Forecasting the Lok Sabha: Can Non-Representative Surveys add Value? Working Paper

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk

Raymond Duch
raymond.duch@nuffield.ox.ac.uk

Nuffield College, University of Oxford

1 Summary

We attempt to forecast the results of the 2019 Lok Sabha election in India. Our method proposes to merge information from i) proprietary online surveys, re-proportioned via a stratification frame; ii) historical election results; iii) publicly available traditional polls. The goal is to forecast the number of seats that will be obtained by the National Democratic Alliance (NDA, led by Modis BJP), the United Progressive Alliance (UPA, led by Gandhis Congress) and all other parties and alliances. Our online surveys are administered to two different populations, namely the Centre for Experimental Social Sciences (CESS[1]) India online subject pool, and a set of Amazon Mechanical Turk[2] workers who self-select into taking these surveys.

There are two purposes to this exercise: a) to forecast the election result correctly; b) to evaluate the extent to which online non-probability samples can add information to traditional polls, especially in the context of a notoriously difficult-to-poll case such as India. The headline results follow, whilst a discussion of the methodology, as well as a more in-depth breakdown of the results, are presented below subsequently.

2 Results

Figure 1 presents the results of our forecasting effort. We predict that the NDA will retain control of a majority in the Lok Sabha; the point estimate for the number of seats they will receive is 304, based on the average of error-adjusted traditional opinion polls, and our online surveys. With respect to the UPA, we predict them to achieve 119 seats; 120 seats would go to all other parties and alliances. Online surveys and traditional polls disagree most on the UPA seats, with traditional polls expecting up to 20 points higher than the average. In general, online polls favour the NDA compared to traditional polls. Uncertainty around the average forecast suggests there is uncertainty about the extent of the NDA victory, but never puts it in doubt. On the other hand, traditional surveys (adjusted for historical error) assign reasonably small probability (0.13) to an NDA majority loss.

[1] `https://cess-nuffield.nuff.ox.ac.uk`
[2] `https://www.mturk.com`

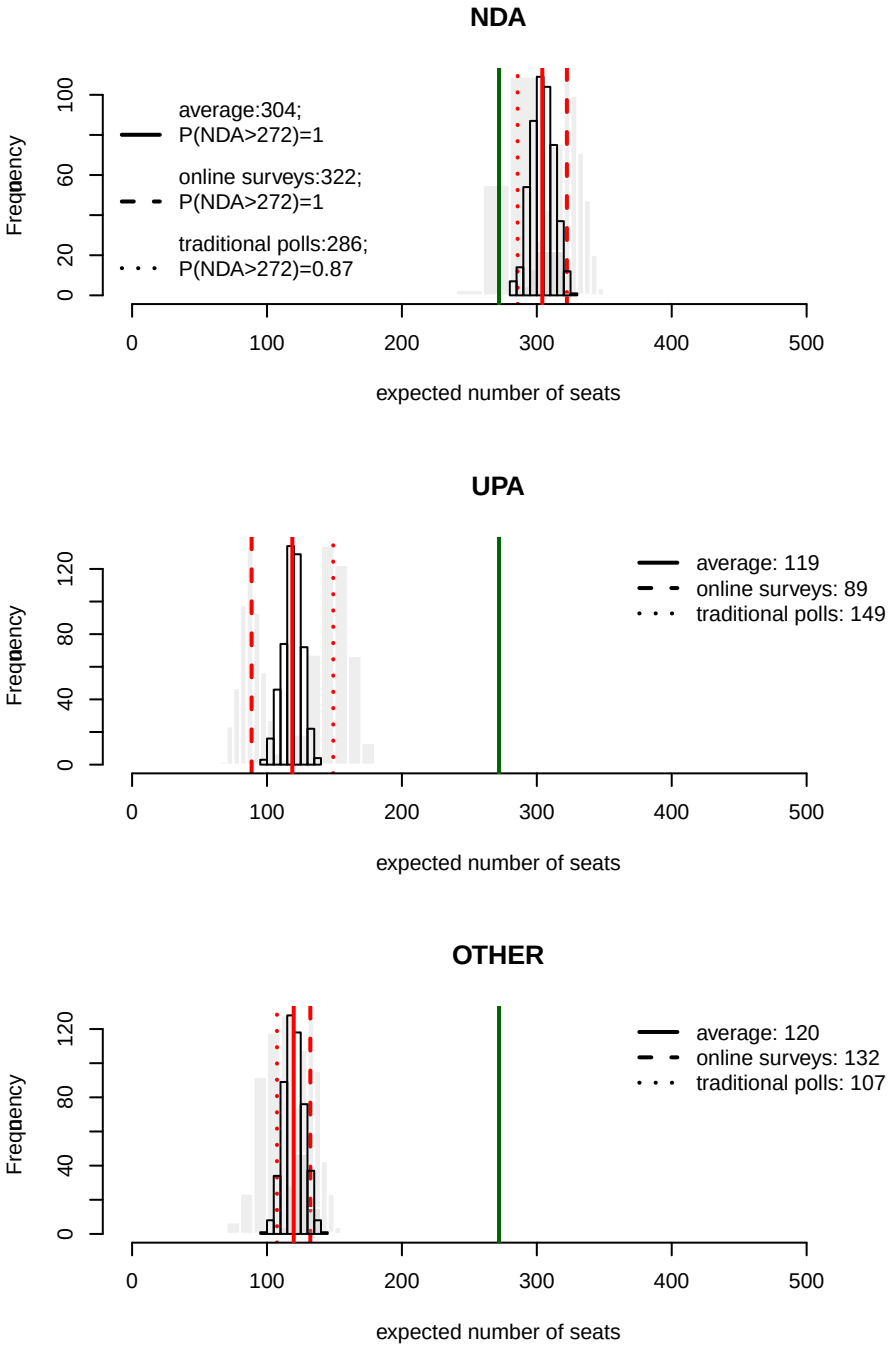

Figure 1: Histogram of predicted seats by party for 500 simulated elections. The faded histograms in the background represent the two components of the forecast, namely our online surveys and publicly available traditional polls. The point estimate is identified by the average of the two sources. The dark green line represents the 272 seats threshold - anything above that gives a majority in the Lok Sabha.

3 Methodology

Our method can be divided into five clearly identifiable steps: i) we estimate the national swing from our online surveys; ii) we re-proportion the swing at the state-level based on the historical relationship between the constituencies within the states and the national swing; iii) we apply this state-level swing to the 2014 election results to obtain 2019 seat estimates; iv) independently, we deploy a Bayesian model to aggregate traditional opinion polls and filter for polling house bias; v) we average online and traditional polls to obtain our forecast point-estimate. In the following paragraphs, each of these steps is described in detail.

3.1 Estimating the National Swing from Online Surveys

We obtain convenience samples of individual Indian voters via online surveys administered through the CESS subject pool and the Amazon Mechanical Turk platform. The subject pool is recruited via Facebook advertisements, and hence we can control its makeup to a degree; however selection effects are such that this must for all intents and purposes be treated as a convenience sample. Participants are asked to provide us with personal characteristics, such as gender, age, educational attainment, income, caste, religion, state of residency etc. Political variables asked include self-reported turnout[3], vote choice[4] and others. We start monitoring on February 20^{th}, roughly 13 weeks before the election date of the 19^{th} of May. Figure 2 shows the sample size per week to election, broken down by voting intention. As of Sunday 12th of May 2019 we have sampled $12,554$ voting intentions, unique by week.

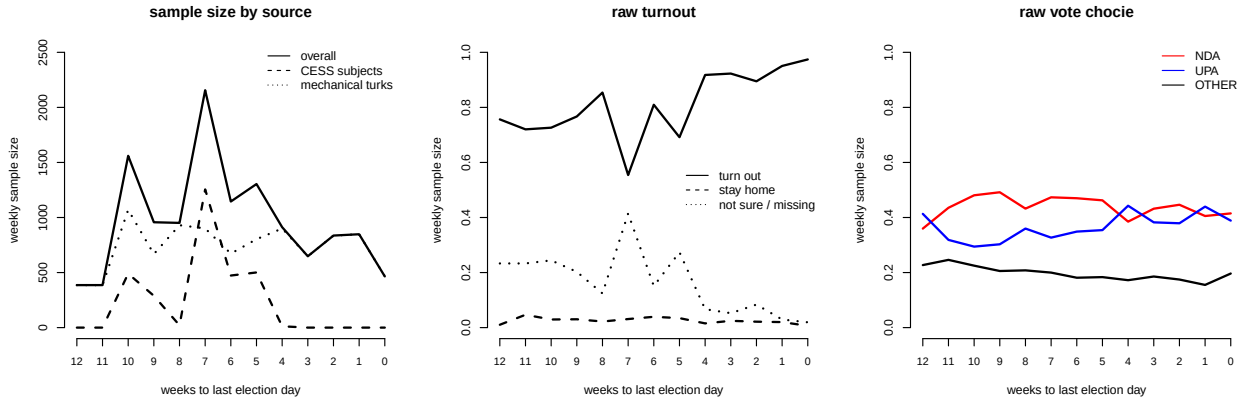

Figure 2: Evolution of the sample over collection period. From left to right: a) sample size by source; b) self-reported turnout percentages; c) vote choice percentages.

[3]Respondents are given the following question: *'This year, the General Election for the Lok Sabha is expected to be held sometime between April and May. Many people have told us they will not vote. How about you, do you think you will vote in the upcoming general election?'*. This question is clearly leading, and designed to curb significant over-reporting. The potential answers include 'Yes'; 'No'; 'Not sure at this point in time'; 'Not heard of upcoming election'. Missing values are treated as akin to 'Not sure at this point in time'.

[4]Respondents are given the following question: *'If Lok Sabha elections were to be held tomorrow, which of the following parties would you vote for?'*. A list of options tailored to the state of residency is then provided, along with write-in options. Missing values here are ignored.

3.1.1 Issues with Online Non-Probability Samples

Two immediate challenges appear: 1) the turnout percentages are too high (this is particularly so for mechanical turks); 2) the baseline vote shares (i.e. the party fixed effect over the campaign) are likely off the mark.

We know the baseline is off by re-proportioning the 2014 results at the constituency level to account for the novel makeup of alliances and party mergers. To summarize: the 2014 vote distribution with 2014 alliances was [NDA = 0.385 ; UPA = 0.232; Other = 0.382]; the 2014 vote distribution with 2019 alliances was [NDA = 0.395; UPA = 0.275; Other = 0.330]. The data used for these calculations, as well as all our historical conjectures and estimations, comes the Lok Dhaba project of the Trivedi Centre for Political Data[5][11]. Given this data, under no circumstances should we expect the 'Other' vote to fall to 20%; nor should we expect the UPA to ever reach 40% in 2019.

Clearly the sample is going to need to be stratified if we are to make credible inference. But successful stratification requires that the vote-choice probabilities of sub-categories of voters be accurately estimated. This is not possible if there is a selection effect into our samples; for instance, if 'Other' voters are significantly less likely to interact with any online medium[6], the voter-categories that are in truth undecided between 'Other' parties and the NDA will erroneously be assumed to 'break' for Modi's alliance. Assuming voter-categories can be sampled in their population proportions conditional on a fixed propensity to interact with the internet, we can simply re-proportion the weekly vote-choice estimates to match the true population proportion, essentially removing this particular assumed form of the selection effect. This being a forecasting exercise, the true population proportions for each alliance are hidden to us, but we know that they will be around the 2014 vote recalculated with 2019 alliances. Hence, to solve the 'alliance baseline' issue in our non-representative sample, for each week we re-sample observed voters, starting from voters of alliances which are most under-sampled, until sample proportions match the 2014 vote shares with 2019 alliances. Any weekly variation in our final stratified estimates will hence be exclusively the product of changing compositions of the alliance electorate, perhaps having 'swing' voter categories becoming more or less likely to vote for a given party. We perform the re-sampling exercise independently for `CESS` subjects and mechanical turks. To limit the degree to which our sample is reliant on the re-sampled data, as well as to avoid issues when target magnitudes would require to overwhelm the weekly data with re-samples (i.e. if we only observed 10 subjects on a given week but the target has 2 decimal places and would require several hundred subjects to be met to these decimals) we set a limit so that the re-sampled proportion of the new dataset can be at most 1/3 (or half of the pre-re-sample dataset). A sensitivity analysis to this procedure will have to be performed to understand the optimal amount of re-sampling, but for now the 1/3 figure is heuristically motivated and provides credible results.

We now focus on solving the issue of extraordinary turnout in our non-probability sample. There are two

[5] `http://lokdhaba.ashoka.edu.in/LokDhaba-Shiny/`

[6] We know this to be true in part because 'Other' parties tend to be caste- and tribe-based, and especially these largely rural voters would have lower access to the internet than other voters

sub-issues here that we must delve into: i) what to do with individuals who are unsure (or do not explicitly say 'Yes' or 'No' but leave the question unanswered); ii) how to correct for over-reporting. In solving (i) we opt to simulate, for each voter whose turnout is unsure or missing, either a 'Yes' or a 'No' with equal probability; this allows us to keep into the sample as large as possible a variety of unique voter categories, and not waste useful information for the purposes of stratification (albeit by imputing a 0 or a 1 at random we'll be shrinking that voter category's propensity to turn-out toward 0.5, something the effects of which will have to be the subject of a sensitivity analysis). Issue (ii) is tricky, in that we expect our sample to be high-turnout due to selection effects (individuals with access to the internet, Facebook accounts or working as mechanical turks will tend to have higher income, education and caste than the average member of the Indian polity). However, the issue of turnout over-reporting is well known; in the US Jackman et al.[10] estimate over-reported turnout in the American National Election Study to be around 13.5% and Pew[8] finds their online panelists over-report turnout to the tune of 17 percentage points compared to the voting age population. We are not aware of any estimate of India-specific propensity to over-report, so we use the 13.5% figure from Jackman et al. (as we suspect the 17% found by pew is at least partly reconcilable with selection effects into their sample, and we have already adjusted for these via re-sampling). Armed with this number, we randomly 'switch off' voters who state they will turn out, until we reach a drop of 13.5% points.

The resulting composition of the re-sampled and turnout-corrected data we will base our predictions for the National Swing on is available in Figure 3. There is a fair bit of overlap between the samples; mechanical turks and Facebook recruited subjects are biased in roughly the same way: they are more male, more young, more wealthy, incredibly more educated, more upper caste and, finally, more from the south (this is particularly so for the mechanical turks, the vast majority of whom come from the state of Tamil Nadu). Assuming we have sampled enough individuals from the relevant voter categories, the pre-stratification bias in the distribution of the sample across these characteristics will not be a problem. The geographical unevenness of the sample may be a problem if it is correlated with turnout or vote choice.

3.1.2 The Stratification Frame

To correctly execute the stratification procedure, we need reliable counts of the number of individuals for each of the sub-categories of voters we are stratifying by. This is difficult to obtain in the context of India; the census cross-tabs which are available are not exhaustive, in that they do not provide us with a satisfactory set of cross-tabs for the variables we think a-priori to be important in this election (more on this later). Moreover, the census counts are old, having last been published in 2011. As such we complement census counts with individual level data from the second wave of the India Human Development Survey[3] (IHDS) a nationally representative survey of $135,986$ voting-age individuals from $42,152$ households, conducted between November 2011 and October 2012. The issue with merging these datasets is that the IHDS is at the individual level, whilst the census cross-tabs are aggregated at a level which is usually unsatisfactory (in that it aggregates over variables we would like to stratify by, making it prohibitive to obtain these counts).

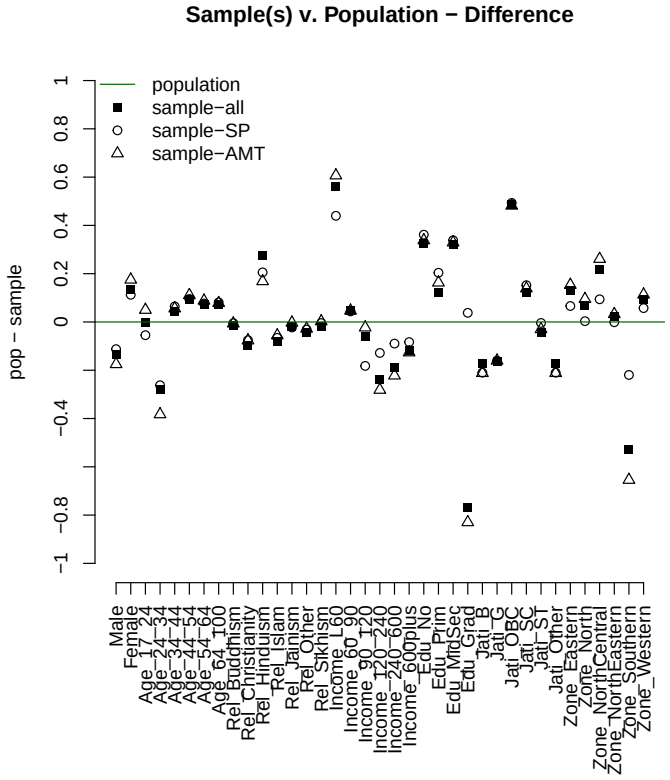

Figure 3: Differences between our population frame and the re-sampled, non-probability samples, in percentages. The percentages that these $pop - sample$ difference are calculated from sum to one by category (i.e. for gender, % male and % female sum to 1; similarly for income and education categories, etc.). Above the dark green line we are under-sampling; underneath it we are over-sampling.

To get around this problem, we deploy the following strategy:

i) we identify the variables for which we want cross-tabs, namely Gender (2 categories), Age (6 categories), Religion (7 categories), Caste (6 categories), Income (6 categories) and Education (4 categories);

ii) we sample at random M^c individuals from the most suitable (i.e. the one that provides counts for the 'deepest' interactions[7].) census cross-tabs file[8]; each individual is sampled from a given unique group in the census tables with probability equal to its groups' proportion within the population;

iii) we stack up the M^i individuals from the IHDS with the M^c individuals sampled from the census; the census sample will lack many characteristics which are fully available in the IHDS (due to the census limitations in the variables cross-tabs); we treat those as 'missing values'. This creates a new individual-level dataset, with $M = M^c + M^i$ rows;

iv) we solve the missing data problem assuming the missing individual level attributes are are 'Missing at Random', and impute these via a random-forest classifier implemented via the packages `ranger`[23] and `missForest`[19]. These algorithms leverage the well known multiple imputation with chained equations[22] paradigm, in combination with random forests[2] which provide exceedingly efficient (as

[7]The best source we find has [Gender, Age, Education] cross-tabs
[8]Indian census cross-tabs are available at `http://www.censusindia.gov.in`

these forests can be grown in parallel) and flexible non-parametric method to impute mixed-type data; [vi)] finally, we count the number of members in each category within this completed dataset; these counts will make up our stratification frame. We will refer to voter categories interchangeably with 'cells', meaning cells within the stratification frame.

We note that for the above strategy to work, the implicit covariance between the many individual-level variables needs to be accurately estimated. In particular, the above strategy relies on the fact the the IHDS contains all the relevant information to estimate in probability the Income, Religion and Caste of an individual of a certain Gender, Age and Education. Another relevant point is the tuning of parameter M^c, the size of the random sample from the census. This should reflect the degree of trust we place in the census data; if we are indifferent between census and IHDS (or any other survey researchers may want to use), we could choose $M^c = M^i$; however if we think the available cross-tabs are more precisely estimated within the census, then we may want M^c to be very large relative to M^i, effectively tuning the stratification frame to be approximately equal to the census for the available cells, and exploiting the estimated individual-level covariance from the IHDS to impute the distribution of the augmented cross-tabs. This is our preferred option, sampling $M^c \approx 7$million . The sensitivity of our final results to this choice of parameters needs to be further assessed; the success of the imputation strategy will also be subject of further scrutiny.

This procedure results in a stratification frame which in principle contains $12,096$ cells. However, many of these cells are of size 0, representing voter categories that do not exist in the Indian population, or they are so rare that they did not make it to the stratification frame due to sampling. As such we are left with $6,611$ voter sub-categories. The acute observer will have noticed that our online sample only contains $12,554$ weekly-unique voters, averaging at about two respondents per cell, which speaks to a prohibitive level of under-powering going on under this set up. The problem of power is also why we opt to exclude state or regional effects, as these quickly blow-up the number of cells, making it impossible for us to estimate voting and turnout intentions for all categories with reasonable certainty.

Fortunately for us, whilst a small proportion of the cells is extremely populated, the vast majority contains only a handful of citizens voters. Of the $6,611$ voter groups in the stratification frame, the most populous 300 make up 82% of the eligible population. The modal profile amongst these large groups [Male, 25-34 years old, Hindu, earning less than 60,000, having completed Middle school, from a Backwards Caste]. These most populated cells involve low income people, which makes the skew of our sample towards higher incomes potentially problematic; only $1,857$ responses from our sample come from members of there top 300 groups. That said, though the full set of category interactions may be rare in our samples, we have many representatives of sub-interactions; for instance, though we many not have many poor people in the sample, we have plenty of young male Hindu middle-school graduates, suggesting that though we many not be powered to estimate the full breadth of interaction effects, we should still be able to make reasonable inference on the relevant groups by exploiting single-variable effects and relevant sub-interactions. Further to aid our cause, we know the poorest in India, which make up a large amount of these large cells, are

significantly less likely to vote[18]. Hence adjusting for the propensity to turn out should further enhance our predictions.

3.1.3 National Swing Estimation

We use a random forest to smooth our sample estimates for turnout and vote choice predicted probabilities for all sub-categories of interest. We opt for a random forest, as opposed to the more classic multilevel regression which is used in many MRP applications[16, 21, 12], due to: a) its efficiency and speed (random forests used with the `ranger`[23] package can be ran in parallel and require only a matter of minutes to analyze millions of data points on a standard machine); b) its ability to estimate complex interaction effects, but still avoid over-fitting thanks to bagging in the estimation procedure and averaging over trees[1]. The specific class of random forests we deploy is a probability machine[15], namely a random forest which calculates probabilities, as opposed to classes, in its terminal nodes. Uncertainty is estimated using the MSPE2 procedure developed by Lu[14], which is an attractive non-parametric estimator of mean squared prediction error, as unlike the Infinitesimal Jackknife[20] (whose properties - consistency, efficiency, etc. - have not been studied in the context of probability machines, to the best of our knowledge) it does not require any further bootstrapping, and exploits the out-of-bag samples from the bagging procedure to calculate the local out-of-bag accuracy of the estimated probabilities. A Normal distribution around the point-estimates provided by the forests is then assumed, with MSPE2 as the variance.

The above procedure can be summarized as follows: for each voter $i = \{1, ..., N\}$ in the training set (our online sample), we predict probabilities of interest by constructing a random forest with $B = 500$ trees, each tree predicting terminal nodes τ_b; two separate forests are ran for turnout $\left(\varphi^T\right)$ and vote choice $\left(\varphi^V\right)$; weights are applied to the vote-choice forest equivalent to the probability of turnout for individuals in the training set, so that the estimated vote distribution is conditional on turnout. The turnout model for each voter in the training set can be described by the following conditional specification:

$$\hat{\mathrm{P}}_i\left(T = 1|\boldsymbol{x}\right) = \varphi^T\left(\boldsymbol{x}\right) = \frac{1}{B}\sum_b^B \tau_b^T\left(\boldsymbol{x}\right); \tag{1}$$

from which we derive the predictive distribution for the voting group of interest $g = 1, ..., N_g$:

$$\mathrm{P}_g\left(T = 1|\boldsymbol{x}\right) \sim \mathrm{N}\left(\varphi^T\left(\boldsymbol{x}\right), (\hat{\sigma}^T_{\mathrm{RMSE2}})^2\right); \tag{2}$$

where $\boldsymbol{x}$ represents the cell characteristics described in the previous paragraphs. From the predictive distribution of turnout, we obtain the turnout weights for all individuals in the training set, $\hat{\mathrm{P}}_i\left(T = 1|\boldsymbol{x}\right)$, which will be used to estimate vote choice conditional on turnout; we also predict turnout probabilities for all the cells in our stratification frame, $\hat{\mathrm{P}}_g\left(T = 1|\boldsymbol{x}\right)$ for the purpose of predicting overall turnout (post-stratification).

Similarly, the vote choice model can be summarized as follows:

$$\hat{\mathrm{P}}_i\left(V=j|T=1,\boldsymbol{x}\right)=\varphi^{V=j}\left(\boldsymbol{x}_i|\hat{\mathrm{P}}_i\left(T=1|\boldsymbol{x}\right)\right)=\frac{1}{B}\sum_{b}^{B}\tau_b^{V=j}\left(\boldsymbol{x}_i|\hat{\mathrm{P}}_i\left(T=1|\boldsymbol{x}\right)\right); \tag{3}$$

with predictive distribution for alliance $j=1,...,3$:

$$\mathrm{P}_g\left(V=j|T=1,\boldsymbol{x}\right)\sim\mathrm{N}\left(\varphi^{V=j}\left(\boldsymbol{x}_g\right),(\hat{\sigma}_{\mathrm{RMSE2}}^{V=j})^2\right); \tag{4}$$

which we use to estimate the vote choice probabilities for each cell, conditional on turnout: $\hat{\mathrm{P}}_g\left(V=j|T=1,\boldsymbol{x}\right)$. Using the common decomposition proposed by Lauderdale et al.[12], we can the estimate the cell vote distribution of individual who will turn out to vote:

$$\mathrm{P}_g\left(V=j,T=1|\boldsymbol{x}\right)=\mathrm{P}_g\left(V=j|T=1,\boldsymbol{x}\right)\times\mathrm{P}_g\left(T=1|\boldsymbol{x}\right) \tag{5}$$

Once we have this quantity, we can just multiply by the number of individuals in the given cell of the stratification frame, Q_g to obtain the estimated national vote share by alliance ν_j; we repeat this process for each week in the monitoring period (of which there are 13). The national level vote share estimates for each week, w = {1,...,13} can then be obtained using the following formula:

$$\nu_{jw}=\frac{\sum_g \mathrm{P}_g\left(V=j,T=1|\boldsymbol{x},W=w\right)\times Q_g}{\sum_g \mathrm{P}_g\left(T=1|\boldsymbol{x},W=w\right)\times Q_g}; \tag{6}$$

similarly, the weekly national turnout percentage estimate ξ can be obtain with:

$$\xi_w=\frac{\sum_g \mathrm{P}_g\left(T=1|\boldsymbol{x},W=w\right)\times Q_g}{\sum_g Q_g}; \tag{7}$$

.

The fruits of this labour as shown in Figures 4 and 5.

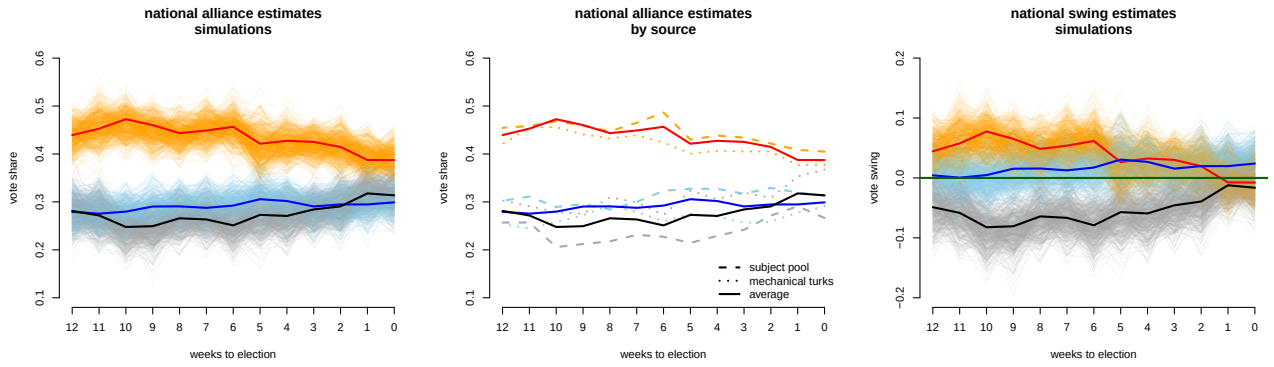

Figure 4: Expected vote share by alliance: the NDA is in red/orange; the UPA in blue/skyblue; the others in black/grey. From left to right: a) 500 simulations of the expected vote share over the 13 weeks monitoring period; b) the breakdown of expected vote share by source (CESS subjects and mechanical turks); c) 500 simulations of the national swing since the 2014 election.

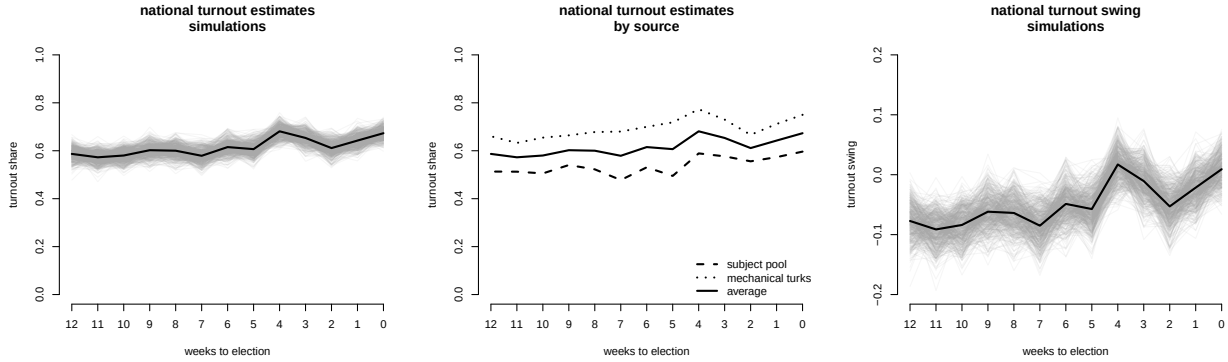

Figure 5: Expected turnout. Same left-right structure as Figure 4.

3.2 Estimating the State Swing Multiplier

Having obtained an estimate for the national swing, we now attempt to convert this to the state-level. As we introduce states into this exercise, we note that we ignore the split of Telangana from Andhra Pradesh in 2014 as we would not have any data to estimate a Telangana-specific effect otherwise.

An alternative to the state-level swing would be to instead apply a Uniform Swing[9], i.e. assume that constituencies in the Lok Sabha would, on average, enjoy the same political change as the National climate would suggest. Then, although the specific constituency may be poorly predicted, the headline share of seats would roughly be accurately predicted. If the assumptions hold, the true swing would be normally distributed across constituencies, leading some constituencies to go the opposite way of the national swing, and others to be even more bullish on the magnitude of that swing; these extremes will appear at roughly the same frequencies, and hence cancel each other out. This tends to work quite well in the US and the UK due to the relatively uniformity of the political climate (meaning political discourse is common across constituencies); the relative homogeneity of constituencies in terms of important socio-economic variables such as class and income; the relative stability of the voting choices. We expect the Uniform Swing to do poorly in India because its polity is so exceedingly heterogeneous across states. On the other hand, we expect constituencies within states to be exchangeable.

Following these considerations, we attempt to estimate a state-level multiplier which will convert the national swing to the state. Our point of departure is the Uniform Swing; for each constituency c in state s, the vote share for alliance j, rho_j, for an election l is calculated as follows:

$$\hat{\rho}_{j,c[s],l} = \rho_{j,c[s],l-1} + \hat{\delta}_{j,l}; \tag{8}$$

where δ is the national swing since the last election, calculated as $\hat{\delta}_{j,l} = \hat{\nu}_{j,l} - \nu_{j,l-1}$. We note that the vote shares estimated won't necessarily add up to 1, but this is not really a problem - in single member plurality constituencies, such as those for the Lok Sabha, what we care about is the party rank (and specifically who comes first), so we can safely ignore this inconsistency.

Based on our previous discussion, we know the estimates $\hat{\rho}$ calculated in this manner will likely be

inappropriate. We then presuppose the existence of a parameter $\psi_{s,j}$, the state level multiplier, such that:

$$\hat{\hat{\rho}}_{j,c[s],l} = \rho_{j,c[s],l-1} + \psi_{s,j}\hat{\delta}_{j,l}, \tag{9}$$

where $\hat{\hat{\rho}}$ is the forecast vote share based on the state-level swing.

One intuitive way to estimate ρ is by looking back in history at the relationship between the National Swing and the average swing of constituencies in a given state. By using the data from the Lok Dhaba project[11], we re-calculate the historical constituency-level vote share according to the 2019 alliances structure, as well as the national swing for all years since 1989. We do not include years prior to this since the BJP was not the force that is has become, and we can assume a structural break before the 1989 election that allows us to ignore previous results (from the Congress-dominated era). Due to the unorthodox estimation setup (a fixed intercept and a hierarchical parameter to estimate) we fit a Bayesian model using the software JAGS[17], whose unique flexibility allows us to specify the structure of any graphical model, and estimate the unobserved nodes via gibbs sampler[5]. The full conditional specification of the model is as follows:

$$\begin{aligned} \rho_{j,c[s],l} &= \rho_{j,c[s],l-1} + \psi_{s,j}\hat{\delta}_{j,l}; \\ \psi_{s,j} &\sim \mathrm{N}(0, {\sigma_\psi}^2); \\ \sigma^\psi &\sim \mathrm{Unif}(0,5); \end{aligned} \tag{10}$$

where σ^ψ is given a non-informative uniform prior as suggested by Gelman et al.[4], and information is pooled across states and parties. The hierarchical prior is specified in the hope that, by applying shrinkage to the state-alliance estimates of the multipler, it will improve out-of-sample forecasting.

The results of this model are available in Figure 6. The dotted green line represents the national swing (i.e. when $\psi_{s,j} = 1$). Predictably, hardly any state as an opposite relationship with the National mood; a state is at most exaggerating the swing ($\psi_{s,j} > 1$), or is uncorrelated to it ($\psi_{s,j} = 0$). We note that the vote share forecast when ($\psi_{s,j} = 0$) reverts back to the previous election result, something which should have a stabilizing influence on our forecast. Finally, we note that we can use this plot in reverse: we could for instance poll only the states which are historically most representative of the national swing for each party, and ignore all others; this could save significant amount of resources. The disadvantage of our approach is that it is based on the historical relationship between the national mood and the state; if this relationship suffers a structural break, we won't be able to adapt.

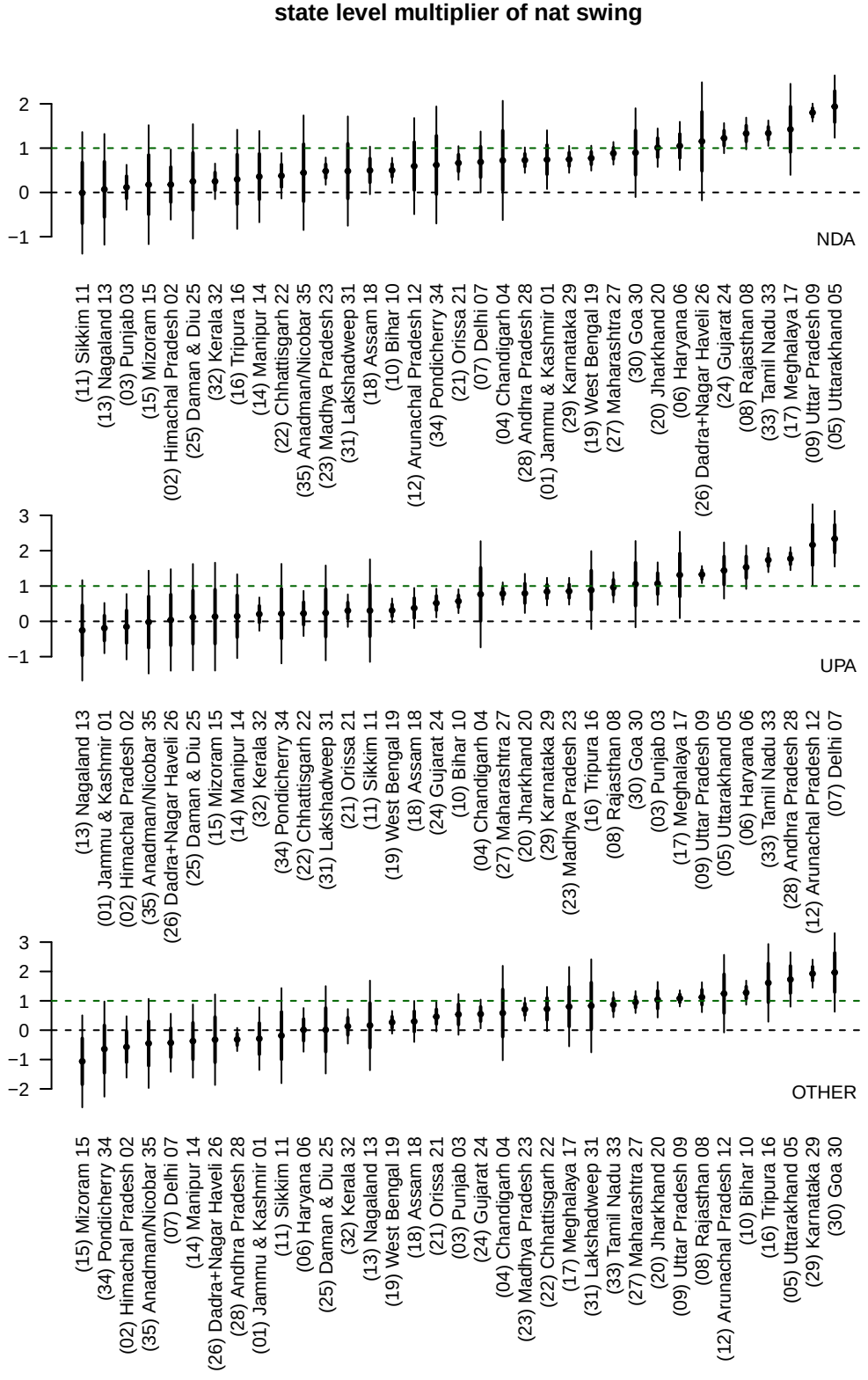

Figure 6: Graphical representation of the state level multiplier of the national swing with uncertainty bounds (2 standard deviations). The dotted green line represents the National Swing.

3.3 Modeling Traditional Opinion Polls

Notwithstanding the substantive modeling efforts to make representative inference from our non-representative online sample, we may still be concerned about idiosyncrasies due to the 'original sin' of selection effects. We leverage the literature on model averaging by Graefe et al.[6] and opt to average our online forecast with other sources. Already, our online forecast is the product of a simple average of two online sources, namely the `CESS` subject pool and the mechanical turks; we note that simple average has been shown is most circumstances to provide better results than conceptually complex Bayesian averages[7]. Hence we opt to average our online forecast with a forecast based on traditional opinion polls. We obtain publicly available opinion polls for the 2009[9], 2014[10] and 2019[11] Indian elections on `Wikipedia`. We focus on polling which provides the headline National seat distribution. This type of polling is the most widely available, going back to previous elections. It is of interest to have comparable polls, ideally from overlapping houses, over as many elections as possible (in our case, three), as we aim to estimate house bias over multiple elections, as well as the general polling error. State-level polls are available for the 2014 and 2019 elections, but given the object of our forecast is the headline number of seats, not the state-level breakdown, we do not leverage those. Moreover, given how wildly wrong the polls were in 2014 (underestimating the NDA victory dramatically), we needed a third benchmark to make meaningful inference over the expected polling error, and this would not have been available at the state level. The total number of polls available to use before the announcement of the 2019 results is 42, over the course of these three elections.

The model we choose to aggregate the polls is a version of the dynamic Bayesian forecasting model by Linzer[13], which seeks to use random walks to propagate uncertainty over days in which polls are absent. Our version is somewhat less complex, as it does not involve hierarchical pooling of state and national level polls; on the other hand we augment the model by enabling estimation of random walk parameters and house effects over multiple elections. We introduce house-bias 'catchers' - namely random effects by polling house, that we net away from the forecast. The model's full conditional specification follows. Expected seat counts P for alliance $j = \{1, 2, 3\}$, elections $l = \{2019, 2014, 2009\}$, polling house $h = \{1, ..., H\}$, with $w = \{1, ..., W\}$ weeks left in the campaign can be assumed to follow a normal distribution, with unstructured variance σ_{P^2} and mean $\mu_{j,l,w,h}$:

$$\begin{aligned} \mathrm{P}_{j,l,w,h} &\sim \mathrm{N}\left(\mu_{j,l,w,h}, \sigma^2_{\mathrm{P}}\right); \\ \sigma_{\mathrm{P}} &\sim \mathrm{Unif}(0, 100); \\ \mu_{j,l,w,h} &= \phi^*_{j,l,w} + \beta^*_j + \zeta^*_{j,h}; \end{aligned} \tag{11}$$

the mean is a linear function of the true daily expected seats $\phi^*_{j,l,w}$, the general expected error by alliance, β^*_j, and the house-specific bias $\zeta^*_{j,h}$. The unstructured variance is given a non-informative prior. The daily

[9] `https://en.wikipedia.org/wiki/2009_Indian_general_election#Opinion_polling`
[10] `https://en.wikipedia.org/wiki/Opinion_polling_for_the_2014_Indian_general_election`
[11] `https://en.wikipedia.org/wiki/Opinion_polling_for_the_2019_Indian_general_election`

expected mean is fixed to sum to 543, so that we can use it directly as our unbiased forecast; this effectively re-scales the house effects $\zeta^*_{j,h}$ to be the number of extra seats predicted for a given party by a given house, on-top of the true count, after considering the general alliance specific error β_j. The true expected counts over the campaign are connected via a random walk, with change parameter σ_η estimated by pooling information from all alliances across all elections. This can be done via reverse random walk for 2009 and 2014, because we know the election results; for 2019 we deploy a forward random walk informed by the same change parameter, essentially projecting the election-week result to the the expected seat count on the day of the last poll, plus estimated uncertainty based on weekly changes in expected counts. This model is shown in its full conditional form below:

$$\begin{aligned}
\phi_{j,l,w} &= 543 \times \frac{\eta_{j,l,w}}{\sum_j \eta_{j,l,w}}; \\
\eta_{j,1,w} &\sim \mathrm{N}(\eta_{j,l,w+1}, \sigma^2_\eta); \\
\eta_{j,l,w} &\sim \mathrm{N}(\eta_{j,l,w-1}, \sigma^2_\eta), \quad \forall \quad l = \{2,3\}; \\
\sigma_\eta &\sim \mathrm{Unif}(0, 100).
\end{aligned} \tag{12}$$

Finally we specify priors for the bias effects. What is being estimated here is the average systematic error over the course of three electoral campaigns, by party. This is decomposed into a general systematic error β and a house specific deviation from that error, ζ. If we are to make inference on house performance, the general alliance error must also be considered a house-specific feature, as it is plausible that some houses do not suffer from this bias, and their ζ will then merely work to cancel out the general bias; as such, to rank pollster performance we should use $\lambda^*_{j,h} = \beta^*_j + \zeta^*_{j,h}$.

These effects are not explicitly dynamic; if a given polling house was very biased in 2009 and 2014, we can expect it to be estimated to be very biased 2019 (the magnitude of the bias being roughly that of the average of the previous two elections). Some thought was given to identification constraints, to aid convergence and interpretation of the coefficients. We opted to use a sum to zero constraint over parties for both general and house specific error; this ensures that each house makes a coherent forecast (i.e. one that adds up to 543 seats). For the house-specific effects, we further use a sum to zero constraint over all houses. This has an impact on the scale of β_j, as this effect becomes the bias of the average pollster, plus the systematic general error. ζ^* is estimated as the difference between the new house bias and the average of the houses, net of the general forecasting error. Had we not used this latter constraint, the effect would have been measured as the difference between the new house and the expected true counts ϕ; this would have led to an unsatisfactory forecast as the new houses would have not influenced the estimation of ϕ, but rather be estimated merely as the difference between ϕ, estimated according to the error from pollsters whose bias is well known, and the forecast from the specific pollster. In our formulation, ϕ is estimated after accounting for the new average of

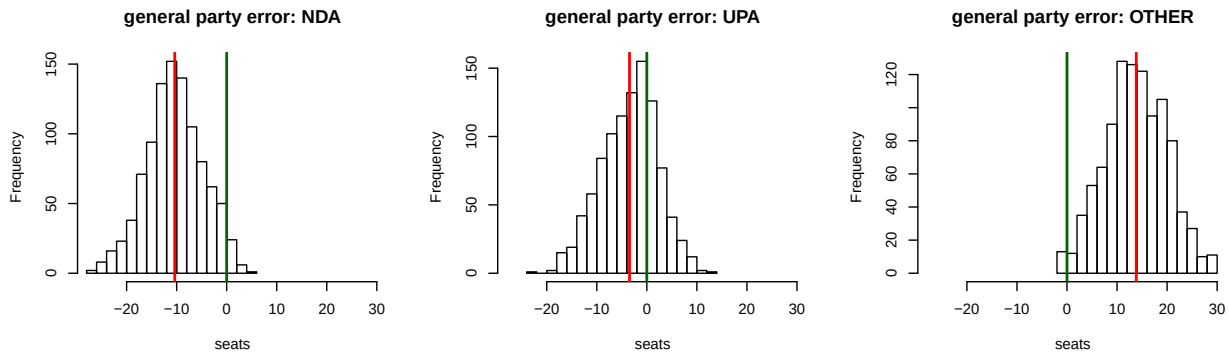

Figure 7: Average error on the true seat counts per alliance, conditional on average house-bias, pooled across elections.

the houses, which is informed by the new polls. The full conditional specifications of the priors on β follows:

$$\begin{aligned} \beta_j^* &= \beta_j - \left(\frac{1}{J} \sum_j \beta_j; \right) \\ \beta_j &\sim \mathrm{N}\left(0, \sigma_\beta^2\right); \\ \sigma_\beta &= 5. \end{aligned} \tag{13}$$

Similarly, the house effects are assigned the following prior:

$$\begin{aligned} \zeta_{j,h}^* &= \zeta_{j,h} - \left(\frac{1}{H} \sum_h \frac{1}{J} \sum_j \zeta_{j,h} \right); \\ \zeta_{j,h} &\sim \mathrm{N}\left(0, \sigma_\zeta^2\right); \\ \sigma_\zeta &= 5. \end{aligned} \tag{14}$$

An important feature of these hierarchical priors is the information we input on the standard deviations σ_ζ and σ_β, which are set to 5 seats. Though this may seem arbitrary, it has useful implications for our forecast: we do not expect the polling error to be as severe as 2014, and having only 2014 and 2019 to measure the expected error inevitably biases it towards the 2014 result. Hence we leverage the full potential of Bayesian reasoning to input subjective priors and potentially improve our out-of-sample forecast. Note that this prior choice does not change the direction of the bias, it just shrinks the point estimate of the more extreme likelihoods.

Figure 7 displays the expected systematic alliance errors based on the data at our disposal, whilst Figure 8 shows the house bias ($\lambda_{j,h}^* = \beta_j^* + \zeta_{j,h}^*$), enabling us to rank worst and best pollsters according to their past results, and providing an expectation for the coming election.

On average, polls can be expected to underestimate the NDA performance by roughly 10 seats, and over-estimate the 'Other' parties by roughly the same amount. It's unclear whether the UPA is significantly under-estimated, but if so it is by negligible amounts. The general systematic bias, overall, is quite small

(plus or minus 10 seats), but could make the difference between expecting a NDA majority as opposed to a hung parliament in 2019. Figure 8 reveals a high degree of heterogeneity across houses with respect to their specific biases. If we rank pollsters by expected mean absolute seats error, across alliances, we obtain the following: [CSDS (13), Karvy (13), CVoter (14), VMR (14), Nielsen (14), India Today (17), CNX (20), VDP Associates (21), Hansa Research (26)], where the figure in brackets represents the average absolute error in seats.

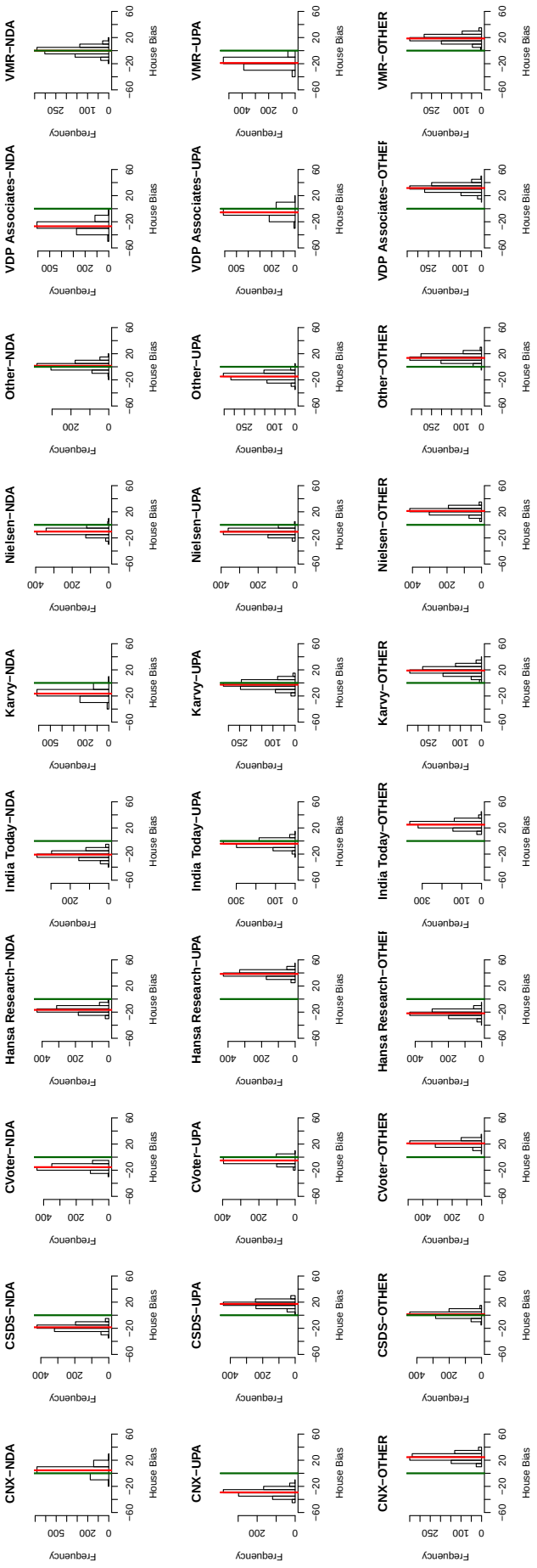

Figure 8: Expected polling house bias.

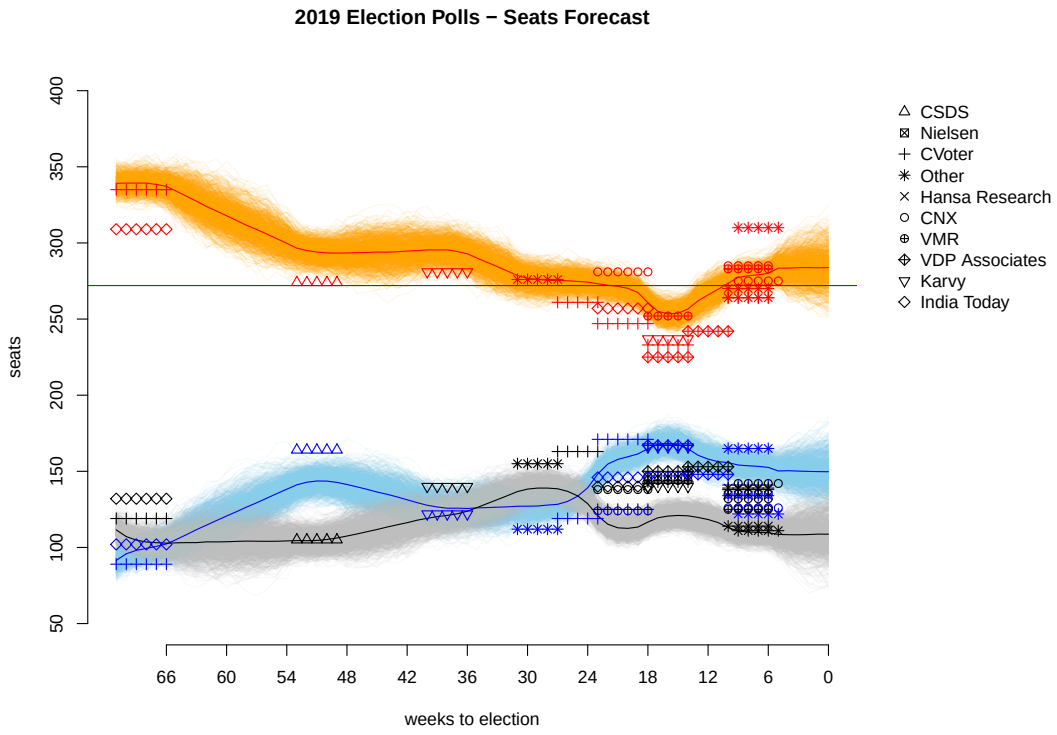

Figure 9: Expected number of seats by alliance for the 2019 Lok Sabha election, net of house bias.

Figures 9, 10, 11 show the fitted model over each respective election campaign. The monitoring period for each election begins 72 weeks before the last election day; the first week of each election is always initialized with the results of the previous election, as 72 weeks is plenty of time for the random walk to 'forget' those results, and it enables a rough estimation of the support between the last election and the first available poll. The 72 weeks figure is chosen as that is the earliest we have polls available for in 2019.

There is no doubt that 2019 is going to be quite different from the previous elections, based simply on the number of polls and differences in the players involved. From Figure 9 it is clear that, once one nets the expected house bias from the equation, the polls suggest the NDA should retain a majority. Should this forecast be wrong, critics will argue historical performance of polls should have been discounted altogether; however, in absence of detailed information about methodology, and in the context of extreme uncertainty, we consider the historical performance of pollsters a reasonable benchmark for our forecasts.

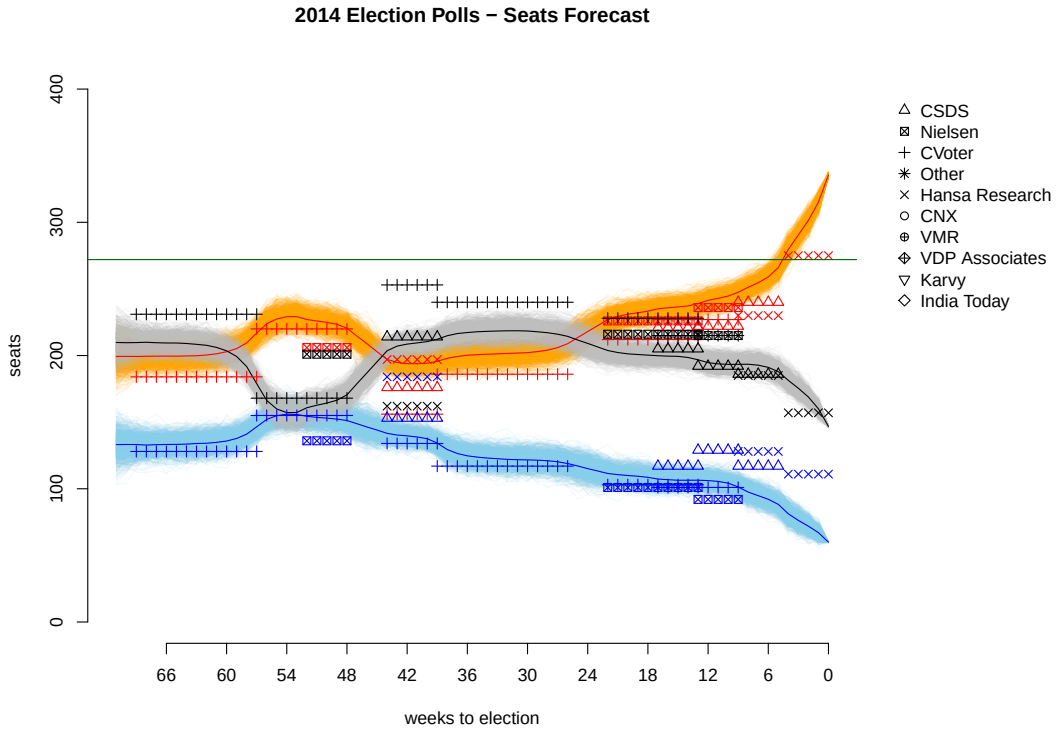

Figure 10: Expected number of seats by alliance for the 2014 Lok Sabha election, net of house bias.

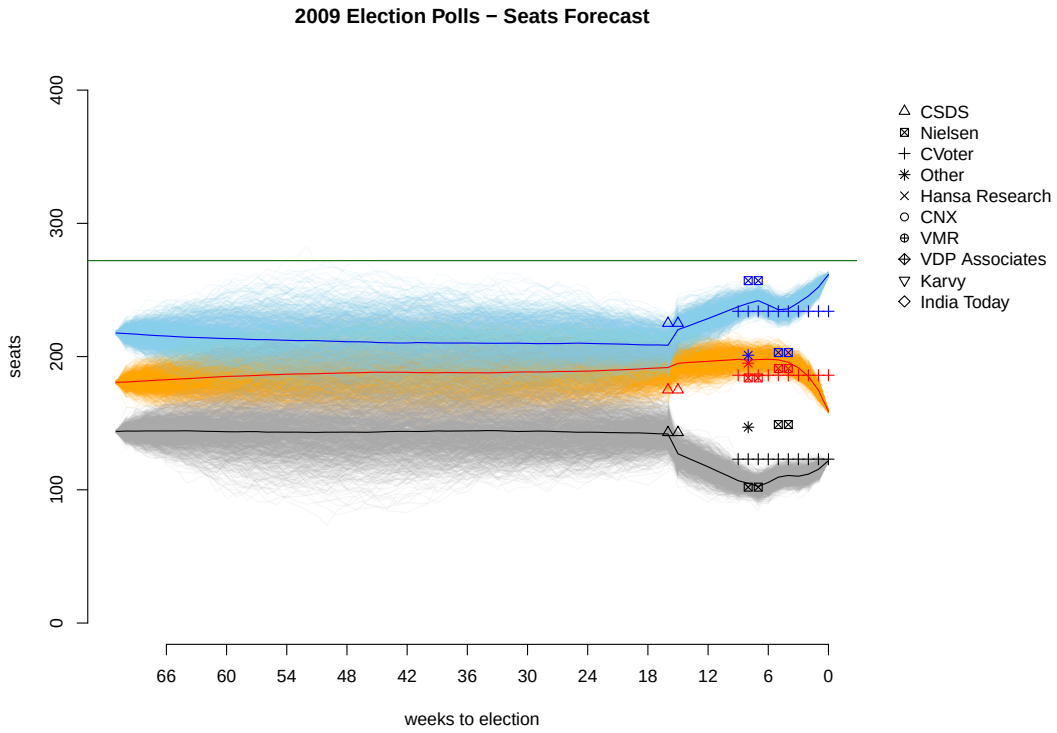

Figure 11: Expected number of seats by alliance for the 2009 Lok Sabha election, net of house bias.

References

[1] Susanto Basu and Brent Bundick. Uncertainty shocks in a model of effective demand. *Econometrica*, 85(3):937–958, 2017.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Sonalde Desai, Reeve Vanneman, and National Council of Applied Economic Research. India Human Development Survey-II (IHDS-II) 2011-12. *ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]*, 2015-07-31.

[4] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[5] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[6] Andreas Graefe, J Scott Armstrong, Randall J Jones Jr, and Alfred G Cuzán. Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54, 2014.

[7] Andreas Graefe, Helmut Küchenhoff, Veronika Stierle, and Bernhard Riedl. Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951, 2015.

[8] Igielnik, Ruth and Scott Keeter and Rachel Weisel. Commercial voter files and the study of u.s. politics. Technical report, Pew Research Center, 2018.

[9] Simon Jackman. The predictive power of uniform swing. *PS: Political Science & Politics*, 47(2):317–321, 2014.

[10] Simon Jackman and Bradley Spahn. Why does the american national election study overestimate voter turnout? *Political Analysis*, pages 1–15, 2019.

[11] Francesca R Jensenius and Gilles Verniers. Studying indian politics with large-scale data: Indian election data 1961 to today. *Studies in Indian Politics*, 5(2):269–275, 2017.

[12] Benjamin E Lauderdale, Delia Bailey, YouGov Jack Blumenau, and Douglass Rivers. Model-based pre-election polling for national and sub-national outcomes in the us and uk. Technical report, Working paper, 2017.

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

[14] Benjamin Lu. *Constructing Prediction Intervals for Random Forests*. PhD thesis, Pomona College, 2017.

[15] James D Malley, Jochen Kruppa, Abhijit Dasgupta, Karen G Malley, and Andreas Ziegler. Probability machines. *Methods of Information in Medicine*, 51(01):74–81, 2012.

[16] David K Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.

[17] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria., 2003.

[18] Eswaran Sridharan. Class voting in the 2014 lok sabha elections. *Economic & Political Weekly*, 49(39):72–76, 2014.

[19] Daniel J Stekhoven and Peter Bühlmann. Missforestnon-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

[20] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

[21] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[22] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[23] Marvin N Wright and Andreas Ziegler. Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.