Statistical Methods for Election Forecasting

Timothée Stumpf Fétizon [timstf@gmail.com]

January 31, 2020

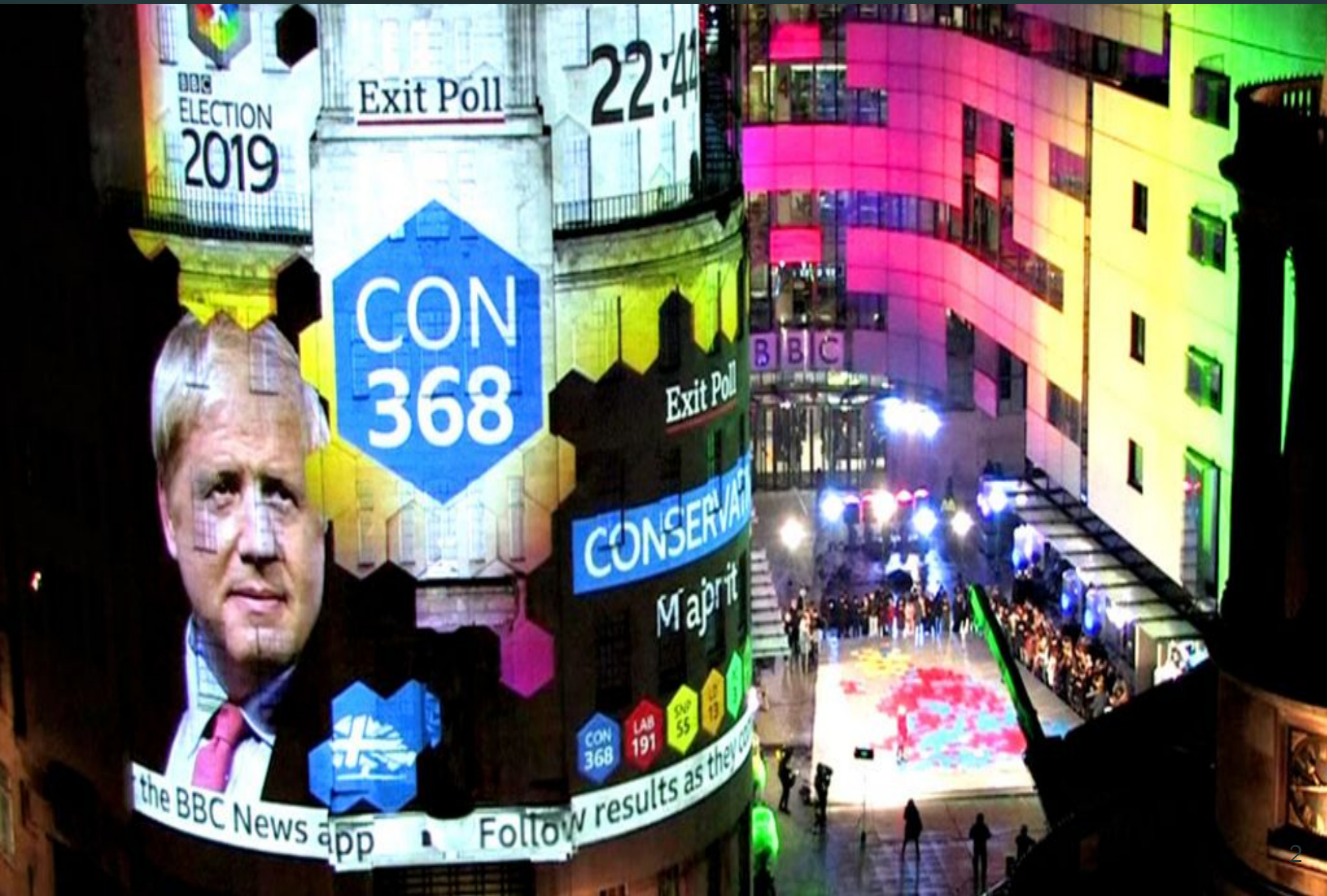
Barcelona GSE Data Science Center

Executive Summary.

Abstract: We design an election forecasting engine that delivers probability statements for arbitrary electoral scenarios. It accommodates data at different levels of granularity, an evolving party landscape and local seat assignment rules. We use the engine to forecast recent Spanish general elections.

Motivation

Mainstream electoral forecasting ignores uncertainty.



Pollsters overstate confidence when acknowledging uncertainty.

Pollsters assume trivial sampling models, e.g. for a given party:

$$\underbrace{v}_{\text{vote count}} \mid \underbrace{p}_{\text{voting propensity}}, \underbrace{N}_{\text{sample size}} \sim \text{Binomial}[N, p] \tag{1}$$

Implying sampling variance of the vote ratio:

$$\text{Var}[v/n] = p(1-p)/N \leq 1/(4N) \tag{2}$$

Pollsters commonly use $1/(4N)$ to compute confidence intervals. If this were accurate, could run large sample poll and save cost of election!

Information of heterogeneous quality needs to be combined.

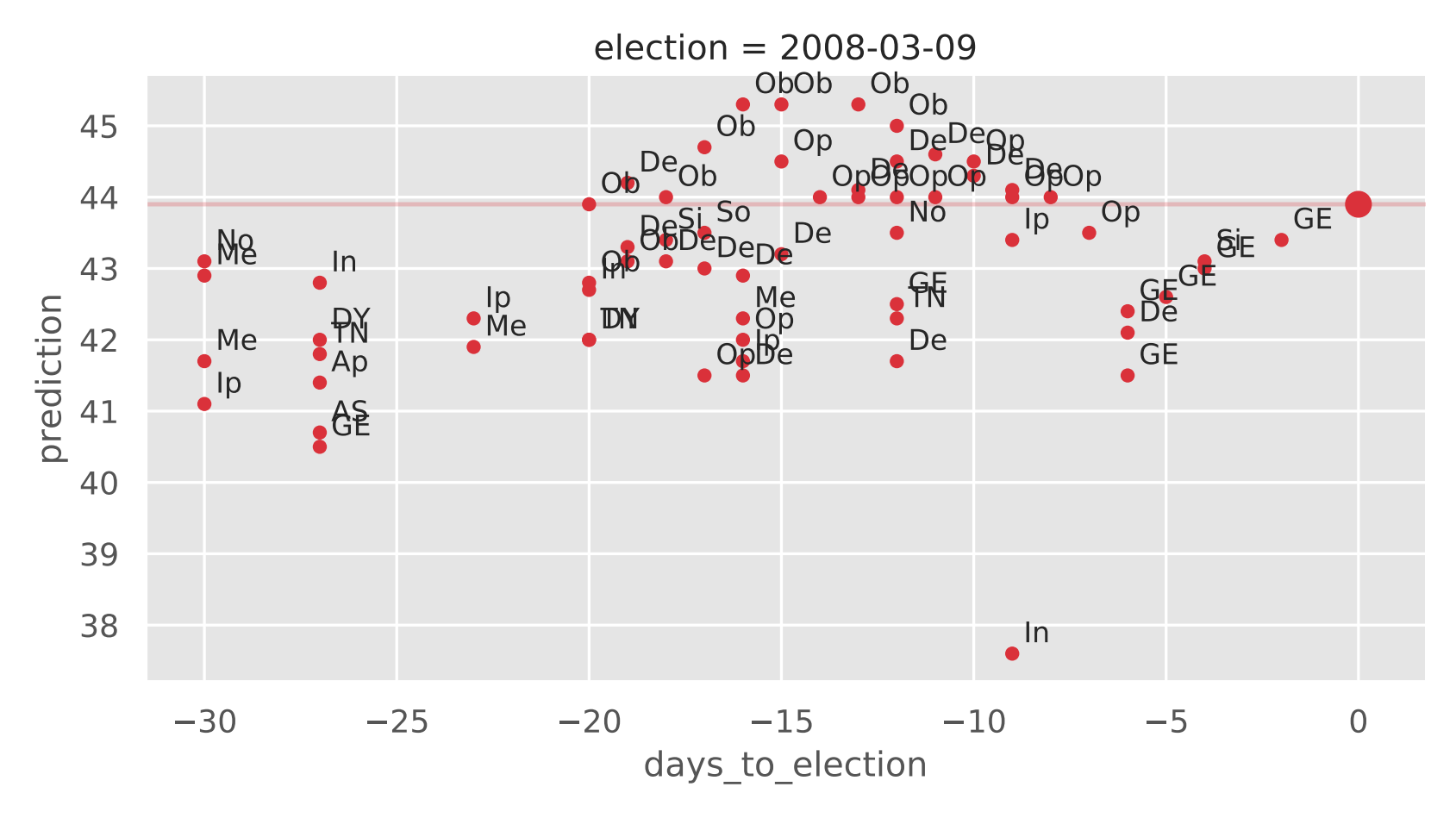


Figure 1: Polls published up to 30 days before the 2008 Spanish general election. How do we assess and merge them?

Approach

Generate a predictive distribution over seats.

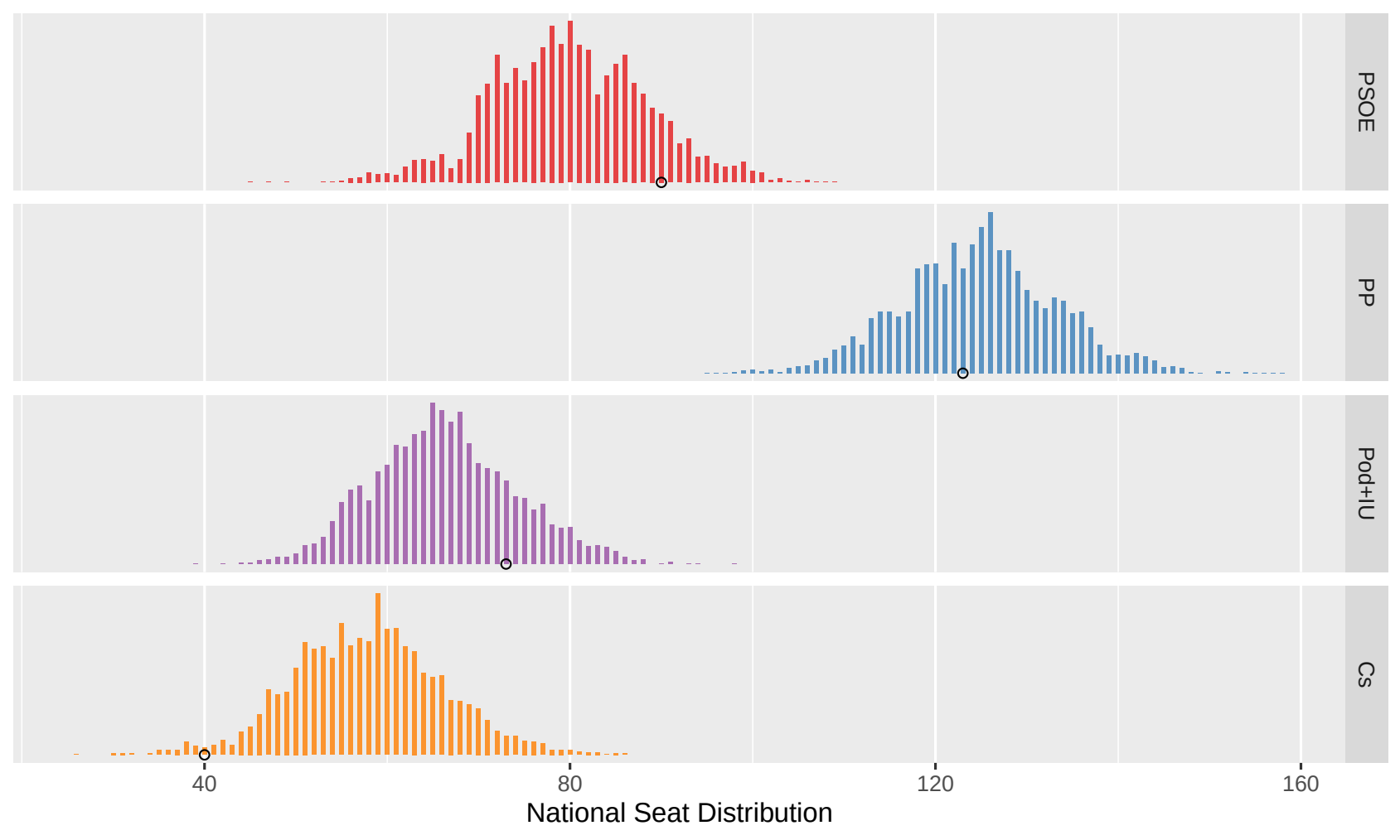


Figure 2: Marginals of predictive seat distribution for 2015 Spanish general election.

Where seat assignment is local, model locally.

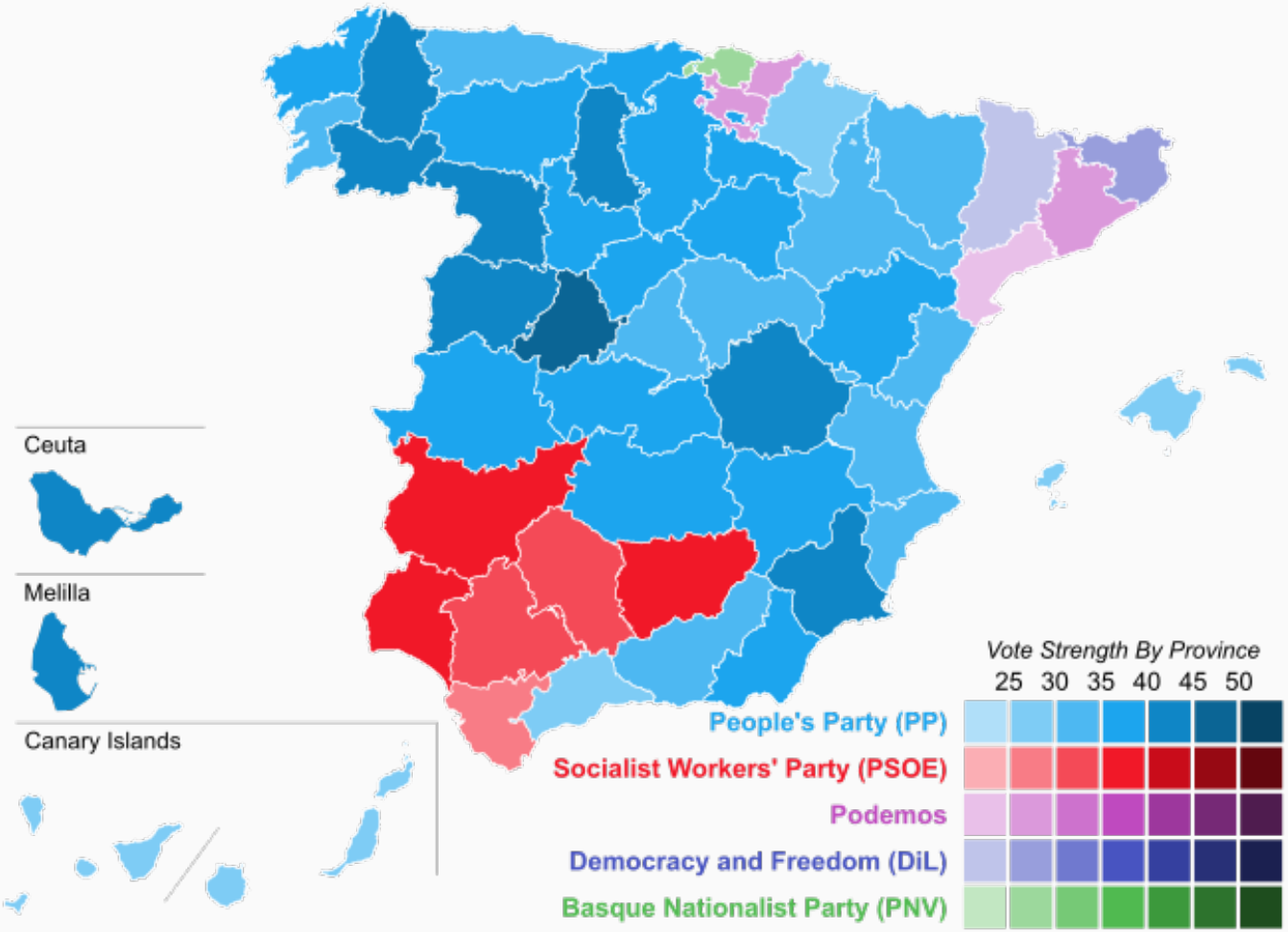


Figure 3: Electoral map of 2015 Spanish general election. Parties that concentrate their vote need fewer votes per seat!

Use distinct models to address different geographical levels.

We observe

- Plentiful low-resolution data (generic national polls).
- Sparse high-resolution data (large-sample survey microdata).

Province increment model by Gelman and Lock (2010) decomposes problem of generating local forecasts:

$$f(\underbrace{v_j}_{\text{vote province } j} \mid \underbrace{P}_{\text{lo-res}}, \underbrace{Q}_{\text{hi-res}}) = \int_0^1 f(v_j|v, Q) f(v|P)\,\mathrm{d}v \tag{3}$$

- $f(v|P)$ models national vote.
- $f(v_j|v, Q)$ models local discrepancies.

Given local votes, seat allocation is deterministic!

National Vote Forecasting

Given the generative model, the predictive follows.

Formulate a parametric model such that independently:

$$p_k \sim f(\cdot| \underbrace{\theta}_{\text{params}}, v) \tag{4}$$

Given priors over v and θ, obtain the joint model:

$$f(v, P, \theta) \propto \underbrace{f(P|\theta, v)}_{\text{likelihood}} \times \underbrace{f(v|\theta)}_{\text{result prior}} \times \underbrace{f(\theta)}_{\text{param prior}} \tag{5}$$

$$\propto f(v|\theta) \times f(\theta) \times \prod_k f(p_k|\theta, v) \tag{6}$$

Finally, average over parameter uncertainty:

$$f(v|P) \propto \int f(v, P, \theta)\, \mathrm{d}\theta \tag{7}$$

Devise a generative model for polls.

Decompose prediction error of each poll:

$$p_k = \underbrace{v_{l[k]}}_{\text{result elec } l} + \underbrace{\gamma_{m[k]}}_{\text{bias pollster } m} + \underbrace{\delta_{l[k]}}_{\text{bias elec } l} + \underbrace{\epsilon_{l[k]}(t[k])}_{\text{drift } t \text{ days before elec } l} + \underbrace{\eta_k}_{\text{noise}} \tag{8}$$

$$\eta_k \sim \mathrm{N}\left[0, \sigma_\eta^2\right] \tag{9}$$

- ϵ_l is a continuous-time stochastic process that allows for sentiment shifts during sampling period
- v_l is observed for training elections and unobserved for the test election.

Use hierarchical priors to regularize.

Pool all replicated parameters to a common prior:

$$v_l \sim \mathrm{N}\left[\mu_v, \sigma_v^2\right], \qquad \gamma_m \sim \mathrm{N}\left[0, \sigma_\gamma^2\right], \qquad \delta_l \sim \mathrm{N}\left[0, \sigma_\delta^2\right] \tag{10}$$

In Stan simulations, location hyperparameters are commonly assigned a Cauchy prior and scale hyperparameters a half-Cauchy prior, e.g.

$$\mu_v \sim \mathrm{St}[0, 1, 1], \qquad \sigma_v \sim \text{Half-St}[0, 1, 1] \tag{11}$$

Different priors are possible for the sentiment drift process.

The prior on the sentiment process controls its smoothness.

Desiderata for a prior $f(\epsilon_l)$ are:

- Sample paths should be (almost surely) continuous.
- Smoothness induces trending, which is desirable given that news do not instantly propagate.
- Computational cost should be linear in length of sampling period.

Integrated Brownian Motion fulfills those requirements. Definition:

$$\epsilon_l(t) = \sigma_\epsilon \int_0^t w_l(s)\,\mathrm{d}s \qquad (0 < s \leq t) \tag{12}$$

$$\epsilon_l(0) = 0 \tag{13}$$

where w_l are independent Brownian motions. Hyperprior on σ_ϵ will be decisive in setting the stability of the process!

Output for training election sentiment process.

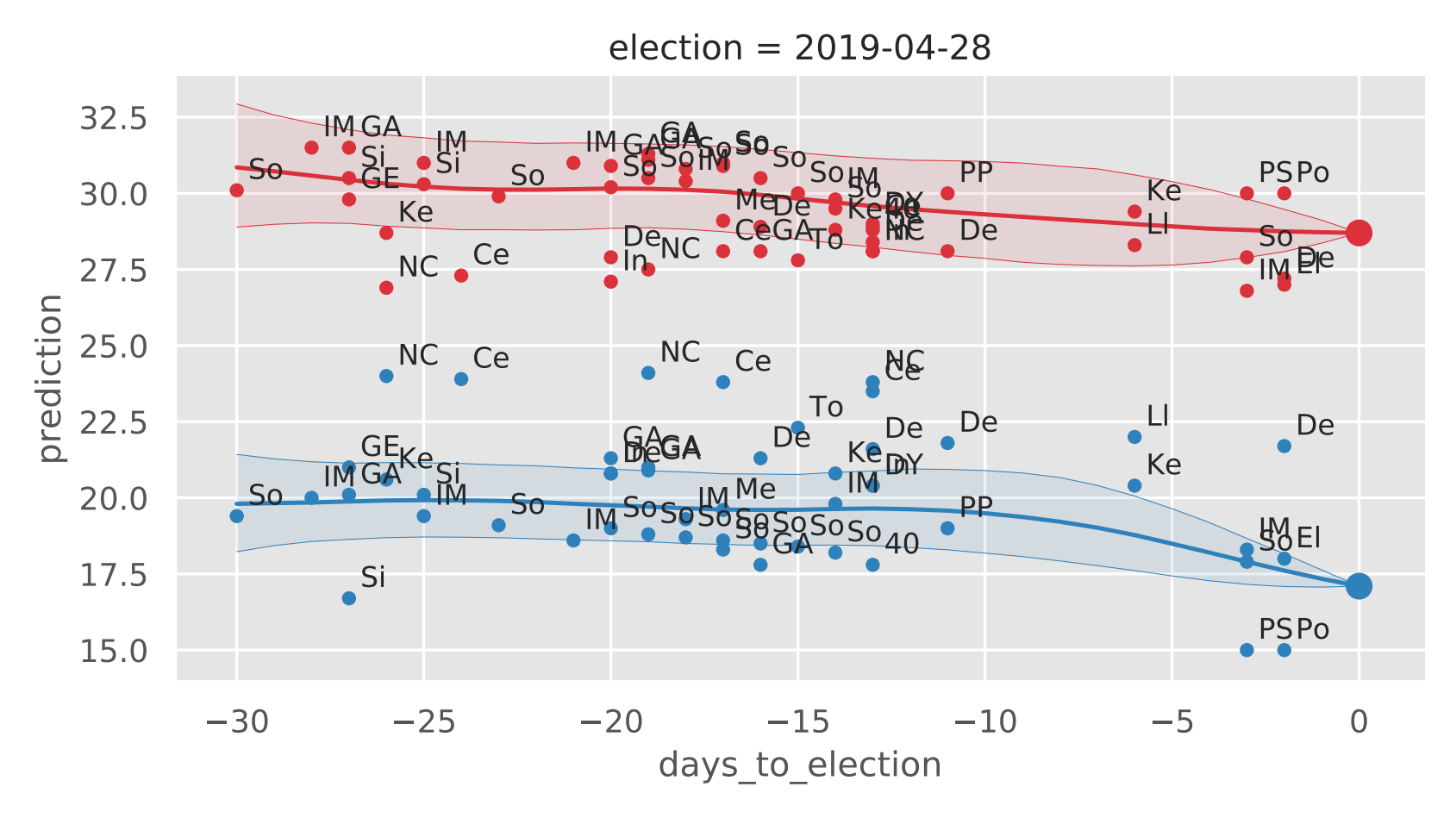


Figure 4: Inferred sentiment process for the April 2019 Spanish general election (in sample).

Output for test election sentiment process.

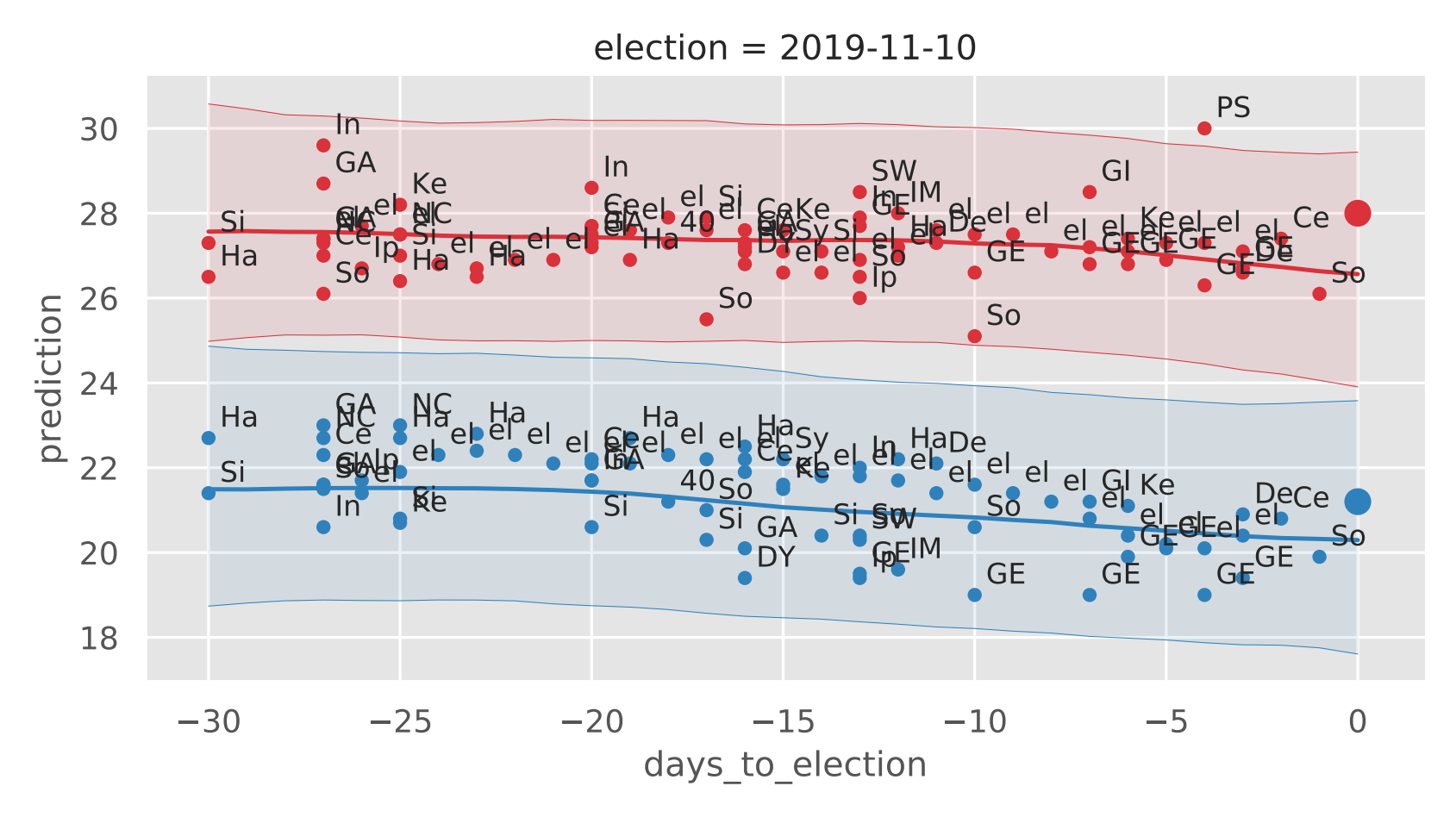


Figure 5: Inferred sentiment process for the November 2019 Spanish general election on election day (out of sample).

Output for test election result.

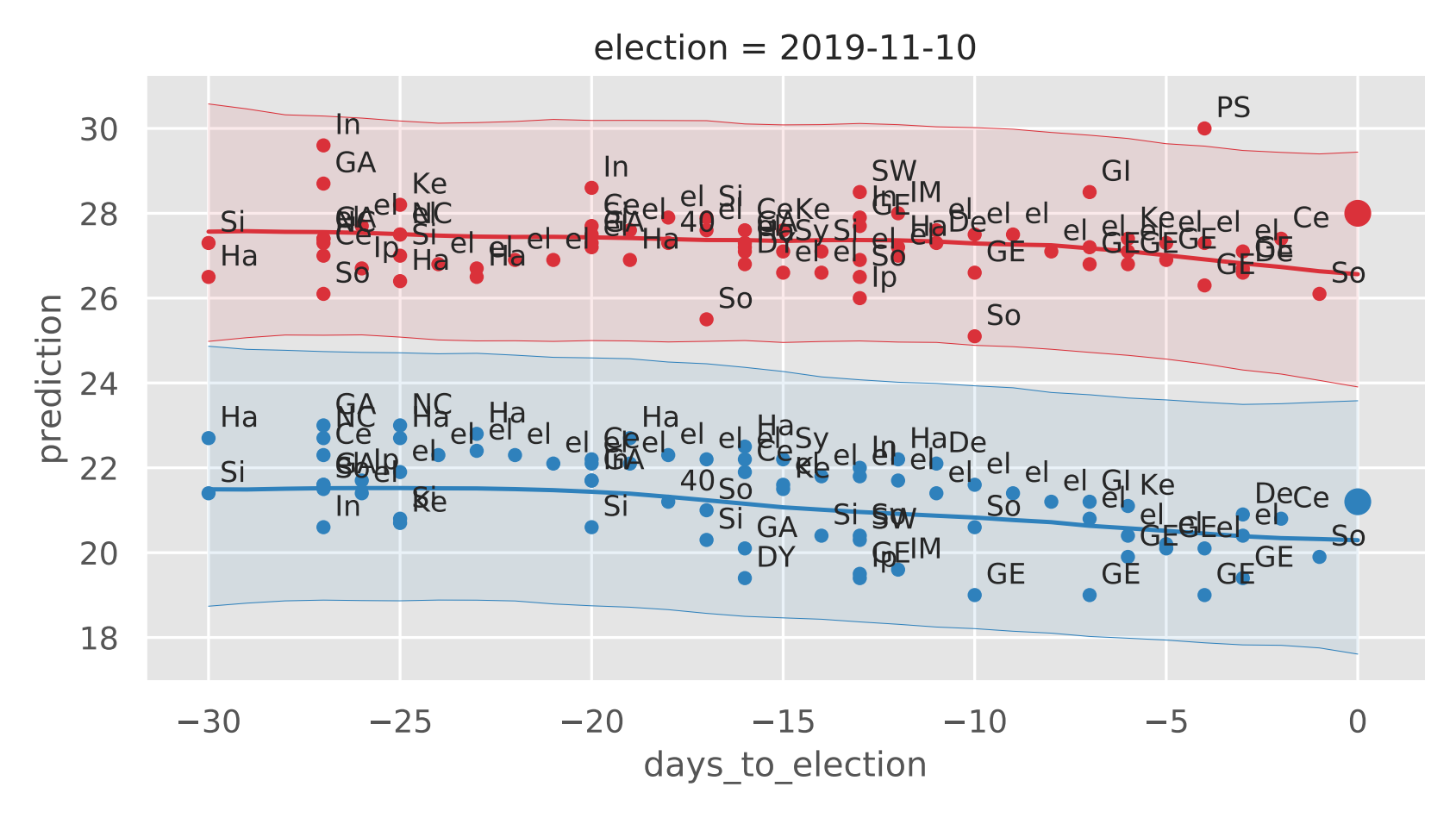


Figure 6: Inferred sentiment process for the November 2019 Spanish general election on election day (out of sample).

Local Vote Forecasting

Output for test election sentiment process.

Occasionally, pollsters publish their microdata. Can we improve on their cooking?

	study_id	province_id	municipality_size	education	...
1	3117	1	city	secondary	...
2	3117	1	city	tertiary	...
3	3117	1	city	secondary	...
4	3117	1	city	secondary	...
..	...	...	...	...	...

Table 1: Survey microdata published by CIS a couple weeks prior to a Spanish general election.

Post-stratification is a time-tested survey technique.

Survey responses consist of tuples may be grouped into *strata* according to categorical variables $\boldsymbol{x}_h$:

$$\left(\underbrace{v_n}_{\text{vote indicator}} , \underbrace{h[n]}_{\text{stratum respondent } n} \right) \tag{14}$$

A simple vote estimate consists of the sample mean:

$$\bar{v} = N^{-1} \sum_{n=1}^{N} v_n \tag{15}$$

If exact population frequencies $f(h)$ are available from some source (e.g. census), define the *post-stratified* estimator:

$$\tilde{v} = \sum_{n=1}^{N} f(h[n]) v_n / N_{h[n]} = \sum_{h=1}^{H} f(h) \times \underbrace{v_h}_{\text{votes stratum } h} / \underbrace{N_h}_{\text{size stratum } h} \tag{16}$$

Post-stratified estimates are resistant to some survey biases.

Suppose $(v_n, h[n])$ are sampled according to:

$$f(v_n|h[n]) \times g(h[n]), \qquad g(h[n]) \neq f(h[n]) \tag{17}$$

Then the sample mean estimator may be biased (from the *tower property*):

$$\mathrm{E}[\bar{v}] = N^{-1} \sum_{n=1}^{N} \mathrm{E}[v_n] = N^{-1} \sum_{n=1}^{N} \sum_{h=1}^{H} g(h) f(v_n|h) = \sum_{h=1}^{H} g(h) f(v_n|h) \tag{18}$$

The post-stratified estimator is still unbiased (again from the *tower property*):

$$\mathrm{E}[\tilde{v}] = \sum_{h=1}^{H} f(h)\, \mathrm{E}[v_h/N_h] = \sum_{h=1}^{H} f(h) f(v_n|h) = f(v_n) \tag{19}$$

Holt and Smith (1979) give broad conditions where $\tilde{v}$ outperforms $\bar{v}$ even in the absence of sampling bias.

Smoothing stratum counts reduces variance.

We have been using empirical frequencies v_h/N_h to estimate $f(v_n|h[n])$. Noisy where N_h is small! Instead, smoothen by estimating parametric logit model:

$$v_n| \underbrace{x_n}_{\text{categorical scalar}} \sim \text{Bernoulli}[p(x_n)] \tag{20}$$

$$p(x_n) = \text{expit}[\alpha + \beta_{x_n}] \tag{21}$$

As usual, use hierarchical priors:

$$\beta_{x_n} \sim \text{N}\left[0, \sigma_\beta^2\right] \tag{22}$$

Non-stratified estimates.

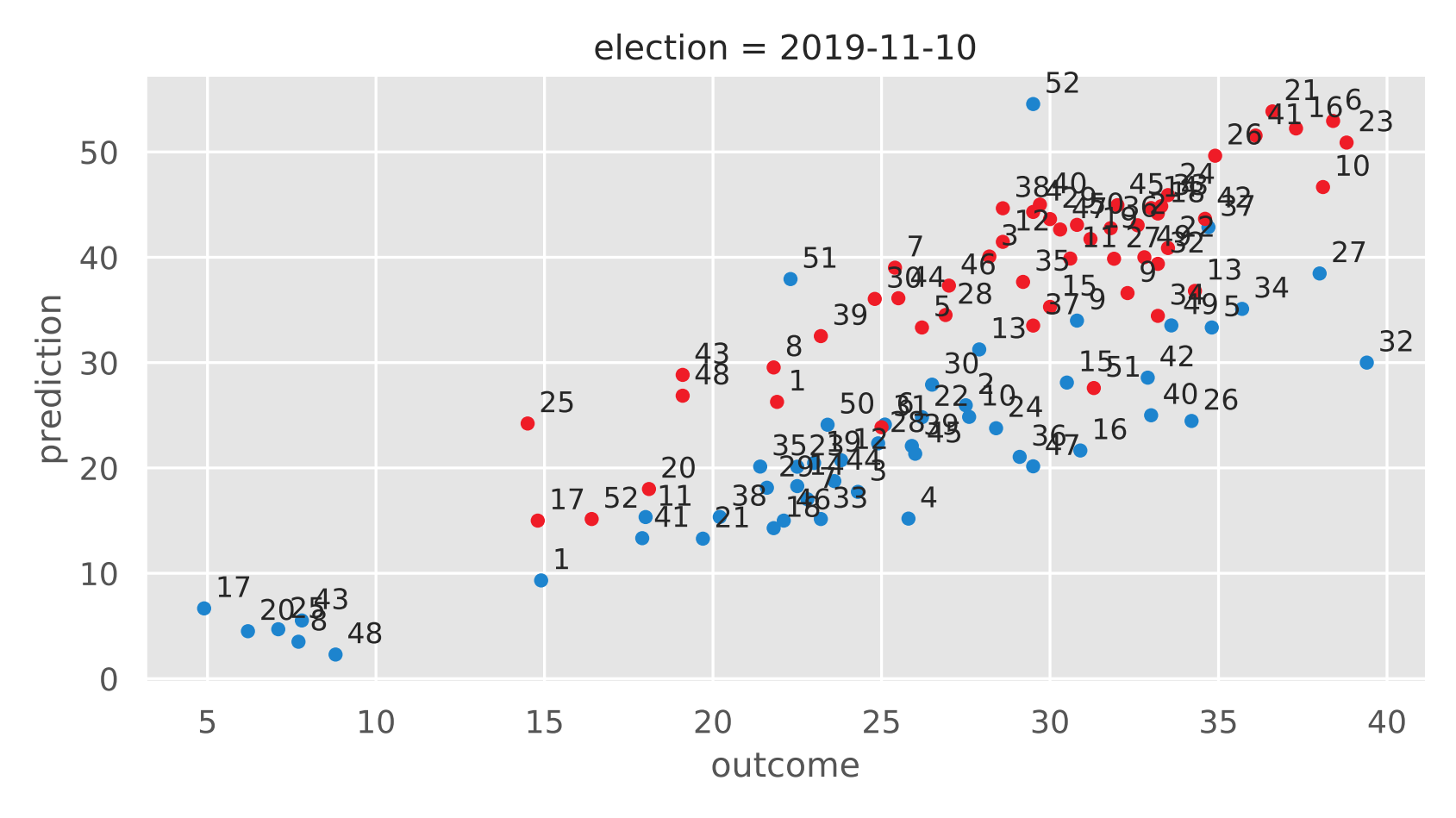


Figure 7: Raw microdata estimates for the 2019 Spanish general election.

Post-stratified estimates.

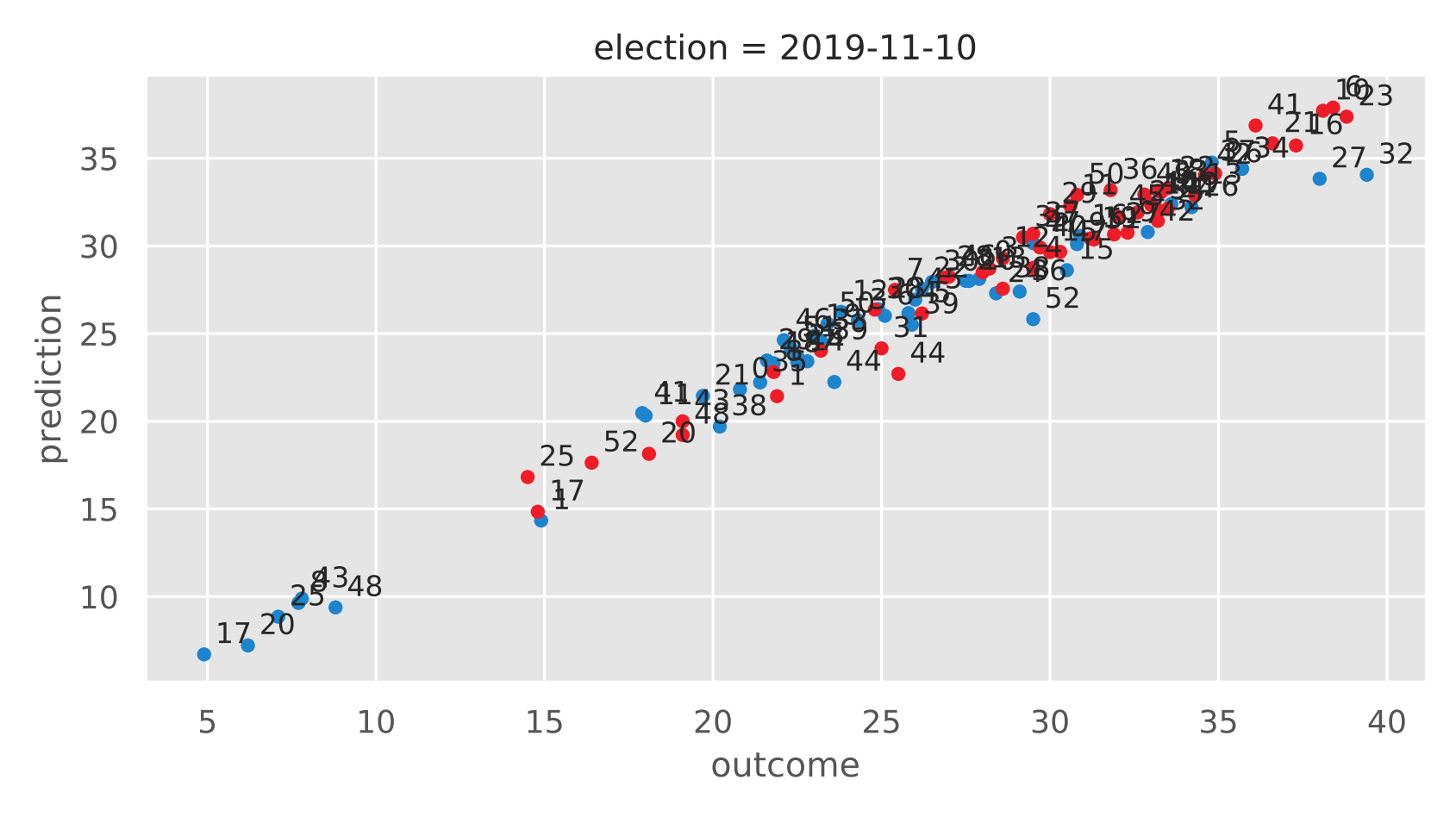


Figure 8: Post-stratified estimates for the 2019 Spanish general election using variables `province` and `previous vote`.

Final touches.

Formulate a parametric model such that independently:

$$\underbrace{q_{kj} - q_k}_{\text{est prov inc poll } k} \sim f(\cdot|\ \underbrace{\phi}_{\text{params}}, v, v_j) \tag{23}$$

Proceed with a simpler version of the national polls model, e.g.

$$q_{kj} - q_k = \alpha_j + \beta(v_j - v) + \eta_k, \qquad \eta_k \sim \mathrm{N}\left[0, \sigma^2_{j[k]}\right] \tag{24}$$

Results

Was it worth it?

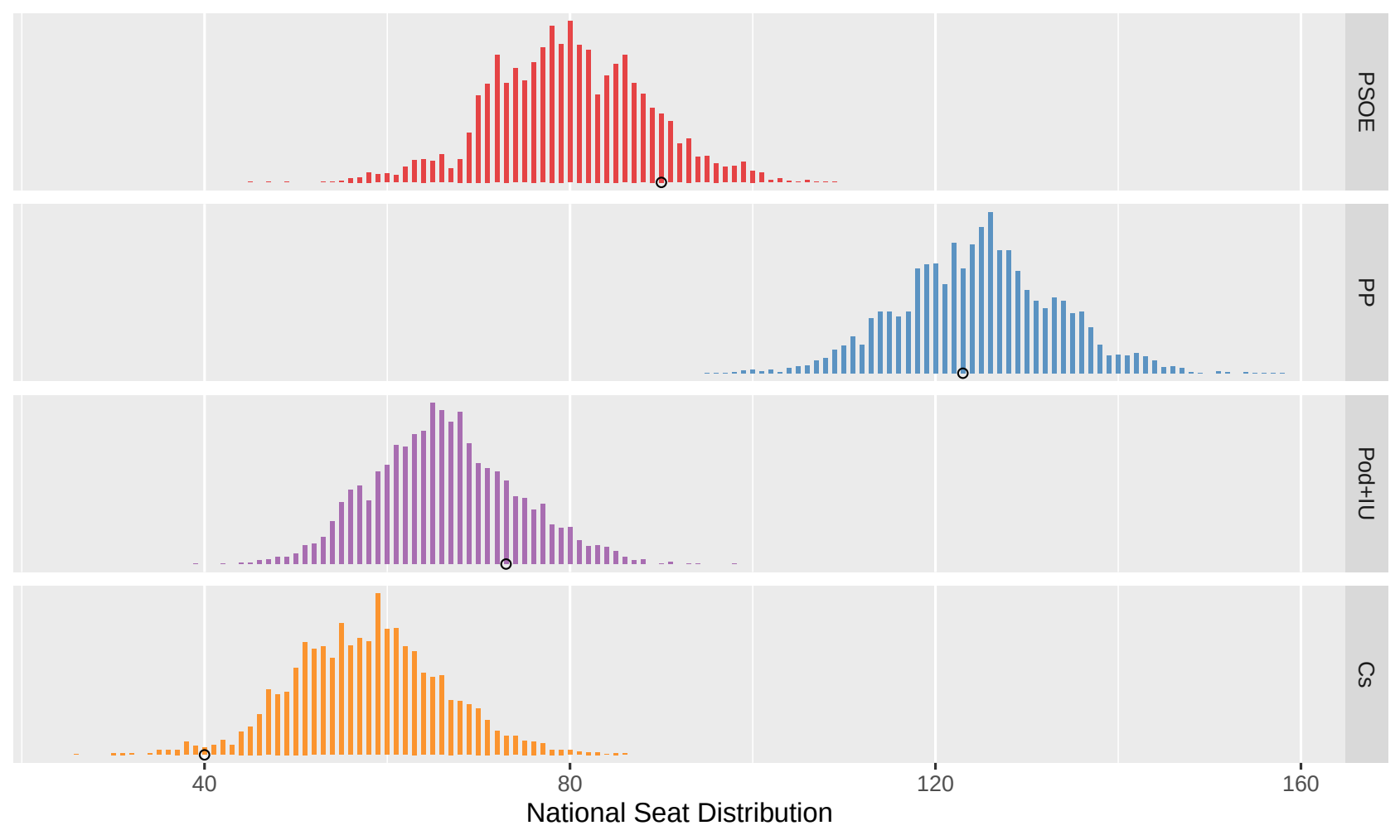


Figure 9: Marginals of predictive seat distribution for 2015 Spanish general election and outcome (dot).

What's next?

- Generate forecast for recent elections with updated methodology.
- Conduct rigorous comparison for different time horizons (t - 5 days, t - 10 days) with competing methods.
- Open questions: Best way to model sentiment process? Induce negative correlation between parties?