When the Winner Comes Third: Simulating Candidates' Winnability With Inaccurate Polls

Daniel Marcelino and Alejandro Tapias

5 de agosto de 2014

Where did the pollsters go wrong?

Candidates	Actual	Average Estimates 3 Weeks *(n=10)*
Russomanno	**18.84**	29.08
Serra	**26.83**	20.84
Haddad	**25.28**	18.25
Others	**23.63**	24.53
Undecideds	–	07.30

Where did the pollsters go wrong?

		Average Estimates	
Candidates	Actual	3 Weeks *(n=10)*	1 Week *(n=5)*
Russomanno	**18.84**	29.08	24.56
Serra	**26.83**	20.84	22.48
Haddad	**25.28**	18.25	19.90
Others	**23.63**	24.53	26.06
Undecideds	–	07.30	07.00

Where did the pollsters go wrong?

	Pollster			
Candidates	Datafolha *n=2*	Ibope *n=2*	Veritá *n=1*	VoxPopuli *n=1*
Russomanno	+	+	+	+
Serra	-	-	*	-
Haddad	-	-	-	-
Others	+	+	+	*

How do political scientists predict elections?

Four traditions are common in the literature

- (1) Economic vote models
- (2) Electoral cycles models
- (3) Models using prediction markets
- (4) Models that use polling data as the primary predictors

How do political scientists predict elections?

Four traditions are common in the literature

- (1) Economic vote models
- (2) Electoral cycles models
- (3) Models using prediction markets
- (4) Models that use polling data as the primary predictors

How do political scientists predict elections?

Four traditions are common in the literature

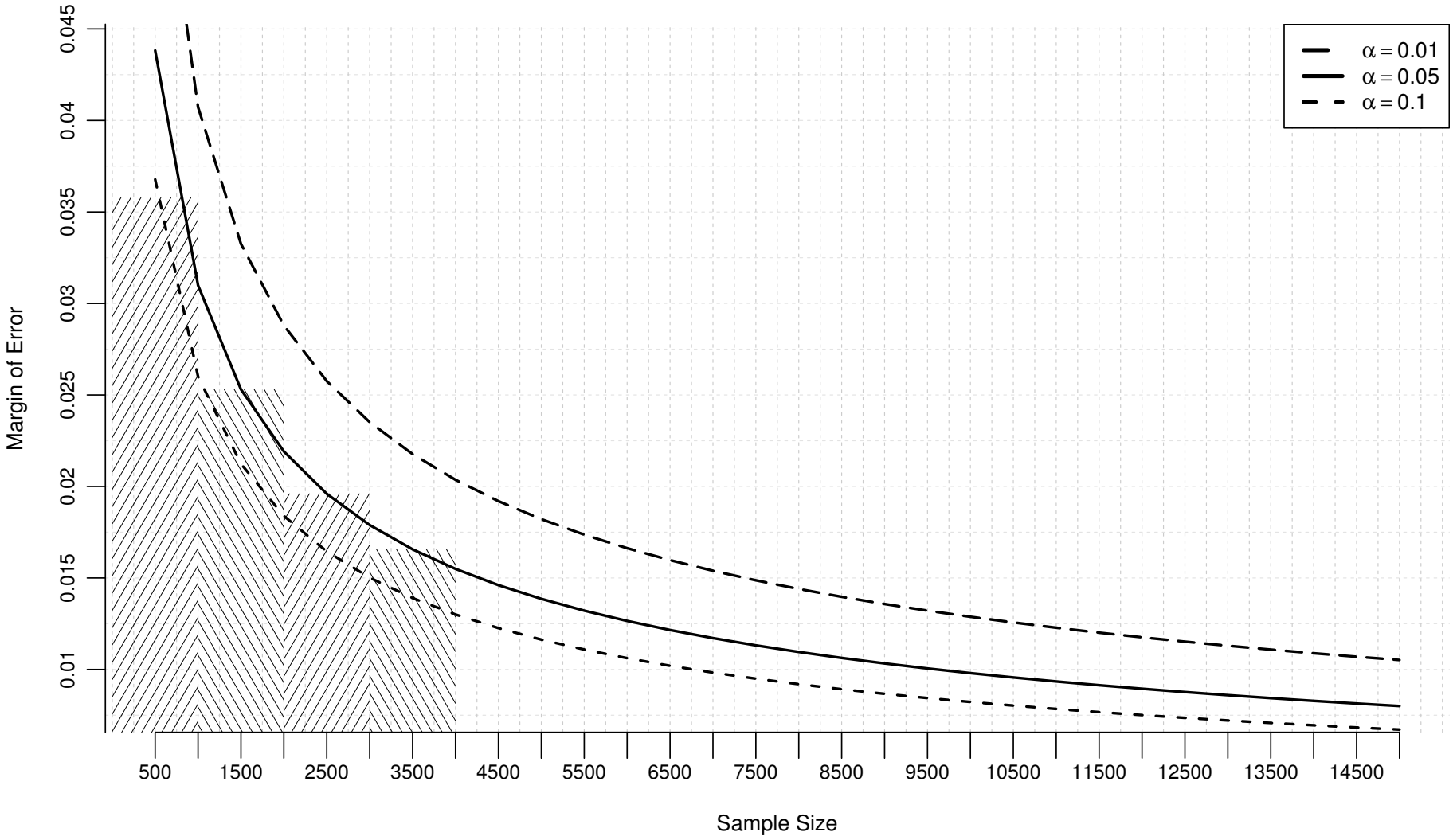
- (1) Economic vote models
- (2) Electoral cycles models
- (3) Models using prediction markets
- (4) Models that use polling data as the primary predictors

How do political scientists predict elections?

Four traditions are common in the literature

- (1) Economic vote models
- (2) Electoral cycles models
- (3) Models using prediction markets
- (4) Models that use polling data as the primary predictors

Political polls: The size matters



A poll is likely to be wrong, yet...

House effects

- Rounding
- Non-response bias
- Wording and Ordering
- Mode bias

A poll is likely to be wrong, yet...

Context: local elections in Brazil

- (-) High number of candidates (12)
- (-) Local elections = polls shortage (28)
- (-) Few pollsters (4)
- (-) Poor sampling designs
- (+) Face-to-face surveys
- (.) Political system features may cause high volatility

How can we cope with irregular and inaccurate polls to fit regular political support?

The Method

Bayesian inference

- The estimand parameters are considered random variables, but these are still related to one another.

Example

If a candidate X at t_1 had 28% of the popular vote, it is very likely that at t_2 he will be alos close to 28% once t_1 and t_2 are close to one another in time. Therefore, if one know θ for X at t_1, this information would change your beliefs about the likely values for X at t_2. Moreover, given this information we would like to know the probability of X winning the election.

The Method

Bayesian inference

- The estimand parameters are considered random variables, but these are still related to one another.

Example

If a candidate X at t_1 had 28% of the popular vote, it is very likely that at t_2 he will be alos close to 28% once t_1 and t_2 are close to one another in time. Therefore, if one know θ for X at t_1, this information would change your beliefs about the likely values for X at t_2. Moreover, given this information we would like to know the probability of X winning the election.

The Method

Bayesian inference

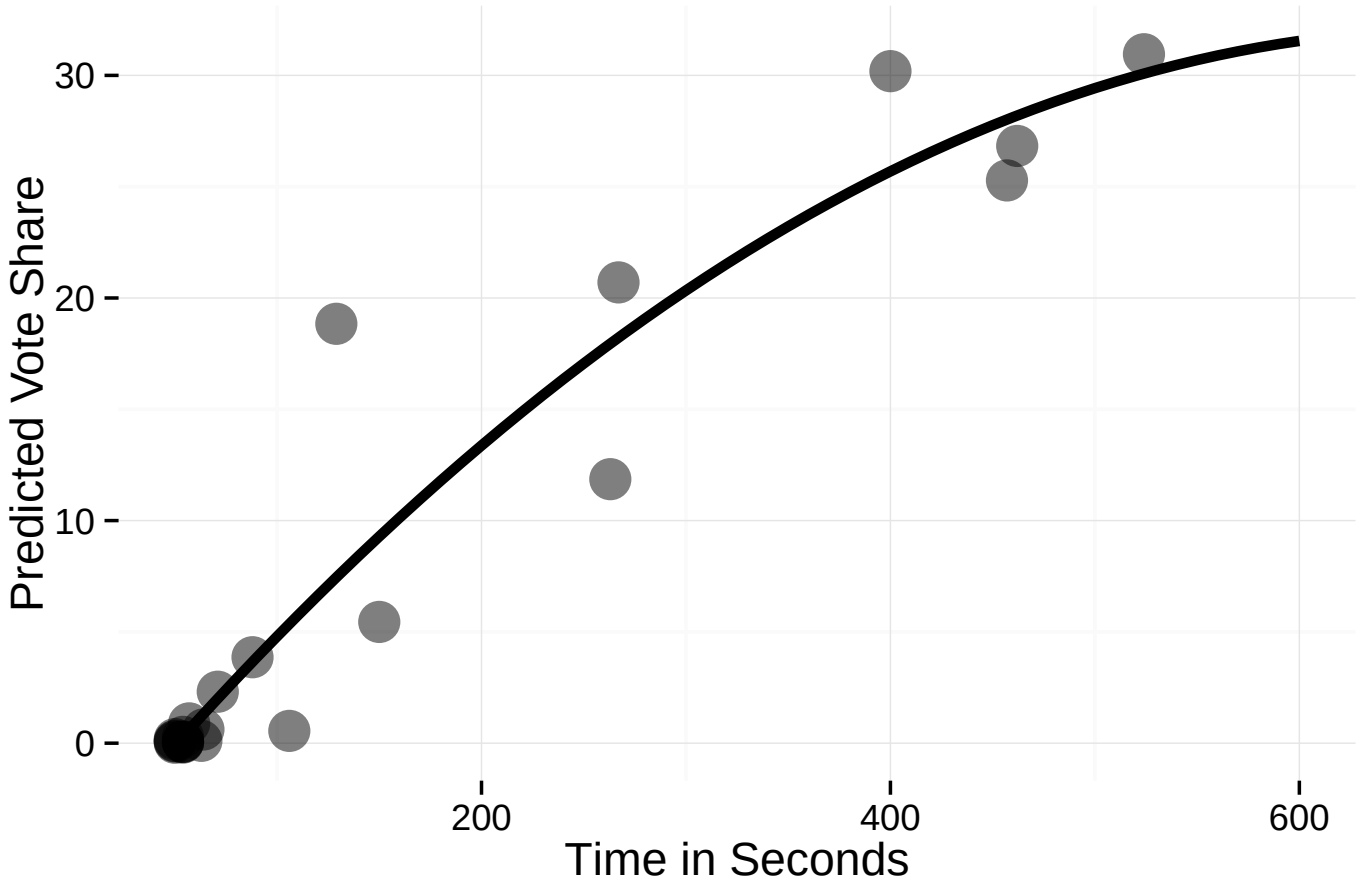
- The estimand parameters are considered random variables, but these are still related to one another.
- Incorporate data from various sources as well as uncertainties associated with the data.
- Prior distribution $\rightarrow$ Posterior distribution.

The Method

Bayesian inference

- The estimand parameters are considered random variables, but these are still related to one another.
- Incorporate data from various sources as well as uncertainties associated with the data.
- Prior distribution $\rightarrow$ Posterior distribution.

Priors: Advertising slots



Prior Distribution

Candidates	α_i	Prior mean	Prior var	Prior sd
Serra	280	0.28	0.101	0.317
Haddad	270	0.27	0.099	0.314
Russomanno	170	0.17	0.071	0.266
Others	280	0.28	0.101	0.317
Total	1,000	1.00		

$$(y_{t1:k}) \sim Multinomial(n, \alpha_{t1:k}).$$

Evidence: Polling data

A poll by Vox Populi with 1,000 voters conducted roughly 15 months ahead the election (13^{th} July 2011) gave this:

- Serra: 26%
- Russomanno: 14%
- Haddad: 2%
- Others: 58%

Posterior Distribution

Candidates	$\alpha_i + y_i$	Posterior mean	Posterior var	Prior sd
Serra	540	0.27	0.066	0.256
Haddad	290	0.15	0.041	0.203
Russomanno	300	0.15	0.044	0.209
Others	580	0.43	0.082	0.286
Total	2,000	1.00		

$$p(\alpha_{t1:k}|y_{t1:k}) \sim Dirichlet(b_{t1:k} + y_{t1:k}).$$

$$p(\alpha_{t1:k}) = \frac{\Gamma(b_{t1:k})}{\Gamma(b_{t1:k})} \alpha_{t1:k}^{b_{t1:k}-1} \ldots, \alpha_{tk}^{b_{tk}-1}$$

The Model

Weighted average

Each poll has its own precision: $p = 1/\sigma^2$.
DataFolha of n=3,959
Ibope of n=1,204

$$\bar{y}^*_{di} = \frac{p_D y_D + p_I y_I}{p_D + p_I} \tag{1}$$

The Model

Predicting vote intentions

Ignorance about θ can be expressed by making the prior precision small. That is, by making prior variance σ_0^2 large.

$$y_i \sim N(\mu_i, \sigma_i^2) \tag{2}$$

The Model

Predicting vote intentions

Given that polls lack precision:

$$\mu_i = \alpha_{ti} + \delta_{ji} + \Delta \tag{3}$$

where δ_j is the bias of polling firm j, an unknown parameter to be estimated. Δ is an unknown parameter to be estimated of event change.

The Model

Predicting vote intentions

To model change in vote intentions, we use a random-walk model as that

$$\alpha_t \sim N(\alpha_{t-1}, w^2), t = 1, \ldots, T \tag{4}$$

where w^2 is a linear interpolation component that detects event discontinuity (before vs. after campaign advertising on TV).

The Model

Predicting vote intentions

With an uniform distribution of prior beliefs, that is before we see any polling data:

$$\alpha_{ti} \sim Uniform(l, u) \tag{5}$$

where l and u denotes lower and upper limits for the range of plausible electoral outcome for a candidate.

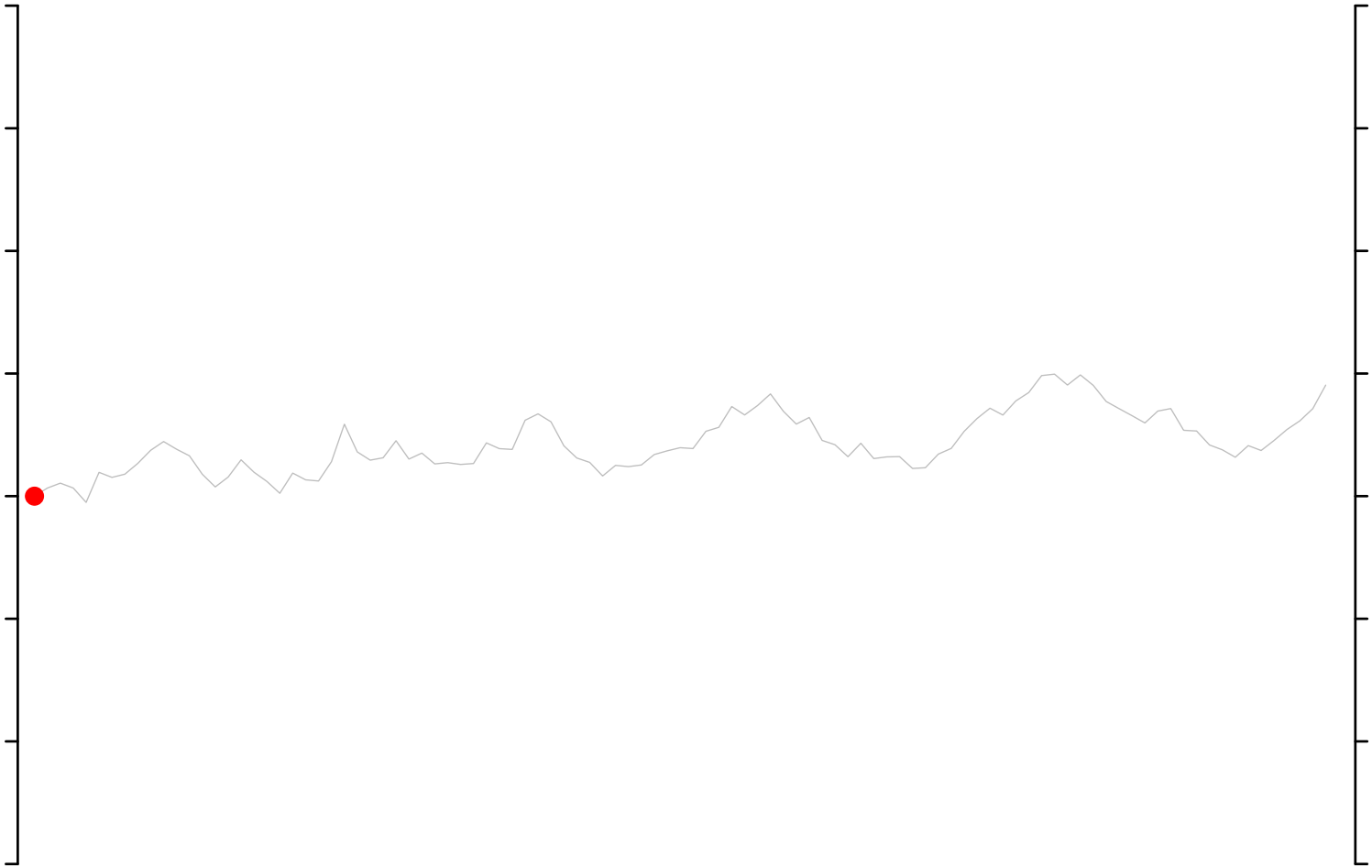
Random-walk model (drunkard's walk)

Candidates are the drunkards

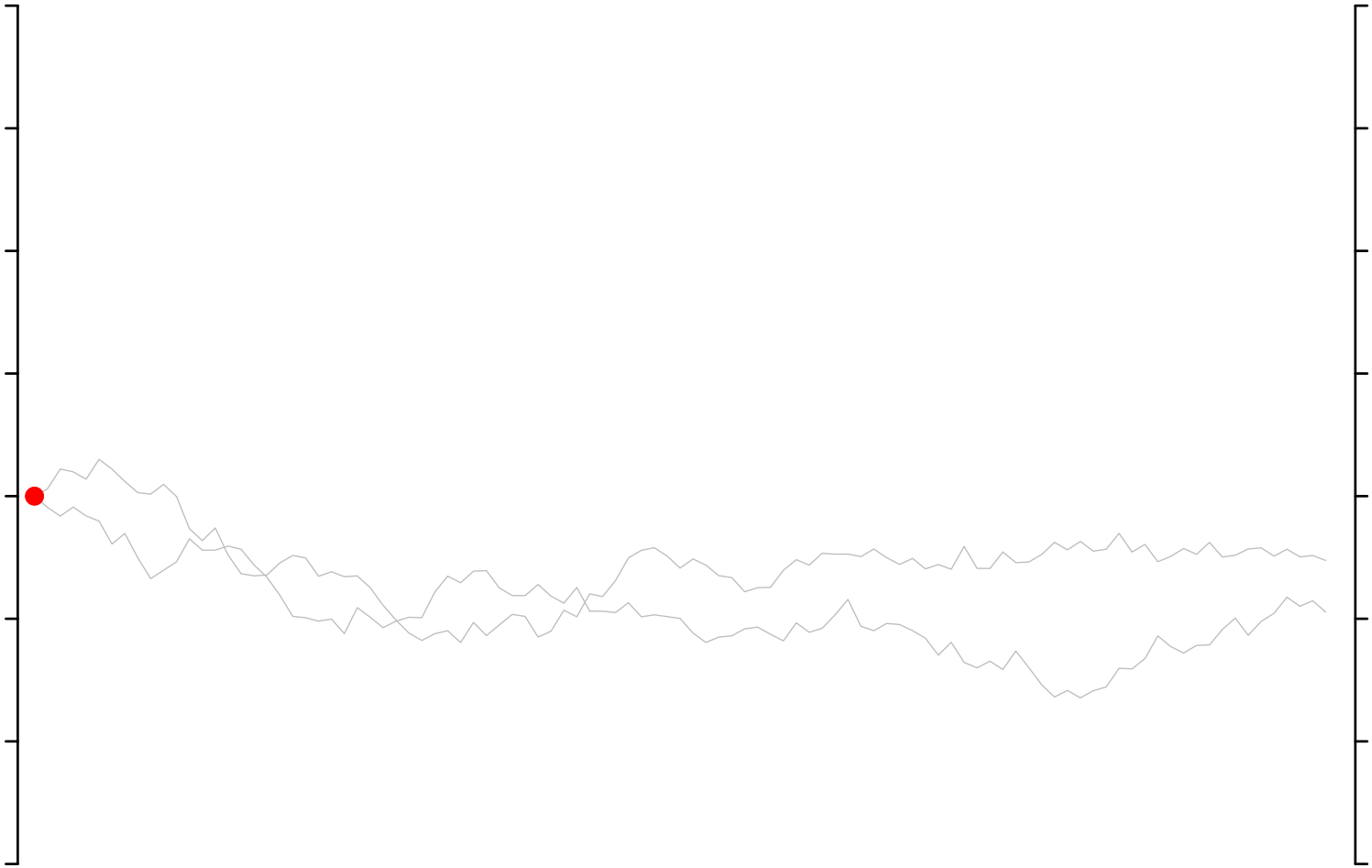
- Stagger left(right) = gain(lose) support.
- Noisy signals = opinion polls.
- Kalman filtering: Learn about likely path given polling data.

We know which bar you left

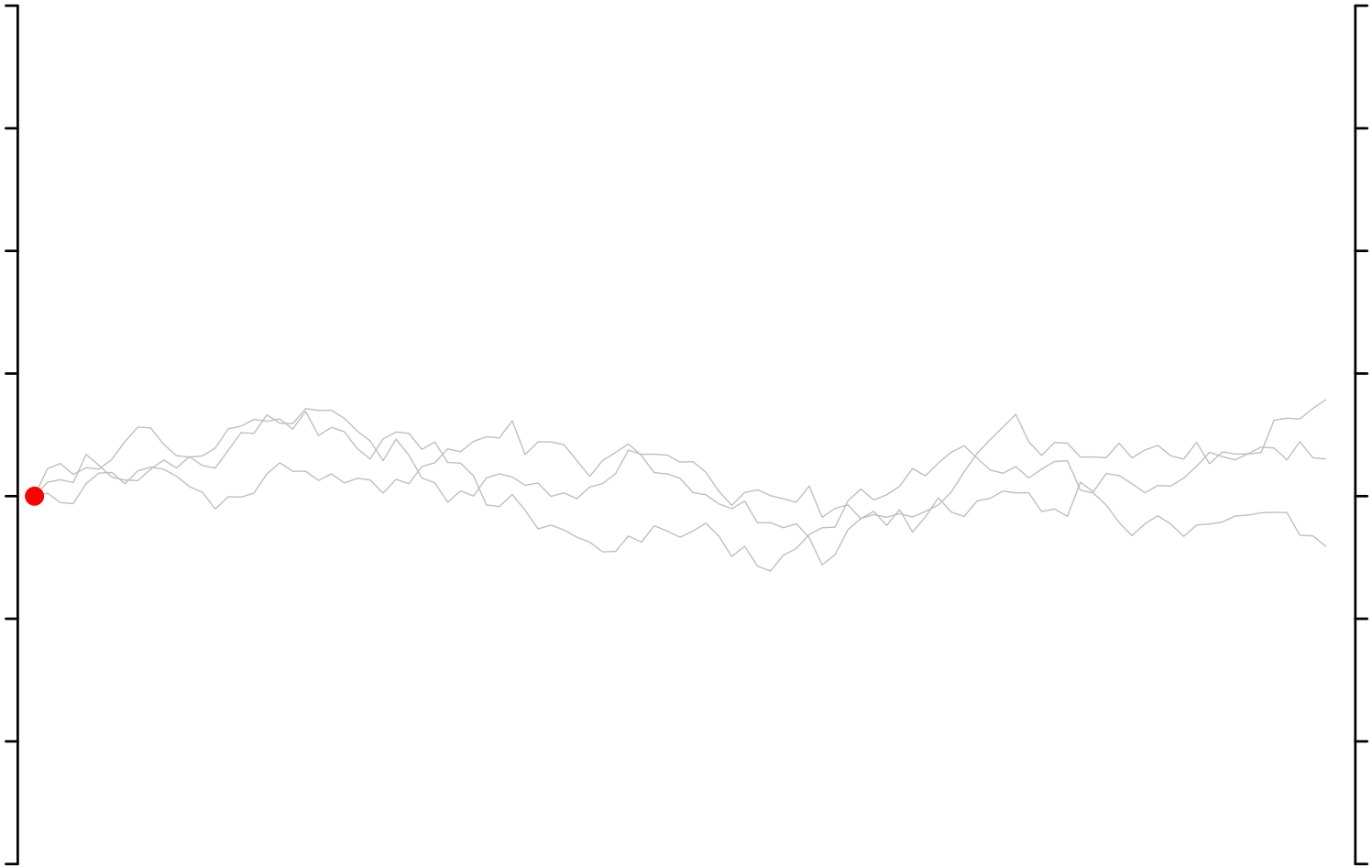
We know the direction of travel



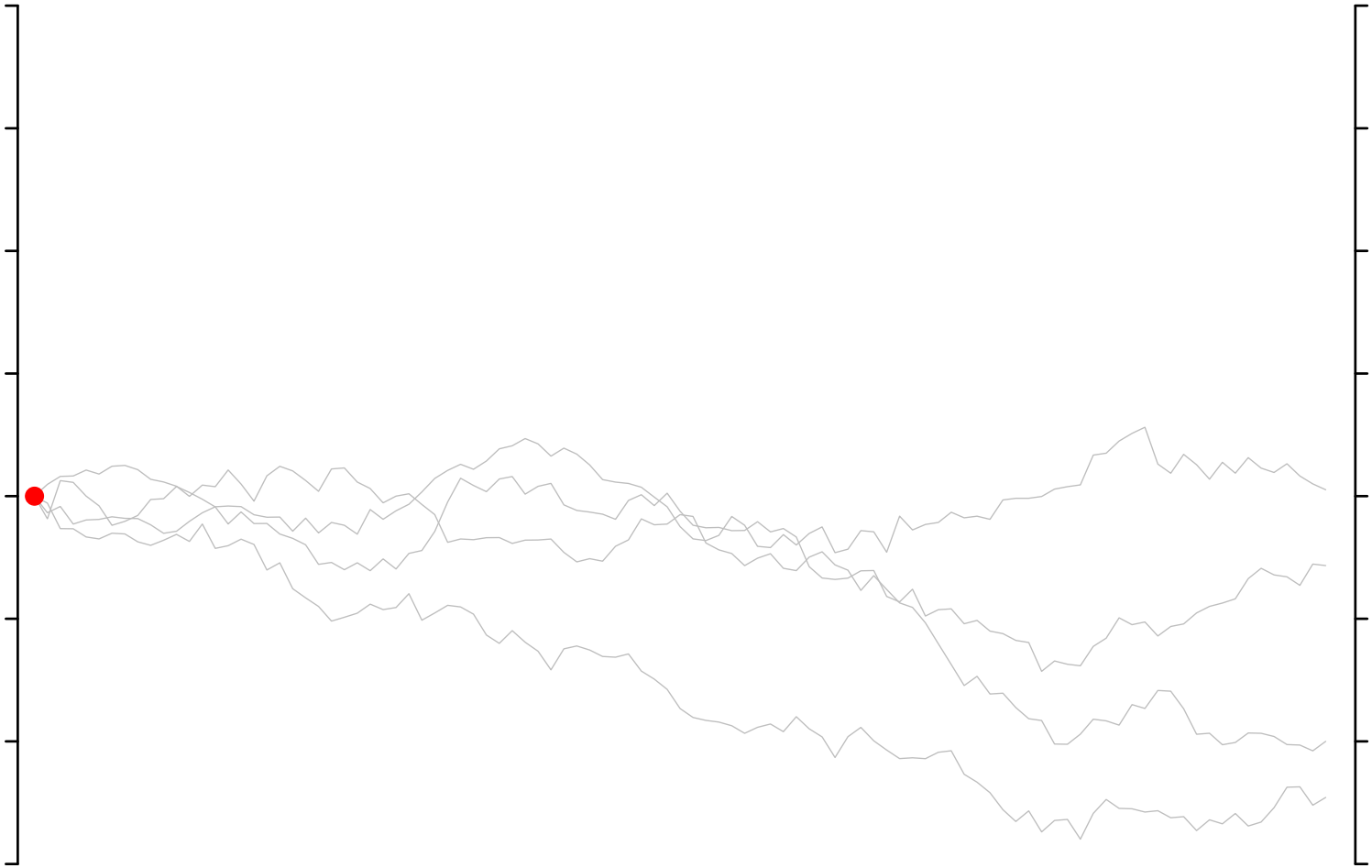
We don't know whether you staggered left or right



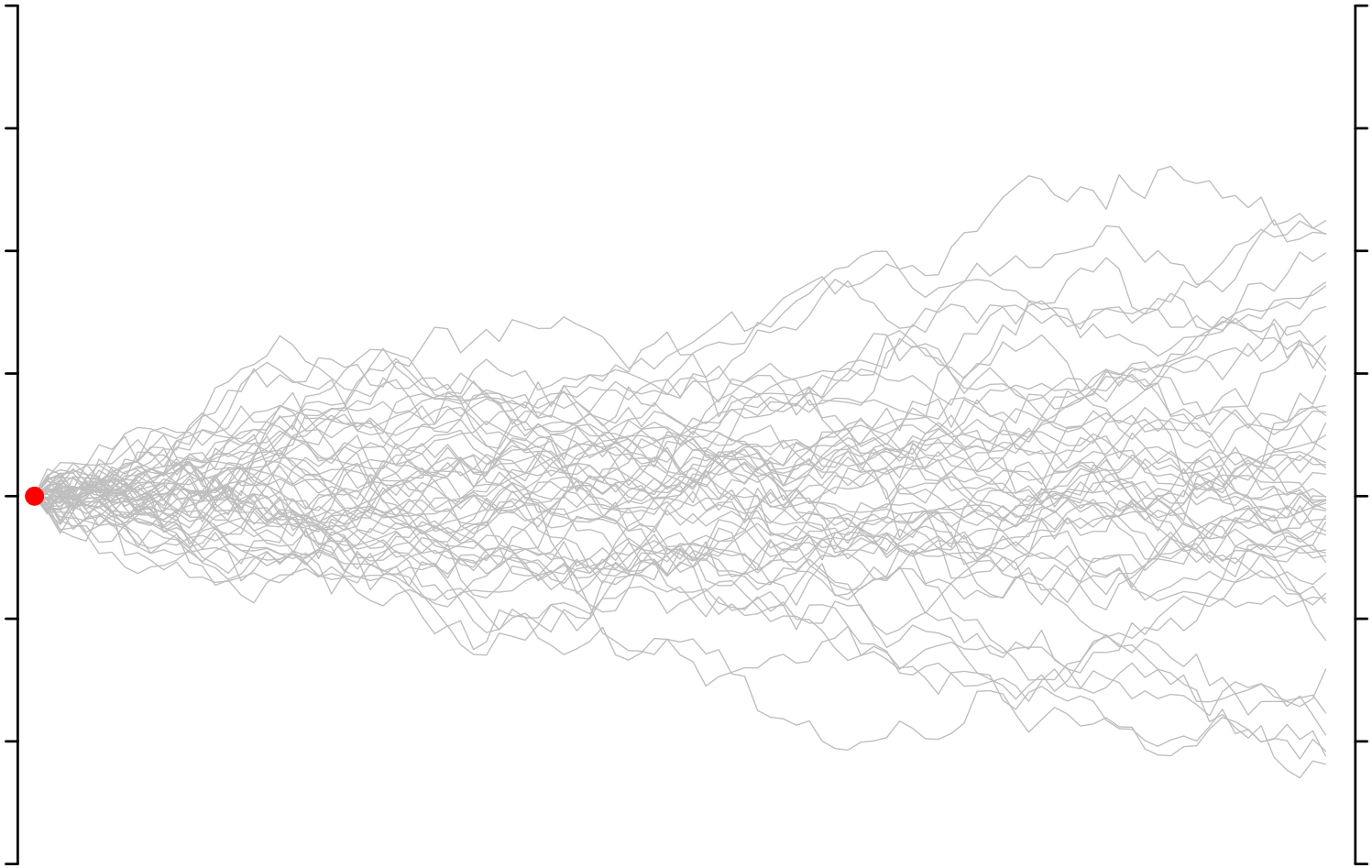
We don't know whether you staggered left or right



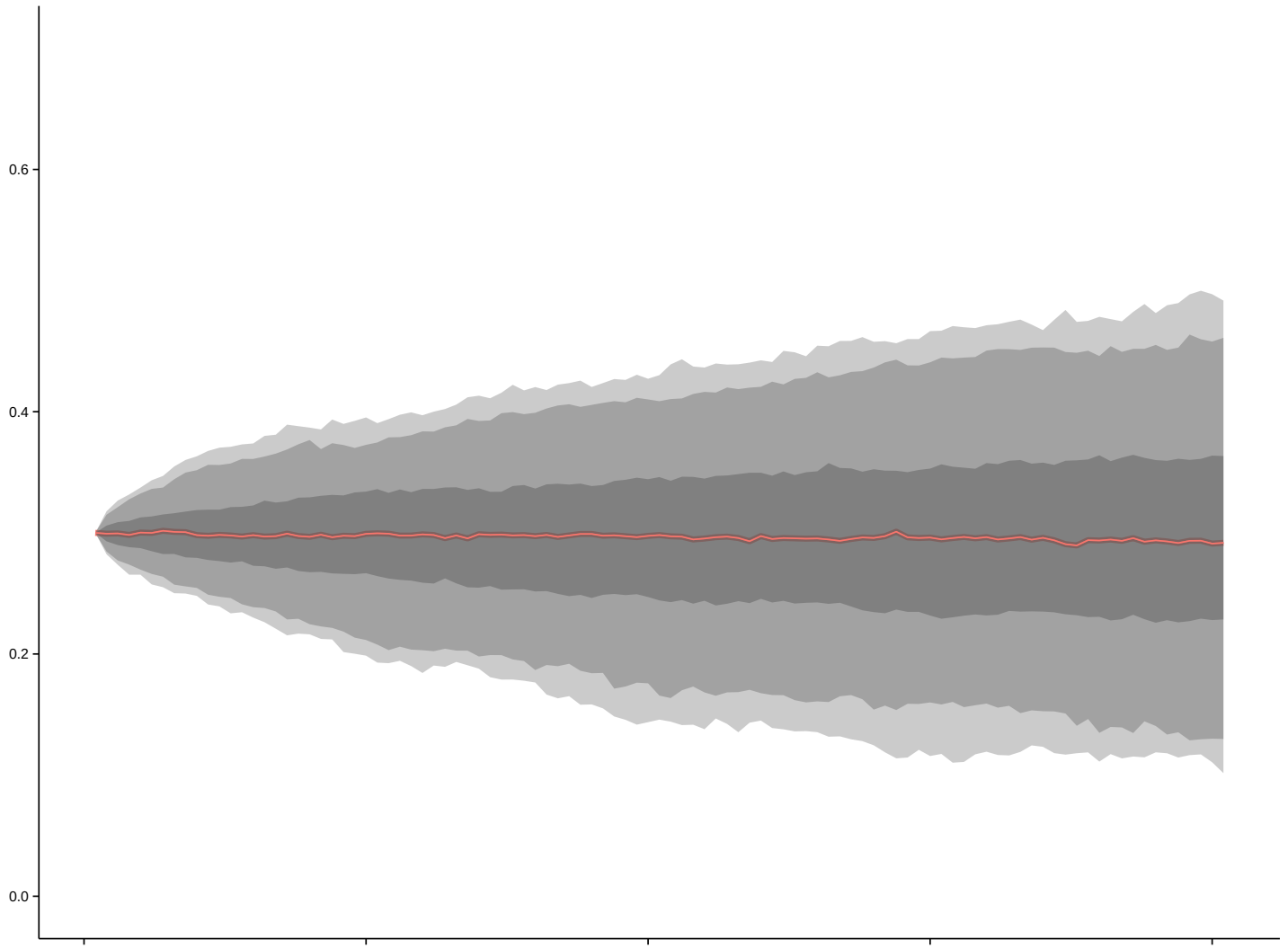
We don't know whether you staggered left or right



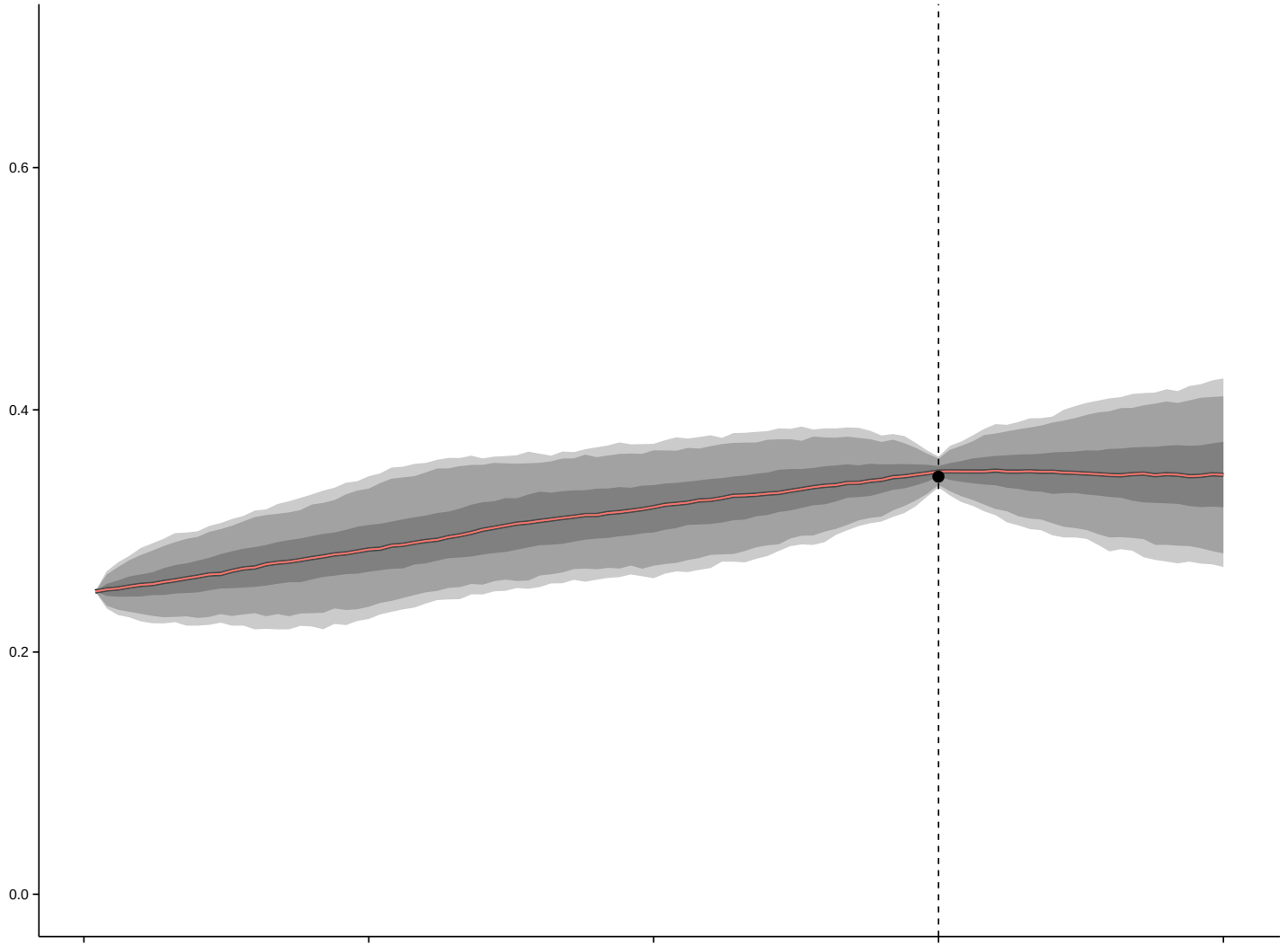
We don't know whether you staggered left or right



We have a belief how a candidate would fare on the election before it takes place



Polls are noisy signals, but we can "learn" about most likely deviations given these signals



Computation

Computation details

- Software WinBugs (OpenBugs)
- The MCMC sampler was run on a single chain with an adaptation period (burn-in) of 100,000 iterations, followed by 500,000 iterations in which every 500^{th} draw was kept for the analysis.
- The resulting data set is a pooled sample of 1,000 valid cases (elections).

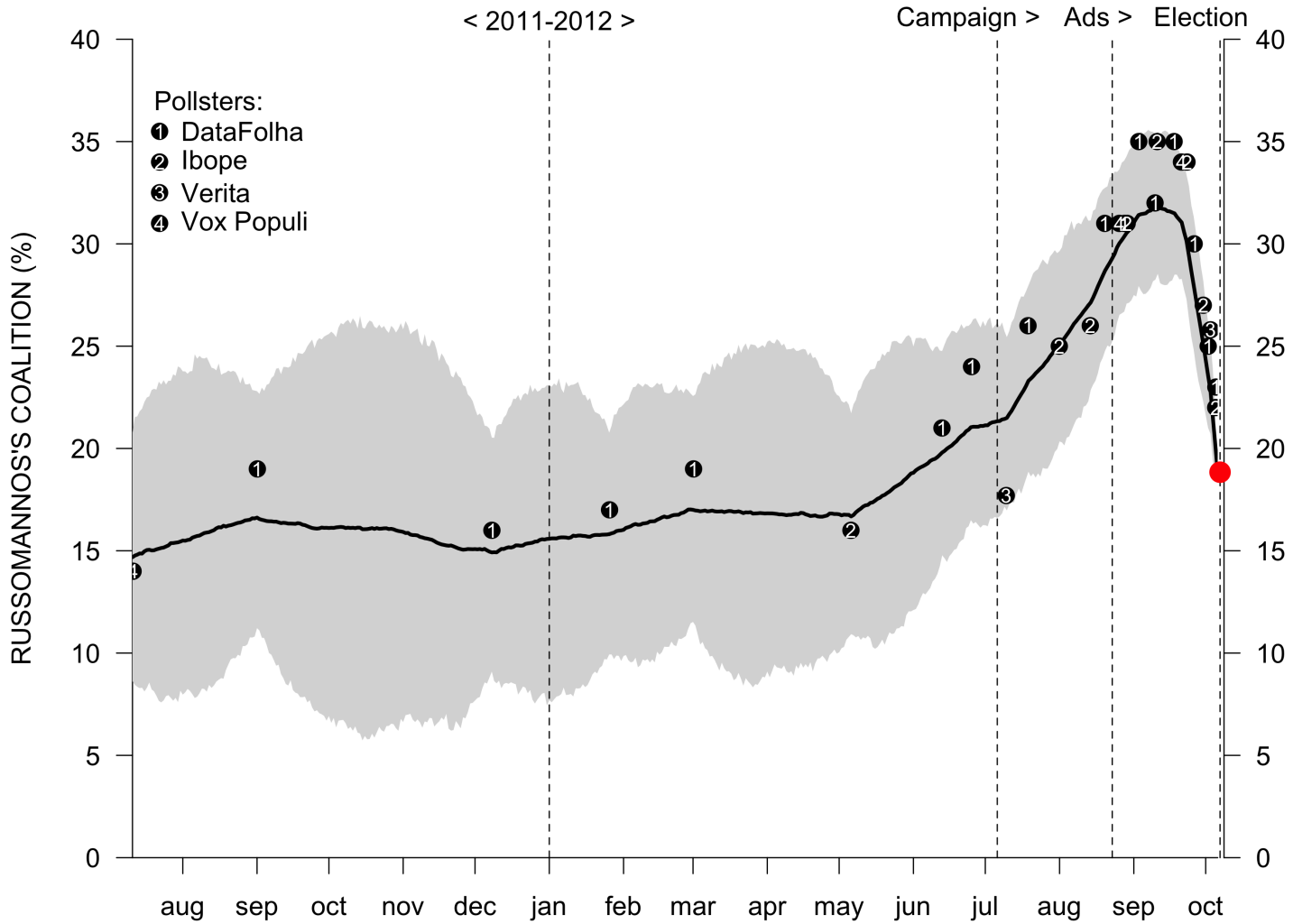
Average estimates for the house effects parameters

	Russomanno (PRB)			Serra (PSDB)			Haddad (PT)		
Pollters	Estimate	2.5%	97%	Estimate	2.5%	97%	Estimate	2.5%	97%
Datafolha	**3.98**	2.00	5.89	**-2.14**	-3.78	-0.39	**-5.40**	-7.09	-3.77
Ibope	**3.51**	1.50	5.69	**-4.52**	-6.34	-2.53	**-4.88**	-6.65	-3.00
Veritá	3.00	-0.09	6.17	-1.03	-3.77	1.97	**-3.60**	-6.16	-1.21
VoxPopuli	**3.37**	0.84	5.52	**-3.58**	-5.56	-1.38	**-4.75**	-6.97	-2.90

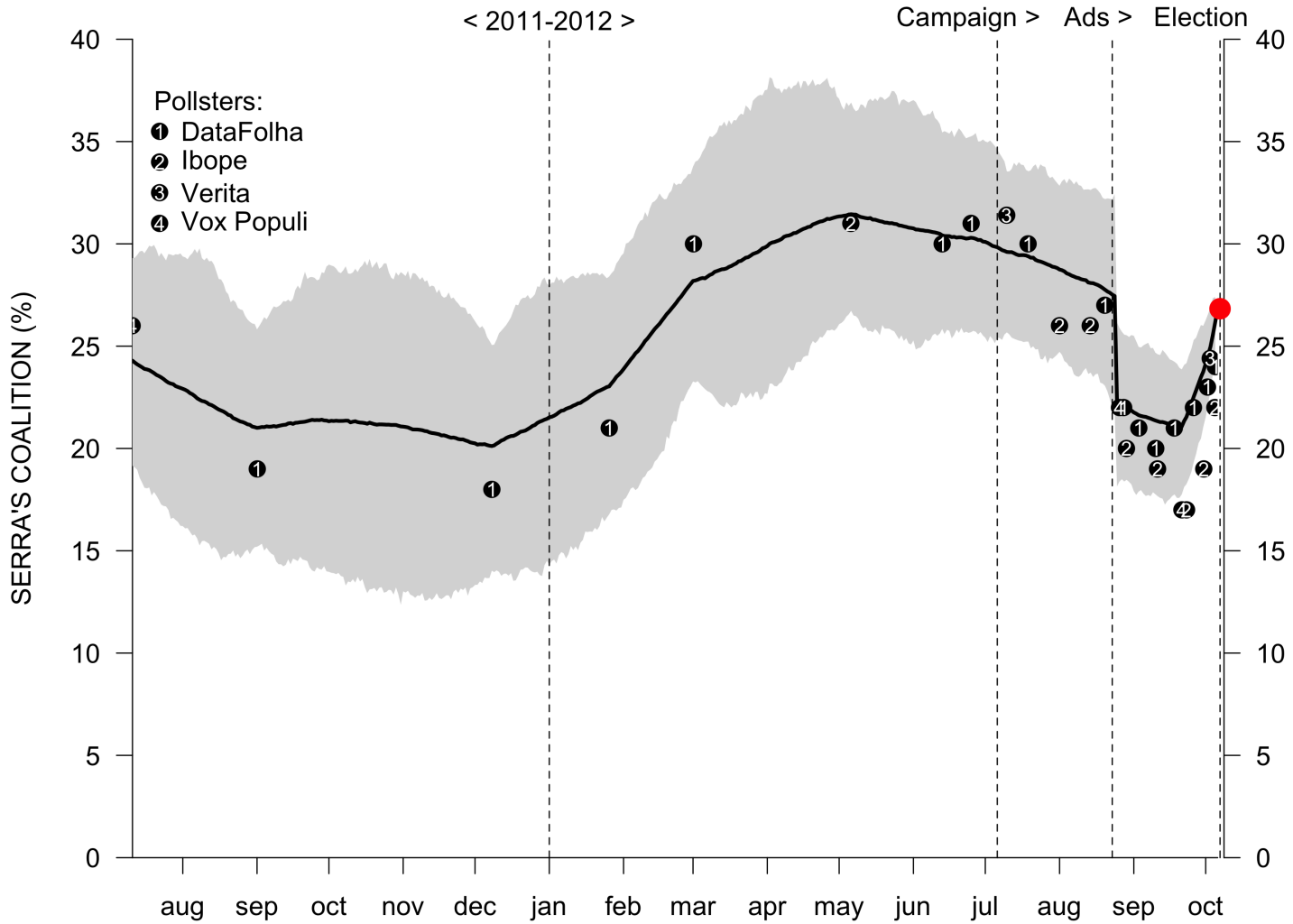
Simulation Results

Candidates	Actual	Average Estimates Last day *(n=1,000)*
Russomanno	**18.84**	20.20
Serra	**26.83**	26.18
Haddad	**25.28**	24.32
Others	**23.63**	24.10
Undecideds	–	05.02

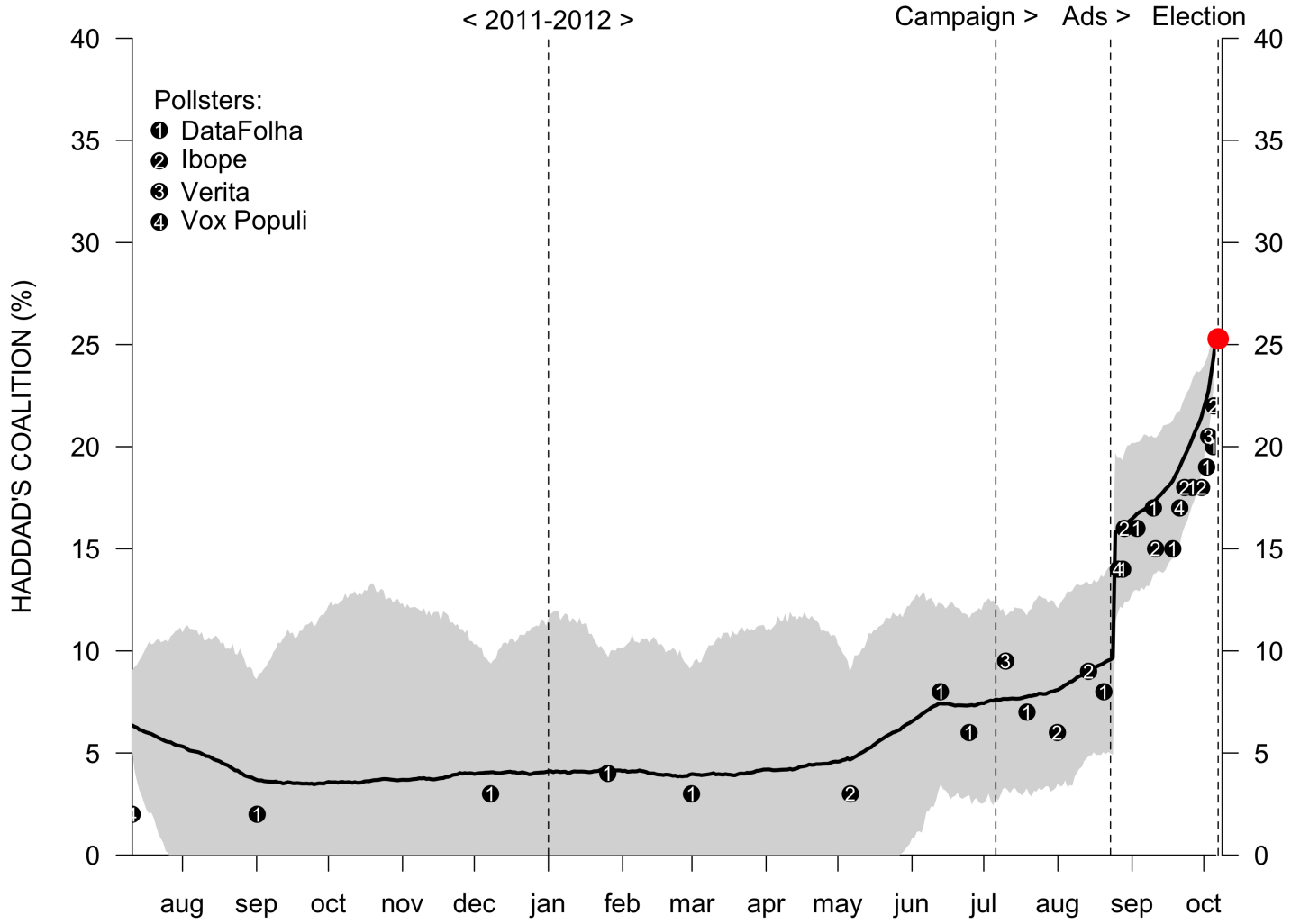
Simulation Results: Share and pointwise for Russomanno (PRB)



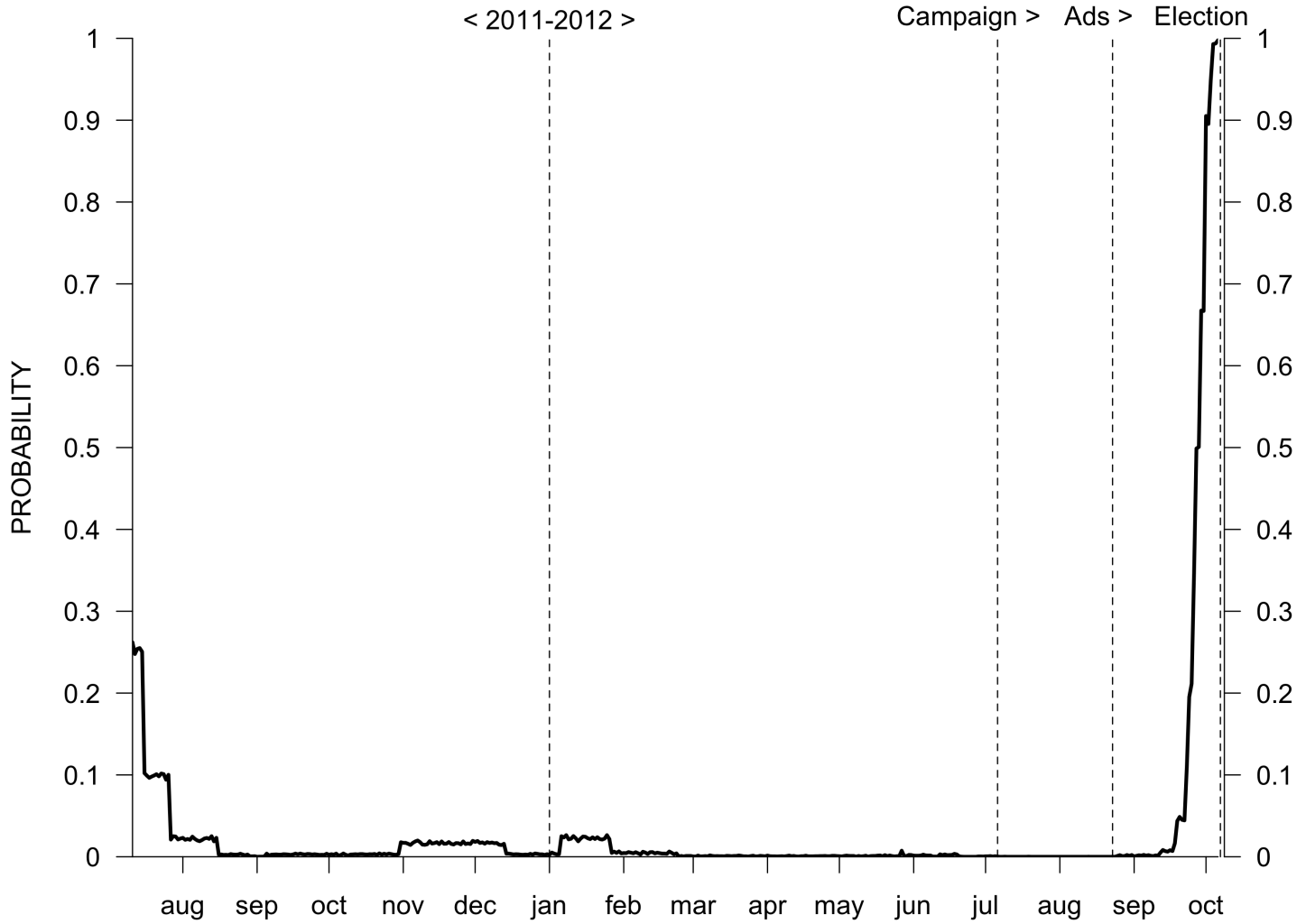
Simulation Outcomes: Share and pointwise for Serra (PSDB)



Simulation Results: Share and pointwise for Haddad (PT)



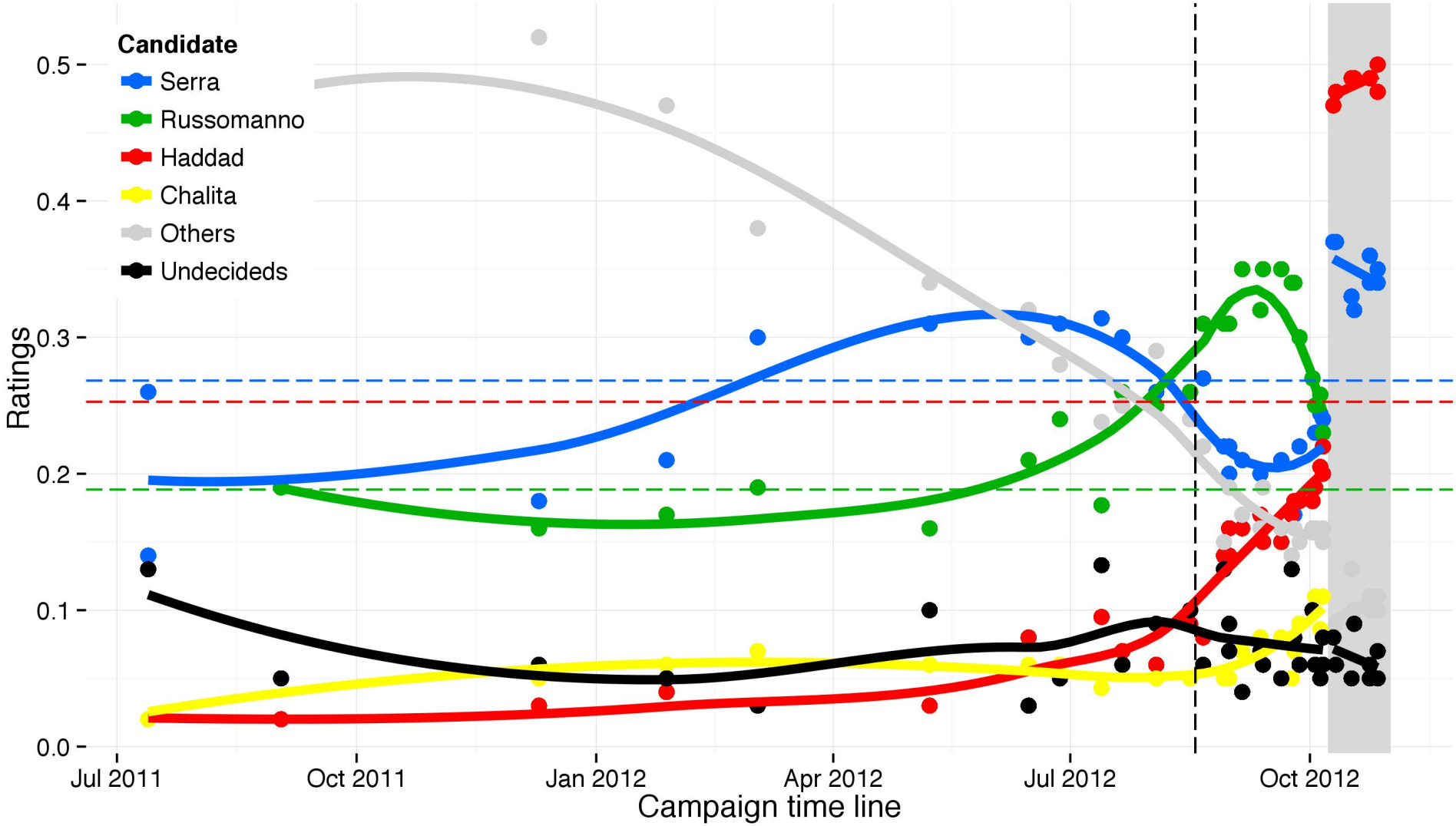
Simulation Results: Probabilities of Haddad (PT) beat Russomanno (PRB) and advance in the runnoff



Conclusions

- In Brazil as everywhere polls lack precision. Precision is mainly affected by two sources: sample size and house effects. After account for them, we could improve the predictions; consequently the information about the election.
- In Brazil, the institution of campaign advertising on TV and radio may cause significant breaks in vote intention, which needs to be modeled accordingly, otherwise, a violation of the linearity assumption may occur.

Where did the pollsters go wrong?



Polls fielded over the last 3 weeks to the election

		Average error				
Candidates	Actual	Mean *n=10*	Datafolha *n=4*	Ibope *n=4*	Veritá *n=1*	VoxPopuli *n=1*
Russomanno	**18.84**	29.08	9.41	10.66	6.96	15.16
Serra	**26.83**	20.84	-4.33	-7.58	-2.43	-9.83
Haddad	**25.28**	18.25	-7.28	-6.03	-4.78	-8.28
Others	**23.63**	24.53	6.09	8.49	7.19	2.23
Undecideds		7.30	5.75	8.00	5.00	13.00

Actual vote is in bold face.

Polls fielded over the week before the election

Candidates	Actual	Mean *n=5*	Average error: Datafolha *n=2*	Average error: Ibope *n=2*	Average error: Veritá *n=1*
Russomanno	**18.84**	24.56	5.15	5.66	6.96
Serra	**26.83**	22.48	-3.33	-6.33	-2.43
Haddad	**25.28**	19.90	-5.78	-5.28	-4.78
Others	**23.63**	26.06	-2.05	-3.05	-4.75
Undecideds		7.00	6.00	9.00	5.00

Actual vote is in bold face.