Predicting elections with emerging political parties [1]

José G. Montalvo Omiros Papaspiliopoulos
Universitat Pompeu Fabra-ICREA
Timothée Stumpf-Fétizon (UPF)

[1] Financial support from the Spanish Ministerio de Economia y Competitividad, the Barcelona GSE Research Network, and the Government of Catalonia is gratefully acknowledged.

Abstract

We propose a new methodology for predicting election results that combines respondent-level survey data and national polls within a Bayesian synthesis framework. This methodology is largely motivated by the specific challenges of forecasting elections with the participation of new political parties. This situation is especially relevant in post-2008 European elections. The increasingly frequent competition of emerging parties, that enter political competition in electoral markets traditionally contested by two large parties, creates important challenges for classic methods of electoral prediction based on historical data. However, our methodology can also be applied to forecast elections without new contenders. We illustrate the advantages of our methodology using the 2015 Spanish Congressional Election.

1 Introduction

Forecasting in social sciences is a challenging endeavor. Probably one of the most challenging exercises in this respect is the forecasting of election results. Most of the literature on election forecasting, including its methodological underpinning, has focused on two-party political systems, a "winner-take-all" system for the Electoral College and democracies with a long history of past elections. Instead, in this paper we develop a methodology most appropriate for elections where new parties enter the electoral competition between two consecutive elections, under a D'Hondt system for allocation of parliamentary seats, and where the vast majority of available opinion polls predict at the national level whereas the seats are allocated at province level.

The scientific approach to electoral forecasting relies mostly on three alternative methodologies: the statistical modelling approach; the use of polls, either voting intention surveys or party sympathy surveys; and political prediction markets based on bets for the candidates[1]. The statistical modelling approach consists of predicting election results from historical and socioeconomic data, an example is the simple "bread and peace" model of Hibbs (2008)[13]. In stable political systems it is known that national election votes are highly predictable from fundamentals[2] while polls are very variable but contain useful information, specially close to the election day. *Increments*, that is the results of electoral provinces relative to the national result, have been observed to be even more predictable using a statistical model.

Our methodology is based on the statistical synthesis of what we are going to call the fundamental model[3] and opinion polls. The fundamental model is trained on "deep" microdata obtained in the form of preelectoral surveys and it is corrected using post-stratification[4] based on census data. It returns probabilistic predictions of voting intention that apply at province level. On the other hand, national opinion polls are also modelled statistically to produce probabilistic aggregated forecasts of voting intention[5] at the national level. The modelling is used to account for *house effects*, the varying quality of polling methodologies, as well as time-trending that takes place as the election times approaches. We then use the *Hamiltonian Monte Carlo* en-

[1] See Lewis-Beck (2005). Recently there have been attempts to use social media and, in particular, Twitter, to predict elections. Using Twitter has been found to be a poor forecasting strategy (Gayo-Avello 2012). Murthy (2015) shows that tweets are more reactive than predictive.

[2] e.g. Gelman and King (1993).

[3] Some authors refer to this model as the statistical or the econometric model. Since we are going to treat the polls with a statistically based approach that will be integrated with the fundamental model we prefer our chosen terminology.

[4] see e.g. Chapter 14 of Gelman and Hill (2007).

[5] In fact Nate Silver uses the term "snapshot" to refer to the aggregation of polls.

gine *Stan*[6] for model estimation and apply Bayesian updating to synthesize the fundamental and pollster models.

The methodology is illustrated on, and largely has been motivated by, the specific challenges of forecasting elections with the participation of emerging parties. After the beginning of the financial crisis many new parties were created in European countries to capitalize on the discontent of voters with the policy reaction to the economic crisis. Dennison and Pardijs (2016)[6] identify 45 "insurgent" parties in Europe, many of them just a few years old, that come across the political spectrum from extreme left to extreme right. These new political contenders usually enter political competition in electoral markets traditionally contested by two large parties, and they create important challenges for classical electoral forecating methods. While there is evidence of decreasing rates of Americans voting for different parties in succcessive presidential elections[7] the recent European experience is very different: insurgent parties hold 1,329 seats in 27 EU countries, which correspond to 18.3% of the total seats of their parliaments.

The methodology developed in this paper is applied to the national Spanish elections of 2015. The political landscape in Spain is complicated by the existence of numerous political parties with non-trivial representation in certain parts of the country (the so-called nationalist parties, e.g. in Catalonia or the Basque country), the fact that a handful only elections have taken place since the restoration of democracy in the country in 1977 after decades of dictatorship, and that electoral polling is not as extensive as in older democracies (e.g. the USA or the UK). Moreover, polling is hardly available at higher spatial resolutions than national. However, by far the biggest challenge in the 2015 elections is that two new political parties ended up taking more than 30% of the parliamentary seats when they had no political representation in the previous parliament.

This article outlines the proposed methodology, which is novel in the construction of the fundamental model, the pollster aggregation model, and their synthesis. We have opted for a *prospective approach*, according to which we attempt to produce forecasts using increasingly sophisticated approaches, hence showcasing the limitations of simpler approaches and the necessity for the one we advocate. To a large extent the order of methods follows the historical evolution of our work on this forecasting challenge as it has been documented on a blog we maintained during the few months prior to the 2015 Spanish national elections, `http://558project.blogspot.com/`.

Our methodology is inspired by a strand of the literature on forecasting electoral outcomes by combining polls and election forecasts. The specifics

[6] See Carpenter, Gerlman et al. (2015).

[7] Smidt (2015).

of these combinations depend on the objective of the exercise. Park et al. (2004)[21] use a multilevel regression model and post-stratification to obtain state level estimates from national polls. They validate their methodology by comparing their predictions for each state with the actual outcomes of the 1988 and 1992 Presidential elections. Lock and Gelman (2010)[18] use a Bayesian model to perform combine polls with forecasts from fundamentals. They merge a prior distribution, obtained from previous election results, with polls to generate a posterior distribution over the position of each state relative to the national popular vote - we will call this approach an "increment model" later on[8]. The objective of this procedure is not to produce a forecast for the national vote but to develop a methodology that separates national vote from states' relative positions which can be very valuable for individual state forecasts.

Our approach to forecasting elections shares some features with Nate Silver's method[9], but also with other popular approaches to electoral predictions, e.g. Votamatic. However, the specific methodology we use to obtain the fundamental model, the aggregation of the polls and the synthesis of both is quite different from Silver's approach. For example, we do not perform ad hoc corrections to the fundamental model. Instead, the weighting of fundamental model and aggregated polls (what Siver refers to as the "now-cast" or "snapshot") arises organically from the Bayesian framework. The weights change over time depending on the information content of the fundamental model and the polls updates. In addition, we synthesize polls using a statistical model that decomposes the forecasing error by pollster in previous elections into different sources of uncertainty.

The outline of the article is as follows. Throughout we use the 2015 Spanish Congressional Elections as a running example but have developed the material into generic methodological sections and sections specific to the Spanish elections. Section 2 describes the political context of the 2015 Spanish Congressional Election. Section 3 develops the fundamental model, Section 5 the pollster model and Section 7 carries out the synthesis of the two models. Sections 4,6 and 8 show results of applying these models and methods to the 2015 Spanish elections. Section 9 concludes. The Appendix explains the type of data we have used to produce our forecasts and the notation used in the formulae.

Code	Party	Ideology	2011 Result
PSOE	Partido Socialista Obrero Espanol	Center-left	0.288
PP	Partido Popular	Right-wing	0.446
Pod	Podemos (including IU)	Left-wing	N/A
C's	Ciudadanos	Center-right	N/A

Table 1: Spanish parties active at the national level in the 2015 elections.

2 The Spanish 2015 Congressional Election

Since the end of the dictatorship in 1977 Spanish politics was characterised by the alternation in government of two political parties: PP (popular party, conservative) and PSOE (socialists); see Table 1 for the main contenders and their characteristics. Some other small and regional parties also participated in the elections but the two largest parties accounted for 75% to 85% of the vote. In the 2015 Electoral Campaign there were at least four large parties because of the emergence of two new national parties: Podemos (radical left) and C's (Ciudadanos, liberal). Podemos and C's had no seats in previous Spanish parliaments[10], whereas in the 2015 elections they ended up with 69 and 40 respectively, out of 350 in total. This structural change is one of the most challenging issues in predicting the results of the 2015 Spanish Congressional Election and, in general, in any electoral contest where the emergence of new and large political parties change the electoral environment with respect to previous elections[11].

The dissatisfaction of a sizeable part of the population with the measures of austerity applied initially by the PSOE government since 2010 lead to a popular demonstration that occupied the center of Madrid during several weeks. This social movement was named 11-M since their assamblies began May 11, 2011. In March 11, 2014 this movement crystallized in a new political party named Podemos, which quite fast got the support, in polls, of 15% of the likely voters. Podemos was initially marketed as the Spanish Syriza[12]. The leaders of Podemos came mostly from Political Science university departments. Some of them had been members of anticapitalism parties in the radical left position of the spectrum. Although in their program for the first election they competed, the European elections of 2014, they included

[8] For the national popular vote they use the model of Hibbs (2008).

[9] For details on Silver's method visit `http://fivethirtyeight.com/`.

[10] Podemos did not even exist at that time.

[11] Another challenging situation for electoral forecasting in the Spanish context took place in 2004 when a terrorist attack took place in Madrid during the last week before the electoral date when no polls are allowed to be run. See Montalvo (2012).

[12] Syriza, or the Coalition of the Radical Left, is the Greek party that won the 2015 legislative election.

the repudiation of public debt and the nationalization of many industries, their positions evolved later as to avoid extreme policies and try to escape from the radical left tag that they had from the beginning.

In addition, the conservative policies of PP, the corruption associated with conservative politicians and the lack of internal regeneration in the party led to the birth of a new liberal party called Ciudadanos (C's). This party was founded in 2006 but was initially geographically concentrated in Catalonia.

Both Podemos and C's appear in the CIS[13] polls as of July 2014. In contrast with Podemos, the support for Ciudadanos was only 0.9% in early 2014, but built up quickly. In July of 2015 polls showed a tie between these two new political contenders while the sum of the two largest political parties has gone down to 50%. Figure 1 depicts the strength of different parties by province in the 2014 European elections.

The primary challenge from a modelling perspective is that Podemos and Ciudadanos have not inherited their electorate from a distinct previous political movement. On the contrary, they are cannibalizing parties with similar ideologies. The following sections describe the modelling alternatives considered to generate a predictive method for the 2015 Spanish Congressional election and the difficulties imposed by the emergence of these new political parties.

2.1 The Spanish electoral system

The Spanish government is appointed by the *Congreso de los Diputados* which consists of 350 representatives. Each of the 52 Spanish provinces[14] elects its own representatives from its seat contingent according to the local electoral outcome. Thus, as in US presidential elections, the popular vote at the national level is not decisive. Furthermore, the allocation of seats at the province level is proportional, as opposed to the *winner-takes-all* rule that most US states apply in presidential elections. The seat allocation is determined by the *D'Hondt method* and is most easily understood in terms of the equivalent *Jefferson method*, which we may frame in terms of finding the market-clearing point in the market for seats[15].

The Jefferson method is used to find the "price" in votes per seat at which the "demand" for seats by parties equals the available budget. Thus, a

[13] Center for Sociological Research (CIS) a publicly sponsored institution that runs the official polls; see also the Appendix.

[14] Provinces and their official codes are listed on `http://www.ine.es/en/daco/daco42/codmun/cod_provincia_en.htm`.

[15] Udina and Delicado (2005) use data on Spanish elections to show the forecast bias of pre-electoral polls when they convert votes into seats using D'Hondt's rule.

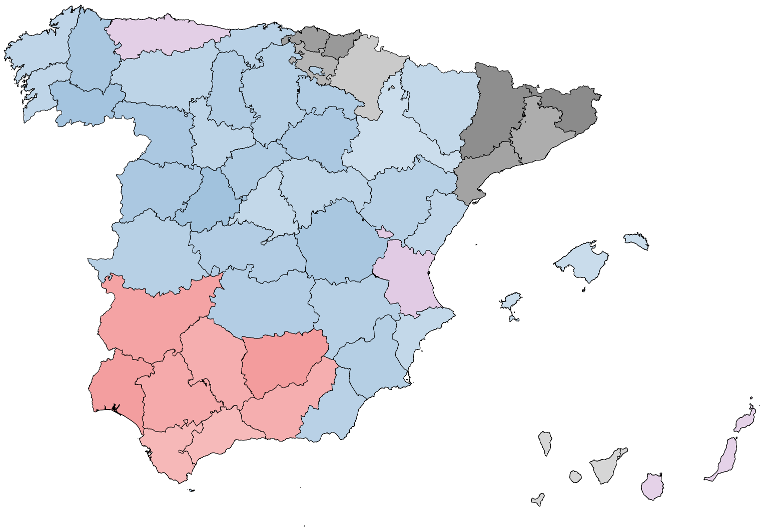

Figure 1: Map of Spanish provinces colored by strongest party in the 2014 European elections and degree of dominance, darker shades corresponding to stronger dominance. Legend: PSOE (red), PP (blue), Podemos+IU (purple), others (gray).

simple iterative algorithm consists of increasing the price per seat until the aggregated demand for seats equals the fixed supply. Then, each party obtains the number of seats it can afford at the equilibrium price.

Since seats are an indivisible good, a party may just fall short of being able to buy an additional seat, with the remainder going to waste. This will occur in every province a party runs in. Thus, given a fixed national vote, it is preferable to have a geographically concentrated electorate. This applies to the regionalist parties in Catalonia and the Basque country.

In the Spanish case, there is an additional rule which states that parties must obtain at least 3% of votes in a given province to take part in the allocation. Otherwise, their votes are disregarded. This acts as an additional penalty on smaller parties whose electorate is spread out across the nation.

3 Fundamental models with emerging parties

We present different approaches of increasing complexity to building a fundamental model and discuss their weaknesses, building up to the models we find most suitable in situations of strong emerging political parties. To put things in perspective, we discuss how these approaches apply to the Spanish 2015 elections.

3.1 Historical models

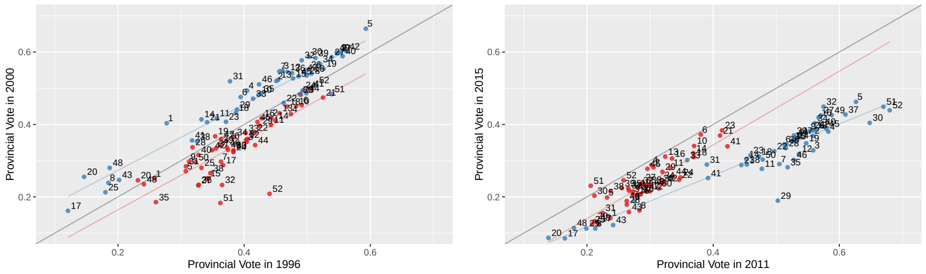

Figure 2: Scatterplot of lagged vote share vs current vote share in 2000 (left) and 2015 (right) relative to previous result, plus robust linear regression line. Legend: PSOE (red), PP (blue). In grey the 45^o line. The labels refer to the INE province code.

Historical models use time series of predictors to run regressions for forecasting the vote proportion or by political party at province level. In stable political systems it is known that the national outcome is highly predictable from fundamentals and past results. This was the case in the Spanish political system until before 2015. For example, Figure 2 (left) plots the electoral result of the 2000 election versus that of the 1996 election for the PP (blue) and the PSOE (red) in each province of Spain, numbered according to the standard postcode coding of Spanish provinces. The picture is similar in other elections prior to 2015. The results positions of provinces relative to the national average are particularly well predictable. This manifests itself through regression lines that are almost parallel the 45^o line, and motivates the use of "increment" models, as we discuss in the following section. Historical models typically include predictors such as unemployment rates, growth of personal disposable income, lagged electoral outcomes, presidential incumbency, regional trends, presidential approval, presidential home advantage (or the corresponding adjustment for party leader home province), partisanship or ideology of each state/district, etc.[16] The first fundamental model we entertain is a historical one, and we use a slightly more principled Bayesian hierarchical specification[17] that includes as predictors the lag share of each party, an incumbency indicator, the rate of unemployment, change of per capita GDP at the national level and the change of per capita GDP at the regional level. The hierarchical model is natural in our context since the data that feed the model are available at different spatial resolutions, see the Appendix.

The performance of the model in a training set of previous Spanish elections was quite good. However, its usefulness and predictive ability for the 2015 Congressional Election was questionable apriori. To start with, such a model cannot provide predictions for parties with no competitive participation in previous elections. Additionally, even when making predictions for PP and PSOE, the traditional political players in Spain, it is unlikely that the model estimated under completely different political environment would have any applicability in new reality; this is already visible in Figure 2 (right) where it is shown that the pattern observed in past elections has indeed been broken in the 2015 elections.

3.2 Increment models

We have already discussed that the positions of electoral districts (provinces) relative to national average are predictable from past results and fundamentals, see e.g. the exit polling approach of Curtice and Firth (2008)[5]. We refer to the difference (or ratio, in an alternative transformation) between province and national result as an "increment" and the models that try to

[16] For instance Campbell (1992, 2008), Gelman and King (1993), Klarner (2008), Fair (2009) or Hummel and Rothschild (2014).

[17] see e.g. Part 2A of Gelman and Hill (2007).

predict those as “increment models”. As any political system, the Spanish one exhibits a degree of partisanship which is persistent across time; this is shown in Figure 2. Such models aim at predicting the *province increment* $\boldsymbol{v}_{tj} - \boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is the national vote over parties in election t and $\boldsymbol{v}_{tj}$ is the vote in province j. If $\boldsymbol{v}_{tj} - \boldsymbol{v}_t$ is persistent over time, we can use the model to predict the future increments $\tilde{\boldsymbol{v}}_j - \tilde{\boldsymbol{v}}$. These models have to be combined with estimates of the national result for each party in order to produce election forecasts at province level, hence they should be thought as component of a larger forecasting machine. The type of data used to train an increment model could vary. For example, historical data could be used, as with the historical models described earlier, but this would be useless when making predictions for emerging parties. Instead, survey data could also be used.

3.3 Our methodology

Given the argumentation in the previous sections, the realities of multiparty parliamentary systems with seat allocation at province level, and the data that are typically available in many countries, the basic characteristics of our fundamental model are driven by the following considerations:

- It should return predictions at the provincial level.
- Point forecasts are almost meaningless if we do not attach a degree of confidence to them. Hence, the methodology should allow probabilistic forecasts of different scenarios. Moreover, when using systems with multiple stages, probabilistic modelling allows us to postpone aggregation. This often leads to more accurate point forecasts.
- Bayesian inference in general, and *hierarchical modelling* specifically, is the natural framework to combine unbalanced data from different sources and at different levels of aggregation. It also delivers the probabilistic forecasts that we require[18].
- Voter choice is fundamentally not binary in the 2015 Congressional Election by contrast, for instance, with the US Presidential Election or previous Spanish Congressional Elections. Therefore, binary choice models are insufficient.
- In many countries, and in Spain in particular, there is little polling at the regional level, and next to none at the provincial level, where seats are allocated. Accordingly, we cannot rely on polls to inform us about the vote at the provincial level, where it actually matters.

[18] Stegmueller (2013) concludes that when using multilevel models the Bayesian approach is more robust and generates more conservative tests than the frequentist approach.

- The drastic change of political scenery with emergence of strong new parties renders historical models insufficient for prediction.

To forecast the territorial distribution of sentiment we use data on individual respondents in preelectoral surveys. In Spain, these are carried out by the government-sponsored research center CIS[19]. These allow us to estimate the relationship between geographical and demographic characteristics, and voters' choice. The downside of these datasets is that the sample size in some provinces is very low, leading to noisy estimates. Furthermore, their sample may be biased. We can correct for these issues by stratifying the respondents into disjoint groups and modelling those groups' response behavior. Then, we cross-reference our sample with Census data to ascertain how many people belong to each group. This approach provides a forecast for parties for which we have little electoral history. In the literature, this approach is known as post-stratification, see e.g. [21]. We refer to each of these disjoint groups as a *stratum.* Strata are defined according to a set of demographic charateristics, i.e. age, gender and education. We denote the characteristics of stratum n in the census by $\tilde{\boldsymbol{c}}_n$. For each of these strata, the census includes an elevation factor w_n which may be interpreted as the weight of the stratum. Furthermore, let $\tilde{\boldsymbol{v}}_n$ refer to that stratum's latent electoral choice. On the basis of the survey data, we may estimate a regression function $\boldsymbol{\mu}$ that maps any vector $\tilde{\boldsymbol{c}}_n$ to the vector of probabilities over electoral choices $\boldsymbol{\mu}(\tilde{\boldsymbol{c}}_n) = \mathrm{E}[\tilde{\boldsymbol{v}}_n|\tilde{\boldsymbol{c}}_n]$. We use these weights to compute our aggregate prediction for a circumscription's vote $\tilde{\boldsymbol{v}}$:

$$\mathrm{E}[\tilde{\boldsymbol{v}}] = \frac{\sum_n w_n \boldsymbol{\mu}(\tilde{\boldsymbol{c}}_n)}{\sum_n w_n} \tag{1}$$

In the full Bayesian treatment, we also obtain a predictive distribution $f(\tilde{\boldsymbol{v}})$ over outcomes, which we will use to combine the different models[20].

We restrict the regression function $\boldsymbol{\mu}$ to a multinomial logit specification:

$$\boldsymbol{\mu}(\tilde{\boldsymbol{c}}_n) = \mathrm{softmax}\left[\boldsymbol{\alpha} + \sum_l \boldsymbol{\beta}_{j_l[n]}\right] \tag{2}$$

The softmax function is a multivariate generalization of the logistic function. $\boldsymbol{\beta}_{j_l}$ is the coefficient pertaining to level j of the categorical factor l and n indexes strata. We follow the standard practice of setting all coefficients of the pivot category ("other parties") to 0 for simpler interpretation.

[19] see the Appendix.

[20] The parameter distribution is estimated with survey data and, therefore, it corrsponds to a model of respondent behavior (i.e., sentiment) which is not necessarily equal to the model for the actual voting behavior on election day.

In keeping with the multilevel regression framework, we pool each factor's levels to a common prior:

$$\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\beta}_{jl} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}[\boldsymbol{\sigma}_l]^2), \quad \boldsymbol{\sigma}_l \overset{iid}{\sim} \mathrm{half} - \mathrm{N}(0, 1) \tag{3}$$

where $\mathrm{diag}[\boldsymbol{\sigma}_l]$ is the diagonal matrix with diagonal elements corresponding to $\boldsymbol{\sigma}_l$. The priors regularize estimates where our sample size is low. For example, some provinces may contribute only with a low number of respondents to the sample, but the partial pooling will compensate by shrinking the estimate to the factor's common mean. We defaulted to standard half-normal hyperpriors after some sensitivity analysis.

4 Learning the fundamental model for the 2015 Spanish elections

To train the fundamental model sentiment model for the 2015 Spanish Congressional Election, we use the 2015 CIS preelectoral survey[21]. This results in a total sample of about 17,452 respondents. We drop all respondents that did not report their voting intention from the sample, which amounts to assuming that their voting intention data is *missing at random*. We include factors for the province, size of the municipality, gender, age, education and labor market activity of the respondent. The categories of each variables are described in Table 2.

The nominal number of parameters in these multilevel models is in the hundreds, hence we resort to Markov chain Monte Carlo methods to sample posterior and predictive distributions. We run 4 chains in *Stan* with 2000 iterations each, half of which we discard. The results of the estimation of the Bayesian multinomial logit are summarized in Figure 3.

In practice, because responses may not reflect behavior at the voting booth accurately, and because of the possibility of sentiment shifts between the survey and the election, we inflate the uncertainty about the constants $\boldsymbol{\alpha}$, reflecting uncertainty about the national vote. In practice, this corresponds to multiplying MCMC draws of $\boldsymbol{\alpha}$ by 1.5. The choice of that factor is partially motivated by computational constraints that arise when combining the fundamental model with the polls.

To interpret estimates correctly, note the following:

- Since the model is overparametrized, some parameters are weakly identified. This manifests in wide marginal distributions. These overstate

[21] As we noted in previous sections this surveys already contain questions about the two new parties.

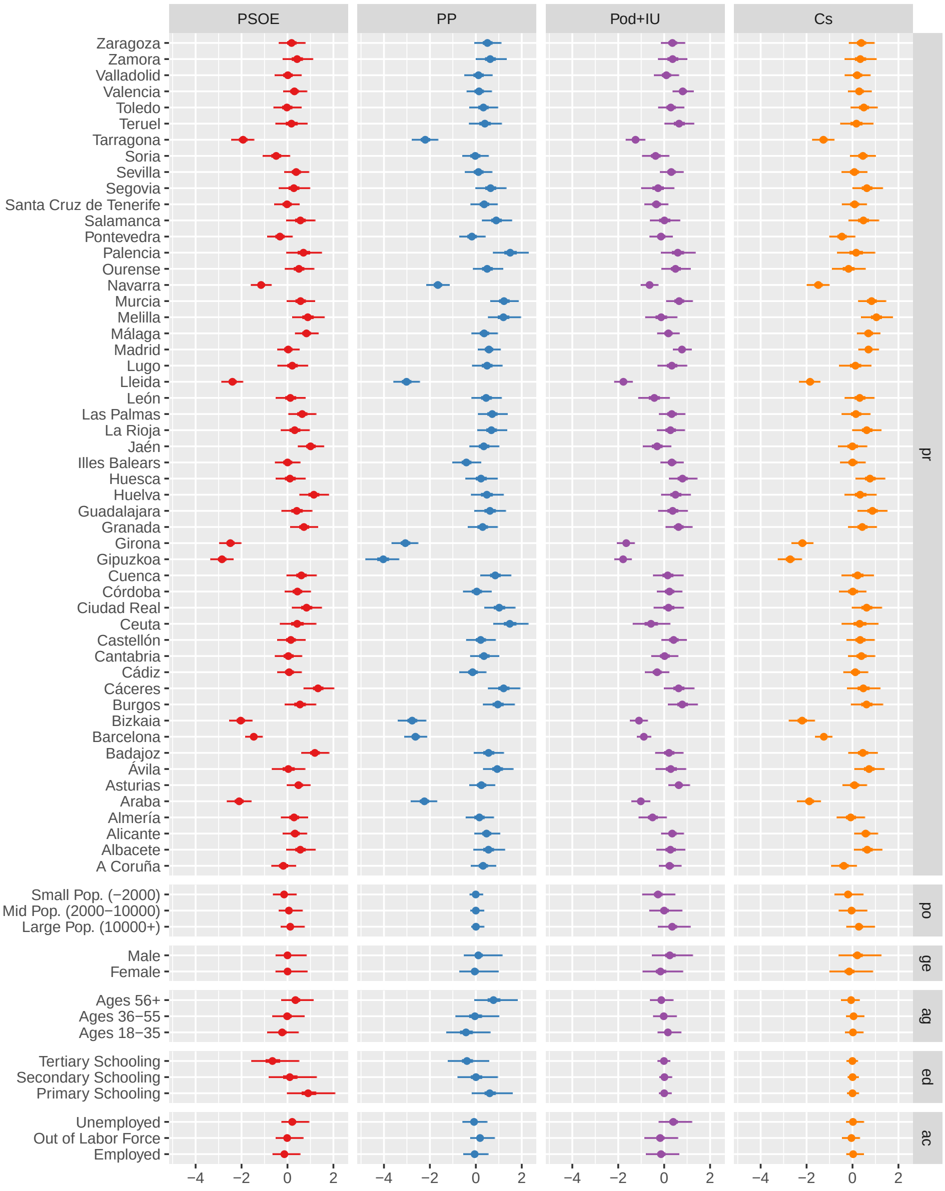

Figure 3: $\boldsymbol{\beta}_{jl}$ marginal distributions (sentiment model level coefficients): median (point), 50 percent credibility interval (thick line) and 95 percent credibility interval (thin line).

uncertainty since the parameters are highly correlated: once a parameter has been fixed, the uncertainty resolves.

- As we fix the parameters of the pivot party (essentially consisting of regional parties) to 0, all estimates must be interpreted *relative* to these parties. Therefore, a positive intercept estimate for PSOE implies that

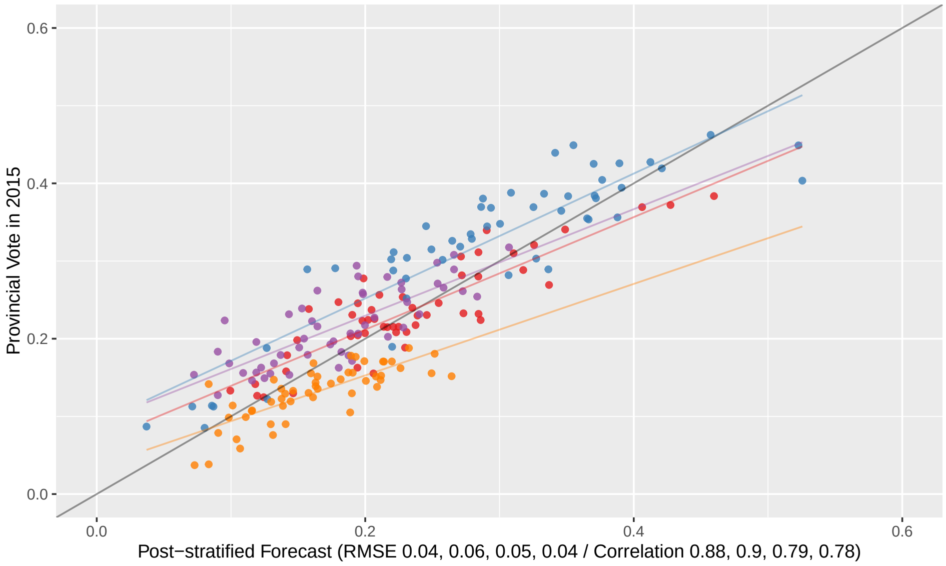

Figure 4: Scatterplot of post-stratified point estimates vs. outcomes and regression line. Statistics are listed in the usual party order. MSE is computed as the average squared difference between the mean prediction and the result over provinces. Legend: PSOE (red), PP (blue), Podemos+IU (purple), C's (orange).

the average respondent is more likely to vote for PSOE than regional parties.

The province effects in Figure 3 indicate that PP has the most variable territorial distribution, while Podemos is fairly constant. PSOE is strong in Andalucia and Extremadura and fairly weak in Catalonia and the Basque Country. PP has its strongest base in Castille and Murcia, but is extremely weak in Catalonia and the Basque Country.

As to the other factors, Podemos and C's are slightly more urban, while the other parties' support does not vary along that dimension. PSOE and PP mostly appeal to uneducated voters. Labor market activity is mostly irrelevant after controlling for other factors.

Figure 4 illustrates point predictions of the fundamental model and how they compare against the actual 2015 election results.

5 Polls model

The post-stratified estimates are reliable as far as the provincial vote *relative* to the national vote is concerned, but they they suffer from shortcomings when it comes to forecasting the national vote. The sentiment model uses survey data which are collected some months before the general elections. Thus, respondents could change their minds in the meantime, which the sentiment model would miss.

Polls, on the other hand, are published until shortly before the elections[22]. Furthermore, pollsters try to adjust their predictions for the different response biases that their polls may suffer from. By tracking and assimilating polls with the fundamental model, we can improve the accuracy of the national prediction, but at the same time that of the province results by more accurately estimating an "intercept".

The next task is aggregating polls. The simplest possibility would be just to average the latest period (one week, two weeks, one month). This local averaging might be carried out using overlapping or non-overlapping windows of time. Forecasting can then be done only under the assumption that there is not going to be a change in public opinion from that time period to the election day. This approach is followed frequently in the mass media. Figure 5 shows the effect of smoothing using a LOESS smoother for the 2015 Spanish elections.

Exploratory analysis reveals that the uncertainty about the election result close to election day by far exceeds the sampling uncertainty. Averaging multiple polls does not eliminate the excess uncertainty. Furthermore, we sometimes observe sharp trending close to election day, even after prolonged periods of stability. Following (and extrapolating) the trend usually takes us closer to the election result than simple averaging. Figure 6 shows for the 2015 Spanish elections how the declared margin of error in the polls, usually given as the inverse of the square root of the sample size, tends to underestimate the true uncertainty. Furthermore, using a linear trend brings us closer to the election result. Finally, averaging over polls does not eliminate the error.

5.1 Our methodology

Using polls comes with its own set of challenges. Rolling averages, like the ones depicted in Figure 5, do not provide a direct measure of uncertainty, which is essential to building a probabilistic model. Furthermore, conditional on the true sentiment, raw polls are *not* independent nor identically distributed for a variety of reasons:

- Polls by the same pollster may exhibit the same systematic bias across elections. For example, some pollsters are subject to political influence,

[22] in Spain a week before, but in Andorra it is allowed to publish polls regarding the Spanish elections up to a day before

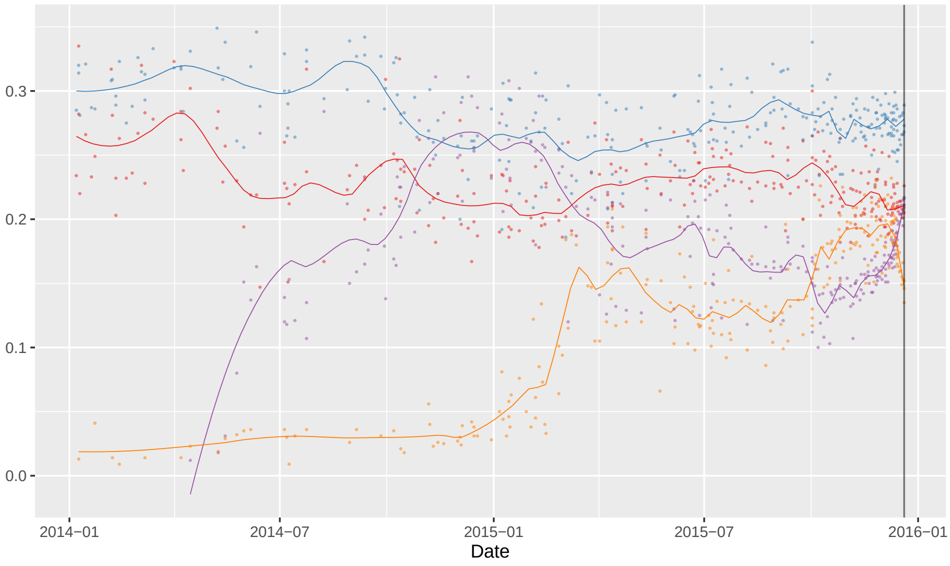

Figure 5: Polling before the general election of 2015, with LOESS smoother. Legend: Legend: PSOE (red), PP (blue), Podemos (purple), C's (orange). The election day is marked by the vertical solid line.

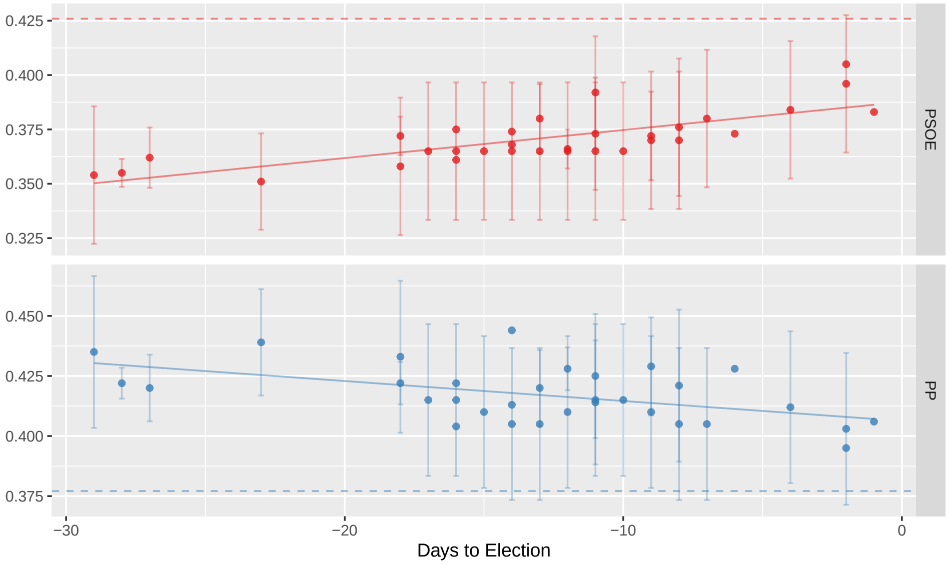

Figure 6: Polling before the general election of 2004. The solid line is a linear trend (OLS), the dashed line is the election result and the error bars correspond to the margin of error reported by the pollster.

which may lead them to systematic bias.

- Polls preceding the same election may suffer from systematic bias across pollsters. This may be due to common methodological flaws and pollsters manipulating their polls to conform with the fold.
- Some pollsters' methodology may be superior, leading to lower error variance.
- Subsequent polls may be trending up or down.

To overcome these shortcoming we develop a model to synthesize polls' results. Let $\boldsymbol{p}_k$ denote a poll's predictions, $\boldsymbol{v}_{t[k]}$ the election result corresponding to poll k. Our model decomposes the error into different sources of uncertainty:

$$(\boldsymbol{p}_k - \boldsymbol{v}_{t[k]}) \sim \mathrm{N}(\boldsymbol{\gamma}_{j[k]} + \boldsymbol{\delta}_{t[k]} + d_k \boldsymbol{\epsilon}_{t[k]}, \boldsymbol{\Sigma}_{j[k]}) \tag{4}$$

where $\boldsymbol{\gamma}_j$ is the time-invariant bias of pollster j, $\boldsymbol{\delta}_t$ is the pollster-invariant bias in election t, d_k corresponds to how many days before the election poll k was published and $\boldsymbol{\epsilon}_{t[k]}$ is the pollster-invariant strength of the trend in a given election. $\boldsymbol{\epsilon}_{t[k]}$ decays as election day approaches, but $\boldsymbol{\delta}_t$ applies to all polls until the election[23].

We allow the covariance matrix of the residual distribution, $\boldsymbol{\Sigma}_{j(k)}$, to vary by pollster. As in the fundamental model, we pool factor levels to a common prior:

$$\boldsymbol{\gamma}_j \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma), \quad \boldsymbol{\delta}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\delta), \quad \boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon) \tag{5}$$

6 Learning the pollster model for the 2015 Spanish elections

We use the pollster model we have described in the 2015 Spanish elections. We work with the results of 157 electoral polls published before the Congressional Elections of 1996, 2000, 2004, 2008 and 2011. This set corresponds to the subset of polls published up to 30 days before a Congressional Election. Hyperpriors are set in accordance with Stan reference priors. We sample from the models using *Stan*, running 4 chains with 2000 iterations each, half of which we discard. Figures 7, 8 and 9 depict the marginal distribution of pollster bias, election bias and election trend respectively.

Estimated pollster biases $\boldsymbol{\beta}$ are generally consistent with political expedience. For example, the pollster *Sigma dos*, which mostly provides polls for

[23] Litzer (2013) builds a dynamic Bayesian forecasting model using a reverse random walk prior that shrinks towards the prior distribution as it moves backwards from the Election Day. See Lock (2010) for evidence of trending close to election day.

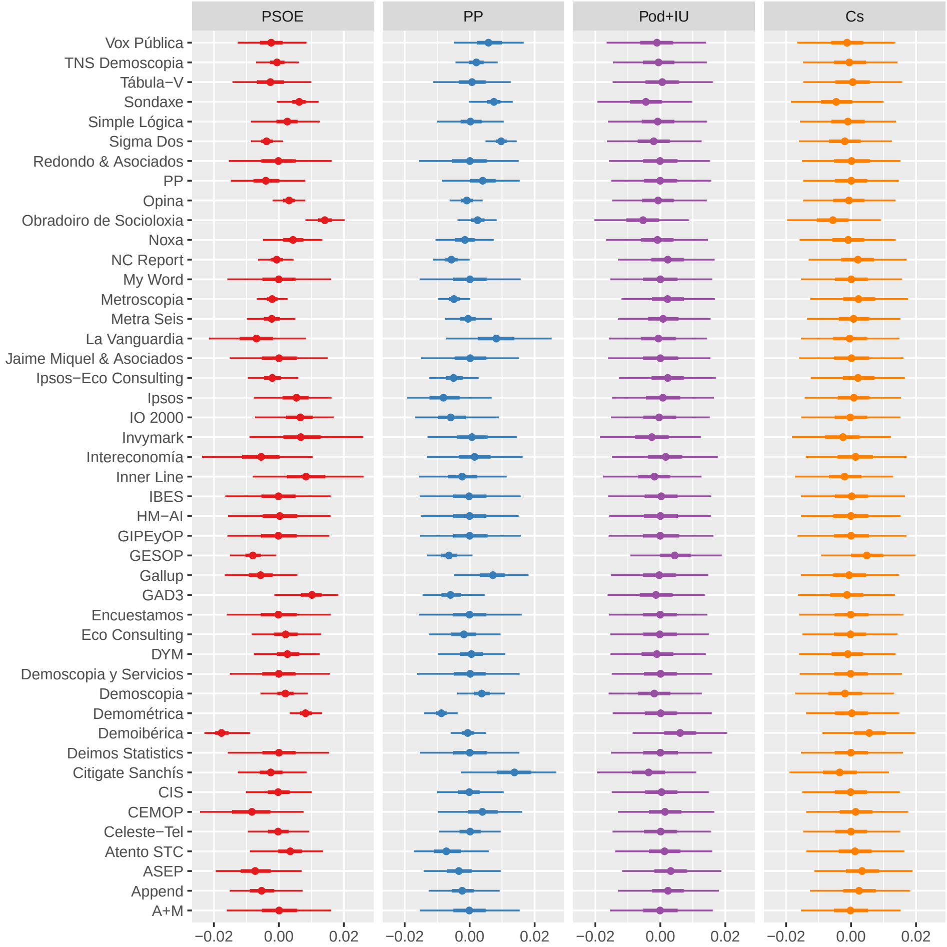

Figure 7: γ_j marginal distributions (pollster bias): median (point), 50 percent credibility interval (thick line) and 95 percent credibility interval (thin line). Positive values imply that the pollster is overestimating.

the right-leaning newpaper *El Mundo*, has a consistent bias in favor of the Popular Party. Election biases $\boldsymbol{\delta}_t$ are large, with pollsters collectively missing the PSOE-PP differential by 7 percentage points, calling into question the quality of Spanish polling and the predictability of Spanish elections in general. Estimated trend effects $\boldsymbol{\epsilon}_t$ are large in many elections, which confirms that some trend adjustment is necessary even within the last 30 days. Finally, election biases seem to coincide in sign and magnitude with trends, especially in the 2004 elections, but we deem our sample to be too small to draw further conclusions.

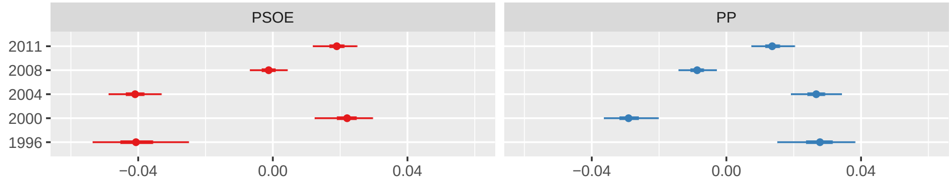

Figure 8: $\boldsymbol{\delta}_t$ marginal distributions (election bias): median (point), 50 percent credibility interval (thick line) and 95 percent credibility interval (thin line). Positive values imply that the pollster is overestimating.

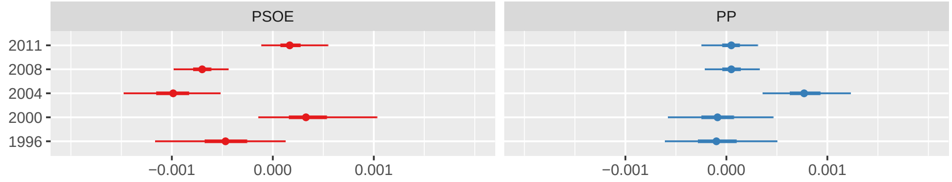

Figure 9: $\boldsymbol{\epsilon}_t$ marginal distributions (election trend): median (point), 50 percent credibility interval (thick line) and 95 percent credibility interval (thin line). Positive values imply that polls are trending down.

7 Evidence Synthesis

To generate our forecast, we need a method that combines and weights the output of the fundamental model (which in Bayesian terms will be the *prior*) and the polls model (*likelihood*). The prior will be denoted by $f(\tilde{\boldsymbol{v}})$, whereas the likelihood is denoted by $f(\tilde{\mathbf{P}}|\tilde{\boldsymbol{v}})$. In probabilistic terms, $f(\tilde{\mathbf{P}}|\tilde{\boldsymbol{v}})$ is the likelihood of observing a new set of polls $\tilde{\mathbf{P}}$ given an election result $\tilde{\boldsymbol{v}}$. The likelihood only operates at the national level, so this is the level the evidence synthesis has to operate. On the other hand, the prior $f(\tilde{\boldsymbol{v}})$ is obtained by aggregating province level forecasts. We obtain the posterior predictive distribution by applying Bayes' law:

$$f(\tilde{\boldsymbol{v}}|\tilde{\mathbf{P}}) \propto f(\tilde{\mathbf{P}}|\tilde{\boldsymbol{v}})\ f(\tilde{\boldsymbol{v}}) \tag{6}$$

Computationally, we perform the Bayesian evidence synthesis by using our MCMC samples from prior and likelihood through *importance sampling*[24]. We use the pollster model to produce estimates of the likelihood $f(\tilde{\mathbf{P}}|\tilde{\boldsymbol{v}})$, hence provide weights to samples generated according to the prior $f(\tilde{\boldsymbol{v}})$, i.e., the fundamental model. Finally, prior samples are redrawn with probabilities equal to the re-normalised weights. Notice that $\tilde{\boldsymbol{v}}$ is implied by our predictions at the province level. Therefore, resampling according to the national vote also updates our forecast at the province level, which is our ultimate goal.

[24] See Chapter 2 of [17].

8 Synthesised predictions for the 2015 Spanish elections

Figure 10 shows how the synthesis operates in the 2015 election. Incorporating polls strongly improves the PP and Podemos prediction, while the C's prediction becomes slightly worse. This is due to pollsters systematically overestimating C's until election day.

Figure 11 shows the predictive seat distribution for the largest five parties. The result is close to the predictive mode for PSOE, PP and Podemos. The result of C's is in the left tail of the predictive distribution. Figure 12 shows point predictions versus actual results and it is directly comparable to Figure 4.

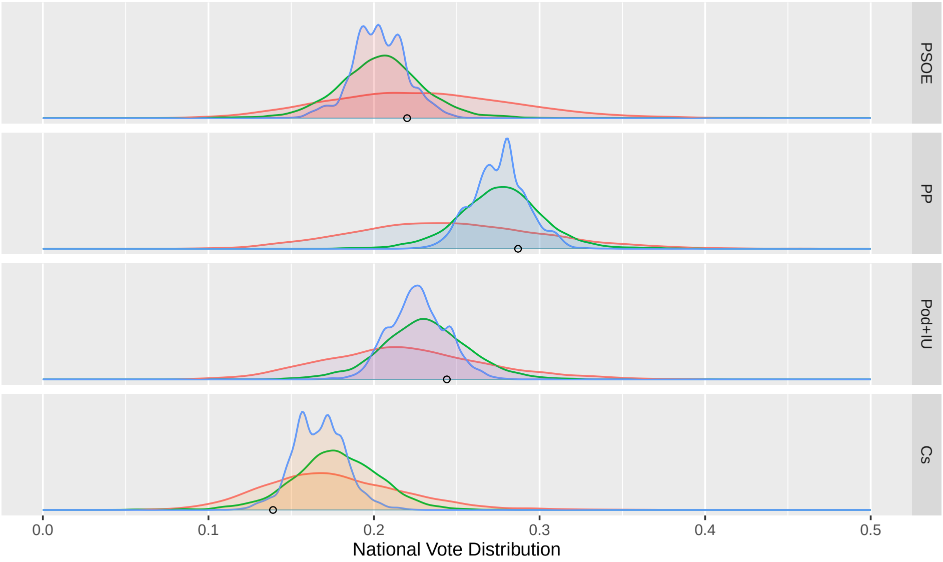

Figure 10: Predictive national vote distribution: fundamental model (red), polls model (green), synthesis (blue). The dots represent the election result.

9 Conclusions

This paper has proposed a methodology to forecast electoral outcomes using the result of the combination of a fundamental model and a model-based aggregation of polls. We propose a Bayesian hierarchical structure for the fundamental model that synthesizes data at the provincial, regional and national level. We use a Bayesian strategy to combine the fundamental model

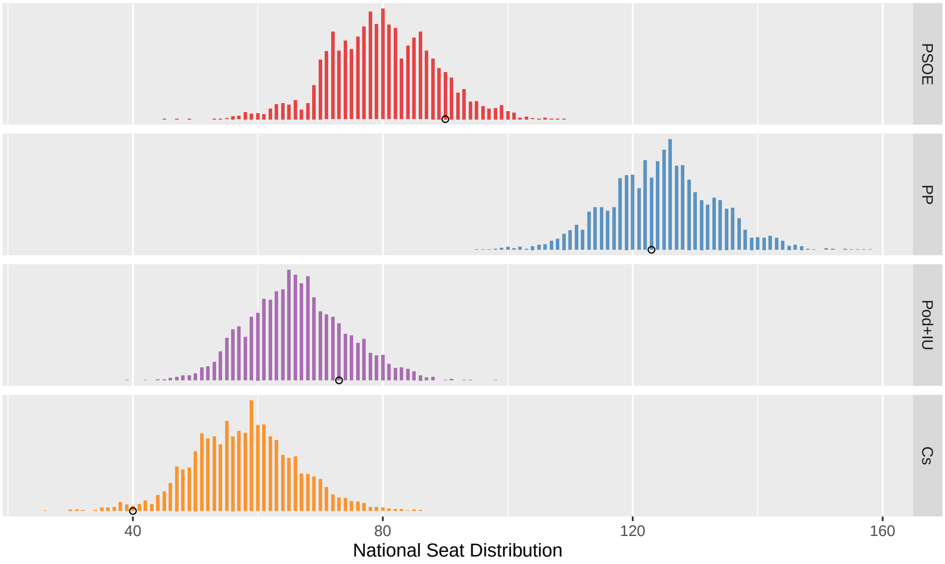

Figure 11: Predictive seat distribution and election result (black dot).

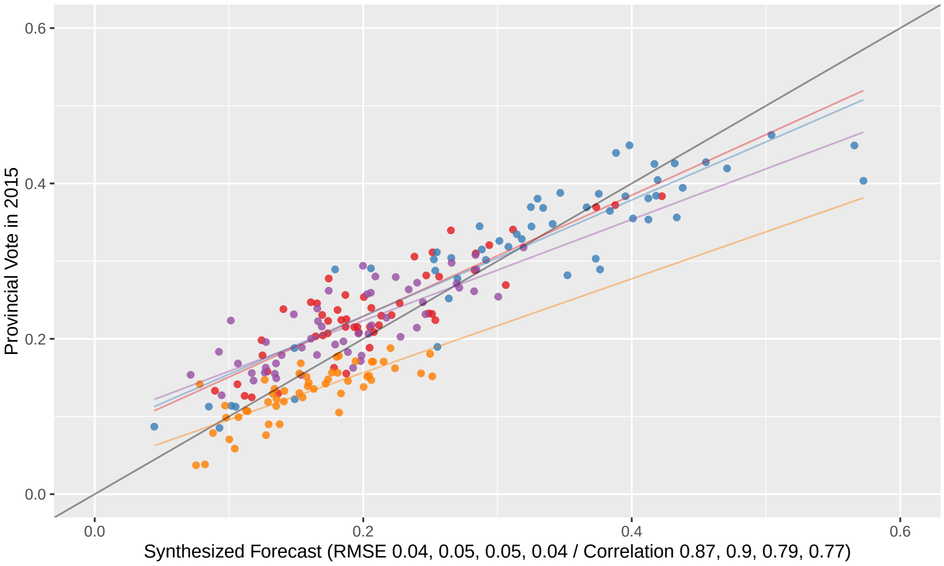

Figure 12: Scatterplot of point predictions vs. outcomes and regression line. Statistics are listed in the usual party order. MSE is computed as the average squared difference between the mean prediction and the result over provinces. Legend: Legend: PSOE (red), PP (blue), Podemos+IU (purple), C's (orange).

with the information coming for recent polls. This model can naturally be updated everytime new information, for instance a new poll, becomes available. This methodology is well suited to deal with increasingly frequent situations in which new political parties enter an electoral competition, although our approach is general enough to accomodate any other electoral situation. We illustrate the advantages of our method using the 2015 Spanish Congressional Election in which two new parties ended up receiving 30% of the votes.

A Input Data

For the survey model, we use the 2015 CIS preelectorals (CIS study number 3117). The study is openly available on `http://www.cis.es/` and includes 17452 respondents. Data was collected from October 27th to November 16th, 2015.

To train the polls model, we use 157 polls published within 30 days of the 1996, 2000, 2004, 2008 and 2011 Congressional Elections. You may consult these polls in the respective Wikipedia articles:

- `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_1996`
- `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_2000`
- `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_2004`
- `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_2008`
- `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_2011`

Furthermore, to generate predictions, we use 51 polls published within 30 days of the 2015 Congressional election. You may consult these polls on `https://en.wikipedia.org/wiki/Opinion_polling_for_the_Spanish_general_election,_2015`.

B Hierarchical modelling notation

Hierarchical modelling notation is a convienient way of describing models that include a lot of categorical variables as regressors. Our hierarchical modelling notation follows the standard set by Gelman and Hill in *Data Analysis using Regression.*

Consider this brief explanation of the notation. Let $\{1, \ldots, I\}$ index a set of observations and $\{1, \ldots, J\}$ be the indices of the levels of a categorical factor. Then, the notation $j[i]$ refers to a map $\{1, \ldots, I\} \mapsto \{1, \ldots, J\}$ which links each observation to its respective factor level. For instance, if the factor is gender, male has index 1, female index 2 and observation 1 is female, then $j[1] = 2$.

If $\boldsymbol{\beta}$ is the vector of coefficients pertaining to the levels of some factor, we can use hierarchical modelling notation to retrieve components of that vector. In keeping with our example, $\beta_{j[1]} = \beta_2$ is the coefficient of the gender of observation 1, which is equivalently the coefficient of the female level of the gender factor.

We may express this equivalently using dummy variables, but hierarchical modelling notation tends to be more concise. For example, consider a simple regression model with one categorical factor. In dummy notation, we write $y_i = \beta_0 + \sum_j \beta_j x_{ij} + \epsilon_i$. In hierarchical modelling notation, we just write $y_i = \beta_0 + \beta_{j[i]} + \epsilon_i$.

C Survey model factors

Table 2 provides the full list of categorical factors included in the survey model.

Factor	Code	Levels
Voting Intention	1	PSOE
	2	PP
	3	Podemos, En Comú Podem, En Marea, IU
	4	Ciudadanos
	5	Others
Province	1-52	INE Province Code
Municipality Population	1	less than 2000 inh.
	2	between 2000 and 10000 inh.
	3	more than 10000 inh.
Gender	1	Male
	2	Female
Age	1	18 to 35 y.o.
	2	36 to 55 y.o.
	3	more than 56 y.o.
Education	1	Primary or less
	2	Secondary
	3	Tertiary
Activity	1	Employed
	2	Unemployed
	3	Out of the labor force

Table 2: Factors and their categories. These categorical features define 8424 distinct strata, or 162 distinct strata per province.

References

[1] Campbell, J. (2008), ”Evaluating U.S. Presidential Election Forecast and Forecasting Equations”, *International Journal of Forecasting*, 24, 259-271.

[2] Campbell, J. (1992), ”Forecasting the Presidential Vote in the States”, *American Journal of Political Science*, 36, 386-407.

[3] Carpenter, Bob, et al. (2016), "Stan: A Probabilistic Programming Language." *Journal of Statistical Software.*

[4] Colomer, J. (2007), "What other sciences look like", *European Journal of Political Science*, 6, 134-142.

[5] Curtice, J. and Firth, D. (2008), "Exit polling in a cold climate: the BBC–ITV experience in Britain in 2005", *J. R. Statist. Soc. A*, 171, 509–539.

[6] Dennison, S. and D. Pardijs (2016), "The world accourding to Europe's insurgent parties", European Council on Foreign Relations ECFR-WP-181.

[7] Fair, R. (2009), "Presidential and Congressional Vote-Share Equations", *American Journal of Political Science*, 55-72.

[8] Fraile, M. and M. S. Lewis-Beck (2010), "Economic voting in Spain: a 2000 panel test", *Electoral Studies*, 29 (2), 210-220.

[9] Gayo-Avello, D. (2012), "No, You cannot predict elections with Twitter", *Internet Computing, IEEE*, 16(6),91–94.

[10] Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press

[11] Gelman, A. and G. King (1993), "Why are American Presidential Election Campaign Polls So Variable When Votes Ares So Predictable", *British Journal of Political Science*, 23, 409-451.

[12] Graefe, A. Armstrong, J.S., Jones, R. and A. Cuzan (2014), "Combining forecasts: an application to elections," *International Journal of Forecasting*, 30 (1), 43-54.

[13] Hibbs, D. (2008), "Implications of the bread and peacemodel for the 2008 US presidential election," *Public Choice*, 137, 1-10.

[14] Hummel, P. and D. Rothschild (2014), "Fundamental Models for Forecasting Elections at the State Level", *Electoral Studies*, 35, 123-139.

[15] Lewis-Beck, M. (2005), "Election Forecasting: Principles and Practice," *British Journal of Politics and International Relations*, 7, 145-164.

[16] Linzer, D. (2013), "Dynamic Bayesian Forecasting of Presidential Elections in the States", *Journal of the American Economic Association*, 108 (501), 124-134.

[17] Liu, J.S. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer.

[18] Lock, K. and A. Gelman (2010), "Bayesian Combination of State Polls and Elections Forecasts", *Political Analysis*, 1-12.

[19] Montalvo, J. G. (2012), "Re-examining the evidence on the electoral impact of terrorist attacks: the Spanish election of 2004" *Electoral Studies*, 31, 96-106.

[20] Murthy, D. (2015), "Twitter and elections: are tweets preditive, reactive or a form of buzz?" *Information, Communication and Society*, 18 (7), 816-831.

[21] Park, D.K., Gelman, A. and J. Bafumi (2010), "Bayesian Multilevel Estimation with Poststratification: State Level Estimates from National Polls," *Political Analysis*, 12, 375-385.

[22] Smidt, C.D. (2015), "Polarization and the Decline of the American Floating Voter," *American Journal of Political Science*, 1-17.

[23] Stegmueller, D. (2013), "How Many Countries for Multilevel Modelling? A Comparison of Frequentist and Bayesian Approaches," *American Journal of Political Science*, 57 (3), 748-761.

[24] Udina, F. and P. Delicado (2005), "Estimating Parliamentary Composition Through Electoral Polls", *Journal of the Royal Statistical Society*, 168 (2), 387-399.

[25] Walsh, E., Dofin, S. and J. DiNardo (2009), "Lies, Damn Lies, and Pre-Election Polling", *American Economic Review*, 99 (2), 2-13.