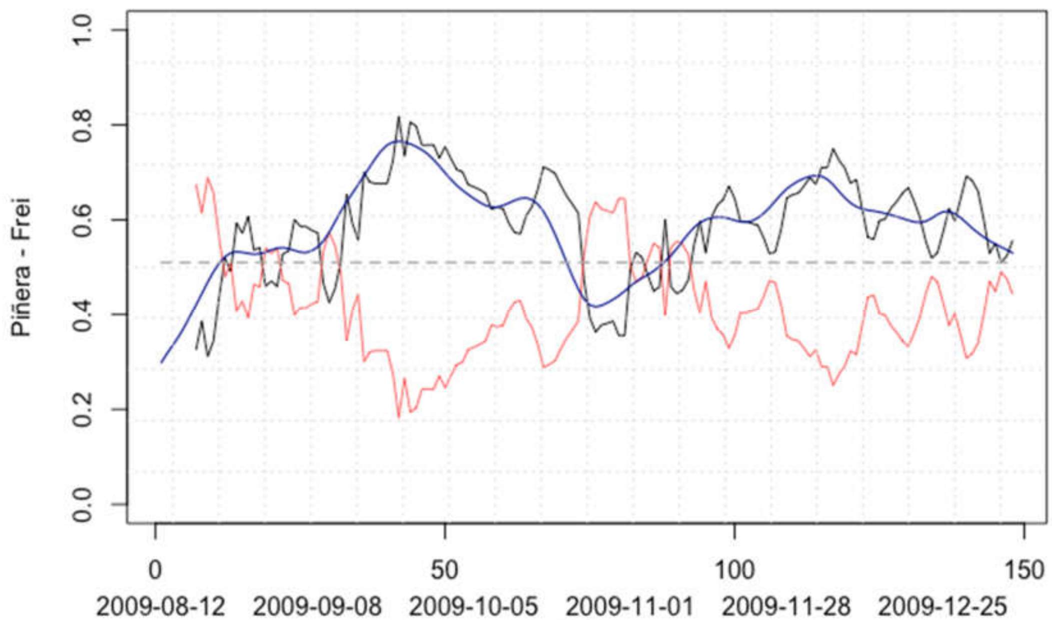

Figure 3. Forecasting results Piñera – Frei 2010.

In the third modelling (Table 4), we work with the 2013 presidential campaign between Michelle Bachelet and Evelyn Matthei. Three ARIMA models were applied: (1,0,1), (2,1,0) and (3,1,0), with sigma2 values suitable for the modelling process. In this case, the standard error is less variable than in the two previous cases. The ARIMA modelling for the Hodrick-Prescott smoothed series has the lowest standard error and offers the best forecast, differing by only 1.03% from the final election result of 62.17% for Bachelet. Unlike the previous case, in this modelling the moving average is no more accurate than the series without transformation, which had an error of 1.78% with the final result. The confirmation remains that the best model for this type of forecasts is for a series smoothed by Hodrick-Prescott, which also remains at a very low standard error.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

Table 4. Forecasting results election 2013 between Michelle Bachelet and Evelyn Matthei.

Variables	Time Series	ARIMA	Sigma ARIMA model	p-value Box Test by Ljung-Box	Average forecasting result	Election result	Difference between Forecast and Election result	Standard Error
Bachelet – Matthei 2013 (Relative)	Normal	(1,0,1)	0.0156638	0.864565	0.6395573	0.6217	0.01785731	0.1307915
	Moving Average	(2,1,0)	0.0006456679	0.9911486	0.6905575	0.6217	0.06885749	0.04694867
	Hodrick-Prescott	(3,1,0)	4.422524e-07	0.8184402	0.6113527	0.6217	0.01034734	0.004601662

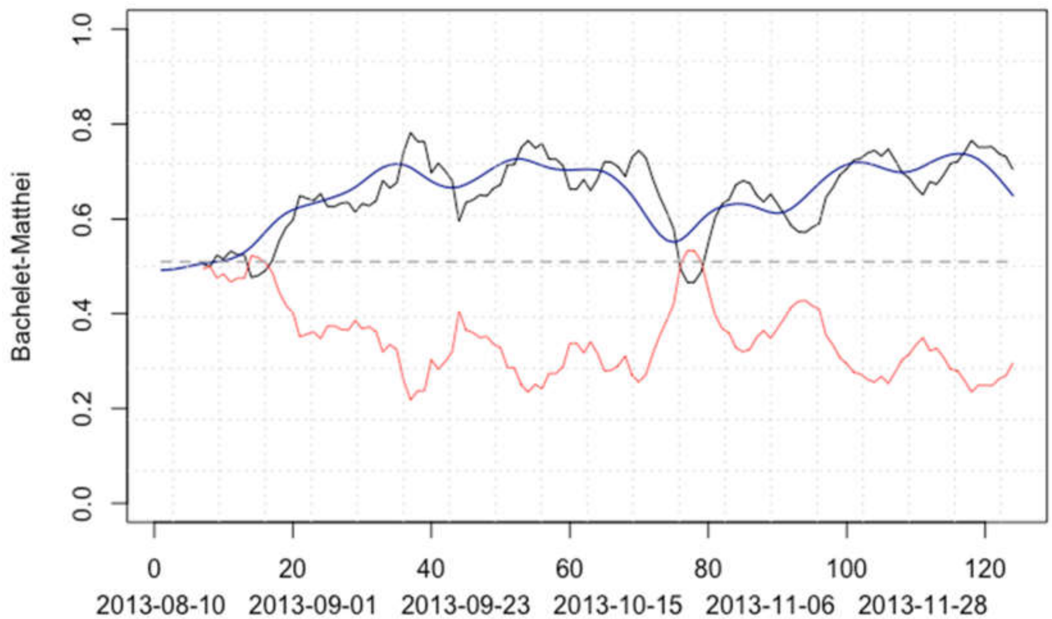

Figure 4. Forecasting results Bachelet – Matthei 2013.

In the fourth modelling (Table 3), we work with the 2017 presidential campaign between Sebastián Piñera and Alejandro Guillier. Three ARIMA models were applied: (1,1,1), (1,1,0) and (4,1,0), with sigma2 values suitable for the modelling process. Of all the modelling, this is the least accurate, where the best model is by Hodrick-Prescott, which differs from the final result by 6.61%, which was favourable to the candidate Sebastián Piñera. What is interesting is that despite not being accurate, it predicts the winner and overestimates its influence rather than modelling indicatively that Guillier's option would win. In other words, in this case the model is not totally accurate in the percentage result but still indicates the winning option effectively.

Table 5. Forecasting results election 2017 between Sebastián Piñera and Alejandro Guillier.

Variables	Time Series	ARIMA	Sigma ARIMA model	p-value Box Test by Ljung-Box	Average forecasting result	Election result	Difference between Forecast and Election result	Standard Error
Piñera – Guillier 2017 (Relative)	Normal	(1,1,1)	0.01176181	0.7434989	0.6652471	0.5458	0.1194471	0.1121784
	Moving Average	(1,1,0)	0.0005553617	0.8413091	0.6335689	0.5458	0.08776888	0.04116279
	Hodrick-Prescott	(3,1,0)	3.32157e-07	0.6613639	0.611959	0.5458	0.06615895	0.00387286

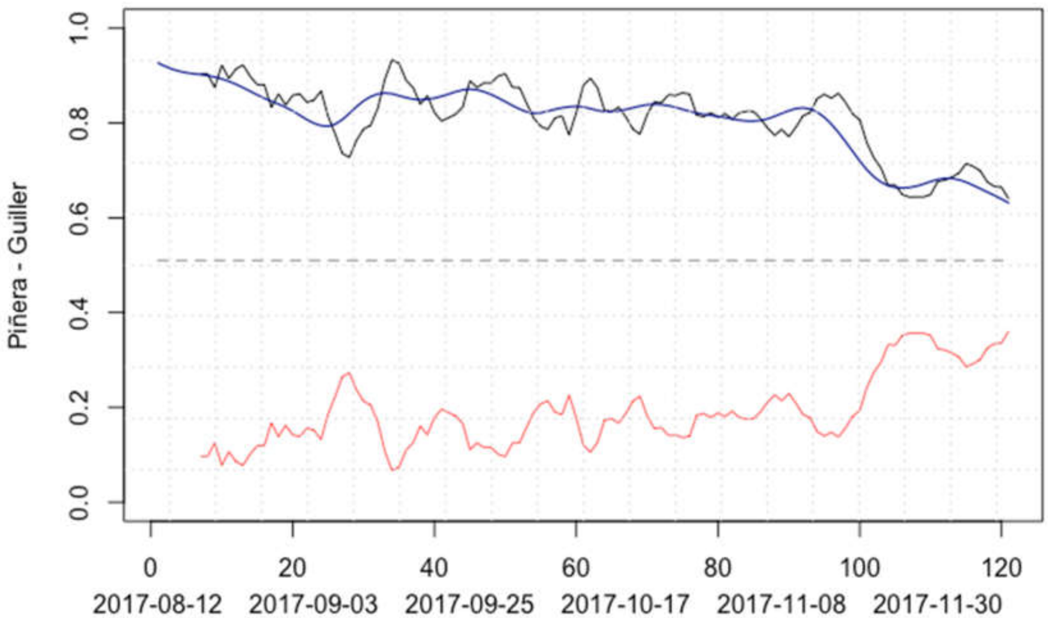

Figure 5. Forecasting results Piñera - Guillier 2017.

Finally, Table 6 indicates the outcome for the 2021 presidential election between Gabriel Boric and José Antonio Kast. This modelling is the only one that presents a forecast that did not point to the definitive winner of the election, since the series without transformation gave Kast as the winner when Boric actually won. However, the Hodrick-Prescott modelling presents a forecast that only differs from the actual result by 0.48%, with a standard error of 0.49%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

Table 6. Forecasting results election 2020 between Gabriel Boric and José Antonio Kast.

Variables	Time Series	ARIMA	Sigma ARIMA model	p-value Box Test by Ljung-Box	Average forecasting result	Election result	Difference between Forecast and Election result	Standard Error
Boric – Kast 2021 (Relative)	Normal	(2,0,0)	0.01887498	0.8072142	0.4623903	0.5564	0.0940097	0.140894
	Moving Average	(1,0,0)	0.000893124	0.2021044	0.5102202	0.5564	0.04617981	0.04325613
	Hodrick-Prescott	(2,2,3)	5.784626e-07	0.9772097	0.5612801	0.5564	0.004880149	0.004917933

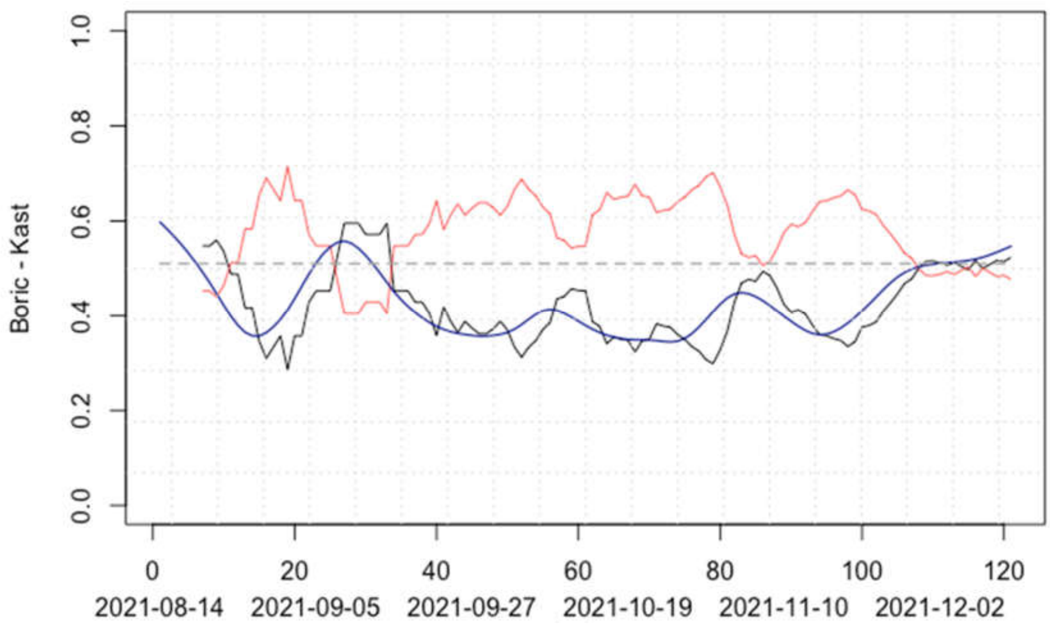

Figure 6. Forecasting results Boric – Kast 2021.

4. Discussion and Conclusions

After applying the modelling to generate forecasts, it can be argued that the use of Google Trends to identify the candidates most likely to win in Chile is highly effective. The following table allows us to evaluate in summary the total of the forecasts developed. Undoubtedly, the most effective and accurate mechanism is by smoothing with the Hodrick-Prescott technique, averaging a difference with the final result of 1.8% (Table 7), a result inflated by the error in the case of the election between Sebastián Piñera and Alejandro Guillier in 2017. This is indicative that to achieve greater precision, specific filtering mechanisms can be sought that are temporally placed on what was being discussed on social media and what was being searched on Google during the election, in order to discern with greater understanding which keywords should be excluded from searches.

Table 7. Evaluación de resultados de los pronósticos según elección y modelo ARIMA utilizado.

Election	Model	Assertion on winner
Bachelet – Piñera 2006	Normal ARIMA:(0,1,1)	Yes
Bachelet – Piñera 2006	Moving Average ARIMA:(2,1,3)	Yes
Bachelet – Piñera 2006	Hodrick-Prescott ARIMA:(2,1,2)	Yes
Piñera - Frei 2010	Normal ARIMA:(0,0,1)	Yes
Piñera - Frei 2010	Moving Average ARIMA:(1,0,0)	Yes
Piñera - Frei 2010	Hodrick-Prescott ARIMA:(4,1,0)	Yes
Bachelet – Matthei 2013	Normal ARIMA:(1,0,1)	Yes
Bachelet – Matthei 2013	Moving Average ARIMA:(2,1,0)	Yes
Bachelet – Matthei 2013	Hodrick-Prescott ARIMA:(3,1,0)	Yes
Piñera – Guillier 2017	Normal ARIMA:(1,1,1)	Yes
Piñera – Guillier 2017	Moving Average ARIMA:(1,1,0)	Yes
Piñera – Guillier 2017	Hodrick-Prescott ARIMA:(3,1,0)	Yes
Boric – Kast 2021	Normal ARIMA:(2,0,0)	No
Boric – Kast 2021	Moving Average ARIMA:(1,0,0)	Yes
Boric – Kast 2021	Hodrick-Prescott ARIMA:(2,2,3)	Yes
Asserted?	NO	7%
	YES	93%
Difference with results	Average Normal Serie	6,86%
	Average Moving Average	5,19%
	Average Hodrick Prescott	1,80%

In the result analysis, out of 15 models, only 1 model failed to identify the winner, i.e., for this analysis, 93% of the models do identify the winner of the election. Possibly, the application of other search cleaning strategies, associated with exclusionary keywords, could help to reduce the probability of errors. However, the model is still effective when three techniques are applied simultaneously to assess which one might be providing information that confounds the interpretation of the forecasts. In any case, all smoothed assessments, whether by moving average or Hodrick-Prescott, were successful in indicating who would win the election.

These results allow us to contribute to the international literature on the predictive electoral value of Google search trends. The assumption that could explain this predictive capacity is that people search for information on Google to inform their voting decision and in doing so allow us to record with good accuracy which of the electoral options is generating the most interest among the population. Google Trends also offers the possibility to explore trends within each search in order to apply both filters and also to identify the topics associated with the searches that people are most interested in.

In Chile, Google penetration is significant, so the question arises as to whether this forecasting strategy would be applicable to other nations where there is less internet access or, conversely, whether in a nation with much greater internet access coverage the model would gain or lose the predictive capability it has shown in the modelling shown here. There is also the question of the ability to scale this type of search while maintaining good predictive results. In Chile, Google Trends allows the interest aroused by the words searched for to be separated by region, so that a specific study can be carried out for each territory. In this case, no such test has been carried out. A very good predictive capacity has been proven and one of the pending tasks is to move from a national analysis to specific regions or cities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0092.v1

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Anokye, R., Acheampong, E., Owusu, I., & Obeng, E. I. (2018). Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. *COGENT SOCIAL SCIENCES*, *4*(1). https://doi.org/10.1080/23311886.2018.1461544

Arias-Loyola, M. (2021). Evade neoliberalism's turnstiles! Lessons from the Chilean Estallido Social. *Environment and Planning A: Economy and Space*, *0*(0), 0308518X2199783. https://doi.org/10.1177/0308518X21997832

Askitas, N. (2015). *Calling the Greek Referendum on the nose with Google Trends*. https://ideas.repec.org/p/rsw/rswwps/rswwps249.html

Bilal, M., Asif, S., Shainila, Y., & Afzal, U. (Eds.). (2018). 2018 Pakistan General Election: Understanding the Predictive Power of Social Media. In *2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*.

Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2022). Analyzing Political Polarization on Social Media by Deleting Bot Spamming. *Big Data and Cognitive Computing*, *6*(1), 3. https://doi.org/10.3390/bdcc6010003

Chauhan, P., Sharma, N., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), 2601–2627. https://doi.org/10.1007/s12652-020-02423-y

Chin, C., & Wang, C. (2021). A new insight into combining forecasts for elections: The role of social media. *Journal of Forecasting*, *40*(1), 132–143. https://doi.org/10.1002/for.2711

Digital in Chile: All the Statistics You Need in 2021. (2022). DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2021-chile

Graefe, A., & Armstrong, J. S. (2012). Predicting elections from the most important issue: A test of the take-the-best heuristic. *Journal of Behavioral Decision Making*, *25*(1), 41–48. https://doi.org/10.1002/bdm.710

Gujarati, D. N., & Porter, D. C. (2009). Basic of Econometric, Fifth Edition. In *Econometrics*.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The **forecast** Package for *R*. *Journal of Statistical Software*, *27*(3). https://doi.org/10.18637/jss.v027.i03

Litterman, R. B. (1986). Forecasting With Bayesian Vector Autoregressions—Five Years of Experience. *Journal of Business & Economic Statistics*, *4*(1), 25–38. https://doi.org/10.1080/07350015.1986.10509491

Lizardo, C. G. de. (2011). *Introducción a la econometria aplicada*. Editorial Trillas.

Lui, C., Metaxas, T., & Mustafaraj, E. (2011). On the predictability of the U.S. elections through search volume activity. *IADIS International Conference*, *1*(1). https://repository.wellesley.edu/object/ir153

Ma-Kellams, C., Bishop, B., Zhang, M. F., & Villagrana, B. (2018). Using "Big Data" Versus Alternative Measures of Aggregate Data to Predict the U.S. 2016 Presidential Election. *Psychological Reports*, *121*(4), 726–735. https://doi.org/10.1177/0033294117736318

Massicotte, P., & Eddelbuettel, D. (2022). *Package 'gtrendsR'*. https://cran.curtin.edu.au/web/packages/gtrendsR/gtrendsR.pdf

Mavragani, A., & Tsagarakis, K. (2019). Predicting referendum results in the Big Data Era. *Journal of Big Data*, *6*(3), 1–20. https://doi.org/10.1186/s40537-018-0166-z

Mayol, A. (2019). *Big Bang: Estallido social 2019*. Catalonia. https://www.libreriacatalonia.cl/product/big-bang-estallido-social

Prado-Román, C., Gómez-Martínez, R., & Orden-Cruz, C. (2021). Google Trends as a Predictor of Presidential Elections: The United States Versus Canada. *American Behavioral Scientist*, *65*(4), 666–680. https://doi.org/10.1177/0002764220975067

Salazar, G. (2020). *Acción Constituyente* (1st ed.). Tajamar Ediciones.

Schmidbauer, H., Rösch, A., & Stieler, F. (2018). The 2016 US presidential election and media on Instagram: Who was in the lead? *Computers in Human Behavior*, *81*, 148–160. https://doi.org/10.1016/j.chb.2017.11.021

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5), 528–543. https://doi.org/10.1108/IntR-06-2013-0115

Skoric, M. M., Liu, J., & Jaidka, K. (2020). Electoral and Public Opinion Forecasts with Social Media Data: A Meta-Analysis. *Information*, *11*(4), 187. https://doi.org/10.3390/info11040187

Sousa Lira, J. M., Salgado, E. G., & Beijo, L. A. (2020). ARIMA models as an alternative to predict the diffusion of the ISO 14001 standard in Europe. *Journal of Environmental Planning and Management*, *63*(2), 275–286. https://doi.org/10.1080/09640568.2019.1577721

Stockton, D. J., & Glassman, J. E. (1987). An Evaluation of the Forecast Performance of Alternative Models of Inflation. *The Review of Economics and Statistics*, *69*(1), 108. https://doi.org/10.2307/1937907

Trapletti, A., & Hornik, K. (2022). *tseries: Time Series Analysis and Computational Finance*. https://CRAN.R-project.org/package=tseries

Trevisan, F. (2014). Search engines: From social science objects to academic inquiry tools. *First Monday*. https://doi.org/10.5210/fm.v19i11.5237

Trevisan, F., Hoskins, A., Oates, S., & Mahlouly, D. (2018). The Google voter: Search engines and elections in the new media ecology. *Information, Communication & Society*, *21*(1), 111–128. https://doi.org/10.1080/1369118X.2016.1261171

TURELIST. (2022). How Many People Use Google. *TrueList*. https://truelist.co/blog/how-many-people-use-google/

Yasseri, T., & Bright, J. (2014). *Can electoral popularity be predicted using socially generated big data?* (arXiv:1312.2818). arXiv. http://arxiv.org/abs/1312.2818