A heteroskedastic model for estimating house effect from Italian pre-electoral poll data

Domenico De Stefano
DiSPeS, University of Trieste, Italy - ddestefano@units.it

Francesco Pauli
DEAMS, University of Trieste, Italy - francesco.pauli@deams.units.it

Nicola Torelli
DEAMS, University of Trieste, Italy - nicola.torelli@deams.units.it

Abstract

We consider pre-electoral polls for Italian general elections of 2013 with the aim of shedding light on pollsters behavior. We compare vote share prediction variability across parties and pollsters and model how this variability changes over time. Estimates of a Bayesian hierarchical model showed that a large portion of the total variability of vote share predictions is explained by the so-called house effect. Furthermore we noted that the variability of vote share predictions slightly reduces over time as the election day approaches.

Keywords: hierarchical Bayesian model; house effect; Italian elections; spline.

1. Introduction

Pre-electoral polls are performed routinely by a number of pollsters, particularly during the months previous to elections, the main purpose being forecasting election results (Hylligus, 2011). In De Stefano *et al* (2013) poll results for 2013 Italian general election are analyzed for the purpose of understanding pollsters behavior. It is in fact well known that pre-electoral polls suffer from non-sampling errors, mainly due to imperfect frames, non-responses, and other forms of selection bias. Some of these non-sampling errors affect the different pollsters to different extents and they might also be more relevant than sampling errors. In order to correct for biases due to non-sampling errors, the pollsters implement diverse ad hoc adjustments, including expert opinions, to obtain the final estimates.

These variations in methodologies and designs produce the so called house effects (Wlezien & Erikson, 2007), which are pollster biases and, since they adopt markedly different techniques, have a non negligible variability. As a result, if we consider a sample of predictions coming from many pollsters, the variability of such predictions is partly due to sampling variability, partly due to house effects variability.

The model considered in De Stefano *et al* (2013), which adopts an approach similar to Linzer (2013) and Jackman (2005), showed the relevance of the house effect phenomenon. To deepen the understanding of this mechanism we extend the model in De Stefano *et al* (2013) to allow for heteroskedasticity of house effects.

2. Data and model

We observe the vote shares of pre election polls for the eight main parties (see table **??**) from the 5th of January to the 23rd of February 2013, provided by 14 pollsters, a total of 89 observations is available. Data up to the 4th of February were obtained from the governmental site where all polls which are published or broadcasted for the general public must be communicated (`www.sondaggipoliticoelettorali.it`). The remaining 15 most recent polls were obtained from informal sources as in this period the release of results of polls to the general public is forbidden by law, thus, these data are less reliable. In figure **??** poll results for one party are depicted. The frequency with which the different pollsters perform polls is quite variable, ranging from 19 to only 4.

Let y_{tsp} be the result of the survey made on day t by house s on party p and n_{ts} be the number of respondents. We assume a Gaussian distribution for y_{tsp} and assume it is the sum of three components: time (trend), pollsters bias (house effects) and a residual. Let then

$$y_{tsp} = \pi_{tp} + b_{tsp}(m_{sp} + \varepsilon_{tsp}), \tag{1}$$

Coalition	Party	Actual vote share
Left	Partito Democratico (PD)	25.43
	Sinistra Ecologia Libertà (SEL)	3.2
	Other (2 parties)	0.92
Right	Il Popolo Della Libertà (PdL)	21.56
	Lega Nord	4.09
	Other (7 parties)	3.53
Center	Scelta Civica Con Monti Per L'Italia	8.3
	Unione Di Centro (UdC)	1.79
	Other (1 party)	0.47
Non-aligned	Movimento 5 Stelle Beppegrillo.It (M5S)	25.56
	Rivoluzione Civile (Riv Civ)	2.25
	Other (29 parties)	2.9

Table 1: Italy's parties in 2013 elections with actual vote shares.

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 various non sampling errors. 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))

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

$$\nu_{tp}|\nu_{-t,p}\zeta \sim \mathcal{N}\left(\frac{1}{2}(\nu_{t-1,p} + \nu_{t+1,p}), \frac{\zeta^2}{4}\right), \quad t = 2, \ldots, T-1 \tag{3}$$

$$\nu_{Tp}|\nu_{-T,p}\zeta \sim \mathcal{N}\left(\nu_{T-1,p}, \zeta^2\right) \tag{4}$$

the coefficient $b_{tsp} = \sqrt{\frac{\pi_{tp}(1-\pi_{tp})}{n_{ts}}}$ is introduced in order to ease comparisons, the other two elements of (**??**), m_{sp} and ε_{tsp}, are then expressed in units of s.d. and, as such, are directly comparable across parties and polls.
The term m_{sp} represents the house effect of pollster s for party p, it is assumed that its variance depends on the party,

$$m_{sp}|\tau_p \sim \mathcal{N}\left(0, \tau_p^2\right) \tag{5}$$

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,

$$\varepsilon_{tsp}|\sigma_s^2 \sim \mathcal{N}\left(0, \sigma_s^2\right) \tag{6}$$

For the variances ζ, τ_p and σ_s a half normal hyperprior with high variance is used.
The model comprises three sources of variation, time (π_{tp}), house effects (m_{sp}) and residual (ε_{tsp}). Conditional on π_{tp}, the latter two can also be interpreted as the variability between pollsters and that within each pollster, respectively.
Our focus is on the house effects variability as measured by the superpopulation variances τ_p^2 and σ_s^2 and the finite population variances (fp-variances in what follows), that is the variances of the model predictions of m_{sp} and ε_{tsp} (Gelman, 2005). Finite population variances are the most relevant quantity to describe the phenomenon.
The main conclusion which this model allows is that the house effects have a relevant role on the total variability of the detrended vote share prediction. This is shown by the comparison of the fp-variances of m_{sp} and ε_{tsp} in figure **??** where their posterior distributions are summarized by the HPD 95% credibility intervals and the medians. As it is not immediately clear what to expect were the house effects not in place, we estimated the model on data simulated from the model itself assuming $m_{sp} = 0$. The lighter colored lines represent the fp-variances obtained by estimating the model on simulated data. We note that in this case the fp-variances of ε_{tsp} are higher than the other.

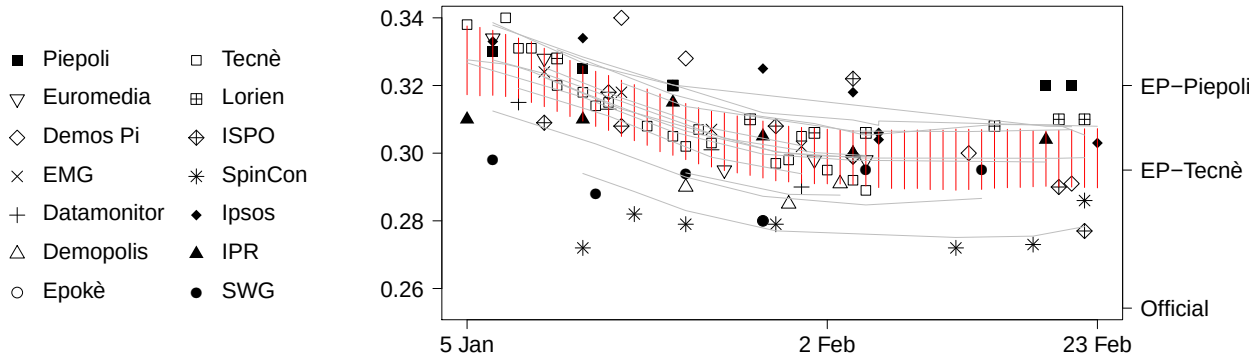

Figure 1: Poll results and model estimates for *Partito democratico (PD)*, red vertical lines represent credibility intervals for π_{tp}, gray lines represents pollster-specific predictions, on the right vertical axis, official results and exit polls (EP) are reported

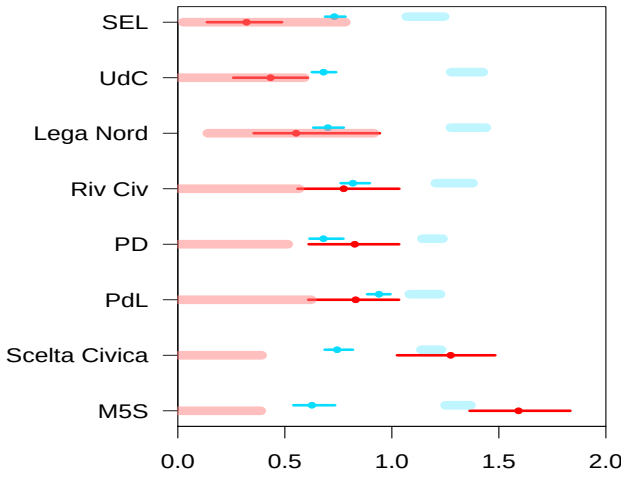

Figure 2: Finite population variances for m_{sp} (red) and ε_{tsp} (blue), dark red (blue) represents prediction from the data, light red (blue) represents prediction on data simulated under a no house effect scenario

A thorough discussion of the results is in De Stefano *et al* (2013), where a multi-year comparison is also performed.
This model assumes that the variability of house effects is constant over time, there are, however, reasons to believe that the variance may diminish as election day approaches. On the one hand, the number of undecided reduces, thus increasing the effective sample size on which the percentages are calculated, this could lead to a reduced weight of house biases in determining the final estimate. On the other hand, it is possible that pollsters correct themselves according to the results of the others. Furthermore, in the first period some results may be intentionally distorted as a mean of propaganda. All this circumstances might lead to shrinkage of the pollsters estimate toward a common value.
In order to model these phenomena we generalize the model to house effects which vary smoothly over time. The smoothness requirement appears reasonable and is also needed to avoid over-fitting.

3. Heteroskedasticity of house effects
The model comprises a common trend $f_p(t)$ analogous to π_{tp} in (**??**). House effect is modeled as a time varying parameter $g_{sp}(t)$ rather than m_{sp} in (**??**), which is constant over time.
In formulas, model (**??**) becomes

$$y_{tsp} = \text{logit}^{-1}(f_p(t)) + b_{tsp}(g_{sp}(t) + \varepsilon_{tsp}) \tag{7}$$

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

$$g_{sp}(x) = \sum_{k=1}^{K} m_{spk} B_k(x) \tag{8}$$

$$f_p(x) = \sum_{k=1}^{K} \nu_{pk} B_k(x) \tag{9}$$

where $B_k(\cdot)$ is a B-spline basis. A time varying finite population variance for the house effects can then be computed based on estimates of $g_{sp}(t)$.
Model (**??**) is estimated excluding the last period (those of the non official polls, to avoid having a period of time with no estimates in the middle). Predictions were obtained for all parties with both models. In figure **??** we depict predictions for two parties comparing the homoskedastic and the heteroskedastic models. A slight tendency to converge toward a common vote sharing is seen especially if we look at the pollsters exhibiting most extreme predictions.
Similarly to what has been done with the homoskedastic version of the model we then compute a "finite population variance" for each value of t and for each party using the estimates of $g_{sp}(t)$, the results are shown in figure **??**. We see in figure **??** that the variances are decreasing (in three out of eight cases) or constant, consistent with what expected.

4. Conclusions
According to our results, pollsters behavior is widely heterogeneous, to the point that house effects represent the largest portion of the total variability of the detrended vote share predictions.
It has also been detected that these effects reduce their magnitudes as election day approaches, although such shrinkage effect is limited in size with respect to the overall variability of the phenomenon.

References

De Stefano D., Pauli F. & Torelli N. (2013). A hierarchical Bayesian model for house effects in pre-electoral polls. Proceedings of the 8th Conference on Statistical Computation and Complex Systems, Milano.

Gaetan, C. & Grigoletto, M. (2004). Smoothing sample extremes with dynamic models. *Extremes*, *7*, 221–236.

Gelman, A. (2005). Analysis of variance–why it is more important than ever. *The Annals of Statistics 33*(1), 1–53.

Hillygus, D. S. (2011). The evolution of election polling in the united states. *Public Opinion Quarterly 75*(5), 962–981.

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science 40*(4), 499–517.

Linzer, D. A. (2013). Dynamic bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association 108*(501), 124–134.

Wlezien, C. & R. S. Erikson (2007). The horse race: What polls reveal as the election campaign unfolds. *International Journal of Public Opinion Research 19*(1), 74–88.

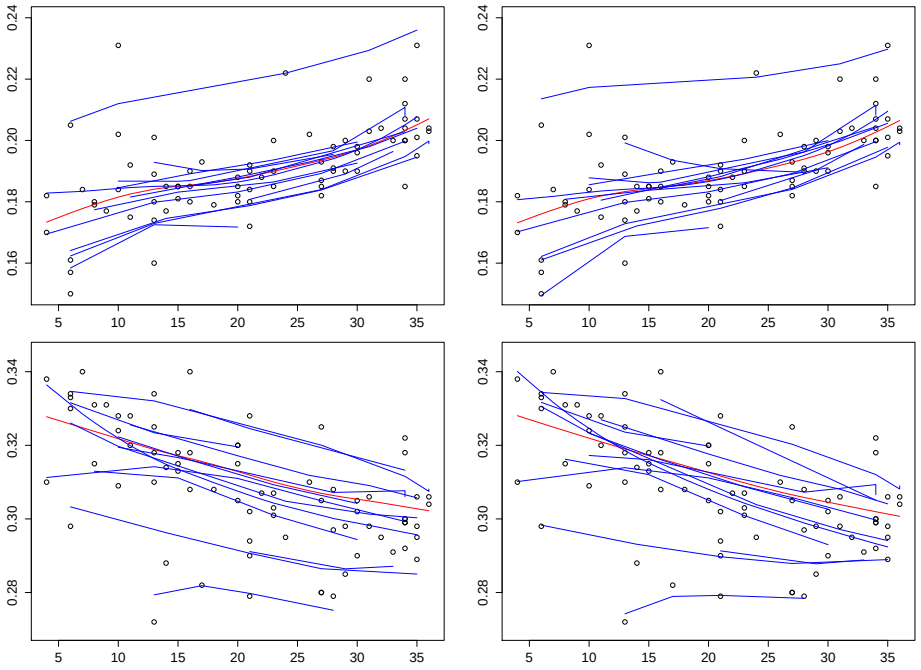

Figure 3: Predictions according to homoscedastic model (left) and heteroskedastic one (right) for Pdl (top row) and PD (bottom row), blue lines represent predictions according to different pollsters, red line is the estimate of π

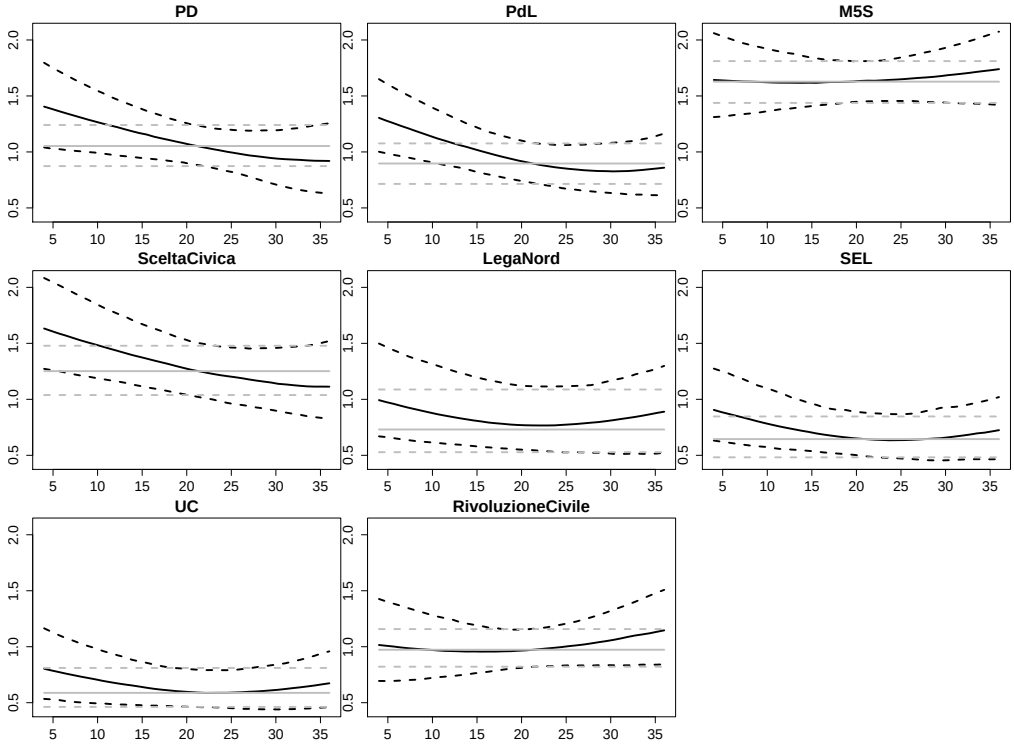

Figure 4: Estimated time dependent h.e. f-p variances