PREELECTORAL POLLS VARIABILITY: A HIERARCHICAL BAYESIAN MODEL TO ASSESS THE ROLE OF HOUSE EFFECTS WITH APPLICATION TO ITALIAN ELECTIONS

BY DOMENICO DE STEFANO[1,a], FRANCESCO PAULI[2,b] AND NICOLA TORELLI[2,c]

[1]*Department of Political and Social Science, University of Trieste,* [a]*ddestefano@units.it*
[2]*Department of Business, Economics, Mathematics, and Statistics, University of Trieste,* [b]*francesco.pauli@deams.units.it,* [c]*nicola.torelli@deams.units.it*

It is widely known that preelectoral polls often suffer from nonsampling errors that pollsters try to compensate for in final estimates by means of diverse ad hoc adjustments, thus leading to well-known house effects. We propose a Bayesian hierarchical model to investigate the role of house effects on the total variability of predictions. To illustrate the model, data from preelectoral polls in Italy in 2006, 2008 and 2013 are considered. Unlike alternative techniques or models, our proposal leads: (i) to correctly decompose the different sources of variability; (ii) to recognize the role of house effects; (iii) to evaluate its dynamics, showing that variability of house effects across pollsters diminishes as the date of election approaches; (iv) to investigate the relationship between house effects and overall prediction errors.

1. Introduction. The number of publicly released preelection polls has grown dramatically over the years in all modern countries. Preelection polls have long played an important role in the conduct and study of elections, especially in U.S. presidential elections. They are essentially used for three different purposes: forecasting election outcomes, understanding voter behavior and planning political campaign strategy (Hillygus (2011)).

It is widely known that preelectoral polls suffer from nonsampling errors which may differ between pollsters (Worcester (1996)). The use of different methodologies and remedies to deal with nonsampling errors result in the so-called house effects (HEs) (Wlezien and Erikson (2007)) which are biases associated with each pollster. Due to the presence of HEs, the variability of the estimates from a group of pollsters is higher than what sampling variability would imply (and higher than the variability around trend of the estimates of a single pollster). We focus on investigating the characteristics of HEs: their role in determining the variability of polls and their dynamics. For instance, one can conjecture that the variability of HEs can decline as the election day approaches, as it has been noted in the U.S. elections by Linzer (2012), Moore (2008) and in AAPOR (2017), among others.

In order to investigate the characteristics and the role of HEs, we specify a Bayesian hierarchical model (described in Section 4), which allows for the variability of polls quantifying the contribution of HEs to the total variability using a decomposition of the variation in the data, similar to the classical ANOVA, proposed by Gelman (2005). We apply this model in the Italian general election case using data gathered from preelection polls carried out in 2006, 2008 and 2013 by various pollsters (Section 3).

Bayesian hierarchical models have been already used to pool the results of multiple polls in order to obtain improved predictions (Silver (2010)). They have also been used to explicitly allow for HEs (as a group effect) by Jackman (2005), who modelled polls for the Australian federal election of 2004, and by Hanretty (2013), who modelled polls for the 2013 Italian general election. Linzer (2013) also adopts a similar approach in order to dynamically combine vote shares for presidential elections across U.S. states.

Key words and phrases. Finite population variance, posterior predictive check, ANOVA, preelectoral polls.

All these models need to be reformulated and expanded if the aim is evaluating the role of some sources of inaccuracies of polls. The proposed model that is able to directly measure and analyze (with respect to its size, dynamics and relationships with other errors) the role of HEs. With respect to previous models, in our specification we are able to avoid unnecessary and partly unjustified model assumptions: in particular, we do not need to consider electoral results (true vote share on election day) to estimate HEs size. In our opinion that choice requires unrealistic assumptions on the pattern of variation of the electoral preferences in time.

Using the Italian elections as case study, unlike the other approaches our model leads to correctly decompose the different sources of variability. In particular, we are able to evaluate the house effect magnitude and its dynamics.

It is worth noting that, unlike other approaches, our goal is not to pool pollster results to estimate party's vote shares, rather, we are interested in properly model variance components, with special attention to the HE behavior.

2. House effects in preelectoral polls. Polls are severely affected by nonsampling errors. While a complete list of the sources of bias would be out of the scope of this paper (see Worcester (1996)), it is worth noting that some of the biases are likely to affect all pollsters equally, while others are likely to be pollster-dependent (see Sturgis et al. (2016) and AAPOR (2017) for an extensive discussion of sources of errors affecting polls in the U.K. and U.S. general elections, respectively). Consider, for example, some of the main sources of nonsampling error broadly ordered by degree of specificity: nonresponse bias, an imperfect frame, questionnaire design issues and weighting procedures. Item and unit nonresponses, typically due to the tendency not to disclose less socially acceptable votes or inconsistent behaviour (i.e., changing votes), are quite severe and will generally affect all pollsters equally. Differences in pollster results may be attributable to modes in which questions are administered (telephone, Internet or face-to-face). The imperfect frame issue, which can be related to the modes of interview administration, can affect all pollsters but to different extents; some may use mobile phone numbers, whereas others may limit sample selection to fixed phone lists. In fact, pollsters do estimate different parameters, as they, de facto, refer to different frame populations. Questionnaire design issues are mainly related to the question wording/ordering and to the presence of filter questions on whether the interviewee plans to actually go to vote and are essentially pollster-specific.

On the other hand, even those nonsampling errors that affect all pollsters equally (such as nonresponse) are susceptible to imply different biases in their final estimates, inasmuch as pollsters make different adjustments in order to compensate for them. In general, in fact, there is no universal method or gold standard to which to refer; on the contrary, the nature and the details of the adjustments, which in most cases amount to using a weighting strategy while others are "ad hoc," depend on the expertise of each pollster and are specific to them, who, in most cases, do not reveal the details. A good example is the use of weighting strategies to try to compensate biases arising from imperfect frames and uncontrolled nonresponses. A common solution is to weigh the results, taking into account past votes or other politically related preferences of the respondent, asked during the interview. Each pollster has its own methods for collecting those data and for producing the weights (it is also worth noting that an additional error might be introduced due to the risk that the past vote is incorrectly reported, especially when the last election was held long before). Moreover, it is very common to use fine poststratification weights with criteria that differ across pollsters and to use population totals that do not represent the actual population counts. All these reasons lead to weighting schemes that are strongly dependent on the pollster.

Overall, it is to be expected that different errors and different corrections to the same errors interact to produce survey estimates that, depending on the pollster, are systematically more or less favourable to particular parties (the HEs).

In principle, HEs are systematic deviations with respect to the true vote share. In practice, they have been usually estimated based on the deviations with respect to an overall estimate, which is obtained by combining (averaging) multiple polls, which is not necessarily an unbiased estimate of the true vote share (for an insightful review of different pooling strategies and their respective trade-offs, see Pasek (2015)). For instance, Erikson and Wlezien (1999) and Wlezien and Erikson (2007) modeled poll results using regression analysis with dummy variables for pollsters to obtain a pooled estimate, keeping into account that single polls are affected by pollster bias. The coefficients of the house dummy variables are a measure of the magnitude of each pollster bias with respect to the reference pollster. A similar method is employed by Panagopoulos (2009) and Silver (2010).

Other authors (Jackman (2005), Pickup and Johnston (2007), Pickup and Johnston (2008)) take a different approach and use electoral results (true vote share on election day) to estimate the true vote share on polls days; the HEs are then estimated based on the deviation of polls results with respect to the estimated true vote share in a manner which is broadly similar to the regression model described above. (It is worth to note that this latter approach, although more appealing because it refers to the true vote share, entails making strong assumptions on the pattern of variation of the electoral preferences in time.)

Despite all these papers acknowledge the role of house effect and estimate them, at least as a nuisance parameter when the final aim is to forecast the election results, none of them give an explicit measure of the contribution of HEs as a source of variation on the overall vote prediction. We deem this a relevant aspect, and we propose a model capable of assessing the role of HEs in determining the variability of polls results and to also to compare such role across parties and time.

A precise quantification of HEs, even within the limits of how precise a quantification can be, may offer new insights on the phenomenon. As outlined in Section 5.2, it allows us to investigate their dynamics, in particular, whether they get smaller as election day approaches. Moreover, we can relate the magnitude of HEs in a particular election with the overall prediction error of that election results (Section 5).

3. Data. We consider vote shares of preelection polls for the Italian general elections of 2006, 2008 and 2013. Preelection polls play a significant role in Italian parliamentary elections. The number of published polls has consistently increased in the last decade. For example, for the 2001 parliamentary elections, 55 polls were published in the month of April. This number almost doubled for the 2006 elections to 104 polls in the month of May and tripled for the 2008 elections to 151 polls in the month of April (Gasperoni and Callegaro (2008)).

Although explaining the Italian political system and the histories of parties is beyond the scope of the present paper and beyond the expertise of the authors, it is worth pointing out some general features before describing the data in detail. Vote shares are referred to the "Camera dei Deputati" (low chamber). For the other branch of parliament, no national polls are held because elections are made on a regional basis. We did not go back in time before 2006 because of changes in the electoral system that would make any comparison with previous elections extremely dubious.

The parties that stood for the elections of 2006, 2008 and 2013 are listed in Table 1, where two things are worth noting. First, a relatively high number of parties are excluded (and listed as "Others"), as because of their size they are ignored in most, if not all, national polls. For each year we considered the subset of parties that was allowed for by all pollsters (eight in

TABLE 1

List of parties for which poll results were available in the three elections. The subdivision in coalitions is based on official allegiances (note that some parties change coalitions across years); actual vote share attained at the election is reported for each year; minor parties are grouped distinguished by coalition, and the numbers of minor parties are reported in brackets (Source: Ministero dell'Interno, Ufficio IV—Servizi Informatici Elettorali (2006), Ministero dell'Interno, Ufficio IV—Servizi Informatici Elettorali (2008), Ministero dell'Interno, Ufficio IV—Servizi Informatici Elettorali (2013))

		2013	2008	2006
Left	Partito Democratico (PD)	25.43	33.18	–
	L'Ulivo (Ul)	–	–	31.27
	Sinistra Ecologia Libertà (SEL)	3.2	–	–
	Rifondazione Comunista (RifCom)	–	–	5.84
	Di Pietro Italia Dei Valori (DiP/IdV)	–	4.37	2.3
	Other	0.92 (2)	–	10.4 (10)
Right	Il Popolo Della Libertà (PdL)	21.56	37.38	–
	Forza Italia (FI)	–	–	23.72
	Alleanza Nazionale (AN)	–	–	12.34
	Unione Di Centro (UC)	–	–	6.76
	Lega Nord (LN)	4.09	8.3	4.58
	Other	3.53 (7)	1.13 (1)	2.33 (8)
Center	Scelta Civica Con Monti Per L'Italia (SC)	8.3	–	–
	Unione Di Centro (UC)	1.79	5.62	–
	Other	0.47 (1)	–	–
Not aligned	Movimento 5 Stelle Beppegrillo.It (M5S)	25.56	–	–
	Rivoluzione Civile (RC)	2.25	–	–
	La Sinistra L'Arcobaleno (SA)	–	3.08	–
	La Destra – Fiamma Tricolore (DF)	–	2.43	–
	Other	2.9 (29)	5.64 (22)	13.19 (12)

2013, seven in 2008 and 2006). This implies that the data are not completely compositional: the vote shares from a poll may not add up to 1, due to the fact that we ignore the share of minor parties. Second, the lists for the three voting rounds are quite different. New political entities arose in 2013 (such as "Movimento 5 stelle Beppegrillo.it", "Scelta Civica Con Monti Per L'Italia") and some other parties underwent less substantial changes. For instance, this is the case for "L'Ulivo" that can be loosely identified with "Partito Democratico;" "La Sinistra L'Arcobaleno" that became "Sinistra Ecologia Libertà." Finally, "Forza Italia" and "Alleanza Nazionale" merged into "Il Popolo Della Libertà."

The characteristics of electoral polls carried out in Italy by different pollsters—each pollster adopting the same survey methods over the course of the electoral campaign—share many common features according to what is described in the notes accompanying each poll the results of which are published. Pollsters usually claim that a stratified sample design is adopted, while a by far less common alternative is to use a panel design. Actually, a version of a quota sampling scheme is adopted instead of a proper probability sample design. It is not clear which sampling frame is used, but CATI is still the most common mode of interview, so the sampling units are selected from lists of telephone numbers or, less commonly, by random digit dialing. The use of CAWI has become more frequent recently, and in some cases a mixed CATI and CAWI mode is used.

For 2013, we observed the vote shares of preelection polls for the eight main parties (see Table 1) from January 5 to February 23 (the election was held February 24), provided by 14 pollsters; 89 observations are available. Data up to February 4 were obtained from the governmental site where all polls that are published or broadcast for the general public must

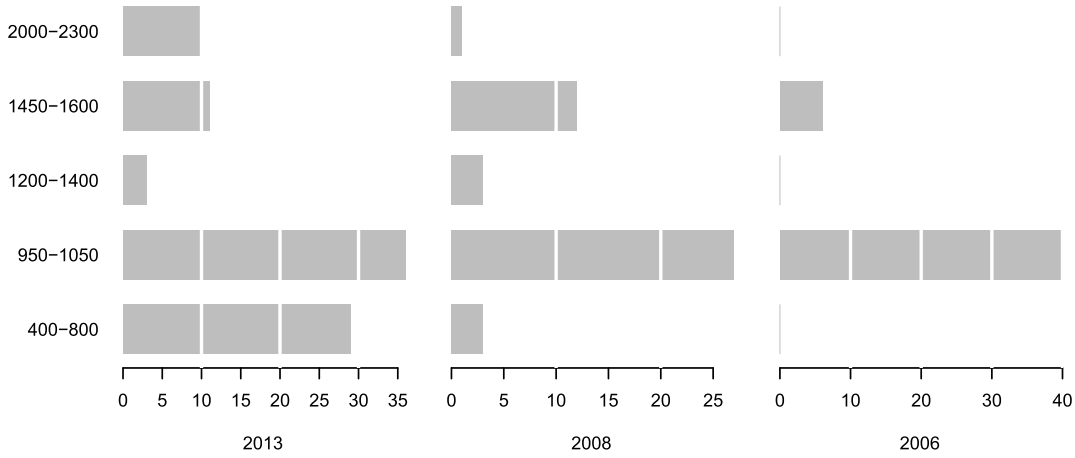


FIG. 1. *Sample sizes of polls for each election* (*absolute frequencies by classes of sample sizes*).

be communicated (Presidenza del Consiglio dei Ministri—Dipartimento per l'Informazione e l'Editoria (2015)).

The remaining 15 most recent polls were obtained from informal sources, as in this period the release of polls results to the general public is forbidden by law; therefore, these data may be seen as less reliable. The sample size ranges from a minimum of 600 up to 2500, with the large majority employing samples of about 1000 units (see Figure 1). The frequency with which the different pollsters carry out polls is quite variable, ranging from 19 to only one (see Figure 2). It is worth noting that this implies that the shares of the eight parties do not add up to one. In fact, the total share of the remaining 38 parties was 7.82% in actual election results. The total of the shares attributed to the parties in the polls goes from 84 to 96.5.

We depict poll results for all parties by pollster in Figure 3. One can notice systematic differences between some of the pollsters.

As far as 2006 and 2008 are concerned, the data structure is the same (see Table 1 for the parties involved and Figure 2 for the frequency with which each pollster carries out polls), except from the fact that all polls are obtained from the official governmental site (so, no polls for the 15 days preceding the elections are available). For 2008, seven parties and eight pollsters are considered, for a total of 46 polls held between February 11 and March 25 (the election was held on April 13). Sample size ranges from 400 to 2000; however, most polls (23) used 1000, and 10 used 1500. For 2006, we consider 46 observations for seven parties and six pollsters, the polls were carried out from January 5 to March 22 (the election was held on April 9), sample size in 2006 was almost invariably 1000. Similar to what happens for 2013, the shares of the parties do not add up to one (see Table 1).

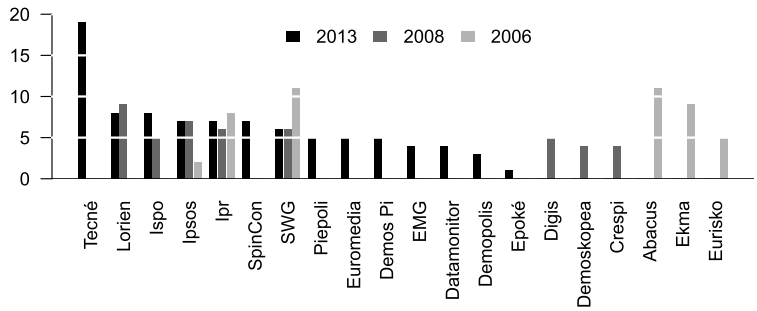


FIG. 2. *Distribution of polls among pollsters across years* (*absolute frequencies*).

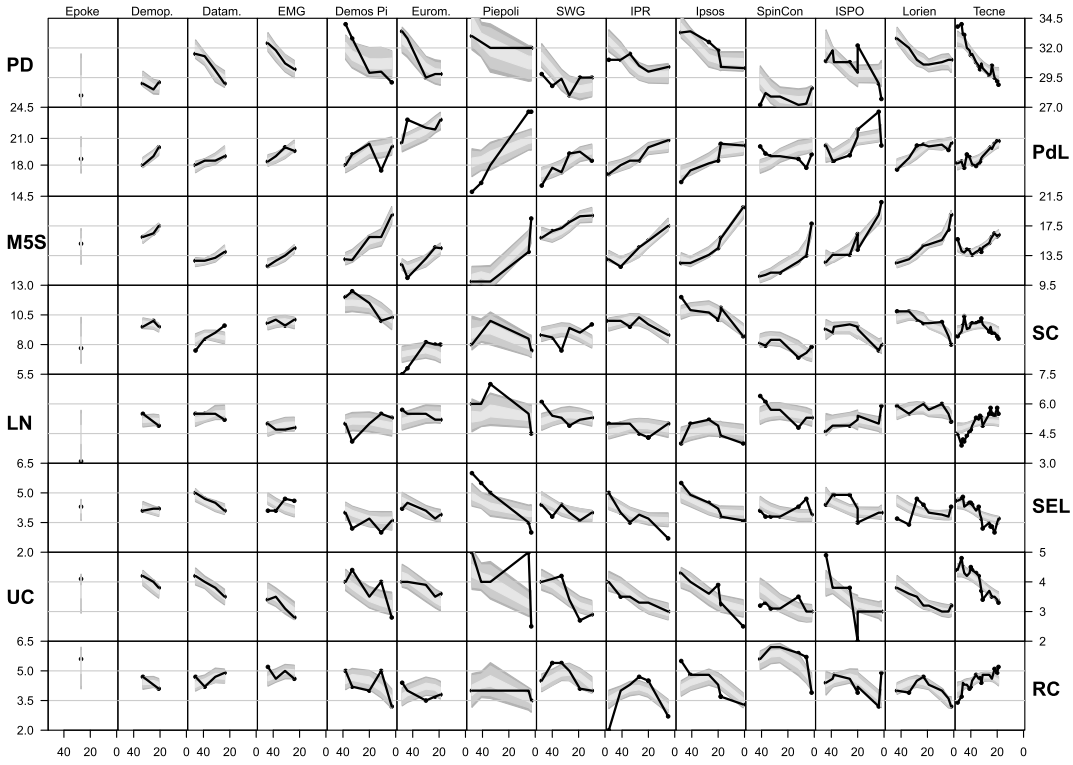


FIG. 3. *Poll results in* 2013 *for each party* (*row*, *note that y-axis*, *showing predicted vote shares*, *differ between parties*, *abbreviations are expanded in Table* 1) *and pollster* (*column*, *ordered by increasing number of polls*) *with reference bands for posterior distributions of trend plus house effect. On the x-axis we report the days to election.*

4. The model. Let y_{tsp} be the vote share estimated in the survey made on day t by house s for party p, and let n_{ts} be the number of respondents.

Then, let

$$y_{tsp} = \pi_{tp} + b_{tsp}(m_{sp} + \varepsilon_{tsp}), \tag{4.1}$$

where π_{tp} represents the true proportion of voters for party p on day t plus an unknown bias, common to all pollsters, due to some of the factors specified in Section 2, m_{sp} represents the HE of pollster s for party p, ε_{tsp} is the random variation within a pollster and the coefficient b_{tsp} is introduced to allow for heteroscedasticity.

In order to ensure that π_{tp} is in the [0, 1] interval, we consider the reparametrization $\pi_{tp} = \text{logit}^{-1}(\nu_{tp})$ and specify a random walk on ν_{tp} (this is kind of a discrete version of a spline Gaetan and Grigoletto (2004) and, as it has been shown by Rue and Held (2005), is equivalent to a Gaussian Markov random field). Let then

$$\nu_{1p}|\nu_{-1,p}, \zeta \sim \mathcal{N}(\nu_{2,p}, \zeta^2), \tag{4.2}$$

$$\nu_{tp}|\nu_{-t,p}\zeta \sim \mathcal{N}\left(\frac{1}{2}(\nu_{t-1,p} + \nu_{t+1,p}), \frac{\zeta^2}{2}\right), \quad t = 2, \ldots, T-1, \tag{4.3}$$

$$\nu_{Tp}|\nu_{-T,p}\zeta \sim \mathcal{N}(\nu_{T-1,p}, \zeta^2), \tag{4.4}$$

where ν_{-t} stands for the vector ν without the tth element. Using this as a prior specification for ν amounts at using a partially improper prior (Speckman and Sun (2003), Yue, Speckman and Sun (2012)). The coefficient $b_{tsp} = \sqrt{\frac{\pi_{tp}(1-\pi_{tp})}{n_{ts}}}$ is introduced in order to allow for

heteroscedasticity and to ease comparisons; the other two elements of (4.1), m_{sp} and ε_{tsp}, are then expressed in units of standard deviation and, as such, are directly comparable across parties and polls.

The term m_{sp}, which represents the HE of pollster s for party p, is assumed that its variance depends on the party,

$$m_{sp}|\tau_p \sim \mathcal{N}(0, \tau_p^2). \tag{4.5}$$

Finally, for the residuals ε_{tsp}, the random variation within a pollster, we assume that the variance is pollster-specific, reflecting the fact that different sampling and adjustment strategies may imply different variabilities,

$$\varepsilon_{tsp}|\sigma_s^2 \sim \mathcal{N}(0, \sigma_s^2). \tag{4.6}$$

For the variances ζ, τ_p and σ_s a half-normal hyperprior with high variance is used.

Conditional on π_{tp}, the model comprises two sources of variation, house effects (m_{sp}) and residual (ε_{tsp}), that can also be interpreted as the variability between pollsters and that within each pollster, respectively.

The roles of m_{sp} and ε_{tsp}, as sources of variation within the model, are best seen by comparing two conditional distributions of y_{tsp}. In fact, $b_{tsp}^2\sigma_s^2$ is the variance of the distribution of y_{tsp}, conditional on pollsters and π_{tp}, thus excluding the variability between pollsters,

$$y_{tsp}|\pi_{tp}, m_{sp}, \sigma_s, \tau_p \sim \mathcal{N}(\pi_{tp} + b_{tsp}m_{sp}, b_{tsp}^2\sigma_s^2), \tag{4.7}$$

while the variance of y_{tsp} conditional on π_{tp} only is given by the sum $b_{tsp}^2(\sigma_s^2 + \tau_p^2)$,

$$y_{tsp}|\pi_{tp}, \sigma_s, \tau_p \sim \mathcal{N}(\pi_{tp}, b_{tsp}^2(\sigma_s^2 + \tau_p^2)). \tag{4.8}$$

The amount of variability can be measured by the variances τ_p^2 and σ_s^2. We follow Gelman (2005) in distinguishing between the latter two, called super-population variances, and the finite population variances (fp-variances in what follows): the variances of the model predictions of m_{sp} and ε_{tsp}. Fp-variances are the most relevant quantities to describe the phenomenon; we refer our conclusions to the set of actually observed pollsters and parties, rather than to a generic population of pollsters. Moreover, this choice allows us to compare the two sources of variability for each party, notwithstanding the fact that the super-population variance of the residuals ε_{tsp} is assumed to vary across pollsters and not parties, unlike the super-population variance of m_{sp}.

Similar models have been adopted by Jackman (2005) and Hanretty (2013) as well as by Pickup and Johnston (2007) and Pickup and Johnston (2008) (see Section 1). However, a major difference between those models and our proposal is that we do not need to consider the true vote share of each party (i.e., their actual election results), a choice that requires additional and partly unjustified model assumptions. Furthermore, we are also able to overcome some limitations of these existing models. The main advantages of our model are: (1) by employing the b_{tsp} coefficient, we make the estimates of the HEs directly comparable across parties (this is not a problem in Jackman (2005) and Pickup and Johnston (2007) where only one party is considered); (2) by introducing the parameter σ_s^2, we allow for the poll's variance to be different than that expected under the random sampling assumption; (3) by introducing the parameter τ_p^2, we allow the variability of the random effects to be estimated and to differ between parties. (Note that posterior distributions of σ_s^2 and τ_p^2 effectively show heterogeneity across pollsters and parties, respectively.) Also, note that, in order to allow for a changing variance of the house effects (Erikson, Panagopoulos and Wlezien (2004)), one might explicitly model differences in variances of HE by letting the variance of m_{sp} to depend on s.

Linzer (2013) does not estimate HEs, as his objective is to obtain a prediction combining pollsters results, not analyze pollsters behaviour, and so "Correcting for overdispersion by estimating firm-specific effects is impractical because most pollsters only conduct a very small share of the surveys."

Estimation is performed using STAN (Carpenter et al. (2017), Stan Development Team (2016)) within R (R Core Team (2015)). Results are based on four parallel chains of length 5000.

5. Results. One of the results of the model is a prediction of the share of votes for each party averaged across pollsters. This is not, however, our focus, as we are interested in the sources of variability rather than the vote share prediction. In fact, we expect the inherent biases of the pollsters not to cancel out, and so we expect the pooled prediction to be biased as well. Vote share predictions, according to each pollster, are reported in Figure 3.

The main focus is on understanding the role of HEs on the total variability of the detrended vote share prediction. In model terms this entails comparing the fp-variances of m_{sp} and ε_{tsp}, whose posterior distributions are summarized by a 95% credibility intervals (high posterior density) and the medians in Figure 4. The fact that fp-variances of m_{tsp} are similar or higher than the fp-variances of ε_{tsp} is evidence of the fact that HEs play a relevant role in all three time periods. In fact, the rectangles represent the fp-variances obtained by estimating the same model on data simulated from the model itself, assuming $m_{sp} = 0$ (more precisely, we simulate $\tilde{y}_{tsp} \sim \mathcal{N}(\hat{\pi}_{tp}, \hat{\pi}_{tp}(1-\hat{\pi}_{tp})/n_{ts})$ with $\hat{\pi}_{tp}$ equal to the posterior mean, according to the model estimate on the original data), in which case fp-variances of ε_{tsp} are higher than those of m_{sp}.

It should be noted that the HEs are even larger when new parties arise (as in 2008 and, even more dramatically, in 2013); newer parties imply a higher variability of the adjustments, as it was to some extent expected and probably related to the common procedure of weighting with respect to past votes. This effect has been highlighted by Durand (2008) in the French political elections for the polling results of the National Front presidential candidate, Jean-Marie Le Pen, where pollsters underestimated his result in 2002, adjusting the forecasting in 2007, according their past behaviors and, consequently, overestimating his performance. As a consequence, the general picture in 2006, when there were fewer novelties with respect to the previous elections, exhibits a less pronounced role of m_{sp} variability.

An examination of the posterior medians of m_{sp} for 2013 (Figure 5, panel a) shows again that HEs have different magnitudes for different parties and also reveals that there are some parties for which specific pollsters exhibit relatively strong biases (pro or con). By looking at

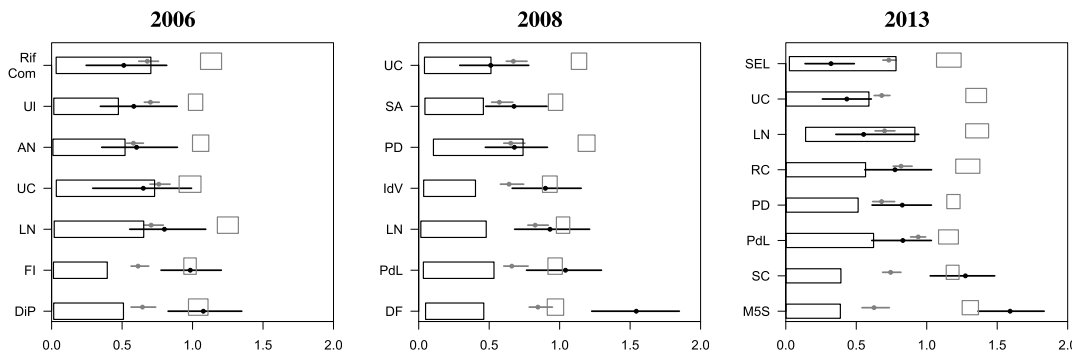
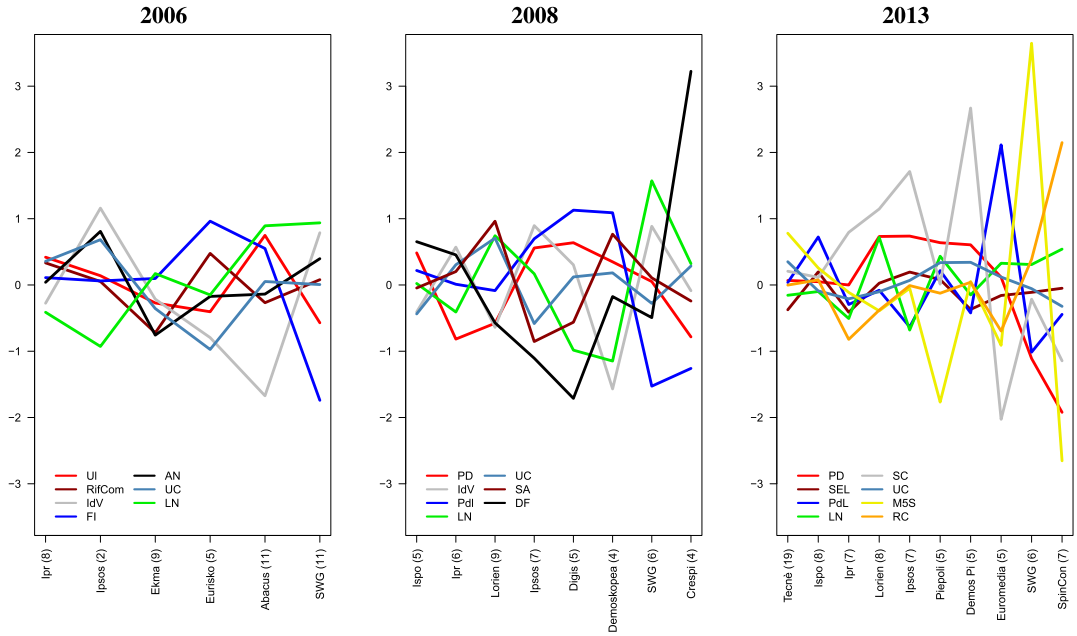
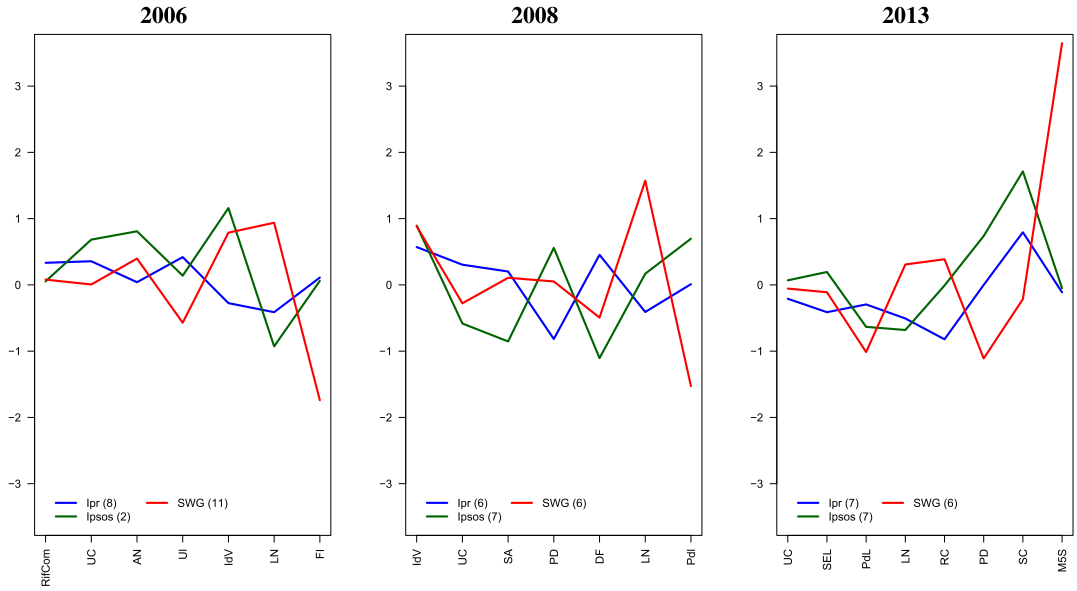


FIG. 4. *Credibility intervals* (95%) *for finite population variances for* m_{sp} (*black segments*) *and* ε_{tsp} (*gray segments*), *rectangles represent c. i. for finite population variances for* m_{sp} (*black rectangles*) *and* ε_{tsp} (*gray rectangles*) *estimated on data simulated under a no house effect scenario.*



(a)



(b)

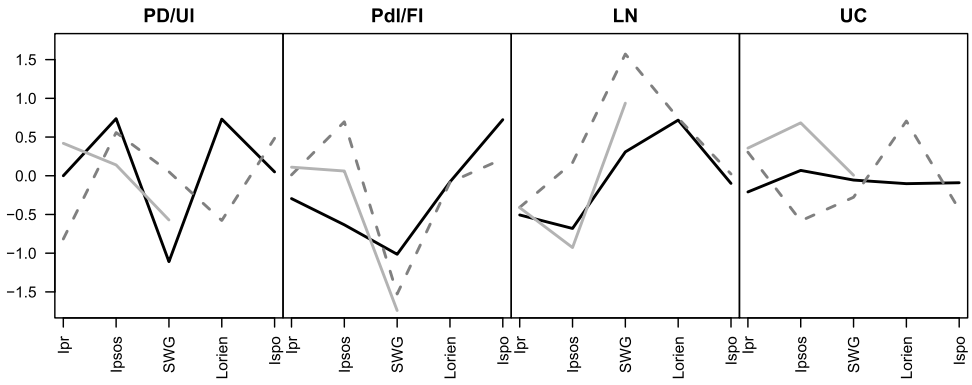
FIG. 5. *Posterior medians for m_{sp} (y-axis), ordered by variability, on the x-axis: Name of pollsters (panel a) and name of parties (panel b). In brackets, number of polls performed. In panel b, only the pollsters that appear in all three elections are depicted.*

Figure 5, panel b, it becomes clearer how HEs differ pollsterswise and, in particular, major differences arise when some parties are considered, especially SWG show a peculiar bias with respect the other two pollsters and with a greater extent in 2013.

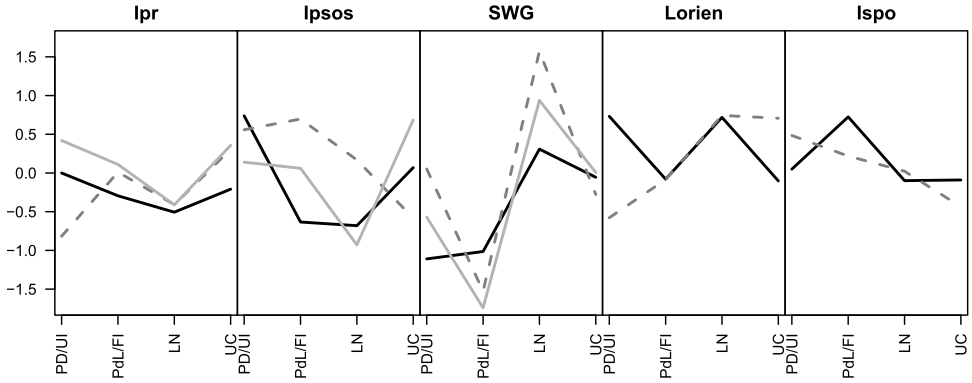
Since the HE are modeled as a random effect (in the hierarchical model specification), their estimates are shrunk toward zero and that this shrinkage is greater for those pollsters for

which less observations are available (who performed less polls). Thus, the fact that Ipsos in 2006 has smaller HE may be due to the fact that it has performed few polls with respect to the others. The fact that Tecnè in 2013 is the pollsters with the smallest HE, on the contrary, is a strong indication that they employ less (house) adjustments than the other: the shrinkage effect is lower for Tecnè since Tecnè performed two to four times the polls performed by the others. For all other pollsters the number of polls performed is fairly similar, thus the shrinkage effect is more or less the same, and the comparisons can be made ignoring the number of polls.

A different perspective on HEs is given in Figure 6 which compares, across years, the biases toward those parties (panel a) and of those pollsters that appear in at least two elections (panel b). From Figure 6, panel a, it can be noted that the HE exhibits a quite similar pattern for the PdL/FI and LN parties across pollsters and years, whereas it appears quite different for the PD/UI party especially between 2006 and 2008 (because of the Ipr polls) and between 2008 and 2013 (because of the Lorien polls). Moreover, looking from the pollster perspective (Figure 6, panel b), it is interesting to note that the pattern of HE of SWG is relatively similar in the three years: estimating higher shares for LN and lower shares for PdL in all three elections. For Ipr the effects are similar but for PD in 2008. For the other two pollsters the patterns are not consistent across the years.



(a)



(b)

FIG. 6. *House effects* (*y-axis*: *median of* m_{sp}): *Compared by pollsters* (*x-axis*) *across years and parties* (*panel a*); *and compared by parties* (*x-axis*) *across years and pollsters* (*panel b*). *Black line*, 2013; *dashed line*, 2008; *grey line*, 2006. *Only those parties and pollsters that appear in at least two elections are depicted.*

It is worth comparing our results with those of Hanretty (2013), keeping in mind that the comparison should be made with caution as the datasets used only partially overlap (Hanretty (2013) considers polls carried out in 2012, which we ignored, but did not consider unofficial polls circulating in the two weeks immediately before elections). However, the most significant house effects detected by Hanretty (2013) are confirmed by our analysis. In addition, the pollsters with greater HEs are the same (Euromedia, SWG, SpinCon).

5.1. *Model checking*. In order to assess the quality of the model, we employ the methodologies proposed by Gelman, Meng and Stern (1996) and that are further discussed in Gelman (2003) and Gelman et al. (2014) to extend classical goodness of fit procedures in the Bayesian paradigm.

This entails comparing observations y^{obs} to their model base predictive distribution $p(y|\theta)\,p(\theta|y^{\text{obs}})\,d\theta$.

In practice, having a sample $\theta^{\text{rep}(k)}$, $k = 1, \ldots, K$ from the posterior distribution obtained using an MCMC procedure, one obtains a sample of replicated data according to the predictive distribution by simulating $y^{\text{rep}(k)}$ from $p(y|\theta^{\text{rep}(k)})$ for each $k = 1, \ldots, K$ and then compares y^{obs} and y^{rep} graphically or using appropriate statistics $T(y)$ and discrepancy measures $D(y;\theta)$ (which measure the distance between the observations and the model). A synthetic measure of the disagreement between the model and the data is the so-called posterior predictive p-value (PPP), defined theoretically as $P(T(y^{\text{rep}}) > T(y^{\text{obs}})|y^{\text{obs}})$ (or $P(D(y^{\text{rep}},\theta) > D(y^{\text{obs}},\theta)|y^{\text{obs}})$) and from the MCMC sample and the replications $y^{\text{rep}(k)}$ by $\frac{1}{K}\#\{T(y^{\text{rep}(k)}) > T(y^{\text{obs}})\}$ (or $\frac{1}{K}\#\{D(y^{\text{obs}};\theta^{\text{rep}(k)}) > D(y^{\text{rep}(k)};\theta^{\text{rep}(k)})\}$). A PPP value near 0 or 1 suggests a lack of fit. The interpretation of the precise value is made difficult by the fact that the distribution of the PPP under the null hypothesis (the model is correct) is not necessarily uniform but, depending on the quantity it is based on, may be more concentrated around 0.5 (Gelman (2013)).

The most obvious comparison is that between observed data y^{obs}_{tsp} and the reference distributions

$$\int p(y_{tsp}|\theta)\,p(\theta|y^{\text{obs}}) = \int \mathcal{N}\big(\pi_{tp} + b_{tsp}(m_{sp}), b^2_{tsp}\sigma^2_s\big)\,p(\theta|y^{\text{obs}})\,d\theta.$$

Different graphical comparisons may be envisaged, observed values may be compared to the predictive distribution for each party and pollster or we can depict (a synthesis of) the posterior predictive distributions of $y^{\text{obs}}_{tsp} - y^{\text{rep}}_{tsp}$ as a (sort of) residual plot.

The standard way to assess model quality is a residual plot; residuals can be obtained from a hierarchical Bayesian model using as a pointwise estimate the median of the posterior predictive distribution of y_{tsp}, the plot of the (standardized) residuals with respect to time, shown in Figure 7, does not reveal issues with the model. Given the hierarchical nature of the model, it is also relevant to inspect the residuals conditional on the party and the pollster; this is depicted in Figure 8. In this case, instead of standardizing the residuals, we depict the raw residuals and their variability bands. Ideally, residuals should not have any trend, and zero should be included in almost all the reference regions; inspection of Figure 8 suggests that this conclusion is tenable with a possible exception for the prediction of PdL from Piepoli where the increasing trend is, to some extent, underestimated.

One aspect of the data is the difference between poll houses; this is explicitly allowed for in the model thanks to the m_{sp} effects. A useful statistic to check model fit in this respect is the sum of absolute values of differences between pairs of houses in contemporary polls; that is, we consider the statistic

$$T_{tp}(y) = \sum_{s=1}^{14} \sum_{v=s+1}^{14} |y_{tsp} - y_{tvp}| \tag{5.1}$$

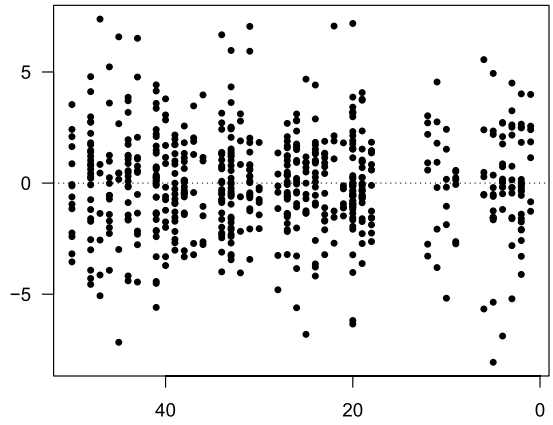


FIG. 7. *Plot of standardized residuals* (*y-axis*) *with respect to time* (*x-axis*: *days to election*). *Residuals are the difference between posterior medians of the predictive distributions*.

for $t \in \mathcal{T}_{sv}$, where $\mathcal{T}_{sv}$ is the set of times for which contemporary polls from pollsters s and v are available. In practice, 52 pairs of pollsters have contemporary polls in a total of 84 instances. PPPs corresponding to (5.1) are reported in Figure 9; they measure the extent to which the model describes the variability among pollsters (i.e., the variance of house effects) in the data. If the model systematically underestimated (overestimated) such variance, we would expect the PPP to be low (high). Overall, the PPP in Figure 9 do not suggest major discrepancies; however, the shape of the plot for some parties (in particular SC) may be suggestive of the fact that the variance of HEs reduces as election day approaches and is

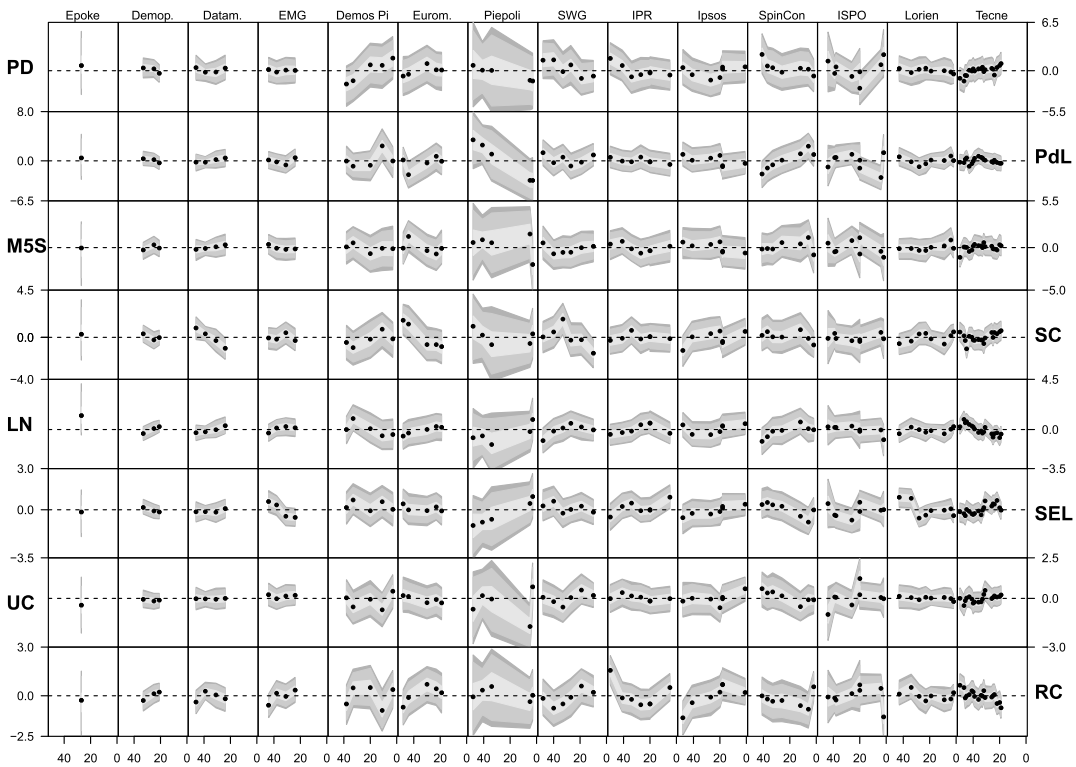


FIG. 8. *Posterior distributions of* $y^{\text{obs}}_{tsp} - y^{\text{rep}}_{tsp}$ (*y-axis*) *for year* 2013 *elections*, *reference regions for probabilities* 50% 90% 95%; *dots represent median values* (*x-axis*: *days to election*).

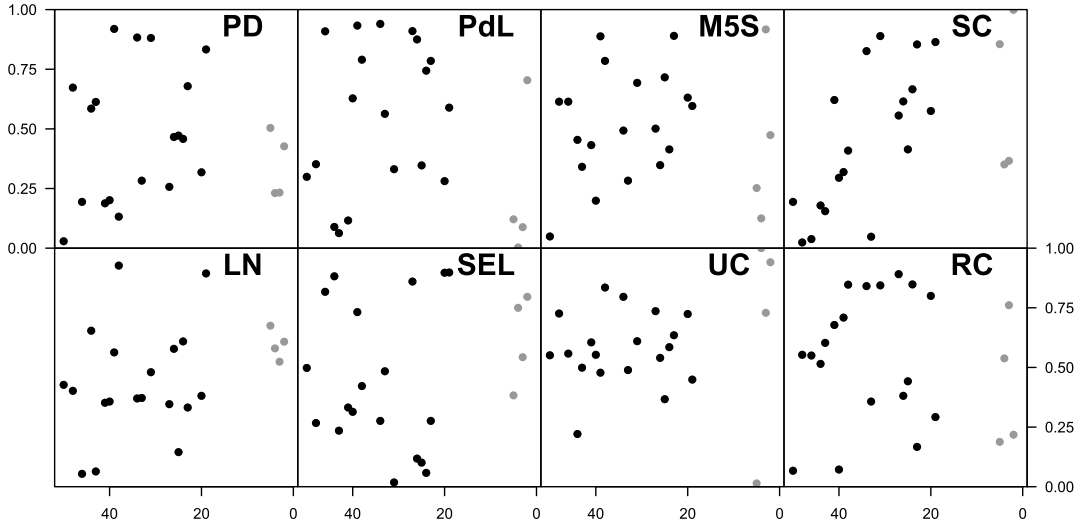


FIG. 9. *Posterior predictive p-values* (*y-axis*) *for the mean absolute difference between pollsters at fixed times* (*black*, *official polls*; *gray*, *unofficial polls*), *x-axis*: *days to election*.

further discussed in Section 5.2. We note in pass that we take this merely as suggesting further analysis, performed in Section 5.2, we do not claim PPP values in Figure 9 to have any value as evidence.

One aspect of the data, which is not (explicitly) modeled according to our specification, is the correlation between parties shares which is expected to be positive due to the almost compositional nature of the data (remember that the shares do not sum to one, due to the existence of other minor parties that are sometimes ignored, and sometimes considered as a whole in reporting polls). It is then relevant to compare observed correlations and model correlations. We also probed our model fit as far as the correlation between parties share is concerned by considering principal components analysis. For details, see Section 3 of the Supplementary Material (De Stefano, Pauli and Torelli (2022)).

Overall, according to our goodness of fit checks, the model appears adequate, with the greater deviations occurring for the smaller parties.

5.2. *Heteroscedasticity of house effects*. The variability of HEs measures to which extent the poll houses disagree in estimating the shares of a party beyond sampling variability.

In specifying the model, we assumed the variability of HEs to be constant over time; there are, however, reasons to believe that the variance may diminish as the election day approaches. Such an effect has been noted in the U.S. election by, among others, Linzer (2012) and Moore (2008). A number of explanations have been put forward for this phenomenon (Lavrakas et al. (2008)). For instance, opinions stabilise, and the number of undecided voters decreases, thus increasing the effective sample size on which the percentages are calculated. This should lead to the reduced weight of house biases in determining the final estimate. In addition, it is possible that pollsters correct themselves according to the results of others (Blumenthal (2008), Blumenthal (2014)). Finally, in the first period some results may be intentionally distorted as a means of propaganda. That some (or all) of the above mechanisms were in place for the polls of the 2013 election is suggested by the posterior predictive checks on the statistics (5.1). Therefore, we modified the model to allow for time-varying HEs and computed a finite population variance based on such estimates.

In formulas we modify model (4.1) by letting house effects vary smoothly with time, that is, (4.1) becomes

$$y_{tsp} = \text{logit}^{-1}\big(f_p(t)\big) + b_{tsp}\big(g_{sp}(t) + \varepsilon_{tsp}\big), \tag{5.2}$$

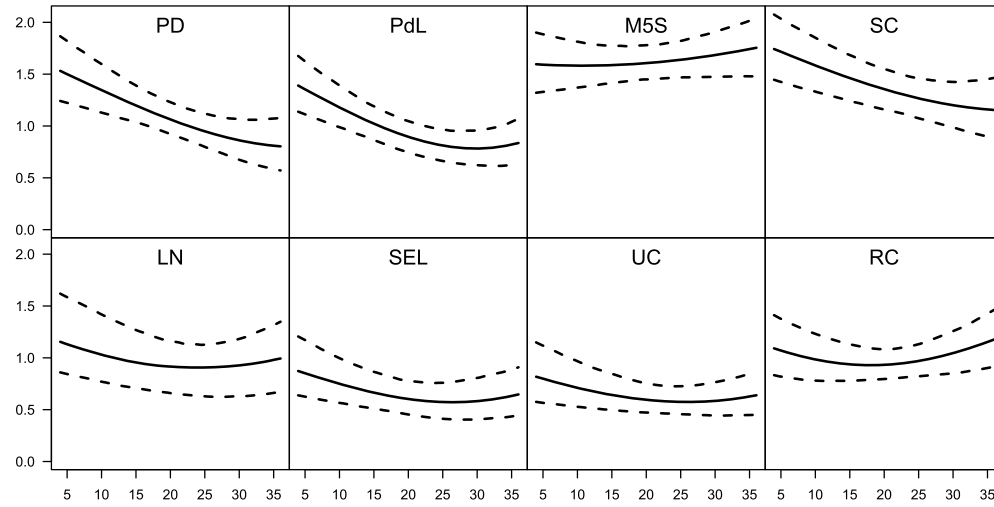


FIG. 10. *Estimated time-dependent fp-variances* (*y-axis*) *for house effects* (*x-axis*: *time*).

where $f_p(t)$ and $g_{sp}(\cdot)$ are spline functions, specified in a standard way; that is,

$$g_{sp}(x) = \sum_{k=1}^{K} m_{spk} B_k(x), \tag{5.3}$$

$$f_p(x) = \sum_{k=1}^{K} \nu_{pk} B_k(x), \tag{5.4}$$

where $B_k(\cdot)$ is a B-spline basis (or any other basis, possibly different for g and f functions).

Model (5.2) is estimated on the data for 2013, excluding the last period (those of the unofficial polls, mainly to avoid having a period of time with no estimates in the middle).

Similar to what has been done with the homoscedastic version of the model, we then compute a fp-variance for each value of t and for each party using the estimates of $g_{sp}(t)$; the results are shown in Figure 10. The estimates with credibility bands in Figure 10 suggest that the variance decreases for PD, PdL, SC, while for the other parties the variation does not appear significant (a constant line would lie within the credibility bands).

It is to be noted that also the other error component (ε_{tsp}) could conceivably have a time dependent pattern (where the error decrease in magnitude as the election day approaches). This would lead to a decreasing variance. The model could be extended to allow for such an effect if deemed appropriate; however, for the data at hand the analysis of the residuals (Figures 7, 8) does not reveal such a pattern.

6. Discussion and concluding remarks. House effects are known to affect political polls; however, since they are not directly observable, it is difficult to assess their relative contribution to the variability of polls estimate.

We propose to model polls results, using a Bayesian specification, which allows to estimate the HE for each party and each pollster (even those who published one or two polls, thanks to the Bayesian nature of the model) and to disentangle the contributions of HEs and sampling to the variability of poll results for each party involved. Thanks to that, we can draw conclusions not only on the relative importance of HEs variability but also shed light on other aspects of the phenomenon.

By adopting our model, we can quantify HEs for each pollster and party. We can then compare these HEs and conjecture which pollsters make more limited use of the post survey

adjustments, that is, rely more on sampling results. We can also point out whether some pollsters have a bias toward a specific party (the model can not distinguish the motives behind the bias but only detect it). By adapting the model, we can rigorously investigate HE time dynamics which may be useful to assess whether pollsters modify their behaviour. Finally, we can relate the magnitude of HEs in a particular election with the overall prediction error of that election results.

In the paper we run the model on a particular case study, that is, the Italian general elections in 2006, 2008 and 2013, considering only the vote share for the Camera dei Deputati. There are many peculiarities in the Italian system that make this case quite interesting in studying both the magnitude and the dynamic of the HE. For instance, the large number of parties and coalitions, compared to other systems, and, more interestingly, the huge differences across elections in terms of dissolution of old parties and creation of new political entities.

According to our model, we found that, in these elections, HEs played a major role in the overall variability and that the residual variability is less than expected under a no HE scenario, implying that nonsampling error is much more relevant than sampling error. Moreover, the fact that HEs are highly variable across pollsters, despite the fact that sources of nonsampling error are fairly common, suggests that they are inadequate corrections of such errors. Looking at HEs dynamics, we observe, particularly in the 2013 elections, a tendency in the HE variances to decrease as the election day approaches which allows us to conclude that after pollsters converge toward a given estimate according to the results of the others. This is in line with the statement reported in AAPOR (2017), where the authors affirm that polls done further from the election day contain more errors. It could be interesting to evaluate to what extent this shrinking is also a consequence of the electorate preference evolution. However, disentangling different influences on preference trends over time is beyond the scope of the proposed model, and it would require stronger assumptions or richer datasets. Another interesting finding is that HEs are even larger when new parties arise (as in 2008 and, even more dramatically, in 2013), which is coherent with the fact that the (possibly inadequate) information coming from past votes, often used by pollsters, has a crucial role in determining HEs. It has also been noted a relation between the magnitude of the HEs and the overall prediction error: in particular, in the 2013 scenario we observe the largest HEs and also the largest prediction errors. Therefore, there is scope for improving preelectoral polls as predictions of actual vote shares by improving the treatment of nonsampling errors, perhaps by using more statistically sound remedies to compensate for systematic bias sources.

Finally, we wanted to highlight that the proposed model is useful not only for preelection polls but also whenever the purpose is to combine survey results from different organizations, as for government approval, policy preferences and other contexts in order to disentangle the sources of bias affecting them.

Acknowledgments. The authors would like to thank the anonymous referees and the Associate Editor for their valuable comments that strongly improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplementary material for "Preelectoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections"
The supplementary material reports additional ex-ploratory analysis results and details on the posterior predictive p-values for variability and for correlation among parties.

REFERENCES

AAPOR (2017). An Evaluation of 2016 Election Polls in the US. Available at https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.

BLUMENTHAL, M. (2008). More on the 'Convergence Mystery'. Available at http://www.pollster.com/blogs/more_on_the_convergence_myster.php?nr=1, accessed 2018-07-10.

BLUMENTHAL, M. (2014). Polls, forecasts, and aggregators. *PS Polit. Sci. Polit.* **47** 297–300.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

DE STEFANO, D., PAULI, F. and TORELLI, N. (2022). Supplement to "Preelectoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections." https://doi.org/10.1214/21-AOAS1507SUPP

DURAND, C. (2008). The polls of the 2007 French presidential campaign: Were lessons learned from the 2002 catastrophe? *Int. J. Public Opin. Res.* **20** 275–298.

ERIKSON, R. S., PANAGOPOULOS, C. and WLEZIEN, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opin. Q.* **68** 588–601.

ERIKSON, R. S. and WLEZIEN, C. (1999). Presidential polls as a time series: The case of 1996. *Public Opin. Q.* **63** 163–177.

GAETAN, C. and GRIGOLETTO, M. (2004). Smoothing sample extremes with dynamic models. *Extremes* **7** 221–236. MR2143941 https://doi.org/10.1007/s10687-005-6474-7

GASPERONI, G. and CALLEGARO, M. (2008). Un miglioramento immeritato? La capacità predittiva dei sondaggi preelettorali e le elezioni politiche del 2008. *Polis* **22** 483–506.

GELMAN, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Stat. Rev.* **71** 369–382.

GELMAN, A. (2005). Analysis of variance—why it is more important than ever. *Ann. Statist.* **33** 1–53. MR2157795 https://doi.org/10.1214/009053604000001048

GELMAN, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electron. J. Stat.* **7** 2595–2602. MR3121624 https://doi.org/10.1214/13-EJS854

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Taylor & Francis, London.

HANRETTY, C. (2013). The 2013 Italian Election: A Poll-Based Forecast SSRN Scholarly Paper No. ID 2214605 Social Science Research Network.

HILLYGUS, D. S. (2011). The evolution of election polling in the United States. *Public Opin. Q.* **75** 962–981.

JACKMAN, S. (2005). Pooling the polls over an election campaign. *Aust. J. Polit. Sci.* **40** 499–517.

LAVRAKAS, P., TRAUGOTT, M., BLUM, M., ZUKIN, C. and DRESSER, D. (2008). The experts reply on the poll convergence mystery. *Surv. Pract.* **1**.

LINZER, D. A. (2012). Pollsters May Be Herding. Available at http://votamatic.org/pollsters-may-be-herding/, accessed 2018-07-10.

LINZER, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *J. Amer. Statist. Assoc.* **108** 124–134. MR3174607 https://doi.org/10.1080/01621459.2012.737735

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2006). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=09/04/2006&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2008). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=13/04/2008&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2013). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=24/02/2013&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MOORE, D. (2008). Evaluating the 2008 pre-election polls—the convergence mystery. *Surv. Pract.* **1**.

PANAGOPOULOS, C. (2009). Polls and elections: Preelection poll accuracy in the 2008 general elections. *Pres. Stud. Q.* **39** 896–907.

PASEK, J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opin. Q.* **79** 594–619.

PICKUP, M. and JOHNSTON, R. (2007). Campaign trial heats as electoral information: Evidence from the 2004 and 2006 Canadian federal elections. *Elect. Stud.* **26** 460–476.

PICKUP, M. and JOHNSTON, R. (2008). Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls. *Int. J. Forecast.* **24** 272–284.

PRESIDENZA DEL CONSIGLIO DEI MINISTRI—DIPARTIMENTO PER L'INFORMAZIONE E L'EDITORIA (2015). Il Sito Ufficiale dei Sondaggi Politici ed Elettorali. Available at http://www.sondaggipoliticoelettorali.it/, accessed 2018-07-10.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

SILVER, N. (2010). Pollster Ratings V4.0: Methodology. Available at http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html, accessed 2018-07-10.

SPECKMAN, P. L. and SUN, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90** 289–302. MR1986647 https://doi.org/10.1093/biomet/90.2.289

STAN DEVELOPMENT TEAM (2016). RStan: The R interface to Stan.

STURGIS, P., BAKER, N., CALLEGARO, M., FISHER, S., GREEN, J., JENNINGS, W., KUHA, J., LAUDERDALE, B. and SMITH, P. (2016). Report of the inquiry into the 2015 British general election opinion polls.

WLEZIEN, C. and ERIKSON, R. S. (2007). The horse race: What polls reveal as the election campaign unfolds. *Int*. *J*. *Public Opin*. *Res*. **19** 74–88.

WORCESTER, R. (1996). Political polling: 95% expertise and 5% luck. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **159** 5–20.

YUE, Y. R., SPECKMAN, P. L. and SUN, D. (2012). Priors for Bayesian adaptive spline smoothing. *Ann*. *Inst*. *Statist*. *Math*. **64** 577–613. MR2880870 https://doi.org/10.1007/s10463-010-0321-6