Title: Measuring and Forecasting Public Opinion with Non-Representative Samples

Candidate Name: Roberto Cerina

College Name: Nuffield College

Degree: Doctor of Philosophy in Sociology

Term of Submission: Trinity

Year of Submission: 2020

Abstract

In this doctoral thesis, I reflect on methods to leverage non-representative samples to measure and forecast public opinion in the run-up to an election. The thesis is made up of three separate papers, dealing with a number of issues around the topic of convenience samples; these issues include a) how and where to collect non-representative data, and how to enhance them with external, auxiliary sources; b) how to construct reliable and deep stratification weights, when these are not immediately available from the census, and when no single dataset contains an exhaustive set of the desired post-stratification information; c) how to optimize learning models to automate variable selection, produce accurate out-of-sample predictions, and improve fitting speed. Findings from this thesis suggest unrepresentative samples can be used to obtain real-time readings of public opinion, at a fraction of the cost of traditional survey methods, provided a number of non-trivial and computationally demanding model adjustments are implemented to smooth and re-weight the samples.

Measuring and Forecasting Public Opinion with Non-Representative Samples

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy in Sociology



Word Count: 52,000

To my family: it's a long way from Latina to Oxford, but carrying you with me eased the load.

To Fij: you lightened the path when it got dark; now we take the next steps together.

Contents

Page No.

Chapter I: Introduction **3**

References **8**

Chapter II: Measuring Public Opinion via Digital Footprints **12**

1 Introduction **13**

2 The Virtual Sample and Post Stratification Data **14**
2.1 The Texas "Virtual" Sample 15
2.2 Matching Sample to the Post Stratification Data 20

3 Area Estimates of Partisan support **25**
3.1 Variable Selection and Imputation . . . . .
3.2 Likelihood Prediction Models 28
3.3 Voter Turnout Estimation . . . . .
3.4 Estimating Vote Probabilities 31
3.5 Aggregation: Vote and Seat Prediction 32

4 Results: Benchmarking Estimates **33**

5 Digital Challenges **41**

References **45**

Appendix **49**

Chapter III: Polling India via Regression and Post-Stratification of Non-Probability Samples **58**

1 Introduction **59**

2 Background **61**
2.1 Expectations of Regression and Post-Stratification in India 63

3 Data **64**
3.1 Constructing a Stratification Frame 65
3.1.1 Stacking 69
3.1.2 Multiple Imputation 71
3.1.3 Iterative Proportional Fitting 72
3.2 Working with Convenience Samples 74
3.2.1 Turnout Training Data 76
3.2.2 Severity of Non-Representation 77

4 Modeling Behaviour **80**

5 Results **87**

6 Discussion **97**

References **102**

Appendix **110**

Chapter IV: Learning Voting Behaviour from Synthetic Surveys **122**

1 Introduction **123**

2 Framework Overview **128**
2.1 Phase I: Generating a Synthetic Population 132
2.1.1 From GLA to PC . . . . .
2.1.2 Augmentation with Political Variables 133
2.1.3 Augmentation with Area-Level Covariates 135
2.1.4 Non-Parametric MICE 136
2.1.5 Raking to Known Area-level Marginals 139
2.1.6 Dealing with an Out-dated Census 140
2.2 Phase II: Generating Synthetic Surveys 141
2.2.1 Surveying the Synthetic Population . . . . .
2.2.2 Collecting Publicly Available Polls . . . . .
2.2.3 Generating Synthetic Survey Samples 143
2.3 Phase III: Learning, Predicting and Aggregating 148
2.3.1 Prediction in Regression and Post-Stratification . . . . .
2.3.2 Choosing a Learner . . . . .
2.3.3 Models 151
2.3.4 Tuning Strategy 152
2.3.5 Meso-level Aggregation 158

3 Results and Evaluation **159**

4 Discussion **170**

References **176**

Appendix **186**

Chapter V: Rejoinder **234**

References **239**

Chapter I: Introduction

Can non-probability samples replace random-digit dial as the go-to- method to measure and predict public opinion? This thesis argues that, provided appropriate model-based adjustments and satisfaction of some minimum requirements, unrepresentative samples of individuals can rival the accuracy and reliability of random samples obtained via established methods, for a fraction of the cost. This is a collection of three separate papers, which are connected through the common aim of measuring public opinion via non-probability samples, and in particular digital data (online subject-pools, social-media trace data, online crowd-sourcing workers, etc.).

Motivation. The motivation behind this exploration lies in the concrete trends that have changed public opinion research in the last decade: i) it has become, and it is increasingly so, expensive and impractical to obtain true probability samples from a population of interest; ii) it has become, and it is increasingly so, practical and inexpensive to measure opinion of groups which are not representative of the population at large, on a number of different online platforms; iii) advances in methodology and computational power have enabled analysis of mass-scale individual-level data for the purpose of producing fine-grained predictions of behaviour for small groups within a population of interest - this is often referred to as *micro-* or *meso-targeting*, depending on the underlying level of analysis. We can talk about these as incentives motivating us to move us away from probability samples, and embrace non-probability samples; in what follows we briefly discuss their validity and their implications.

Questioning the reliability of random-digit-dial (RDD) polls has become commonplace in the literature, and for good reason: Pew Research Center (Keeter et al. 2017) measures a drop in average response rate per year of 27 percentage points since 1996, down to a mere 9% in 2016. Though the average accuracy of RDD polls in the realm of voter intention has remained stable (Prosser & Mellon 2018, Jennings & Wlezien 2018), selection effects into RDD polls are significant: whilst on most measures of interest we observe consistency between RDD surveys and population benchmarks, the highly politically engaged and the highly educated are over-represented (Keeter et al. 2017) in RDD polls. This selection has contributed to recent high profile polling failures, such as the 2016 US election, which saw lack of adjustment for educational attainment as

a key problem, leading to poor performance in key mid-western states (Kennedy et al. 2018). As has been pointed out by Salganik (Salganik 2019) and others, the issue at hand is not whether, in an ideal world, we would prefer non-representative samples over RDD: we know from Yeager et al. (Yeager et al. 2011) and others that high-quality RDD polling outperforms moderately sized non-probability samples under most circumstances; rather, the issue is that presuming we can obtain true probability samples with RDD has become too strong an assumption. Hence, conditional on having to leverage model-based techniques to adjust for selection effects, researchers are faced with a trade-off: do we opt for high-cost, low-frequency, small-N RDD samples, endowed with moderate but predictable selection effects?; or alternatively do we sample low-cost, high-frequency, large-N non-probability samples, characterized by more complex and less understood selection effects?

This thesis argues that it is not only reasonable, but significantly more effective when weighting practical cost-bias trade-offs, to use non-probability samples for the purpose of monitoring public opinion during an election campaign. It is further argued that sampling from multiple sources can increase measurement and prediction accuracy in the context of elections (Graefe et al. 2014). As such, four sources of non-representative data are considered in this thesis: Facebook digital trace data; Facebook-recruited online survey respondents; Amazon Mechanical Turk (AMT) workers; and the YouGov-British Election Study online subject pool. Within Facebook, we explore quota sampling which leverages the advertising interface provided on the platform, which has been shown to produce representative samples under some circumstances (Boas et al. 2020, Zhang et al. 2020), as well as samples measured unobtrusively by monitoring behaviour on a selection of political pages and matching observed behaviour to Voter Registration Files[1]. With respect to AMT, though previous studies have shown some limited use for public opinion research (Goel et al. 2015) due to arguably unacceptable average errors against population benchmarks, this thesis argues AMT samples are still valuable. Due to the different selection effects into this human-intelligence marketplace, they are likely to capture signal from populations which are not easily reached under other online methods; it will be shown that in the case of India we are able to get a more representative sample

[1]These are large individual level repositories of information including socio-economic, demographic and behavioural information for registered voters. These are famously available in the US, and usually unavailable in other applications.

in terms of income on AMT than by recruiting Facebook users.

The final incentive to move to non-probability samples comes by way of developments in the methodology of prediction models and computational power. Researchers have always been interested in micro- and meso-level behaviour: in 1965, Pool et al. (de Sola Pool et al. 1965) were pushing boundaries by deploying a rudimentary form of meso-targeting, dividing the US electorate into 480 voter categories, and using national opinion polls to inform their voting behaviour. Fast-forward to 2004 and Park et al. (Park et al. 2004) were introducing Multilevel Regression and Post-stratification (MRP), a modeling leveraging Bayesian hierarchical modeling to produce stable out-of-sample predictions, expanding the number of voter categories analyzed to $3,264$, and enabling accurate area-level estimation of voting behaviour from only nationally representative samples. More recently, Ghitza and Gelman (Ghitza & Gelman 2013) further innovated by estimating multiple interaction effects, highlighting the importance of capturing non-linearities in our predictions. The current state-of-the-art has moved MRP to a new paradigm heavily influenced by the machine-learning literature, namely Regularized Prediction and Post-stratification (RPP) (Kennedy & Gelman 2019, Goplerud et al. 2018, Bisbee 2019, Ornstein 2019, Leemann et al. 2020), where the social scientist's traditional role in model specification is replaced by algorithms which search for the optimal summary of a covariate space, compromising between internal and external validity of predictions. These methodological innovations become influential and widespread thanks to the powerful computational tools at the disposal of researchers, often at no monetary cost: the increase in availability to researchers of large-scale computing power through platforms such as `Google Colab`[2], `Amazon Web Services`[3] and other means, as well as the sheer power of modern laptop and desktop computing, has enabled the kind of work presented in Wang et al. (Wang et al. 2015), which analyzes $750,148$ Xbox survey responses under the MRP paradigm outlined above; Jackman et al. (Jackman et al. 2019), which applies a robust version of MrP to around $800,000$ Australian Voter Advice Applications; and Wang et al. (Wang et al. 2019), who leverage massive training sets to accurately predict demographic characteristics of `Twitter` users. The papers presented in this thesis contribute to, and expand the horizons of,

[2] `https://colab.research.google.com`
[3] `https://aws.amazon.com`

the regression and post-stratification literature, by leveraging novel data sources, and proposing novel modeling techniques for smoothing micro-data and obtaining deeper stratification frames.

Ethics. The author is not oblivious to the fact that recent scandals, in particular the infamous Cambridge Analytica (CA) fiasco (Cadwalladr & Graham-Harrison 2018), have highlighted ethical concerns relative to the collection and use of personal data, and especially digital trace, in pursuit of electoral prediction and micro-targeting. On the one hand, it is worth mentioning that the autonomy-related concerns around the CA scandal are largely overblown, as noted by recent empirical findings (Nickerson 2019): the sort of digital influencing campaign conducted by CA, built around the idea that some psychological profiles are more influenceable than others and could be 'tipped-over' by specific kinds of political advertisements, could not have moved a significant number of voters in the 2016 election based on existing evidence. Other kinds of more subtle influencing - such as the imperceptible bias in search-engine algorithms (Epstein & Robertson 2015) have been shown to be far more dangerous to the political autonomy of potential voters. On the other hand, autonomy issues around micro-targeting and political campaigns remain looming: the fact that CA did not have the 'right formula' does not mean said formula is inconceivable, and this may have ethical implications for the work presented here. Readers' main ethical concern with this research is likely to lie in the first of the three papers, where digital traces are harvested for making political predictions, and matched to voter-registration records. Ethical objections to this kind of practice - when deployed on a mass scale, or for profit, which this academic thesis does not do - have been loud in recent years: in the seminal book *'The Age of Surveillance Capitalism'* (Zuboff 2019), Zuboff's radical critique of the digital technology industry raises the spectre of a world where everything, from our most intimate thoughts - including our voting propensities - to our physical and psychological ailments, is recorded and studied en-masse. Her privacy objection then collapses into an autonomy critique: our behaviour is altered, our brains incentivized to engage in activities which go against our interest and beyond our will. We guard against Zuboff's critiques as follows: i) less than $10,000$ individuals are ultimately matched to a voter registration file, a sample size smaller than any given wave of the British Election Study online panel; ii) we limit our

collection to Facebook pages which can be characterized as public spaces, namely official pages of candidates for political, created and managed with the intent to encourage public conversations and attract digital attention; iii) we do not use bots or crawlers for data collection, effectively self-imposing a limit on the potential scope of our activity from the outset via increasing the costs associated with collection a given digital trace; iv) we further limit ourselves to observation, and do not link our predictions to experiments which could alter a user's vote-choice - though we should note these mass-scale experiments altering political outcomes are not uncommon in the scientific literature (Bond et al. 2012, Enríquez et al. 2019). Of further interest to the first paper is the use of voter registration files to expand the information content of digital traces. It should be noted that the mere use of these files does not imply an increase in predictive power over traditional surveys - and hence does not suggest 'special' violations of autonomy are rendered possible by this technology - as the most predictive data sources tend to be those in the public domain - or those easily attainable wit traditional polling practices (Nickerson & Rogers 2014); we should further note that traditional polling methods are still affected by substantial privacy concerns (Bulmer 1979). We contend that, given that micro-data augmentation with voter-registration files is a common practice in political campaigns (Nickerson & Rogers 2014), there is an academic interest in studying the degree to which these sources can enhance digital signals, and it is on these basis that we consider the matching performed in the first paper ethical.

Papers reported in this thesis, which required novel data collection involving human subjects went through multiple steps of ethical review; the first and second paper both obtained ethical clearance from the Central University Research Ethics Committee (CUREC). The third paper in this thesis used British Election Study micro-data which was publicly available and collected by third-parties, or aggregate data which did not have any privacy or otherwise ethical implications, and hence did not require to be submitted to further ethical scrutiny.

As a final note, I would like to highlight that this short discussion does not do justice to the project of 'an ethics of digital data in election forecasting', which deserves its own doctoral thesis, though I am satisfied that the research presented here does not raise insurmountable ethical issues; where some ethical questions arise, these are typically contained and reasonably justified by academic interest.

Thesis Structure. The rest of the thesis proceeds as follows: three papers are presented in order of completion; each paper is introduced a title-page containing an abstract; each papers ends with an appendix, typically enclosing plots, figures and extra details which did not fit in the original paper. The three papers have separate reference sections, typically placed before the appendix. At the end of the third paper, a rejoinder is provided to summarize the findings of the thesis with a 'bird's eye' point of view and point to avenues for future exploration. Note that the first paper has been published in the *International Journal of Forecasting*; the second paper has recently been submitted to the *Royal Statistical Society - Series A* journal; and the third paper will eventually be submitted to this latter outlet.

References

Bisbee, J. (2019), 'BARP: Improving Mister P using Bayesian Additive Regression Trees', *American Political Science Review* **113**(4), 1060–1065.

Boas, T. C., Christenson, D. P. & Glick, D. M. (2020), 'Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics', *Political Science Research and Methods* **8**(2), 232–250.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E. & Fowler, J. H. (2012), 'A 61-million-person Experiment in social influence and political mobilization', *Nature* **489**(7415), 295–298.

Bulmer, M. (1979), *Censuses, surveys and privacy*, Macmillan International Higher Education.

Cadwalladr, C. & Graham-Harrison, E. (2018), 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach', *The guardian* **17**, 22.

de Sola Pool, I., Abelson, R. P. & Popkin, S. L. (1965), *Candidates, issues and strategies: a computer simulation of the 1960 and 1964 presidential elections*, Vol. 33, Massachusetts Institute of Technology Press.

Enríquez, J., Larreguy, H., Marshall, J. & Simpser, A. (2019), 'Online political infor-

mation: Facebook ads, electoral saturation, and electoral accountability in Mexico', *Working paper* .

Epstein, R. & Robertson, R. E. (2015), 'The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections', *Proceedings of the National Academy of Sciences* **112**(33), E4512–E4521.

Ghitza, Y. & Gelman, A. (2013), 'Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups', *American Journal of Political Science* **57**(3), 762–776.

Goel, S., Obeng, A. & Rothschild, D. (2015), 'Non-representative surveys: Fast, cheap, and mostly accurate', *Working Paper* .

Goplerud, M., Kuriwaki, S., Ratkovic, M. & Tingley, D. (2018), 'Sparse Multilevel Regression (and Poststratification (sMRP))', *Working Paper* .

Graefe, A., Armstrong, J. S., Jones Jr, R. J. & Cuzán, A. G. (2014), 'Combining forecasts: An application to elections', *International Journal of Forecasting* **30**(1), 43–54.

Jackman, S., Ratcliff, S. & Mansillo, L. (2019), 'Small area estimates of public opinion: model-assisted post-stratification of data from voter advice applications', *Working Paper* .

Jennings, W. & Wlezien, C. (2018), 'Election polling errors across time and space', *Nature Human Behaviour* **2**(4), 276–283.

Keeter, S., Hatley, N., Kennedy, C. & Lau, A. (2017), 'What low response rates mean for telephone surveys', *Pew Research Center* **15**, 1–39.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D. et al. (2018), 'An evaluation of the 2016 election polls in the United States', *Public Opinion Quarterly* **82**(1), 1–33.

Kennedy, L. & Gelman, A. (2019), 'Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample', *arXiv:1906.11323* .

Leemann, L., Broniecki, P. & Wuest, R. (2020), 'Improved multilevel regression with post-stratification through machine learning (autoMrP)', *Working Paper* .

Nickerson, D. (2019), '*Why I was not Worried about Cambridge Analytica*', `https://cess-nuffield.nuff.ox.ac.uk/events/workshop/social-media-workshop`.

Nickerson, D. W. & Rogers, T. (2014), 'Political campaigns and big data', *Journal of Economic Perspectives* **28**(2), 51–74.

Ornstein, J. T. (2019), 'Stacked regression and poststratification', *Political Analysis* pp. 1–9.

Park, D. K., Gelman, A. & Bafumi, J. (2004), 'Bayesian multilevel estimation with post-stratification: State-level estimates from national polls', *Political Analysis* **12**(4), 375–385.

Prosser, C. & Mellon, J. (2018), 'The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses', *Government and Opposition* **53**(4), 757–790.

Salganik, M. J. (2019), *Bit by bit: Social research in the digital age*, Princeton University Press.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flock, F. & Jurgens, D. (2019), Demographic inference and representative population estimates from multilingual social media data, *in* 'The World Wide Web Conference', pp. 2056–2067.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A. & Wang, R. (2011), 'Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples', *Public opinion quarterly* **75**(4), 709–747.

Zhang, B., Mildenberger, M., Howe, P. D., Marlon, J., Rosenthal, S. A. & Leiserowitz, A. (2020), 'Quota sampling using Facebook advertisements', *Political Science Research and Methods* **8**(3), 558–564.

Zuboff, S. (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*, Profile Books.

Chapter II:

Measuring Public Opinion via Digital Footprints

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Abstract

Do digital traces accurately reflect individual preferences? Can signals from social media be used to measure public opinion? This paper provides evidence in favour of these hypotheses. We test a regression and post-stratification strategy that combines samples of digital traces with a stratification frame containing individual-level socio-economic data, in order to generate area forecasts of the outcome social phenomena of interest. In our example, we forecast the two-party vote of Democrats and Republicans in the 2018 Texas congressional district and Senate election. Our implementation assumes we can observe, and sample, individuals signaling their preference by favoring one virtual location over another. In our case, visiting Democrat versus Republican Facebook pages during the election campaign. Over the course of seven weeks preceding the mid-term elections we generate vote share forecasts which do not use any traditional survey data as input. Our results indicate that individuals leave digital traces that reflect their preferences.

1 Introduction

Online social media increasingly dominates political discourse. An increasingly large percentage of the population leaves distinct digital traces that signal their political preferences. A challenge is mapping these digital traces to the general population and predicting behavior. We propose one method for accomplishing this goal and implement a "proof of concept" demonstrating that observing and recording digital traces can be the basis for predicting political behavior in the population. With recorded digital traffic to candidate Facebook pages, we forecast the 2018 mid-term elections in the state of Texas. Election forecasts are an ideal vehicle for assessing whether digital traces are a useful tool for measuring public opinion because the outcome of interest, the party or candidate vote share, is observed after the election. We assess the accuracy of digital trace forecasts by comparing actual outcomes in the 36 Texas congressional races (and one Senate race) with our predictions. Texas was selected for this "proof of concept" because it has a large number of diverse districts of varying competitiveness.

The post stratification dataset for generating the forecasts is an enumeration of all registered voters in the state of Texas. Our innovation is to construct a "virtual" sample consisting of voter preferences unobtrusively measured by observing who visits Republican and Democratic candidate Facebook pages. Using basic socio-demographic characteristics of these Facebook users, we map their revealed party preferences to the cells in the post stratification dataset. We then implement a random forest to model preferences and generate predicted vote share for the parties in each congressional district. The essay provides important insights into how digital trace can inform models predicting human behavior. It also highlights the challenges associated with collecting these digital trace data.

There are two important features of the prediction and post-stratification approach we implement. First, our post stratification dataset contains individual records for a significant proportion of the population. These are publicly available voter registration and vote history records. Secondly, our sampling frame is unique in that it is a virtual sample of digital traces – we populate the sample with individuals we observe interacting with either a Republican or Democrat Facebook page. Rather than measuring self-reported preferences, we measure revealed preference.

The cells in our post stratification dataset reflect the individual characteristics that we believe determine vote choice. By generating a sufficiently large sample of virtual partisan decisions we can, in principle, estimate the partisan composition of each of these population cells. We do this by matching our observed virtual decision makers to the cells in the population frame – so, for example, the digital trace of a young black Republican auto-worker could be matched to an appropriate cell in our state of Texas population frame. We use the `R` (Team et al. 2013) package `fastLink` (Enamorado et al. 2018) to match our virtual sample of digital traces to the voter registration list, based on names, sex and city of residency; which resulted in a success rate of approximately 53% percent.

We demonstrate that "digitally" revealed preferences are strongly correlated with behavior. Most individuals observed visiting either a Republican or Democrat candidate's Facebook page voted in either a Republican or Democratic primary. Our sample's digital partisanship is strongly correlated with the partisan primary in which they voted, as well as their imputed partisanship status on their voter registration record.

The partisan composition of individual cells in our sampling frame is informative to the extent that the likelihood of interacting with political content on social media is independent of voting preferences within the voting groups. A comparison of the social media partisanship of our virtual sample cells to the recorded partisanship of our population cells suggests that our independence assumption is roughly accurate for partisan voters. This independence allows us to generate an informative distribution of Republican and Democratic partisans within each cell of the state post stratification dataset.

For each cell in the post stratification dataset we estimate a vote probability based exclusively on the digital trace of individuals visiting candidates' Facebook pages. The estimation is a novel variation of Multilevel Regression with Post-stratification (MRP) (Wang et al. 2015, Buttice & Highton 2013, Ghitza & Gelman 2013, Lauderdale et al. 2020). We employ a Probability Machine (Malley et al. 2012) which is a random-forest based algorithm implemented via the `ranger` (Wright & Ziegler 2015) package. We then weight these probabilities by the likelihood of individuals turning out and aggregate them to obtain estimates of district and state level support.

The pre-election vote share predictions for both Senate and congressional districts

performed satisfactorily when benchmarked against actual election results. The digital trace forecast called races correctly for the Senate and all but 3 congressional districts (7, 32 and 23). Absolute error was high due to significant attenuation bias, something to be expected in MRP applications (Lauderdale et al. 2020), but which was likely exacerbated by our sampling procedure, modeling choices and high levels of noise in the digital sample. We correctly-called twenty-nine of the thirty-two contested congressional races and one Senate race, coming close to matching the performance of `FiveThirtyEight` (Silver 2018*a*, **?**)[4], though again matched less well in terms of absolute error. Donald Trump's 2016 vote share at the congressional district level gives a lower absolute error, as a predictor of two-party Republican vote share, than either the digital trace forecast or `FiveThirtyEight`.

These results suggest that a virtual sampling strategy is promising. Forecasts based on sampling revealed preferences from digital activity on social media perform similarly to those based on conventional sampling methods. The essay describes how to implement a digital trace forecast and proceeds as follows: We first describe how we generated a sample of partisan digital traces of eligible Texas voters. A subsequent section indicates how we matched the virtual sample to the Texas population frame. A third section describes the random-forest estimation of Republican and Democrat voting probabilities for eligible voters in all 36 Texas congressional districts and the Senate. This is followed by our week-by-week mid-term election forecasts which are benchmarked against conventional polling forecasts. We conclude with a discussion of the challenges associated with using digital trace data to forecast population behavior.

2 The Virtual Sample and Post Stratification Data

Research designs explicitly incorporate prediction and post-stratification strategies with the aim of improving the consistency and precision of estimated outcomes for subsets of populations. The technique is frequently employed for estimating political preferences and outcomes for sub-units of national populations (Park et al. 2004). In fact there are quite diverse applications for MRP techniques (Downes & Carlin 2020).

The method has four fundamental components: 1) a post stratification dataset

[4]We benchmark against `FiveThirtyEight`'s polls-only model, and compare our weekly estimates to their daily estimates averaged by week.

that defines the individual-level categories that predict the outcome of interest; 2) an individual-level model that includes these categorical variables and, possibly, aggregate-level variables that predict the outcome of interest; 3) a sampling frame that indicates how the individual-level observations will be collected; and 4) a post-stratification algorithm. The goal is to map an estimated quantity (such as the likelihood of voting Republican) to each of the cells in the population frame (high educated black women in congressional district 32, for example). Simply multiplying this estimated likelihood times the population for this cell gives us the estimated Republican and Democrat voters (for this cell).

Predicted outcomes for each cell are generated with coefficients estimated from individual-level models. Current practice is to estimate these individual-level coefficients with relatively large survey samples of the population of interest (Park et al. 2004, Kiewiet de Jonge et al. 2018). Some evidence suggests that the performance of post-stratification estimation varies significantly depending on the features of these survey samples (Buttice & Highton 2013). On the other hand, there is evidence that post-stratification performs relatively well with unconventional sampling frames. A case in point is the MRP forecast of the 2012 election outcomes based on an unconventional convenience Xbox user sample (Wang et al. 2015).[5] In fact, individual-level estimates from highly skewed sampling frames are the basis for many MRP studies (Downes & Carlin 2020, Zhang et al. 2015). A key factor in the success of these MRP estimations is sample size: MRP estimation needs to be powered in order to accurately estimate vote likelihoods of sub-categories. A smaller sample, generated in the same manner, is more likely to violate the ignorability condition and hence less likely to correctly calibrate rare sub-categories of voters. Hence, unconventional sampling frames as the basis for MRP can work so long as the sample sizes are very large.

2.1 The Texas "Virtual" Sample

The MRP techniques invite the exploration of novel sampling strategies on which to base post-stratification estimation. In this spirit, we propose a non-probability quota sample

[5] Wang et al. administered a daily opinion survey to a very large sample of Xbox users. Their sample was extremely skewed: it over-represented Romney partisans, the young, the highly educated, and third-party voters. Despite this, their MRP estimates closely tracked the published polls, outperforming them in predictive power within the last few weeks of the campaign.

that consist of social media subscribers. A "virtual" sample is entirely unobtrusive. Individuals are never asked to report their preferences or behavior. Rather their choices on social media are observed, unobtrusively. In our case, we observe Facebook users commenting on partisan Facebook pages.[6] In this particular digital sampling frame we restrict ourselves to Facebook pages – our simple requirement from these pages is that they attract political partisans and there is evidence to this effect (Bakshy et al. 2015).[7]

Our digital sampling strategy consisted in research assistants (RAs)[8] monitoring the Facebook pages of all Texas Republican and Democratic candidates for congress and the Senate – a detailed description of the system followed by the RAs is available in Algorithm 1 in the Appendix, and summarized as follows: a total of 68 candidate Facebook pages were monitored (see Table $A1$ in the appendix for details on these pages); for the eight-week period leading up to the election RAs were tasked on a daily basis to collect up to 30 digital traces from the relevant pages. The RAs would collect information on Facebook users who left a digital trace expressing support on a candidate's page, where digital support was deemed to come from a like, love or explicitly positive comment under a post published on the candidate's page. Whether a comment classified as 'explicitly positive' was left up to the RAs to decide based on their qualitative judgment. RAs were also asked to note, in a `Qualtrics` survey instrument, basic information regarding these Facebook users – in particular, Facebook ID, gender, and current address and home address if available. These Facebook users are categorized as partisans based on the pages they express support for and then are added to our virtual convenience sample.[9] The assumption here is that visiting a candidate's Facebook page and expressing a positive sentiment is a strong signal of voting intention. Again, there are recent findings suggesting that Twitter sentiment is correlated with

[6]The U.S. has about 140 million Facebook subscribers and Texas has about 17 million. In both cases these represent about 43 percent of their respective populations `https://www.internetworldstats.com/stats26.htm`.

[7]There is also evidence that sentiment expressed on candidate Facebook pages correlates with election outcomes although this is of less relevance to the MRP predictions we are generating in this forecasting exercise (Zhang 2018).

[8]Research assistants collected on average just over 325 traces per day, and were paid roughly \$12 per hour. The average cost per digital trace was \$0.4, for a total of just over \$6,000 over 48 days. Up to four different RAs were used on any given collection day.

[9]Our research assistants were instructed only to consider Facebook users of voting age, i.e., 18 years or older. It is of course possible that we collected social media profiles of individuals who were not eligible to vote – for example, non-citizens or individuals below the age of 18. Virtually all of these ineligible voters would be screened out of the sample once these Facebook profiles were matched to the Texas voter files. The possible very rare exception would someone who died or was convicted of a felony after the vote registration data were collected.

partisan voter registration data (Barberá et al. 2015) and that Facebook likes predict reported vote choice (Kristensen et al. 2017). Post-stratification estimation is based on this partisan virtual sample.

A virtual sample can consist of any digital trace that reflects a choice amongst different virtual destinations or options. It could be, for example, liking an Uber versus Lyft Facebook page or responding positively to competing Instagram campaigns. We argue that these novel sampling frames can be the basis for quite robust MRP-like estimations. The 2018 mid-term congressional and Senate elections in Texas provided an opportunity to evaluate this alternative MRP estimation strategy. Over the course of the 7 weeks preceding the November 6, 2018 mid-term election day we collected $15,683$ digital traces. Our innovation is to map voting probabilities estimated from a "virtual" Texas sample to a well-defined Texas population frame.

Our Texas population frame has $38,880$ cells. For each of the 36 Texas congressional election districts we disaggregate the population by registration-file partisanship (three categories), gender (two categories), age (six categories), ethnicity (five categories) and education (six categories). Our initial step, though, is to match individuals who visit a particular partisan FB page to one of $1,080$ cells in the overall Texas population frame (prior to disaggregation to the 36 congressional districts). Hence, the "partisanship" of each cell is determined by the frequency with which we observe individuals, with a particular set of characteristics, visiting a partisan FB page.

The degree to which sampled partisan user counts in each cell reflect true cell preferences depends on whether the partisan members of the cell are represented in our quotas at the same rate they are in the population. For this to be the case we would need to assume that the conditional propensity to interact with social media is independent of partisanship. This is an independence assumption that is unlikely to be strictly observed in practice. In order to directly evaluate this assumption we would need to know "true cell preferences" or the true conditional propensities to interact with social media. We know neither and determining the precise magnitude of these deviations is beyond the scope of this particular project.

Of course, most sampling frames, convenience or not, will violate, to some extent, this independence assumption. The best we can do here is benchmark our forecast to these other conventional forecasts. To the extent that our digital trace sampling strategy

generates similar forecasts as those employing more conventional sampling strategies, it is reasonable to conclude that violations of the independence assumption by digital trace sampling is no worse than the violations occurring with more conventional samples.[10]

A second concern is that the size of the virtual quotas is large enough to capture relevant cell differences in partisanship. We tune our sample size to ensure this is true for large categories in the voting population, under a probability-sampling scenario (a limiting best-case for our application). To tune sample size we assume a sampling distribution for cell counts of partisan voters of the following form:

$$\boldsymbol{n} \sim \text{Multinomial}(p_{1,r=1}, ..., p_{G,r=1}, p_{1,r=0}, ..., p_{G,r=0}, N); \tag{1}$$

where $\boldsymbol{n}$ represents the vector of sampled counts for voters belonging to cell g, where $g = 1, ..., G$, and expressing a preference for party r, where $r = \{0, 1\}$ (indicating support for the Republican party). This set up enables us to leverage *worst-case-scenario* sampling for a multinomial distribution following the recommendations of Thompson (Thompson 1987). We need to specify the nature of our population frame; in our case, indicating the relevant socio-economic categories in our population frame that predict vote choice.

We set the sample size to ensure that we can have sample frame cells that represent at least 2.5 percent of the population[11]. In order to ensure a probability of at least 0.9 that all estimates of the multinational parameters are within 0.025 of the population

[10] Earlier versions of this essay experimented with more "shallow" versions of this independence assumption. In one simulation we assumed that the marginal distribution of the two-party votes was 50-50. We then proceeded to re-sample within each cell accordingly to ensure this marginal likelihood was preserved. By balancing the marginal voting likelihood, the within-cell distribution of votes is as-if independent of the propensity to interact with social media. On balance these variations on the independence assumption did not generate significantly different results and in some cases reduced the accuracy of the forecasts.

[11] By setting the expected distance from the population parameter $d = 0.025$ with a given probability $(1 - \alpha)$ we are essentially saying that if a given group's true proportion is 0.01, there is a good chance that the sample estimate for this quantity will be zero; hence our sample is only powered to accurately estimate group proportions which are around 2.5% of the population. This rough equivalence between d and the size of the smallest groups we are powered to estimate results from the tiny size of the true population proportions; as the groups approach zero, we need d to also approach zero to accurately capture their population levels. Moreover, given we are sampling to eventually derive Republican and Democrat voting likelihoods within small groups (say those that make up 0.1% of the population), small sampling errors can lead to huge swings in within-group voting likelihoods, and ideally we'd want d to be smaller than the group size. Note that 2.5% is significantly higher than $1/1080 \approx 0.1\%$ of the population, which would allow our sample to be theoretically tuned to accurately estimate every single voter sub-groups. Instead we opt to be powered for the largest voting groups, here defined as groups making up 2.5% or more of the Texas voting population, and assume the smaller groups will be randomly distributed between the two major parties.

proportions, we need a sample size of at least $N = 1610$. Note here that 2.5% refers to the whole Texas population, as our cells will be collected at the state-wide level. Although we make use of sampling theory to inform a-priori expectations regarding sample-size for desired power, this is still a non-probability convenience sample due to the sampling strategy and nature of the digital locations.

Our sampling strategy is dynamic, designed to calibrate weekly changes in voter preferences. Accordingly, we generate a weekly "virtual" sample of the Texas electorate. *Weeks-to-election* is the relevant time unit. Our monitoring of the election begins 7 weeks before election-day, hence monitoring the period starting on the 18^{th} of September and ending on the 6^{th} of November. Assuming a worst case scenario where each multinomial parameter is time-independent (and hence we need completely new information on every time period to estimate these accurately), we sample N as above for every week within our monitoring frame, for a total of $N = 11,270$ individuals. The effective weekly sample size is further inflated with respect to the Thompson number to account for sample loss as a result of poor matching with the voter registry. While the assumption that the multinomial parameters are time-independent is certainly too strong, it will roughly be true for *swing* groups of voters. Treating each week as an independent sample should ensure the minimum sample size necessary to represent swing voters with reasonable precision. And stable partisans should be estimated precisely with this sampling strategy.

Procedures were implemented to avoid bias from systematic day-of-the-week effects on clustering of preferences[12]. We spread the collection of partisan voters evenly over the days of the week[13]. Initially we sampled, daily, 5 digital traces for each of the 68 FB pages (the total number of congressional and senatorial candidates). This generated a sample of $2,380$ traces per week. As the campaign progressed we increased the sample size in order to promote higher collection rates for swing voters from a subset of the pages. The figure above is however only "potential" given that Facebook pages are not active at the same rate. If no post is published on a given day, no user will have the opportunity to interact with the page. This creates a situation similar to non-response in

[12]For instance it is plausible that Fridays are a particularly bad day for the incumbent President's party, as the "Friday news dump" may lead to heightened coverage of scandals or critique of partisan policy.

[13]The supplementary material contains Routine 1 that describes the data collection steps in detail

an opinion poll. If this behavior is systematically correlated with voter characteristics used to estimate cell voting likelihood it can severely bias the estimates. Moreover, inactivity selectively reduces sample size, and makes it less likely for us to capture rare voter-types which may be over-represented in inactive pages.[14]

It is possible the same individuals leave multiple digital traces per week. For the following reasons we opt to consider each of these "draws" as a separate individual: i) we are not interested in individual-level predictions, but category-level; ii) some sub-categories are rarely sampled on social media, so this avoids wasting information; iii) some sub-categories are "swinging" during the campaign, and hence the same individual may leave different digital traces; iv) and because of post-stratification weighting, repeated traces will only be a problem if they are a consistently large outlier for the given cell.

In total, we sampled $15,683$ digital traces; 53.6% were Democratic and 46.4% Republican. Our conjecture is that observing an individual's digital trace is a particularly robust measure of preference. It is an unobtrusive revealed preference. And while, as we have seen in this section, constructing a virtual sample of these choices is challenging we believe that in many cases the benefits outweigh the costs. In order to implement the MRP, a necessary and challenging step is to match the virtual sample to the population.

2.2 Matching Sample to the Post Stratification Data

We propose a novel strategy for matching our virtual sample with specific population frame cells – a critical step in generating the post-stratification estimates. Our virtual sample includes a name, geographic location, possibly some limited demographic information and of course their partisan digital trace – whether they were observed on a Democratic or Republican congressional candidate Facebook page.[15] In order to match these digital preferences to the Texas population frame we need richer socio-demographic information on the individuals.

[14]Most pages we sample from correspond to a congressional district. This was done to ensure the interacting users would be expressing their preference over the congressional vote, as opposed to the senatorial or gubernatorial ballot. We did not include a congressional district variable in our likelihood estimation. If any district-candidate page is systematically less active than that of their district rival, district co-variates would capture artificial variation due to heterogeneous page-activity rates, and activity could be confounded with voting preferences.

[15]This information was collected systematically on a daily basis by a research assistant who was provided with a data collection protocol.

In our case, we get these richer profiles by matching our social media sample to the state-wide Texas L2 voter registration file.[16] This file has individual-level data for over 13 million registered Texas residents, ranging from voting history to socio-economic and demographic characteristics. Matching to the L2 file allows us to identify potentially non-registered voters and to generate turnout weights based on historical voting patterns. These administrative records of voters will have missing and incorrect data. A recent Pew Research study assesses the completeness and accuracy of the various variables included in these commercial voter registration files. (Igielnik et al. 2018) While there are clearly some short-comings to these data, for the purposes of post stratification they represent one of the best resources available.

There are other strategies for obtaining this socio-demographic information for individuals in the digital trace sample. Ideally, these predictive characteristics should also be collected unobtrusively and inexpensively. But having observed the outcome variable unobtrusively, one option would be to conduct conventional, obtrusive, surveys to obtain this information.

We obtain three variables from the Facebook individual profiles that can match to the L2 voter history file. Name is one variable. We use the `R` (Team et al. 2013) package `humaniformat` (Keyes 2016) to parse the Facebook names and obtain individual entries for *First Name*; *Last Name*; *Middle Name* and *Name Suffix*; this allows us to exploit the high level of name detail of our voter list counterparts. Secondly, we match Facebook profiles to the *City of Residency* record in L2. The closest social media counterpart to this variable is the *Current City* entry. When this is missing we use *Home Town* as a proxy. The location and name variables are recorded with considerable error. The social media source data has spelling mistakes and inaccuracies. These are further exacerbated by entry mistakes by research assistants who manually collect these characteristics. Because of the noise in these data we do not attempt perfect-matching on these variables. The final variable we match on is *gender*, which is unambiguously dichotomous in both datasets.

With the `R` package `fastLink` (Enamorado et al. 2018) we match our virtual sample to the voter registration list, based on names, sex and city of residency. `fastLink` leverages a probabilistic linkage Bayesian Mixture model, where similarity between variables

[16] L2 is a major supplier of voter data; see `https://www.l2political.com` for details.

can be a function of the Jaro-Winkler (JW) string-distance. It assigns a probability of match to each row in the first dataset, with respect to each row in the second. The model imputes missing values from the posterior distribution which enables us to match records even if entries for a given variable are missing in one or both datasets (in which case the similarity score between the remaining match-variables will determine the success of the match). The package is flexible enough to allow for mixed perfect-imperfect matching strategies; for instance, we match perfectly on *Gender*, and via string-distance on *First name*, *Last Name*, *Middle Name*, *Name Suffix*, *City of Residency*. For computational reasons, we set a high string-distance match threshold; namely, two strings are considered matched if the JW similarity is larger or equal to 0.99 (where 1 is an exact match). This allows us to match while ignoring minor inaccuracies. The overall matching threshold is set to the default 0.85, i.e. if the posterior probability of a match is equal to, or above, 0.85, the function returns the row index of a plausible match on the voter registration file, otherwise it does not.

From our social media sample of $15,683$ digital traces, we successfully match $8,278$ traces from $4,475$ users to an entry in the voter-registration – this is a success rate of 52.8%. The matching uncertainty can be quantified by varying the matching threshold from 0.75 to 0.99; this changes the number of matches from $8,739$ to $7,322$, highlighting $1,417$ borderline cases. The False Discovery Rate (FDR), which is the proportion of false matches having overall posterior matching probability higher than the given threshold, is well controlled at 1.2%, suggesting the resulting matches are of high quality[17]. Holding the FDR close to zero is important – given our social media sample is quite small, we cannot afford to have the matching introduce further noise.

This match rate of 53 percent is respectable considering the identification information for each of our FB records can be quite noisy. Pew (Igielnik et al. 2018) reports the rate at which their American Trends Panel is matched to five different commercial voter files. In this case the identification information for each record is much cleaner. They report

[17]We do not report the False Negative Rate (FNR) calculated by the `fastLink::summary` function because the FNR is calculated relative to all potential pair comparisons, which leads to numbers quickly approaching 100% as the datasets become more imbalanced and less overlapping. The alternative `fastLink:confusion` function was computationally not feasible because the sheer size of our potential matches number in the billions. The function would in principle provide an FNR value relative to the reference population, which in our case is the social media sample. A visual inspection of the matches, as well as the outcomes of our predictions based on the matched sample, suggest the matching was high quality.

match rates of between 50 and 79 percent. A concern here is that we cannot assume that the unmatched records are "missing at random". As Pew (Igielnik et al. 2018) suggest in their analyses of commercial data files, the unmatched records in their case were systematically younger, more mobile, less politically engaged, and less partisan. Hence our inability to match higher proportions of our digital sample could introduce bias into our vote preference estimations.

To assess whether digital traces accurately reflect political preferences, we assess whether those digital traces matched to the L2 voter file have a digital partisanship that a) corresponds to the imputed partisanship available in the L2 voter file; and b) matches the partisanship of the primary in which these individuals voted (if they cast a primary ballot).[18] Of those subjects with Republican digital traces, 71 percent were identified Republicans in the L2 database; a similar 83 percent of subjects who have a Democratic digital trace are Democrats in the L2 files. 88 percent of Republican digital traces voted in a Republican primary and 91 percent of Democratic digital traces voted in Democratic primaries. There is a high correlation between digital traces and partisan measures from our commercial voter registration files.

In a similar exercise Pew (Igielnik et al. 2018) find that between 62 and 85 percent of their American Trends Panelists who self identify as Democrat are identified as Democrat in the five commercial files they evaluate. For the Republicans, between 55 and 71 percent of those self-identifying in the ATP were identified as Republicans in the commercial files. And this congruence in the partisanship of our digital trace data and registration records is consistent with other similar efforts to measure partisanship and ideology with Facebook and Twitter digital traces (Bond & Messing 2015, Barberá 2015). Hence, we are very comfortable concluding that digital traces represent revealed partisan preferences.

The frequency with which individuals leave partisan digital traces over the course of the monitoring period is also an indicator of partisanship. Those leaving multiple, as opposed to a single, digital traces for a party are more likely to be matched to the same party in the voter registration file. Out of the $2,130$ matched voters that had Republican digital traces, 574 (around 27%) liked/loved/expressed support for a Republican page more than once during the monitoring period. Of the $2,392$ matched voters that had

[18]The L2 voter file includes an imputed partisanship that is estimated by the commercial providers.

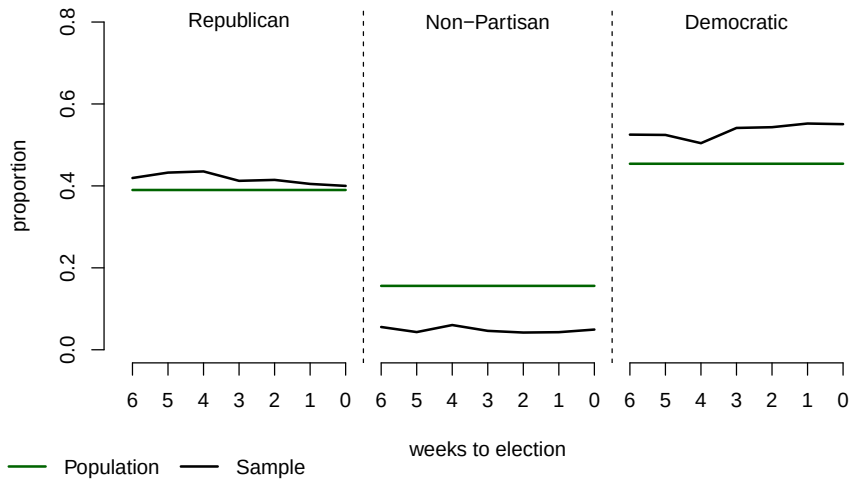


Figure 1: Population v. Sample comparison: partisanship.

Democrat digital traces, 716 (around 30%) liked/loved/expressed support for a Democrat page more than once. Amongst digital Republican supporters, the probability of being a Republican on the L2 file is 6.5% higher amongst individuals who left multiple pro-GOP digital traces, compared to individuals who left only one; similarly, amongst digital Democrats it is 4.5% more likely for individuals who left multiple pro-Democratic digital traces to be a partisan Democrat on the L2 file. Multiple likes seem to provide a cleaner signal of support, suggesting the frequency of digital traces can be leveraged in future work to obtain cleaner signals of support[19].

Our final sample size is $8,278$ digital traces. This is slightly below the Thompson number. Holding probability constant at 0.9 (that all estimates of the multinational parameters are within 0.025 of the population proportions), we are powered (assuming equal sample size across weeks), at the weekly level, to estimate voter categories that are larger than 2.9 percent of the total population.

Figure 1 and Figures A2 to A5 in the Supplementary Appendix compare the sociodemographic profile of our matched sample to that of the population at large. With respect to *Partisanship*, we overestimate the percentage of Democratic and Republicans in the population and understate the percentage of Independents. This is an understandable result of sampling from partisan Facebook pages. We underestimate the

[19]It was rare for voters to like both Democratic and Republican pages: we found that, out of the $4,475$ matched users, only 47 liked posts on both pages. These appear to be overwhelmingly Democrats (70%), although the sample is too small to make meaningful inference. In principle, these voters could have been 'window shopping' - visiting and liking multiple pages as part of their preference formation process. The number of voters who liked both sets of pages is quite small, suggesting that such a learning process is uncommon amongst users in our sample.

young and overestimate the old and slightly over-represent women. Our virtual sample over-represents whites and slightly under-represents minorities. The education profile of the virtual sample is quite close to that of the population.

Just to recap, we matched a small convenience sample of individuals visiting partisan Facebook pages to the L2 database of Texas registered voters. We assess the extent to which convenience samples of this type are similar, on key variables, to the actual population of interest. Our conclusion is that reasonably calibrated virtual samples can be quite representative of the overall population.

3 Area Estimates of Partisan Support

A critical step in the classic Multilevel Regression with Post-stratification (MRP) approach is to estimate a model of the outcome of interest based on a survey conducted with a sample of the relevant population (Wang et al. 2015, Buttice & Highton 2013, Ghitza & Gelman 2013, Lauderdale et al. 2020). Coefficient estimates from these survey-based models facilitate the area-level estimations. In our case, the area estimates are of partisan support for each of the 36 congressional districts in Texas. We do not rely on survey responses to estimate our model of vote preference. Vote preference is an unobtrusively measured digital trace. And the explanatory variables in the vote choice model are all individual-level measures obtained from the matched L2 population file.

Our novel prediction and post-stratification estimation strategy for the 36 Texas congressional districts has three crucial steps: i) the identification of voter characteristics from the L2 file that would be used in the vote choice prediction model; ii) estimation, at the voter-category level, of vote choice conditional on the probability of vote turnout; iii) the weighting of these estimates by cell counts and summing over the area of interest to recover estimates of support.

3.1 Variable Selection and Imputations

Our MRP implementation relies on the L2 data base to provide all of the information required for constructing the classic population frame and also most of the information necessary for estimating the vote choice and turnout models. We rely on the virtual sample primarily to provide the measure of vote preference. Once this virtual sample

is matched to the L2 data base we can then use the rich set of L2 variables to predict their revealed vote preference.

From L2 we construct a core data set, $\boldsymbol{X}$. Variable selection for $\boldsymbol{X}$ reflects the unique estimation strategy we propose here. We predict vote choice and turnout with a machine learning estimation strategy; specifically, we use random forests for both vote-choice and turnout. Hence, we impose no a priori functional form for the likelihood function for either models. We do, on the other hand, propose a slate of prospective variables that we anticipate are correlated with electoral behaviour. As Duch and Stevenson (Duch & Stevenson 2008) point out, there is considerable agreement amongst voting behavior scholars as to what set of variables best predict electoral behaviour. These core variables include age, gender, imputed religion, household income, race, community size, partisan profiles (primary and past vote history, for example), and education. And these are predictors typically incorporated in U.S. vote choice models (Lewis-Beck et al. 2008).

We include in $\boldsymbol{X}$ additional variables from the rich L2 file – in total approximately 50 variables. Many of these $\boldsymbol{X}$ will serve two other purposes in the estimation – impute missing values and predict voter turnout. Our search for these 50 candidates was motivated by the literature cited above – in particular, cells were built with the variables mentioned above, while strong correlates with these were included to motivate imputation and turnout models. Table A2 in the appendix presents the full range of variables introduced in each model of electoral behaviour and in the imputation model. Note that our vote choice model is at the cell-level, while our turnout model is at the individual level. For the purpose of producing turnout estimates like in Figure 6, we simply sum the individual-level turnout probabilities across all the registered voters; the introduction of the roughly 50 individual level variables is justified by the desire to obtain better individual-level turnout propensities, and hence improve area-level estimates of turnout. In summary, the selection of these additional variables was guided by how well we anticipated they would predict turnout and perform in the imputation model.

For each h registered voter in data set $\boldsymbol{X}$ we have $\boldsymbol{x}_h$. Missing values in $\boldsymbol{X}$ are imputed with a random-forest multiple-imputation strategy implemented via the packages `ranger` (Wright & Ziegler 2015) and `missForest` (Stekhoven & Bühlmann 2012). These provide a flexible non-parametric framework for imputing mixed-type data. Imputation error is estimated via calculating the Out of Bag (OOB) error for each imputed covari-

ate, at each iteration. OOB error is roughly zero for all imputed variables suggesting that, if the assumption that data is *Missing at Random* holds, the observed data contains nearly all the information needed to complete individual records. This results in a completed dataset $\boldsymbol{X}^{\mathrm{I}}$ with voter-specific characteristics as $\boldsymbol{x}_h^{\mathrm{I}}$. From $\boldsymbol{X}^{\mathrm{I}}$ we derive two datasets: $\boldsymbol{Z}^{\mathrm{I}}$, indexed by h as above and $\boldsymbol{X}^{\mathrm{I}^+}$, indexed by expanded voter identifier h^+ .

Post Stratification Dataset. All observations from the L2 voter registration file are included in the population frame. Hence, $\boldsymbol{Z}^{\mathrm{I}}$ is of the same length as $\boldsymbol{X}^{\mathrm{I}}$. The covariates in $\boldsymbol{Z}^{\mathrm{I}}$ are modified. First, they are reduced in number. Secondly, the covariate values are recoded so they uniquely identify voter-categories of interest. This results in a feature space, m, that is simply the product of the number of covariates and covariate values. We define the voter category C_g, for categories $g = 1, ..., G$ as a unique realization of the set of variables which compose $\boldsymbol{Z}^{\mathrm{I}}$, i.e. $C_g = \{Z_1 = z_1, ..., Z_m = z_m\}$; then for a voter h in the registry, $h \in g$ if $\boldsymbol{z}_h^{\mathrm{I}} = C_g$. Table A2 in the appendix shows the voter characteristics chosen, along with their size in absolute numbers and as a proportion of the population. The feature space must be shrunk to m to ensure we have enough power to accurately represent the resulting categories' vote likelihood (recall from Thompson (Thompson 1987) we are powered to capture groups of voters which make up roughly 2.9% of the registered electorate). The resulting voting categories are then used to characterize the cells of our population frame.

The voter categories $\boldsymbol{C}$ will not include a geographical identifier, such as congressional district number, as these will be systematically correlated with non-response. Nevertheless, the cells of our stratification frame will be defined by the interaction between categories $\boldsymbol{C}$ and congressional districts; we will then allow the district-category counts in our population frame (i.e. the number of individuals from category g which inhabit district $D = d$ for $d = 1, ..., 36$) to weight our category-exclusive (non-district) likelihood and produce different district-category predictions. We note again that two districts with exactly the same demographic composition, under this set-up, will have the exact same predictions.

Turnout data. We do not have a digital trace measure of turnout likelihood. To estimate voter turnout probabilities we therefore rely on individual level historical voting

behavior from the voter-registration file. To model the voting record of individuals over multiple years, the voter registration dataset is "expanded" such that each voter is "observed" over each of these years. $\boldsymbol{X}^{\mathrm{I}^+}$ is the expanded version of the completed voter sub-space, deployed for estimating turnout probabilities; it is constructed by using the L2 covariate "voted in year ____" as outcome variable, transforming $\boldsymbol{X}^{\mathrm{I}}$ into *long* format since, for any individual, we may observe their turnout in multiple elections.

3.2 Likelihood Prediction Models

The likelihoods predicting vote choice will employ these two datasets, $\boldsymbol{Z}^{\mathrm{I}}$ and $\boldsymbol{X}^{\mathrm{I}^+}$, with complete imputed values. We will omit the imputation overscript "I" and simply refer to relevant completed datasets as $\boldsymbol{X}$ and $\boldsymbol{Z}$. We estimate the joint probability of an individual in category g voting for the Republican candidate ($R = 1$) and turning out on election-day ($T = 1$) (Lauderdale et al. 2020):

$$\mathrm{P}_g\left(R = 1, T = 1 | C\right) = \mathrm{P}_g\left(R = 1 | T = 1, C\right) \times \mathrm{P}_g\left(T = 1 | C\right) \tag{2}$$

We again rely on the `ranger` package to implement a probability machine (Malley et al. 2012) composed of B^T trees. The terminal node of each tree will represent the relative frequency of respondents having the relevant voter-category characteristic. This is a random forest tuned to have probabilities in the terminal nodes of each of its trees, and standardized estimates producing an output-matrix with elements between zero and 1, and rows summing to 1. Random forests improves MRP estimation in two respects. First, the forest estimation outputs the best[20] non-linear function of its inputs without us imposing a-priori functional specifications. If the input is not *important* (Janitza et al. 2018) it will be consistently ignored by each decision tree. Second, from a practical standpoint, ranger-implemented forests are faster than almost any other machine available – given the size of our population frame, this was an important consideration.

3.3 Voter Turnout Estimation

Conditioning vote probabilities on the likelihood of turning out in an election is a critical element of area forecasts of election outcomes. Recent elections, such as the 2016 U.S.

[20] According to minimization procedure of an out-of-bag prediction error

Presidential race, have highlighted the importance of estimating accurate turnout probabilities (Kiewiet de Jonge et al. 2018). Efforts to model turnout probabilities typically rely on self-reported vote (either history or intention). Their advantage is that they capture the short-term temporal dynamics in turnout propensities. The Kiewiet et al (Kiewiet de Jonge et al. 2018) turnout model based on 2016 pre-election polls better predicted the demographic composition of voter turnout in the Presidential election than did models based on the 2012 Current Population Survey data. Given the importance of turnout composition for the 2016 Presidential election this explains in large part their ability to accurately predict electoral vote wins in all but one state.

Our turnout model relies exclusively on the L2 Texas voter registration data files. Voter registration files provided by state election authorities do not rely on reported turnout (Keeter et al. 2016). They avoid the well-known over-reporting of turnout in surveys (Burden 2000, McDonald 2014, Keeter et al. 2016) that can bias estimated turnout probabilities.[21] And there is evidence to this effect. Pew Research, for example, compared 2016 Presidential vote predictions weighted by turnout probabilities based on self-reports compared to commercial voter files. Predictions based on commercial voter registration data outperform those that employ self-reported voting turnout weights (Igielnik et al. 2018). These administrative files also have their drawbacks – examples include missing and erroneous data and the purging of the historical voting record of voters (Presser 1990, McDonald 2014). Another disadvantage is that they are historical and hence do not help us estimate current, real-time, changes in the turnout propensities of key demographic categories.

We estimate, the probability of turnout, $\mathrm{P}_h\left(T=1|\boldsymbol{x}\right)$, for all individuals in the complete voter file $\boldsymbol{X}$, where T is a random variable taking value 1 if the individual casts a ballot on election-day and 0 otherwise. P_h, is used as observation-weight in the estimation of the vote-choice likelihood, for each of our matched social media voters. Adjustments are taken to exclude voters who would have not been eligible to vote in earlier years; time-dependent characteristics are adjusted to reflect the year of the outcome variable; time-independent characteristics, such as race or sex, but also income or education, are assumed to stay constant over the years.

[21] Of particular concern is that this over-reporting reflects systematic sampling biases or variations in social desirability that are correlated with the demographic predictors that are included in models of turnout probability.

To estimate P_h, we train a random forest on the expanded turnout dataset $\boldsymbol{X}^+$; the forest's estimate of the probability of turning out on election day conditional on voter characteristics and being registered is defined as follows:

$$\hat{\mathrm{P}}_{h^+}\left(T=1|\boldsymbol{x}^+\right)=\varphi^T\left(\boldsymbol{x}^+_{h^+}\right)=\frac{1}{B^T}\sum_b^{B^T}\tau_b^T\left(\boldsymbol{x}^+_{h^+}\right); \tag{3}$$

where φ^T represents the point estimate of a probability machine trained to predict turnout probabilities, and whose value is the average of B^T probability trees τ_b^T. Having trained the model, we use it to output a prediction for the the probability of turning out for each member of the voting population:

$$\hat{\mathrm{P}}_h\left(T=1|\boldsymbol{x}\right)=\varphi^T\left(\boldsymbol{x}_h\right); \tag{4}$$

We extract the turnout probabilities of voters in our matched social media sample $\hat{\mathrm{P}}_s\left(T=1|\boldsymbol{x}\right)$ indexed by $s=1,...,S$ where $s\subset h$. This quantity will be used as the observation-weights in our vote-choice model, effectively conditioning that distribution on turnout.

We calculate the empirical error distribution using the MSPE1 procedure from Lu (Lu 2017); this is designed to calculate the global mean-squared prediction error. We then assign a Normal distribution to the prediction error, and use the Root Mean Squared Predictive Error (RMSPE) as an estimator of the variance. Though the Normal distribution does not characterize the empirical distribution perfectly, it is useful to obtain reasonable prediction intervals; hence we describe the predictive distribution of turnout as follows:

$$\mathrm{P}_h\left(T=1|\boldsymbol{x}\right)\sim\mathrm{N}\left(\varphi^T\left(\boldsymbol{x}^{\mathrm{I}}_h\right),(\hat{\sigma}^T_{\mathrm{RMSE1}})^2\right). \tag{5}$$

The vote choice likelihoods in the area forecasts will be estimated for the g voter-categories. The probability of vote choice in these likelihoods will be conditional on the probability of turnout: $\mathrm{P}_g\left(T=1|C\right)$. P_g is the category-level probability of turning out on election day, which will be needed to calculate Equation 2. It is derived from P_h. We simply average across the voter categories identified by $\boldsymbol{C}$ to obtain category-level

estimates of turnout probabilities, as follows:

$$\mathrm{P}_g\left(T=1|C\right)=\frac{1}{\sum_{h\in g}1\left(\boldsymbol{z}_h=C_g\right)}\sum_{\substack{h\in g\forall \boldsymbol{z}_h=C_g;\\ \boldsymbol{z}_h\in\boldsymbol{x}_h}}\mathrm{P}_h\left(T=1|\boldsymbol{x}\right); \tag{6}$$

where the outcome quantity is the average across a number of simulations from the predictive distribution of $\mathrm{P}_h\left(T=1|\boldsymbol{x}\right)$.

3.4 Estimating Vote Probabilities

To estimate vote-choice probabilities we specify the outcome variable as the partisan digital trace R_s of voters $s=1,...,S$ where $R_s=1$ if the trace is Republican and 0 if Democrat. In total we have 8,278 individual traces from our digital sample that are matched to the voter registration data. With this matched sample, we train a probability machine as above to estimate the probability of voting Republican, conditional on the individual turning out and the set of their voter characteristics:

$$\hat{\mathrm{P}}_s\left(R=1|T=1,\boldsymbol{z}\right)=\varphi^R\left(\boldsymbol{z}_s|\hat{\mathrm{P}}_s\left(T=1|\boldsymbol{x}\right)\right) \tag{7}$$

Here, $\boldsymbol{Z}$ includes an identifier for each week to election $W=w$ for $w=6,...,0$; an identifier $L=l$, where $l=\{1,0\}$ indicating whether the trace at hand belongs to a state-wide or district level election; and the digital trace $R=r$ where $r=\{1,0\}$, which is our outcome variable. These new identifiers are also added to the categories $\boldsymbol{C}$ used in the area forecasts. Note that the category-conditional turnout estimated previously is assumed to be constant across W and L.

Our next step in generating the area forecast is to assign a vote choice probability to each of the 1,080 cells in the Texas population frame. The trained forest is used to provide category-level predictions; for category g such that $s\in g$ if $\boldsymbol{z}_s=C_g$:

$$\hat{\mathrm{P}}_g\left(R=1|T=1,C\right)=\varphi^R\left(C_g\right). \tag{8}$$

Again, we estimate the global OOB Root Mean Squared Error MSPE1 procedure and encounter an empirical distribution which is approximated by a Normal probability

distribution function:

$$\mathrm{P}_g\left(R=1|T=1,C\right) \sim \mathrm{N}\left(\varphi^R\left(C_g\right), (\hat{\sigma}^R_{\mathrm{RMSE1}})^2\right); \tag{9}$$

This section proposes a novel approach to the model estimation stage of classic MRP area estimation. First, the outcome variable, vote preference, is observed, unobtrusively, rather than self-reported. Second, these observed choices by individuals are matched directly to the population frame that includes an extensive set of explanatory variables, again not self-reported, that are employed in the prediction model. Third, the probabilities of vote choice for each of the population cells is estimated with a random forest machine that imposes no a-priori functional form on the model specification. This generates the two quantities needed to estimate the cell-level joint probability of voting Republican and turning out, as specified in Equation 2.

3.5 Aggregation: Vote and Seat Predictions

The last step in the area estimation process is to generate congressional candidate vote predictions for each of the 36 congressional districts and vote predictions for the overall Senate race in the state of Texas. Accordingly, we have two kinds of aggregation to perform: district level and state-wide. The state-wide category level population counts are just the number of voters belonging to each category in the voter registration file: $Q_g = \sum_h 1(\boldsymbol{z}_h = C_g)$. The state-wide area estimate for the vote share of the Republican party in the Senate election, with W weeks left in the campaign, is calculated as follows:

$$V^R_w = \frac{\sum_g \mathrm{P}_g\left(R=1,T=1|C,L=1,W=w\right) \times Q_g}{\sum_g \mathrm{P}_g\left(T=1|C,L=1,W=w\right) \times Q_g}. \tag{10}$$

At the district level we count the number of voters in the intersection between a given category and a given district $D = d$, $\forall d = 1, ..., 36$: $q_{gd} = \sum_h 1(\boldsymbol{z}_h = C_g \cap D_h = d)$. The district level estimates are the product of the following calculation:

$$v^R_{dw} = \frac{\sum_g \mathrm{P}_g\left(R=1,T=1|C,L=0,W=w\right) \times q_{gd}}{\sum_g \mathrm{P}_g\left(T=1|C,L=0,W=w\right) \times q_{gd}}. \tag{11}$$

For each of the 36 Texas congressional districts this gives us a predicted vote (and vote share) for the Democratic and Republican candidates.

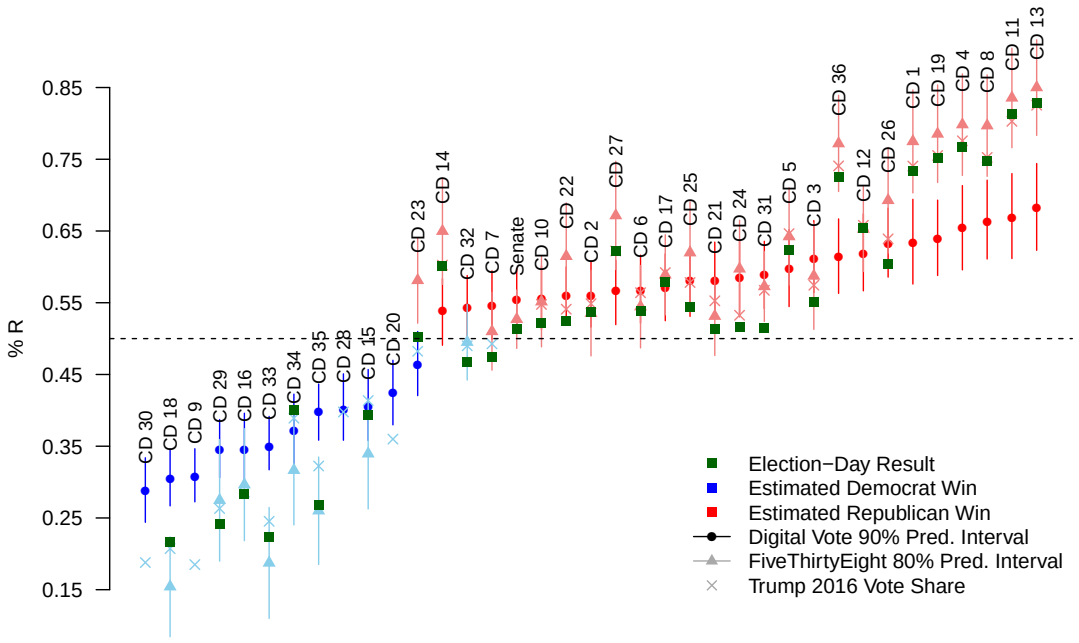


Figure 2: Election-week estimates of support for the Republicans, by area of interest; a comparison with `FiveThirtyEight.com` and actual election results is provided.

4 Results: Benchmarking Estimates

Figure 2 presents the Digital Vote Senate and congressional predictions for the election week period. Thirty-two of the 36 congressional districts fielded candidates from both major parties; 28 of these were seen as not particularly competitive prior to the election[22] – and in these cases our calls match the actual election outcome. There were four a-priori competitive districts: Districts 7, 32, 23 and 21; District 7, and elected Democrats while we predicted Republican wins; District 23 elected (by a very small margin) a Republican whilst we predicted a Democratic win. We called District 21 correctly for the Republicans. The senate was also correctly predicted by our method. Of the 32 contested congressional races, we called 29 correctly. We also called the Senate race correctly. On balance this suggests our method is able to predict pluralities in most circumstances.

Of particular interest is how the Digital Vote weekly predictions for each of the Texas congressional districts compared to traditional survey estimates of party vote share. `FiveThirtyEight.com` aggregates public opinion surveys being conducted during this period and publishes daily forecasts of the vote shares for the Republican and

[22]Some races, such as district 24 and 31, ended up being more competitive than anticipated, as can be seen by the large error that `FiveThirtyEight` collected on these.

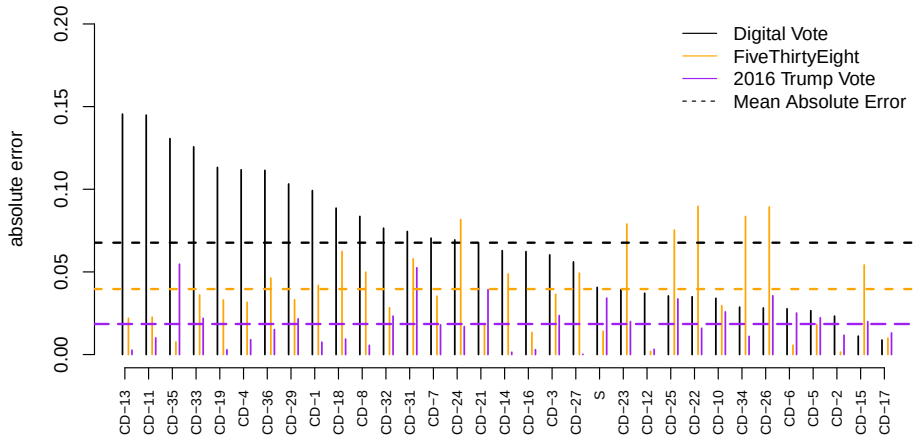


Figure 3: Mean Absolute Error: FiveThirtyEight, Trump 2016 and Digital Vote.

Democratic congressional and Senate candidate contests in Texas. The FiveThirtyEight "light" model is based on a proprietary aggregation of the latest public opinion polls. We focus on the "light" version of the site's forecasts because it has no other information beyond polls and hence provides a more direct comparison to our method that emulates opinion polls using a digital sample.

The election week comparisons in Figure 2 suggest that the FiveThirtyEight aggregated results perform similarly to ours in the most competitive races: they call incorrectly call Districts 7 and 23, but correctly call Districts 32 and 21, as well as the Senate race. In terms of headline number of correct calls, they perform better than us by a single seat, with 30 contested congressional races and the Senate called correctly. Where our predictions fall short is in comparing Mean Absolute Errors (MAE).

Figure 3 summarizes the MAE for the two methods over the 36 congressional districts and the Senate seat. It also includes the MAE for a hypothetical forecast based on the 2016 two-party Trump share. These are based on the differences between the estimated two-party vote shares and actual vote share observed on election day.

There are two important messages in Figure 3. First, overall, prediction and post-stratification forecasts based on our virtual sampling strategy have high Mean Absolute Error (MAE). Over all 32 districts where both major parties fielded candidates, the MAE is about 6.8 percentage points. The virtual sampling strategy does not perform as well as the FiveThirtyEight forecasts when we only consider Mean Absolute Error

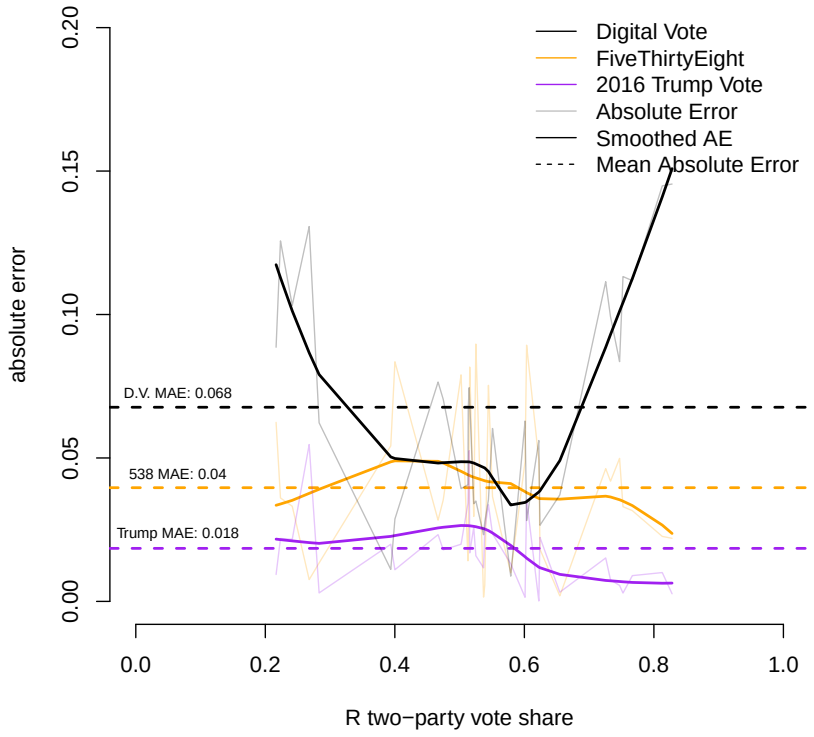


Figure 4: Mean Absolute Error, ordered by size of the Republican two-party vote share.

– their overall average is about 4 percentage points. FiveThirtyEight predictions are closer to the election-day vote share outcomes than those of Digital Vote. As mentioned earlier, the 2016 Trump vote obtains the lowest MAE.

Second, the Digital Vote suffers from significant attenuation bias. This is clear in Figure 4 where we plot the absolute error ordered by size of the Republican vote share. A rough U-shape emerges around the state-level predicted mean, clearly suggesting that absolute error is increasing where the voting population is less representative of the state average. Note though that the Digital Vote does perform significantly better around the state's predicted mean, suggesting if we could address the attenuation bias, our performance would improve dramatically.

Sub-category estimates. Also of interest is the accuracy of the category-level predictions for our $1,080$ sub-categories of voters. We compare the category-level predictions derived from our sample of $8,278$ digital traces to those derived from regular surveys, holding modeling assumptions constant. We use two publicly available surveys from the *Texas Politics Project*[23] as benchmarks. In each poll $1,200$ respondents were surveyed; the first poll was conducted between the 7^{th} and the 15^{th} of June; the second between the 15^{th} and the 21^{st} of October. These surveys are only relevant to the Senate race. We note that this survey sample, though it includes two separate polls, is heavily underpowered to accurately estimate the $1,080$ cells in our model. Nevertheless, a comparison with our sample should show reasonable convergence on average, allowing for large sampling variability. To keep modeling assumptions as constant as possible, we use the reported voting history to condition vote-choice on turnout. The comparison is made between estimates for the 2^{nd} *week-to-election*, which is when the latest data is available for the surveys. The surveys included measures that had categories corresponding to those in Digital Vote, for Partisanship, Race, Age and Education. The exception is 'Vocational Training', which is an education category not surveyed by these polls. We can therefore produce comparisons for 900 voter sub-categories. Figure 5 compares estimate Republican votes derived by the digital trace to those observed in the Texas Project surveys.

Observations falling exactly on the 45 degree $y = x$ line represent sub-categories for

[23]The Texas Politics Project's polling data archive is available at: `https://texaspolitics.utexas.edu/polling-data-archive`.

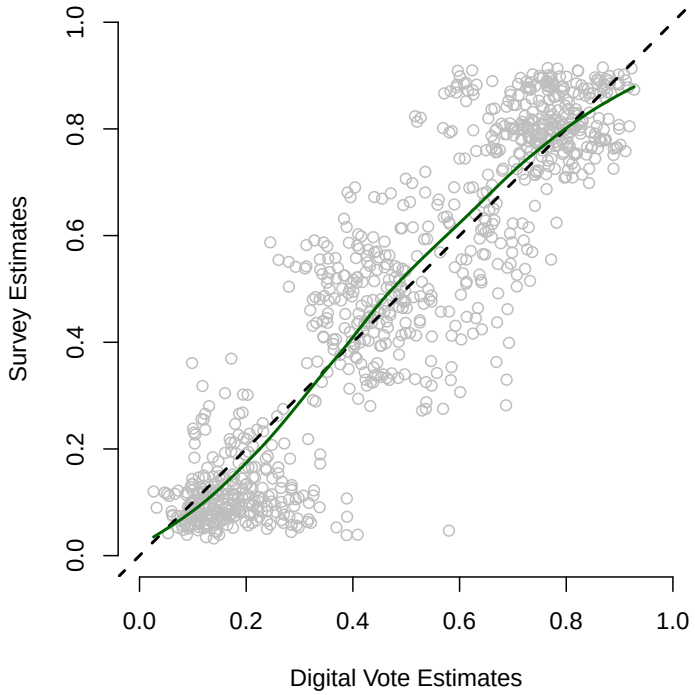


Figure 5: Voter sub-categories comparisons: Surveys v. Digital Vote.

which the Texas Survey Project and Digital Vote predictions were identical. The plot indicates that Digital Vote matches closely the Texas Survey estimates; a slight over-estimation of the Republican vote seems to be concentrated around more Democratic constituency. This is consistent with the results of the area-estimates comparison in Figure 2. A LOESS curve fit to the data closely tracks the $y = x$ line suggesting that, at the sub-category level, digital traces reflect actual voting intentions.

Turnout. We estimate turnout probabilities from voting histories in the voter-registration file. Turnout as percent of registered voters in the Texas 2018 mid-term election was 53%; this represents an increase in excess of 19 percentage points over the previous 2014 mid-term election. As a result our turnout model that relied primarily on historical administrative data underestimated voting turnout. At the state-level we predicted 40.4% of registered voters to turnout while the actuals were 53.0%. Figure 6 benchmarks our district turnout forecasts against actual turnout statistics for the 36 districts and the overall state tally. The LOESS curve we fit to the data summarizes the strong correlation between our forecasted and actual turnout in the 36 districts. But for all districts actual turnout rates for 2018 were consistently higher than those predicted by our turnout model. Hence the LOESS curve lies above the 45 degree line. We also plot the

actual 2018 district turnout percentages against turnout in the previous 2014 midterm election. Historical mid-term turnout significantly under-estimates the actuals for 2018. Consistent with the historically high 2018 mid-term turnout, we see in Figure 6 that turnout in the 2016 Presidential election is in fact a good predictor of 2018 turnout.

In our model, predicted vote choice, for any cell in the stratification frame, is conditioned on the likelihood of turning out to vote. While our turnout model performs poorly this is particularly problematic if there is systematic under-estimated turnout for particularly cells in the stratification frame (say, young female African Americans). We have some evidence from Texas Senate race exit polls suggesting this was not the case:[24] For every 100 self-identifying Republicans who cast their ballot in the Senate race, 83 Democrats turned out. In our digital sample, we predicted that 77 Democrats would show up to vote for every 100 Republicans who turned out. The digital sample had a small over-estimation of Republican turnout; but far short of the 19 points turnout difference with 2014. Our conjecture is that the sharp rise in turnout was reasonably balanced across groups predisposed to vote Republican versus Democrat.

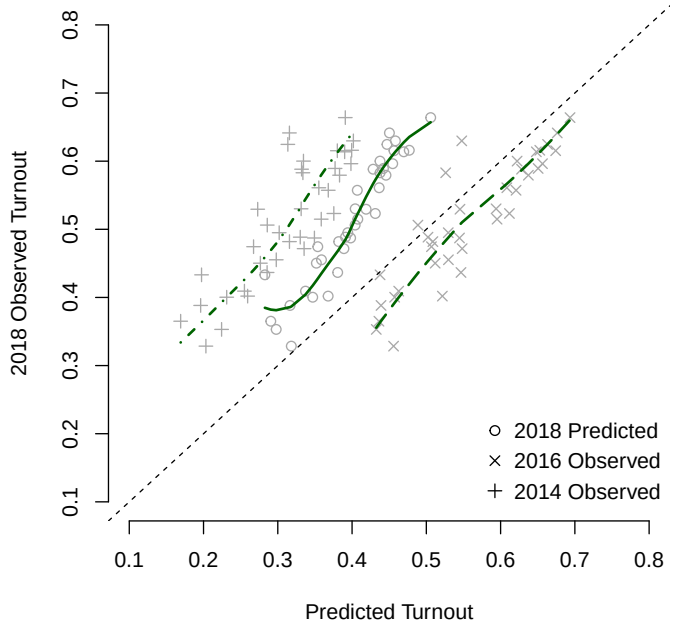


Figure 6: 2018 Turnout Predictions: Digital Vote versus 2018, 2016 and 2014 Actual Turnout.

[24] https://www.washingtonpost.com/graphics/2018/politics/voter-polls/texas.html

Costs and Benefits. Generating our virtual sample of digital traces required monitoring visits to candidate Facebook pages and extracting relevant information from these pages. On balance the costs of generating virtual samples are comparable to those of conventional surveys. Of interest here is the cost of achieving comparable levels of predictive accuracy using our novel digital trace procedure relative to conventional survey methods. One metric is simply to ask what effective sample size would be required to generate the predictive accuracy of the Digital Vote and FiveThirtyEight estimates. Accordingly, we compare our sample size with that of hypothetical probability samples calibrated to achieve the same level of accuracy. We generate these hypothetical comparisons by taking the observed point estimate from either Digital Vote or FiveThirtyEight and treating this as an informed prior for the purposes of calculating the sample size.

The sample size calculation for a proportion, in our case the share of the two-party vote going to the Republicans, is $n = \left(\frac{z_{\alpha/2}\sqrt{\pi(1-\pi)}}{d} \right)^2$ with π set to a point estimate, α set to 0.05 and d representing the desired margin of error. For each of the competitive races we forecast (where both parties were fielding candidates) we set d to equal the actual absolute errors from Figure 3, and calculate the sample size that independent probability surveys would have needed to achieve the same level of accuracy we did in each race. We generate these hypothetical sample sizes using the errors from both the Digital Vote and FiveThirtyEight. Sample size increases exponentially as the desired margin of error decreases – for example, the sample size required for a margin of error of 0.005 is 40,000. In some instances both our digital forecast and FiveThirtyEight have errors smaller than this implying sample sizes that could exceed half-a-million units. Hence, we cap the true effective sample size at $n = 9600$, or namely the effective sample size for a poll with a desired margin of error of $d = 0.01$ under the worst-case-scenario $\pi = 0.5$. This lower-bound value for d is chosen to represent the maximum level of a-priori desired accuracy. Table A3 in the Supplemental Appendix presents the hypothetical sample sizes for each district. Summing over the sample size needed for each district to achieve the observed level of accuracy, we find that independent probability surveys would have had to sample over $70,000$ individuals to achieve FiveThirtyEight's level of accuracy, and around $30,000$ individuals to achieve ours. FiveThirtyEight's level of accuracy is just under 2.5 times that of the Digital Vote, in terms of hypothetical sample size.

A conservative metric for assessing relative costs are those polls specific to Texas that FiveThirtyEight included in their model. These Texas polls conducted during the monitoring period would most strongly inform the FiveThirtyEight Texas congressional and Senate seat forecast (Silver 2018*b*). These consisted of a total of 22 house surveys of over 11,000 respondents; 26 Senate surveys of over 25,000 individuals.[25] FiveThirtyEight's aggregated estimates are less than 2 and a half times more accourate than our digital forecasts, for the Texas midterm district and Senate elections (the $30,000$ versus $70,000$ level of accuracy from Table A3). If we just consider FiveThirtyEight's house surveys then Digital Vote achieves this comparable level of precision with 4,475 unique Facebook users which is about 40 percent of the respondents in these 22 house surveys.

A potential benefit of these digital forecasts is that they provide added information to conventional forecasting strategies. An election-week regression of the results for the 32 contested congressional races and the Senate on the FiveThirtyEight polls-based predictions and our own forecast reveals a small, insignificant and negative coefficient on the Digital Vote. We conduct a similar regression estimation substituting simply the Trump 2016 vote for the FiveThirtyEight polls-based predictions – we do not add additional information to the Trump-based forecast (see Table A4 in the Appendix). We also simply averaged our forecast with other available predictions on the assumption that forecasts may capture different information and hence will contribute to reducing errors; as a result, predictive power would be increased (Graefe et al. 2014). In fact, averaging either the Trump vote or the FiveThirtyEight forecast with our own does not make the forecast significantly better.

This forecasting exercise provides three insights into the costs and benefits of forecasting election outcomes with digital trace data. Digital trace data generate reasonably accurate predictions of congressional district vote shares. Secondly, the effective digital trace sample sizes required to attain these levels of precision are comparable to those of conventional surveys. Thirdly, the incremental information content of these digital predictions is limited. An interesting question is whether changes in design and estimation might enhance the predictive value added of digital trace approaches.

[25] These figures refer to the number of surveys conducted during our monitoring period.

5 Digital Challenges

This project explore strategies for incorporating unobtrusively measured digital traces into models predicting behavior. Our proof of concept is the vote for congressional and senate candidates in the 2018 mid-term elections in Texas. This is an ideal application because the accuracy of predictions can be calibrated against post-election outcomes. Digital traces predict pluralities at the district level; and individual level digital traces are highly correlated with actual voting behavior. On the other hand, our digital trace predictions incorporate little additional information above and beyond conventional forecast methods.

There are methodological challenges associated with collecting and modeling digital trace data. We briefly reflect on what can be done to enhance the precision of digital trace forecasting.

The adoption of a random forest estimation approach contributes to the precision of our forecasts. It allows us to minimize the number of parameters used to fit the model - each tree does not split at a given covariate unless the split reduces the out-of-bag mean squared error. At the same time though the tree forecasts are averaged to produce the forest prediction - effectively applying shrinkage toward the overall mean. MRP models though typically shrink coefficients towards the overall mean which tends to produce attenuation (Lauderdale et al. 2020). On the other hand, random forests are capable of identifying and modeling non-linear multi-way interactions (Denisko & Hoffman 2018), suggesting that our estimation strategy does capture relevant deep voter category effect. Hence, the estimation algorithm itself is not contributing to attenuation bias.

Attenuation results from omitting district-specific effects in our estimation procedure, hence precluding a-priori the identification of local heterogeneity. A solution to the attenuation bias is to include more information about the areas being forecast in the estimation; this is something that is regularly done in MRP applications to conventional surveys, and it has been shown to improve area estimates (Hanretty et al. 2018, Buttice & Highton 2013, Lax & Phillips 2013, Ghitza & Gelman 2013).

We can improve the random forest estimation by including in the model complex interaction effects between area and demographics. For example, we would want to understand how the correlation between gender and digital vote preference differed in

rural Republican-dominated districts compared to urban Democrat-dominated districts. This would require sampling larger amounts of area-specific digital trace data than we do in this exercise. As we pointed out earlier, our individual-level model of digital vote preferences is only sufficiently powered at the overall state level – hence, for example, for the gender effect we do not differentiate between rural Republican and urban Democrat districts. Our sampling based exclusively on official candidate pages limits the volume of digital activity we can sample. A larger overall sample-size of district-specific digital traces would have allowed us to capture district effects.

While it may not be feasible to collect sufficient digital traces for a specific district, it might be feasible to group districts into very homogeneous categories – lets say very non-competitive rural Republican districts in the U.S. Assuming access to commercial voter registration files for the entire United States, one could then estimate vote preference equations for these "typical" U.S. congressional districts.

This would entail monitoring political preference expressed on a broader range of "API-friendly" social media (for example, Twitter) and matching the digital traces to a voter registration file so we learn, amongst other things, their congressional district. This strategy will generate limited information for some districts. They could be pooled with "high" information districts which are most 'similar' (in terms of demographic, geographic, political and cultural attributes) to the district we are interested in. We can consider these voters exchangeable conditional on some district-similarity parameter, allowing us to obtain exponentially more information at the district level.[26] These improvements alone, subject to resources and the availability of social-media data, would significantly reduce attenuation bias.

By improving the quality of the digital trace sample we can reduce absolute error and the attenuation bias of our forecasts. There are four considerations here. One aim is simply increasing the size of the digital trace sample – including more individuals in the sample. Executing this in a cost-effective, and ideally automated, fashion is challenging.

A second factor that significantly contributes to the quality of the digital trace sample is complete and accurate information on the individuals in the sample. We were fortunate to have the complete voter registration data base for Texas and hence only

[26] A similar strategy was used by `FiveThirtyEight` to allow for information obtained via traditional opinion polling to flow from frequently polled districts to those less frequently polled.

needed basic information from Facebook accounts in order to match digital traces to the population frame along with their complete demographic profiles.[27] In most cases (certainly outside of the U.S.), researchers will not have this information. A challenge then will be obtaining a sufficiently complete demographic profile of the digital traces in order to match them to the population frame and estimate a reasonably well-specified prediction model.

A third consideration is the precision by which digital traces are matched to demographic profiles (commercial voter registrations, for example) and the population frame. The effective sample size, and its quality, increases with more precise matches. The matching algorithm contributes enormously here but also access to diverse population files can increase the matching of digital traces to demographic profiles. [28] The adoption of a two-stage matching procedure, one to 'auxiliary files' to clean the digital trace as much as possible, and finally to the L2 files, should be explored in further work.

Finally, as was alluded to earlier, we can improve the quality of the digital trace measure itself by measuring an individual's partisan digital trace on multiple social media platforms. This would require observing the digital trace left by the same individual on different social media such as Facebook, Twitter, or WhatsApp but also multiple partisan digital traces by the same individual on a particular social medium over time. An important challenge here would be matching individuals to their various publicly available social media accounts. But as our analyses demonstrate this could be useful because the volume of an individual's digital traces seems to correlate with strength of partisanship.[29]

Conclusion. We propose a novel MRP estimation strategy that combines samples of these digital traces with a population frame that has extensive individual-level socio-economic data in order to generate area forecasts of the outcome variable of interest.

[27] As Pew (Igielnik et al. 2018) points out though there is considerable variance in the completeness of these commercial files.

[28] Just to illustrate, in the Texas case we were unable to match a large number of digital traces to the L2 data file. We might have been able to match some of these to the Texas Appraisal District Files, assuming these had information which was consistent with L2, but slightly different so that it would have been closer to the Facebook data. For example, Appraisal District files might have had a slightly different permutation of name spelling and address specificity. With the additional demographic information from the Appraisal District Files it would ultimately become easier to match to the L2 file and therefore increase the overall digital trace sample size.

[29] Although this should not detract from the important advantage of the Facebook medium because Facebook has in effect standardized the 'support' procedure – a like or love is a much clearer signal of support than a retweet, or a friendship.

Our implementation assumes we can observe, and sample, individuals signaling their preference by favoring one virtual location over another. The digital trace in our case is visiting a Democrat versus Republican Facebook page during the election campaign. We face a challenge in measuring the socio-demographic characteristics of the individuals leaving these digital traces. We implement one strategy – matching individuals to quite exhaustive voter registration files – but there are a number of alternatives here. We demonstrate that even a relatively small virtual sample can be quite representative of the overall population. Finally, we train a random forest machine to estimate the probability of voting Republican, conditional on individual-level data from the complete voting history and registration data for Texas. Over the course of seven weeks preceding the mid-term elections we generate vote share forecasts for all 36 congressional seat contests (including "potential" votes if either party does not field a candidate) and for the Senate race. The forecasts do not use any survey results as input. The predicted vote pluralities are comparable to other conventional methods. Primarily because of attenuation bias, the absolute errors of our forecasts are greater than those of other methods. By optimising sampling to be theoretically representative at the district level and by incorporating district-specific variables in our modeling, we could significantly reduce attenuation bias and significantly improve forecast precision.

References

Bakshy, E., Messing, S. & Adamic, L. A. (2015), 'Exposure to ideologically diverse news and opinion on Facebook', *Science* **348**(6239), 1130–1132.

Barberá, P. (2015), 'Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data', *Political analysis* **23**(1), 76–91.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. (2015), 'Tweeting from left to right: Is online political communication more than an echo chamber?', *Psychological science* **26**(10), 1531–1542.

Bond, R. & Messing, S. (2015), 'Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook', *American Political Science Review* **109**(1), 62–78.

Burden, B. C. (2000), 'Voter turnout and the national election studies', *Political Analysis* pp. 389–398.

Buttice, M. K. & Highton, B. (2013), 'How does multilevel regression and poststratification perform with conventional national surveys?', *Political analysis* **21**(4).

Denisko, D. & Hoffman, M. M. (2018), 'Classification and interaction in random forests', *Proceedings of the National Academy of Sciences* **115**(8), 1690–1692.

Downes, M. & Carlin, J. B. (2020), 'Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study', *Biometrical Journal* **62**(2), 479–491.

Duch, R. M. & Stevenson, R. T. (2008), *The economic vote: How political and economic institutions condition election results*, Cambridge University Press.

Enamorado, T., Fifield, B. & Imai, K. (2018), 'Using a probabilistic model to assist merging of large-scale administrative records', *Social Science Research Network (SSRN): 3214172* .

Ghitza, Y. & Gelman, A. (2013), 'Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups', *American Journal of Political Science* **57**(3), 762–776.

Graefe, A., Armstrong, J. S., Jones Jr, R. J. & Cuzán, A. G. (2014), 'Combining forecasts: An application to elections', *International Journal of Forecasting* **30**(1), 43–54.

Hanretty, C., Lauderdale, B. E. & Vivyan, N. (2018), 'Comparing strategies for estimating constituency opinion from national survey samples', *Political Science Research and Methods* **6**(3), 571–591.

Igielnik, R., Keeter, S. & Weisel, R. (2018), 'Commercial voter files and the study of u.s. politics', *Pew Research Center* pp. 1–58.

Janitza, S., Celik, E. & Boulesteix, A.-L. (2018), 'A computationally fast variable importance test for random forests for high-dimensional data', *Advances in Data Analysis and Classification* **12**(4), 885–915.

Keeter, S., Igielnik, R. & Weisel, R. (2016), 'Can Likely Voter Modes be Improved? Evidence from the 2014 U.S. House Elections', *Pew Research Center* pp. 1–32.

Keyes, O. (2016), 'A human name parser for R', `https://github.com/Ironholds/humaniformat`.

Kiewiet de Jonge, C. P., Langer, G. & Sinozich, S. (2018), 'Predicting state presidential election results using national tracking polls and multilevel regression with poststratification (MRP)', *Public Opinion Quarterly* **82**(3), 419–446.

Kristensen, J. B., Albrechtsen, T., Dahl-Nielsen, E., Jensen, M., Skovrind, M. & Bornakke, T. (2017), 'Parsimonious data: How a single facebook like predicts voting behavior in multiparty systems', *PloS one* **12**(9), 1–12.

Lauderdale, B. E., Bailey, D., Blumenau, J. & Rivers, D. (2020), 'Model-based preelection polling for national and sub-national outcomes in the US and UK', *International Journal of Forecasting* **36**(2), 399–413.

Lax, J. R. & Phillips, J. H. (2013), How should we estimate sub-national opinion using MRP? Preliminary findings and recommendations, *in* 'annual meeting of the Midwest Political Science Association, Chicago'.

Lewis-Beck, M. S., Norpoth, H., Jacoby, W. & Weisberg, H. F. (2008), *The American voter revisited*, University of Michigan Press.

Lu, B. (2017), Constructing Prediction Intervals for Random Forests, PhD thesis, Ph. D. thesis, Pomona College.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. & Ziegler, A. (2012), 'Probability machines', *Methods of information in medicine* **51**(01), 74–81.

McDonald, M. (2014), 'Voter turnout. united states election project', `http://www.electproject.org/`.

Park, D. K., Gelman, A. & Bafumi, J. (2004), 'Bayesian multilevel estimation with post-stratification: State-level estimates from national polls', *Political Analysis* **12**(4), 375–385.

Presser, S. (1990), 'Can context changes reduce vote over-reporting?', *Public Opinion Quarterly* **54**(4), 586–593.

Silver, N. (2018*a*), 'Forecasting the race for the house', `https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/`. [Online; posted 06-November-2018].

Silver, N. (2018*b*), 'Forecasting the race for the senate', `https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate/`. [Online; posted 06-November-2018].

Stekhoven, D. J. & Bühlmann, P. (2012), '`missForest` — non-parametric missing value imputation for mixed-type data', *Bioinformatics* **28**(1), 112–118.

Team, R. C. et al. (2013), 'R: A language and environment for statistical computing'.

Thompson, S. K. (1987), 'Sample size for estimating multinomial proportions', *The American Statistician* **41**(1), 42–46.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Wright, M. N. & Ziegler, A. (2015), '`ranger`: A fast implementation of random forests for high dimensional data in C++ and R', *arXiv:1508.04409* .

Zhang, X. (2018), 'Social media popularity and election results: A study of the 2016 taiwanese general election', *PloS one* **13**(11), e0208190.

Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J. & Croft, J. B. (2015), 'Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system', *American journal of epidemiology* **182**(2), 127–137.

Appendix

Sampling

Algorithm A1 A description of the relevant steps in our sampling routine. The page space ρ included 68 partisan pages; $N(M)$ was initially set to 5 for all pages, and later increased to 30 for the subset of pages we believed to be most informative of "swing" voters; relevant collected characteristics included *Digital Partisanship*; *Facebook Name*; *Current City*; *Home Town*; *Gender*. M was defined as any random pick of the $N(M)$ *likes*, *loves* or *explicitly positive comments* on the relevant page, for the given day.

procedure GETUSERINFO
    let ρ a vector of relevant social media pages
    let N be a counting operator
    let $M(\rho_{O_d})$ be the subset of users being collected per page, according to daily order O_d
    let **collect** be a function of the input user, with output z, the set of relevant characteristics

    # FOR EACH WEEK LEFT IN THE CAMPAIGN
    for w in $W, ..., 0$ **do**

        # FOR EACH DAY WITHIN THE WEEK
        for d in $7, ..., 1$ **do**

            # SAMPLE AT RANDOM A COLLECTION ORDER
            O_d = **sample**(**from** = $1, ..., N(\rho)$, **size** = $N(\rho)$, **replacement** = `FALSE`)

            # FOR EACH PAGE IN THE PAGE SPACE
            for O_d in $1, ..., N(\rho)$ **do**

                # VISIT THE LATEST DAILY POST ON THE PAGE
                goto ρ_{O_d}

                # FOR EACH EXPLICITLY PARTISAN USER IN $M(\rho_{O_d})$
                for i in $1, ..., N(M(\rho_{O_d}))$ **do**

                    # COLLECT INFORMATION FROM THEIR PUBLIC ABOUT PAGE
                    z_i = **collect** $M_i(\rho_{O_d})$

                end for
            end for
        end for

    end for

end procedure

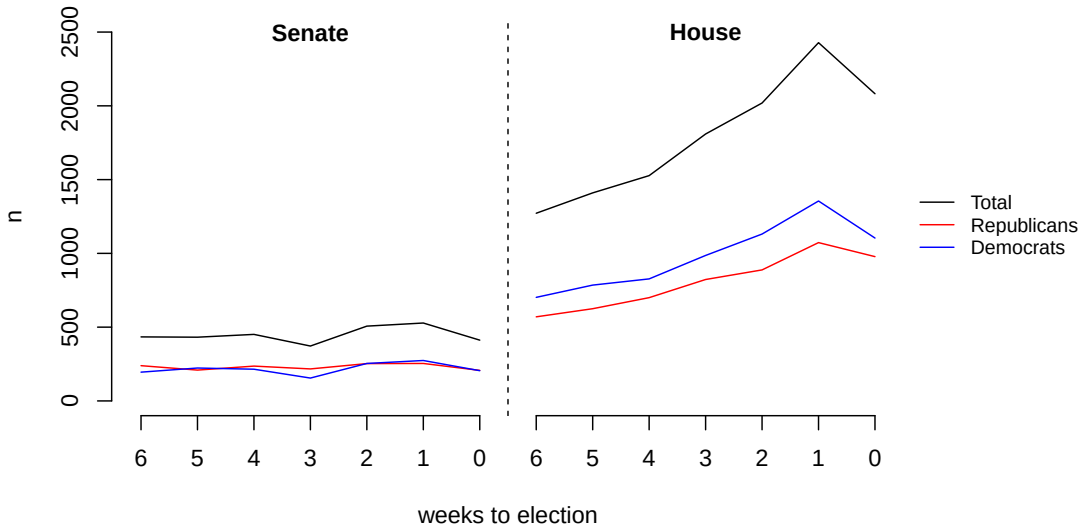


Figure A1: Sample size of digital traces collected by party and election-type, over the monitoring period. The increasing trend in collection for the house is due to increasing collection for "swing" districts, in the hope of better filling competitive cells. Weaker numbers for "election week" are due to lower number of days available for collection

Table A1: Sample summary by page.

Page Address	Party	Election Type	District ID	n	% Total
https://www.facebook.com/betoorourke/	Dem	US Senate	0	1521	9.70
https://www.facebook.com/tedcruzpage/	Rep	US Senate	0	1615	10.30
https://www.facebook.com/vote4McKellar	Dem	US House	1	19	0.12
https://www.facebook.com/RepLouieGohmert/	Rep	US House	1	125	0.80
https://www.facebook.com/toddlittonforcongress/	Dem	US House	2	230	1.47
https://www.facebook.com/CrenshawforCongress/	Rep	US House	2	192	1.22
https://www.facebook.com/Lorieburchforcongress/	Dem	US House	3	98	0.62
https://www.facebook.com/VanForTexas/	Rep	US House	3	44	0.28
https://www.facebook.com/KrantzforCongress/	Dem	US House	4	224	1.43
https://www.facebook.com/RepRatcliffe/	Rep	US House	4	168	1.07
https://www.facebook.com/lancegoodenfortexas/	Rep	US House	5	114	0.73
https://www.facebook.com/JanaLynneSanchezforUSCongress/	Dem	US House	6	196	1.25
https://www.facebook.com/wright4congress/	Rep	US House	6	141	0.90
https://www.facebook.com/LizzieForCongress/	Dem	US House	7	786	5.01
https://www.facebook.com/johnculberson/	Rep	US House	7	598	3.81
https://www.facebook.com/stevenforcongress/	Dem	US House	8	26	0.17
https://www.facebook.com/BradyforTexas/	Rep	US House	8	148	0.94
https://www.facebook.com/repalgreen/	Dem	US House	9	105	0.67
https://www.facebook.com/siegelfortexas/	Dem	US House	10	205	1.31
https://www.facebook.com/MichaelMcCaulTX/	Rep	US House	10	106	0.68
https://www.facebook.com/mike.conaway/	Rep	US House	11	136	0.87
https://www.facebook.com/VanessaAdiaTX12/	Dem	US House	12	198	1.26
https://www.facebook.com/RepKayGranger/	Rep	US House	12	111	0.71
https://www.facebook.com/gregsagan2018/	Dem	US House	13	140	0.89
https://www.facebook.com/ThornberryForCongress/	Rep	US House	13	72	0.46
https://www.facebook.com/adrbell/	Dem	US House	14	136	0.87
https://www.facebook.com/WeberForTexas/	Rep	US House	14	41	0.26
https://www.facebook.com/votevicente/	Dem	US House	15	124	0.79
https://www.facebook.com/westley4Congress/	Rep	US House	15	167	1.06
https://www.facebook.com/voteforveronica/	Dem	US House	16	140	0.89
https://www.facebook.com/Seeberger1ForCongress/	Rep	US House	16	204	1.30
https://www.facebook.com/RickKennedyforCongress/	Dem	US House	17	153	0.98
https://www.facebook.com/BillFloresForCongress/	Rep	US House	17	161	1.03
https://www.facebook.com/CongresswomanSheilaJacksonLee/	Dem	US House	18	105	0.67
https://www.facebook.com/Ava-for-Congress-526965147465733/	Rep	US House	18	122	0.78
https://www.facebook.com/miguellevario19/	Dem	US House	19	195	1.24
https://www.facebook.com/JodeyArrington/	Rep	US House	19	142	0.91
https://www.facebook.com/JoaquinCastroTX/	Dem	US House	20	143	0.91
https://www.facebook.com/ChipRoyforCongress/	Dem	US House	21	230	1.47
https://www.facebook.com/KopserforCongress/	Rep	US House	21	235	1.50
https://www.facebook.com/KulkarniforCongress/	Dem	US House	22	225	1.43
https://www.facebook.com/PeteOlsonTX/	Rep	US House	22	126	0.80
https://www.facebook.com/GinaOrtizJones/	Dem	US House	23	755	4.81
https://www.facebook.com/HurdForCongress/	Rep	US House	23	747	4.76
https://www.facebook.com/JanMcDowellDemocrat/	Dem	US House	24	179	1.14
https://www.facebook.com/RepKennyMarchant/	Rep	US House	24	150	0.96
https://www.facebook.com/JulieForTexas/	Dem	US House	25	192	1.22
https://www.facebook.com/RepRogerWilliams/	Rep	US House	25	186	1.19
https://www.facebook.com/LinseyFaganTx/	Dem	US House	26	218	1.39
https://www.facebook.com/michaelcburgess/	Rep	US House	26	175	1.12
https://www.facebook.com/EricHolguin/	Dem	US House	27	222	1.42
https://www.facebook.com/CloudforCongress/	Rep	US House	27	161	1.03
https://www.facebook.com/repcuellar/	Dem	US House	28	53	0.34
https://www.facebook.com/SylviaRGarcia/	Dem	US House	29	147	0.94
https://www.facebook.com/aronoffforcongress/	Rep	US House	29	112	0.71
https://www.facebook.com/CongresswomanEBJtx30/	Dem	US House	30	89	0.57
https://www.facebook.com/MJforTexas/	Dem	US House	31	72	0.46
https://www.facebook.com/judgecarter/	Rep	US House	31	103	0.66
https://www.facebook.com/ColinAllredTX/	Dem	US House	32	819	5.22
https://www.facebook.com/petesessions/	Rep	US House	32	429	2.74
https://www.facebook.com/MarcVeasey	Dem	US House	33	52	0.33
https://www.facebook.com/billups4congress/	Rep	US House	33	227	1.45
https://www.facebook.com/UsCongressmanFilemonVela/	Dem	US House	34	31	0.20
https://www.facebook.com/profile.php?id=100011088094658&fref=mentions	Rep	US House	34	65	0.41
https://www.facebook.com/LloydDoggettTX/	Dem	US House	35	156	0.99
https://www.facebook.com/Davidsmallingforcongress/	Rep	US House	35	12	0.08
https://www.facebook.com/daynasteele36/	Dem	US House	36	227	1.45
https://www.facebook.com/RepBrianBabin/	Rep	US House	36	137	0.87

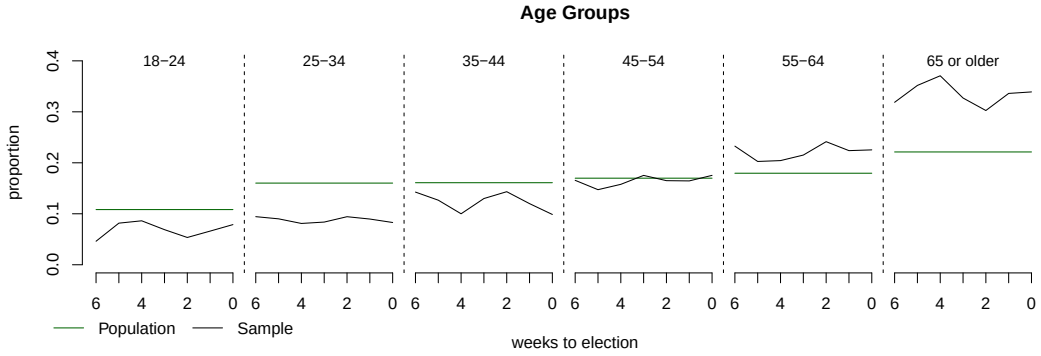


Figure A2: Population v. Sample comparison: age.

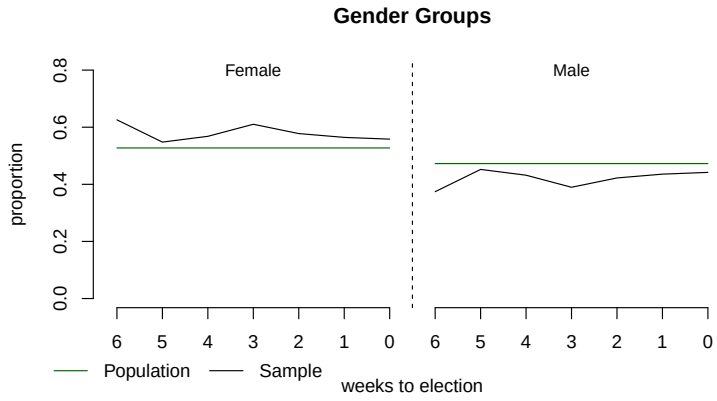


Figure A3: Population v. Sample comparison: gender.

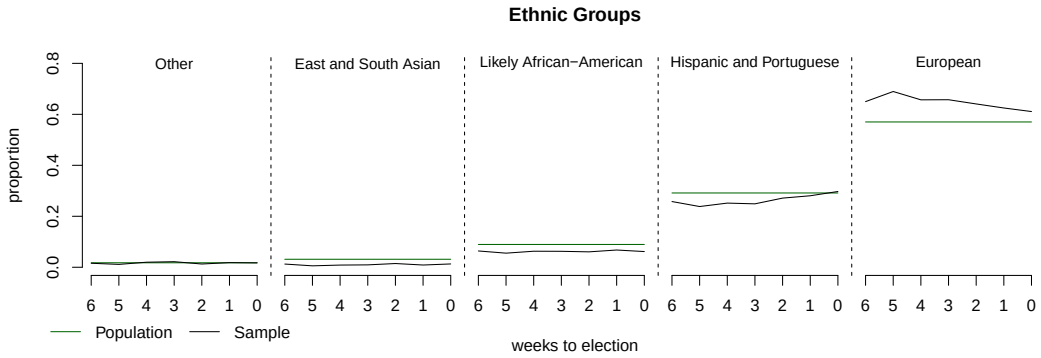


Figure A4: Population v. Sample comparison: ethnicity.

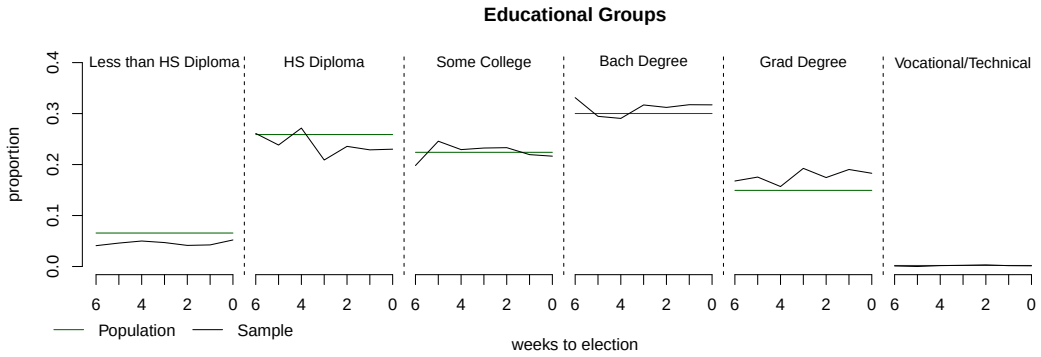


Figure A5: Population v. Sample comparison: highest level of education.

Variable Selection

Table A2: Summary table of the reduced covariate-space $\boldsymbol{X}$.

Variable_Names	Class	Vote Choice Cell	P($\boldsymbol{x}$ = NA)	Median/Mode
US_Congressional_District	factor		0.00	21
Voters_Gender	factor	✓	0.00	F
age_cat	factor	✓	0.00	65_or_older
EthnicGroups_EthnicGroup1Desc	factor	✓	0.06	European
CommercialData_Education	factor	✓	0.38	Bach Degree - Extremely Likely
income_cat_000	factor		0.03	(80,100]
Religions_Description	factor		0.51	Protestant
ElectionReturns_G16_Cnty_Margin_Trump_R	numeric		0.00	9
ElectionReturns_G12_Cnty_Margin_Obama_D	numeric		0.00	-11
ElectionReturns_G12PrecinctTurnoutAllRegisteredVoters	numeric		0.00	53
ElectionReturns_G14PrecinctTurnoutAllRegisteredVoters	numeric		0.00	32
ElectionReturns_G16PrecinctTurnoutAllRegisteredVoters	numeric		0.00	62
ElectionReturns_G12PrecinctTurnoutDemocrats	numeric		0.00	49
ElectionReturns_G14PrecinctTurnoutDemocrats	numeric		0.00	27
ElectionReturns_G16PrecinctTurnoutDemocrats	numeric		0.00	58
ElectionReturns_G12PrecinctTurnoutIndependentsAllOthers	numeric		0.00	24
ElectionReturns_G14PrecinctTurnoutIndependentsAllOthers	numeric		0.00	7
ElectionReturns_G16PrecinctTurnoutIndependentsAllOthers	numeric		0.00	34
ElectionReturns_G12PrecinctTurnoutRepublicans	numeric		0.00	69
ElectionReturns_G14PrecinctTurnoutRepublicans	numeric		0.00	48
ElectionReturns_G16PrecinctTurnoutRepublicans	numeric		0.00	76
General_2016_11_08_reg	factor		0.24	Y
age_cat_2016_11_08	factor		0.00	65_or_older
General_2014_11_04_reg	factor		0.22	N
age_cat_2014_11_04	factor		0.00	45-54
General_2012_11_06_reg	factor		0.39	Y
age_cat_2012_11_06	factor		0.00	45-54
CommercialData_BookBuyerInHome	numeric		0.78	2
CommercialData_AreaMedianEducationYears	numeric		0.07	12
CommercialData_EstHomeValue	numeric		0.04	175308
CommercialData_EstimatedAreaMedianHHIncome	numeric		0.07	71854
CommercialData_HHComposition	factor		0.76	1 adult Male & 1 adult Female
CommercialData_OccupationGroup	factor		0.46	Retired
Parties_Description	factor	✓	0.00	Democratic
VotingPerformanceEvenYearGeneral	numeric		0.10	60
VotingPerformanceEvenYearPrimary	numeric		0.16	0
VotingPerformanceMinorElection	numeric		0.08	0
Voted_in_Primary_2018_03_06	numeric		0.41	0
Voted_in_Primary_2016_03_01	numeric		0.42	0
Voted_in_Primary_2014_03_04	numeric		0.53	0
Voted_in_Primary_2012_05_29	numeric		0.57	0
Voted_in_Primary_2018_03_06_party	factor		0.80	R
Voted_in_Primary_2016_03_01_party	factor		0.71	R
Voted_in_Primary_2014_03_04_party	factor		0.87	R
Voted_in_Primary_2012_05_29_party	factor		0.87	R

Results

District	Digital Vote Error	Digital Vote n	FiveThirtyEight Error	FiveThirtyEight n
Senate	4.06	575	1.42	4,782
CD-1	9.92	91	4.19	382
CD-2	2.32	1,752	0.15	9,555
CD-3	6.03	251	3.65	698
CD-4	11.18	69	3.16	618
CD-5	2.65	1,312	1.82	2,665
CD-6	2.77	1,230	0.57	9,527
CD-7	7.04	192	3.53	769
CD-8	8.36	123	4.99	250
CD-9				
CD-10	3.4	819	2.96	1,084
CD-11	14.49	41	2.27	1,028
CD-12	3.71	658	0.19	8,707
CD-13	14.55	39	2.2	1,012
CD-14	6.29	242	4.87	368
CD-15	1.12	7,441	5.42	293
CD-16	6.22	224	1.34	4,455
CD-17	0.88	9,415	1	9,300
CD-18	8.86	104	6.24	129
CD-19	11.32	69	3.3	594
CD-20				
CD-21	6.75	205	1.8	2,945
CD-22	3.5	773	8.97	113
CD-23	3.93	618	7.9	150
CD-24	6.94	194	8.17	139
CD-25	3.55	744	7.53	160
CD-26	2.82	1,126	8.93	103
CD-27	5.61	300	4.93	349
CD-28				
CD-29	10.32	81	3.32	696
CD-30				
CD-31	7.45	168	5.79	280
CD-32	7.65	163	2.83	1,198
CD-33	12.57	55	3.61	450
CD-34	2.87	1,088	8.35	119
CD-35	13.07	54	0.76	7,393
CD-36	11.14	73	4.63	316
Total		30,289		70,627

Table A3: sample size needed in independent probability surveys to obtain the levels of accuracy observed. Errors are reported in percentage points. The table is empty where the election was not contested. A maximum effective sample-size bound is set to $n = 9600$. Note: decimal changes in area estimates due to simulation uncertainty can lead to increases or decreases of thousands of effective units, hence caution in interpreting these numbers is needed. Some degree of simulation uncertainty persists, as the number of simulations that we used to characterize uncertainty is relatively small - only 500; this was capped because we calculated turnout at the individual level, and it was exceedingly computationally expensive to produce simulations for each of the 13 million Registered voters. Note that inference over average error, average bias at the category level, and success in calling area-level results is robust to this uncertainty.

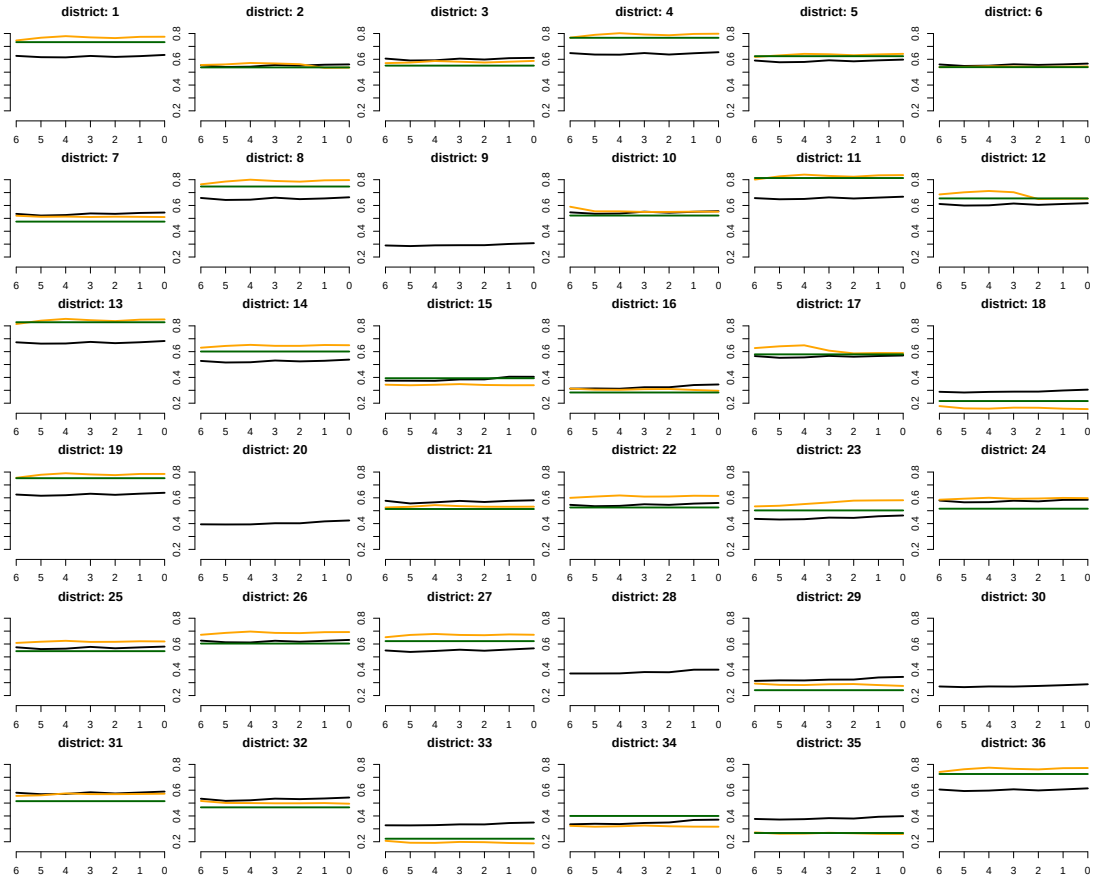


Figure A6: Congressional District Forecasts: FiveThirtyEight (orange lines) and Digital Vote (black lines) against election results (in green) over the campaign.

Information Test

	Dependent variable:		
	2018 Election Result		Change from Trump 2016
	(1)	(2)	(3)
538 Predictions	0.868***		
	(0.097)		
Digital Vote	0.002	−0.333***	
	(0.172)	(0.075)	
2016 Trump Vote		1.223***	
		(0.049)	
538 Pred. Change from Trump 2016			0.045
			(0.074)
Digital Vote Pred. Change from Trump 2016			−0.095**
			(0.044)
Constant	0.049	0.042**	−0.017***
	(0.046)	(0.017)	(0.003)
Observations	33	33	33
R^2	0.962	0.994	0.188
Adjusted R^2	0.959	0.993	0.134
Residual Std. Error (df = 30)	0.034	0.014	0.016
F Statistic (df = 2; 30)	376.969***	2,317.410***	3.480**
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A4: Regression table confronting the question: does the Digital Vote add any novel insights to polls based forecasts? The insignificant coefficient in model (1) suggests no new information is present. No new information is added relative to the 2016 Trump Vote (2). No new information on the change since 2016 is present at the district level (3).

Chapter III:

Polling India via Regression and Post-Stratification of Non-Probability Samples

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Abstract

We present a fully integrated methodology to poll the Indian electorate, and show its power by deploying it in the run-up to the 2019 Lok Sabha election. Our findings show our methodology is promising, performing significantly better than all established polling outlets in the country. This paper contributes to the literature on opinion polling in challenging settings - of which India is an important example - by demonstrating that reasonably powerful samples can be obtained cheaply using novel technologies, namely social media and crowd-sourcing platforms. We further show how these samples are biased, and describe source-specific selection. Our contributions extend to the literature on stratification frames as we propose a novel methodology to obtain deep and reliable estimates of the joint-distribution for the polity of interest, in setting where: a) random micro-data samples of the census are not available; b) public census tables are shallow and outdated; c) political variables are not included in the available joint distribution; d) nationally representative surveys are available, but they are heavily biased at the area level. Our methodology provides robust weights for deep cells, by integrating multiple sources of population data. Finally, to the literature on probability machines, we contribute by proposing a beta-binomial model which enables accounting of prediction and sampling uncertainty via Monte Carlo simulations, leveraging assumptions about the predicted probabilities and their variance which are only mildly approximate. We further build on literature about the hyper-parameters of Random Forests as probability machines, and show that large values of `mtry` produce high-quality predictions. We hope our findings will be picked up by the Indian polling industry, as appetite for more regular and more reliable polling grows in the fast-growing democracy.

1 Introduction

Voting behaviour in India has long been the subject of academic scrutiny: Eldersveld and Ahmed (Eldersveld & Ahmed 1978) first exported the 'National Election Study' model, pioneered in the US (Heath et al. 2005), to the world's largest democracy; Butler et al. (Butler et al. 1995) analyzed the emergent patterns of swing over the first 40 years of democratic rule, introducing a framework of analysis which persists in current scholarship (Heath 2015). More recently, the advent of the internet and advanced technological tools has enabled the spread of large-scale, historical administrative data, which has opened the door to more fine-grained analyses of the Indian vote (Jensenius & Verniers 2017). Despite this substantial tradition in the study of electoral behaviour, Indian pre-election opinion polls have notoriously performed poorly as a tool for prediction of electoral outcomes (Rai 2014, Solomon 2016). Typically listed as reasons for these failures are the sheer size and heterogeneity of the Indian electorate, which can badly reduce statistical power at the area-level; lack of transparency in the polling industry with respect to survey design and modeling of seat projections; differential access to technology - translating to differential probabilities of inclusion in surveys - across socio-demographic groups.

This paper seeks to deliver a theoretically sound, practically expedient and easily attainable framework for opinion researchers to analyze pre-election attitudes in India, under non-representative samples of voters. We apply our methodology to convenience samples gathered in the run-up to the 2019 Lok Sabha election. Multilevel Regression and Post-stratification (MrP) (Park et al. 2004) is a popular technique for dealing with non-representative samples which has gained traction in recent years. Most interestingly for our application, the Wang et al. (Wang et al. 2015) Xbox study shows that extremely unrepresentative, but plentiful, surveys of online gamers can be smoothed and re-weighted via MrP to accurately predict election outcomes in the US. In this paper, we will show that smaller and further less representative samples can be used to obtain estimates of the national mood which fall within acceptable margins from the truth.

There are three key elements to the execution of an MrP model: a) a *stratification frame*: a vector of counts for a set of *cells*, or mutually-exclusive and exhaustive voter

categories, in the population of interest; b) *training data*: an individual-level dataset which enables the cross-referencing of demographic profiles with outcome-behaviour of interest - in our case, voting behaviour; b) a *learner* - an algorithm which can best summarize the relationships implied by the individual-level dataset and coherently project these onto voter-categories we have not yet seen in the sample - also known as 'out-of-sample' predictions. This paper contributes to the literature on each of these key elements, by innovating the frames, training-data and learners needed to generate accurate readings of public opinion.

We propose a novel treatment of stratification frames, motivated by the challenges faced in the Indian case-study. Whilst in traditional settings such as the UK or the US we would rely on census micro-data - random 1% samples of the decennial census - to generate *deep* and *precise* frames, we cannot do so in India, as the micro-data is not readily available. To optimize the depth - precision trade-off in our Indian stratification frame, we propose a novel Data Integration approach (Carrig et al. 2015), which allows the simultaneous estimation of counts from multiple complementary sources, such as census tables and auxiliary surveys. This innovation is similar in implementation, though motivated by substantially different reasoning, to Cerina et al (Cerina et al. 2020), and contributes to the literature dealing with the augmentation of stratification frames (Leemann & Wasserfallen 2017).

We further innovate the training-data economy for Indian behavioural data by collecting and assembling panels from two unorthodox online populations for which recruitment is cheap and easy, as is largely streamlined by third-party platforms: Amazon Mechanical Turks workers and Facebook users. We contribute to the literature on online non-probability sampling in India, by showing persistent bias in geographic distribution, gender, education and income, as well as evidence for differential online selection effects based on sampling medium.

Finally, we replace the Bayesian multilevel regression learner with Random Forests (Breiman 2001*a*) (RFs). This is motivated by three considerations: i) we sought to produce real-time estimates of the vote, hence engaged in real-time updating of our forecast; this required daily fitting and de-bugging of our predictions, a feat which proved arduous for high-dimensional parametric Bayesian models, which can take dozens of hours to run

for moderate sample sizes and a medium-sized parameter space. Random forests can grow trees in parallel, and deploy memory-optimized tree-growing algorithms (Wright & Ziegler 2015), making them vastly superior on this metric; ii) we sought to remove modeler-bias variable selection, which tends to be a feature of MrP models, especially when it comes to choosing area-predictors (Leemann et al. 2020) and interacting area-predictors with individual-level variables (Lauderdale & Blumenau 2019); iii) we sought to deploy an estimator that would apply regularization to our predictions; following Hastie et al. (Hastie et al. 2009) we sought to implement a bagging estimator, which would pay a price in terms of bias in favour of shrinking prediction variance.

The rest of the paper is outlined as follows: Section 2 sketches a picture of the 2019 election, outlining some of the general challenges which make this an important, albeit difficult, exercise in psephology; Section 3 presents the various data sources we use at each step of our estimation procedure, exploring the characteristics of our convenience samples; in the same section, we further explore the nature of our custom-built stratification frame, outlining a systematic procedure to obtain reliable cell-level counts. Section 4 outlines models to estimate cell-level turnout and voting behaviour; the results of our predictions are reported in Section 5. Finally, we provide a discussion of limitations and outline areas of future work in Section 6.

2 Background

The 2019 election in India sees two major alliances facing off: the National Democratic Alliance (NDA) led by incumbent Prime Minister Narendra Modi, and the United Progressive Alliance (UPA) led by opposition leader Rahul Ghandi. The country elects members of parliament via a first-past-the-post system, and the coalition which has a majority of the 543 contested seats in the Lok Sabha[30] usually nominates a prime minister. In 2014 the NDA captured a large and unforeseen victory, winning 336 seats, whilst the UPA heavily under-performed, only winning 59 seats. Polling error on the NDA result was enormous, averaging at an under-estimation of around 88 seats for the last three published polls, which happened also to be the most accurate (in relative

[30] A trivial number seats are assigned by the president.

terms). We should note here the oddity of Indian pre-election polling, in that vote shares are hardly ever published, hence making it impossible to disentangle polling-error from seat-projection error. Nevertheless, is is reasonable to assume pollsters are using some variation around the uniform-swing model (Solomon 2016); in this paper we show that this simple model has plenty of predictive power in the Indian context, as was noted by Butler et al. (Butler et al. 1995), and we can hence deduce that polling error must be a sizeable contributor to the error of published seat projections.

All things considered, it is perhaps not surprising that monitoring public opinion in India is a difficult task. Beyond the typical challenges we face in countries with richer traditions in public opinion research, including herding, selection effects, bad turnout models etc., it is worth citing a few figures to highlight the complexity of the Indian problem. On the eve of the 2019 election, around $2,300$ parties were registered to compete[31]. Over 900 million individuals were eligible to vote, and around 600 million were expected to take part. These voters are spread over a surface area around 13 times larger than the UK, within the confines of 36 states and territories, many of which have population sizes larger than the largest countries in the world; famously, if the state of Uttar Pradesh were a country, it would be the 5^{th} most populous country in the world. Within each state, voters speak some subset of the 23 officially recognised languages, and uncountable number of unofficial dialects. According to the World Bank, around 65% of the population lives in rural areas[32]. To cope with the logistics associated with ensuring each member of this large and diverse electorate has the opportunity to cast their ballot, the election is usually ran over multiple weeks, each constituency assigned to a *phase*, meaning different parts of the countries vote at different times, and campaign effects can sway public opinion during the voting period. The 2019 election was scheduled to be completed in 7 phases, spanning over 5 weeks.

In this context, it is logistically and statistically difficult to obtain representative samples of voters: surveys have to be designed in multiple languages; rural voters may only be reached by face-to-face interviews due to lack of communications technology; alliances involve local parties and hence in-depth knowledge of local politics is necessary.

[31] `https://economictimes.indiatimes.com/news/elections/lok-sabha/india/india-now-has-2293-political-parties-149-registered-between-february-march/articleshow/68451605.cms?from=mdr`

[32] `https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=IN`

We suspect past polling failures can be in part attributed to selected samples and failures to adjust these selection effects. This paper makes the point, which has been made by Salganik (Salganik 2019) and others, that the issue at hand is not whether, in an ideal world, we would prefer non-representative samples over representative ones; rather, the issue is that, given the context, presuming we can obtain true probability samples may be too strong an assumption. Hence we need to develop a methodology to ensure the measures we obtain from these convenience samples are accurate enough for us to confidently make inference. This paper contributes in developing this methodology.

2.1 Expectations of Regression and Post-Stratification in India

To the best of our knowledge, this is the first time the regression and post-stratfication framework is deployed in such limiting conditions. It is therefore important to outline a-priori reasons for why we think this methodology can work, to justify our effort. Firstly we must set out evaluation metrics: typical MrP focuses on small-area estimation, and especially politically relevant areas such as parliamentary districts, though we know a-priori this will be prohibitive to do in India, due to expected selection effects into online media which will preclude access to large swaths of important local demographics. This is a common problem in the literature on polling in India, which is typically resolved by producing estimates at the national level (Solomon 2016), as it is easier correct for non-probability samples at the aggregate level, and then applying a Uniform Swing projection (Butler et al. 1995) to predict headline seat numbers and 'call' the election.

In the context of small-area estimation it is known that MrP will work provided we have a number of 'scope conditions': i) strong area-level covariates are included in the regression model (Buttice & Highton 2013, Hanretty et al. 2018); ii) the proportion of the total variance in the outcome variable which is driven by between-state heterogeneity is high (Buttice & Highton 2013); iii) the area-level sub-samples are 'large enough' (typically 100 units per area, but dependent on model complexity and sub-sample heterogeneity) (Lauderdale & Blumenau 2019) to capture intricate area-level signals. More generally, as it pertains to all regression and post-stratification applications, and not just to small-area estimation, the most important assumption which must be satisfied here is *ignorability* (Jackman et al. 2019): conditional on the set of covariates in a prediction model, the residual variance in the outcome must be uncorrelated with response

or non-response to our survey. A-priori, it would seem we would surely fail this test: collecting samples of the Indian population online will carry unobservable selection effects in a country with highly heterogeneous exposure to technology. Nevertheless, we have examples in the literature of highly selected samples yielding reasonable estimates after extensive modeling: the Xbox study (Wang et al. 2015) yielded national-level results which were about 0.5% off the national polling average for the 2012 US election, from a population of 'gamers', and in particular gamers who selected into a particular platform - the Xbox. Similar results have been obtained on samples from Mechanical Turks (Goel et al. 2015) and Facebook users (Cerina & Duch 2020). Hence, based on previous work, whilst acknowledging that ignorability will fail, we have reason to believe the introduced bias will be moderate, and by using a robust modeling framework which encourages regularization, we can obtain reasonable national estimates of the vote. Finally, it is worth noting that the key recent drivers of the Indian vote are reasonably well known (Heath 2015): we expect that controlling for 'Caste', 'Income', 'Past Vote' and a variety of regional factors correlated with rurality and access to technology will take us reasonably close to being able to 'ignore' the systematic bias in residuals.

3 Data

In this section we present the data used in this regression and post-stratification excercise. There are two data-objects of primary interest: a stratification frame and a convenience micro-data sample. Unlike many other applications in the regression and post-stratification literature (see Gelman and Little (Gelman & Little 1997), Lax and Phillips (Lax & Phillips 2009) Warshaw and Rodden (Warshaw & Rodden 2012), Buttice and Highton (Buttice & Highton 2013), to cite a few), we cannot rely on off-the-shelf stratification frames derived from the census, for reasons to be outlined below. As such, we have to deploy a strategy to piece together a new and exhaustive frame from scratch. With respect to the convenience sample, we rely on multiple modes of collection, namely searching for subjects amongst Facebook users and Amazon Mechanical Turks. There is evidence that relying on multiple loci for data collection can lead to more robust estimates of effects of interest (Duch et al. 2020), due to the differential selection effects canceling out over loci; furthermore, we rely on the well established

literature in aggregating election forecasts (Graefe et al. 2014, 2015), and expect that averaging loci-specific selection effects will improve the mean-squared predictive error of our predictions. A full list of datasets used in this regression and post-stratification pipeline is available in Table 1.

	IHDS	Census	INES	Lok Dhaba	Facebook Users	Mechanical Turks
Dates	Nov. 2011 - Oct. 2012	Apr. 2010 - Mar. 2011	Apr. - May 2014	Historic	Feb. - Apr. 2019	Feb. - Apr. 2019
Sampling	Probability	Probability	Non-Probability	Electorate	Non-Probability	Non-Probability
Depth	Micro	Micro[33]	Micro	Macro	Micro	Micro
N	135,984	3,800,213	22,295	Electorate	1,633	5,807
Demographic	✓	✓	✓		✓	✓
Political			✓	✓	✓	✓
Strat. Frame	✓	✓	✓	✓		
T. Model			✓	✓		
V. Model				✓	✓	✓

Table 1: A summary table of data sources (columns) by some meta-characteristics (rows), including: dates of collection; representative quality; depth; sample size; whether they include demographic or political variables; whether they are leveraged to build the stratification frame and/or the voting behaviour models. T. and V. in the leftmost column stand for Turnout and Vote-Choice respectively. Note that the Lok Dhaba is a repository of Lok Sabha results over time.

3.1 Constructing a Stratification Frame

A stratification frame is a vector of counts, where each element represents a *cell*. Cells are defined as mutually exclusive sets of individual characteristics which uniquely identify a *type* of interest; examples of individual characteristics include gender, age, religious affiliation, caste status, educational attainment, household income etc. Hence an example of the information contained in a stratification frame would be that there are 1,209 Hindu females from a backwards caste of age between 35 and 45, with no formal education, living in a household which makes less than Rs 60,000 per year.

We define the 'depth' of a cell by the number of variables which define it - so for instance a cell representing the number of individuals who are characterized by the variable-set 'Gender: Male & Caste: OBC' is *shallower* than one defined as 'Gender: Male & Caste: OBC & State: Tamil Nadu & Education: College Degree & Age: 25 to 34'. We note here that some of the variables of interest in this study - of which 'Caste' is the ultimate example - could have been defined to obtain deeper cells. The

[33]Though the census tables enumerate the entire population of the country at the macro-level, we sample at random from its joint distribution a smaller micro-data sample of size noted in the table, which we use later to integrate the IHDS.

Indian Human Development Survey, which provides the backbone of our stratification frame, and which we used to generate the survey questions for our respondents, provides multiple definitions for this variable[34]. Our definition of 'Caste' is the broad - reflecting 6 main categories: 'Brahmin'; 'Forward / General (except Brahmin)' , 'Other Backward Castes (OBC)', 'Scheduled Castes (SC)', 'Scheduled Tribes (SC)'; 'Other'. This is the shallowest definition provided by the IHDS, which we chose to conform to. We chose this shallower definition as we were a-priori uncertain about the translation from 'Jati' - i.e. the surname or clan-name which typically denotes the caste an individual belongs to, of which there are hundreds of thousands of variants, each corresponding to a slightly different broader cast in each State - and Caste. Jati is what is typically asked in a survey, because an individual may not know his broader caste status, and hence it was left up to us to manually code Jatis to caste; we did so to the best our our ability by consulting government websites and other formal and informal online resources. Aggregating upward makes it more likely that classification errors between the caste categories would cancel out, reducing the risk of systematic bias in classification, and hence improving the odds of meaningfully capturing the effect of caste on the vote. More generally, whenever we encountered multiple definitions for the same variable across multiple sources, we aggregated the deeper source up, in order to recover the same level of analysis. We should note that the importance of the variable 'Caste' in our predictions, as highlighted by the Variable Importance Metric (VIM) in Figure A16 in the Appendix (coded as 'Jati') suggests we accurately detected its essence.

Beyond the depth of any individual variable such as 'Caste', the depth of cells in our stratification frame was chosen according to: a) the effective heterogeneity in behaviour of interest across cells, i.e. does adding a candidate variable increase our ability to distinguish cross-cell behaviour?; b) the statistical power we are endowed with to learn cross-cell differences from our sample of behavioural micro-data, typically a function of the size and diversity of the sample. Stratification frames have to contain counts which enable researchers to obtain unbiased estimates of the true probabilities of inclusion of each cell in the population at large. Two qualities of the stratification frame are of concern: its depth and its precision. A deep frame is one that contains cells constructed

[34] `http://ihds.umd.edu/social-groups`

via interactions of large numbers of variables; a precise frame is one where cell-counts are estimated with low uncertainty.

Highly precise frames can usually be obtained via a country's census. A drawback of this source is that it is usually limited in its depth: the most extensive Indian census cross-tabs[35] provide reasonable depth for geographic distribution, and its interaction with age, gender and educational attainment; other subsets of the above mentioned characteristics are available, but none which we could consider exhaustive. In data-rich environments such as the United States or the United Kingdom, census micro-data, usually around 1% of the anonymized individual-level census responses, is available for researchers to use, enabling the construction of deep and precise frames. In the absence of census micro-data, we can rely on large representative surveys to construct our stratification frame; in India, we can use the second wave of the India Human Development Survey (Desai et al. 2015-07-31) (IHDS) a nationally representative survey of 135,986 voting-age individuals from 42,152 households, conducted between November 2011 and October 2012, as a starting point. There are two drawbacks to the use of this data: first, it does not include any political variables, such as past turnout and vote-choice behaviour, which are known to be highly predictive of current political behaviour in typical applications of model-based pre-election opinion polling (Lauderdale et al. 2020); second, the IHDS is not representative at the area-level. In the following paragraph we highlight the importance of the stratification frame being accurate at the deep-cell level - hence at the area-level - even when the ultimate forecast quantity of interest belongs to the aggregate level.

Though we are not producing a forecast at the area level, rather we are concentrating on national vote-shares and headline seat-numbers, it is still necessary that our frame be accurate at the state-level: a residual effect of state-level heterogeneity is a well known factor in Indian political behaviour (Heath 2015, Chakrabarti & Roy 2007), hence we have a strong preference for including state-level variables in our cell-definition. Though we end-up aggregating over states in our final forecast, we borrow insights from the literature dealing with forecasting of aggregate economic indicators from disaggregated measures (Zellner & Tobias 2000, Marcellino et al. 2003, Hubrich

[35]Indian census data is available at `http://www.censusindia.gov.in`

2005, Castle & Hendry 2010, Hendry & Hubrich 2011, Lütkepohl 2011) to form a prior that an aggregate-level forecast obtained by combining multiple disaggregate forecasts will be at least as accurate as one dealing solely with aggregate-level variables. Hence it becomes important for our stratification frame to be as representative as possible of deep-cell levels, even when we are aggregating over these. We note that after a point, the returns to making our cells deeper become negative - that is if our predictions become too noisy, which can happen as the statistical power to capture marginal effects decreases with cell-size, aggregation will lead to worse predictions than a non-noisy, direct forecast of the aggregate quantity. We deploy bagged regression trees (Hastie et al. 2009) models in Section 4 which can perform automatic variable selection and heavy regularization, hence ensuring prediction noise levels are controlled.

Considering the content of the last two paragraphs, the need for a strategy to a) improve the representative quality of the IHDS at the deep-cell level, and b) augment its covariate-space to include political variables, emerges. To improve area-level cell density, Lauderdale et al. (Lauderdale et al. 2020) propose to use Iterative Proportional Fitting (IPF) (Deming & Stephan 1940), also known as 'raking', to re-weight the data according to known area-level frequencies, whilst preserving the underlying interactive structure implied by the original micro-data. Leeman and Wasserfallen (Leemann & Wasserfallen 2017) propose to augment the covariate-space of the frame by using nationally representative auxiliary surveys containing overlapping covariates with the existing stratification frame, as well as augmentation-variables of interest; they achieve augmentation by re-weighting the auxiliary survey to match the known area-level distribution of the overlapping covariates, and recording the resulting adjusted distribution of the augmentation variables, for each area of interest. Building on these methods, we propose a novel methodology, which leverages data integration via multiple imputation (Carrig et al. 2015).

A final note is needed to guide the researcher in choosing 'how deep' to go when building the stratification frame to post-stratify their predictions. It is our belief that one should go 'as deep as possible'- that is one should include as many variables it is computationally feasible to include in defining a given cell - provided the modeling

framework adopted incorporates model-selection. As we will see in later paragraphs, we adopt a Random Forest algorithm to learn about cell-level propensities to vote. In the dichotomous node-splitting framework implemented within each tree composing the forest, variables will only affect the predictions if and only if it they are meaningful in better classifying the training outcomes. For example, if 'Caste' is the only variable that matters in distinguishing NDA and UPA voters in our training examples, even if our cells are several variables deep, the predictions between a model that only includes the variable 'Caste' in the covariate space, and one that includes 'Caste' along with all the other available variables, will be identical. This happens because variables which do not lead to improvement in performance, defines for example by the Meas Squared Error at each split-point - are not used to make a prediction. So there is no detriment, as far as we can tell, to the accuracy of the resulting point estimates from including large number of covariates in the prediction models - and hence no drawback in stratifying according to the deepest possible frame.

3.1.1 Stacking

We take the IHDS micro-data as the starting point to build our stratification frame. We select a subset of the IHDS such that we can obtain cell-counts for cells defined by {*State - Zone - Gender - Age - Educational Attainment - Family Role - Marital Status - Rurality - Religion - Literacy - Caste - Yearly Household Income*}; note that these cells are deeper than necessary, but we select a surplus of variables to help with multiple-imputation, as we will see in later paragraphs.

An initial concern is that the dataset is not representative at the area-level; this issue presents itself in two ways: i) the marginal distribution of each variable by state is unrepresentative of the true population; ii) the underlying interactive structure across the variables, i.e. the way in which educational attainment is correlated with caste, will be spurious at the state-level. A further concern is that the IHDS does not include political variables such as past vote-choice and turnout behaviour. We seek to solve all three of the issues outlined above by engaging in a large data-integration exercise, merging the IHDS with data from the decennial census; the Indian National Election Study (INES) (Team 2014) and the Lok Dhaba[36] (Jensenius & Verniers 2017) archive

[36] `https://tcpd.ashoka.edu.in/projects/`

of past election results.

First, we focus on updating the underlying interactive structure of the IHDS variables. We want the underlying joint distribution at the state-level to match that of the population. Our population targets for the joint distribution are derived from the deepest decennial census tables we can find, namely tables containing counts for cells defined as {*State - Zone - Gender - Age - Educational Attainment*}; tables containing caste, arguably the most predictive variable for vote-choice a-priori, were not as deep, failing to contain a breakdown by age and gender; moreover, the caste classification used in the census was different from that used in the IHDS, and hence was not immediately reconcileable. We could in principle directly rake the IHDS to these census targets, though a problem emerges - namely some of the voter-categories that exist in the census do not exist in the IHDS. This is due to some of these cells being relatively rare in the population, and hence not featuring in the relatively small random sample collected by the IHDS. To avoid instability in raking weights (Norman 1999), we choose to populate these missing cells using a stacking procedure: first, we extract a 0.5% random-sample of the population from the census tables - that is a micro-data sample of cells of size $3,800,213$, hence converting the census tables from macro- to micro-data, ensuring this new sample is large-enough to include the largest cells from every state; second, we stack this micro-data sample below the IHDS, treating variables which are present in the IHDS but not in the census sample as 'missing'. This leads to a new augmented dataset of size $3,936,197$, with a swath of missing data for records coming from the census; the resulting joint-distribution for the overlapping variables is now dominated by the census, hence parallels the known distribution of the population, as desired. Note that the price for correcting the joint distribution is that we now have a large amount of missing values; we deal with this problem at the end of this Section.

Having corrected the underlying interactive structure to the best of our ability, we now deal with augmenting the covariate space with political variables, and in particular turnout and vote-choice from 2014. The post-election National Election Study (NES) from 2014 is the perfect dataset to do so. It contains cells defined as {*State - Zone - Gender - Age - Religion - Caste - Educational Attainment - Yearly Household Income - 2014 Turnout - 2014 Alliance Vote*}. Once again, we perform the same stacking

procedure, treating non-overlapping variables as entirely missing; we append the NES to the augmented dataset, hence generating a novel object boasting 3958492 records, and once again a significant number of completely missing values. Table A2 in the Appendix shows a subset of the stacked dataset.

3.1.2 Multiple Imputation

We now seek to solve our missing-data problem to generate a complete micro-data sample of the population, from which to extract deep and precise cell-counts. To do so, we rely on the Multiple Imputation via Chained Equations (MICE) (Van Buuren & Oudshoorn 1999) paradigm. Following Carrig et al. (Carrig et al. 2015), we note two pre-requisites for the MICE procedure to produce meaningful imputed values for the stacked object: first, that the values be 'missing at random' (Rubin 2004), a fact which we believe to be plausible for each of the underlying datasets; second, that the three datasets from which it is constructed share a data generating process. We assume this is true a-priori, given that all three datasets are designed to be representative of the true population. Looking deeper at the nature of the three datasets, the NES stands out as being the most problematic for the common DGP assumption: whilst the dataset is largely representative of demographics, it is almost exclusively a sample of likely voters, with 91% of respondents reporting to have cast a ballot in 2014, highlighting important selection-effects. To overcome this problem, we rake the NES data to the known 2014 election results, at the state level, hence reconvering the correct vote-choice and turnout marginals at the state-level; we then proceed to take a bootstrap sample of the dataset, using the standardized raking weights as the sampling probabilities. The resulting dataset can then reasonably be thought of sharing a DGP with the IHDS and the census, though deeper selection issues relating to the NES are to be expected.

MICE can be extraordinarily computationally expensive, as it seeks to model a series of full-conditional distributions over multiple iterations. `missForest` (Stekhoven & Bühlmann 2012) is a MICE-based algorithm designed to deal with high-throughput data, and hence more ameanable to processing our large-dataset. `missForest` uses Random Forests (Breiman 2001*a*) to predict missing values at each MICE step. It is preferable to parametric MICE as it performs automatic variable selection and regular-

ization; can deal with mixed-data types; and has been shown to out-perform traditional imputation models under multiple settings (Starkweather 2014). To further improve the speed of our imputation model, we rely on the `ranger` (Wright & Ziegler 2015) package for the `R` (Team et al. 2013) statistical programming language to train the random forests at each MICE step; the `ranger` algorithm leverages a `C++` backbone to optimize memory-handling and processing of Random Forests, enabling a fast implementation in high-dimensional settings. The `missRanger` (Mayer 2018) package enables to implement `missForest` with the superior `ranger` algorithm. We can control a number of hyper-parameters in `missRanger`; we limit ourselves to increasing the number of trees to $1,000$ (`num.trees = 1000`) , a number we deem to be large-enough to preclude Monte-Carlo effects in the imputation estimates, whilst keeping other relevant parameters as the `ranger` default; for instance, each tree is fit to a complete bootstrap-sample of the original data (`sample.fraction = 1`, `replace = TRUE`); no maximal depth is imposed on each tree (`max.depth = NULL`); and all available (complete and incomplete) covariates are used as split-candidates.

To make the imputation more robust, we further introduce a substantial state-level predictor, including past election results at the state-level from the Lok Dhaba archive, as well as demographic characteristics derived from the census, and a spatial matrix indicating for each state the distance from any other state - in order to account for a degree of local-smoothing (Hanretty et al. 2018). The total number of variables in the area-predictor is 175, of which 35 are distances from other states; 32 are past-vote variables, including things like past vote by alliance at the state and zone level, as well as past turnout, for the last three elections; and the rest are zone- and state- level aggregation of individual-level demographics.

The algorithm takes approximately 12 hours to converge after 6 iterations, on a `r4.4xlarge` AWS EC2 instance[37].

3.1.3 Iterative Proportional Fitting

Having performed multiple imputation, we now have a complete dataset which we can aggregate to obtain cell-level counts. An important problem is that the dataset remains

[37]Details on Instance types are available on the AWS website: `https://aws.amazon.com/ec2/instance-types/`

rather unrepresentative of the population at the state-level, as it suffers from a) the selection issues affecting the distribution of the NES; b) the lack of statistical power in the IHDS and NES, leading to weak state-level imputations of IHDS-specific and NES-specific variables; c) some level of noise as a result of implementing MICE on mostly-missing features. Figures A1 and A2 in the Appendix show the distribution of 2014 voting behaviour at the zone and state levels described by the post-imputation stratification frame. It comes as no surprise that the state-level imputation is poor, though we can gain some confidence by observing acceptable zone-level errors, ranging between 7 and 9 points on average for the two main parties; correlation around Turnout is reasonably high, though clearly the imputed turnout is larger than the true 2014 turnout across the board, revealing perhaps further evidence of NES selection effects and over-reporting. It is worth under-scoring we do not need these imputations to be perfect at the area-level, as we will eventually aggregate over these; nevertheless, we want our frame to be as representative of the true population as possible at the deep-cell level to improve the robustness of our forecast via disaggregation, and the plots cited above betray a significant degree of noise in the frame's deep levels.

To improve the representation of the dataset at the deep-cell level we employ IPF, as suggested by Lauderdale et al. (Lauderdale et al. 2020). The IPF will iteratively re-weight each individual in our large post-imputation micro-frame, until these conform to a target set of distributions. We choose population targets for variables whose precision is highest - namely variables coming directly from the census tables, and known election results. The resulting targets are a) the joint distribution of state, age, gender and education; b) the joint distribution of state and 2014 voting-behaviour. We implement IPF via the `R` package `anesrake` (Pasek & Pasek 2019). We rake areas independently, in order to speed-up convergence and avoid having the algorithm reconcile distributions at different levels; it is therefore possible the algorithm could change the relative sizes of the states by significantly up- or down-weighting any given one. To ensure the states remain consistently sized, we re-scale the weights to match the size of each state implied by the census. Figures A3 and A4 in the Appendix show the resulting match between the observed 2014 voting behaviour and our post-imputation and post-raking frame, at the zone and state-level respectively. Typically, state-level 2014 voting behaviour in the frame is now exactly what we know it should be. The state-level turnout is still

somewhat poor - though vastly ameliorated compared to its pre-raking counterpart; this is largely a result of the limited turnout training data we have, due to the NES being largely a sample of voters. Though we tried to expand the baseline proportion of non-voters in the training data by taking a re-weighted bootstrap sample, we are still left with a large number of state-level non-voters which are not present in the NES, and as a result they are also missing from the post-imputation frame. This limits the re-weighting options of the IPF algorithm, effectively precluding a perfect match. Nevertheless, the match is perfect at the zone-level, and close-to-perfect at the state-level for the major parties.

To conclude the construction of a representative stratification frame for the 2019 Lok Sabha election, we aggregate the weights at the cell-level; this reduces the size of the object to keep in memory by reducing the number of rows. Table 2 displays a subset of the completed frame. We note that the frame is representative of 2011 demographics and 2014 voting behaviour. This may be unsatisfactory, given the fast-evolving demographics of the country, and possibly severely our-dated; we are however constrained by data-availabily. The question of how to update deep and precise stratification frames remains open in the literature, and will the the subject of future work.

Weight	States	Zones	Gender	Age_cat	Religion	Education level	Caste	Vote 14	···
131.92	(01) Jammu & Kashmir 01	(01) North 01	(1) Male 1	(54-64]	(2) Muslim 2	(01) No Formal Edu 01	(3) Other Backward Castes (OBC) 3	UPA	···
175.26	(01) Jammu & Kashmir 01	(01) North 01	(2) Female 2	(44-54]	(2) Muslim 2	(01) No Formal Edu 01	(3) Other Backward Castes (OBC) 3	UPA	···
345.93	(01) Jammu & Kashmir 01	(01) North 01	(1) Male 1	(17-24]	(2) Muslim 2	(03) Middle or Secondary 03	(3) Other Backward Castes (OBC) 3	OTHER	···
982.97	(01) Jammu & Kashmir 01	(01) North 01	(1) Male 1	(24-34]	(2) Muslim 2	(03) Middle or Secondary 03	(3) Other Backward Castes (OBC) 3	NDA	···
127.06	(01) Jammu & Kashmir 01	(01) North 01	(2) Female 2	(24-34]	(2) Muslim 2	(03) Middle or Secondary 03	(3) Other Backward Castes (OBC) 3	UPA	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48.04	(35) Anadman/Nicobar 35	(04) Eastern 04	(2) Female 2	(54-64]	(1) Hindu 1	(03) Middle or Secondary 03	(2) Forward/General (except Brahmin) 2	OTHER	···
161.28	(35) Anadman/Nicobar 35	(04) Eastern 04	(1) Male 1	(34-44]	(1) Hindu 1	(01) No Formal Edu 01	(2) Forward/General (except Brahmin) 2	NDA	···
125.88	(35) Anadman/Nicobar 35	(04) Eastern 04	(1) Male 1	(24-34]	(1) Hindu 1	(05) Some Graduate or Higher 05	(2) Forward/General (except Brahmin) 2	NDA	···
35.73	(35) Anadman/Nicobar 35	(04) Eastern 04	(1) Male 1	(24-34]	(1) Hindu 1	(03) Middle or Secondary 03	(2) Forward/General (except Brahmin) 2	OTHER	···
40.57	(35) Anadman/Nicobar 35	(04) Eastern 04	(1) Male 1	(24-34]	(1) Hindu 1	(02) Primary or Lower 02	(3) Other Backward Castes (OBC) 3	OTHER	···

Table 2: Subset of the complete stratification frame, aggregated at the cell-level.

3.2 Working with Convenience Samples

Having described how we obtain representative stratification weights for a number of voter-categories of interest, we now move on to describe the behavioural training data we use to make our cell-level predictions. We collect micro-level convenience samples of Indian voters using two modalities: a first group comes from the Centre for Experimental Social Science (CESS)[38] online subject pool; a second group is reached via the Amazon

[38] https://cess-nuffield.nuff.ox.ac.uk

Mechanical Turk platform. The CESS online subject pool is recruited via Facebook advertisements, and hence its makeup is partially controlled to be closer to the population of interest; on the other hand, selection effects into the Facebook pool are such that this must for all intents and purposes be treated as a non-probability convenience sample.

Participants are asked to provide us with personal characteristics to fill in the cells of our stratification frame, as well as their current and past political behaviour. Political-behaviour questions include recalled 2014 behaviour, self-reported likelihood of 2019 turnout[39] and 2019 vote choice[40] and others.

We start monitoring on February 20^{th}, roughly 13 weeks before the election date of the 19^{th} of May; we conclude monitoring on April 10^{th}, the day before the beginning of the 6 weeks long election. Figure A5 in the Appendix shows the sample size per week to election, broken down by source and voting intention. Note that each coalition we are examining is composed of a myriad of parties; Table A1 in the Appendix addresses our allocation of parties to each coalition[41]. As of the 11^{th} of April 2019, we have sampled $6,785$ voting intentions from $5,807$ users[42]; Mechanical Turks contributed $5,152$ responses from $4,227$ users; the Facebook-recruited subject pool contributed $1,633$ responses from $1,580$ users.

Much like the stratification frame, one initial hurdle in online survey data is missing data. These are generated when an individual responds with a *'don't know'* or *'prefer not to answer'* to any of our survey questions; for 2014 voting behaviour, we further have missing values from February 20^{th} to March 7^{th}, as we had not yet devised a method to post-stratify by past-vote, and hence had initially not asked past-behaviour questions. With the exception of caste, missing values on demographic covariates are rare - typically under 1% of the sample, with age-category being the worst offender at

[39] 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'.

[40] 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 are treated as akin to 'Not sure at this point in time'.

[41] The allegiances of a number of small parties was shifty throughout the election, and researchers could disagree as to their coalition allocation. Fortunately the alliance-changing parties tend to be small and ignorable at the national level, though they may well contribute to small discrepancies between this work and others.

[42] A unique user is identified using their IP address.

1.5% missing. 10% of the respondents have missing caste, as a result of there being some surnames or caste sub-groups (*Jati*) which we could not match to any of the 6 broader caste-categories used by the IHDS[43]. 12% of 2019 turnout responses and 10% of 2019 vote-choice responses are missing, whilst respectively 32% and 37% of responses had missing values for 2014 turnout and voting behaviour.

We deploy a similar strategy as we did in dealing with missing values for our stratification frame, by assuming our `NA`s are missing at random, and using multiple imputation with chained equations, to again impute these via random-forest. Note we included an area-level predictor to make the imputations more robust, and otherwise used the same hyper-parameters we did for the imputation stratification-frame imputation model. We assess the success of the multiple imputation procedure using the Out of Bag (OOB) error provided to us by `missForest`, available in Table A3 in the appendix. We can skip over the OOB error for demographics, as there were so few missing values for these variables that the imputation will not make much of a difference. With respect to caste, 74% of OOB observations are correctly classified. 71% of 2014 turnout observations, and 83% of 2014 vote-choice are correctly classified; in contrast, 80% of 2019 turnout and 82% of 2019 vote-choice are correctly predicted. Overall, to the extent that the OOB error is a good indicator of imputation quality, the chained random forests achieve reasonable classification-rates, leading us to have confidence that the complete imputed dataset is a plausible draw from the space of voters we could have included in a convenience sample.

3.2.1 Turnout Training Data

Looking back at Figure A5, it is clear that our convenience samples have a turnout problem: after imputation, only 4% of the 2019 turnout responses are negative, suggesting the presence of both selection effects and heavy over-reporting. To obtain a better measure with which to train a turnout model, we opt to use the unadulterated 2014 NES, and assume that turnout behaviour across demographics will be stable between 2014 and 2019; this assumption worked for Lauderdale et al. (Lauderdale et al. 2020) in the UK and US, and seems plausible a-priori in India too. Though the NES also suffers from turnout selection and over-reporting, as we saw during the construction of

[43]The 6 case categories use are: *Brahmin; Forward/General (except Brahmin); Other Backward Castes (OBC); Scheduled Castes (SC); Scheduled Tribes (ST); Others*

the stratification frame, the NES has a larger proportion of non-voters (9%); it is a more representative (to be explored in the next section) and larger (see Table 1) survey, and was performed closer to the 2014 election, hence less likely to suffer from recall bias. It too suffers from a small number of missing values: demographic variables are almost entirely complete, with the exception of educational attainment (5% missing) and caste (9% missing); 2014 turnout is close-to-complete, with a mere 0.5% missing, whilst 12% of 2014 vote-choice is also missing. Again we deploy `missRanger` to complete the survey, and display the relative imputation accuracy, as measured by OOB errors, in Table A3. The imputation quality seems to be worse for the NES, perhaps a reflection that there is more variance to explain in the survey. OOB education is correctly classified 71% of the time, whilst Caste imputation is more noisy - doing just better than a coin-toss at 57% correct classification rate. Fortunately, the proportion missing for these variables is small, hence we can accept noisy estimates. 2014 voting behaviour is better predicted, with 2014 turnout being correctly classified 95% of the time, whilst 2014 vote-choice is is correctly allocated 67% of the time. All in all, considering the relatively small number of missing values, and the accuracy of the imputation algorithm with respect to turnout, we are confident in using the complete dataset to train our turnout model.

Having dealt with missing values, we still feel it's important to somehow account for turnout over-reporting. We make a blanket assumption of uniform over-reporting, assume the sample is otherwise representative, and correct the turnout values of the completed 2014 NES by randomly 'turning off' individuals who declare they have gone to the polls in 2014, repeating the process until we reach a percentage turnout equal to the observed 2014 national turnout - namely around 66.5%. If our assumptions are correct, and turnout is stable across demographics over time, predictions from a model trained on this completed and corrected NES 2014 data should produce be reliable estimates of 2019 turnout.

3.2.2 Severity of Non-Representation

In this section we assess the degree to which our convenience samples, and the NES, are representative of the voting-age population. Throughout, the reader is advised to keep in mind that the reference population figures are themselves modeled, as they are directly derived from the stratification frame, and hence subject to error. We will seek to develop

ways to account for uncertainty in building stratification frames in future work. Figures A6 to A14 in the Appendix provide an overview, obtained by plotting the distributions of a select number of variables in the stratification frame (our target population), against their sample counter-parts. Note that sampling constraints based on zone, age and gender were deployed by CESS in the later stages of sampling for the Facebook pool, which enabled for a more representative pool compared to the Mechanical Turks, though these constraints seem to have been more effective on geography than gender or age.

With respect to geographic distribution across the samples (Appendix - Figures A7 and A8), we can confirm findings (Boas et al. 2020, Dietrich & Winters 2015) regarding the massive over-representation of two states, namely Tamil Nadu and Kerala, which together make up only 9.8% of the voting age population in our stratification frame, and yet represent over 80% of Mechanical Turks; geographic over-representation in the Facebook sample is not severe, partly as it is controlled by sampling constraints, and partly because the geographical distribution of Facebook users is more representative of the population (Boas et al. 2020). The distribution of states in the 2014 NES mirrors that of the population.

The distribution of gender (Appendix - Figure A9) is realistic in the NES, whilst both convenience samples over-represent males - with Mechanical Turks being further skewed toward males than Facebook users. With respect to religious affiliation (Appendix - Figure A10), our sample is reasonably representative of the population, perhaps with the oddity of Christians being over-represented amongst Mechanical Turks. With respect to age (Appendix - Figure A13), we note that whilst the NES is representative of the population at large, the convenience samples are biased: Mechanical Turks are overwhelmingly between 25 and 35 years old, whilst Facebook userks are even younger, exibiting over-representation of all under 35s. Caste (Appendix - Figure A11) seems somewhat problematic: perhaps unsurprisingly, Brahmins make up a plurality in both our convenience samples, and we heavily over-represent individuals in the 'Other' caste category - which include people who do not subscribe to caste system, refused to be labeled or explicitly selected an 'Other' option - a response significantly more common amongst Facebook users. We also note that individuals from 'Scheduled Tribes' are under-represented. The distribution of caste in the NES, on the other hand, seems to over-represent Backwards Castes. The NES best approximates the population in terms

of yearly household income (Appendix - Figure A12), followed by the Mechanical Turk sample, with the Facebook sample being least representative. Whilst Mechanical Turks are richer than the voting-age population, Facebook users are overwhelmingly high-income, with 59% of the sample clustered in the top-two income categories, in contrast with Mechanical Turk concentration in the top-two brackets of 16% and a reference population concentration of 1.6%. We find the most severe bias in the education levels (Appendix - Figure A14): whilst the NES over-samples individuals with no formal education, but otherwise displays a reasonable distribution of education levels, the two convenience samples are both made up largely of graduate-educated individuals: 92% of our Mechanical Turks have some graduate education, in contrast with a striking 95% of our Facebook sample and a mere 8.9% of the population at large.

Regarding voting behaviour in the last election (Appendix - Figure A6), it's unsurprising that the 2014 NES best approximates the vote, though again significantly over-reporting turnout levels. Individuals from the Facebook group are significantly more likely to have voted for the NDA in 2014, and significantly less likely to have voted for the UPA; Mechanical Turks are somewhat more likely to have voted NDA in the last election, though they are also more likely to have voted for the UPA - and significantly less likely to have voter for a third alliance. Turnout behaviour is exceedingly un-representative for both convenience samples, with 97% of all respondents saying they turned out to vote, compared to 66% of eligible voters in 2014, reflecting both a higher degree of political engagement due to their demographics, but almost certainly a high degree of over-reporting.

In summary, excluding the NES, we are faced with two non-representative samples with similar biases: these samples have attracted a public which is more male, younger, higher-status, highly educated, and highly politically engaged. These characteristics are consistent with common selection effects across these sampling mediums, which are to do with online literacy and English language proficiency. Where they differ is in the medium-specific selection effects, namely the work-environment provided by the Mechanical Turk platform, and the leisure and connectivity provided by Facebook. These modes of use provide incentives which select for different people: Turks are more diverse in their income structure, and less diverse in their age-cohort; they are also less

diverse geographically, reflecting the uneven penetration of desktop internet access. The Facebook pool is more geographically diverse and does a better job at filling multiple age-cohorts, as well as being more representative of Gender. Overall, these findings are consistent with, and reinforce, previous studies on the nature of Indian online sampling (Boas et al. 2020, Charnysh et al. 2015, Dietrich & Winters 2015, Antin & Shaw 2012).

4 Modeling Behaviour

Having built a reliable post-stratification frame, and having collected relevant training data for turnout and vote-choice, we now seek to model behaviour at the cell-level. The ultimate goal of this excercise is to aggregate post-stratified cell-level behaviour; in order to present our aggregation formula, we first define some terminology: let $i \in 1, \ldots, N$ be an index of cells, where each cell represents a row in our stratification frame; let X denote the covariate characteristics which uniquely identify a cell, and $\mathrm{N}_i(X)$ be the number of individuals belonging to cell i in the population; finally, let $\Pr_i(T \mid X)$ be the turnout probability of each cell, conditional on the cells' characteristics, and $\Pr_i(V \mid T, X)$ be the vote-choice density of each cell, conditional on turnout and the cells' characteristics. Then, we can calculate out national point estimate for turnout, $\hat{\tau}$, and our national estimate of vote-choice conditional on turnout, $\hat{\nu}$, as follows:

$$\hat{\tau} = \frac{\sum_i \hat{\Pr}_i(T \mid X) \times \hat{\mathrm{N}}_i(X)}{\sum_i \hat{\mathrm{N}}_i(X)}; \tag{12}$$

$$\tag{13}$$

$$\hat{\nu} = \frac{\sum_i \hat{\Pr}_i(V \mid T, X) \times \hat{\Pr}_i(T \mid X) \times \hat{\mathrm{N}}_i(X)}{\sum_i \hat{\Pr}_i(T \mid X) \times \hat{\mathrm{N}}_i(X)}. \tag{14}$$

Note that, whilst $\hat{\Pr}_i(T|X)$ is estimated using exclusively the 2014 NES data, $\hat{\Pr}_i(V|T, X)$ is actually taken to be the simple average over our convenience samples $s \in \{\text{MechTurk}, \text{fb}\}$:

$$\hat{\Pr}_i(V \mid T, X) = \frac{\sum_s \hat{\Pr}_i(V \mid T, X, S = s)}{\sum_s \mathbb{1}\{S = s\}}. \tag{15}$$

Typically we are interested in uncertainty around these point estimates. We distinguish two relevant types of uncertainty: parametric uncertainty, which deals with uncertainty about the point-estimates of conditional turnout and vote choice distributions; and

sampling uncertainty, which, conditional on a given estimate of the conditional densities, accounts for variance across random draws of the turnout and vote-choices for each cell. A sensible way to fully characterize the uncertainty in our model is to simulate the election D times, where separate random draws are indexed by $d = 1, ..., D$, according to a probabilistic model:

$$T_{d,i} \sim \text{Binomial}\left(\Pr_{d,i}(T \mid X), \hat{\text{N}}_i(X)\right); \tag{16}$$

$$V_{d,i} \sim \text{Multinomial}\left(\Pr^{\star}_{d,i}(V = j \mid T, X, S), T_{d,i}\right); \tag{17}$$

$$\Pr_{d,i}(T \mid X) \sim \text{Beta}\left(\mu_i^T, \sigma_i^T\right); \tag{18}$$

$$\Pr_{d,i}(V = j \mid T, X, S) \sim \text{Beta}\left(\mu_i^{V=j}, \sigma_i^{V=j}\right); \tag{19}$$

$$\Pr^{\star}_{d,i}(V = j \mid T, X, S = s) = \frac{\Pr_{d,i}(V = j \mid T, X, S)}{\sum_j \Pr_{d,i}(V = j \mid T, X, S)}; \tag{20}$$

where relevant probability densities are assumed to behave according to a series of independent Beta distributions, each fully specified by a mean and a variance parameter; vote choice quantities are also assumed to be independent, and normalized a-posteriori - more on this in the next paragraph. The objective of our estimation strategy is to obtain estimates for μs and σs. By aggregating each set of simulations d as per the aggregation formulae described at the beginning of this paragraph, we can recover an empirical density for aggregated vote-choice and turnout, and hence make probabilistic statements about the election results.

Multivariate Outcomes. The parametric density $\Pr_i(V \mid T, X)$ is multivariate in nature - that is, we are interested in predicting the vote-choice probabilities for individual i with respect to each option $j = \{\text{NDA}, \text{UPA}; \text{Other}\}$. To estimate these quantities we rely on the *one v. all* (Galar et al. 2011, Rifkin & Klautau 2004) strategy for predicting choice-probabilities for each category. This involves normalization of the resulting scores into probabilities via either standard normalization, or a more complex multivariate link

function such as *softmax* (Goodfellow et al. 2016). Softmax is especially relevant if we encounter corner-probabilities, as it will tend to add weight to probabilities close to 1, and down-weight probabilities close to 0, relative to the other components one is normalizing over. The context where it is most useful is in multi-layer neural networks for classification, where it works by increasing the probability of extracting strong and definitive signals over noise. In our Lok Sabha study, given we are dealing with a single layer of predictions for three alliances with similar levels of popularity, we have no reason to think employing softmax normalization will provide any advantage over standard normalization. Hence we resolve to normalize the predicted mean of the vote-choice densities as shown in the Equation 20.

Choosing a Learner. Traditional regression and post-stratification applications resolve to estimate $\Pr_i(V \mid T, X)$ and $\Pr_i(T \mid X)$ using multilevel regression (Hanretty 2019*b*) (Mr). Mr is endowed with a number of useful properties: i) *regularization* - Mr is preferable to simple regression is its ability to *regularize* predictions (Gelman 2018), also known as 'borrowing of strength'. This quality is particularly useful when data is scarce at the micro-level of interest, and predictions can rely on the global characteristics of the data, as is often the case for pre-election surveys; ii) *interpretability* - these parametric models are not black-boxes, and size and direction of effects is always clear to the modeler; furthermore, they align conceptually with the hierarchical nature of the data, allowing the researcher to think about modeling level-specific heterogeneity; iii) *functional exactness* - if a modeler knows the true functional form of the data generating process, they can select the optimal number and structure of additive parameters, as well as the optimal subset of the covariate space that reflect the truth; iv) *tractable uncertainty* - uncertainty estimates around coefficients and predictions are known, as is the functional form of the parameters, hence enabling the estimation of prediction intervals. In Bayesian settings, Mr can be estimated using Monte-Carlo methods (Gelman et al. 2013) for moderate sample-sizes, which enable the use of simulations for making inference under uncertainty.

Irrespective of these useful properties, we have found in practice that it can be difficult to work with Mr in the context of pre-election modeling: i) Mr models require fine-tuned *model specification*, and predictions can be hugely sensitive to these. This

sensitivity precludes the use of MrP in settings where not much is known about the underlying functional form; for instance whilst Lauderdale et al. (Lauderdale & Blumenau 2019) manage to leverage cross-level interactions to produce relatively accurate estimates of vote-choice in the 2019 UK election, a similar model would be difficult to produce for democracies where the literature on said cross-level interactions is thinner or inconclusive; we believe India to be such a case; ii) excessive regularization can lead to *attenuation bias* at the aggregate level - though Hanretty et al. (Hanretty et al. 2018) have shown that area-level attenuation bias can be reduced by including a substantive area-level predictor, degrees of said bias persist at other individual-levels, at the discretion of the modeler. As an example, we can consider whether to have an 'Age' random effect; if the random effect is applied even when the underlying data provides enough evidence that each age-category effect can be precisely estimated under independence, then the shrinkage-effect applied by multilevel modeling will lead to attenuation-bias at the age-level; iii) in our opinion, the most severe infractions of Mr models come in the context of computational speed and scalability concerns. Monte-Carlo estimation of Bayesian Multinomial models *does not support large N* - state-of-the-art for UK take 12 hours, for around $100,000$ responses.

The MrP literature is not oblivious to these challenges: sparse regression modeling (Goplerud et al. 2018) has been proposed as a potential solution to enforce automated variable-selection, whilst variational inference (Kucukelbir et al. 2017, Goplerud 2020) has been proposed to improve scalability, at the cost of typically biased uncertainty estimates. Regression and post-stratification practitioners have also turned to the machine learning literature: Bayeisan Additive Regression Trees (BART) (Bisbee 2019) have been shown to out-perform regular Mr in some applications; Random Forests (Cerina & Duch 2020) have been used to deal with high-dimensional settings; stacking (Ornstein 2019) of several non-linear learners shows promise, though it too suffers from unfeasible computational speed.

It is worth noting that there are plausible criticisms to the use machine-learning methods which deploy automatic variable selection, adaptive shrinkage and non-linear modeling: i) in the context of public opinion polling, some practitioners argue that *non-linear effects are limited* - due to typically small sample sizes and relatively well understood additive functional forms - hence the departure from existing practice is not

fully justified given the context; ii) loss of *interpretability*: though some learners - such as Classification and Regression Trees (CART) (Breiman et al. 1984) - retain intuitive interpretation, and the literature has developed to provide measures of importance and significance of the input features (Strobl et al. 2008, Mentch & Hooker 2016) for more complex algorithms, most require a leap of faith into different degrees of black-box thinking; iii) superficial treatment of *prediction uncertainty*: with some notable exceptions due to recent advances in the statistical understanding of bagging estimators (Wager & Athey 2018, Athey et al. 2019, Lu & Hardin 2019), cross-validation (Kohavi et al. 1995) is the industry-standard for obtaining uncertainty estimates around machine-learning predictions. This puts excessive reliance on the training sample to house as deep and detailed a picture of the population of interest - a plausible assumption under big-data conditions, which quickly breaks down for relatively small convenience samples.

Driven by these considerations, we choose to employ Random Forests (RF henceforth) (Breiman 2001*a*,*b*) as the backbone of our models to estimate $\Pr_i(V \mid T, X)$ and $\Pr_i(T \mid X)$: RFs are a *bagging* algorithm (also known as *bootstrap aggregation*), as well as an *ensemble* method (Hastie et al. 2009): the former refers to its quality to iteratively take a large number of bootstrap samples of the training records and features to fit ideally uncorrelated CARTs; the latter to its predictions being a simple average of said bootstrapped CARTs. RFs are optimized for speed, especially in their `ranger` implementation (Wright & Ziegler 2015), as they can rely on parallel processing and other computational tricks to several trees of optimal depth. RFs are amongst the most consistent performers in terms of their ability to produce high-quality predictions in various competitions, with some researchers claiming *'the random forest is clearly the best family of classifier'* (Fernández-Delgado et al. 2014), though this is contested elsewhere in the literature (Wainberg et al. 2016).

We take the RF predictor for the mean behaviour of a given cell i, μ_i, to be the the average of B regression trees, each fit to a random sample of training set $\mathrm{T}(y_o, x_o), \forall o = 1, ..., O$, where o indexes the respective training observations, accord-

ing to a set of random parameters $\Theta = \{\theta_1, ..., \theta_B\}$:

$$\mu_i = \frac{1}{B} \sum_b \phi_{i,b}(x, \theta); \tag{21}$$

where $\phi_{i,b}$ stands for the prediction of component tree b for cell i.

Hyper-Parameter Tuning. RFs are ensembles of CARTs, and as such are governed by numerous hyper-parameters which describe the tree-growing mechanism. We use the same set of hyper-parameters for both our turnout and vote-choice predictions, as we have no a-priori reason to believe these would require separate adjustment. We set `num.trees = 1500`, hence growing a large-enough number of trees to ensure no Monte-Carlo effects are present in our estimates. We set the number of candidate features to split, `mtry = p`, where p is the total number of available features, hence imposing that all features be considered for splits in each tree, effectively transforming our RF to an ensemble of bagged CARTs. We do this as we are mindful of attenuation-bias, and we seek to relax the shrinkage effect of the forest. Furthermore, considering recent work in the UK (Lauderdale & Blumenau 2019) on the importance of cross-level interactions and non-linear effects, we prioritized a model which allows for these interactions to emerge in each tree, to the extent that they explain residual variance; if trees are precluded from considering these interactions due to any subset of the relevant variables being randomly out-of-bag, each tree will be a weaker predictor, and the ensemble will be weak as a result. Note further that the recent literature on growing probability-forests (Olson & Wyner 2018) explicitly suggests a large `mtry` - and in particular `mtry = p` - outperforms other tuning choices in some key UCI machine-learning repository datasets (Bache & Lichman 2013). In principle, we could have set `mtry` to any arbitrarily large number; however we found the conceptual equivalency with bootstrapped trees more attractive, as they represent a distinct, well-defined modeling strategy in the literature (Breiman 2001*b*, Hastie et al. 2009). On the other hand, having each tree consider the same covariate space may increase the correlation across trees, and decrease the ensemble predictive gains (Hastie et al. 2009); as a result, we reduce the size of each bootstrapped sample, `sample.fraction =` $\frac{1}{3}$ and impose sampling without replacement, `replace= FALSE`, such that each record in each bootstrap sample is unique, and the samples are small-enough that they are extremely likely to be highly heterogeneous amongst themselves.

Note that this will also help reduce fitting time. No limit is explicitly imposed on the depth of each tree, `max.depth = NULL`, nor on the minimum size of terminal nodes `min.node.size = 1`, again in an effort to encourage the discovery of deep, non-linear patterns.

Prediction Error. Wager and Athey (Wager & Athey 2018, Athey et al. 2019) have shown that Random Forest predictions are asymptotically Gaussian in typical regression settings, and provide a consistent estimator of the asymptotic variance via the infinitesimal jackknife procedure (Wager et al. 2014). Lu and Hardin (Lu & Hardin 2019) have instead developed a consistent estimator of the conditional prediction error for RFs, leveraging the OOB observations in each tree. The simple version of their procedure is highly intuitive: for every cell in the test-set, they go through each tree in the forest and record its out-of-bag neighbours - observations which were randomly not chosen to train the given tree, but were predicted to be in the same terminal-node as the cell-observation; they then calculate the error for each of these neighbours, as their true observed outcome is available to match against the prediction for each tree, and take a weighted average of this error based on the number of times each of these observations is an out-of-bag co-habitant of the test-cell. The resulting score is the conditional MSPE.

We should note here that, to our knowledge, no estimator for the prediction error of probability machines - i.e. models which output predictions in the range $[0, 1]$ - has been thus far presented in the literature. The Lu & Hardin estimator does not impose a functional form over the distribution of the error term, hence it can be directly employed to parametrize non-gaussian distributions with $[0, 1]$ support, which are also indexed by some measure of MSPE. The Beta distribution is such a density, as it can be fully parametrized by its mean and variance, as suggested in Equations 18 and 19 earlier in this section.

It follows from the definition of the Beta distribution that: i) $\mu_i^T \in (0, 1)$ and $\mu_i^{V=j} \in (0, 1)$; ii) $(\sigma_i^T)^2 < \mu_i^T(1 - \mu_i^T)$ and $(\sigma_i^{V=j})^2 < \mu_i^{V=j}(1 - \mu_i^{V=j})$. In our application, the first set of conditions regarding the mean holds for the entirety of our RF predictions; the second set - the conditions relating to the variance parameter - are more problematic. The effect the truncation has on the distribution of mean turnout and mean

vote-choice across cells can bee seen from Figure A15 in the Appendix. The red dots in the plot indicate predictions whose MSPE exceeds the beta-imposed limit; we can see that a) these are a small proportion of the total cells, for each predicted quantity, and b) that the limit imposed by the distribution is typically close to the estimated MSPE, and only really a problem for means closer to the corners of the distribution, hence we do not expect the correction to the limit to have significant consequences on our estimates. Having assumed this functional form, we can simulate $D = 500$ draws from our probabilistic model, and obtain credible uncertainty estimates around our cell-level behavioural predictions, which can be aggregated by each simulation draw to preserve uncertainty at the national-level.

A final note on the estimation of the MSPE: in practice, the `forestError` package for `R` with which the MSPE estimator is implemented, is exceedingly computationally expensive. To overcome computational costs, we deploy a highly practical strategy: we randomly sample 1% of the cells in our stratification frame - an arbitrary amount large-enough to contain a reasonable amount of heterogeneity, but small enough to limit the computational burden - and run the `forestError` algorithm to derive their conditional MSPE; we then use these predicted erorrs as training data to estimate the error of the remaining cells, via RF regression implemented by `ranger`, using the default hyper-parameters. This provides an approximate measure of uncertainty for each cell in the stratification frame; note that our practical problem is to obtain reasonable confidence bounds around national-level estimates of vote-share, and as a result we can afford to be imprecise at the cell-level, as we expect errors to cancel-out by aggregating simulations at the national level.

5 Results

In this section we present the results of our estimation procedure. We first outline the behavioural predictions focusing on different geographical levels of aggregation, and show promising levels of accuracy are obtainable from cheap and unrepresentative surveys such as those treated in this paper, depending on the statistical power underpinning the desired level of analysis. We then apply a uniform swing model (Butler & Stoke 1974)

to obtain predictions about seat-counts, and show our predictions outperform other available polling. Note that the election unfolds over 5 weeks, and we stop monitoring right before the beginning of voting, hence our point estimates refer to our predictions roughly 5 weeks out from the final election day.

It is worth underscoring that the outcomes of interest for the evaluation of our prediction framework are exclusively at the national level: the headline number of seats assigned to each major alliance in the Indian parliament after the election, and the vote share for the two largest alliances (UPA and NDA). Headline seats are important because they tell us who will govern after the election; it is commonplace for the literature on election forecasting to focus on headline seats (Whiteley et al. 2011, Fisher et al. 2011, Fisher 2015, Ford et al. 2016), though it is less common for pollsters to do so. On the other hand, Indian opinion polls typically release seats-projections exclusively, and though the exact methodology they use to translate polls to seats is not transparent, we have some evidence that they - like us - use uniform swing models (Solomon 2016). Hence our only option to recover an informative comparison with the Indian polling industry is by comparing seats generated via a Uniform Swing model informed by national vote share estimates. All that said, a small number of pollsters did release vote-share projections for certain states - notably the most populous state of the country, Uttar Pradesh, received more extensive scrutiny. We chose not to focus on these comparisons as a point of evaluation of the prediction framework for a number of reasons: i) we knew from the start we were under-powered a-priori to model state-level outcomes of interest. We know from other research dealing with regression and post-stratification (Lauderdale et al. 2020, Lauderdale 2019, Hanretty 2019) that we need about 100 respondents per area of interest on average for modeling to be successful; moreover, we need these area-level sub-samples to be diverse enough to inform area-level interactions. We are much below the 100 people-per-state mark, and moreover our sample is dramatically selected - as we have shown in previous sections; ii) a district-level stratification frame was not feasible, as there exists - to our knowledge - no dataset with joint demographic information at the district level. We could have modeled these district-level features outright, but this would have lead us further to one extreme of the depth-precision trade-off (namely favouring depth in spite of dramatic loss of precision); iii) we knew coming into this

work that the Uniform Swing model was a valid option in India, thanks to seminal work by Butler, Lahiri and others (Butler et al. 1995, Lahiri & Roy 1984) and hence designed a strategy that would exploit this model. In the following paragraphs, we will touch on sub-national performance of our model, and show that as we aggregate up to the national level, model performance increases; we take this as evidence in favour of our conjectures with respect to the lack of sub-national power in our sample, and respective lack of sub-national precision in our stratification frame.

National Behaviour. We begin with our predictions about National-level behaviour. Figure 1 shows the proportion of individuals who would turn out to vote over the course of the voting period; these predictions are made using the NES 2014 as training data, correcting for over-reporting and assuming stable turnout dynamics over demographics. Our prediction is only 0.6% away from the truth - our point estimate is 66.9% against an observed turnout of 67.5% - suggesting our assumptions about the 2014 NES were useful. The three standard deviation prediction uncertainty over the mean is just over ± 1, which is rather narrow. Figure 1 shows the national vote share predictions over

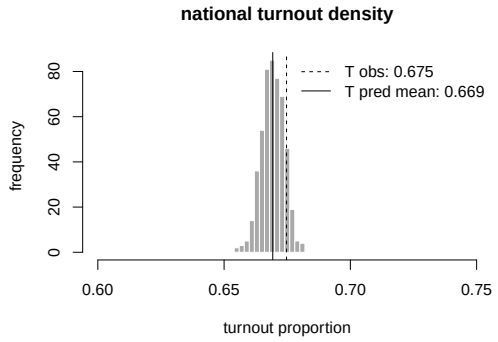


Figure 1: Predicted distribution of turnout at the national level.

the course of the campaign; after the beginning of voting, we stop monitoring and predictions are projected forward according to a random walk. At five weeks to the end of election, the predicted point estimate for the NDA is 41.6%, against an observed value of 45.0% - contributing to a prediction error of 3.4% points; the prediction for the UPA is seldom perfect, with a point estimate of just under 30.6% against an observed vote share of just under 31.3%, contributing to a prediction error of around 0.5%; finally, the remaining error on the 'Other' alliances is fully specified by the error on the two larger alliances, as they have to sum to one, accounting for just under a 4% over-estimation.

On average, we have a prediction error of 2.6%, which is within the range of a standard polling error in western democracies, obtained via traditional polling methods (Jennings & Wlezien 2018). With respect to uncertainty, again we see a three standard deviation interval of ±1 - far too narrow for comfort. If we project our estimates to the last day of the election, according to a random walk, the intervals for NDA, UPA, and 'Other' alliances expand respectively to ±8, ±7 and ±3 points, indicating significant scope for movement over the voting period, according to the behavioural variance observed up to the beginning of voting.

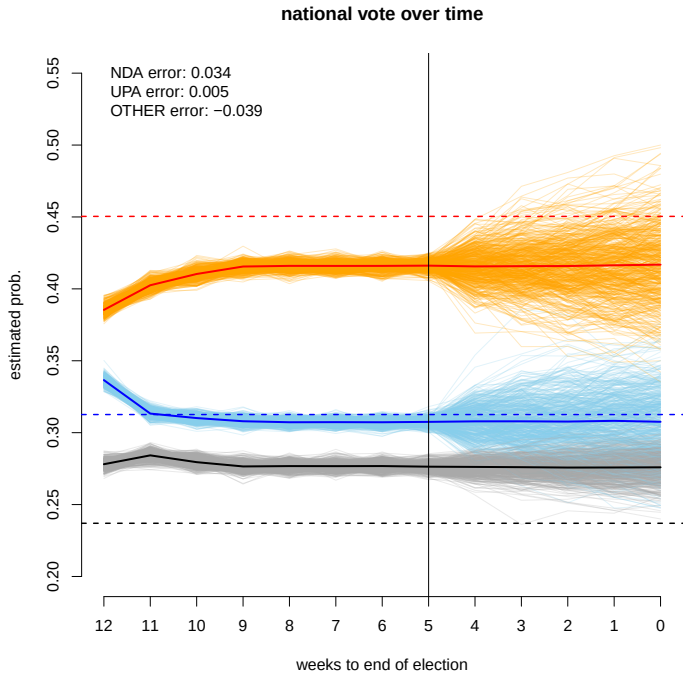


Figure 2: National vote share for the three major alliances over the course of the campaign; monitoring stops before the beginning of voting, and vote-share after this point

Feature Importance. Measures of the 'importance' of each feature can be used to perform hypothesis testing (Strobl et al. 2008); they can further work as a check that the inner-workings of black-box models are sensible. We use a permutation *variable importance* (VIM) (Altmann et al. 2010, Janitza et al. 2018) metric as a check on the mechanistic of the RF algorithm, to see if the notions known from theory of voting - i.e. the significance of the relationship between age and turnout - are being learned and leveraged in some way by the algorithm; we further use VIM to rank variables according to their predictive contribution. The VIM score is calculated by measuring the change

in average OOB predictive accuracy when each feature is rendered more stochastic via random permutation; larger changes in predictive accuracy reflect the outsized impact of the predictor in question. Non-parametric hypothesis testing can then be performed by simulating a distribution of the predictors' VIM under the null hypothesis that the predictor is 'unimportant', and recording the percentage of null-scores that are greater than the observed VIM. The null distribution is obtained by randomly permuting the response variable P times, and recording the associated 'null importance' of the features for each permutation. When strong predictors are correlated, such that there can be clusters of highly predictive correlated features, the Altmann et al. (Altmann et al. 2010) permutation VIM distributes importance across correlated features, and assigns significant p-values to all of the predictive features in the group. This suggests we should not interpret the effects independently, but rather comment on the total effect with their group-cohabitants. Figure A16 in the Appendix presents the most significant feature importance scores for each of our behavioural models.

Beginning with our turnout model, it is clear that age is playing a pivotal role, both at the individual-level and the aggregate level: aggregate-level age variables, and their correlates (i.e. % children in household) score the highest importance, along with individual-level age. Religion and caste seem to play some role, with multiple caste-related state-level variables registering some significance. Aggregate-level turnout behaviour, and individual-level alliance of choice, also play a significant role in these turnout predictions.

When it comes to vote-choice, the first thing we note that that 2014 behaviour is the most predictive variable for 2019 behaviour across all alliances. Caste is also a significant determinant of vote-choice, whilst religious affiliation plays a key role in the UPA vote, perhaps indicating the relatively outsized Muslim affiliation with the alliance. State-level caste-composition and age profiles also show significance. Many distance variables are significant across alliances, indicating cross-state correlation, geographical patterns in the vote and a degree of local smoothing. Weeks to the end of the election play a significant role for all alliances, evidence of significant campaign effects. It is further striking that there seem to be many more significant predictors of the vote for the

NDA than for other alliances, with the 'Other' group being the least predictable (low VIM magnitude per variable and little significance across these), perhaps a sign of the heterogeneity of the 'Other' category, which is a derivative and ad-hoc conglomeration of very different regional and national parties.

Sub-national Behaviour. Typical MrP applications tend to have as primary objective the estimation of behavioural outcomes at the sub-national level (Hanretty 2019*b*). Our application is different, in that we use regression and post-stratification simply as a way to produce more robust national-level forecasts from non-representative surveys. Note that we could not produce constituency-level results directly from the model, as we had no constituency-level stratification frame to post-stratify against, due to the general lack of availability of constituency-level cross-tabs. Nevertheless, we may still want to test the performance of our method at the sub-national level, to get a sense of the degree to which relatively small and unrepresentative surveys can be smoothed to predict sub-national behaviour in this high-heterogeneity scenario. We select two geographical levels of interest: zones and states; there 6 zones and 35 states in our application[44]. Figure 3 and Figure A17 (in the Appendix) compare our predictions at the sub-national level with the observed election results.

Overall, sub-national accuracy is poor. We expected this to be the case, especially given the highly skewed and unrepresentative geographic distributions of our survey data. Zone-level predictions outperform state-level ones significantly, showing reasonably high cross-zone correlations for the two major alliances. In particular, state-level forecasts for the NDA vote are correlated with the observed outcomes at 0.97, suggesting a reasonable degree sub-national signal was present in the pro-NDA survey response. The poor performance of our predictions as we get to deeper geographical levels suggests a lack of area-level statistical power is partly to blame. In circumstances where there is high cross-area correlations, as is often the case in US or UK sub-national estimation, area-specific statistical power is less of an issue (Buttice & Highton 2013) than it is in India, where cross-state heterogeneity in political behaviour can be dramatic.

[44]Though Telangana formally seceded from Andhra Pradesh in 2014, we ignore this split and still consider these two states as one. This allows us to avoid having to adjust census and IHDS population estimates for a state which at the time was not recorded.

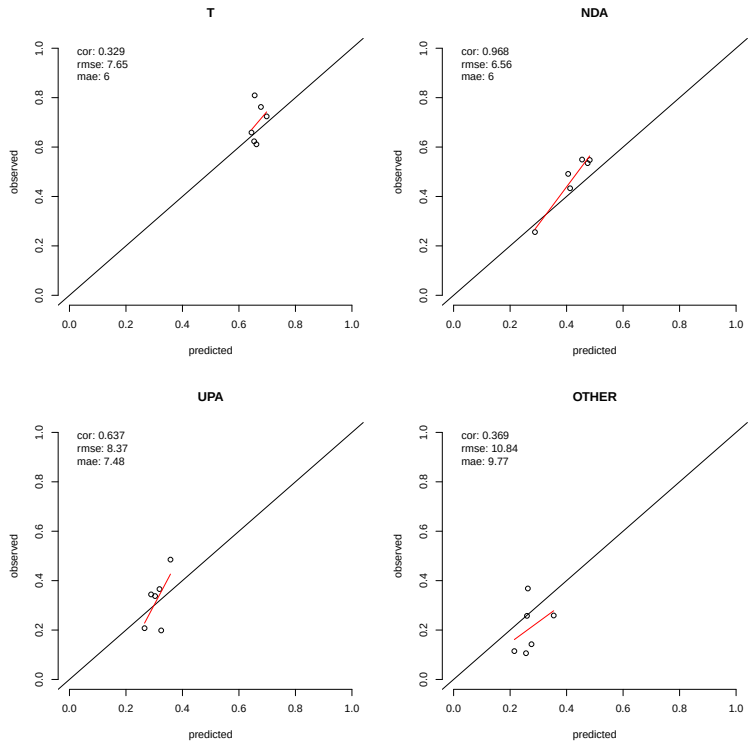


Figure 3: Zone-level predictions v. observed zone-level 2019 behaviour.

Seats Projection. The final and most important step in our polling effort is in 'calling' the election, i.e. assigning a predicted number of seats to each alliance, and measuring the probability of obtaining a majority in the Lok-Sabha. To do this, we convert our national-level vote-share estimates into seats by applying a 'uniform swing' (Butler & Stoke 1974, Jackman 2014). The basic premise of this model is that every constituency will 'swing' with the national vote, such that a national change from the last election in the vote-share of each party will reflect an identical change at the constituency level. More formally, we can estimate the 2019 constituency-level vote share of a given party as follows:

$$\hat{\delta}_{j,2019} = \hat{\nu}_{j,2019} - \nu_{j,2014}; \tag{22}$$

$$\hat{\lambda}_{j,2019,c} = \lambda_{j,2014,c} + \hat{\delta}_{j,2019}; \tag{23}$$

where $\delta_{j,t}$ indicates the change in vote share for a given party, for an election held at time t; $\lambda_{j,t,c}$ is the vote share for party j at time t in constituency c, for $c = 1, ..., C$.

Under this model, a specific constituency may be poorly predicted, but the headline share of seats tends to be accurate; this is because the true constituency swings

tend to be realizations of a Normal distribution with the area-level swing as its mean. We can expect some constituencies to go the opposite way of the area-level swing, and others to be swinging more heavily toward that trend; these extremes will appear at roughly the same frequencies and hence cancel each other out. At the national level, the uniform swing model 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. Though India is a far more heterogeneous country, there is evidence that a uniform swing may have substantial predictive power (Butler & Stoke 1974, Lahiri & Roy 1984).

We would like to compare our estimates of the election result to of state-of-the-art Indian election polling. Indian polls typically release seat counts as opposed to vote-shares, hence we can perform a direct comparison with our seats predictions. In the context of India, the most common opinion polls which are published on a relatively regular basis before the beginning of the election period, are seats-projection polls. The methodology for these polls is largely a black-box. We obtain publicly available opinion polls for the 2019 election from the relevant Wikipedia page[45], which was being constantly updated and scrutinized, and represented the most reliable source of polling reports throughout the election. For each poll reported on the Wikipedia page, we try to distinguish between the polling company and the news organization which commissioned the poll; where no details on the polling company are available, we assume the news organization did the polling in-house. Pre-election polling stopped at around 40 days before voting begun; hence we compare our estimates at 5 weeks until the end of the election, with the last published poll of each of 10 polling houses which completed their last published poll at least a year before the last day of voting in 2019. Amongst the polling houses, we can identify three broad groups: 6 of the houses in question, namely 'CNX', 'Jan Ki Baat', 'VMR', 'News Nation', 'CVoter' and 'Zee 24 Taas' polled the campaign continuously, and published their last poll in the final week before voting begun - within 50 days to the end of the election; 3 pollsters, namely 'VDP Associates', 'Karvy' and 'India Tody',

[45] https://en.wikipedia.org/wiki/Opinion_polling_for_the_2019_Indian_general_election

polled less regularly and published their last projections between 140 and 80 days to the end of the election; finally, 'CSDS' and 'ABP News' released their last projection between a full year and 200 days to the end of the election. Figure 4 presents our seats projections, and Table 3 describes how our error rates compare to other pollsters.

House	Days Left	NDA Pred.	Error	UPA Pred.	Error	Other Pred.	Error	Abs. Error	NDA - UPA Pred.	Error
Cerina & Duch 2019	39	327	**27**	83	**16**	133	**−44**	**87**	244	**11**
2014 Result		309	45	76	23	158	158	226	233	22
CNX	43	275	79	126	**−27**	142	-53	159	149	106
Jan Ki Baat	45	310	**44**	122	-23	111	**−22**	**89**	188	**67**
VMR	49	283	71	135	-36	125	-36	143	148	107
News Nation	49	270	84	134	-35	139	-50	169	136	119
CVoter	49	264	90	141	-42	138	-49	181	123	132
Zee 24 Taas	49	264	90	165	-66	114	-25	181	99	156
VDP Associates	80	242	112	148	-49	153	-64	225	94	161
Karvy	109	237	117	166	-67	140	-51	235	71	184
India Today	139	257	97	146	-47	140	-51	195	111	144
ABP News	200	276	78	112	-13	155	-66	157	164	91
CSDS	353	274	80	164	-65	105	**−16**	161	110	145

Table 3: Seat projections and errors across pollsters. In bold are the best performing pollsters for each column; the Other-error column has two bold quantities, given that the CSDS poll which is the best performing was so outdated, it was worth highlight a second-best performing poll which was more recent and comparable.

Firstly, it is immediately clear that our uniform-swing projections outperforms other pollsters: our poll has the lowest absolute error , 87 seats - with the Jan Ki Baat poll coming a close second with 89 wrongly assigned seats; a gulf of 54 seats stands between second and third place. Importantly, our methodology vastly outperforms other pollsters on the NDA-UPA lead, missing the mark by a mere 11 seats, when compared to the next-best-performing Jan Ki Baat poll, which misses the mark by 67 seats. This is the most important prediction quantity, as it typically directly translates to government formation. At the alliance level, our method is the best performing on both NDA and UPA vote shares, with Jan Ki Baat coming in second for the NDA vote, and CNX for the UPA vote: the former misses the NDA share by 44 seats, compared to our 27 seats error, whilst the latter fails to forecast the UPA seat share by 27 seats, compared to our 16. The 'Other' seat share is more troubling for our methodology: our 44 seats error places us 5^{th} in this category - though note this forecast is typically less consequential for the overall outcome. Finally, we note again that that uncertainty around the seats predictions is too narrow: the NDA forecast ranges 20 seats, between 319 and 339; the UPA forecast ranges 15 seats, between 73 and 88; and the 'Other' forecast ranges 16 seats, between 125 and 141. An indication of the over-confidence of the forecast is that, even if our poll is the best performing, none of our intervals include the observed results.

Part of this over-confidence is simply due to the uniform-swing assumption: even if we do incorporate variance in the swing, this typically understates constituency-level variance from the last election - hence each constituency's swing is actually more uncertain than the uniform swing model would suggest.

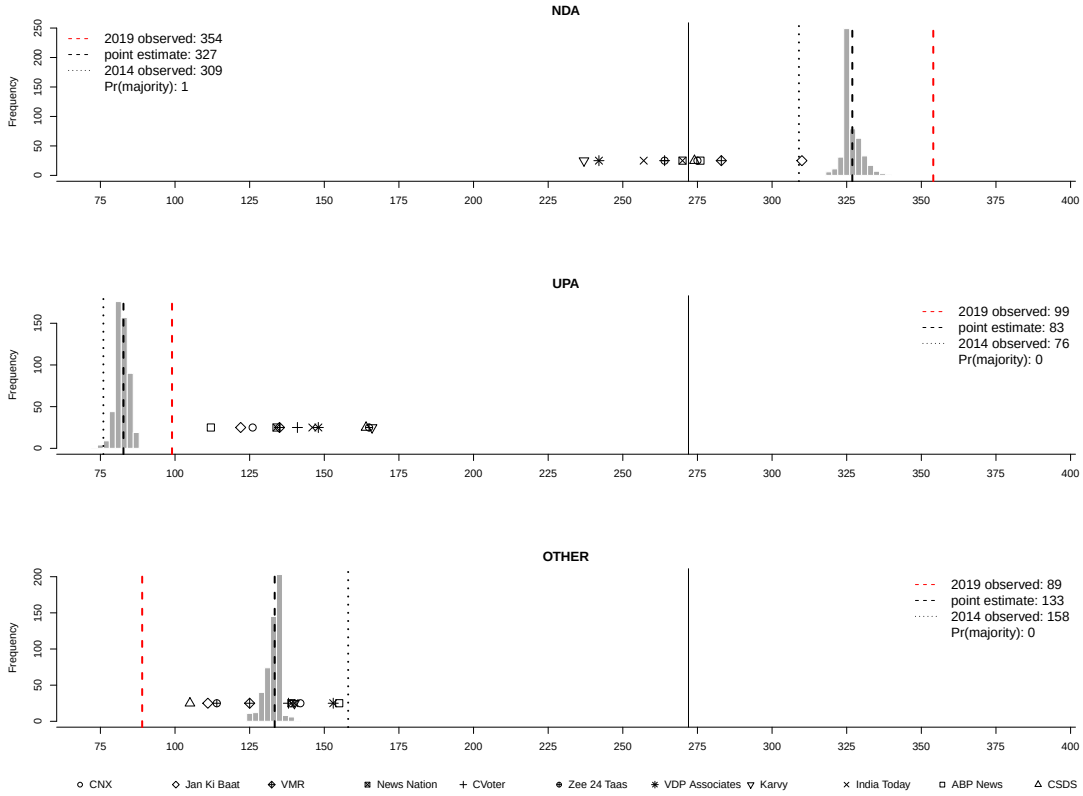


Figure 4: Seats projections at 5 weeks till the end of voting.

6 Discussion

This paper presents a fully integrated and exhaustive methodology to poll the Indian electorate. We note here that this methodology is absolutely general, in that every choice is guided by some general principles which could be deployed in India just as well as they could in Indonesia, the UK, Italy or Taiwan. We summarize the general principles which guide our choices below:

a) with respect to *study design*: i) prepare the study to be as 'deep' (in terms of cells) as your computational hardware and your stratification frame precision allows; ii) incorporate variable selection in the prediction models (Goplerud et al. 2018, Bisbee 2019, Ornstein 2019); iii) tune parameters according to general principles to do with the nature of the problem, in our case probability estimation (Malley et al. 2012, Olson & Wyner 2018) as opposed to regression or classification, rather than the specific application (India v. UK v. US);

b) with respect to the building of a *stratification frame*: i) pick a backbone dataset, for which you can put a high degree of trust on the accuracy of the cell-counts (Lauderdale et al. 2020); for this 'backbone' data, privilege a probability sample over recent one; ii) decide on a set of auxiliary datasets including 'key variables' (typically political covariates, not available in census data) you want to augment the backbone by, being mindful that this will require modeling, and hence carries the risk of introducing noise. If anywhere in the framework, here is where parsimony guaranteed by a 'human in the loop' can be of value. Here we privilege recent samples over probability samples: cross-tabs from these will define the cross-tabs in our stratification-frame, and political whims can change drastically over joint voter characteristics in a matter of weeks; iii) integrate the auxiliary data with the backbone via models which enable for the generation of credible joint distributions - which are key in closely-fought elections (India models that do not include the interaction between 'Caste' and 'Past Vote', for instance, would be scarcely informative). Note that the insistence on correctly estimating joint distributions has not been considered strictly necessary in previous literature on this topic (Leemann & Wasserfallen 2017, Ornstein 2017); nevertheless, in light of recent developments in the literature dealing with the importance of deep interactions in accurate pre-

dictions from regression and post-stratification models (Lauderdale & Blumenau 2019, Lauderdale 2019) we have come to believe these cross-tabs are essential to the success of regression and post-stratification approaches; iv) rake to the most recent, informative, known area-level voting intentions (Lauderdale et al. 2020) (typically these will be the results from the most recent election before the one under study);

c) with respect to *data collection*: i) generate a survey instrument only after having designed the stratification frame (to avoid inconsistencies between predicted categories); ii) collect samples from as many *different* modes as resources allow (Duch et al. 2020) - the more different the sources (in terms of the residual explanation for why a given individual, after accounting for controls, is present in a given source) - the more likely the averaging of information from these modes will result in better predictions, as mode-specific errors will tend to cancel out to the extent that they are uncorrelated (Graefe et al. 2014). Note further that simple averages of predictions derived from different sources will tend to outperform averages which include complex modeled weights (Graefe et al. 2015).

On the building of stratification frames specifically, we note that there is a tension between precision and depth of frames. We take this into consideration throughout Section 3.1, as our data integration effort is a delicate balance between introducing new variables such as past-vote and past Turnout, integrating the datasets these variables are derived from (e.g. the Election Study with IHDS with Census), and doing so in a conservative way to maximise the chances that our stratification frame is precise where it matters – which for our purposes is ultimately at the national level, and partially at the state-level. We took a number of practical steps to ensure the costs of this trade-off were minimized:

1. Steps taken to maximise depth: i) Integration of IHDS with Census cross-tabs to improve area-level representation; ii) Integration of Election Study and IHDS to include political variables in stratification frame;

2. Steps taken to maximise precision: i) Using Random Forests for imputation to introduce heavy smoothing via bagging and minimize noise; ii) Raking integrated

frame to past election results; iii) Raking the Election Study to observed 2014 turnout before integration with other datasets.

Substantively, this paper has made a number of contributions, outlined and summarized in what follows. Areas for further research are highlighted for each of the literatures this paper contributes to.

To the literature on opinion polling in limiting settings - of which India is an important example - where it is challenging to obtain representative samples at reasonable costs, we have shown that reasonably powerful samples can be obtained cheaply using novel technologies, namely social media and crowd-sourcing platforms. We have further shown how these samples are biased, and described source-specific selection: Facebook users tend to be more upper caste, richer, and younger; Mechanical Turks are more diverse in terms of caste, more middle-class, overwhelmingly working age, and overwhelmingly from Tamil Nadu and Kerala - in what remains a mystery of the literature on Indian Mechanical Turks; both groups are much more highly educated and richer than the vast majority of the Indian polity. Future work in this area should aim at further investigating these selection biases, comparing these with traditional samples and novel social-media.

To the literature on stratification frames, we have proposed a novel methodology to obtain deep and reliable estimates of the joint-distribution for the polity of interest, in setting where: a) random micro-data samples of the census are not available; b) public census tables are shallow and outdated; c) political variables are not included in the available joint distribution; d) nationally representative surveys are available, but they are heavily biased at the area level. Our methodology provides robust weights for deep cells, by integrating multiple sources of population data. Future work should investigate sustainable ways to account for uncertainty in the estimated weights, which are likely to contribute heavily to the final uncertainty in voting-behaviour predictions. It is worth noting that one of the reasons why our prediction intervals end up being too narrow is due to the simple statistical fact that our predictions are weighted averages across thousands of cells; by the simple logic of the central limit theorem, averaging over hundreds of cells leads to shrinking variance in the order of hundreds of times. Hence incorporating realistic uncertainty around the weights should relax the shrinkage effect. Another area of further investigation should be how to continuously update and upkeep

long series of cell-weights. The MrP literature has not yet dealt with the problem of over-time robustness of the weights, largely due to it being overwhelmingly US and UK centric, and hence having been able to leverage frequently updated official estimates.

To the literature on probability machines, we propose a beta-binomial model which enables accounting of prediction and sampling uncertainty via Monte Carlo simulations, leveraging assumptions about the predicted probabilities and their variance which are only mildly approximate. We further build on literature about the hyper-parameters of probability machines, and show that large values of `mtry` produce high-quality predictions. Future work on probability machines should focus on developing a sampling theory for probability outputs. It is worth spending a few lines on the nature of the probability output we have predicted, and relate it to more traditional regression settings. In typical logistic regression, we estimate conditional probabilities by assuming the existence of a latent dimension generating behaviour, which we cannot observe, but on which we can make inference under some standard assumptions about the nature of the latent-noise. Aggregation at the cell-level then entails a weighted average over the inverse logit of this latent quantity - which will range in $(0, 1)$. This approach is qualitatively and quantitatively different from that employed by the RF, or probability-machines in general: a predictive model of individual behaviour - typically a dichotomous variable - is directly learned; individuals are clustered in groups, and the average of the behaviours within each group is used to make predictions. Herein lies the fundamental difference: RFs model behaviour, whilst logistic regression models underlying propensities. At a high-enough level of aggregation, these will be indistinguishable; nevertheless, this difference has important implications for the theory of model uncertainty around the probabilities predicted in our paper. We have attempted to provide an approximate solution, which required correcting a small proportions of our cell-level predictions to conform to a Beta support. We hope to explore a more formal connection with traditional probability sampling theory in future work.

Ultimately, our work provides evidence in favour of the hypothesis that, given appropriate modeling, cheap and unrepresentative samples available through novel and unconventional means can be used for the large-scale study of opinion worldwide. As the coverage of social-media samples expands along with technological and internet ac-

cess across the developing world, we can expect these samples to become increasingly valuable to researchers and the global public as a whole. This technological expansion, coupled with modern challenges to traditional polling practices such as lowering response rates, promises to change the polling industry - we believe for the better. The coming preponderance of unrepresentative samples also means heavier reliance on modeling; as such, it is important to lay out a principled 'practitioners' guide' for the many modeling choices pollsters have to make in order to extract the desired signal. This paper contributes to this goal by providing some initial insights for the digital-age pollster.

References

Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. (2010), 'Permutation importance: a corrected feature importance measure', *Bioinformatics* **26**(10), 1340–1347.

Antin, J. & Shaw, A. (2012), Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the us and india, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', pp. 2925–2934.

Athey, S., Tibshirani, J., Wager, S. et al. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178.

Bache, K. & Lichman, M. (2013), 'UCI machine learning repository', `http://archive.ics.uci.edu/ml`.

Bisbee, J. (2019), 'BARP: Improving Mister P using Bayesian Additive Regression Trees', *American Political Science Review* **113**(4), 1060–1065.

Boas, T. C., Christenson, D. P. & Glick, D. M. (2020), 'Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics', *Political Science Research and Methods* **8**(2), 232–250.

Breiman, L. (2001*a*), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L. (2001*b*), 'Using iterated bagging to debias regressions', *Machine Learning* **45**(3), 261–277.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Butler, D., Lahiri, A. & Roy, P. (1995), *India Decides: Elections 1952-1995*, Books & Things.

Butler, D. & Stoke, D. (1974), *Political change in Britain: Basis of electoral choice*, Springer.

Buttice, M. K. & Highton, B. (2013), 'How does multilevel regression and poststratification perform with conventional national surveys?', *Political analysis* **21**(4).

Carrig, M. M., Manrique-Vallier, D., Ranby, K. W., Reiter, J. P. & Hoyle, R. H. (2015), 'A nonparametric, multiple imputation-based method for the retrospective integration of data sets', *Multivariate behavioral research* **50**(4), 383–397.

Castle, J. L. & Hendry, D. F. (2010), 'Nowcasting from disaggregates in the face of location shifts', *Journal of Forecasting* **29**(1-2), 200–214.

Cerina, R., Baio, G., Duch, R. & Fisher, S. (2020), Learning Voting Behaviour from Syntheric Surveys, *in* 'MrP Conference', Departments of Statistics and Political Science and Institute for Social and Economic Research and Policy at Columbia University.

Cerina, R. & Duch, R. (2020), 'Measuring public opinion via digital footprints', *International Journal of Forecasting* .

Chakrabarti, R. & Roy, J. (2007), 'Effect of redrawing of political boundaries on voting patterns: evidence from state reorganization in india', *Social Science Research Network (SSRN): 1008094* .

Charnysh, V., Lucas, C. & Singh, P. (2015), 'The ties that bind: National identity salience and pro-social behavior toward the ethnic other', *Comparative Political Studies* **48**(3), 267–300.

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics* **11**(4), 427–444.

Desai, S., Vanneman, R. & National Council of Applied Economic Research (2015-07-31), 'India Human Development Survey-II (IHDS-II) 2011-12', *ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research* .

Dietrich, S. & Winters, M. S. (2015), 'Foreign aid and government legitimacy', *Journal of Experimental Political Science* **2**(2), 164–171.

Duch, R., Laroze, D., Robinson, T. & Beramendi, P. (2020), 'Multi-modes for detecting experimental measurement error', *Political Analysis* **28**(2), 263–283.

Eldersveld, S. J. & Ahmed, B. (1978), *Citizens and politics: Mass political behavior in India*, University of Chicago Press.

Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014), 'Do we need hundreds of classifiers to solve real world classification problems?', *The journal of machine learning research* **15**(1), 3133–3181.

Fisher, S. D. (2015), 'Predictable and unpredictable changes in party support: a method for long-range daily election forecasting from opinion polls', *Journal of Elections, Public Opinion & Parties* **25**(2), 137–158.

Fisher, S. D., Ford, R., Jennings, W., Pickup, M. & Wlezien, C. (2011), 'From polls to votes to seats: Forecasting the 2010 british general election', *Electoral Studies* **30**(2), 250–257.

Ford, R., Jennings, W., Pickup, M. & Wlezien, C. (2016), 'From polls to votes to seats: Forecasting the 2015 british general election', *Electoral Studies* **41**, 244–249.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H. & Herrera, F. (2011), 'An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes', *Pattern Recognition* **44**(8), 1761–1776.

Gelman, A. (2018), 'Regularized prediction and poststratification (the generalization of mister p)', `https://statmodeling.stat.columbia.edu/2018/05/19/regularized-predictionpoststratification-generalization-mister-p`.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Gelman, A. & Little, T. C. (1997), 'Poststratification into many categories using hierarchical logistic regression'.

Goel, S., Obeng, A. & Rothschild, D. (2015), 'Non-representative surveys: Fast, cheap, and mostly accurate', *Working Paper* .

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep learning*, MIT press.

Goplerud, M. (2020), 'Fast and accurate estimation of non-nested binomial hierarchical models using variational inference', *arXiv:2007.12300* .

Goplerud, M., Kuriwaki, S., Ratkovic, M. & Tingley, D. (2018), 'Sparse Multilevel Regression (and Poststratification (sMRP))', *Working Paper* .

Graefe, A., Armstrong, J. S., Jones Jr, R. J. & Cuzán, A. G. (2014), 'Combining forecasts: An application to elections', *International Journal of Forecasting* **30**(1), 43–54.

Graefe, A., Küchenhoff, H., Stierle, V. & Riedl, B. (2015), 'Limitations of ensemble bayesian model averaging for forecasting social science problems', *International Journal of Forecasting* **31**(3), 943–951.

Hanretty, C. (2019*a*), '2019 General Election MRP predictions', `https://www.survation.com/2019-general-election-mrp-predictions-survation-and-dr-chris-hanretty/`.

Hanretty, C. (2019*b*), 'An introduction to multilevel regression and post-stratification for estimating constituency opinion', *Political Studies Review* **18**(4), 630–645.

Hanretty, C., Lauderdale, B. E. & Vivyan, N. (2018), 'Comparing strategies for estimating constituency opinion from national survey samples', *Political Science Research and Methods* **6**(3), 571–591.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Heath, A., Fisher, S. & Smith, S. (2005), 'The globalization of public opinion research', *Annu. Rev. Polit. Sci.* **8**, 297–333.

Heath, O. (2015), 'The BJP's return to power: mobilisation, conversion and vote swing in the 2014 Indian elections', *Contemporary South Asia* **23**(2), 123–135.

Hendry, D. F. & Hubrich, K. (2011), 'Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate', *Journal of business & economic statistics* **29**(2), 216–227.

Hubrich, K. (2005), 'Forecasting euro area inflation: Does aggregating forecasts by hicp component improve forecast accuracy?', *International Journal of Forecasting* **21**(1), 119–136.

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

Jackman, S., Ratcliff, S. & Mansillo, L. (2019), 'Small area estimates of public opinion: model-assisted post-stratification of data from voter advice applications', *Working Paper* .

Janitza, S., Celik, E. & Boulesteix, A.-L. (2018), 'A computationally fast variable importance test for random forests for high-dimensional data', *Advances in Data Analysis and Classification* **12**(4), 885–915.

Jennings, W. & Wlezien, C. (2018), 'Election polling errors across time and space', *Nature Human Behaviour* **2**(4), 276–283.

Jensenius, F. R. & Verniers, G. (2017), 'Studying indian politics with large-scale data: Indian election data 1961–today', *Studies in Indian Politics* **5**(2), 269–275.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Ijcai', Vol. 14, Montreal, Canada, pp. 1137–1145.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2017), 'Automatic differentiation variational inference', *The Journal of Machine Learning Research* **18**(1), 430–474.

Lahiri, A. K. & Roy, P. (1984), 'Assessing swings in multi-party systems: The Indian experience', *Electoral Studies* **3**(2), 171–189.

Lauderdale, B. E. (2019), 'How YouGov's 2019 General Election model works', `https://yougov.co.uk/topics/politics/articles-reports/2019/11/27/how-yougovs-2019-general-election-model-works`.

Lauderdale, B. E., Bailey, D., Blumenau, J. & Rivers, D. (2020), 'Model-based pre-election polling for national and sub-national outcomes in the US and UK', *International Journal of Forecasting* **36**(2), 399–413.

Lauderdale, B. E. & Blumenau, J. (2019), Constructing and assessing seat level estimates. Reading the 2019 Election Polls: Event by the London School of Economics, Department of Methodology, and the British Polling Council.

Lax, J. R. & Phillips, J. H. (2009), 'Gay rights in the states: Public opinion and policy responsiveness', *American Political Science Review* **103**(3), 367–386.

Leemann, L., Broniecki, P. & Wuest, R. (2020), 'Improved multilevel regression with post-stratification through machine learning (autoMrP)', *Working Paper* .

Leemann, L. & Wasserfallen, F. (2017), 'Extending the use and prediction precision of subnational public opinion estimation', *American journal of political science* **61**(4), 1003–1022.

Lu, B. & Hardin, J. (2019), 'A unified framework for random forest prediction error estimation', *arXiv:1912.07435* .

Lütkepohl, H. (2011), 'Forecasting nonlinear aggregates and aggregates with time-varying weights', *Jahrbücher für Nationalökonomie und Statistik* **231**(1), 107–133.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. & Ziegler, A. (2012), 'Probability machines', *Methods of information in medicine* **51**(01), 74–81.

Marcellino, M., Stock, J. H. & Watson, M. W. (2003), 'Macroeconomic forecasting in the euro area: Country specific versus area-wide information', *European Economic Review* **47**(1), 1–18.

Mayer, M. (2018), 'Using `missRanger`', `https://cran.r-project.org/web/packages/missRanger/vignettes/vignette_missRanger.html`.

Mentch, L. & Hooker, G. (2016), 'Quantifying uncertainty in random forests via confidence intervals and hypothesis tests', *The Journal of Machine Learning Research* **17**(1), 841–881.

Norman, P. (1999), 'Putting iterative proportional fitting on the researcher's desk'.

Olson, M. A. & Wyner, A. J. (2018), 'Making sense of random forest probabilities: a kernel perspective', *arXiv:1812.05792* .

Ornstein, J. T. (2017), 'Subnational Public Opinion Estimation Using MrsP', *Working Paper* .

Ornstein, J. T. (2019), 'Stacked regression and poststratification', *Political Analysis* pp. 1–9.

Park, D. K., Gelman, A. & Bafumi, J. (2004), 'Bayesian multilevel estimation with post-stratification: State-level estimates from national polls', *Political Analysis* **12**(4), 375–385.

Pasek, J. & Pasek, M. J. (2019), 'Package `anesrake`'.

Rai, P. (2014), 'Fallibility of opinion polls in India', *Economic and Political Weekly* pp. 13–17.

Rifkin, R. & Klautau, A. (2004), 'In defense of one-vs-all classification', *Journal of machine learning research* **5**(Jan), 101–141.

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, Vol. 81, John Wiley & Sons.

Salganik, M. J. (2019), *Bit by bit: Social research in the digital age*, Princeton University Press.

Solomon, S. (2016), 'Polling one billion: measuring public opinion in the world's largest democracy', *LSE South Asia Centre* .

Starkweather, J. (2014), 'A new recommended way of dealing with multiple missing values: Using `missForest` for all your imputation needs', *Benchmarks RSS Matters July* .

Stekhoven, D. J. & Bühlmann, P. (2012), '`missForest` — non-parametric missing value imputation for mixed-type data', *Bioinformatics* **28**(1), 112–118.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008), 'Conditional variable importance for random forests', *BMC bioinformatics* **9**(1), 307.

Team, L. (2014), 'National Election Study 2014', `https://www.lokniti.org/national-election-studies`.

Team, R. C. et al. (2013), 'R: A language and environment for statistical computing'.

Van Buuren, S. & Oudshoorn, K. (1999), *Flexible multivariate imputation by `mice`*, Leiden: TNO.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wager, S., Hastie, T. & Efron, B. (2014), 'Confidence intervals for random forests: The jackknife and the infinitesimal jackknife', *The Journal of Machine Learning Research* **15**(1), 1625–1651.

Wainberg, M., Alipanahi, B. & Frey, B. J. (2016), 'Are random forests truly the best classifiers?', *The Journal of Machine Learning Research* **17**(1), 3837–3841.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Warshaw, C. & Rodden, J. (2012), 'How should we measure district-level public opinion on individual issues?', *The Journal of Politics* **74**(1), 203–219.

Whiteley, P., Sanders, D., Stewart, M. & Clarke, H. (2011), 'Aggregate level forecasting of the 2010 general election in britain: The seats-votes model', *Electoral Studies* **30**(2), 278–283.

Wright, M. N. & Ziegler, A. (2015), '`ranger`: A fast implementation of random forests for high dimensional data in C++ and R', *arXiv:1508.04409* .

Zellner, A. & Tobias, J. (2000), 'A note on aggregation, disaggregation and forecasting performance', *Journal of Forecasting* **19**(5), 457–457.

Appendix

Party Allocation to Alliance

Table A1 displays our allocation of parties to alliances; this was constructed based on the results found on `Wikipedia`, namely pages `https://en.wikipedia.org/wiki/List_of_National_Democratic_Alliance_members` and `https://en.wikipedia.org/wiki/List_of_United_Progressive_Alliance_members`, as well as unfolding news stories during the campaign. Parties not mentioned in the table were allocated to the 'Other' category. Note that there are several other parties which belong to these alliances, many of which have been created uniquely for the 2019 election. Where these residual parties have no chance of being competitive in any district, they are ignored, as they will not impact our analysis.

Party	Alliance
ajsu party	NDA
all india anna dravida munnetra kazhagam	NDA
all india n.r. congress	NDA
apna dal (soneylal)	NDA
asom gana parishad	NDA
bharath dharma jana sena	NDA
bharatiya janata party	NDA
bodoland peoples front	NDA
desiya murpokku dravida kazhagam	NDA
indigenous people's front of tripura	NDA
janata dal (united)	NDA
lok jan shakti party	NDA
manipur people's party	NDA
mizo national front	NDA
naga peoples front	NDA
nationalist democratic progressive party	NDA
national people's party	NDA
pattali makkal katchi	NDA
rashtriya loktantrik party	NDA
rashtriya samaj paksha	NDA
republican party of india (a)	NDA
shiromani akali dal	NDA
shivsena	NDA
sikkim democratic front	NDA
tamil maanila congress (moopanar)	NDA
united democratic party	NDA
bahujan vikas aaghadi	UPA
communist party of india	UPA
communist party of india (marxist)	UPA
communist party of india (marxist-leninist) (liberation)	UPA
dravida munnetra kazhagam	UPA
hindustani awam morcha (secular)	UPA
indian national congress	UPA
indian union muslim league	UPA
jammu & kashmir national conference	UPA
jan adhikar party	UPA
janata dal (secular)	UPA
jharkhand mukti morcha	UPA
jharkhand vikas morcha (prajatantrik)	UPA
karnataka pragnyavantha janatha party	UPA
kerala congress (m)	UPA
nationalist congress party	UPA
peace party	UPA
rashtriya janata dal	UPA
rashtriya lok dal	UPA
rashtriya lok samta party	UPA
revolutionary socialist party	UPA
swabhimani paksha	UPA
telangana jana samithi	UPA
telugu desam	UPA
viduthalai chiruthaigal katchi	UPA
vikassheel insaan party	UPA

Table A1: Our allocation of parties to alliances.

Micro-Data for Integration

Source	Source_ID	States	Gender	Family_Role	Age_cat	Marital_Status	Rurality	Religion	Literacy	Education_level	Income_level	Jati	Zones	2014 Turnout	2014 Alliance_Vote
IHDS	1	(01) Jammu & Kashmir 01	(1) Male 1	(01) Head 1	(54-64]	(1) Married 1	(0) rural 0	(2) Muslim 2	(0) No 0	(01) No Formal Edu 01	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(01) North 01	NA	NA
IHDS	2	(01) Jammu & Kashmir 01	(2) Female 2	(02) Wife/Husband 2	(44-54]	(1) Married 1	(0) rural 0	(2) Muslim 2	(0) No 0	(01) No Formal Edu 01	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(01) North 01	NA	NA
IHDS	3	(01) Jammu & Kashmir 01	(1) Male 1	(03) Son/Daughter 3	(17-24]	(2) Unmarried 2	(0) rural 0	(2) Muslim 2	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(01) North 01	NA	NA
IHDS	4	(01) Jammu & Kashmir 01	(1) Male 1	(03) Son/Daughter 3	(24-34]	(1) Married 1	(0) rural 0	(2) Muslim 2	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(01) North 01	NA	NA
IHDS	5	(01) Jammu & Kashmir 01	(2) Female 2	(04) Child-in-Law 4	(24-34]	(1) Married 1	(0) rural 0	(2) Muslim 2	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(01) North 01	NA	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	.	⋮	⋮	⋮	⋮	⋮
IHDS	135980	(34) Pondicherry 34	(2) Female 2	(02) Wife/Husband 2	(44-54]	(1) Married 1	(1) urban 1	(1) Hindu 1	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(06) Southern 06	NA	NA
IHDS	135981	(34) Pondicherry 34	(1) Male 1	(03) Son/Daughter 3	(24-34]	(2) Unmarried 2	(1) urban 1	(1) Hindu 1	(1) Yes 1	(05) Some Graduate or Higher 05	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(06) Southern 06	NA	NA
IHDS	135982	(34) Pondicherry 34	(1) Male 1	(03) Son/Daughter 3	(24-34]	(2) Unmarried 2	(1) urban 1	(1) Hindu 1	(1) Yes 1	(05) Some Graduate or Higher 05	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(06) Southern 06	NA	NA
IHDS	135983	(34) Pondicherry 34	(1) Male 1	(01) Head 1	(44-54]	(1) Married 1	(1) urban 1	(1) Hindu 1	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(06) Southern 06	NA	NA
IHDS	135984	(34) Pondicherry 34	(2) Female 2	(02) Wife/Husband 2	(34-44]	(1) Married 1	(1) urban 1	(1) Hindu 1	(1) Yes 1	(03) Middle or Secondary 03	(03) R 90,000 - R 1,19,999 03	(3) Other Backward Castes (OBC) 3	(06) Southern 06	NA	NA
Census	1	(28) Andhra Pradesh 28	(1) Male 1	NA	(44-54]	NA	(0) rural 0	NA	NA	(01) No Formal Edu 01	NA	NA	(06) Southern 06	NA	NA
Census	2	(29) Karnataka 29	(1) Male 1	NA	(64-100]	NA	(0) rural 0	NA	NA	(01) No Formal Edu 01	NA	NA	(06) Southern 06	NA	NA
Census	3	(19) West Bengal 19	(1) Male 1	NA	(34-44]	NA	(0) rural 0	NA	NA	(01) No Formal Edu 01	NA	NA	(04) Eastern 04	NA	NA
Census	4	(08) Rajasthan 08	(1) Male 1	NA	(44-54]	NA	(1) urban 1	NA	NA	(05) Some Graduate or Higher 05	NA	NA	(01) North 01	NA	NA
Census	5	(33) Tamil Nadu 33	(1) Male 1	NA	(34-44]	NA	(0) rural 0	NA	NA	(04) Higher Secondary 04	NA	NA	(06) Southern 06	NA	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	.	⋮	⋮	⋮	⋮	⋮
Census	3800209	(19) West Bengal 19	(2) Female 2	NA	(54-64]	NA	(1) urban 1	NA	NA	(03) Middle or Secondary 03	NA	NA	(04) Eastern 04	NA	NA
Census	3800210	(06) Haryana 06	(2) Female 2	NA	(54-64]	NA	(1) urban 1	NA	NA	(02) Primary or Lower 02	NA	NA	(01) North 01	NA	NA
Census	3800211	(01) Jammu & Kashmir 01	(1) Male 1	NA	(17-24]	NA	(0) rural 0	NA	NA	(02) Primary or Lower 02	NA	NA	(01) North 01	NA	NA
Census	3800212	(23) Madhya Pradesh 23	(1) Male 1	NA	(24-34]	NA	(1) urban 1	NA	NA	(02) Primary or Lower 02	NA	NA	(02) North-Central 02	NA	NA
Census	3800213	(20) Jharkhand 20	(1) Male 1	NA	(24-34]	NA	(1) urban 1	NA	NA	(03) Middle or Secondary 03	NA	NA	(04) Eastern 04	NA	NA
NES	1	(28) Andhra Pradesh 28	(1) Male 1	NA	(54-64]	NA	NA	(1) Hindu 1	NA	(02) Primary or Lower 02	(01) Less than R 60,000 01	(2) Forward/General (except Brahmin) 2	(06) Southern 06	1	NDA
NES	2	(28) Andhra Pradesh 28	(2) Female 2	NA	(44-54]	NA	NA	(1) Hindu 1	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(5) Scheduled Tribes (ST) 5	(06) Southern 06	1	OTHER
NES	3	(28) Andhra Pradesh 28	(2) Female 2	NA	(34-44]	NA	NA	(1) Hindu 1	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(2) Forward/General (except Brahmin) 2	(06) Southern 06	1	NDA
NES	4	(28) Andhra Pradesh 28	(2) Female 2	NA	(64-100]	NA	NA	(1) Hindu 1	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(2) Forward/General (except Brahmin) 2	(06) Southern 06	1	NDA
NES	5	(28) Andhra Pradesh 28	(1) Male 1	NA	(64-100]	NA	NA	(1) Hindu 1	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(2) Forward/General (except Brahmin) 2	(06) Southern 06	1	NDA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	.	⋮	⋮	⋮	⋮	⋮
NES	22291	(28) Andhra Pradesh 28	(2) Female 2	NA	(34-44]	NA	NA	(3) Christian 3	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(4) Scheduled Castes (SC) 4	(06) Southern 06	1	OTHER
NES	22292	(28) Andhra Pradesh 28	(2) Female 2	NA	(24-34]	NA	NA	(4) Sikh 4	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	NA	(06) Southern 06	1	OTHER
NES	22293	(28) Andhra Pradesh 28	(2) Female 2	NA	(34-44]	NA	NA	(4) Sikh 4	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(3) Other Backward Castes (OBC) 3	(06) Southern 06	1	UPA
NES	22294	(28) Andhra Pradesh 28	(1) Male 1	NA	(17-24]	NA	NA	(7) Other 7	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	NA	(06) Southern 06	1	NA
NES	22295	(28) Andhra Pradesh 28	(2) Female 2	NA	(17-24]	NA	NA	(7) Other 7	NA	(01) No Formal Edu 01	(01) Less than R 60,000 01	(3) Other Backward Castes (OBC) 3	(06) Southern 06	1	NDA

Table A2: Subset of the stacked dataset derived from appending the NES and the Census micro-data sample to the IHDS. Missing values are indicated as NA

Pre- and Post-Raking Stratification Frame

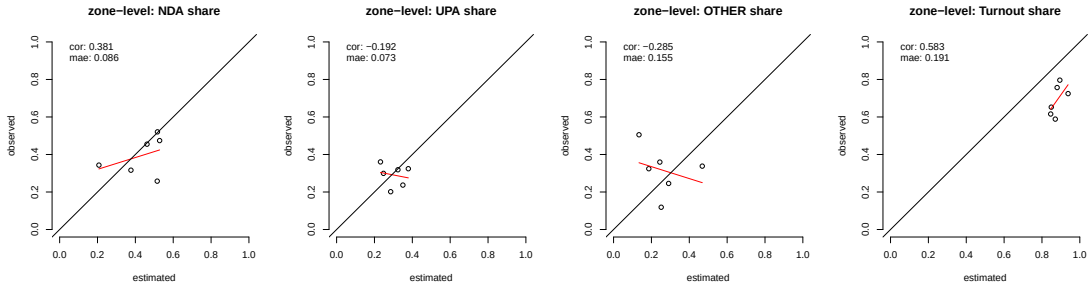


Figure A1: Post-imputation, pre-raking plot of voting distribution implied by the stratification frame at the Zone level.

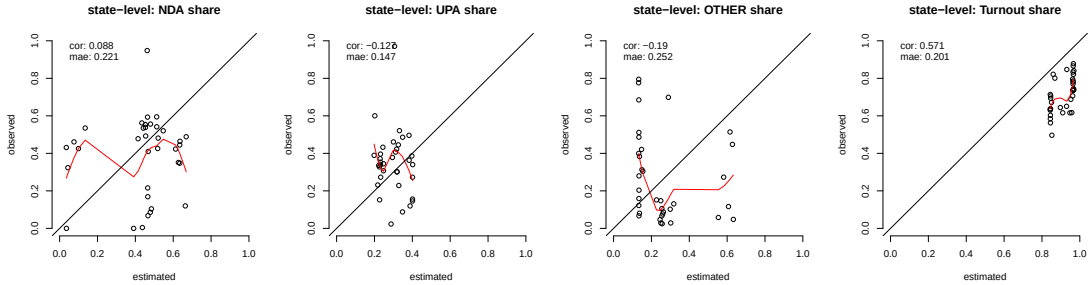


Figure A2: Post-imputation, pre-raking plot of voting distribution implied by the stratification frame at the State level.

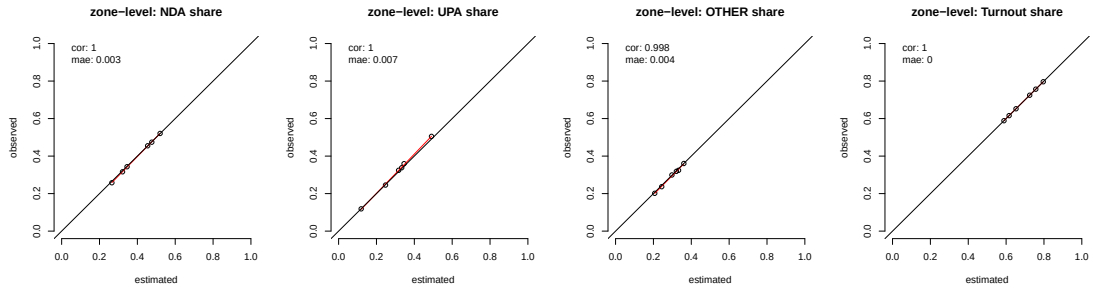


Figure A3: Post-imputation, post-raking plot of voting distribution implied by the stratification frame at the Zone level.

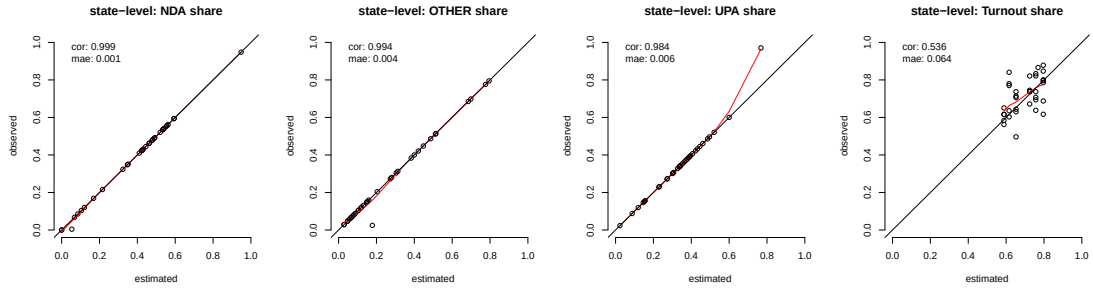


Figure A4: Post-imputation, post-raking plot of voting distribution implied by the stratification frame at the State level.

Convenience Samples

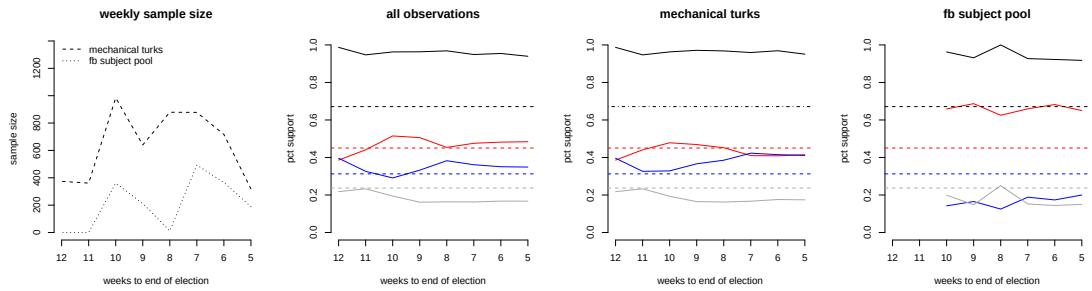


Figure A5: A graphical summary of the samples' evolution over the course of the collection period. In the top-left panel, the sample size over-time is disaggregated by mode; in the rightmost three panels, we see how raw voting behaviour looks on the aggregate, as well as by mode.

Variable	Convenience Sample	2014 NES
Religion	0.16	
H.Inc.	0.56	
Gender	0.02	
Education		0.29
Age_cat	0.34	0.58
Caste	0.26	0.43
2014 Vote Choice	0.17	0.33
2014 Turnout	0.29	0.05
2019 Vote Choice	0.18	
2019 Turnout	0.20	

Table A3: Out-of-bag error derived from the `missRanger` multiple imputation procedure, by variables and training sample. The OOB score ranges $[0, 1]$, indicating the proportion of miss-classified OOB observations.

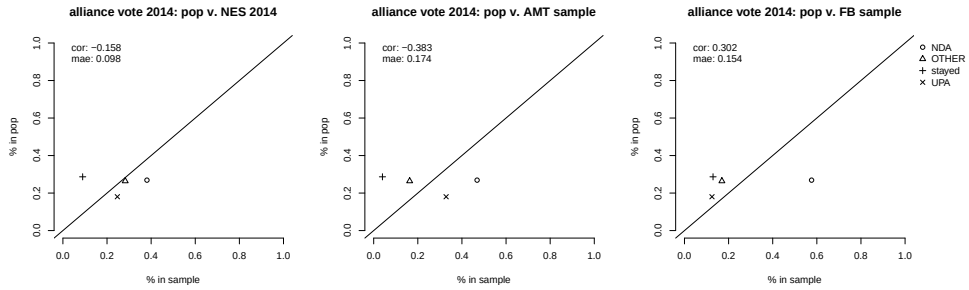


Figure A6: Micro-data samples compared with modeled population: 2014 Alliance Vote.

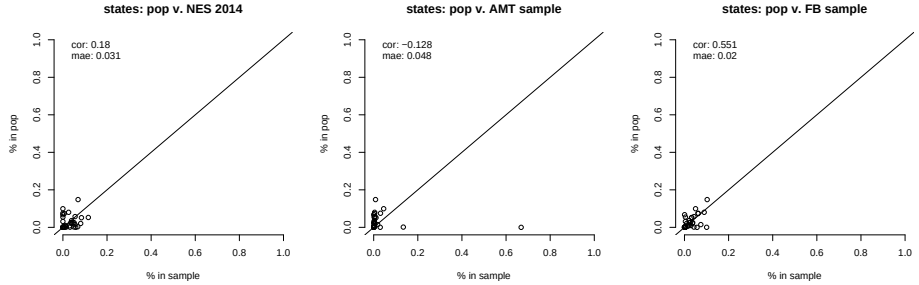


Figure A7: Micro-data samples compared with modeled population: States.

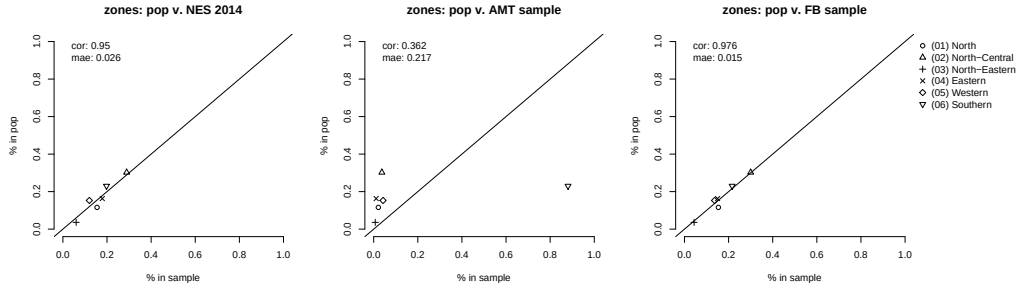


Figure A8: Micro-data samples compared with modeled population: Zones.

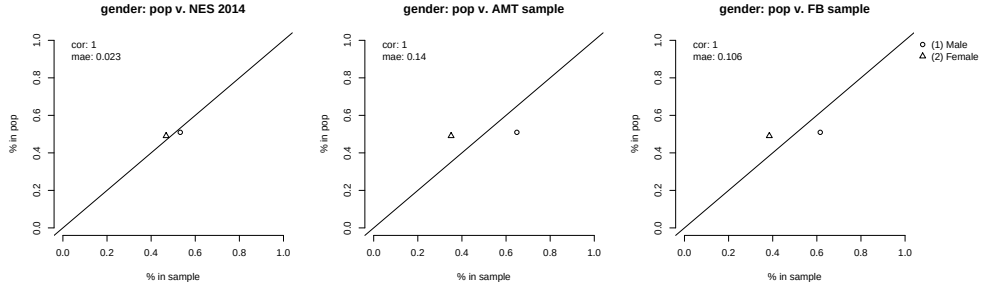


Figure A9: Micro-data samples compared with modeled population: Gender.

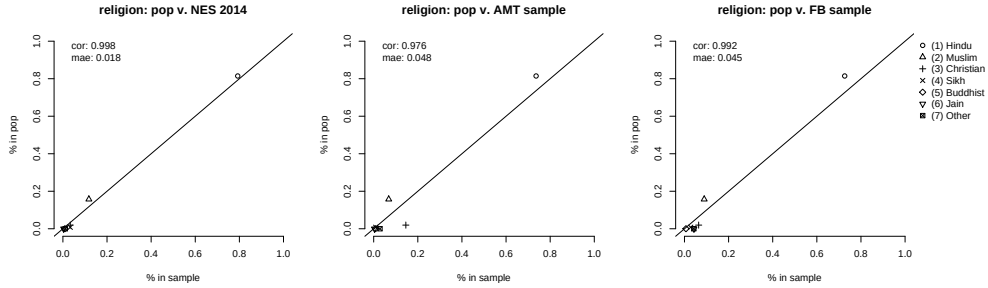


Figure A10: Micro-data samples compared with modeled population: Religion.

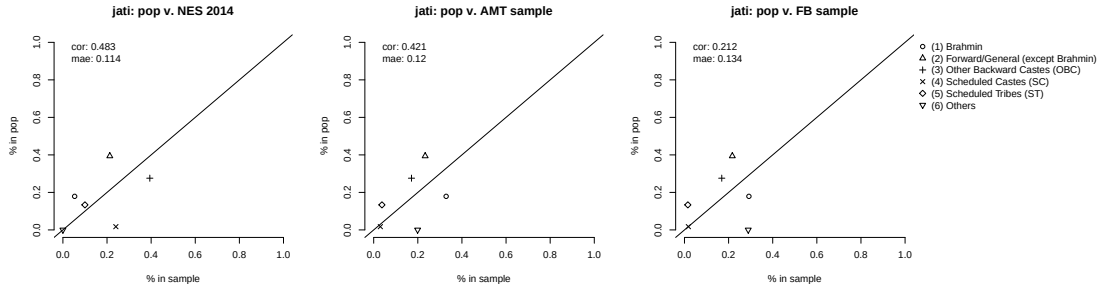


Figure A11: Micro-data samples compared with modeled population: Caste.

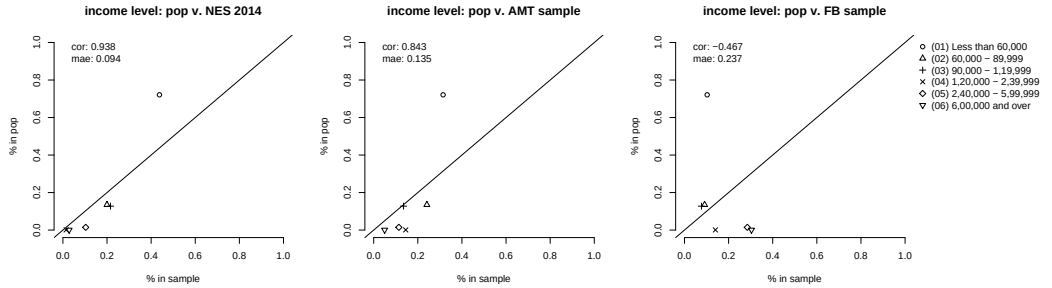


Figure A12: Micro-data samples compared with modeled population: Yearly Household Income.

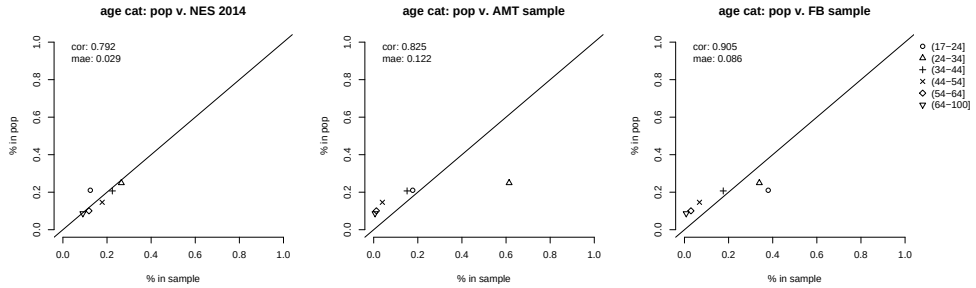


Figure A13: Micro-data samples compared with modeled population: Age Groups.

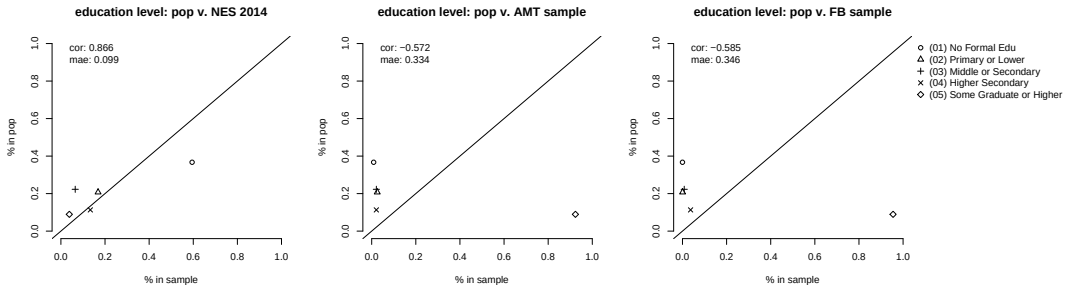


Figure A14: Micro-data samples compared with modeled population: Educational Attainment.

Modeling Behaviour

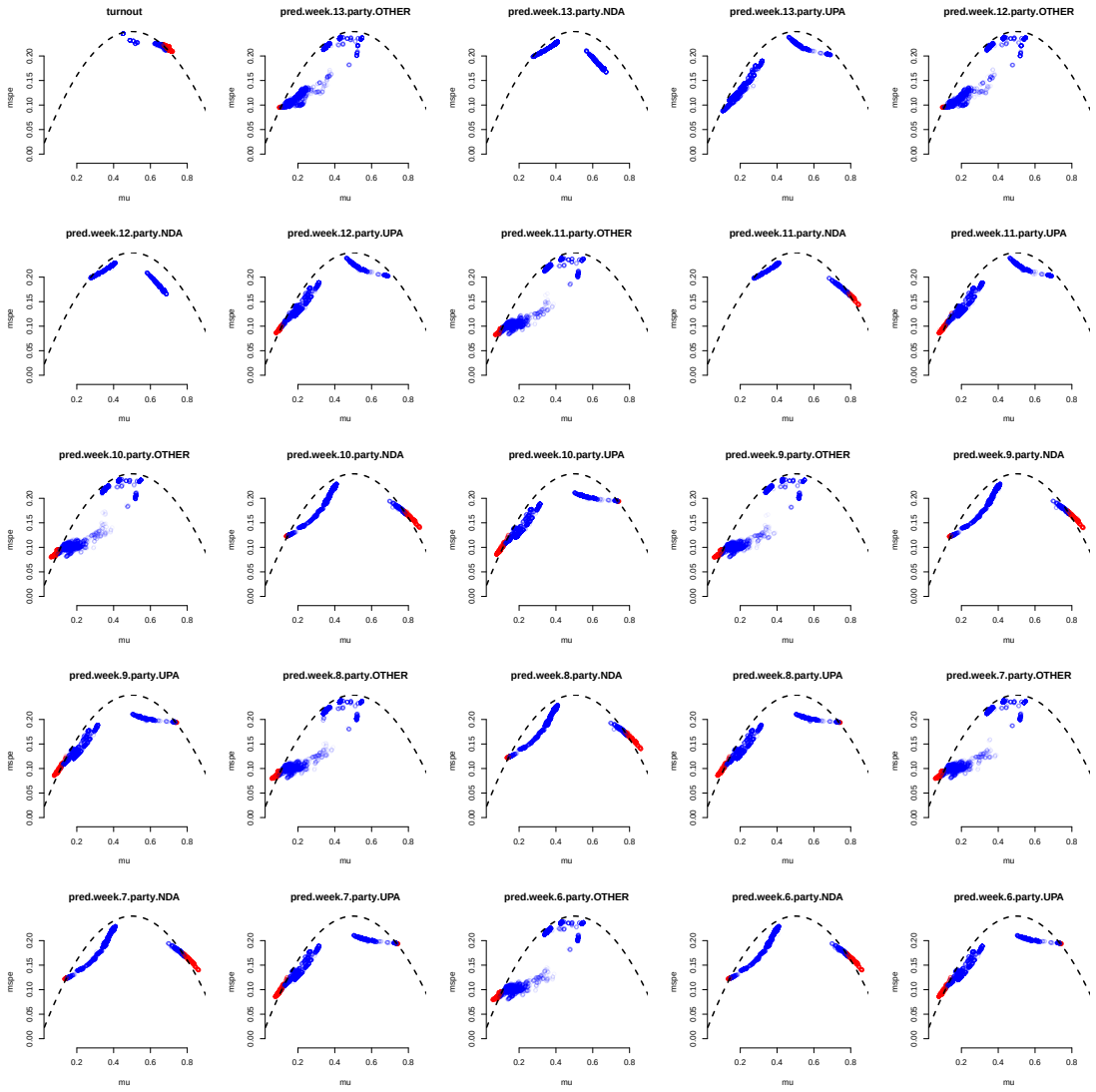


Figure A15: Predicted means and variances, against the variance limits required for conformity with the Beta distribution; red dots indicate non-conformity, and the $\mu(1-\mu)$ limit is shown by the dotted line.

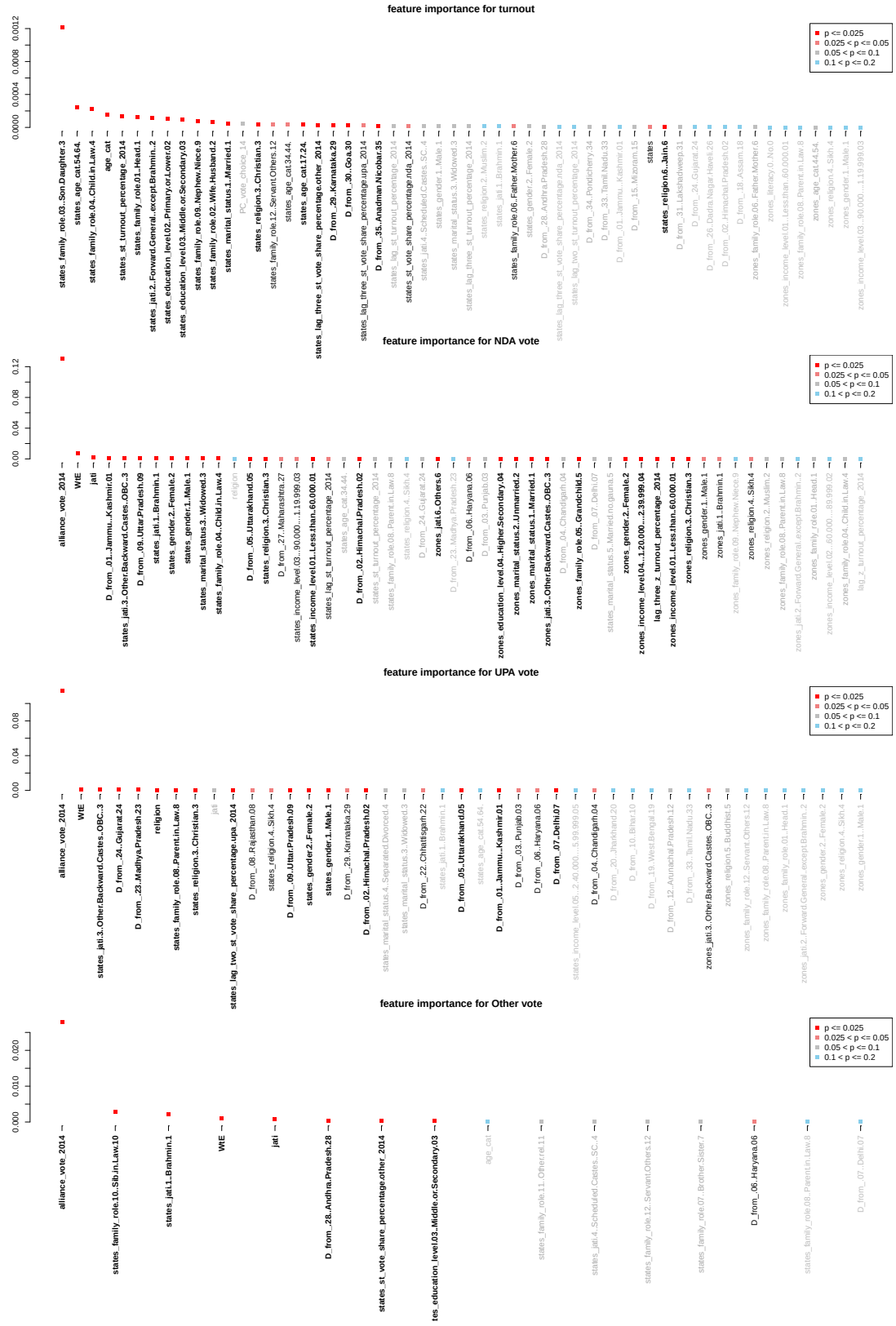


Figure A16: Variable Importance scores, and their significance. Color-coding indicates p-values, with red pointing to elevated significance, and vice-versa blue. Axis labels of highly significant variables are bolder, whilst low-significance variable labels are faded. State-level features are prefixed by 'states', whilst zone-level variables are prefixed by 'zone'; no prefix is attached to individual-level variables. Variable prefixed by 'D' are centroid distances from to given state.

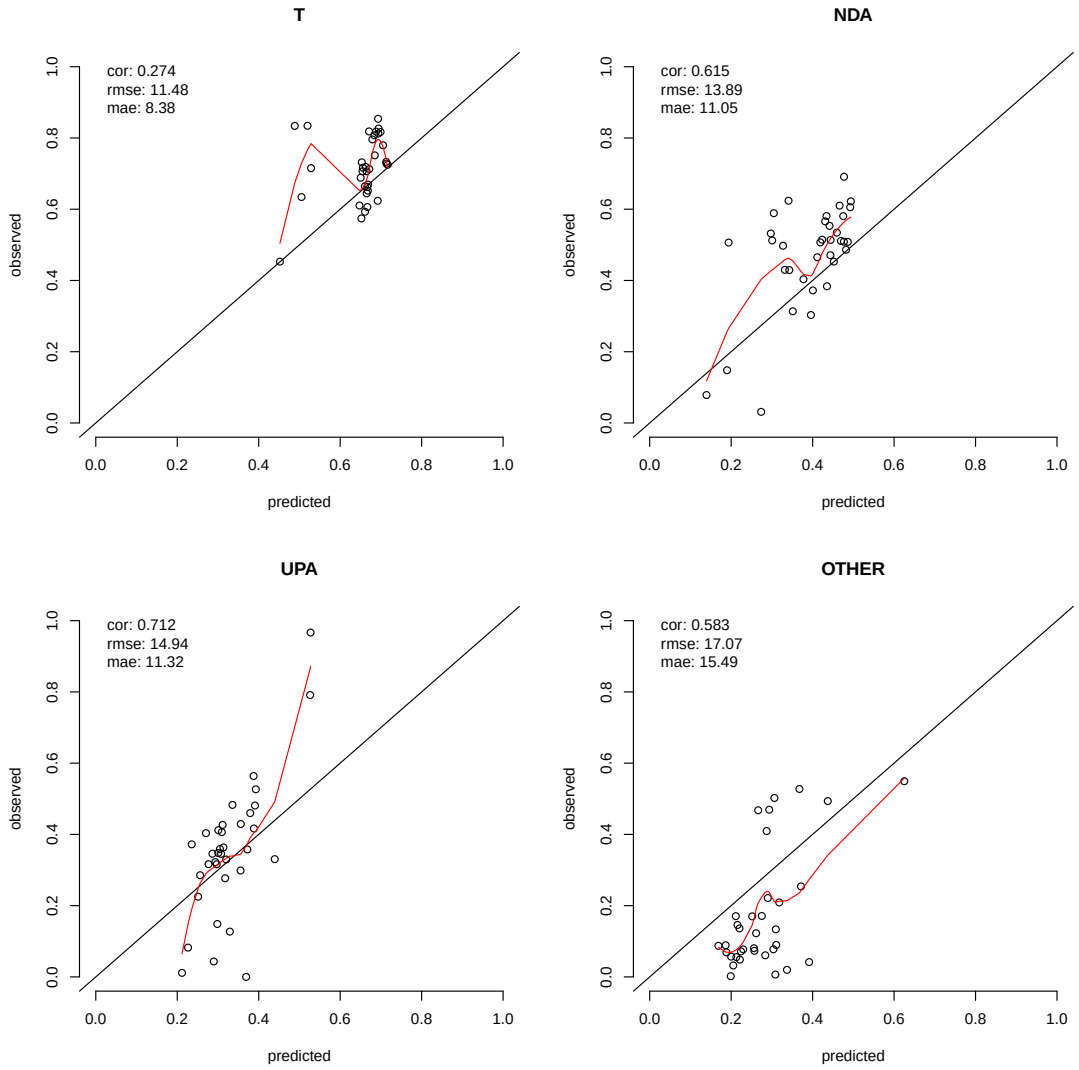


Figure A17: State-level predictions v. observed state-level 2019 behaviour.

Chapter IV:

Learning Voting Behaviour from Synthetic Surveys

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Gianluca Baio
g.baio@ucl.ac.uk
Department of Statistical Science, University College London

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Stephen Fisher
stephen.fisher@trinity.ox.ac.uk
Department of Sociology, University of Oxford

Abstract

This paper contributes to the literature on regression and post-stratification for the purpose of monitoring public opinion during an election campaign; we focus on the 2019 UK General Election, a case-study for which MrP predictions have been shown to be highly accurate. We produce three distinct contributions: i) we present 'Data Integration + MICE', a novel strategy to complete stratification frames and generate post-stratification targets for deeply interacted covariates which are typically not present in traditional stratification frames; ii) we are the first in this literature to apply the state-of-the-art `xgboost` learning algorithm, an extremely scalable and highly customizeable tree gradient booster, to generate predictions for cells in our stratification frame; we further develop an original tuning strategy applicable to the electoral context many advanced democracies, which leverages the well established *uniform swing* hypothesis as a tuning prior; iii) we present a strategy to update out-dated and non-representative micro-data with publicly available summaries of opinion polls published during an election campaign, and show these updates can massively improve the usefulness of publicly available, but typically out-dated, micro-data, for the purpose of up-to-date meso-level inference. Our contributions increase scalability of MrP applications and broaden the set of training data which can be used to make inference in the run-up to an election.

1 Introduction

Multilevel Regression and Post-stratification (MrP) (Gelman & Little 1997, Park et al. 2004) leverages micro, or individual level, data to produce accurate meso, or group level, inference. MrP requires three ingredients: 1) an individual-level dataset, generally augmented with aggregate-level covariates; 2) a learner of the kind of multilevel regression (Gelman & Hill 2006), or any other learner with desirable *regularizing* properties; 3) a stratification frame, storing the frequency of each *cell*, or categorical group of interest, in the population. Each of these ingredients comes with its own set of challenges: micro-level data can often be unrepresentative, outdated, and suffer from measurement error; learners can over-fit unrepresentative data or excessively reduce prediction variance, leading to attenuation-bias; stratification frames can be incomplete, missing key predictors for the application at hand. Table 1 summarizes how the current MrP literature deals with these challenges.

Provided a solution is given to each of the three challenges set out above, MrP has been shown to produce accurate area-level estimates in high-uncertainty applications. Most notably, MrP has captured headlines in the UK (Sabbagh 2019, Travis 2017), where large polling organizations such as YouGov (Lauderdale 2019) and Survation (Hanretty 2019) have successfully measured public opinion at the parliamentary constituency level in the run-up to the 2017 and 2019 General Elections[46].

This paper provides three distinct novel contributions to the MrP literature: i) we present 'Data Integration + MICE', **a novel strategy to complete stratification frames and generate post-stratification targets for deeply interacted covariates** which are typically not present in traditional stratification frames. We see this as a direct improvement on the existing 'MrsP' framework; ii) we are the first in this literature to apply the state-of-the-art `xgboost` (Chen & Guestrin 2016) learning algorithm, an extremely scalable and highly customizeable tree gradient booster, to generate predictions for cells in our stratification frame; we further develop **an original tuning strategy applicable to the electoral context of many advanced democracies,**

[46]Note that these commercial applications only produce estimates for the 632 parliamentary constituencies in Great Britain, excluding Northern Ireland due to fundamental differences in political dynamics.

which leverages the well established *uniform swing* model (Butler & Stoke 1974) as a tuning prior; iii) we present **a strategy to update out-dated and non-representative micro-data with publicly available summaries of opinion polls published during an election campaign**, and show these updates can massively improve the usefulness of publicly available, but typically out-dated, micro-data, for the purpose of up-to-date meso-level inference.

Beyond these three main innovations, two minor contributions to the literature further deserve to be highlighted: we are the first to spell-out, from start to finish, how to allocate parliamentary constituencies to census data at the Grouped Local Authority (GLA) level; we provide a 'data' contribution, by gathering the cross-tabs from 2019 election polls in a single dataset.

We measure the utility of our innovations in the UK electoral context, where commercial applications under optimal circumstances have been highly successful. The elements of commercial MrP applications which are difficult to reproduce in research environments are the following: i) *micro-data sample size.* The final YouGov 2019 MrP estimates leveraged $105,612$, high-quality individual-level observations, a portion of which ($14,838$ observations) came from a fully representative sample, whilst the remainder was sampled from the YouGov online panel. This kind of data gathering exercise could cost hundreds of thousands of £; ii) *time relevance.* YouGov's entire sample for their final MrP estimates was concentrated around the 7 days up to and including December 10^{th}, date in which the YouGov estimates were published - when the election was on December 12^{th}. Typical public micro-data for research will tend to be outdated, staggered over long time frames, and the researchers would have to make decisions about modeling adjustments to implement dynamic estimates and forecast future behaviour. In the specific case of the 2019 General Election, the most recent publicly available micro-data was wave 16 of the British Election Study (BES) online panel (Fieldhouse et al. 2019), gathered respectively in June of 2019, and hence desperately outdated; iii) *data & model selection.* MrP solutions dealing with voting behaviour and constituency-level estimation allow data from multiple sources to come together at various stages: in the building of stratification frames, auxiliary micro-data is generally used to further disaggregate cells derived from the census (see Leemann & Wasserfallen's MrsP (Leemann

& Wasserfallen 2017)), and constituency-level past election results are leveraged to increase the representative quality of the frame (Lauderdale et al. 2020). At the modeling stage, turnout training-data can be different from training-data for vote choice (Lauderdale et al. 2020); further to the question of modeling choices, the functional forms that made the 2019 models successful, and to which MrP applications are heavily sensitive, are fast-evolving and largely to the discretion of the researcher. For instance, Hanretty et al. (Hanretty et al. 2018) show that a substantial area predictor is key to successful estimates, though the literature agrees that researchers should *"...optimiz[e] an MRP model for a particular research question"* (Warshaw & Rodden 2012, Buttice & Highton 2013), leaving researchers with the task to make theoretically motivated decisions, often without a ground-truth to evaluate their veracity. Having motivated the inclusion of certain individual and area-level variables does not end model selection: in presenting YouGov's method for the 2019 election, Lauderdale & Blumenau (Lauderdale & Blumenau 2019) highlighted the need for cross-level interaction effects, such as between individual-level past vote choice and constituency level variables. The takeaway from this paragraph is that there are a number of critical choices left to the discretion of the researcher; there is uncertainty as to which successful choices made in commercial applications; and the success of the estimates is at the mercy of theory and professional discretion.

In response to these challenges, we propose a novel framework to allow scholars in constrained research environments (constrained by funding, data, area-expertise etc.) to obtain meso-level estimates which can rival the accuracy of high-stakes commercial applications. Our framework pools together low-quality data at the micro-level, and contemporary summary statistics available at the macro-level, to generate synthetic samples of up-to-date micro-data; we then use these to train a learner which performs variable selection without the need for user input.

The paper proceeds as follows: in section 2, we present the three phases that characterize the framework, and give a broad overview of the rationale, data and methods we are calling on at each step. Section 3 presents an evaluation of the modeling strategy and the synthetic micro-data, where we are able to disentangle the two contributions

by leveraging up-to-date pre-election survey micro-data published after the election in question; direct comparisons with ground-truth, YouGov's pre-election forecast, and simple uniform swing are also presented. Finally, section 4 summarizes the findings, and proposes ways in which this methodology can add value beyond psephology.

Type	*Paper*	*Micro-Sample*	*Strat. Frame*	*Learner*
i	Park et al. (Park et al. 2004)	Probability	Original	Multilevel Reg.
i	Buttice & Highton (Buttice & Highton 2013)	Probability	Original	Multilevel Reg.
i	Lax & Phillips (Lax & Phillips 2013, 2009)	Probability	Original	Multilevel Reg.
i	Warshaw & Rodden (Warshaw & Rodden 2012)	Probability	Original	Multilevel Reg.
i	Ghitza & Gelman (Ghitza & Gelman 2018)	Probability	Original (Voter File)	Multilevel Reg.
i	De Jonge et al. (Kiewiet de Jonge et al. 2018)	Probability	Original (Voter File)	Multilevel Reg.
ii	Leemann et al. (Leemann et al. 2020)	Probability	Original	Machine Learning (Ensemble)
ii	Ornstein (Ornstein 2019)	Probability	Original	Machine Learning (Stacked Ensemble)
ii	Bisbee (Bisbee 2019)	Probability	Original	Machine Learning (Bayesian Additive Regression Trees)
iii	Leemann & Wasserfallen (Leemann & Wasserfallen 2017)	Probability	Synthetic (Modeled)	Multilevel Reg.
iii	Ornstein (Ornstein 2017)	Probability	Synthetic (Modeled)	Multilevel Reg.
iii	Hanretty et al. (Hanretty et al. 2018)	Probability	Synthetic (Raked)	Multilevel Reg.
iv	Goplerud et al. (Goplerud et al. 2018)	Probability	Synthetic (Modeled)	Machine Learning (Sparse Regression)
v	Wang et al. (Wang et al. 2015)	Non-Probability (Xbox Opt-In Survey)	Original	Multilevel Reg.
v	Goel et al. (Goel et al. 2015)	Non-Probability (AMTurk Opt-In Survey)	Original	Multilevel Reg.
vi	Cerina & Duch (Cerina & Duch 2020)	Non-Probability (Social Media)	Original (Voter File)	Machine Learning (Random Forest)
vi	Jackman et al. (Jackman et al. 2019)	Non-Probability (Online Opt-In Survey)	Original	Machine Learning (Bayesian Additive Regression Trees)
vi	Gao et al. (Gao et al. 2019)	Non-Probability	Original	Machine Learning (Structured Bayesian Priors)
vii	Lauderdale et al. (Lauderdale et al. 2020)	Non-Probability (Online Panel)	Synthetic (Modeled + Raked)	Multilevel Reg.
vii	Lauderdale et al. (Lauderdale 2019)	Non-Probability (Online Panel)	Synthetic (Modeled + Raked)	Multilevel Reg.
viii	'Learning Voting Behaviour from Synthetic Surveys'	Non-Probability (Online Panel + Synthetic)	Synthetic (Modeled + Raked)	Mchine Learning (Gradient-Boosted Trees)

Table 1: A selection of 20 published papers, outlining the 'state of the art' in terms of how authors handle the three ingredients of MrP applications. Type viii has never been tackled in the literature so far, with this paper being the first to do so.

2 Framework Overview

Our novel framework is characterized by three phases: *Phase I* deals with generating a synthetic population, characterized by a joint likelihood over multiple variables of interest which is ideally indistinguishable from that of the true population of interest; *Phase II* approaches the problem of merging static micro-level information from the synthetic population with dynamic macro-level information from auxiliary data in a systematic manner; *Phase III* concerns the training and testing of a machine learning model to predict variables of interest (in our case, probabilities of turnout and vote-choice by party) for each member of our synthetic population, and further aggregating these to make predictions at the meso level (i.e. the area level or another group level).

Before delving into the details of our framework, it is worth highlighting differences with other research which has developed similar strategies to deal with similar problems. With respect to building stratification frames (our Phase I), *'Model-based pre-election polling for national and sub-national outcomes in the US and UK'* (Lauderdale et al. 2020) served as the backbone of this piece of work; it is this paper that, in dealing with the 2016 US election and the 2017 UK election, as well as the 2016 Brexit referendum, first introduced the idea of 'augmenting' census-provided stratification frames. It is unfortunate that the paper does not spell-out how to go about this beyond a very superficial description in Table 1; we attempt to fill-in the gaps in this paper, which deals with the 2019 UK election and is therefore faced with a slightly different set of challenges with respect to out-datedness of the census - for instance. We also learned from work by YouGov (Lauderdale & Blumenau 2019, Lauderdale 2019) and Survation (Hanretty 2019) which was developed simultaneously to this paper; here the main findings which we apply to our own research include a) the importance of allowing a model to learn individual-area interactions; b) the understanding that running these complex models via STAN (Carpenter et al. 2017) can be computationally intensive, and one has to be ready to make some simplifying assumption with respect to the exploration of the posterior predictive distribution; this is what motivated us to move to a machine learning approach, which we saw as more computationally efficient and equally capable of approximating the necessary functional forms. The rest of the thoughts in this paper, with respect to regression and post-stratification solutions, are original.

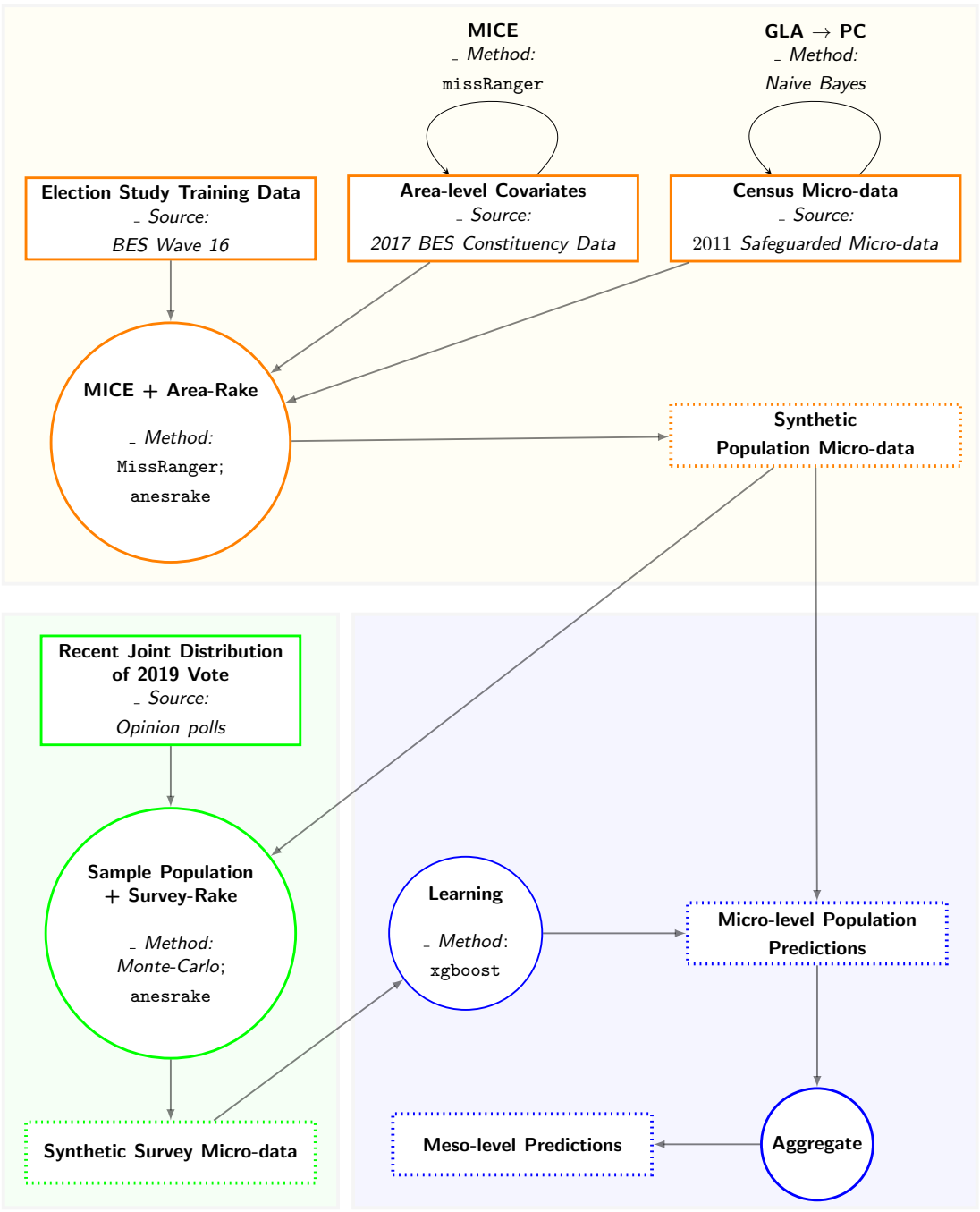


Figure 1: Schematic visualization of the proposed framework: Phase I (orange); Phase II (green) and Phase III (blue). Data are shown in rectangular boxes, and the source we use in the framework is reported beneath the data-type; if the data is generated within the framework - and hence is predicted with uncertainty - the rectangle is dotted to underscore the data is not completely 'solid'. Circular nodes indicate the applications of functions to the data, and the specific package or method used is noted beneath the function name.

Thinking Un-parsimoniously. Throughout this framework we pay little attention to the idea of 'parsimony' in our modeling choices. Though parsimonious approaches have been known to perform better or equally to un-parsimonious ones when properly tuned by a modeler with ample knowledge of the phenomena at hand, especially when it comes to voting behaviour (Lewis-Beck & Stegmaier 2014, Fisher 2015, Graefe et al. 2015, Abramowitz 2008), the philosophy we embrace in this work is one of relative ignorance about the underlying phenomena (voting behaviour in the 2019 election); all that we know is that voting behaviour models are complex (Lauderdale et al. 2020); and that they can involve a dramatic number of non-linear effects (Lauderdale & Blumenau 2019), including interactions across multiple levels of analysis, the nature of which may be clear to subject experts but unknown to us. As such, we use prediction algorithms which are able to select optimal models (i.e. the right set of variables and parameters to replicate the systematic elements of behavioural outcomes' data generating process) based on objective criteria derived directly from the data - for example loss functions based on out-of-bag predictions. Given our a-priori ignorance, any plausible set of covariates could be important in discerning vote-chocie - and hence any such set of variables must be included in the analysis. This is a situation akin to Genome Wide Association Studies (Cantor et al. 2010), where the relevance of any genetic variant to the outcome of interest is unclear a-priori, and therefore all variants have to be brought to bare simultaneously. It is no coincidence that the methods used in this paper have been developed to deal directly with GWAS (Wright & Ziegler 2015, Stekhoven & Bühlmann 2012).

A critic might say that the data generating process of the vote is well known, and all that is necessary to avoid the large computational expense of this framework is to read the literature and design models accordingly. Our response to this is simply that our method need not be tailored to any specific application: whilst the true model of voting behaviour might be known, the data generating process of other measurable quantities of interest which can be studied at the micro- and meso- levels, such as COVID-19 susceptibility; psychological attitudes; propensities to be exposed and believe fake-news; etc. may be completely unknown. It is in order to speak to this general public that we privilege non-parsimonious and heavily computational approaches. To the extent that we seek to summarize data before analysis, for example by reducing the area-level

covariates to the top 100 principal components later in the framework, it is simply to enable computation under reasonable time-frames - it is not driven by theoretical needs or wants.

2.1 Phase I: Generating a Synthetic Population

The orange part of Figure 1 presents a schematic visualization of the data and operations conducted during Phase I of our framework. In what follows we provide a brief outline and motivation of each element of Phase I.

Much like traditional MrP applications, our framework needs a stratification frame to post-stratify our individual-level estimates and obtain aggregate-level predictions. These are typically derived by aggregating census-micro data, where this is available; in the context of the 2019 general election in Great Britain, we use the publicly available UK Safeguarded Microdata[47], a large, representative, individual level sample at the *Grouped Local Authority* (GLA) geographic level. There are three shortcomings to this dataset for our purposes: i) the dataset is at the wrong geographical area of interest, and there is no precise mapping of GLAs to PCs, as boundaries overlap both ways; ii) it does not contain any political variables of interest, such as past General Election (GE) vote or Brexit referendum vote, which have been shown to be the most predictive variables for future voting behaviour in applications of MrP to previous elections in GB (Lauderdale et al. 2020); iii) the 2011 census micro-data is extremely out-dated for the 2019 election, and may no longer accurately represent the relevant distributions of interest at the area level. Our solutions to these shortcomings motivate the schematic in Figure 1.

2.1.1 From GLA to PC

First we deal with converting GLAs in the original census micro-data to PCs, the geography of interest. The strategy, developed in detail in the Appendix, involves a two-step procedure: first we calculate the approximate probability that each individual belongs to a given census Output Area (OA), the lowest non-broken geography in common between GLAs and PCs; second, we sample at random from the candidate OAs according to these probabilities. The result is a vector OAs, which can be uniquely mapped to PCs. Note that the census micro-data is from 2011, and so are all the relevant lookups and auxiliary data we leverage for this mapping.

Though in principle it is possible to fully preserve matching uncertainty, by directly

[47] `https://www.ons.gov.uk/census/2011census/2011censusdata/censusmicrodata/safeguardedmicrodata`

storing the probabilities of belonging to each OA (and therefore each PC) for each census record, we prefer to work with a single random sample of OAs. This is due to the following reasons: we are not interested in the individual-level variance in assignment; the cross-PC variance in demographics is well preserved by the random sample; and the excessive computational expense of storing probabilities of belonging to hundreds of thousands of OAs per record.

The Appendix further contains an evaluation of the GLA to PC allocation, from which we can show that, though not perfect, our strategy for assigning PCs to observations in our census micro-data is principled, accurately reproduces the ordering of constituency over these key variables, and provides reliable estimates for each of these variables at the constituency level, as indicated respectively by high correlation and low MAE. By improving on random assignment, we effectively relieve convergence pressure from the raking algorithm applied at the end of Phase I; this is achieved by making it more likely for there to be a weight-vector that can re-construct the desired marginals, whilst preserving the respective odds-ratios. Beyond facilitating raking, accurate assignment of PCs at the individual level is also likely to preserve some of the PC-specific individual-level correlations, which may become important in the prediction modeling of Phase III.

2.1.2 Augmentation with Political Variables

To overcome the lack of political variables, we integrate the census micro-data with two auxiliary datasets: a micro-level sample which contains at least a subset of the census variables, and an area-level dataset containing area-level predictors. First, the area-level data derived from from the BES[48] is merged to the census micro-data by PC, hence augmenting the records with relevant PC-level information. Second, we obtain the latest publicly available auxiliary micro-data, namely the BES online panel wave 16, and also augment this dataset with area-level covariates. Finally, we perform *data integration*, a two-step procedure of stacking and multiple-imputation, laid out by Carrig et al. (Carrig et al. 2015). Multiple imputation is always performed using the chained-equations method (Van Buuren & Oudshoorn 1999), which enables us to leverage the

[48] `https://www.britishelectionstudy.com/data-object/2017-bes-constituency-results-with-census-and-candidate-data`

full conditional distributions of our incomplete data[49].

Carrig et al. indicate the assumption of a 'common data-generating process' must hold for this sort of data-integration to produce informative results. This needs to be carefully justified in our case: the BES online panel notoriously over-samples from the highly politically engaged population, and as a result under-samples non-voters; moreover, the slim number of non-voters that are sampled are not representative of the general population of non-voters (Mellon & Prosser 2017). Hence if we rely exclusively on the BES online panel for estimating the political variables of interest we are going to end up with estimates that are biased in the same direction. The post-integration dataset will undergo significant re-weighing via raking, though as Mellon and Prosser show (Mellon & Prosser 2017), this will not make the predicted voting population more representative. Secondly, there is the issue of time-of-sampling: the census is from 2011, whilst the BES panel is from June 2019. Imputing variables using the 2019 BES essentially transforms the 2011 representative census micro-data, to a large online-sample from 2019; it remains representative of the 2011 demographics, but the full conditional distribution of the political variables will be entirely derived from the recent online panel. Hence we have to think of this integrated dataset as a massive non-representative online panel. It follows that the 'common data generating process' assumption is dubious to say the least.

The key to accepting this assumption for our application is that it does not have to be true - it just has to be useful. We want the resulting integrated dataset to be as close as possible to the true 2019 population, in all its demographic and political features; the online BES panel is the most recent survey available to us which can inform us on 2019 political behaviour - which we will need in Phase II - so we have no choice but to hold on to these, as we have no equally recent and large representative samples we can use freely. What can we do to lesser the amount of online selection bias faced by these surveys, as we use them to impute political variables?

We propose two strategies to temper the bias of the online panel: i) we integrate two further surveys with the census micro-data: the 2017 and 2015 face-to-face BES

[49]Note that the area-level covariates have missing values, especially in the aggregate demographics (such as % with a college degree by Scottish PCs, for instance), and hence have to undergo a round of MICE prior to data-intergation. This is done separately, before stacking the datasets, to ensure the area-level imputed values are consistent, as we do not want individual-level variables to influence the constituency-level imputed values.

surveys. These have the great advantage of sporting a validated turnout measure, at the cost of being outdated and not having any information relative to the 2019 election; these surveys are more representative, collected through a very different sampling design, and hence do not suffer from the same selection bias as the online panels. Stacking these with the BES can be useful to moderate the the 2015 and 2017 conditional likelihoods toward more realistic levels, and improve predictions for 2019 political variables. The stacking diversifies the 2015 and 2017 voting population in the training data. To the extent that one of the issues with the BES online panel is the lack of diversity in non-voter types, and noting the persistence in non-voting behaviour across demographics between elections (Lauderdale et al. 2020), we can partially remedy that by introducing non-voters from the face-to-face surveys;

ii) the second strategy in our toolkit is to choose an imputation model which applies a significant amount of regularization. This can help by squeezing imputed estimates of behaviour toward the global mean, and hence shrinking high-voting-likelihood types toward the average voting likelihood, and vice-verse. The exact degree of shrinkage will depend on the learner we choose to predict the missing values. Table A1 in the Appendix displays the result of our stacking effort: a novel dataset is produced from stacking four separate samples: the census micro-data; the BES online panel; the 2015 and 2017 face-to-face surveys. We must now complete the integrated data by imputing missing values.

2.1.3 Augmentation with Area-Level Covariates

We can see from Table A1 a subset of the area-level predictor. We want these features to be included in the dataset and considered in the multiple imputation model to improve imputation accuracy by exploiting correlations across areas; much like in standard multilevel models, we expect the area-level covariates to selectively reduce the shrinkage effect applied by the underlying MI model. A key question is how extensive the area predictor should be - both here and later in Phase III; details on provenance and nature of these features are available in the Appendix. We seek to shrink the area-predictor for computational ease during the MI procedure; in doing so, we want to avoid making any theory-driven variable selection, and preserve as much of the information in the predictor as possible. We resolve to keep all the political variables as they are (excluding any

one of perfectly co-linear candidates); we then perform a Principal Component Analysis (PCA) (Wold et al. 1987) on all the census, APS-derived, and geographic variables, and keep only the first 100 scaled principal components, effectively summarizing the non-political area predictor in 100 variables[50] in a strategy similar to that suggested by Leeman et. al (Leemann et al. 2020). Note: we perform MI on the area level separately, prior to the data integration + MI procedure, to avoid introducing nonsensical individual-level variance in the missing values of the area-level covariates.

2.1.4 Non-Parametric MICE

After having generated the stacked dataset presented in Table A1, we seek to perform Multiple Imputation with Chained Equations (MICE). For an introduction to the general MICE algorithm, and a clarification on the notation we use in this paragraph, see section A3 in the Appendix. Note that the integrated data is awash with complexities: mixed-data types; millions of records to impute; hundreds of features to select from for outlining an imputation model; massive imbalance between missing and observed counts for auxiliary variables; unrepresentative training data. We seek a statistical solution that can leverage the MICE framework to best tackle these challenging circumstances, with computational speed fit for debugging in time-sensitive applications; `missForest` (Stekhoven & Bühlmann 2012), a MICE algorithm with leverages Random Forests (RFs) (Breiman 2001*a*) to impute missing values at each iteration, is such a solution. An introduction to the classic Random Forests algorithm is available in the Appendix. It's sufficient to know that a RF is a non-parametric algorithm that seeks to summarize a set of weak Classification and Regression Trees (CART) (Breiman et al. 1984) to only learn from consistently predictive features of random sub-samples of a training set; in the bias-variance trade-off (Hastie et al. 2009) that characterizes learners, RFs privilege variance reduction at the cost of more biased (regularized) predictions (Zhang & Lu 2012). Pseudo Algorithm A1 in the Appendix is a version of the original `missForest` from Stekhoven (Stekhoven & Bühlmann 2012).

RFs output regularized predictions, by averaging over multiple weak learners; they perform automatic variable selection, relieving us of model-selection responsibilities; they can be exceedingly fast, and easily handle large datasets, because: i) the multi-

[50] These explain a cumulative 93% of the total variance in area-level covariates.

ple weak learners can be estimated in parallel; ii) each weak learner only deals with a small subset of features; iii) learned forests can be stored, making prediction largely a matter of sub-setting the test-data. We note that `missForest` has a tried and tested track-record of accuracy in biology and related fields (Stekhoven & Bühlmann 2012, Starkweather 2014, Shah et al. 2014, Waljee et al. 2013, Hariz et al. 2017), though its application to electoral studies is relatively new (see (Cerina & Duch 2020) for one such application). There a number of algorithms which can implement RFs on `R` (Team et al. 2013), the statistical programming language we use throughout this paper; the one acting as backbone to the traditional `missForest` is the `randomForest` (Breiman et al. 2011) package. `ranger` (Wright & Ziegler 2015) is a novel version of this algorithm based on C++ and optimized for memory and speed, provides an intuitive `R` package (Wright et al. 2017); we implement `missForest` using this faster algorithm, via the `missRanger` package (Mayer 2018,*a*). Finally note that the sheer magnitude of the missing values - which now artifically number in the millions - in the stacked data popses a problem for the estimation; we have reason to believe that the algorithm may fail to converge or converge onto 'noise'. To overcome this challenge we provide weights to the input data to reflect the prior belief about the influence we want each dataset to have on the imputation of a missing value. The chosen setup then encodes the following behaviour: trees grown on variables which are present in both datasets will have an equal probability of training on either, effectively smoothing predictions; trees grown to predict variables which are missing in one dataset, but not in the other, will be trained fully on the non-missing dataset, and vice-versa. More details on this are available in the Appendix, which also contains an evaluation of the Imputation procedure based on an Out of Bag (OOB) comparison. The OOB error for political variables is around 5%, with 'Brexit Vote 2016' and 'Political Attention' being particularly well predicted, with OOBs below 1%. In general, the OOB error is very small, suggesting the imputation is high quality. Note that this does not mean our resulting complete, integrated synthetic dataset is represnetative: low average OOB means that our synthetic records could have plausibly come from the population of auxiliary datasets, under the imputation model; it does not imply that there are plausible relative to a ground-truth. We will have to re-weight these observations to obtain a better approximation of the underlying but unobserved 2019 census population. What the imputation does accomplish however is to

better approximate the true joint distribution of the variables, which is only marginally accounted for via other methods such as MrsP (Leemann & Wasserfallen 2017).

Improving on MrsP. We see this *data integration + non-parametric MICE* approach as an alternative to MrsP (Leemann & Wasserfallen 2017) motivated by a series of weaknesses in this original apprach, which we detail below:

i) *ignoring joint-distributions of augmentation variables.* Critically, MrsP does not provide a way to incorporate joint distributions of variables to augment the micro-data by, in the absence of joint area-level tables. Take for instance 'turnout' and 'vote-choice', where it's impossible to obtain joint area-level tables since vote-choice is never observed for individuals who don't turn out. MrsP will generate independent population targets for the two variables in the stratification frame, based on the joint-distribution of the available cells within the corrected survey. Provided the correlation between these back-bone demographics and the augmentation variables is strong in the auxiliary survey, this approximation may be good enough, and we will observe realistic patterns of correlation between vote-choice and turnout, even in absence of explicit joint modeling. However, as the number of auxiliary variables increases, it becomes increasingly likely that the cross-correlation between these auxiliary variable will matter, and especially so if the initial cells are not very predictive.

   Hence we propose an alternative approach that approximates the joint distribution of both initial cells and all the auxiliary variables simultaneously via MICE.

ii) *reliance on probability surveys.* MrsP has to rely on high-quality, representative auxiliary surveys, as it leverages strictly the observed cell-level distributions of the re-weighted auxiliary survey. This limits the number of auxiliary variables researchers can contemplate to those for which one has representative surveys for, which automatically excludes novel and exciting possibilities such as social-media or purchasing data. Modeling can help in increasing the number of potential auxiliary datasets to those for which we have no representative surveys through regularization or explicit accounting for selection biases. Regularization can further

help in instances where the cell-level samples within the chosen auxiliary survey are small, as it can allow for borrowing of strength across cells.

Our approach can work with both probability and non-probability samples as it allows simultaneous integration of multiple surveys, which can acquiesce the biases of any given one, and exploits regularized modeling to better control the signal-to-noise ratio; it further is able to add value to small cells, or cells that are under-represented in the survey, by borrowing of strength.

iii) *model-selection.* A final limitation of MrsP - and most of the existing MrP literature in general - is that it requires the researcher to make a-priori choice relative to which variables they think will most benefit the stratification frame. As we have said before, the success of MrP procedures can be hugely sensitive to these choices. As such, we prefer a machine-learning approach that allows accounting for the joint distribution of as many variables auxiliary variables is possible to measure, and as is computationally feasible, to be included; this relaxes the theoretical constraints on variable selection, albeit constraints relative to desired speed of estimation and computational burden remain.

2.1.5 Raking to Known Area-level Marginals

After data-integration of the census and auxiliary datasets is completed, we need to ensure the resulting augmented dataset respects the true marginal distributions for some key variables which are known to us, and which are important to predict the area-level quantities of interest-namely turnout and vote-choice. For instance, we make an assumption that the proportion of people who voted conservative in the June 2017 GE in a given area must be the exact same in December 2019. The post-integration marginals for the 2017 conservative vote may be out of sync with the observed vote due to a number of reasons, including a severely unrepresentative auxiliary sample at the individual level, and excessive or insufficient pooling in the MICE predictions. To ensure the observed and predicted data match, we employ *iterative proportional fitting* (Deming & Stephan 1940), also known as *raking*, to control-totals (e.g. the observed area-level marginals for the 2017 conservative vote). Raking enables us to re-weight our observations to ensure the marginal distributions in our dataset match the desired control-totals; we

can rake to multiple control-totals simultaneously, i.e. we can demand weights for a given observation to be consistent with both 2017 conservative vote and 2016 brexit vote, for instance. Note however that the more control-totals we demand our weights agree to, the more likely that these are inconsistent and in disagreement with each other, or with the existing joint distributions, leading to convergence problems (Battaglia et al. 2009). This will be particularly true if we are raking to control-totals from different time periods, as is the case in our example. Details on the raking algorithm, including scope conditions for its success, are provided in the Appendix.

2.1.6 Dealing with an Out-dated Census

MICE and raking will enable us to deal with the outdated nature of the augmented census-microdata, as our 2011 census frame will have imputed political variables which are consistent with a 2019 individual-level auxiliary sample, and will be re-weighted to conform to various key area-level marginal distributions from 2017 and 2016. Indeed, Leemann & Wasserfallen's simple version of the MrsP procedure (Leemann & Wasserfallen 2017) shows that it is enough to get these key area-level marginal distributions right for the resulting stratification frame to be valid. We note however that, whilst these may be empirically sufficient fixes to the outdated nature of the frame for most applications, they do not guarantee that the resulting frame conforms to the true joint distribution of the population at the area-level, over key political and socio-demographic variables, at the time of the 2019 election. The limits of the frame are evident as we only have a small number of recent distributions to rake to, and the individual-level auxiliary data we use to augment the census comes from an un-representative online panel. This is a limitation of this application, and other MrP applications, which calls for the development of methodologies to update the full joint-distribution of census-microdata. Said methods should be the subject of further inquiry.

To conclude our presentation of Phase I, we note that typical MrP applications aggregate the data to generate a stratification frame at the *cell* level; we do not, as the number of individual-level variables we augment our census micro-data with, and define our hypothetical cells by, is large enough that aggregating the individual sample over cells leads to little to no reduction in the rows of the dataset. The outcome of Phase I

is a Synthetic Population dataset at the individual-level.

2.2 Phase II: Generating Synthetic Surveys

The green areas in Figure 1 correspond to a simple schematic visualization of the operations and data coming into Phase II. The objective here is to generate a sample of micro-data which is representative of the population of interest not just in the predictors for voting behaviour, as the raked synthetic population is presumed to be, but also in the current turnout and voting patterns.

2.2.1 Surveying the Synthetic Population

Unlike traditional MrP applications, our framework makes use of population microdata for more than mere post-stratification. Note that our synthetic population already contains an initial guess at the turnout and vote-choice probabilities for each individual in the population, based on the likelihood of the auxiliary dataset, re-weighted to conform to the area-level control-totals after raking. Hence, assuming the Synthetic Population we have generated in Phase I is a good enough approximation of the true population of interest, we can survey it: by sampling at random from this population we can learn about vote choice for the 2019 election conditional on turnout, at the area level, by group, or more generally at the meso-level (the subject of our inquiry). Sampling further eases the computational burden of having to perform statistical operations on the entire synthetic population. A naive estimate of vote-share by area or group could be obtained by simply taking the summary statistics of these synthetic-population samples (after appropriate power calculations); however, these are obviously out-dated, as they are ultimately derived from the BES, and would add little more to population-weighted estimates from the original panel wave 16.

Hence we need a way to update the full joint distribution of 2019 vote-choice in our synthetic survey samples to reflect current preferences; to do so, we leverage the dynamic information coming from publicly available opinion polls.

2.2.2 Collecting Publicly Available Polls

An example of the information we obtain from opinion polls is displayed in Figure 2. The data collection rule is simple: we collect marginal and joint distributions for the 2019

GE vote, exclusively for voters which are likely to turn-out, for every publicly available poll released between November 1^{st} and December 12^{th}; we only collect data relative to GB, ignoring Northern Ireland. We do not collect marginal and joint distributions of turnout from the polls due to resources constraints - meaning we will rely on the BES wave 16 turnout question for our application. Though this will no doubt impact our estimates negatively, as turnout attitudes in December of 2019 are unlikely to resemble June of the same year, we note that updating vote-choice should be enough to show marked improvements in the quality of the out-dated micro-data. Future work will focus on examining the impact of updating turnout and vote-choice jointly. All pollsters release marginal distributions for likely voters, but the choice of which joint distributions to release, and the way in which the attributes of interest are defined, varies wildly between pollsters. Table 2 displays which attributes we were able to collect 2019 vote distributions for, by pollster[51]. We were able to collect at least 1 observation of the joint distribution of 2019 vote for 53 attributes, from 63 different polls by 13 different pollsters, along with the marginal distribution of 2019 vote over 9 major parties[52] for each poll.

As we use this polling data to update the voting likelihoods observed in the synthetic samples, problems arise however as sampling design is variable across polling companies, as are the types of weights incorporated in the tables they publish. As a result, in an uncertain and chaotic environment relative to the data generating process of the polls, we make the following simplifying assumptions: i) *random sampling*: we assume that the marginal and shallow-joint distributions are the results of truly random samples from the population of voters in GB. We know this assumption to be violated, as many pollsters re-weights their samples to account for known selection effects and/or demographic imbalances; nevertheless, under the assumption that the weighting mechanisms each pollsters uses are realistic, it is reasonable to assume a true random sample would have given rise to the same tables; ii) *equivalent turnout propensities*: we assume that our guess at turnout for each individual in the synthetic population will be equivalent to

[51]Note that some of the missing attributes by pollster are due to inconsistencies across pollsters: for example, *Region* and *Age* variables may be categorized differently by pollster; the data-cleaning rule we use is to retain the most frequent categorization where possible; re-aggregate to recover the most frequent categories where the pollsters provide more disaggregated distributions; and drop infrequent and irreconcilable categories entirely.

[52]To be exact, 8 major parties + the *other* category.

that made by the pollsters; moreover, we assume all pollsters make the same guess. Again we know this assumption to be violated, as different pollsters engage in different mechanisms to determine turnout propensities at the demographic level. But under the assumption that the polls are reliable, they should all be showing equivalent quantities after accounting for random sampling; hence the assumption is reasonable.

2.2.3 Generating Synthetic Survey Samples

Under the assumptions outlined above, we sample at random from the synthetic population conditional on turnout for each poll in our database; we sample exactly N individuals per poll, where N is the size of the poll conditional on turnout - i.e. the size of the likely voters in each poll. Note that to do this we will have to make our own best-guess at the individual-level turnout propensities in the population[53] ; for now assume we have made a turnout prediction (in terms of probability) for each individual in the population, and we are using these probabilities to sample likely voters from the synthetic population. Note further that polling generally takes a number of days, as is evident from Figure A11, which shows the days-in-field per polling company over the course of the election campaign. We don't actually know the distribution of the observations per day, so we make a further simplifying assumption and take the number of observations for each poll to be the same for every day the poll was in the field; this means we effectively end up smoothing the polls over time.

Relationship to Ecological Inference. Having obtained a sample of likely voters, we are left with outlining a method to harness polling information and update the marginal and joint distributions of 2019 vote in our synthetic survey samples. From a hierarchical prospective, we can think of data coming from opinion polls as containing marco- or meso-level information; the synthetic survey samples contain micro-level information.

Data integration of this sort is typical of *ecological inference* (Freedman 1999, King et al. 2004), where the objective is generally to make micro- or meso-level inference

[53] Further recall that thanks to our multiple-imputation step, we already have guesses for 2019 turnout in the synthetic population; nevertheless, these guesses are derived from a multiple imputation procedure which was not optimized for obtaining the best predictions for this single variable, but rather for approximating the full joint distribution of the missing and observed variables. Hence we believe we can obtain more accurate guesses using the same training data, but tuning a non-parametric model to predict turnout.

from area-level dynamics. Typical ecological models which guard against the *ecological fallacy* make valid inference by stipulating a robust theory for the functional form of the relationship between the micro and macro levels(See Wakefield (Wakefield 2004) and Richardson et al. (Richardson et al. 2005) for a review of the most common ecological models). More recently, a class of Bayesian Hierarchical models known as *hierarchical related regression* models (Jackson et al. 2008) has been developed to integrate individual and area-level information, by positing a hierarchical modeling structure which relates the macro and micro parameters of interest, and jointly estimating these under a Bayesian framework.

Though related to ecological inference by similar aims and features, our application differs from the literature cited above in important ways: i) we are not interested in parametric inference, but merely in consistent data integration; ii) micro-level accuracy is not of interest, but meso-level accuracy is; iii) we know a-priori that our 2019 synthetic micro-data on voting behaviour is outdated and hence inaccurate, whilst ecological models which integrate micro-level data usually leverage contemporary representative samples.The implications of these considerations follow. The first point is straight forward: we do not need to posit a parametric model for the data generating process that gave rise to both the synthetic micro-data and the polling macro- and meso- data, as we are not interested in inference. Point ii highlights that we can afford to be 'wrong' about the individual-level, so long as we are 'right' at the meso-level of interest. Practically, we integrate synthetic surveys and opinion polls with the ultimate objective to achieve accurate area-level predictions. There are multiple individual-level datasets that may be consistent with the same area-level likelihood, and we are indifferent as to which of these micro-level datasets we end up with as a result of data-integration. Point iii states simply that we desire to heavily discount the marginal distributions from the synthetic micro-data, in favour of their counterparts from the opinion polls, whenever this is possible. This is in contrast from general pooling performed in ecological models, which assumes the micro and macro data sources contribute significantly. Considerations about data-size and computational speed also loom large.

Driven by these practical considerations, we opt to use Iterative Proportional Fitting (Deming & Stephan 1940) as a way to integrate the marginal and shallow-joint distributions we obtain from opinion polls with the micro-data at our disposal.

Raking. We need to ensure each sample from the synthetic micro-data is consistent with the poll it is meant to be representing. We re-weight the samples using IPF. More specifically, we perform two rounds of raking: in the first round we use control-totals at multiple levels - e.g. we have control-totals for the marginal distribution of the vote, as well as the shallow-joint (2-way and 3-way) distributions of 2019 vote by covariates the kind of 2017 GE vote and 2016 Brexit vote. Due to issues relative to approximations in the distributions as a result of missing control-totals and mismatches between sample levels and control-totals, which we will describe in the Appendix, we typically encounter convergence issues in this first effort at raking, which will stop after the maximum number of rounds has been reached, rather than when the control-totals are equivalent to the sample totals. As a result we perform a second raking stage, which seeks to ensure the marginal distribution of the 2019 vote is perfectly in line with the control total. This is because typically the marginal distribution of the poll is thought to be the more important feature of the micro-data which underpin MrP models. No convergence problems are ever encountered at this step.

In summary, we rake the synthetic survey micro-data to the marginal and shallow joint distributions derived from the polls, simultaneously. Some convergence problems are to be expected, as pollsters disagree amongst each other, and with the synthetic sample, and hence it may not always be possible to find optimal weights such that all the constraints are respected. Nevertheless, we find empirically that even when raking terminates prematurely due to convergence issues, the changes in weights are still useful for the purposes of improving the of the synthetic micro-data. Moreover, to ensure the synthetic surveys are at least marginally consistent with the opinion polls, we enforce a second round of raking which only uses the marginal distributions; no convergence problems are encountered here.

The resulting data from Phase II is a set of synthetic survey micro-data, meant to approximate to the best information available the real 63 surveys fielded by the 13 pollsters in our sample.

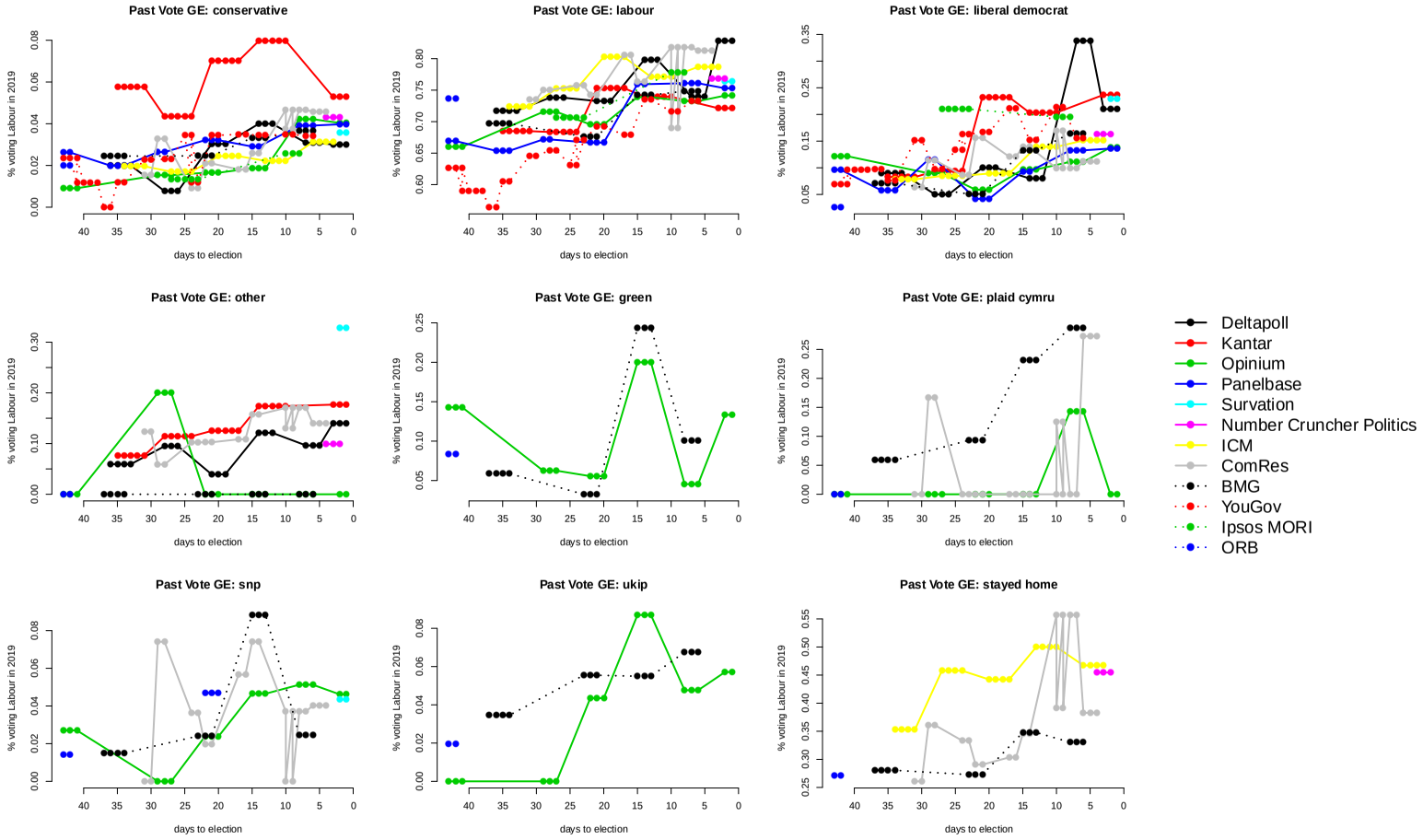


Figure 2: Distribution of Labour vote over the course of the 2019 campaign, by 2017 general election voting groups, derived from public opinion polls. The legend on the bottom-right categorizes the pollsters. Polls fielded over multiple days are assumed have recorded the same information for each day in the field.

Attribute	BMG	ComRes	Deltapoll	ICM	Ipsos MORI	Kantar	Number Cruncher Politics	Opinium	ORB	Panelbase	Survation	YouGov
Age Polls - 2. 19-24	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
Age Polls - 3. 25-34	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Age Polls - 4. 35-44	✓	✓		✓	✓	✓	✓	✓	✓		✓	
Age Polls - 5. 45-54	✓	✓		✓	✓	✓	✓	✓	✓		✓	
Age Polls - 6. 55-64	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Age Polls - 7. 65+	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Approx Social Grade - 1. ABC1	✓	✓	✓	✓	✓				✓	✓		✓
Approx Social Grade - 2. C2DE	✓	✓	✓	✓	✓				✓	✓		✓
Brexit Vote 2016 - leave	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Brexit Vote 2016 - remain	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Brexit Vote 2016 - stayed home	✓			✓		✓	✓	✓	✓	✓		
Educational Attainment - 1. Level1 or None	✓				✓						✓	
Educational Attainment - 2. Level2 or Level3					✓						✓	
Educational Attainment - 3. Level4 or Higher	✓			✓	✓		✓				✓	
Ethnicity - 1. White	✓			✓			✓					
Ethnicity - 2. Non White	✓			✓			✓					
Household Income - 1. £0 - £19,999	✓										✓	
Household Income - 2. £20,000 - £39,999	✓										✓	
Household Income - 3. £40,000+	✓										✓	
Past Vote GE2017 AND Brexit Vote 2016 - conservative AND leave				✓								
Past Vote GE2017 AND Brexit Vote 2016 - conservative AND remain				✓								
Past Vote GE2017 AND Brexit Vote 2016 - labour AND leave				✓								
Past Vote GE2017 AND Brexit Vote 2016 - labour AND remain				✓								
Past Vote GE2017 AND Brexit Vote 2016 - liberal democrat AND leave				✓								
Past Vote GE2017 AND Brexit Vote 2016 - liberal democrat AND remain				✓								
Past Vote GE2017 - conservative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Past Vote GE2017 - green	✓							✓	✓			
Past Vote GE2017 - labour	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Past Vote GE2017 - liberal democrat	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Past Vote GE2017 - other	✓	✓	✓			✓	✓	✓	✓		✓	
Past Vote GE2017 - plaid cymru	✓	✓						✓	✓			
Past Vote GE2017 - snp	✓	✓						✓	✓	✓	✓	
Past Vote GE2017 - stayed home	✓	✓		✓			✓		✓			
Past Vote GE2017 - ukip	✓							✓	✓			
Political Attention - high							✓					
Political Attention - low							✓					
Political Attention - mid							✓					
RGN11NM - 10. Wales	✓	✓	✓	✓			✓	✓	✓		✓	
RGN11NM - 11. Scotland	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓
RGN11NM - 1. North East	✓	✓							✓			
RGN11NM - 2. North West	✓	✓							✓			
RGN11NM - 3. Yorkshire and the Humber	✓	✓							✓			
RGN11NM - 4. East Midlands	✓	✓							✓			
RGN11NM - 5. West Midlands	✓	✓							✓			
RGN11NM - 6. East of England	✓	✓							✓			
RGN11NM - 7. London	✓	✓	✓		✓		✓	✓	✓		✓	✓
RGN11NM - 8. South East	✓	✓		✓					✓			
RGN11NM - 9. South West	✓	✓		✓					✓			
Sex - 1. Male	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sex - 2. Female	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tenure - 1. Owns Outright	✓				✓		✓					
Tenure - 2. Owns with Mortgage or Loan	✓				✓		✓					
Tenure - 3. Rents	✓				✓		✓					

Table 2: Available attributes by pollster.

2.3 Phase III: Learning, Predicting & Aggregating

The bottom-right, blue quadrant of Figure 1 presents the final schematics for Phase III of our framework. In this phase, we aim to learn from our synthetic survey samples, which we will use as training data, and leverage what we learn to predict values of interest for each individual in our synthetic population. We then aggregate individuals up to the area-level, or any other level of interest, and produce our estimates of 2019 voting percentages and turnout for each parliamentary constituency in GB.

2.3.1 Prediction in Regression and Post-Stratification

The traditional MrP procedure requires breaking the population down into small categorical groups of interest, or *cells*; predictions are made for each of these cells' behaviours of interest and then aggregated at the area-level, weighting for the size of each cell relative to the overall population. Our application works in a similar way, with the main difference being we predict individual-level values, rather than cell-level, since individuals in our synthetic population are defined by so many characteristics that aggregating to cells does not provide meaningful reduction in the size of the population dataset.

Generally the individual-level data with which to train our learners to make cell-level predictions will be lacking: samples will tend to be heavily under-powered at the area-level, and even more so at the cell-level; furthermore, MrP applications will often deal with unrepresentative samples (Wang et al. 2015, Goel et al. 2015, Jackman et al. 2019, Cerina & Duch 2020), suffering from important selection bias. In both the circumstances cited above, we may wish to apply *regularization* to the predictions, i.e. we would like to partially pool our cell-level estimates of behaviour towards the global mean; in the case of under-powered samples, this is desirable to the degree that there is high correlation between the frequently observed areas or groups and the rare or missing; in the case of un-representative samples, we may wish to regularize to concentrate on the signal, and filter-out the noise, from the selected training data.

2.3.2 Choosing a Learner

The key quality of a learner that makes it useful for our purposes, therefore, is the degree of *regularization* it applies to the out-of-sample cell-level predictions (Gelman 2018), and the extent to which this can be tailored to the specific characteristics of the application at

hand. It follows that multilevel regression is a popular solution, as it provides reasonably adaptive regularization; its output and inner workings are relatively easy to interpret compared to other learners; and it is amenable to fully-Bayesian estimation via Markov Chain Monte Carlo (MCMC) methods (Gelman et al. 2013).

Despite these considerable qualities, multilevel regression models have some significant drawbacks: i) *model selection*. Similarly to other parametric regression models, extensive model-selection is required; the modeler has to define the functional form of the outcome of interest in detail, choosing which variables to include in the prediction, the relevant interactions, and the hierarchical structure; ii) *attenuation bias*. This is a well recognized phenomena in MrP applications (Lauderdale et al. 2020, Cerina & Duch 2020), and is usually the result of an under-developed area-predictor, and/or excessive shrinkage at the individual level; iii) *estimation speed*. Though recent advances in computational tools have seriously diminished the computational burden that high-dimensional multilevel models pose, there is still a fundamental deficit relative to other machine-learning techniques which can offer similar regularization properties.

Considering briefly point iii, we note that the comparison with machine learning for predictions, where parameters are usually estimated via fast optimization algorithms (Goodfellow et al. 2016) is admittedly unfair, as MCMC has to solve integration problems, which enable us to learn about entire probability distributions (and hence uncertainty) rather than merely estimating some optimal point on a distribution of interest. Nevertheless, it seems to us that current MrP applications share many of the aims and solutions of traditional prediction problems in the machine learning literature, and the focus on integration over optimization may be attributable to the method's highest profile practitioners and initial proponents coming out of the Bayesian tradition. With that in mind, we note that when using the state-of-the-art No-U-Turn sampler (Hoffman & Gelman 2014)[54], implemented in the STAN probabilistic modeling language (Carpenter et al. 2017), to explore the posterior distributions of interest, a fully parametric multilevel model for the 632 constituencies in GB, characterized by all the features explored in Hanretty et al. (Hanretty et al. 2018), can take takes over 11 hours to run for sample sizes around $100,000$ units. These speeds can be somewhat alleviated by the use of Vari-

[54]This is a more sophisticated version of Hamiltonian Monte Carlo (Duane et al. 1987, McElreath 2017, Gelman et al. 2013), which enables faster exploration of the posterior through an adaptive search for optimal step-sizes.

ational Bayes (Blei et al. 2017, Goplerud 2019), which is a promising avenue currently implemented in STAN (Kucukelbir et al. 2017), though we typically pay a price in terms of the accuracy of the uncertainty estimation. Going back to our first two points, namely model-selection and attenuation bias, we note that the two are heavily linked: Hanretty et al. (Hanretty et al. 2018) show that attenuation bias can be relaxed by introducing a meaningful area-level predictor over the individual-level categories, as well as including a local-smoothing prior at the constituency level; Lauderdale & Blumenau (Lauderdale & Blumenau 2019) made the point that cross-level interactions can help selectively relax shrinkage. It follows that we may want a systematic approach to the inclusion of modeling features, one that makes accuracy of high-stakes MrP application reproducible without the a-priori knowledge of the functional form of the phenomena at hand. The literature has started to propose some interesting ideas on this matter: sparse regression (Goplerud et al. 2018) is a relatively simple extension of the existing framework which allows for endogenous model-selection; tree-ensembles such as Bayesian Additive Regression Trees (Bisbee 2019) and Random Forests (Cerina & Duch 2020) have been proposed to deal with issues of variable selection, along with more sophisticated *stacking* (Ornstein 2019), which bring together a disparate number of learners and computes the optimal predictions for each cell in the stratification frame using cross-validation.

Contributing to this growing branch of the regression and post-stratification literature, we propose a novel solution based on ensemble methods, and in particular we leverage `xgboost` (Chen & Guestrin 2016), a gradient boosting algorithm which performs automatic model selection; provides flexible and fine-grained control of regularization via parameter tuning; is immensely scalable in number of records and features; and is incredibly fast-fitting. We provide a detailed review of the method, along with all tree-based learners used in this paper, in the Apppendix.

`xgboost` allows us to control several hyper-parametrs; these include all the standard parameters we can control in a Classification and Regression Tree (CART) or a Random Forest (RF), such as the depth of a tree, the size of the bootstrapped samples α, and the number of features to split at each iteration. We can further influence the number of bootstrapped samples[55], the degree of smoothing across terminal nodes of each tree, and

[55] In `xgboost` these correspond to the number of boosting iterations

the number of terminal nodes - thus the complexity of the learner - per tree. Though these are many parameters to tune, they provide unprecedented control over the learner: `xgboost` is flexible enough to to accommodate a large swaths of functional families of interest. Nevertheless, recent work has shown that xgboost can be highly sensitive to parameter tuning (Bentéjac et al. 2019); as such we must outline a principled tuning strategy, which we do in the following paragraphs.

2.3.3 Models

Before we provide a principled strategy for hyper-parameter tuning, we must define the set of models we are training. Table 3 shows the characteristics of each model we run.

First, we want to prove to the reader that our strategy for predicting meso-level quantities, conditional on some training data, works when the training data is up-to-date. Hence we take the BES waves 17 and 18, which encompass over $70,000$ observations from an online panel obtained between election week and 6 weeks prior, and apply our modeling strategy to these data points; these are models $i.a$ and $i.b$ in the table. Note that these BES waves were published several weeks after the election results were announced, and hence would have not been available to a practitioner in the run-up to the election. Nevertheless, we use these to convince the reader that, if there are any faults in the results of models fit on the synthetic micro-data samples, these do not come as a result of a poor modeling strategy, but are more likely to be as a result of data-quality issues. We will return to this in the next section.

Second, we want to be able to tease-out how much of the hypothetical gain in accuracy as a result of our synthetic micro-data procedure can be attributed to our raking-to-the-polls method, as opposed to the qualities of the original BES 16 data, which serves as the backbone of the synthetic population data from which we sample our synthetic micro-data. We further need a measure of turnout which would have been available prior to the election, as we are suggesting that the synthetic micro-data models would have been available prior to the election; hence we fit a turnout model on the BES 16 data to predict population turnout early on in the campaign. What we just outlined are models $ii.a$ and $ii.b$ from the table.

Finally, we run our vote-choice model on the synthetic micro-data sample - which is conditional on the turnout predictions from model $ii.a$. This is model $iii.b$. Note that

this is the only model for which we have pollster-identifiers, and hence this could in principle be used to make inference about pollsters, their sampling strategies and their accuracy - though we leave these considerations for further research. One challenge with the pollster identifiers is that each pollster will 'spit-out' pollster-specific predictions; hence, as we go to predict vote-choice for our synthetic population, each record in the synthetic population has an hypothetical number of vote-choice predictions equal to the number of pollsters we are analyzing. We consider averaging predictions from multiple pollsters, such that house-biases may cancel out; this requires generating a set of population predictions for each pollster, and then averaging these together. This approach is theoretically sound, but computationally challenging depending on the size of the synthetic population and the speed of model-fitting. It makes it difficult to tune the model with cross-validation, as we would have to run 12 models, one for each pollster, at each search-iteration. A more computationally efficient alternative is to choose a 'champion' pollster a-priori, and make predictions for such a pollster exclusively. We opt for 'YouGov' as our a-priori champion, given their past-election performance and the frequency of their surveying in this election, and produce predictions on behalf of this polling firm.

Model Id.	Training Data	Outcome Quantity	Tuning Cost Function	WtE	Pollster Id.
i.a	BES 17 & 18	Turnout $\in \{0, 1\}$	Out-of-Bag Brier Score	[1,7]	
i.b	BES 17 & 18	Vote Choice $\in \{\texttt{con}, \texttt{lab}, \texttt{ld}, \texttt{grn}, \texttt{brx}, \texttt{pc}, \texttt{snp}, \texttt{oth}\}$	RMSE on Uniform Swing	[1,7]	
ii.a	BES 16	Turnout $\in \{0, 1\}$	Out-of-Bag Brier Score	[26,29]	
ii.b	BES 16	Vote Choice $\in \{\texttt{con}, \texttt{lab}, \texttt{ld}, \texttt{grn}, \texttt{brx}, \texttt{pc}, \texttt{snp}, \texttt{oth}\}$	RMSE on Uniform Swing	[26,29]	
iii.b	Synthetic	Vote Choice $\in \{\texttt{con}, \texttt{lab}, \texttt{ld}, \texttt{grn}, \texttt{brx}, \texttt{pc}, \texttt{snp}, \texttt{oth}\}$	RMSE on Uniform Swing	[1,7]	✓

Table 3: A display of the main characteristics of each of the 5 models we need to tune and run. The training data for the first four models refers to the British Election Study Online Panel Waves. In the 'Tuning Cost Function' column, 'Out-of-Bag' refers to the proportion of the training-data which was left-out in each search-iteration of the cross-validation procedure. WtE stands for 'Weeks to Election', a measure of time to illustrate the temporal breadth of each training sample.

2.3.4 Tuning Strategy

Tuning hyper-parameters is fundamental to the success of models we discuss thus far, and `xgboost` is no exception (Bentéjac et al. 2019). We use a cross-validation (Stone 1974, 1977, Hastie et al. 2009) strategy to tune hyper-parameters and estimate uncertainty; more specifically, we use k-fold cross-validation, which involves splitting the original dataset into k typically non-overlapping samples, fitting the learner to $k-1$ of the

samples and testing the performance of each fit to the test sub-sample. A challenge k-fold cross-validation is choosing the number of folds. An influential contribution by Kohavi (Kohavi et al. 1995) recommends 10 fold cross-validation, though we note there is ongoing discussion in the literature as to the quality of various k-fold strategies for both tuning and uncertainty estimation (Bengio & Grandvalet 2004, Zhang & Yang 2015). Our choice of k will be smaller, sacrificing some of the theoretical benefits in favour of a more practical approach accounting for computational speed, and empirically checking the consistency of our estimates is reasonable across different cross-validation runs.

We focus on a small subset of hyper-parameters which we thought were key to our application, largely governing model complexity and shrinkage, and allow the rest to be set to default; we further restrict the search space to some reasonable a-priori limits; details on these parameters are available in Table 4.

Parameter	Theoretical Range	Prior Range	Regularization ↑
`eta`	$[0, 1]$	$[0.2, 0.8]$	$\to 0$
`max_depth`	$[1, \infty]$	$[1, 10]$	$\to 1$
`min_child_weight`	$[0, \infty]$	$[1, 15]$	$\to \infty$
`subsample`	$[0, 1]$	$[0.2, 0.8]$	$\to 0$
`colsample_bytree`	$[0, 1]$	$[0.2, 0.8]$	$\to 0$
`nrounds`	$[1, \infty]^{56}$	$[1, 150]$	$\to 1$

Table 4: A table summarizing the parameters to be tuned: the prior range represents the subset of the theoretical range we chose to explore; the last columns displays the direction of the regularization effects. Parameter `eta`, also known as *learning rate*, shrinks the impact of each boosting iteration; `max_depth` sets the maximum depth of each tree; similarly, `min_child_weight` sets a hard limit to the minimum size of a terminal node; `subsample` and `colsample_bytree` respectively set the proportion of records and features to be considered at each boosting iteration; `nrounds` governs the number of boosting rounds.

The objective of tuning is to find the best possible combination of hyper-parameters which maximises fit on some out-of-sample metric, the nature of which we shall disucss later in this section. We face numerous challenges here: i) it is unfeasible to examine all possible combinations of hyper-parameters; ii) gradient optimization on the out-of-sample cost function cannot be immediately implemented, as it cannot be guaranteed that the cost function is smooth and convex; iii) we need to ensure our training subset is large enough to capture deep-interactions and non-linear effects, which typically translate to computationally and time-intensive training; iv) because `xgboost` relies heavily on random subsampling of features and records, under a small number of boosting itera-

tions predictions can be relatively unstable. Challenges i and ii are solved by employing a random search approach, where we examine an arbitrary but large upper limit of $1,000$ random combinations of the parameters, and rank these by their relative error on the out-of-sample metric. With respect to iii, we perform a simple 'leave $\frac{1}{2}$ out' cross-validation at each search round; namely, for each search iteration, we train a the model on a random half of the full training set, and test its performance on the other half. This limits the amount of times the model has to be fit compared to k-fold cross-validation, whilst also generating large enough training samples, able to reflect important data complexities. Finally, to increase the consistency of the predictions across separate training and fitting rounds, as well as to improve model performance under the same logic of the random-forest bagged estimator, we select the top 1% of performing models, draw an arbitrary 5 sets of predictions for each[57], and then average across models and draws to generate our final predictions for each quantity[58].

Choosing Cost Functions. We develop models for turnout, as well as vote-choice conditional on turnout. Our training strategy depends on two different cost functions.

For turnout, we seek to minimize the out-of-sample Brier score directly, at each search-iteration. This presents some problematic aspects, as turnout from the BES online panel, whether from recent observations or worse from out-dated waves, is notoriously poorly measured: for the BES wave 16, the weighted turnout average across all observation is 81.5%, whilst for the combined waves 17 and 18 we have 82%[59]. To avoid large over-estimation of turnout, we make two relatively strong assumptions: first, we assume turnout in the upcoming election will be equal to the last election on record - namely the 2017 election; second, we assume *uniform over-reporting*, across voter-types and over time and geographical areas. Armed with these tools, we proceed to correct the original turnout outcome variable by randomly turning-off any voter for whom the

[57]The limit on this number is computational, as it can be costly to keep in memory or save on disk the results of each model fit.

[58]Note that a small degree of stochastic variability persists, in the order of 0.05 to 0.4 decimal points on the RMSE for each predicted party and turnout at the area level, though the performance of the overall RMSE remains highly stable across draws

[59]These numbers are obtained by first completing the respective datasets to impute any missing data via the usual `missRanger` procedure, and then converting the turnout variable to a dichotomous one by sampling in probability for each respondent, according to the following schema: 'Very unlikely that I would vote' is equivalent to a 0 probability of voting; 'Fairly unlikely' corresponds to 0.25; 'Neither likely nor unlikely' corresponds to 0.5; 'Fairly likely' to 0.75, and 'Very likely that I would vote' stands 1 probability of voting.

turnout dichotomous variable is equal to 1, and continue doing so until the weighted average settles to the 2017 turnout value of 68.8 %. We then assume this variable is correctly measuring 2019 turnout preferences, and resolve to minimize the out-of-bag brier-score on this variable at each search-iteration.

With respect to the vote-choice cost function, we can be a little more sophisticated, by encoding the prior expectations of a *uniform swing* at the area level. It is worth exploring this model in some depth, and assessing its relationship with our modeling strategy.

The *uniform swing* model was first postulated by Butler and Stokes (Butler & Stoke 1974, Butler & Van Beek 1990), and more recently has been re-discovered in the US (Jackman 2014). The basic uniform swing model for a given party vote share is presented below:

$$\delta_t = \nu_t - \nu_{t-1}; \tag{24}$$

$$\psi_{t,c} = \psi_{t-1,c} + \delta_t; \tag{25}$$

where ν_t is the national vote share of the party in question at election t; δ_t is the change in that share since the last election; and $\psi_{t,c}$ is the constituency c vote share for that party at election t. The uniform swing has performed well in recent election in GB: using the observed swings for 2015 and 2017, and swinging Scotland separately from England and Wales to account for the differences in political parties and dynamics, one recovers estimates of headline seat numbers that are within 11 and 5 seats from the true seat counts respectively (Fisher 2019). In 2017, the only year for which we have a comparison of an MrP model and the uniform swing, the latter performed exceptionally well at the constituency level, calling 91.8% of constituencies correctly, and doing just slightly worse than the much applauded YouGov model, which called 92.9% of PCs correctly. This is not to say that the MrP model not remarkable, but rather that uniform swing sets a high bar for justifying complex and expensive modeling. We have mentioned that one potential pitfall of MrP models is attenuation bias; models which do not include a constituency-level predictor are significantly subjected to this phenomena, which dramatically narrows the variance of area-level predictions (Cerina & Duch 2020, Hanretty et al. 2018).

We can think of parameter tuning in terms of swing: providing only individual-level features to the learners - or providing hyper-parameters that shrink predictions heavily - makes it more likely to recover that a *proportional swing* model, i.e. a model where the transition probabilities of party voters between elections are constant across constituencies - essentially assuming a uniform swing at the voter-level rather than the constituency-level. McLean (McLean 1973) pointed out that the uniform swing of the sort intended by Butler, namely a uniform change in the vote share from one party to another over two elections across constituencies, is incompatible with constant transition probabilities across PCs, and cites Butler and Stokes to highlight that one must account for local environmental pressures *"...such that the elegant harmony of uniform swing* [*would be*] *restored.* (Butler & Stoke 1974, McLean 1973). This insight, going back to the early $70's$, is perfectly reconcilable with the Hanretty et al. notion that robust area-level predictors are needed to improve prediction accuracy (Hanretty et al. 2018) at the area-level. Moreover, we can reconcile proportional swing to regularization in learning models: higher levels of regularization will lead to a more proportional swing, and vice-versa to whatever swing is supported by the data. Given the recent performance of the uniform swing, we would expect models underpinned by high-powered probability samples to steer toward a uniform swing, a priori; however, the samples we are training our models on are non-probability.

We must therefore steer our vote-choice models to learn uniform-swing-compatible patterns; we do so by calculating the area-level vote predictions at each search-iteration[60], and storing the root mean squared distance between the unifrom swing predictions and the `xgboost` predictions across all parties and all areas.

We have hence outlined a full tuning strategy, involving separate cost functions for turnout and vote-choice. Note however that another valid approach would be to treat turnout model hyper-parameters as simply other tuning parameters contributing to the vote-choice cost-function - hence performing tuning for both turnout and vote-choice in a single shot. This sort of approach is attractive, as it tunes the turnout model directly to maximise the chance of capturing uniform-swing patterns, though it has the significant disadvantage of being quite tasking in computational terms: `R` as a software

[60] This involves training a model on a subset of the training observations; using this model to predict vote-choice values for the synthetic population; aggregating the synthetic population at the area-level.

faces some challenges in dealing with garbage cleaning, and memory storage, and the prediction matrix for the synthetic population data can be as large as 20 GB, making it difficult to repeatedly compute and store in-memory over $1,000$ search-iterations. A corollary issue relates to the model's sensitivity to the interactions between turnout and vote-choice hyper-parameters: had we decided for the one-shot approach, depending on the sensitivity of the model to the turnout parameters, we would have had to increase the number of search-iterations many times over.

Weights. We conclude this section by outlining some weighting choices. In both turnout and vote-choice model, we incorporate sampling weights as provided by the BES or raking, respectively. We incorporate conditionality on turnout by assigning turnout weights to the training observations for vote-choice models. The turnout weights are derived from the predictions of the optimal turnout model for all training observations. To implement discounting of outdated observations, we further assign hyperbolic temporal weights of the following form: $\omega_{it} = t_i/\min(t)$, where t stands for 'weeks to election', and can take value 1 for election-week, up to a large integer for the most out-dated observation in the training set. Such weighting implies that an observation from the most recent week-to-election is given weight equal to 1, whilst observations from the second most recent weeks are given hyperbolically less weight, and so on. These weights are incorporated in the models jointly to the turnout weights and other available sampling weights by simply multiplying those together to form a joint weight. The scale of these weights will not be easily interpretable, though that is not a concern as `xgboost` normalizes the weights and uses them as sampling weights in the formation of the trees. A final weighting-related issue is that of where parties are standing, and where they are not. We simply remove from the training data any individual voting for a given party, if that party is not fielding a candidate in the constituency where the individual resides. This does not account for strategic voting, but should provide some degree of padding against, for example, Brexit Party voters who are unaware their candidate will not stand in their constituency.

2.3.5 Meso-level Aggregation

The final step of our framework is to generate meso-level predictions. The meso-level quantities of interest which one is typically interested in the UK electoral context are constituency-level vote-shares, though we note our strategy enables the examination of a variety of meso-level quantities, such as behaviour by deeply interacted individual-level features (e.g. turnout propensity by age-gender-ethnicity group). Models trained on the synthetic survey micro-data, which are characterized by pollster-identifiers, can further be used to systematically examine differences across pollsters in their sampling of small voter categories of interest. The formulae to calculate the area-level turnout shares and vote shares for parliamentary constituencies $a = 1, ..., 632$ follow. There exist a list of indices for each parliamentary constituency, A_a, such that each member i, for $i = 1, ..., N$, of the synthetic population, belongs to a given area exclusively, $i \in A_a$; all synthetic population records belong to an area, and none belong to more than one. Hence, for vote-choice set defined by $j = \{$`conservative`, `labour`, `liberal democrat`, `green`, `brexit`, `plaid cymru`, `snp`, `other`$\}$, we can recover area level predictions or turnout and vote-choice as follows:

$$\hat{\tau}_a = \frac{\sum_{i \in A_a} \Pr_i(T \mid X) \Pr_i(X)}{\sum_{i \in A_a} \Pr_i(X)}; \tag{26}$$

$$\hat{\nu}_{a,j} = \frac{\sum_{i \in A_a} \Pr_i(V_j \mid T, X) \Pr_i(T \mid X) \Pr_i(X)}{\sum_{i \in A_a} \Pr_i(T \mid X) \Pr_i(X)}; \tag{27}$$

where $\Pr_i(X)$ is the sampling probability of each synthetic-population record, estimated via raking to the known population marginals; $\Pr_i(T \mid X)$ is the turnout probability of each member of the synthetic population, estimated by the respective turnout models; and $\Pr_i(V_j \mid T, X)$ is the probability of voting for party j conditional on turnout, estimated with the respective vote-choice model.

3 Results and Evaluation

We have thus far generated a synthetic population by multiple imputation and data-integration with auxiliary datasets, as well as raking to recent known election results; we have further generated synthetic micro-data by sampling from this synthetic population and raking to some of the marginal and shallow-joint distributions available from public opinion polls; finally, we laid out a strategy for training models of turnout and vote-choice, predict turnout and vote-choice values for our synthetic population records, and aggregating at the desired area level. We now move on to evaluate the performance of our predictions for each of the models outlined in the previous section. We pick three measures to show the local and global performance of the models: first, we analyze the overall performance in terms of total seat numbers by party; this is an important metric, as it is ultimately the only metric that can tell us the outcome of an election: will the Conservatives have a majority, or will there be a hung parliament?. A second and third global measure of accuracy are the national-level vote shares and the Conservative lead over Labour, which is typically a meaningful contributor to the overall seat tally. In terms of a local metric of success, we present the RMSE of the area-level predictions, by outcome of interest. We compare our three sets of predictions with the observed outcomes, published after election-day, and three reference models: the uniform-swing model fitted to the election-week opinion polling average, the YouGov MrP model, and the Survation MrP model. For each metric, we measure and compare error rates by party across models, and further record the gain in accuracy attributable to updating the BES wave 16 with current opinion polls.

Tables 5, 6 and 7 present the results; we will refer back to these in the following paragraphs. We can further explore model-performance at the area level; these are available for model *i* in Figure 9, for model *ii* in Figure A12 in the Appendix, and for model *iii* in Figure 10. Uniform swing area-level predictions are available in Figure A13, in the Appendix. A final meso-level quantity of interest is the demographic breakdown of turnout and vote-chocie predictions. We will evaluate meso-level predictions for the three models at different meso-levels, by comparing the predicted probabilities implied by our models with those derived by a post-election survey by Lord Ashcroft[61]. A

[61] `https://lordashcroftpolls.com/2019/12/how-britain-voted-and-why-my-2019-general-election-post-vote-poll/`

comparison between our synthetic stratification-frame and a 10% random sample from the Survation 2019 MrP frame concludes our evaluation of the methods outlined in the paper.

Model	Training Data	CON	LAB	LIB	GRN	BRX	SNP	PC	Abs. Error
Observed		365	203	11	1	0	46	4	
Uniform Swing	Nat. Polls (WtE = 1)	349	216	17	1	0	46	3	
		(16)	(-13)	(-6)	(0)	(0)	(0)	(1)	36
YouGov 2019 MrP	Proprietary	339	232	15	1	0	41	4	
		(26)	(-29)	(-4)	(0)	(0)	(5)	(0)	64
Survation 2019 MrP	Proprietary	353	220	11	1	1	43	3	
		(12)	(-17)	(0)	(0)	(-1)	(3)	(1)	33
i.a - i.b	BES 17 & 18	361	200	11	1	0	56	3	
		(4)	(3)	(0)	(0)	(0)	(-10)	(1)	18
ii.a - ii.b	BES 16	283	237	36	1	12	59	4	
		(82)	(-34)	(-25)	(0)	(-12)	(-13)	(0)	166
ii.a - iii.b	Synthetic	323	248	8	0	0	53	0	
		(42)	(-45)	(3)	(-1)	(0)	(-7)	(0)	98
	Abs. Synthetic Gain	40	-11	22	-1	12	6	0	68

Table 5: Total seat-tally predictions for each party, compared to the observed results; the seat errors per party are shown below the point estimates, with negative values indicating an over-estimation, and vice-versa. The final row of the table shows the 'Abs. Synthetic Gain', i.e. the gain in absolute error by updating the BES wave 16 with current opinion polls. Expected sums may be off by one or two seats due to rounding.

Evaluating Modeling Strategy. We begin by evaluating the performance of our modeling strategy on the BES Waves 17 and 18. These are waves that were not available prior to the election, and we expect these to be higher quality than our synthetic micro-data. An important reason for this is recency: though many of the synthetic cross-tabs are up-to-date, as they have been raked to the most recent campaign polls, no novel information was added at the constituency level, as only a handful of constituency polls are publicly available. Hence the synthetic micro-data will be characterized by constituency-effects as they were present in June of 2019, when many of the political and economic variables which could affect elections via area-level dynamics were wildly different from the election day in December. Hence, holding modeling strategy constant, we expect models fit to this data to outperform our synthetic micro-data. If this model correctly captures the election results, it suggests the modeling strategy could have done so given appropriate data, providing evidence in favour of the modeling strategy.

The model we fit on BES waves 17 & 18, model i in the tables above, performs remarkably well on two of the four metrics. Table 5 shows this model obtaining the

Model	Training Data	T	CON	LAB	LIB	GRN	BRX	SNP	PC	CON-LAB
Observed		67.4	44.7	32.9	11.8	2.8	2.1	4.0	0.5	11.8
Uniform Swing	Nat. Polls (WtE = 1)		43.5	33.7	11.9	2.6	3.2	3.8	0.2	9.8
			(1.2)	(-0.8)	(-0.1)	(0.2)	(-1.1)	(0.2)	(0.3)	(2)
YouGov 2019 MrP	Proprietary		42.6	33.8	12.4	2.8	2.9	3.4	0.5	8.8
			(2.1)	(-0.9)	(-0.6)	(0)	(-0.8)	(0.6)	(0)	(3)
Survation 2019 MrP	Proprietary		42	34.6	12.7	3	2.9	3.4	0.6	7.4
			(2.7)	(-1.7)	(-0.9)	(-0.2)	(-0.8)	(0.6)	(-0.1)	(4.4)
i.a - i.b	BES 17 & 18	63.2	40.6	27.5	16.4	5.0	3.1	4.1	0.8	13.1
		(4.2)	(4.1)	(5.4)	(-4.6)	(-2.3)	(-1)	(-0.1)	(-0.3)	(-1.3)
ii.a - ii.b	BES 16	60.8	26.7	24.5	20.9	8.5	8.2	4.7	1.3	2.2
		(6.6)	(18)	(8.4)	(-9.1)	(-5.7)	(-6.1)	(-0.7)	(-0.8)	(9.6)
ii.a - iii.b	Synthetic	60.8	38.1	33.6	14.5	3.2	4.4	4.2	0.8	4.5
		(6.6)	(6.6)	(-0.7)	(-2.7)	(-0.4)	(-2.3)	(-0.2)	(-0.3)	(7.3)
	Abs. Synthetic Gain		11.4	7.7	6.4	4.3	3.8	0.5	0.5	2.3

Table 6: National vote share predictions for each party and turnout, compared to the observed results; the error of the predictions is shown below the point estimates. The column 'CON-LAB' displays the conservative lead over labour, a pivotal prediction quantity worthy of separate scrutiny. Note that the uniform swing prediction is entirely derived from the election-week national polls, rather than from aggregation of area-level results.

Model	Training Data	T	CON	LAB	LIB	GRN	BRX	SNP	PC
Uniform Swing	Nat. Polls (WtE = 1)		4.5	4.9	4.6	1.6	5.0	4.5	4.6
YouGov 2019 MrP	Proprietary		4.0	3.9	2.3	0.9	2.8	5.0	2.0
Survation 2019 MrP	Proprietary		4.6	4.8	3.6	1.8	2.9	2.2	1.0
i.a - i.b	BES 17 & 18	5.7	5.3	8.2	6.5	3	3.3	3.4	4.1
ii.a - ii.b	BES 16	7.7	18.3	11.9	11.1	7.3	14.6	5.6	6.8
ii.a - iii.b	Synthetic	7.7	7.3	6.5	5.3	2.3	5.6	4.2	8.8
	Abs. Synthetic Gain		11	5.4	5.8	5	9	1.4	-2

Table 7: Area-level RMSE by model and party or turnout. The final row of the table shows the 'Abs. Synthetic Gain', i.e. the gain in absolute error by updating the BES wave 16 with current opinion polls.

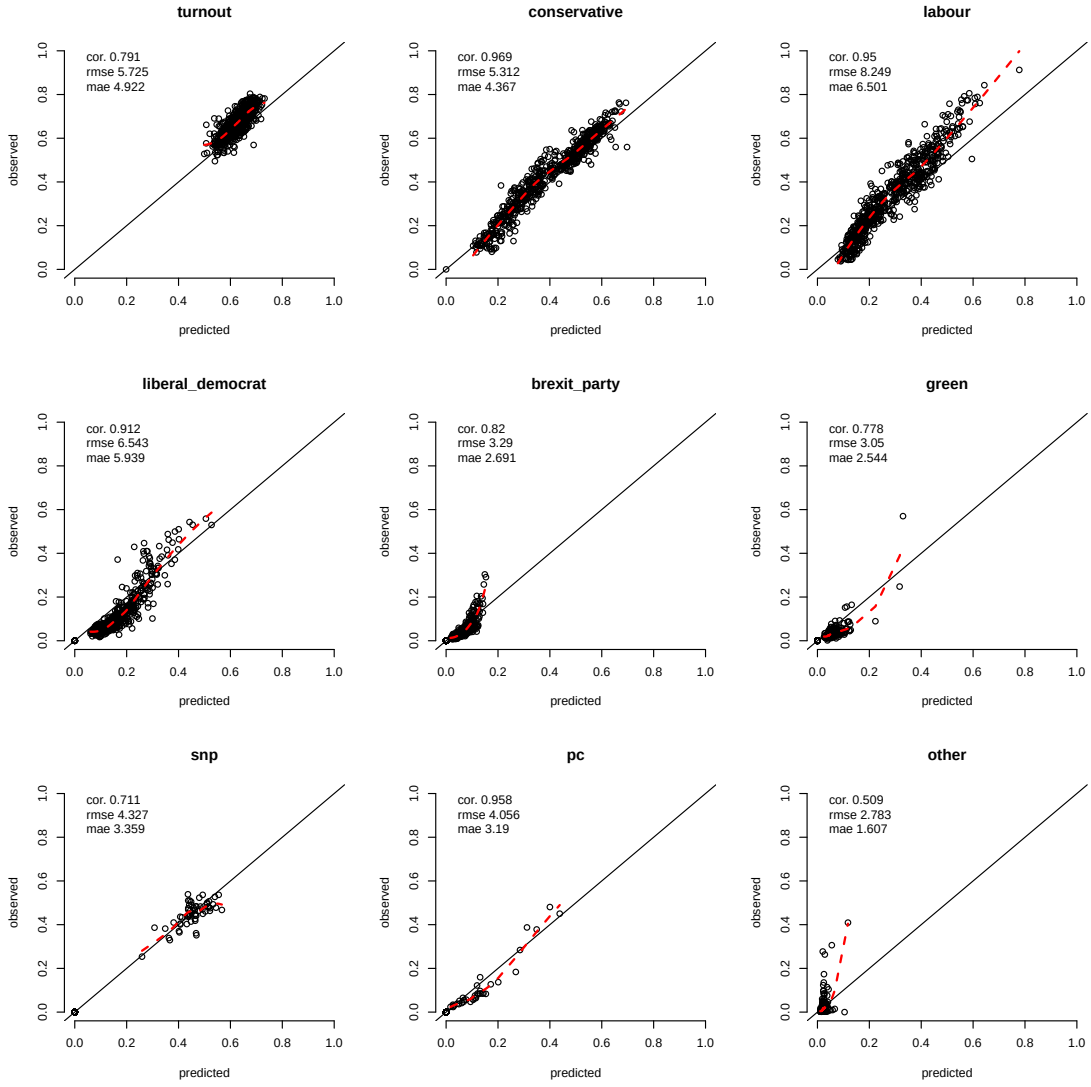


Figure 9: Area-level predictions from the model trained on BES waves 17 and 18.

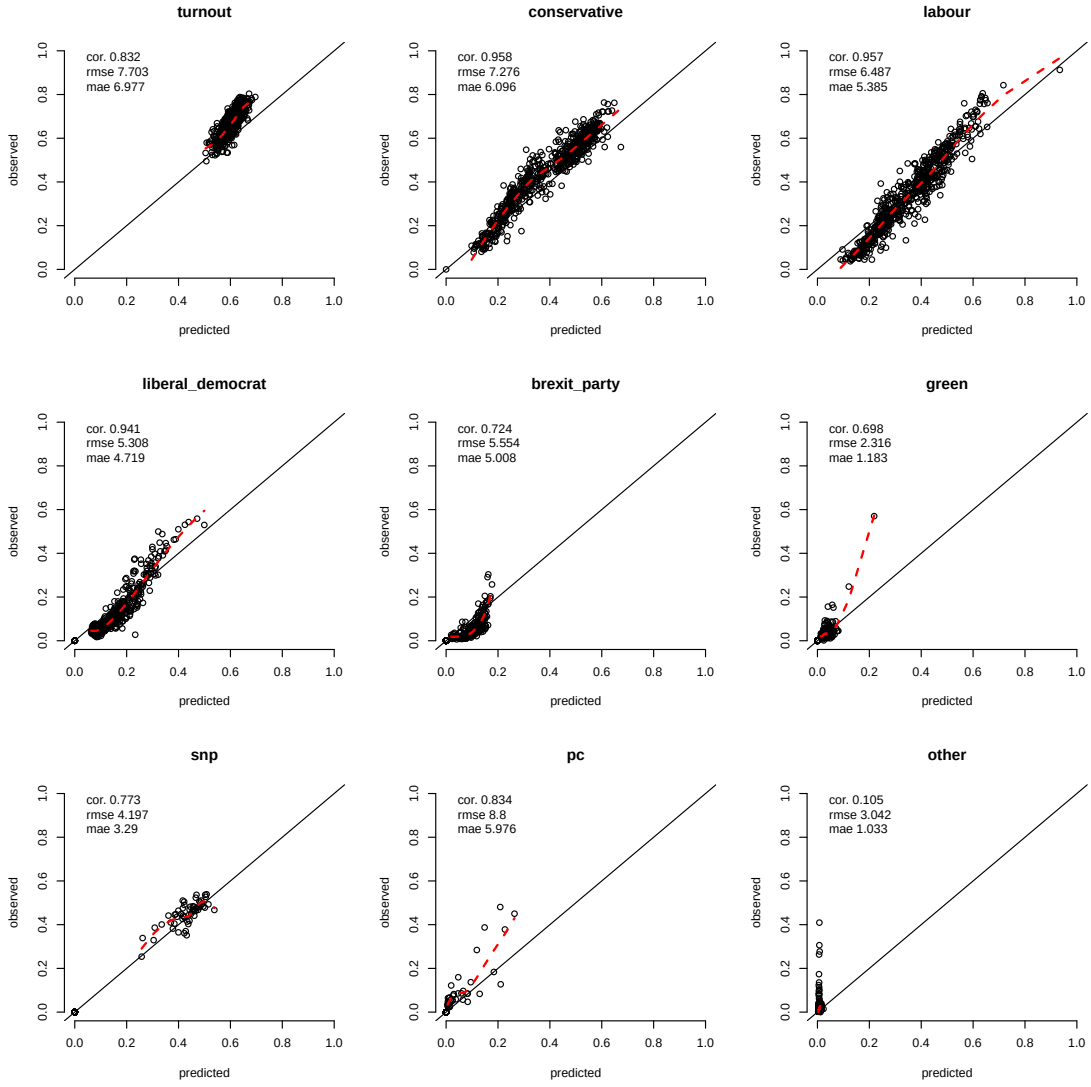


Figure 10: Area-level predictions from the model trained on fully synthetic micro-data.

closest seat-tally predictions for each major party, the lowest absolute seat error overall - 18 seats in total, topping the Survation model by 15 seats - and only seriously missing the mark on the SNP. An important reason for this good performance appears in Table 6, specifically the CON-LAB lead: this has the lowest error on the lead - a mere 1.3 points. On the other hand, this remarkable performance is not met by an equally potent result for the point-estimate of national vote shares: part of the reason for the large CON-LAB lead implied by this model is a severe under-estimation of the Labour party vote - 5.4 points below the true value. This is along with a more moderately severe under-estimation of the Conservative vote (4.1 points) and a large over-estimation of the Lib-Dems (4.6 points). The Green over-estimation is also out of the norm compared to other models (2.3 points), though errors in the national vote shares for Brexit party, SNP and PC are well contained even when compared to the state-of-the-art. Table 7 reveals this performance is consistent at the area-level, with the area-level RMSEs presenting the same pattern as the national level. Digging deeper at the area-level using Figure 9, some important patterns emerge: i) the small under-estimation of the Conservative vote happens almost exclusively in the their 'best' constituencies, suggesting a small degree of attenuation bias; ii) the Labour under-estimation is pervasive, though it worsens for safe-labour seats; iii) we over-estimate the performance of the Lib-Dems for seats where they are particularly weak, and under-estimate when they are strong, suggesting again a degree of attenuation bias; iv) patterns around Plaid Cymru and the SNP are reasonably well captured.

Overall, given the positive performance on the total seat tally and the CON-LAB lead, we deem the modeling strategy largely successful; questions remain over what features of the overall framework would lead to such a large shrinkage-effect for labour, especially when the original (weighted) online survey showed Labour at 32%. We will discuss potential culprits for this behaviour in section 4.

Evaluating the Synthetic Micro-data. Having established that our modeling strategy would have lead to the correct 'call' for this general election, with reasonably high-quality training data, we can now evaluate the effect of updating the BES wave 16 online panel with current opinion polls published over the campaign. We seek to show that our strategy can increase the value of the outdated BES 16 micro-data - the only publicly

available micro-data prior to the election; we further seek to measure whether the extent of this improvement is enough to void the need for collecting more recent micro-data.

Results are mixed: one the one hand, we decisively show the synthesis procedure massively increases the quality of the predictions, holding modeling strategy constant, for every single metric analyzed above: the total seat-tally error drops by 68 seats; the party-level absolute error on the national vote drops by 11.4 points for the conservatives, 7.7 points for Labour and 6.4 points for the Lib-Dems, and around 4 points for the Greens and Brexit Party, with every party benefiting from the synthesis overall. Similar drops in error are recorded for the area-level RMSE, with the exception of Plain Cymru, seemingly the only party to have a worse Synthetic prediction at the constituency level. The weaker gains for small and regional parties are explainable by the size of the cross-tabs in each opinion poll our samples from the synthetic population are raked to: it's not uncommon for polls to have only a handful of Plaid Cymru supporters, making many of the relationships uncovered by the cross-tabs at the party-level seemingly spurious, potentially adding noise to the sample as we rake.

Major gains not-withstanding, the Synthetic model just about fails to predict a Conservative majority, and overall compares poorly to the other models in terms of seat tallies; the national vote-share predictions are overall reasonable, though the major under-estimation of the Conservative vote tarnishes the model's performance on this metric; the RMSE at the area level also proves to be an unflattering comparison, with the major parties collecting RMSEs in the 5 to 8 range - again the Conservative under-estimation being the most severe.

Looking at the area-level plots for the synthetic predictions in Figure 10, similar patters emerge as with the recent micro-data, with a few important differences: i) the under-estimation of the Conservatives is more severe and more pervasive across constituencies; ii) the error on Labour is no longer merely due to general under-estimation but rather more linear attenuation bias - under-estimation in Labour strongholds and over-estimation where Labour is least competitive; iii) the Lib-Dem predictions are ameliorated, reflecting perhaps that the characteristics for which we were able to provide an update via the opinion polls were most meaningful in predicting the Lib-Dem vote, and may not have been properly reflected in the online BES panels; iv) the Plaid Cymru predictions are very poor and lose significant correlation, when compared to the BES

17 & 18 predictions.

Evaluating Turnout Modeling. The final piece of our puzzle is turnout. This was a-priori the hardest part of the prediction effort: a) even in the recent BES 17 and 18 online-panel data, turnout intention is known to be over-reported due to non-negligible selection effects; b) the recent data was unavailable prior to the election, and we real-time the turnout polls, hence we relied exclusively on turnout intentions data from June for a December election - an effort bound to be sub-par. Recall further that we applied a correction to the self-reported turnout measure: we ensured our weighted self-reported turnout in the training micro-data was equal to the past-election turnout (68.8%).

Tables 6 and 7 tell a consistent story about our turnout predictions: predictions trained on the most recent online micro-data out-performed those trained on data from June by over 2% at both national and area levels. At the national level, the turnout prediction error was about 4 points, which is indicative of a meaningful turnout prediction; the score worsens as we look at the area level, though not by a contextually significant amount.

It is intriguing that the direction of the error is positive, meaning we under-estimated turnout by just over 4% in the recent micro-data and close to 7% with the out-dated BES; this suggests we likely over-corrected self-reported turnout, as the uniform over-reporting assumption likely failed as it interacted with the selection effects into the panel. The increase in predicted turnout between BES 16 and waves 17/18 points to more 'activated' voters closer to the election.

Evaluating Demographic Inference. We seek to establish how accurate our prediction modeling was for alternative meso-level quantities of interest, such as demographic group or past vote. We choose the Lord Ashcroft post-election poll as it is large ($13,128$ adults who voted in the 2019 election), highly transparent, freely available, and relies on a representative sample; we further prefer it over other pollsters as we never included Ashcroft polls in our modeling, whilst we did include Ipsos or YouGov polls in the Synthetic micro-data, hence this should provide a more reliable out-of-sample accuracy measure. We seek to evaluate demographic performance of the BES 17/18 model, the BES 16 model, and the synthetic model on two counts: at the aggregate level, over 37 overlapping demographic groups of interest, we examine how close our predictions for

were to what transpires from the Ashcroft polls; at the disaggregated level, for each demographic group we compare predictions from the three models against the Ashcroft ground-truth. These analyses are shown respectively in Figure 11 and Figures A14 to A20 in the Appendix.

The aggregate-level analysis of the performance of each model at the demographic level reveals similar patterns as the area-level analysis with respect to model-comparison: the wave 16 model is unsurprisingly the worse performing, recording large errors on the major parties, reflecting the highly peculiar political situation in June of 2019 with weak Conservative and Labour parties across most demographics, and strong third-party options such as Lib-Dems, Brexit and Greens. The Synthetic model seems to be marginally better at capturing demographic-level dynamics, which is proof that our raking to the polls' cross-tabs had a positive impact. The performance on the Conservative party demographic-level predictions is overall luck-luster, something that again is surprising to us, as the micro-data marginals for both BES 17/18 and Synthetic models reflected a reasonably good estimate of the conservative vote. The large under-estimation of the Labour share in the BES 17/18 model, as well as some of the seemingly large out-sized errors in the Synthetic model's conservative predictions, can be investigated further using the disaggregated plots in the appendix.

Looking at Figures A14 to A20, we can state the Synthetic model typically performs well at the demographic profile level, which is perhaps unsurprising given it can leverage the cross-tabs of many recent polls. Note however that, to the extent that the polls' cross-tabs were correct, it seems the Synthetic predictions are dragged toward the backbone BES 16 predictions, suggesting the underlying interactive structure of polls prior to raking still has a considerable impact on post-raking models. The biggest pitfalls relative to the BES 17/18 model came at the past-vote level, where the under-estimation of the support Labour would gather from its 2017 supporters translated to a significant under-estimation of the Labour vote and an over-estimation of the Lib-Dems and Greens (via those Labour voters who voted Remain in the Brexit referendum) and the Brexit Party (amongst those 2017 Labour voters who voted Leave).

With respect to Turnout, unfortunately there are no representative exit polls or post-election polls which from which we can obtain accurate demographic breakdowns to eval-

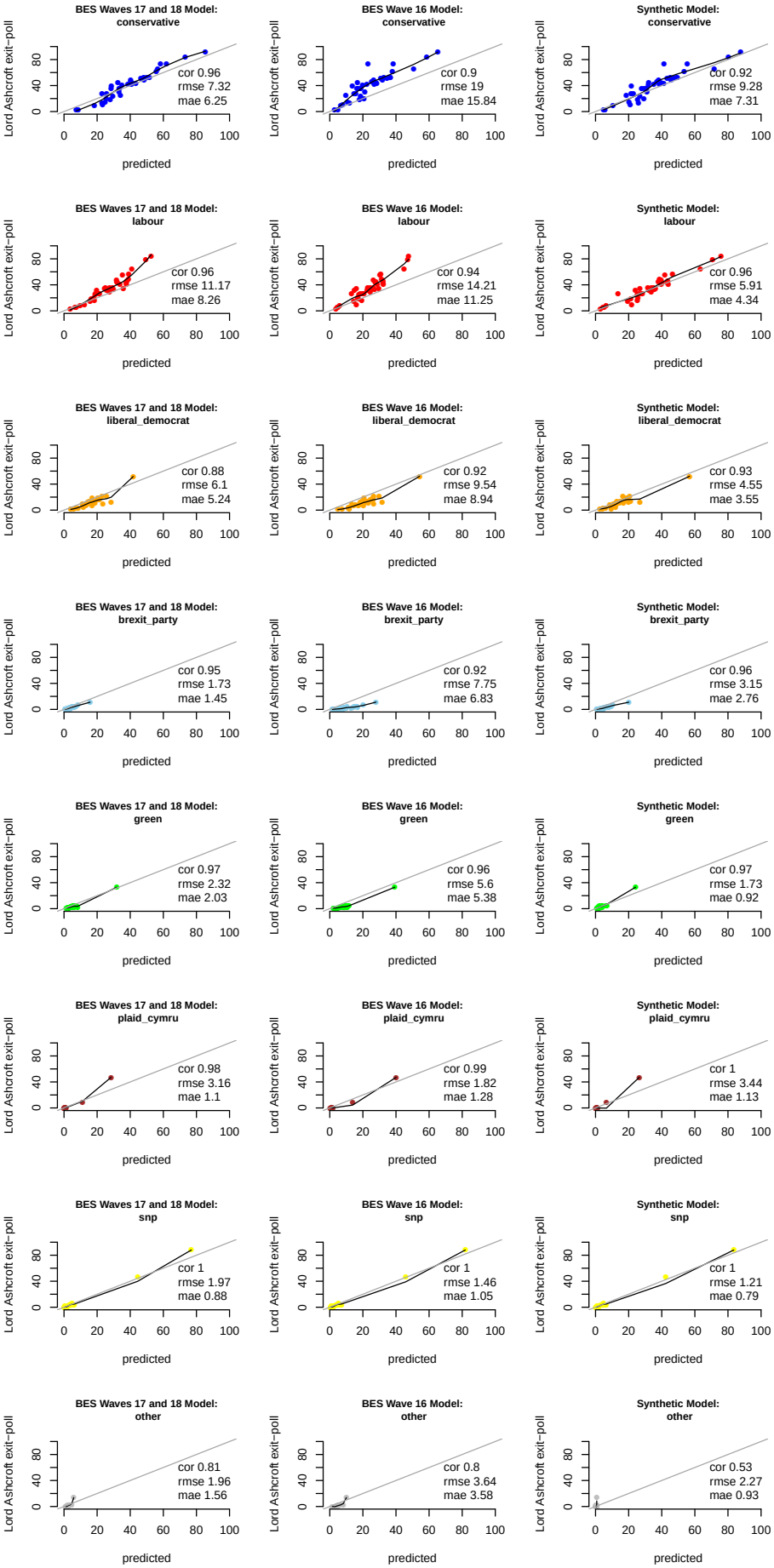


Figure 11: Aggregate demographic vote-choice prediction-accuracy by model and party.

uate our predictions. There is the BES wave 19, though again this is a non-representative online panel, and is notoriously unreliable with respect to turnout propensities. Moreover, commercial MrP results from YouGov and Survation do not publish their turnout predictions. We resolve to compare our turnout predictions at the demographic level with those of the final Survation poll, fielded during the two days prior to the election. We choose Survation due to their being amongst the best performing 2019 election pollsters, as well as having the best performing MrP model. Survation collects turnout likelihood responses between 0 and 10; because of the well known turnout over-reporting problem in traditional pre-election opinion polls (Wells 2019), we use a conservative estimate of turnout: only voters that respond with absolute certainty (a score of 10) are considered to be likely to turn-out. Because the poll is smaller than your typical exit-poll, counting $2,395$ respondents, we are unlikely to have power to make meaningful inference at the demographic level; on the other hand, we can compare average precision across all demographic groups of interest, and treat this as an added measure of turnout prediction accuracy at the meso-level. Figure A21 in the Appendix displays the results of this comparison. It is clear that the Survation polls dramatically over-estimates turnout; we can tell this not only because most of the demographics of interest seem to have turnout larger than the national observed turnout of 67.4 percentage points, but also because we know the exact bias at the 'Country' level. The Survation poll suggests turnout in England, Scotland and Wales to be respectively 78.5, 78.6 and 75.5 points, whilst we know it to have been 67.5, 68.1 and 66.6 respectively, highlighting around a 10 points upward-bias (errors of 11%, 10.5%, 8.9% at the country level). After adjusting the Survation predictions for a 10 point bias across all demographics, we recover a picture which is almost identical to what we observed at the area-level, with the BES 17/18 model obtaining a MAE around 5%, and a BES 16 model MAE at around 7 points. Worthy of note is the higher correlation at the demographic level, just under 0.9 for the BES 16 model and just under 0.85 for the BES 17/18 model, both higher than their area-level counterparts.

We can conclude our turnout predictions substantially under-estimated turnout propensities, though they appear to have done so fairly uniformly. Evidence for this is the smoothing spline fit to the turnout comparison plots in Figure A21, which is almost perfectly parallel to the $y = x$ like. To the extent that the Survation numbers

can serve as a ground-truth for a cross-demographic comparison, if not for a comparison of absolute-levels, it seems our predictions are highly correlated, and the error is constant across demographics. These findings suggest that, though there is evidence over-reporting was not uniform at the individual-level for the selected respondents of the various BES waves, it may have been uniform across demographic groups, suggesting a group-level adjustment of the same form may have performed very well. A posteriori, we can think of a better strategy for correcting for turnout over-reporting - namely leveraging the validated face-to-face BES waves to generate group-level turnout targets, and ensure the micro-data conforms to these via the same 'turn-off' procedure we outlined in this paper. Finally, looking back at the differences in national vote-share predictions, we have little reason to believe the turnout differences are generating much, if any, of the differences between the Synthetic model and the BES 17/18 model fits. This is because the general patterns of turnout correlation at the area level, demographic level, and national errors are very similar; the main difference between the two sets of turnout predictions appears to be that models trained on the BES 16 under-predict turnout by a further 2 points compared to those trained in BES 17/18, uniformly across demographics and areas.

4 Discussion

This paper makes a number of important contributions to the literature on model-based pre-election polling. We have expanded the literature on the building of stratification frames, by providing a detailed account - the first in the literature to our knowledge - of the building of a complex political stratification frame for the UK electorate. Other authors, such as Lauderdale et al. (Lauderdale et al. 2020), have hinted at similar procedures, but our account is by far the most exhaustive. In doing so, we believe to have proposed a novel, theoretically more sophisticated, alternative to MrsP, via our Data Integration + MICE framework; this contribution now enables researchers to obtain arbitrarily deep joint probabilities for the purposes of post-stratification.

We have further expanded the literature on polling aggregation, by using IPF to merge polls at 'deep' demographic levels, and shown conclusively that doing so dramatically increases the quality of the underlying micro-data. Given this, we see no

reason why researchers seeking to predict elections using MrP-like methodologies would not implement IPF to increase the power of the underlying data: would we ever prefer a single poll over an aggregate? We do to MrP what Jackaman (Jackman 2005) and Linzer (Linzer 2013) did to polling aggregation for election forecasting - namely we show, holding sample-size constant, multiple polls are better than one - a truism particularly strong when the underlying micro-data at hand is unrepresentative, as it is often the case in MrP applications.

Our last contribution is to show that, under appropriate tuning rules, `xgboost` can optimize the trade-off between regularization and area-level fitting. Our contribution here is to suggest that hyper-parameter tuning need not be ad-hoc, and can help us bring-in some general theoretical notions from the political science literature - such as the uniform swing. Note that, though our tuning strategy relies on theory in a similar way as the traditional modeler relies on theory for variable selection, our strategy is vastly more adaptive: tuning a model under uniform swing does not preclude features from being included or excluded in a model, depending on the underlying training data.

An important limitation of this work is the lack of accounting for uncertainty, which also limits our ability to make probabilistic predictions. We recognize this is an important limitation, though it was not something we could have possibly tackled in the same paper as the rest of our methodology: the question of how to account for uncertainty at every step of this framework deserves separate scrutiny and the development of a unifying theory. It's worth noting all of the different places in our framework where uncertainty comes in: i) at the assignment of parliamentary constituencies to census micro-data; ii) at the 'Data Integration + MICE' step; iii) as we rake to known constituency-level marginals; iv) as we sample from the synthetic population; v) as we rake to the polling marginals; vi) as we impute missing values in our training micro-data; vii) as we predict turnout; viii) as we predict vote-choice conditional on turnout. We further highlight that accounting for uncertainty throughout the framework carries an important computational load: unless a unifying theory is provided which allows for mathematical estimation of the functional form and standard-errors of predictions, we have to rely on simulations, which increase exponentially at every uncertainty-inducing step; moreover, the computational capacity needed to store hundreds of simulated synthetic popula-

tions, which include several million data points each, is currently unattainable in most research and commercial settings.

Nevertheless, we plan to implement robust accounting for uncertainty in each of the seven sources of uncertainty described above, in future versions of this paper. Our strategies for preserving uncertainty in each source follows:

1. *assignment of parliamentary constituencies to census micro-data*: instead of storing the full stratification frame (which counts roughly 3 million rows), we can store multiple random sub-samples - the size of which should be calibrated to limit distortions to the relative relationships between the size of each cell of interest. An initial approach could be to assume cell-independence, and calibrate the samples according to the Multinomial Thompson tables (Thompson 1987), to ensure the largest cells are preserved in each frame sample. Then we can afford to store a different constituency allocation in each frame-sample, given the reduced computational cost, and preserve allocation uncertainty via simulations;

2. *data integration + mice & sample from the synthetic population conditional on turnout*: multiple imputation is likely to be a large source of uncertainty, due to the large amount of missing data for political variables in the census micro-data. For each cell, we can simulate random cell-counts via predictive mean matching, according to the predictive distribution implied by the `missRanger` algorithm. The challenge here is to choose k, the 'neighbourhood size' parameter on which predictive mean matching relies to sample candidate cell counts, which needs to be tuned. A possible tuning strategy could be to test the algorithm on the complete census micro-data, by randomly setting values to zeroes over multiple iterations of k, and picking the value of k which allows for the minimum loss - a cross-validation approach based on the fully-observed data. After selecting k, and assuming we have already multiple sample-frames from the constituency allocation step, we can simply run a round of MICE on each of the sample-frames, and allow predictive means matching to sample candidate cell counts at random for each sample-frame; the resulting sample-frames will incorporate uncertainty from both the allocation and the imputation steps. We must be careful to choose the number of samples appropriately - this too needs to be tuned - as too few samples will under-represent

the total variance within and between samples.

It then follows that we can then use the same strategy for dealing with uncertainty about sampling from the synthetic population conditional on turnout. As we preserve uncertainty about turnout probabilities from the multiple imputation rounds, we can take a separate set of random samples of likely voters - one for each sample-frame - to rake to polling marginals. Again, care has to be taken to specify a large enough number of initial samples to account for these compounding uncertainties, and there will no doubt be a trade-off between accuracy of uncertainty estimation and computational ease: the more simulation rounds, the more accurate the uncertainty estimates, but the heavier the computational load;

3. *rake to known constituency-level marginals & rake to the polling marginals*: uncertainty about raking weights can be incorporated by formulating the raking problem as a Bayesian model (Si & Zhou 2019); this further enables more flexibility to propagate uncertainty from multiple sources, including the prior uncertainty around the marginal distribution targets themselves - which can be important if the marginal targets are outdated, for instance;

4. *prediction uncertainty*: for the tree-based learners such as the ones used in this paper, and in particular Random Forests, this has been solved by some of the recent literature (Athey & Wager 2019, Athey et al. 2019, Lu 2017, Lu & Hardin 2019). We privilege the Lu & Hardin approach to estimate Mean Squared Predictive Error (MSPE) by leveraging the out-of-bag samples, generated during the construction of each weak learner, as a 'test-set' with which to measure the accuracy of each weak learner, and then smooth this error over learners to obtain the expected average squared loss. This approach holds provided some scope conditions outlined in their work, and it's extremely intuitive. The challenge here is going to be to translate these findings to `xgboost` and gradient-boosted trees in general. It s possible, under some general pre-requisites, to reconcile gradient boosted trees with random forests, by thinking of boosted trees as an ensemble of trees, each trained on the loss of the previous tree (Zhang & Lu 2012, Ghosal & Hooker 2018). In that sense, in both approaches the end result is an average of weak learners; the main difference is that in boosted trees the learners are 'weighted'. Hence

it should be possible to estimate MSPE via appropriately re-weighting the errors associated with each weak learner. This is an important contribution which we will focus on in future work.

Another limitation in this work is the underlying micro-data data which we used to inform our stratification frame. The BES wave 16 data is dramatically outdated and unrepresentative (of 2019 voting intention); it was sampled during a very strange period in British polling history, namely one in which the two major parties had suffered a dramatic slump, a fact that is clearly affecting our estimates throughout the framework. All this considered, it made sense to use this data for the building of the frame, as it was the only one available a-priori. Future work should be concerned with adopting a real-time updating strategy for the frame, which leverages the same survey data one collects for prediction purposes; after all, if the information in the survey is good enough to make forecasts - provided appropriate modeling and adjustments - it is good enough to build stratification frames. On the same topic, one should also be able to update the frame with the latest public-opinion polls, though again this promises to be computationally expensive.

Future research should investigate the various tuning options at our disposal; we trained our vote-choice models to minimize the root mean squared error on the predicted uniform swing constituency-level predictions, but other sorts of training are available. For instance, we could have trained to minimize the CON-LAB lead, at different levels of aggregation. Still in the realm of tuning, future work should look into tuning turnout and vote-choice models simultaneously. This promises to be quite computationally tasking if the underlying training data for turnout and vote-choice are different, though in most cases one should be able to obtain turnout and vote-choice data from the same survey.

The results of this paper are overall promising: a model trained on an unrepresentative online sample of respondents to the most recent BES online panel can accurately forecast the outcome of the election and outperform other pollsters on the CON-LAB lead. This is done without any ad-hoc variable selection, and using a framework which can easily be applied in any other setting where multiple sources of data can be used to build relevant stratification frames and make cell-level predictions. More work remains

to be done to improve accuracy at the area-level, which is not yet industry-standard. The improvements obtained by updating the underlying outdated training data with recent deep polling averages at the demographic level show us at least one way in which to further improve our best estimates.

References

Abramowitz, A. I. (2008), 'Forecasting the 2008 presidential election with the time-for-change model', *PS: Political Science and Politics* **41**(4), 691–695.

Athey, S., Tibshirani, J., Wager, S. et al. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178.

Athey, S. & Wager, S. (2019), 'Estimating treatment effects with causal forests: An application', *arXiv:1902.07409* .

Battaglia, M. P., Izrael, D., Hoaglin, D. C. & Frankel, M. R. (2009), 'Practical considerations in raking survey data', *Survey Practice* **2**(5), 1–10.

Bengio, Y. & Grandvalet, Y. (2004), 'No unbiased estimator of the variance of k-fold cross-validation', *Journal of machine learning research* **5**(Sep), 1089–1105.

Bentéjac, C., Csorgo, A. & Martínez-Muñoz, G. (2019), 'A Comparative Analysis of `xgboost`', *arXiv:1911.01914* .

Bisbee, J. (2019), 'BARP: Improving Mister P using Bayesian Additive Regression Trees', *American Political Science Review* **113**(4), 1060–1065.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017), 'Variational inference: A review for statisticians', *Journal of the American statistical Association* **112**(518), 859–877.

Breiman, L. (2001*a*), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L. (2001*b*), 'Using iterated bagging to debias regressions', *Machine Learning* **45**(3), 261–277.

Breiman, L., Cutler, A., Liaw, A. & Wiener, M. (2011), 'Package `randomForest`', `http://stat-www.berkeley.edu/users/breiman/RandomForests`.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Brick, J. M., Montaquila, J. & Roth, S. (2003), Identifying problems with raking estimators, *in* 'annual meeting of the American Statistical Association, San Francisco, CA'.

Butler, D. & Stoke, D. (1974), *Political change in Britain: Basis of electoral choice*, Springer.

Butler, D. & Van Beek, S. D. (1990), 'Why not swing? measuring electoral change', *PS: Political Science & Politics* **23**(2), 178–184.

Buttice, M. K. & Highton, B. (2013), 'How does multilevel regression and poststratification perform with conventional national surveys?', *Political analysis* **21**(4).

Buuren, S. v. & Groothuis-Oudshoorn, K. (2010), '`mice`: Multivariate imputation by chained equations in `R`', *Journal of statistical software* pp. 1–68.

Cantor, R. M., Lange, K. & Sinsheimer, J. S. (2010), 'Prioritizing gwas results: a review of statistical methods and recommendations for their application', *The American Journal of Human Genetics* **86**(1), 6–22.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of statistical software* **76**(1).

Carrig, M. M., Manrique-Vallier, D., Ranby, K. W., Reiter, J. P. & Hoyle, R. H. (2015), 'A nonparametric, multiple imputation-based method for the retrospective integration of data sets', *Multivariate behavioral research* **50**(4), 383–397.

Cerina, R. & Duch, R. (2020), 'Measuring public opinion via digital footprints', *International Journal of Forecasting* .

Chen, T. & Guestrin, C. (2016), `xgboost`: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics* **11**(4), 427–444.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid monte carlo', *Physics letters B* **195**(2), 216–222.

Elff, M., Gschwend, T. & Johnston, R. J. (2008), 'Ignoramus, ignorabimus? on uncertainty in ecological inference', *Political Analysis* **16**(1), 70–92.

Fieldhouse, E., Green, J., Evans, G., Mellon, J. & Prosser, C. (2019), 'British Election Study Internet Panel Wave 16', `https://www.britishelectionstudy.com/data-object/wave-16-of-the-2014-2019-british-election-study-internet-panel/`. doi:10.15127/1.293723.

Fienberg, S. E. et al. (1970), 'An iterative procedure for estimation in contingency tables', *The Annals of Mathematical Statistics* **41**(3), 907–917.

Fisher, S. (2019), 'The most unpredictable election ever ? Don't believe it', `https://www.prospectmagazine.co.uk/politics/the-most-unpredictable-election-ever-dont-believe-it-polling-numbers-december-12`.

Fisher, S. D. (2015), 'Predictable and unpredictable changes in party support: a method for long-range daily election forecasting from opinion polls', *Journal of Elections, Public Opinion & Parties* **25**(2), 137–158.

Freedman, D. A. (1999), 'Ecological inference and the ecological fallacy', *International Encyclopedia of the social & Behavioral sciences* **6**(4027-4030), 1–7.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Gao, Y., Kennedy, L., Simpson, D. & Gelman, A. (2019), 'Improving multilevel regression and poststratification with structured priors', *arXiv:1908.06716* .

Gelman, A. (2018), 'Regularized prediction and poststratification (the generalization of mister p)', `https://statmodeling.stat.columbia.edu/2018/05/19/regularized-predictionpoststratification-generalization-mister-p`.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Gelman, A. & Little, T. C. (1997), 'Poststratification into many categories using hierarchical logistic regression'.

Ghitza, Y. & Gelman, A. (2018), 'Voter registration databases and mrp: Toward the use of large-scale databases in public opinion research', *Political Analysis* pp. 1–25.

Ghosal, I. & Hooker, G. (2018), 'Boosting random forests to reduce bias; one-step boosted forest and its variance estimate', *arXiv:1803.08000* .

Goel, S., Obeng, A. & Rothschild, D. (2015), 'Non-representative surveys: Fast, cheap, and mostly accurate', *Working Paper* .

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep learning*, MIT press.

Goplerud, M. (2019), 'Fast and accurate estimation of hierarchical models using variational algorithms'.

Goplerud, M., Kuriwaki, S., Ratkovic, M. & Tingley, D. (2018), 'Sparse Multilevel Regression (and Poststratification (sMRP))', *Working Paper* .

Graefe, A., Küchenhoff, H., Stierle, V. & Riedl, B. (2015), 'Limitations of ensemble bayesian model averaging for forecasting social science problems', *International Journal of Forecasting* **31**(3), 943–951.

Hanretty, C. (2017), 'Areal interpolation and the uk's referendum on eu membership', *Journal of Elections, Public Opinion and Parties* **27**(4), 466–483.

Hanretty, C. (2019*a*), '2019 General Election MRP predictions', `https://www.survation.com/2019-general-election-mrp-predictions-survation-and-dr-chris-hanretty/`.

Hanretty, C. (2019*b*), 'EP2019 Results mapped onto Westminster constituencies', `https://medium.com/@chrishanretty/ep2019-results-mapped-onto-westminster-constituencies-8a2a6ed14146`.

Hanretty, C., Lauderdale, B. E. & Vivyan, N. (2018), 'Comparing strategies for estimating constituency opinion from national survey samples', *Political Science Research and Methods* **6**(3), 571–591.

Hariz, N. B., Khoufi, H. & Zagrouba, E. (2017), On combining imputation methods for handling missing data, *in* 'International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems', Springer, pp. 171–181.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Hoffman, M. D. & Gelman, A. (2014), 'The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.', *Journal of Machine Learning Research* **15**(1), 1593–1623.

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

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

Jackman, S., Ratcliff, S. & Mansillo, L. (2019), 'Small area estimates of public opinion: model-assisted post-stratification of data from voter advice applications', *Working Paper* .

Jackson, C., Best, Nicky & Richardson, S. (2008), 'Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1), 159–178.

Johnston, R., Hay, A. & Taylor, P. (1982), 'Estimating the sources of spatial change in election results: a multiproportional matrix approach', *Environment and Planning A* **14**(7), 951–961.

Johnston, R. & Pattie, C. (2000), 'Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting', *Political Analysis* **8**(4), 333–345.

Kiewiet de Jonge, C. P., Langer, G. & Sinozich, S. (2018), 'Predicting state presidential election results using national tracking polls and multilevel regression with poststratification (MRP)', *Public Opinion Quarterly* **82**(3), 419–446.

King, G. (1998), *Unifying political methodology: The likelihood theory of statistical inference*, University of Michigan Press.

King, G., Tanner, M. A. & Rosen, O. (2004), *Ecological inference: New methodological strategies*, Cambridge University Press.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Ijcai', Vol. 14, Montreal, Canada, pp. 1137–1145.

Kolenikov, S. (2014), 'Calibrating survey data using iterative proportional fitting (raking)', *The Stata Journal* **14**(1), 22–59.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2017), 'Automatic differentiation variational inference', *The Journal of Machine Learning Research* **18**(1), 430–474.

Lauderdale, B. E. (2019), 'How YouGov's 2019 General Election model works', `https://yougov.co.uk/topics/politics/articles-reports/2019/11/27/how-yougovs-2019-general-election-model-works`.

Lauderdale, B. E., Bailey, D., Blumenau, J. & Rivers, D. (2020), 'Model-based pre-election polling for national and sub-national outcomes in the US and UK', *International Journal of Forecasting* **36**(2), 399–413.

Lauderdale, B. E. & Blumenau, J. (2019), Constructing and assessing seat level estimates. Reading the 2019 Election Polls: Event by the London School of Economics, Department of Methodology, and the British Polling Council.

Lax, J. R. & Phillips, J. H. (2009), 'How should we estimate public opinion in the states?', *American Journal of Political Science* **53**(1), 107–121.

Lax, J. R. & Phillips, J. H. (2013), How should we estimate sub-national opinion using MRP? Preliminary findings and recommendations, *in* 'annual meeting of the Midwest Political Science Association, Chicago'.

Leemann, L., Broniecki, P. & Wuest, R. (2020), 'Improved multilevel regression with post-stratification through machine learning (autoMrP)', *Working Paper* .

Leemann, L. & Wasserfallen, F. (2017), 'Extending the use and prediction precision of subnational public opinion estimation', *American journal of political science* **61**(4), 1003–1022.

Lewis-Beck, M. S. & Stegmaier, M. (2014), 'Us presidential election forecasting', *PS, Political Science & Politics* **47**(2), 284.

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

Lu, B. (2017), Constructing Prediction Intervals for Random Forests, PhD thesis, Ph. D. thesis, Pomona College.

Lu, B. & Hardin, J. (2019), 'A unified framework for random forest prediction error estimation', *arXiv:1912.07435* .

Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2012), *The BUGS book: A practical introduction to Bayesian analysis*, CRC press.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. & Ziegler, A. (2012), 'Probability machines', *Methods of information in medicine* **51**(01), 74–81.

Mayer, M. (2018*a*), 'Package `missRanger`'.

Mayer, M. (2018*b*), 'Using `missRanger`', `https://cran.r-project.org/web/packages/missRanger/vignettes/vignette_missRanger.html`.

McElreath, R. (2017), 'Markov chains: Why walk when you can flow?', `http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/`.

McLean, I. (1973), 'The problem of proportionate swing', *Political Studies* **21**(1), 57–63.

Mellon, J. & Prosser, C. (2017), 'Missing nonvoters and misweighted samples: explaining the 2015 great british polling miss', *Public Opinion Quarterly* **81**(3), 661–687.

Nielsen, D. (2016), Tree boosting with `xgboost`: why does `xgboost` win "every" machine learning competition?, Master's thesis, NTNU.

Norman, P. (1999), 'Putting iterative proportional fitting on the researcher's desk'.

Oh, H. & Scheuren, F. (1978), 'Some unresolved application issues in raking ratio estimation', *1978 Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC* pp. 723–728.

Ornstein, J. T. (2017), 'Subnational Public Opinion Estimation Using MrsP', *Working Paper* .

Ornstein, J. T. (2019), 'Stacked regression and poststratification', *Political Analysis* pp. 1–9.

Park, D. K., Gelman, A. & Bafumi, J. (2004), 'Bayesian multilevel estimation with poststratification: State-level estimates from national polls', *Political Analysis* **12**(4), 375–385.

Pasek, J. & Pasek, M. J. (2019), 'Package `anesrake`'.

Ramosaj, B. & Pauly, M. (2017), 'Who wins the miss contest for imputation methods? our vote for miss boopf', *arXiv:1711.11394* .

Richardson, S., Jackson, C. & Best, N. (2005), Bayesian hierarchical models in ecological studies, *in* 'Invited Paper for the International Workshop for Statistical Modelling, Sydney'.

Rifkin, R. & Klautau, A. (2004), 'In defense of one-vs-all classification', *Journal of machine learning research* **5**(Jan), 101–141.

Ripley, B. D. (2007), *Pattern recognition and neural networks*, Cambridge university press.

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, Vol. 81, John Wiley & Sons.

Sabbagh, D. (2019), 'The MRP poll predicts a big win for the Tories - is it right?', `https://www.theguardian.com/politics/2019/nov/27/mrp-poll-conservatives-on-course-for-68-seat-majority`.

Schafer, J. L. & Olsen, M. K. (1998), 'Multiple imputation for multivariate missing-data problems: A data analyst's perspective', *Multivariate behavioral research* **33**(4), 545–571.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. & Hemingway, H. (2014), 'Comparison of random forest and parametric imputation models for imputing missing data using `mice`: a caliber study', *American journal of epidemiology* **179**(6), 764–774.

Si, Y. & Zhou, P. (2019), 'Bayes-raking: Bayesian Finite Population Inference with Known Margins', *arXiv:1901.02117* .

Springate, D. A., Ashcroft, D. M., Kontopantelis, E., Doran, T., Ryan, R. & Reeves, D. (2015), 'Can analyses of electronic patient records be independently and externally validated? study 2—the effect of β-adrenoceptor blocker therapy on cancer survival: a retrospective cohort study', *BMJ open* **5**(4).

Starkweather, J. (2014), 'A new recommended way of dealing with multiple missing values: Using `missForest` for all your imputation needs', *Benchmarks RSS Matters July* .

Stekhoven, D. J. (2011), 'Using the `missForest` package'.

Stekhoven, D. J. & Bühlmann, P. (2012), '`missForest` — non-parametric missing value imputation for mixed-type data', *Bioinformatics* **28**(1), 112–118.

Stone, M. (1974), 'Cross-validation and multinomial prediction', *Biometrika* **61**(3), 509–515.

Stone, M. (1977), 'An asymptotic equivalence of choice of model by cross-validation and akaike's criterion', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 44–47.

Team, R. C. et al. (2013), 'R: A language and environment for statistical computing'.

Thompson, S. K. (1987), 'Sample size for estimating multinomial proportions', *The American Statistician* **41**(1), 42–46.

Travis, A. (2017), 'YouGov's poll predicting a hung parliament is certainly brave', `https://www.theguardian.com/politics/2017/may/31/yougov-poll-predicting-hung-parliament-brave`.

Van Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical methods in medical research* **16**(3), 219–242.

Van Buuren, S. & Oudshoorn, K. (1999), *Flexible multivariate imputation by* `mice`, Leiden: TNO.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wakefield, J. (2004), 'Ecological inference for 2× 2 tables', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **167**(3), 385–425.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J. & Higgins, P. D. (2013), 'Comparison of imputation methods for missing laboratory data in medicine', *BMJ open* **3**(8).

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Warshaw, C. & Rodden, J. (2012), 'How should we measure district-level public opinion on individual issues?', *The Journal of Politics* **74**(1), 203–219.

Wells, A. (2019), 'Likelihood to Vote', `https://ukpollingreport.co.uk/faq-turnout`.

Wold, S., Esbensen, K. & Geladi, P. (1987), 'Principal component analysis', *Chemometrics and intelligent laboratory systems* **2**(1-3), 37–52.

Wright, M. N., Wager, S., Probst, P. & Wright, M. M. N. (2017), 'Package `ranger`'.

Wright, M. N. & Ziegler, A. (2015), '`ranger`: A fast implementation of random forests for high dimensional data in C++ and R', *arXiv:1508.04409* .

Zhang, G. & Lu, Y. (2012), 'Bias-corrected random forests in regression', *Journal of Applied Statistics* **39**(1), 151–160.

Zhang, Y. & Yang, Y. (2015), 'Cross-validation for selecting a model selection procedure', *Journal of Econometrics* **187**(1), 95–112.

Appendix

A1 Approximate, Naive Bayesian Assignment of Geography

In Phase I we need a methodology to assign a PC to each census micro-data record. The more accurate the initial assignment, the less likely we are to encounter convergence problems at the raking stage, after the data integration operations. We propose to exploit the lowest level of geography which is common to both GLAs and PCs, namely the census Output Areas. There are in total 227,759 OAs; for individuals in a given GLA, only a subset of the OAs will be plausible matches - we refer to these as *candidate* OAs. We make the assumption that demographics will play a key role in the assignment: a White Muslim Man is more likely to belong to an OA which has many individuals similar to him in some respects. For each GLA, the probability of encountering a given demographic profile in any of the candidate OAs can be written as $\Pr(D = d \mid \text{OA})$, where D represents the vector of possible characteristics under study - in our application D={Sex, Age, Ethnicity, Religion, Social Grade, Employment Status, Employment Sector}, and d is any given realization of those - e.g. d={Male, 25 to 34, White, Muslim, AB, Full-time, Agriculture}. If we know the joint conditional distribution $\Pr(D = d \mid \text{OA})$, we can use Bayes theorem to uncover $\Pr(\text{OA} \mid D = d)$, the probability of belonging to any of the candidate OAs conditional on the record's characteristics, as follows:

$$\Pr(\text{OA} \mid D = d) = \frac{\Pr(D = d \mid \text{OA}) \times \Pr(\text{OA})}{\Pr(\text{D=d})}; \tag{A1}$$

here Pr(OA) is the probability distribution of belonging to one of the candidate OA, and $\Pr(D = d)$ is the joint distribution of the demographics over the entire GLA in question.

Unfortunately the joint conditional $\Pr(D = d \mid \text{OA})$ is not available to us; from the ONS[62] and the Scottish Census[63] we can get access to the relevant marginal distributions of the characteristics in D at the OA level, but the joints are not obtainable. As such, in order to solve Equation A1, we have to make a simplifying assumption and take the

[62] https://www.nomisweb.co.uk/census/2011/bulk/r2_2

[63] https://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html

demographics as independent:

$$\Pr(\text{OA} \mid D = d) = \frac{\prod_k \Pr(D_k = d_k \mid \text{OA}) \times \Pr(\text{OA})}{\prod_k \Pr(D_k = d_k)}; \tag{A2}$$

where $k = 1, ..., K$ indexes the demographic characteristics in D[64].

A final issue to deal with are missing values. Some of the above variables, such as Social Grade or Employment Sector, are collected in the census only for working-age individuals; the age-brackets can vary from '16 and above', to 'between 16 and 74'. As such they will be missing for many individuals, and cannot be used for comparison. When this is true, we calculate $\Pr(\text{OA} \mid D^\star = d^\star)$, where $D^\star \subseteq D$ and varies across individuals depending on their missing-value patterns.

We can then re-write the above GLA-specific probabilities of assignment as $\pi_o = \Pr(\text{OA} = \text{o} \mid D^\star = d^\star)$, for candidate OA index $o = 1, ..., O_g$, where O_g represents the number of candidate OAs for a given $GLA = g$. For each census individual $i = 1, ..., N_g$, resident in a given $GLA = g$ we assign an OA by sampling according to a categorical distribution with the derived probabilities:

$$\text{OA}_i \sim \text{Cat}(\pi_1, \dots \pi_{O_g}). \tag{A3}$$

Because OAs have a unique mapping to PCs, mapping to OAs automatically guarantees the desired mapping to constituencies.

Evaluating Mapping Quality. We can evaluate the quality of the OA-PC assignment strategy described above by testing whether we recover the known 2011 marginal distributions for the key variables above, in each parliamentary constituency. Figure A1 plots the known marginals against the marginals derived by summing over the assigned OAs in the census micro-data. This comparison points to a broadly successful assignment: the average correlation of our PCs derived by OA assignment, compared to the underlying truth known from the BES constituency results and candidate data[65], is 0.9 across the 43 dichotomous variables in our census micro-data, indicating the approxi-

[64] Note: the above probabilities should sum to 1 over OAs; however, because of the independence assumption, they often don't. Before re-sampling, we re-scale the above probabilities to ensure they are consistent

[65] The BES area-level dataset does not contain approximate social grade counts; it further uses a classification for our 'employment sector' variable which is different from ours. Hence we re-derive the marginal counts for these two variables by aggregating the OA values at the PC level.

mate Bayesian assignment can order the PCs accurately. The average Mean Absolute Error (MAE) per constituency is 1.65 percentage points, which is again quite low, indicative of an assignment strategy that comes close to he truth PC composition. Where the assignment comes short is on variables for which the independence assumption is clearly unsatisfactory: working full time over being retired; owning a house outright over owning with a mortgage; for these variables, the errors are around 7% - though note the correlations are well preserved. We can further ask how much accuracy in assignment is gained with this data-intensive procedure over the random assignment of OAs; if a simple random assignment outperforms, or performs similarly to, our approximate Bayesian assignment, then there is a good case to avoid collecting all the external OA-level data, and performing computationally intensive probability sampling. This comparison is available in the Appendix, where Figure A3 shows how completely random assignment of OAs performs in terms of the resulting PC composition, and Figure A2 compares he MAE and correlations from both models. The Bayesian model outperforms random assignment in both metrics: the gain in average MAE is small - 0.6%, though the Bayesian method consistently reaches gains of around 3% over random assignment on socio-economic variables, including Ethnicity, Religion, Social Grade, Tenure and Educational Attainment; the Bayesian method further enables massive accuracy gains in terms of correlation, with a difference of 0.15 correlation points on average.

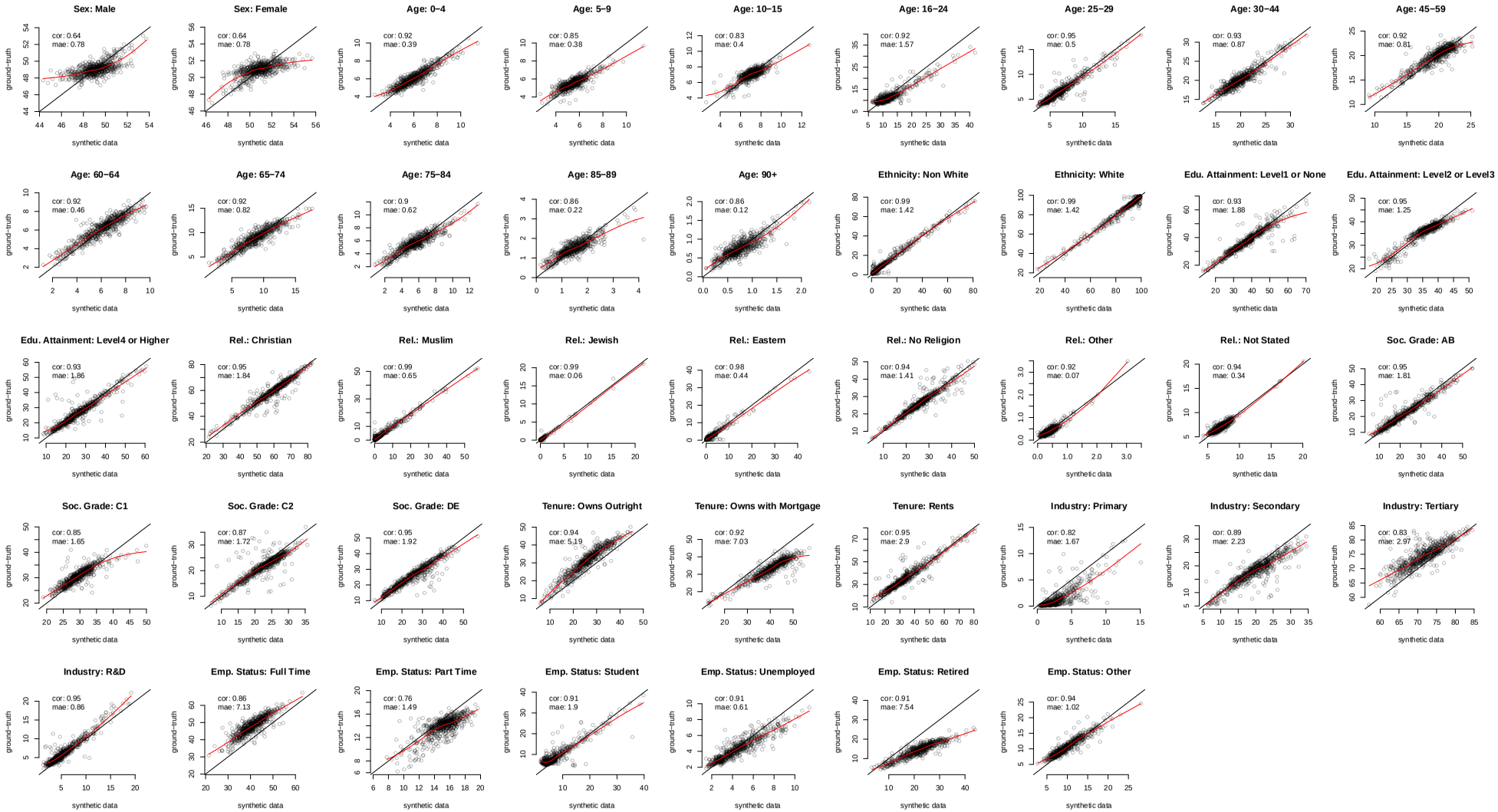


Figure A1: Comparison of known PC marginals v. marginal derived from the approximate Bayesian OA assignment. —: smoothing spline.

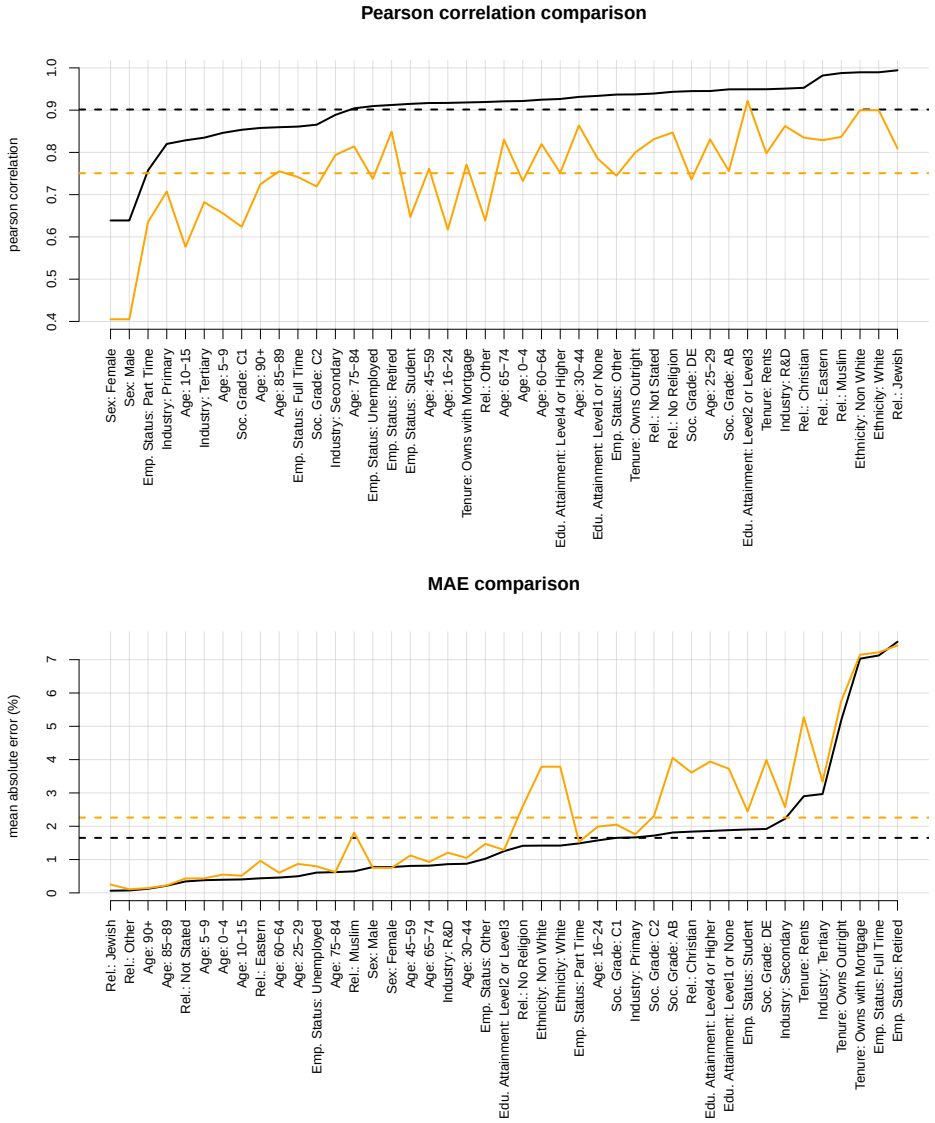


Figure A2: Comparison between PC compositions derived by approximate Bayesian mapping v. completely random assignment.

Black: Bayesian mapping **Orange**: random assignment; — : Observed; **- - -** : Mean.

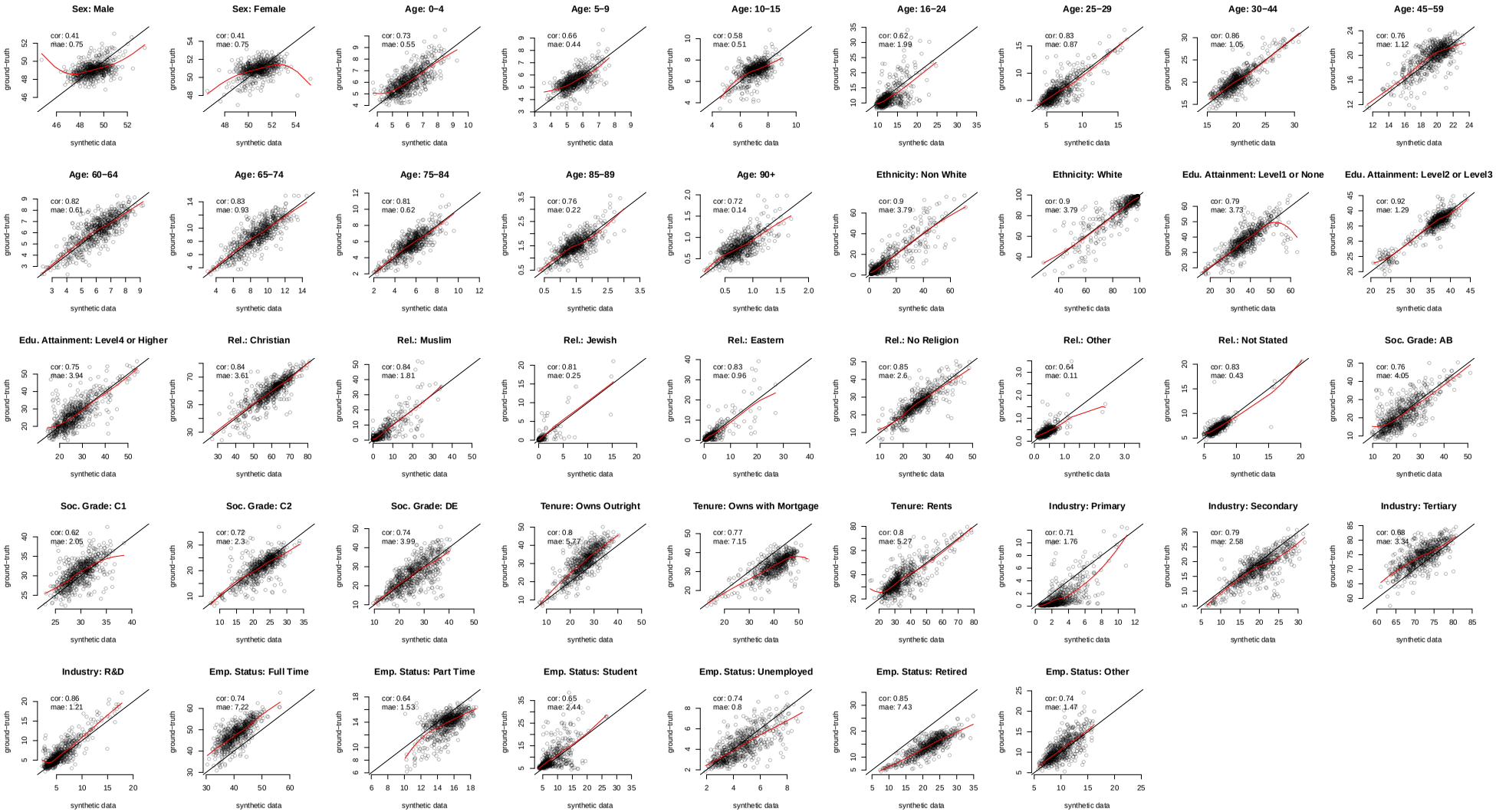


Figure A3: Comparison of known PC marginals v. marginal derived from the completely random OA assignment. —: smoothing spline.

A2 Data Integration + MICE

The *data integration* technique outlined in this section was developed by Carrig et al. (Carrig et al. 2015). Its purpose in Phase I of our framework is to enable the appending of political variables of interest, by which we want to stratify our predictions at a later stage, to the records in the census micro-data. The technique requires us to think about the augmentation of census micro-data as a missing-value problem: we can assume that the census micro-data was generated by the same data generating process as an auxiliary dataset of choice - in our case, the BES online panel wave 16, and append the auxiliary dataset to the incomplete data. The incomplete data is then characterised by a series of missing values, which are assumed to be missing at random, and are amenable to Multiple Imputation (MI).

Area-level Covariates. We can see from Table A1 a subset of the area-level predictor. We want these features to be included in the dataset and considered in the multiple imputation model to improve imputation accuracy by exploiting correlations across areas; much like in standard multilevel models, we expect the area-level covariates to selectively reduce the shrinkage effect applied by the underlying MI model. A key question is how extensive the area predictor should be - both here and later in Phase III. We count $1,024$ candidate variables (excluding PC and Region identifiers), of which 193 come from the BES results dataset[66], and are aggregated at the PC level based on the 2011 census; 756 are the most recent (June 2019 and December 2018) Annual Population Survey (APS) economic variables available; the remaining variables are political. The political variables include past election results by party - vote shares, win indicators and margins - from 2010, 2015 and 2017; the Hanretty approximation of Brexit referendum results and turnout (Hanretty 2017); the Hanretty approximation of the 2019 EU election results (Hanretty 2019*b*); and finally dummies indicating whether a given party is standing in the constituency in the upcoming election. The political variables are going to be the most interesting and predictive in our application, though it's important to also include a substantial non-political predictor to capture any residual effect, as well as novel economic conditions from more recent APS data. To these variables we add

[66]These also include PC-level Social Grade and Employment Sector variables which are re-calculated from the OA datasets.

longitude and latitude for each PC, and further add distance values in kilometers across PCs, to enable some degree of local smoothing as per Hanretty (Hanretty et al. 2018).

We seek to shrink the area-predictor for computational ease during the MI procedure; in doing so, we want to avoid making any theory-driven variable selection, and preserve as much of the information in the predictor as possible. We resolve to keep all the political variables as they are (excluding any one of perfectly co-linear candidates); we then perform a Principal Component Analysis (PCA) (Wold et al. 1987) on all the census, APS-derived, and geographic variables, and keep only the first 100 scaled principal components, effectively summarizing the non-political area predictor in 100 variables[67] in a strategy similar to that suggested by Leeman et. al (Leemann et al. 2020). Note: we perform MI on the area level separately, prior to the data integration + MI procedure, to avoid introducing nonsensical individual-level variance in the missing values of the area-level covariates.

Non-Parametric MICE. We seek to perform Multiple Imputation with Chained Equations (MICE) on the stacked dataset presented in Table A1. For an introduction to the general MICE algorithm, and a clarification on the notation we use in this paragraph, see section A3 in the Appendix.

The integrated data is awash with complexities: mixed-data types; millions of records to impute; hundreds of features to select from for outlining an imputation model; massive imbalance between missing and observed counts for auxiliary variables; unrepresentative training data. We seek a statistical solution that can leverage the MICE framework to best tackle these challenging circumstances, with computational speed fit for debugging in time-sensitive applications; `missForest` (Stekhoven & Bühlmann 2012), a MICE algorithm with leverages Random Forests (RFs) (Breiman 2001*a*) to impute missing values at each iteration, is such a solution. We will introduce RFs in some detail in Section A4.1, but for now it's sufficient to know that a RF is a non-parametric algorithm that seeks to summarize a set of weak Classification and Regression Trees (CART) (Breiman et al. 1984) to only learn from consistently predictive features of random sub-samples of a training set; in the bias-variance trade-off (Hastie et al. 2009) that characterizes learners, RFs privilege variance reduction at the cost of more biased (regularized) predic-

[67] These explain a cumulative 93% of the total variance in area-level covariates.

tions (Zhang & Lu 2012). Pseudo Algorithm A1 is a version of the original `missForest` from Stekhoven (Stekhoven & Bühlmann 2012), modified only slightly to fit with our nomenclature.

Algorithm A1 MICE with Random Forests as the Learner.

Require: $Y = \{Y^{mis}, Y^{obs}\}$ and X

define stopping criterion γ;
make initial guess for missing values: $Y_0^{\star} = \{\hat{Y}_0^{mis}, Y^{obs}\}$;
$o \leftarrow$ vector of sorted indices of columns in Y^{mis} w.r.t. increasing amount of missing values;
set counter: $j = 0$
while not γ **do**
    update counter: $j = j + 1$
    for incomplete variable: k in o **do**
        train RF: $\varphi_{j,k} = f\left(Y_k^{obs} \mid Y_{(j-1,-k)}^{\star}, X, \theta_{j,k}\right)$;
        predict $\hat{Y}_{j,k}^{mis}$ from $\varphi_{j,k}$;
        update training data: $Y_{j,k}^{\star} = \{\hat{Y}_{j,k}^{mis}, Y_k^{obs}\}$;
    end for
    update γ.
end while
return $Z = (Y_{j-1}^{\star}, X)$

The `missForest` stopping rule γ is as follows: the algorithm stops at iteration j, and returns the imputed dataset of the previous iteration $j-1$, if the standardized difference between the newly imputed missing values $\hat{Y}_{j,k}^{mis}$, and the previously imputed $\hat{Y}_{j-1,k}^{mis}$ is positive for both continuous and categorical variables; we refer to Stekhoven (Stekhoven & Bühlmann 2012, Stekhoven 2011) for a more detailed explanation of the criteria. RFs output regularized predictions, by averaging over multiple weak learners; they perform automatic variable selection, relieving us of model-selection responsibilities; they can be exceedingly fast, and easily handle large datasets, because: i) the multiple weak learners can be estimated in parallel; ii) each weak learner only deals with a small subset of features; iii) learned forests can be stored, making prediction largely a matter of subsetting the test-data. We note that `missForest` has a tried and tested track-record of accuracy in biology and related fields (Stekhoven & Bühlmann 2012, Starkweather 2014, Shah et al. 2014, Waljee et al. 2013, Hariz et al. 2017), though its application to electoral studies is relatively new (see (Cerina & Duch 2020) for one such application).

There a number of algorithms which can implement RFs on `R` (Team et al. 2013), the statistical programming language we use throughout this paper; the one acting as backbone to the traditional `missForest` is the `randomForest` (Breiman et al. 2011)

package. `ranger` (Wright & Ziegler 2015) is a novel version of this algorithm based on C++ and optimized for memory and speed, provides an intuitive `R` package (Wright et al. 2017); we implement `missForest` using this faster algorithm, via the `missRanger` package (Mayer 2018,*a*). One key difference between `missRanger` and its predecessor is with respect to the stopping rule γ: `missRanger` stops iterating when the average out-of-bag (OOB) error across all features stops improving. This stopping rule is more intuitive, and optimizes the algorithm for predictive performance over imputation-stability. `missRanger` further permits the use of predictive mean matching (PMM) to provide more realistic estimates of variance for imputed variables; avoid imputation of values which are un-observable (such as a predicted value of 0.33 for a dichotomous variable); and produce multiple imputed datasets to generate a realistic approximation of the density of each imputed value (Mayer 2018*a*). PMM works by replacing the point-estimate for each imputed variable $\hat{Y}^{mis}$ with a random sample from the m nearest neighbours of this observation in the training-set, where m is a parameter to be set by the user. Considering our stacked training set (the stacked BES wave 16 and the 2015 and 2017 face-to-face surveys) has over $50,000$ observations, we set $m = 30$; this is considered large-enough to provide the PMM algorithm with a wide-enough choice of neighbours for each observation in the census micro-data, but small enough to make the imputation still meaningfully cluster around a given point-estimate.

Finally, we grapple with the issue of imbalance between the millions of completely missing values for political variables in the census micro-data, and the mere tens of thousands of records at our disposal for training. `missRanger` will begin by assigning a random draw from the marginal distribution of each of the incomplete variables, to each of the missing values. In typical circumstances, where the percentage missing in a given variable will be between 0 and 30%, the initial value does not matter, as provided the usual conditions are met, after a few iterations the algorithm will converge to the joint distribution of the observed data. However, if the proportion of observed values is less than 1%, as is the case in the integrated data, it is highly likely that the algorithm will either fail to converge, or converge to noise, as the initial values are not informative of the full conditional distributions. We could opt to set sparse imputation-weights for the integrated dataset; say $\iota = \{\iota_{census}, \iota_{auxiliary}\}$ is a vector of weights for each record in the integrated dataset, and it takes value zero for all records in the census micro-data:

$\iota_{census} = 0$; and whatever demographic and/or selection weight associated to the record for the auxiliary samples: $\iota_{auxiliary} = w_{auxiliary}$. Practically, the weights act on the RF procedure by re-defining the probability that each record is used to train each of the weak CART learners that make up the forest; `missRanger` hence normalizes the weights to sum to one, thereby excluding the incomplete census micro-data from the training data, and assigning population-representative probability to observations from the auxiliary dataset. One issue with this method is that the training data, though rich in information for political variables, are completely missing for some of the census variables - such as Employment Sector and Employment Status, which we could not reconcile between census and BES answers. Hence if we leave out the census all together in training, we could be imputing nonsensical values (usually completely at random) for the missing values of variables which are present in the census, but not in the training data. To resolve this we amend weight vector ι to reflect some prior belief about the influence we want each dataset to have on the imputation of a missing value; hence we set $\iota_{census} = \frac{1}{N_{census}} N_{auxiliary}$. This setup then encodes the following behaviour: trees grown on variables which are present in both datasets will have an equal probability of training on either, effectively smoothing predictions; trees grown to predict variables which are missing in one dataset, but not in the other, will be trained fully on the non-missing dataset, and vice-versa.

Evaluating Imputation Quality. To evaluate the quality of the imputation on the integrated data we can leverage the OOB errors estimated for each random forest. `missRanger` automatically calculates the following OOB metrics: for continuous variables, the normalized root mean squared error (NRMSE); for categorical features, the proportion of falsely classified (PFC):

$$\text{NRMSE} = \sqrt{\frac{\text{mean}((Y^{oob} - \hat{Y}^{oob})^2)}{\text{var}(Y^{oob})}}; \tag{A4}$$

$$\text{PCF} = \frac{\text{n. incorrectly classified}}{\text{n. OOB obs}}; \tag{A5}$$

where the *oob* superscript indicates that these metrics are computed strictly for the OOB observation of each forest, and 'mean' and 'var' indicate the empirical mean and variance functions. In the context of the integrated data, it's impossible to measure

imputation accuracy of the political variables for the census micro-data, as there are no OOB observations from this set; we can however estimate how well the algorithm performs on the auxiliary samples, and use this to set our expectations as to performance on the census micro-data. The opposite is true for variables which exist in the census but which we could not reconcile to the BES data.

We set hyper-parameters to the default given by `missRanger` - hence use about $\frac{2}{3}$ of the records and $\sqrt{p}$ of the variables to build each tree, where p is the total number of candidate covariates (more on this is Section A4.1); the number of trees is set to 500, a number believed to be large enough to secure the benefits of bootstrap aggregation. Finally, we set case weights according to ι described above. The algorithm takes around 15 hours to converge after 5 iterations, on a `r4.4xlarge` AWS EC2 instance[68]. Figure A4 displays the OOB errors over all imputed features, along with the percentage missing of each variable in each of the integrated datasets, for a single run of the `missRanger` on the full stacked dataset.

The OOB error for political variables is around 5%, with 'Brexit Vote 2016' and 'Political Attention' being particularly well predicted, with OOBs below 1%. In general, the OOB error is very small, suggesting the imputation is high quality. Note that this does not mean our resulting complete, integrated synthetic dataset is represnetative: low average OOB means that our synthetic records could have plausibly come from the population of auxiliary datasets, under the imputation model; it does not imply that there are plausible relative to a ground-truth. We will have to re-weight these observations to obtain a better approximation of the underlying but unobserved 2019 census population. What the imputation does accomplish however is to better approximate the true joint distribution of the variables, which is only marginally accounted for via other methods such as MrsP (Leemann & Wasserfallen 2017).

A3 Multiple Imputation via Chained Equations

We borrow some useful notation from Van Buuren (Van Buuren 2007): $Y_1, ..., Y_K$ is a matrix of K incomplete variables, with N records each, such that $Y_{i,k}$ is the i^{th} record of the k^{th} variable. Each of these incomplete variables has a portion of records which is

[68] Details on Instance types are available on the AWS website: `https://aws.amazon.com/ec2/instance-types/`

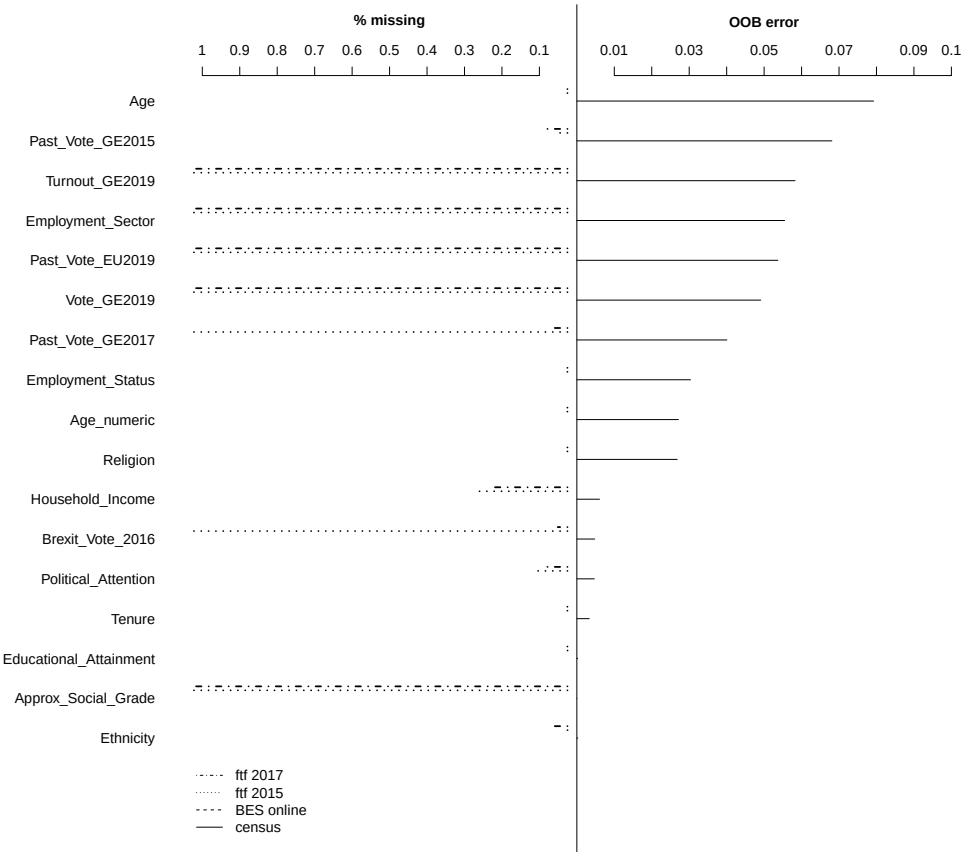


Figure A4: Left panel: percentage missing observations by dataset source. Right panel: out-of-bag error in the aggregated dataset. Note that the variable 'Age numeric' from the auxiliary datasets is included in the imputation merely to help with the imputation of the variable 'Age', which is categorical.

missing, denoted as Y_k^{mis}, and a portion which is observed, Y_k^{obs}. We use Y_{-k} to indicate the matrix of $K-1$ incomplete variables, excluding Y_k. We further let $R_{i,k} = \{1, 0\}$ be a binary indicator of whether a given record is observed or missing, and we let X be $1 \times P$ matrix of fully observed covariates. Finally, $\Theta = \theta_1, ..., \theta_k$ is a list of variable-specific vectors of parameters which drive the data generating process of Y.

As laid out by Rubin (Rubin 2004), the work of an imputation algorithm can be broken down in three distinct phases: *modeling*, which involves the specification of a joint model for $\Pr(Y, X, R \mid \Theta)$; *imputation*, dedicated to deriving the posterior predictive distribution $\Pr(Y^{mis} \mid Y^{obs}, X, R, \Theta)$ of the missing values given the observed data; and *estimation*, which comes down to estimating the parameters Θ so that we may draw samples from them.

We focus briefly on the modeling stage to give an account as to why the so called Full Conditional Specification (FCS) (Van Buuren 2007) may be desirable in a joint-modeling task. The imputation model must be able to reliably approximate the joint likelihood of the incomplete data. Take for instance the variables 'turnout' and 'vote-choice'; both are missing in the census sample, but we know the likelihoods of vote-choice among voters and non-voters and significantly different, and vice-versa; independent models won't do. Ideally we would want to leverage the joint distribution of the two variables, but that generally requires specifying a parametric model linking all the variables which could contribute to this in a joint model, similarly to what is done in full-Bayesian specifications such as those presented in the `BUGS` book (Lunn et al. 2012). These models are highly informative and theory-driven, as well as interpretable, but carry non-ignorable drawbacks in dealing with mixed-data types, can be exceedingly slow, and can be highly sensitive to model specification. We therefore turn to FCS, popularized by Van Buuren under the acronym MICE - Multiple Imputation via Chained Equations; these methods are widely used via the homonymous package (Buuren & Groothuis-Oudshoorn 2010). The basic MICE theory and algorithm is described in what follows.

MICE relies largely on a Gibbs sampler-like estimation paradigm where, provided some regularity conditions are met (Van Buuren & Oudshoorn 1999, Schafer & Olsen 1998, Rubin 2004), sampling from full-conditional distributions for a long-enough number of iterations guarantees that future samples are as-if sampled from a joint distribution of interest. The fundamental idea is that we can obtain $\Pr(Y, X, R \mid \Theta)$

by iterating over conditional models $\Pr(Y^{mis} \mid Y^{obs}, X, R, \Theta)$; Van Buuren proposes a convenient specification of this conditional model: $\Pr(Y_k \mid Y_{-k}, X, R, \theta_k)$. Given we never observe $\Pr(Y \mid X, R = 0, \Theta)$, we have to make the *ignorability* assumption $\Pr(Y|X, R = 0, \Theta) = \Pr(Y|X, R = 1, \Theta)$ to make inference over the missing data. We present the general MICE procedure in pseudo Algorithm A2, similarly to Stekhoven (Stekhoven & Bühlmann 2012), to emphasize the computational component.

Algorithm A2 A general version of the MICE algorithm.

Require: $Y = \{Y^{mis}, Y^{obs}\}$ and X;

make initial guess for missing values: $Y_0^\star = \{\hat{Y}_0^{mis}, Y^{obs}\}$;
for iteration: $j = 1, 2, \ldots J$ **do**
    for incomplete variable: $k = 1, 2, \ldots, K$ **do**
        draw $\theta_{j,k}$ from $\Pr\left(\theta_k \mid Y^\star_{(j-1,-k)}, X\right)$;
        draw $\hat{Y}_{j,k}^{mis}$ from $\Pr\left(Y_k \mid Y^\star_{(j-1,-k)}, X, \theta_{j,k}\right)$;
        update training data: $Y^\star_{j,k} = \{\hat{Y}_{j,k}^{mis}, Y_k^{obs}\}$
    end for
end for
return $Z = (Y_J^\star, X)$

The number of iterations J has to be chosen such that *convergence* of the algorithm is ensured; typically, between 10 and 20 iterations ensure convergence, provided the regularity conditions are met (Van Buuren 2007). The estimation of the parameters Θ is fully Bayesian in the general MICE framework, which enables the user to simulate the full probability distribution of the missing values, hence preserving uncertainty where needed. A version of this algorithm is implemented in the `mice` package (Buuren & Groothuis-Oudshoorn 2010).

ID	Weights	source	PCON	RGN	Sex	Age	Ethnicity	Edu. Att.	⋯	V_GE2019	T_GE2019	Pol. Att.	GE2017	⋯	W17Con	Con17	⋯	PC1	⋯	PC100
1	NA	census	E14000530	8. South East	2. Female	15. 70-74	1. White	3. Level4 or Higher	⋯	NA	NA	NA	NA	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
2	NA	census	E14000530	8. South East	2. Female	11. 50-54	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
3	NA	census	E14000530	8. South East	1. Male	4. 16-18	2. Non White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
4	NA	census	E14000530	8. South East	2. Female	6. 25-29	1. White	2. Level2 or Level3	⋯	NA	NA	NA	NA	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
5	NA	census	E14000530	8. South East	2. Female	9. 40-44	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋯	⋮
2533871	NA	census	W07000080	10. Wales	1. Male	12. 55-59	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
2533872	NA	census	W07000080	10. Wales	1. Male	6. 25-29	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
2533873	NA	census	W07000080	10. Wales	1. Male	9. 40-44	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
2533874	NA	census	W07000080	10. Wales	1. Male	14. 65-69	1. White	1. Level1 or None	⋯	NA	NA	NA	NA	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
2533875	NA	census	W07000080	10. Wales	1. Male	6. 25-29	1. White	2. Level2 or Level3	⋯	NA	NA	NA	NA	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
1	0.9604524	online	E14000530	8. South East	2. Female	11. 50-54	2. Non White	3. Level4 or Higher	⋯	conservative	1	mid	conservative	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
2	1.0844278	online	E14000530	8. South East	2. Female	13. 60-64	1. White	3. Level4 or Higher	⋯	ukip	0	NA	conservative	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
3	1.2815808	online	E14000530	8. South East	1. Male	9. 40-44	2. Non White	2. Level2 or Level3	⋯	green	1	high	labour	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
4	0.9007911	online	E14000530	8. South East	2. Female	5. 19-24	2. Non White	2. Level2 or Level3	⋯	conservative	1	mid	conservative	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
5	1.0400663	online	E14000530	8. South East	2. Female	8. 35-39	2. Non White	2. Level2 or Level3	⋯	NA	1	mid	conservative	⋯	1	55.05618	⋯	-19.36410028	⋯	-1.2363376
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋯	⋮
43136	0.4657323	online	W07000080	10. Wales	2. Female	15. 70-74	1. White	3. Level4 or Higher	⋯	conservative	1	NA	conservative	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
43137	0.3864105	online	W07000080	10. Wales	2. Female	8. 35-39	2. Non White	3. Level4 or Higher	⋯	plaid cymru	1	high	plaid cymru	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
43138	0.7175755	online	W07000080	10. Wales	2. Female	7. 30-34	2. Non White	2. Level2 or Level3	⋯	labour	1	mid	labour	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
43139	1.2089341	online	W07000080	10. Wales	1. Male	7. 30-34	2. Non White	3. Level4 or Higher	⋯	liberal democrat	0	mid	labour	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
43140	0.4775328	online	W07000080	10. Wales	1. Male	16. 75-79	1. White	1. Level1 or None	⋯	conservative	1	NA	conservative	⋯	0	30.19158	⋯	-9.24119809	⋯	-0.8319513
1	0.5681950	ftf_2015	E14000531	5. West Midlands	1. Male	10. 45-49	2. Non White	2. Level2 or Level3	⋯	NA	NA	mid	NA	⋯	1	65.40823	⋯	-7.59031051	⋯	-0.5247825
2	1.1363900	ftf_2015	E14000531	5. West Midlands	1. Male	6. 25-29	2. Non White	2. Level2 or Level3	⋯	NA	NA	low	NA	⋯	1	65.40823	⋯	-7.59031051	⋯	-0.5247825
3	1.1363900	ftf_2015	E14000531	5. West Midlands	1. Male	9. 40-44	2. Non White	3. Level4 or Higher	⋯	NA	NA	mid	NA	⋯	1	65.40823	⋯	-7.59031051	⋯	-0.5247825
4	1.7045840	ftf_2015	E14000531	5. West Midlands	1. Male	12. 55-59	2. Non White	1. Level1 or None	⋯	NA	NA	mid	NA	⋯	1	65.40823	⋯	-7.59031051	⋯	-0.5247825
5	1.7045840	ftf_2015	E14000531	5. West Midlands	1. Male	9. 40-44	2. Non White	2. Level2 or Level3	⋯	NA	NA	low	NA	⋯	1	65.40823	⋯	-7.59031051	⋯	-0.5247825
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋯	⋮
2983	1.7045840	ftf_2015	W07000079	10. Wales	1. Male	5. 19-24	2. Non White	2. Level2 or Level3	⋯	NA	NA	low	NA	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2984	0.5681950	ftf_2015	W07000079	10. Wales	2. Female	18. 85-89	2. Non White	1. Level1 or None	⋯	NA	NA	low	NA	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2985	1.1363900	ftf_2015	W07000079	10. Wales	1. Male	13. 60-64	2. Non White	1. Level1 or None	⋯	NA	NA	low	NA	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2986	0.5681950	ftf_2015	W07000079	10. Wales	2. Female	5. 19-24	2. Non White	2. Level2 or Level3	⋯	NA	NA	low	NA	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2987	1.1363900	ftf_2015	W07000079	10. Wales	2. Female	13. 60-64	2. Non White	1. Level1 or None	⋯	NA	NA	high	NA	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
1	1.1219637	ftf_2017	E14000532	2. North West	2. Female	15. 70-74	2. Non White	1. Level1 or None	⋯	NA	NA	high	conservative	⋯	1	51.01913	⋯	-0.01583402	⋯	0.8789959
2	1.1219637	ftf_2017	E14000532	2. North West	1. Male	16. 75-79	2. Non White	2. Level2 or Level3	⋯	NA	NA	high	conservative	⋯	1	51.01913	⋯	-0.01583402	⋯	0.8789959
3	1.1219637	ftf_2017	E14000532	2. North West	1. Male	8. 35-39	2. Non White	3. Level4 or Higher	⋯	NA	NA	mid	labour	⋯	1	51.01913	⋯	-0.01583402	⋯	0.8789959
4	1.1219637	ftf_2017	E14000532	2. North West	2. Female	10. 45-49	2. Non White	2. Level2 or Level3	⋯	NA	NA	mid	conservative	⋯	1	51.01913	⋯	-0.01583402	⋯	0.8789959
5	1.1219637	ftf_2017	E14000532	2. North West	2. Female	7. 30-34	2. Non White	3. Level4 or Higher	⋯	NA	NA	high	labour	⋯	1	51.01913	⋯	-0.01583402	⋯	0.8789959
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋯	⋮
2190	0.5609818	ftf_2017	W07000079	10. Wales	2. Female	7. 30-34	2. Non White	1. Level1 or None	⋯	NA	NA	low	labour	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2191	0.5609818	ftf_2017	W07000079	10. Wales	1. Male	11. 50-54	2. Non White	3. Level4 or Higher	⋯	NA	NA	high	labour	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2192	1.1219637	ftf_2017	W07000079	10. Wales	2. Female	16. 75-79	2. Non White	1. Level1 or None	⋯	NA	NA	low	conservative	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2193	0.5609818	ftf_2017	W07000079	10. Wales	2. Female	11. 50-54	2. Non White	2. Level2 or Level3	⋯	NA	NA	high	conservative	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953
2194	1.1219637	ftf_2017	W07000079	10. Wales	1. Male	10. 45-49	2. Non White	3. Level4 or Higher	⋯	NA	NA	high	stayed home	⋯	0	29.75402	⋯	-6.77707629	⋯	0.1555953

Table A1: A subset of the full integrated dataset intended to showcase its structure. Solid horizontal lines separate the four different samples which have been stacked. The total number of records in the stacked sample, after purging records with Age 15 or lower, is $2,587,793$.

A4 Iterative Proportional Fitting

Iterative Proportional Fitting (IPF) (Deming & Stephan 1940) is a numerical algorithm that enables maximum-likelihood estimation of entries in contingency tables of varying sizes, subject to a set of marginal constraints (Norman 1999), also known as control-totals. The procedure can also acts as a weighting mechanism for the entries of a contingency table, whereby the adjusted estimates of counts represent the most likely joint distribution under the marginal constraints (Norman 1999). Johnston et al (Johnston et al. 1982, Johnston & Pattie 2000, Elff et al. 2008) first proposed IPF as an estimation algorithm for ecological inference, and then showed that the derived estimates are equivalent to those produced by a log-linear model for estimating the counts of contingency tables (King 1998). They further show the IPF estimates can be derived from a probabilistic model which assumes Laplace's indifference principle - namely a uniform prior over each micro-state that gave rise to the macro- and meso- states which we know; this property of IPF estimates fits with the indifference requirements imposed by our consideration ii above. The existence of a probabilistic generating process for these estimates implies we coul derive prediction intervals for the IPF estimates, though we do not concern ourselves with this at present: there are a myriad of unaccountable sources of variance in this framework, and there is no common data-generating process we can speak of; hence we opt to demonstrate our method by focusing on pont estimates, a feat not totally uncommon in the machine learning literature. Future work should shed light on how to quantify uncertainty in a manner that preserves the utility of the predictions in MrP settings. Though IPF estimates correspond to those produced by a well defined statistical model, they do not require the estimation of any of the model parameters, as the adjusted entries of the contingency table of interest can be imputed directly via the iterative algorithm. In this sense, we can consider this approach non-parametric. The avoidance of parametric estimation carries significant decreases in computational costs.

The Algorithm. Algorithm A3 presents the procedure for the simple case of a 2×2 contingency table. Stopping criterion ψ is theoretically set to realize when convergence is reached, i.e. $\sum_j \hat{p}_{i,j,k} = Q_i$ and $\sum_i \hat{p}_{i,j,k} = Q_j$. In practice (Kolenikov 2014), the stopping rule is generally chosen according to:

a) a function of iteration k:

if $k > m \rightarrow$ **stop**;

where m is some limiting iteration number - a hyper-parameter set by the researcher,

b) a function of the largest change in weights per iteration:

if $\delta_k <= \delta^\star \rightarrow$ **stop**,

where $\delta_k = \max \mid \frac{\hat{p}_k}{\hat{p}_{k-1}} - 1 \mid$ and $\delta^\star$ is some threshold hyper-parameter. Note that δ can be used to detect non-convergence:

for $k > 1 \rightarrow$ **if** $\delta_k > \delta_{k-1} \rightarrow$ **stop**;

c) a function of the discrepancy between the control-totals and the re-weighted observed totals:

if $\Delta_{i,k} > \Delta^\star$ **or** $\Delta_{j,k} > \Delta^\star \rightarrow$ **stop**,

where $\Delta_{i,k} =\mid Q_i - \sum_j p_{i,j,k} \mid$ and $\Delta_{j,k} =\mid Q_i - \sum_i p_{i,j,k} \mid$; $\Delta^\star$ is a pre-specified tolerance level.

Any combination of these stopping rules can be simultaneously implemented in the `anesRake` (Pasek & Pasek 2019) package for `R`. Typically, the algorithm will either converge, never reach convergence, or actively diverge; warnings are usually associated with these outcomes for the benefit of the researcher.

Algorithm A3 The IPF algorithm for a 2×2 contingency table, characterized by rows i and columns j.

Require: $Q = \{Q_i, Q_j\}$ and $\boldsymbol{p}$
define stopping criterion ψ;
set counter: $k = 0$
set initial values: $\hat{p}_0 = p$
while not ψ **do**
  for control totals dimension: q in $\{i, j\}$ **do**
    update counter: $k = k + 1$;
    for row: i in $1, ..., I$ **do**
      for column: j in $1, ..., J$ **do**
        $\hat{p}_{i,j,k} = \frac{\hat{p}_{i,j,k-1}}{\sum_{-q} \hat{p}_{i,j,k-1}} \times Q_q$;
      end for
    end for
    update ψ.
  end for
end while
return $\hat{\boldsymbol{p}}_k$

When the entries of the contingency table are completely unknown, the algorithm can be used to perform ecological inference (Johnston et al. 1982, Johnston & Pattie 2000, Elff et al. 2008), by initializing each cell of the contingency table with a unit-value, and having them converge to the maximum-likelihood estimate under the constraints. Often we will already have entries in the table, along with some known marginal distributions derived from different datasets; here the algorithm can be used for data-integration via re-weighting (Norman 1999, Kolenikov 2014). A useful property of the resulting weights is that they do not alter the odds-ratios - and hence the interaction structure - of the original table (Norman 1999, Springate et al. 2015). This is useful in our application as we have spent quite some effort in getting credible estimates for the interaction structure of the synthetic population data. IPF is known to converge to the maximum likelihood estimates, provided some regularity conditions are met (Fienberg et al. 1970). Battaglia et al. (Battaglia et al. 2009) provide a practical guide to aid convergence in raking problems. They enumerate the following as potential reasons for non-convergence: i) some cells are empty (equal to zero) - this actively ensures non-convergence in most cases (Oh & Scheuren 1978); ii) imposing too many marginal constraints; iii) imposing contradictory constraints; iv) imposing constraints according to 'small' categories - i.e. categories which make up a small proportion of the population; v) imposing constraints on categories for which we have very few survey respondents - e.g. fewer than 5 respondents; vi) if differences between the control-totals and the marginals calculated from the data to be re-weighted are large prior to raking; vii) if the correlation amongst variables in the control-totals is high (Brick et al. 2003).

Raking the Synthetic Population to Known Election Outcomes. The first occasion on which we apply the IPF algorithm in the framework is in Phase I, to make the synthetic population more representative of what we think is the 2019 reference population. To recap, we find ourselves with a completed version of Table A1; because of the outdated census, selection issues with training data, potential for excessive regularization via the random forests and a small degree of mis-allocation of OAs at the PC level, this complete dataset is unrepresentative of the 2019 population. In particular, we are keen to have the synthetic micro-data match the marginal distribution of the 2016 Brexit referendum result, the 2017 General Election results, and the 2019 EU

Election results, as these are the most recent available political variables, and we believe these to be highly important a-priori. Using IPF to ensure the marginals for these variables are respected is akin to assuming the December 2019 voting population will be composed of whichever combination of voters maximises the likelihood of observing the three past election results; hence it will be a plausible mix of the three past elections. Ideally we would weight these by some measure of demographic change, but in absence of up-to-date reliable demographic information we ignore this complexity.

Figures A9, A7 and A5 in the Appendix show the differences in composition of the PCs in our synthetic micro-data compared to the known results. Overall, the PC-level composition of voters implied by our initial guesses for these three variables, which is based exclusively on the imputation model, are moderately correlated with the observed quantities, though mean absolute errors by constituency are high on average, and display significant bias - see each set of plots for details. There are some notable patterns in the predictive accuracy of the initial guess: i) PC-level correlations for *large*, non-geographically-exclusive choices - such as '% Leave' in the 2016 referendum; '% Conservative', '% Labour', and '% Lib. Dem.' in 2017; and the '% Brexit Party' for the EU election, - are generally high, around 0.7, though average errors are high at around 10%, where this number is typically dragged upwards by bias; ii) conversely, small and geographically situated parties such as the 'SNP' and 'Plaid Cymru' are generally poorly predicted, sporting extraordinarily high errors relative to their size - '% SNP' is typically off by just under 17% - as well as relatively low correlation - typical around 0.5 for Plaid Cymru and lower for the SNP; iii) Turnout is poorly predicted in all three elections, always significantly biased upward, with correlations hovering around 0.35, and average errors from 15 to 24 percentage points.

We should underscore that these preliminary guesses, though largely underwhelming, are not unexpected or unaccounted for: we have specified many times that we know our training data is highly selected, but it is the best we could work with which was public and freely available prior to the election. We also note that our working hypothesis is that these results indicate compositional differences in the synthetic micro-data relative to the true underlying population, but our prior is that the joint-distributions are well approximated by the MICE procedure; hence we believe that by re-weighting through raking, which preserves the interactive structure whilst changing the composition of our

sample to match the desired targets, we can achieve a credible synthetic population.

There are some inherent challenges to raking this large unrepresentative sample: i) *migrant voters*: these are voters predicted to vote parties which do not have a control-total in the given constituency - e.g. we have 1,395 2017 Plaid Cymru voters in West Dunbartonshire - Scotland. These 'migrant' voters are especially a problem with the Welsh party, as there are in the training data many individuals who now say they live in Scotland or England, but who previously would have lived in Wales and voted for the Nationalists. We note this group of voters is small - only 3% of the records prior to raking. We know IPF doesn't support raking to zero, so we have to find an alternative solution: we choose to drop 'migrant' voters prior to raking. This removes the possibility of such voter types in the population, though in general they are rare and can be ignored; ii) *simultaneously raking across all* 632 *PCs*: it's computationally unfeasible to have the raking algorithm re-weight the millions of records that make up the pre-raking dataset simultaneously; moreover, including control-totals at the national level may lead to contradictions at the constituency level, and hence is something we want to avoid in order to ensure the raking algorithm converges. We opt to rake each constituency individually, and then re-weight the assigned raking weights to ensure the weighted sum of the observations in a given constituency matches the size of the known electorate[69]. This ensures the relative sizes of the constituencies are respected, whilst speeding up the algorithm and making convergence more likely; iii) *choice of weight-cap*: with `anesrake` we have the choice to set a cap parameter which limits the amount a given observation can be up- or down-weighted per iteration. This is in order to encourage stability across iterations and avoid excessive information loss by massively up-weighting a small number of observations whilst simultaneously setting all others close to zero. We set this parameter to 30 arbitrarily: the raking algorithm has to find a solution which does not change the weights of each observation in the integrated dataset by more than 30 times; this imposes very little constraining in practice, and speeds up the algorithm by reducing the number of iterations needed for convergence. Figures A10, A8 and A6 in the appendix show the marginal distributions of our synthetic population after the IPF algorithm has been applied. It is clear that the raking has worked well,

[69]Here we use the 2017 electorate as a proxy for the December 2019 electorate, hopefully without losing too much information.

though we should note it did not converge for all the constituencies; nevertheless, most of the choices are now correlated close to 1, and different close to 0, with their known marginals.

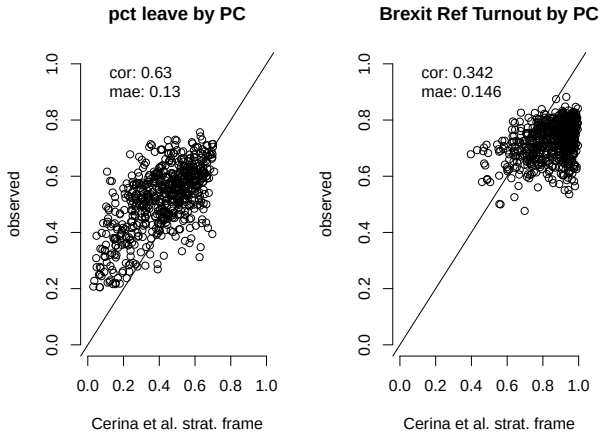


Figure A5: Pre-raking comparison in distribution of 2016 Brexit referendum results between the observed quantities and our stratification frame/synthetic population.

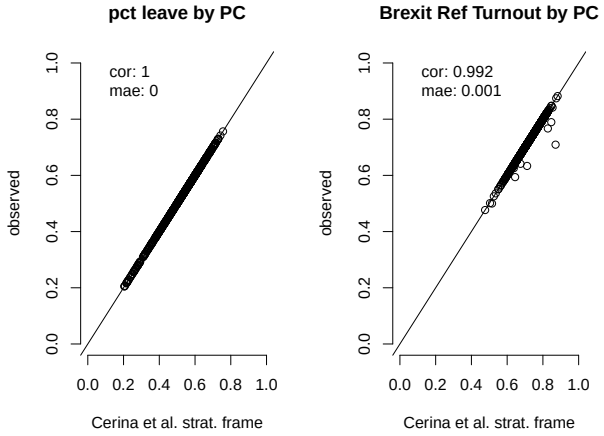


Figure A6: Post-raking comparison in distribution of 2016 Brexit referendum results between the observed quantities and our stratification frame/synthetic population.

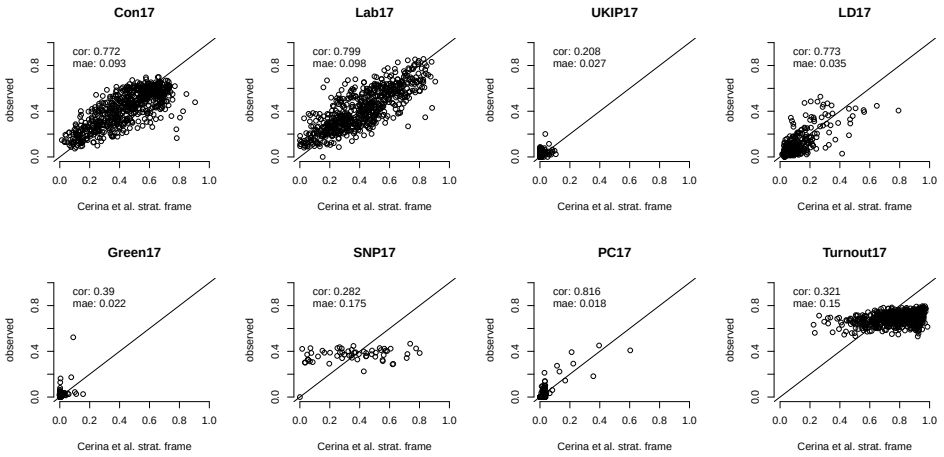


Figure A7: Pre-raking comparison in distribution of 2017 General Election results between the observed quantities and our stratification frame/synthetic population.

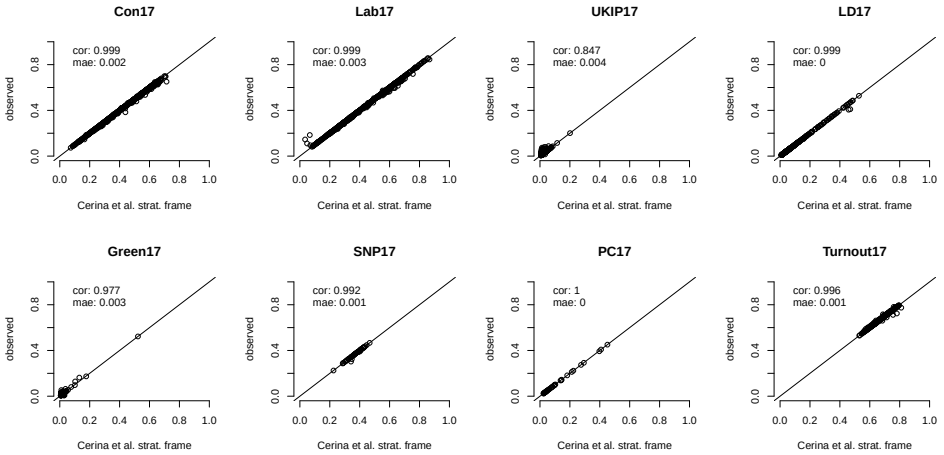


Figure A8: Post-raking comparison in distribution of 2017 General Election results between the observed quantities and our stratification frame/synthetic population.

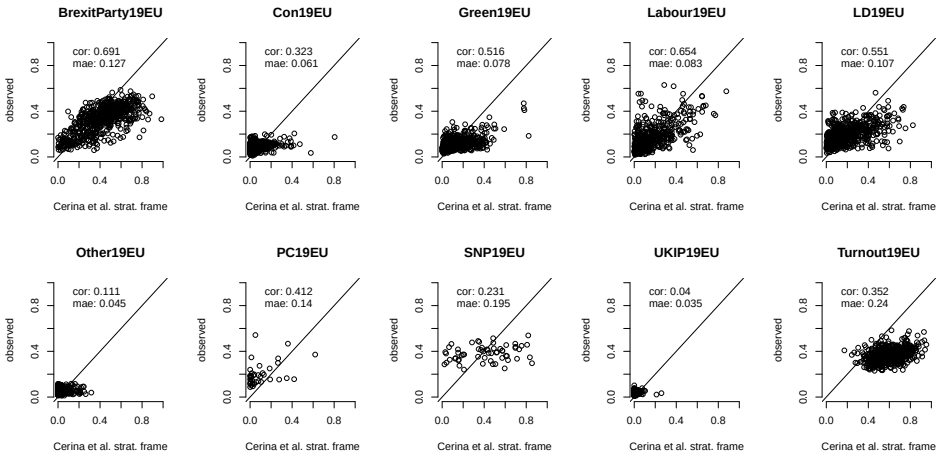


Figure A9: Pre-raking comparison in distribution of 2019 EU Election results between the observed quantities and our stratification frame/synthetic population.

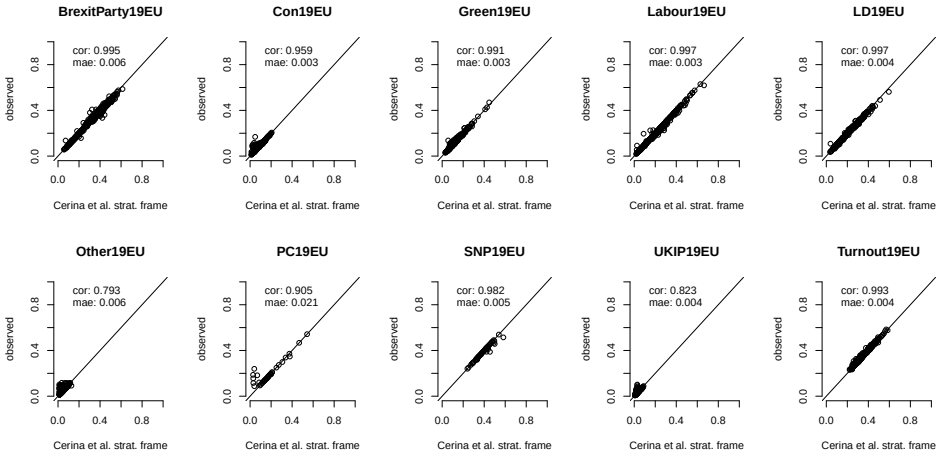


Figure A10: Post-raking comparison in distribution of 2019 EU Election results between the observed quantities and our stratification frame/synthetic population.

Issues with Raking Synthetic Samples to the Polls. The second raking effort we perform in the paper is to rake samples fro the sythetic population to the available distributions from public opinion polls. We have outlined this in some detail in the main text, but it's worth noting some of the issues we faced in executing this re-weighting step.

Figure A11 presents the sampling effort of each pollster over the election campaign. Together with table 2, it is clear we are going to have gaps - missing values - in our control-totals, over the course of the campaign: not every pollster provides us with every control-total, and pollsters are the fielding of polls is sparse over the campaign. We propose to fill the gaps by closest-in-time imputation: if voting distribution $\boldsymbol{p}_{v,h,d}$ for variable $v = 1, ..., V$, polling house $h = 1, ..., H$ and day-to-election $d = D, ..., 0$ - e.g. 2019 vote by 'Age: 25 to 34', for a YouGov poll, in-field 7 days before the election - is missing, we search for a poll released as close as possible to d, for which the 2019 vote distribution by the variable of interest is not missing, and we simply use that as a proxy.

Another problem we encounter are mis-matches between sample levels and control-totals - e.g. we may have no Plaid Cymru voters in our sample, but have control-totals for the party. Conversely, we may have UKIP voters in our sample, but 0 such voters in the control totals. We resolve by dropping all kinds of mismatches from the analysis: first we drop any control-total which is equal to zero; then we drop any control-total for which we have no levels in the sample; finally, we drop any level in the sample for which we have no control-total (say because we dropped this total as it was set to zero). The resulting sample will typically be as much as slightly smaller than that with which we started, where size reductions can get as high as 10% on 'unlucky' samples, but will be perfectly matched to the control-totals at hand. The steps necessary to create synthetic micro-data are summarized below.

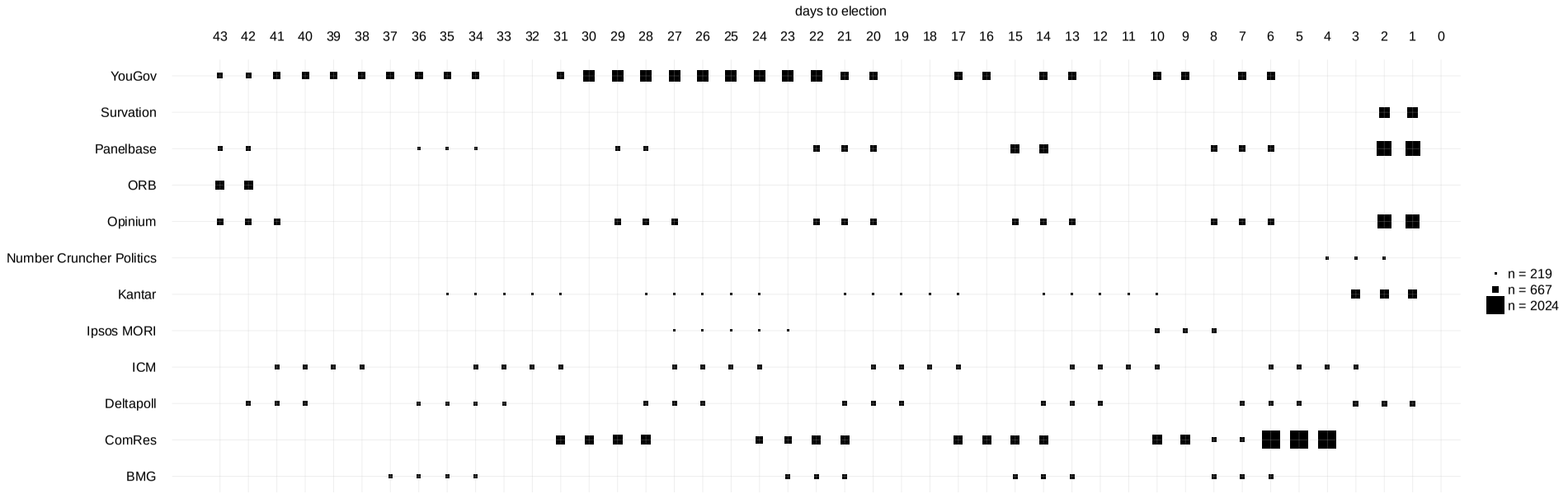


Figure A11: Sampling over the election campaign, by pollster. The sample size of each poll is divided over the in-field days, hence the points represent the average number of people surveyed per day in each poll.

- For each publicly available GB-level opinion poll from November 1^{st} to Decemeber 12^{th}:

  i. sample N likely voters from the synthetic population where N is the number of likely voters in the poll;

  ii. for each variable for which we have 2019 Vote control-totals:

    a. if control-total is missing:

    * perform closest-distribution imputation;

    b. if control-total = 0 or control-total not-in sample:

    * drop control-total;

    c. if sample levels not in control-totals:

    * drop sample levels;

  iii. rake resulting sample to resulting control-totals;

  iv. if frequency of sample levels does not perfectly match control totals:

  – rake again, exclusively to the marginal distribution of the 2019 Vote.

A4.1 Learning

In this section we present the various learning models adopted throughout the framework, which is unified by the use of tree-based learners. All tree-based learners share some basic desirable properties (Hastie et al. 2009): i) *scalability*: they are computationally smart, and able to handle large numbers of input features and records; ii) *handling mixed-data types*: they do not require cleaning of categorical data prior to fitting[70]; iii) *robustness to outliers in input space*: methods for preventing over-fitting, such as *pruning* or *bagging* are easily implemented on trees; iv) *automatic model-selection*: the tree-growing algorithm considers all features and discards those that are weakly predictive. Though simple decision trees such as CARTs (Breiman et al. 1984) can struggle in predictive power compared to other machine-learning methods, random forests (Breiman 2001*a*), and boosted trees (Friedman 2001) in particular, have contributed in making tree-based learners competitive predictive algorithms. These advantages come at the cost of a loss of interpretability relative to typical linear or semi-linear regression methods, though note that this is especially true for tree *ensembles*; standard CARTs are very intuitive and easily interpretable.

There are two major instances of supervised learning in our framework: the first is in Phase I, as we have already seen, during the multiple-imputation procedure, as we seek to learn from some BES training data the political choices of census respondents; the second is in Phase III, learning from the updated synthetic micro-data to make real-time vote-choice predictions for each individual in the synthetic population. Other instances of statistical learning in this paper include the naive-Bayes allocation of OAs to GLAs in Phase I and the raking estimates in Phases I and II; we consider these separately, as they don't fit the supervised learning framework - i.e. they don't rely on the estimation of parameters via the minimization of an loss function over a training set. We have already presented some of the properties of the random forest algorithm used in Phase I, and will formalize these further in this section; the learning objective of Phase III will be further outlined in what follows.

The rest of this section is outlined as follows: first, we provide a general definition for a *learner*; second, we present the simplest tree-learner available to us - CART;

[70] This is true of most tree-based learners, though `xgboost` requires one-hot encoding of categorical variables (Chen & Guestrin 2016).

subsequently, we proceed to outline the two learners we use throughout the framework: random forests and gradient-boosted trees. Throughout, we make heavy use of Hastie et al. (Hastie et al. 2009), and our contribution is primarily to put these methods in the context of our application.

Basic Definition of (Supervised) Learning. Following the exposition in *Elements of Statistical Learning* (Hastie et al. 2009), consider system $Y = f(X) + \epsilon$, where ϵ are stochastic perturbations to the deterministic system of inputs and outputs, such that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Voting behaviour is an example of such a system, where the response to a typical polling question - *If the election were being held today between the blues and the reds, with candidate circle representing the blues and candidate star representing the reds, who would you cast your vote for?* - is the output, and variables dealing with society, politics, demography, geography, cognition and other types belong to the set of inputs; the stochastic error accounts for random perturbations to this system. We seek to learn function f from a training set $T = (y_i, \boldsymbol{x}_i), \forall i = 1, ..., N$, where $\boldsymbol{x}_i$ is a vector of features for individual i of size p, typically a sample of of all the possible sets of inputs and outputs - in our case, often an individual-level opinion poll. As we seek to approximate f, it is useful to define some parameters θ which govern the fit of the function to the training data; these are often chosen to be useful and/or interpretable, e.g. regression coefficients in linear models, or split-points and topology of feature sub-spaces in decision trees. For a given choice of $\theta = \hat{\theta}$, we can produce predictions from our approximated function: $\hat{y}_i = f(\boldsymbol{x}_i; \hat{\theta})$; in this exposition, we will sometime refer to this as $\hat{f}_i$. The choice of parameter is generally driven by the minimization of some loss-function $\mathcal{S}(\hat{y}_i, y_i)$ quantifying the difference between predictions and training targets. The process of choosing the values of θ which minimize the loss-function within the training set is called *supervised learning*; the algorithm which we choose to implement this is called a *learner*.

Classification and Regression Trees. Decision trees for classification and regression - also known as CART (Breiman et al. 1984) - are the simplest kind of tree-based learners. A decision tree looks at the training vector $(y_i, \boldsymbol{x}_i)$, as a series of points on a euclidean space $\mathcal{R}^{p+1}$; it then considers all possible sub-spaces in the feature space $\mathcal{R}^p$, and splits this space into the most useful M sub-spaces $R_1, ..., R_M$; $R_m \subset \mathcal{R}^p$. CART

then approximate the system's functional form as:

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbb{1}(\boldsymbol{x} \in R_m); \tag{A6}$$

where c_m is a constant which optimally summarizes the region in terms of output. In regression trees, this is picked to minimize the sum of squared residuals, whilst in classification trees a measure of information gain, such as the Gini index is optimized. The result will typically be some useful summary of the output in the region - e.g. under regression trees, the mean of the training outputs within the region:

$$\boldsymbol{\iota}_m = \{\boldsymbol{i} \mid \boldsymbol{x} \in R_m\}; \tag{A7}$$

$$\hat{c}_m = \frac{1}{\sum_{i \in \boldsymbol{\iota}_m} \mathbb{1}(\boldsymbol{x}_i \in R_m)} \sum_{i \in \boldsymbol{\iota}_m} y_i. \tag{A8}$$

It is clear that CART's complexity lies in the splitting algorithm: how to choose the optimal set of subspaces $R_1, ..., R_m$. Because optimizing simultaneously across all the possible sub-spaces in $\mathcal{R}^p$ becomes quickly computationally unfeasible, CART proceeds *greedely* - i.e. in s step-wise fashion which considers a subset of the problem at every step, and finds locally optimal solutions, which are passed on as starting points to the next step. CART performs recursive binary partitioning of the spaces according to some variation on pseudo Algorithm A4.

In the case of voting behaviour, and in particular MrP applications, we are interesting the resulting probabilities of voting for a given party for cells in our stratification frame. These can be easily calculated by training separate regression trees for each party, and calculating the proportion of $1s$ in each terminal node. This is an example of the *one v. all* (Rifkin & Klautau 2004) strategy for classifying categorical outcomes or predicting choice-probabilities for each category, and will usually require normalization of the resulting scores into probabilities via either standard normalization, or a more complex multivariate link function such as *softmax* (Goodfellow et al. 2016), where the latter has more desirable properties in more complex models dealing with gradient-based optimization. Trees grown for the purpose of predicting probabilities have some special requirements, some of which are outlined by Malley's *probability machines* (Malley et al. 2012); one of the heuristics provided is requiring a minimum node-size of 10% of the

sample - to obtain more fine-grained estimates of probabilities in the region[71].

Algorithm A4 A simple greedy pseudo algorithm for finding splits - regression tree.

Require: training data: $T = (y_i, \boldsymbol{x}_i), \forall i = 1, ..., N$;
define stopping criterion ω;
set iteration counter: $k = 0$;
set root node R_0 s.t. $\sum_i \mathbb{1}(\boldsymbol{x}_i \in R_0) = N$;
while not ω **do**
    update counter: $k = k + 1$;
    for each node o in R_k **do**:
        for candidate feature $j = 1, ..., J$: **do**
            for candidate split point $s = 1, ..., S$: **do**
                split node: $\nu_1(j, s) = \{X \mid X_j \leq s\}$ and $\nu_2(j, s) = \{X \mid X_j > s\}$;
                count obs. in region: $\boldsymbol{\iota}_1 = \{\boldsymbol{i} \mid \boldsymbol{x} \in \nu_1\}$ and $\boldsymbol{\iota}_2 = \{\boldsymbol{i} \mid \boldsymbol{x} \in \nu_2\}$;
                estimate parameters: $\hat{c}_1 = \underset{i \in \boldsymbol{\iota}_1}{\text{mean}}(y)$ and $\hat{c}_2 = \underset{i \in \boldsymbol{\iota}_2}{\text{mean}}(y)$;
                measure split quality: $\text{RSS}(j, s) = \sum_{i \in \boldsymbol{\iota}_1}(y_i - \hat{c}_1)^2 + \sum_{i \in \boldsymbol{\iota}_2}(y_i - \hat{c}_2)^2$;
            end for
        end for
        select $(j = j^\star, s = s^\star)$ s.t. $RSS(j = j^\star, s = s^\star) = \min_{j,s} RSS(j, s)$;
        store node splits: $r_o = \{\nu_1(j = j^\star, s = s^\star), \nu_2(j = j^\star, s = s^\star)\}$;
    end for
    update ω;
    update $\boldsymbol{R}_k = \boldsymbol{r}$;
end while
return terminal nodes: $\boldsymbol{R}_k$.

Setting stopping rule ω is extremely important to the effectiveness of CART. Very large trees could in theory perfectly classify every training observation, whilst short 'tree stumps' could fail to learn salient features of the system. Again bringing the 2019 election in for context: a shallow tree may miss some of the more complex effects outlined by Lauderdale et al. (Lauderdale & Blumenau 2019), such as interaction effects between constituency-level variables and individual-level variables; a maximally grown tree will likely fit the residual online-selection effects in our synthetic micro-data, such that its predictions will over-state the likelihood of voting for minorities in the UK. One strategy which is applied is generally to *prune* a very large tree, penalizing complexity - a strategy not dissimilar to penalized regression - for details see (Breiman et al. 1984, Hastie et al. 2009, Ripley 2007).

Bias-Variance Trade-off. More generally, recall the bias-variance decomposition of the mean squared error, faced by predictions from a learned function $\hat{f}(\boldsymbol{x}_0)$ evaluated

[71] Though note this is in the context of an ensemble of trees, not a single regression tree.

at a training feature-set $\boldsymbol{x}_0$:

$$\begin{aligned}\text{Err}(\boldsymbol{x}_0) =& E[(Y - \hat{f}(\boldsymbol{x}_0))^2 \mid X = \boldsymbol{x}_0]; && \text{(A9)}\\ & \sigma^2 + \Big(\text{E}[\hat{f}(\boldsymbol{x}_0)] - f(\boldsymbol{x}_0)\Big)^2 + \text{E}\left[\hat{f}(\boldsymbol{x}_0) - \text{E}[\hat{f}(\boldsymbol{x}_0)]\right]^2; && \text{(A10)}\\ & \sigma^2 + \text{Bias}^2(\hat{f}(\boldsymbol{x}_0)) + \text{Var}(\hat{f}(\boldsymbol{x}_0)); && \text{(A11)}\end{aligned}$$

where the first term is the irreducible error, whilst the latter two terms are respectively the squared bias and the variance of the prediction. The *bias-variance trade-off* faced by CART is evident in the above: as the nodes get smaller, which is typically what will happen as we grow deep trees, R_m^0 can become so small as to contain a single data-point y^0; hence the bias term above will cancel out. As the number of points contained in each terminal node decreases, however, the variance across terminal nodes increases, limiting to the irreducible error, as predictions become more different. Conversely, limiting the number of terminal nodes by growing tree-stumps will increase bias, as the same constant will be used to predict large swaths of the data; a reduction in prediction variance follows. Hence un-pruned large CARTs will have a tendency to over-fit the data. Especially because our application deals with non-representative samples, and for other reasons outlined in previous sections, we need a learner that can better handle the bias-variance trade-off.

Random Forests. Beyond pruning, an alternative way to curb over-fitting is to use *bagging* (also known as *bootstrap aggregation*) (Hastie et al. 2009), whose application to CARTs was developed by Brieman (Breiman 2001*b*), and eventually gave rise to the *random forest* (RF) (Breiman 2001*a*) algorithm, which we describe below.

The RF starting point is bagging: under the assumption that training set T is a random sample of inputs and outputs of a given system, we can obtain new training data t^b for $b = 1, ..., B$ by re-sampling records with replacement from T. Typically, any t^b will contain αN records, $\alpha \in (0, 1)$. We can then fit a separate regression tree to each t, and average the predictions to obtain the bagged prediction:

$$\hat{f}_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\boldsymbol{x}_{t^b}); \qquad \text{(A12)}$$

the bagged predictions will differ from the original estimate to the extent that the learner is likely to capture highly non-linear patterns from the data (Hastie et al. 2009). Taking predictions $\hat{f}_b(\boldsymbol{x}_{t^b})$ as B identically - but not independently - distributed random variables with correlation ρ, and further assuming $\text{Var}\left(\hat{f}_b(x_{t^b})\right) = \psi^2 \; \forall \; \boldsymbol{b}$, we can calculate the variance of the average of the variables as:

$$\text{Var}(\hat{f}_{bag}(\boldsymbol{x})) = \rho\psi^2 + \frac{(1-\rho)}{B}\psi^2. \tag{A13}$$

As the number of bags B increases, the variance of the bagging estimator becomes solely dependent on the correlation amongst the predictors trained on each different bag; if these are highly correlated, the gains from bagging are modest. Therefore Breiman brings another element into the algorithm; namely, trees are grown on a random selection of $q < p$ features, typically set to $\sqrt{p}$ by default in most software, including `ranger` (Wright & Ziegler 2015). This usually induces significant reductions in ρ, which maximises the predictive power of the forests. The resulting RF prediction is then:

$$\hat{f}_{rf}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(\boldsymbol{x}_{t^b}, \Theta_b); \tag{A14}$$

where Θ_b summarizes the characteristic of each tree in terms of split variables, cutpoints at each node, and terminal-node values (Hastie et al. 2009). One of the great benefits of RFs is that the various trees which make up the forest can be grown in parallel (Wright & Ziegler 2015), generating massive gains in computational speed compared to other learners. Pseudo Algorithm A5 presents a standard RF growing procedure.

Algorithm A5 A simple pseudo algorithm for growing random forests.

Require: training data: $T = (y_i, \boldsymbol{x}_i), \forall i = 1, ..., N$;
1: **for** bags $b = 1, ..., B$ **do**
2: generate bootstrapped sample t^b:
3: i) sample without replacement $q = \sqrt{p}$ features from T;
4: ii) sample with replacement αN observations from T, $\alpha \in (0, 1)$;
5: **run** Algorithm A4 with t^b as training data;
6: **end for**
7: return ensemble of trees: $\{\boldsymbol{R}^1, ..., \boldsymbol{R}^B\}$;

It can be shown that RFs provide gains in preidctive accuracy exclusively through the variance reduction channel, as bias is typically greater than CARTs (Hastie et al. 2009). Hence, for the purpose of including tree-based learners in the MrP framework, we

can conceptualize the relationship between RFs (and pruned trees) to un-pruned CARTS as similar to that of what multilevel regression and linear regression: they enable us to regularize our predictions. We can think of the hyper-parameters that drive the degree of variance reduction - namely the number of features considered by each tree q, the relative size of sub-samples α, and the number of bags B, in a similar way as we do priors in Bayesian estimation of multilevel models, which have been commonly modified to obtain optimal results in the MrP tradition (Gao et al. 2019): we can tune the hyper-parameters to reflect our prior beliefs about the degree of regularization required, and the correlation across bagged trees. A difficulty is that these parameters typically are less interpretable a-priori compared to parametric priors, and hence are typically tuned through some sort of n-fold cross validation which requires only minimal understanding.

Amongst the advantages of RFs, beyond those common to all tree-based learners and the added gain in predictive accuracy as a result of variance-reduction, is that they have been shown to have desirable properties, such as consistent asymptotic behaviour (Wager & Athey 2018, Athey et al. 2019), and multiple well-behaved estimators of prediction uncertainty exist that do not rely on cross-validation or additional bootstrapping (Lu & Hardin 2019). We do not make use of these properties in present work, as useful uncertainty bounds for our predictions are derived exogenously, later in this section.

We use RFs as the algorithm of choice to perform multiple-imputation in Phase I. The reality is that RFs are biased, and have likely produced over-regularized estimates of missing values - a price we are practically willing to pay for ease of dealing with mixed-data types, computational advantages in dealing with million of records, and because the raking in Phase I will have corrected some of the excessive regularization at the meso-level. Boosting by recursive fitting of RFs has been proposed as a way to curb RF bias (Zhang & Lu 2012), and in particular so called *one-step boosting* - i.e. the fitting of a second RF to the residual of the first - has been shown to have desirable properties and good predictive power (Ghosal & Hooker 2018). In fact, there is preliminary empirical evidence that boosting increases the accuracy of RFs in multiple imputation tasks (Ramosaj & Pauly 2017), though to our knowledge no package is available for `R` to perform this at scale.

Boosted Trees via Approximate Gradient Optimization. We mentioned in the previous paragraph that one-step boosted random forests have shown promise in reducing the RF bias. More generally, we seek to find a learner that preserves some of the variance-reduction obtained by bootstrap aggregation, whilst tempering the relative increase in bias. Gradient-boosted decision trees provide such an attractive solution. We can examine a simple procedure to grow gradient-boosted trees in Pseudo Algorithm A6.

Algorithm A6 A simple pseudo algorithm for growing gradient-boosted trees.

Require: training data: T;
define loss function $\mathcal{S}(y_i, f(\boldsymbol{x}))$;
initialize predictions $\hat{f}_{i,0}(\boldsymbol{x}) \; \forall \; i = 1, ..., N$;
for boosting iterations $k = 1, .., K$ **do**
    for observations $i = 1, ..., N$ **do**
        compute $r_{i,k} = -\frac{\partial \mathcal{S}(y_i, \hat{f}_{i,k-1}(\boldsymbol{x}))}{\partial \hat{f}_{i,k-1}(\boldsymbol{x})}$;
    end for
    run Algorithm A4 with $\boldsymbol{r}_k$ as training outcomes;
    store terminal nodes $\{R_{k,m}\}_1^{M_k}$;
    for terminal nodes index $m = 1, .., M^k$ **do**
        compute optimal node summary $c^{\star}_{k,m}$ to minimize $\mathcal{S}(y_i, \hat{f}_{i,k-1}(\boldsymbol{x}) + c_{k,m})$;
    end for
    update $\hat{f}_{i,k}(\boldsymbol{x}) = \hat{f}_{i,k-1}(\boldsymbol{x}) + \sum_1^{M_k} c^{\star}_{k,m} \mathbb{1}(\boldsymbol{x} \in R_{k,m})$
end for
return: $\hat{f}_{i,boost}(\boldsymbol{x}) = \hat{f}_{i,K}(\boldsymbol{x})$

We focus our discussion on `xgboost` (Chen & Guestrin 2016), a framework for estimating gradient-boosted trees implemented on `R` via the homonymous package. The 'X' in the name stands for 'Extreme', which summarizes the many principled simplifications, both in terms of mathematics as well as computer-system design, which are implemented in the package, and which enable state-of-the-art speed and accuracy on massive datasets (see Nielsen (Nielsen 2016) and Chen and Guestrin (Chen & Guestrin 2016) for a summary of the accomplishments of the framework in various domains.).

The standard tree-boosting problem can be conceptualized as follows: similarly to the random forest aggregated bootstrap, we seek to learn a set of additive CART approximations to a system $f(\boldsymbol{x})$:

$$\hat{f}_{boost}(\boldsymbol{x}) = \sum_{b=1}^{B} f_b(\boldsymbol{x}, \Theta_b); \tag{A15}$$

note that f_b above is a random variable, indicating the space of all potential tree struc-

tures which may be grown. Chen and Guestrin propose that the optimal set of additive tree structures be chosen via optimization of the following objective function:

$$\mathcal{L} = \sum_i \mathcal{S}(\hat{y}_i, y_i) + \sum_b \Omega(f_b(\boldsymbol{x}, \Theta_b)); \tag{A16}$$

where $\mathcal{S}$ is a convex, differentiable loss function indicating the differences between predictions and training data. Ω is a penalizing function acting similarly to a ridge-type penalty:

$$\Omega(f_b(\boldsymbol{x}, \Theta_b)) = \gamma M_b + \frac{1}{2}\lambda \sum_1^{M_b} c_{b,m}^2; \tag{A17}$$

here M_b represents the number of terminal nodes in the tree structure indexed by b; $c_{b,m}$ is the optimal summary of leaf m in tree structure b; γ is a regularization parameter penalizing complexity as intended by tree-depth, and λ helps smooth the predictions - i.e. reduce the variance across terminal nodes.

It's unfeasible to optimize simultaneously over all candidate tree-structures; instead, they proceed greedily. Defining $\hat{y}_i^k$ as the prediction for the i^{th} record produced by the ensemble at the k^{th} boosting iteration, we seek to greedily optimize the following additive objective:

$$\mathcal{L}^k = \sum_i \mathcal{S}(y_i, \hat{y}_i^{k-1} + f_k(\boldsymbol{x}, \Theta_k)) + \Omega(f_k(\boldsymbol{x}, \Theta_k); \tag{A18}$$

$$\simeq \sum_i \left[g_i f_k(\boldsymbol{x}, \Theta_k) + \frac{1}{2} h_i f_k^2(\boldsymbol{x}, \Theta_k) \right] + \Omega(f_k(\boldsymbol{x}, \Theta_k); \tag{A19}$$

where the last line above is the second-order Taylor approximation of the objective, after removing all constants not containing optimization parameters, and where $g_i = \frac{\partial \mathcal{S}(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}^{k-1}}$ - the gradient - and $h_i = \frac{\partial \mathcal{S}(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}^{k-1}}$ - the hessian. At this point the problem becomes familiar: for each iteration we face a setup which is similar to the tree-growing problem we faced in CART. We need to specify the locally-optimal tree structure defined by $\Theta_k^\star$; the candidate structure at a given iteration is characterized by terminal nodes $\{R^k\}_1^{M_k}$, which define partitions of the predictor space implied by the candidate tree, such that we can identify observations $I_m^k = \{i \mid \boldsymbol{x}_i \in R_m^k\}$ which fall in a given terminal node. Recall that for simple regression trees, the optimal predictor at each terminal node was simply the average of the observations in that node; for boosted trees, the

optimal predictor for each terminal node can be shown to be:

$$c^{\star}_{k,m} = -\frac{\sum_{i \in I^k_m} g_i}{\sum_{i \in I^k_m} h_i + \lambda}. \tag{A20}$$

`xgboost` needs a metric to adjudicate between tree structures. For each iteration, the optimal tree structure $f_k(\boldsymbol{x}, \Theta^{\star}_k)$ to greedily add, is that which can split the euclidean space of predictors into regions $\{R^{\star,k}\}_1^{M^{\star}_k}$ which produce optimal values $c^{\star}_{k,m}$, such that the following score-function is minimized:

$$\mathcal{L}^k = -\frac{1}{2}\sum_{m}^{M^k} \frac{(\sum_{i \in I^k_m} g_i)^2}{\sum_{i \in I^k_m} h_i + \lambda} + \gamma M^k. \tag{A21}$$

Note that in practice it is not feasible to evaluate all possible tree structures to find $\Theta^{\star}_k$, so we proceed greedily, as per CART, evaluating the benefits of further splitting candidate trees; instead of using the sum of the squared loss for the proposed nodes, as was the case for simple regression trees, we use the above scoring function to measure information gain for each split; see Chen and Guestrin (Chen & Guestrin 2016) for more details.

We have thus provided the very basic knowledge of `xgboost` that we need in order to tune the hyper-parameters and interpret software output. There are however numerous other intricacies which make `xgboost` a remarkable framework; we will name a few and and refer back to the original paper for more details: i) *fast approximate greedy split-finding for large data* - the algorithm performs a greedy exhaustive search for optimal splits when the this is computationally feasible; when it is not, it is able to detect such instances and implement a *weighted quantile sketch* method to propose split-points, which minimizes the number of relevant split points to test; moreover, when the dataset is massive, the algorithm diffuses operations such as sorting and calculating histograms over multiple cores, which makes optimal use of parallel processing; ii) *sparsity aware split-finding* - `xgboost` is able to handle missing values in the predictors by performing an initial check to evaluate whether assigning the outcome with missing-predictor to the left or right candidate nodes provides maximum information gain, and then sticking to this assigning as a 'default' direction in subsequent splits; iii) *memory optimization* - makes use of fast cache memory for computing gradient, hessian and split-scores, and

uses *sharding* to deal with large datasets when multiple hard-drives are available.

A5 Area-level Results

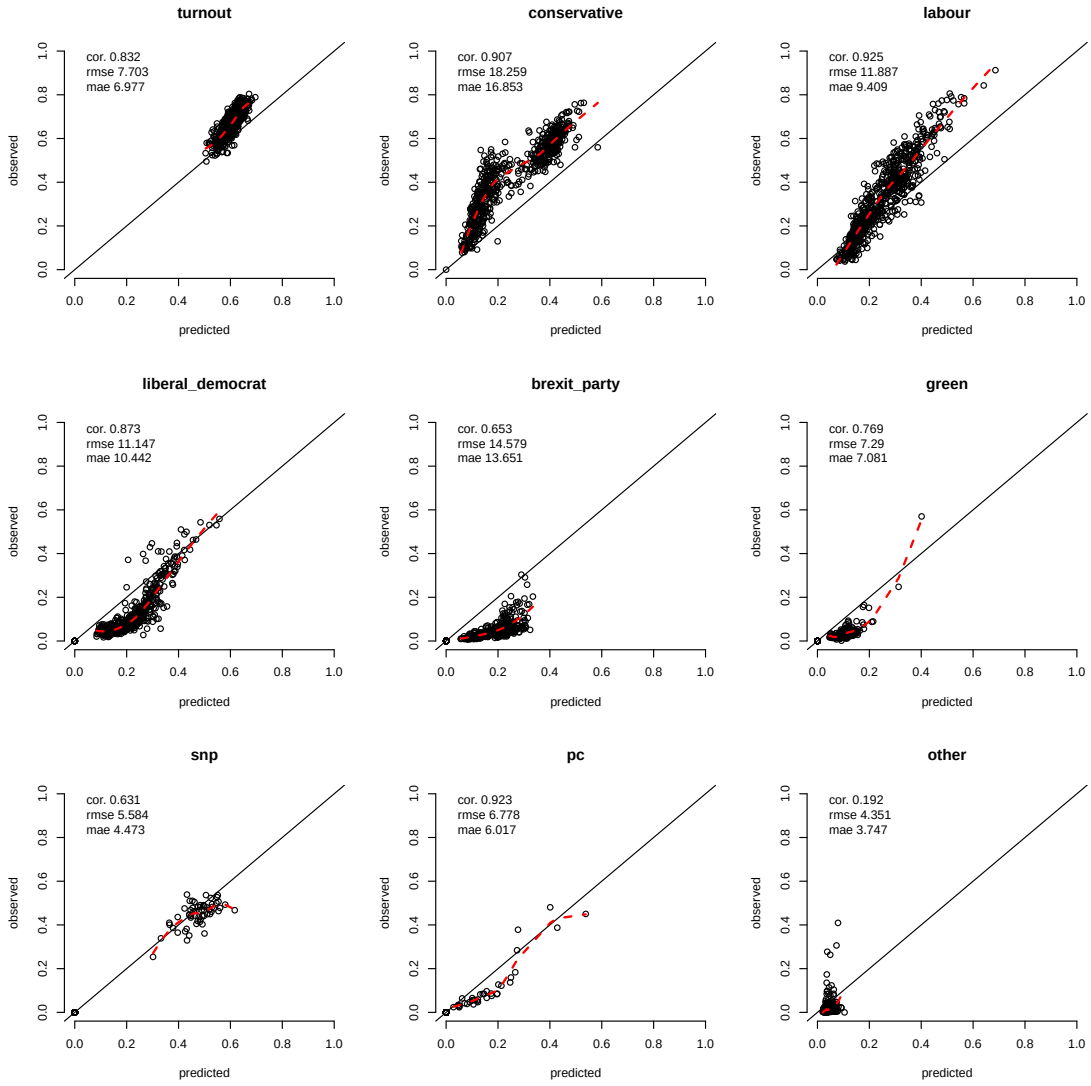


Figure A12: Area-level predictions from the model trained exclusively on data from the June BES wave.

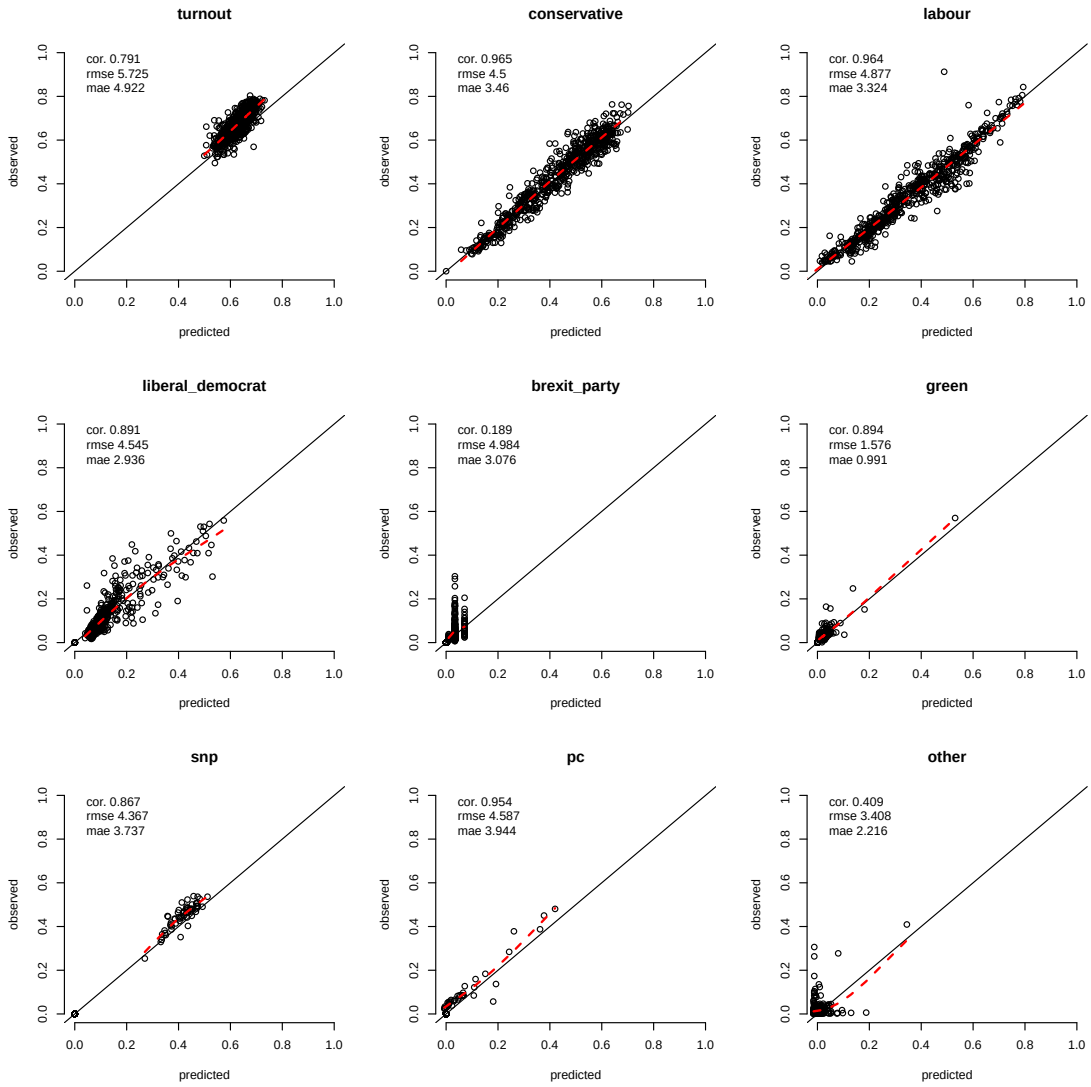


Figure A13: Area-level predictions from the uniform swing model trained on the election-week polling average. Note here that the turnout plot is simply included for consistency, and it shows the plot for predictions fit to the best performing turnout model.

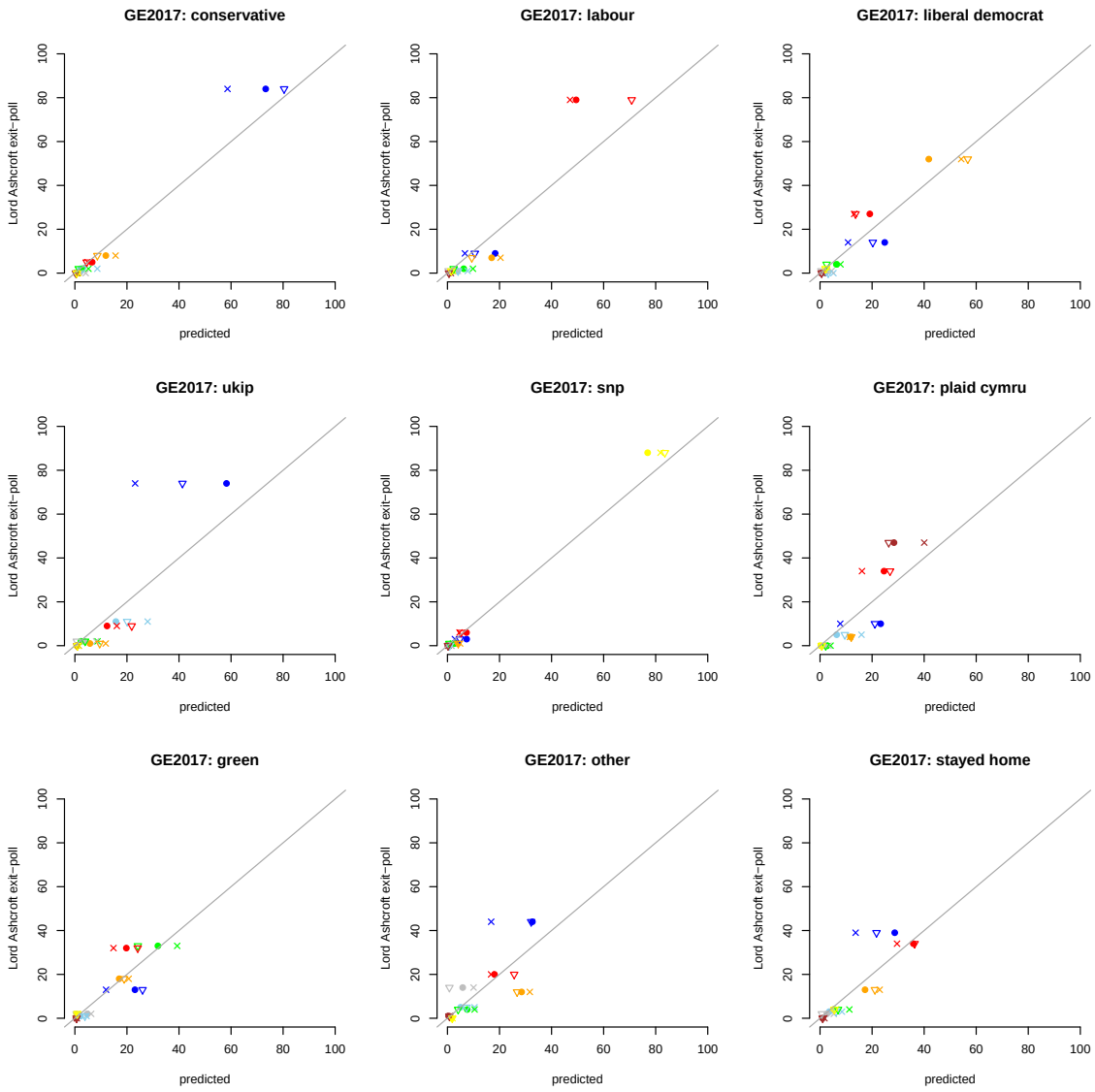


Figure A14: Vote-choice prediction-accuracy by 2017 Vote Choice.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

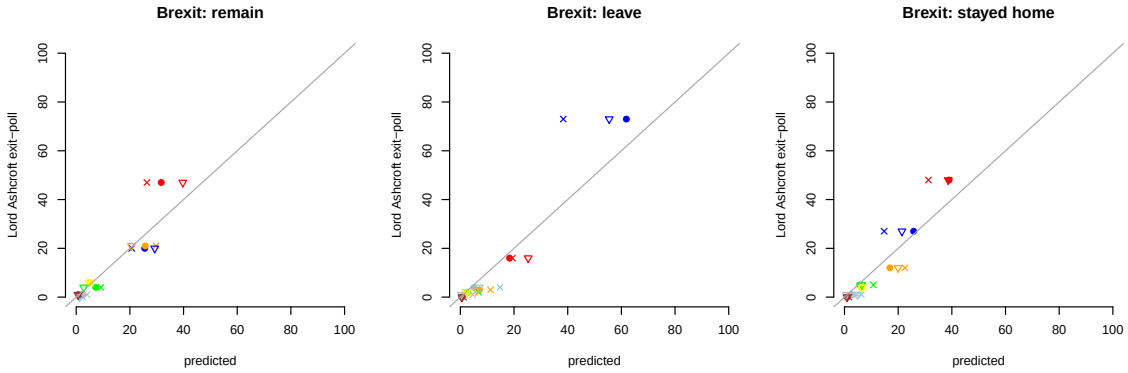


Figure A15: Vote-choice prediction-accuracy by Brexit referendum Choice.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

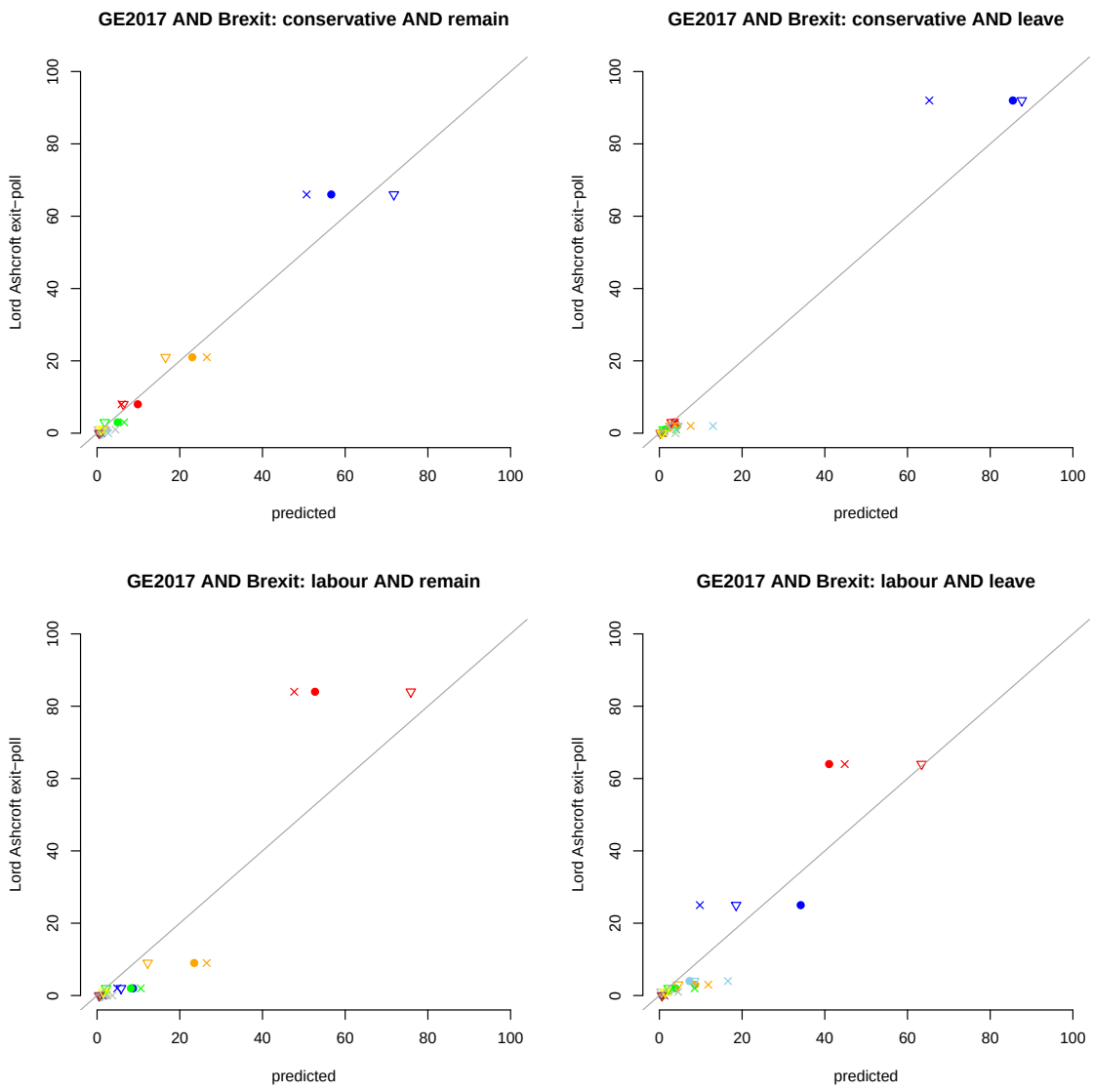


Figure A16: Vote-choice prediction-accuracy by Brexit referendum Choice AND 2017 vote-choice.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

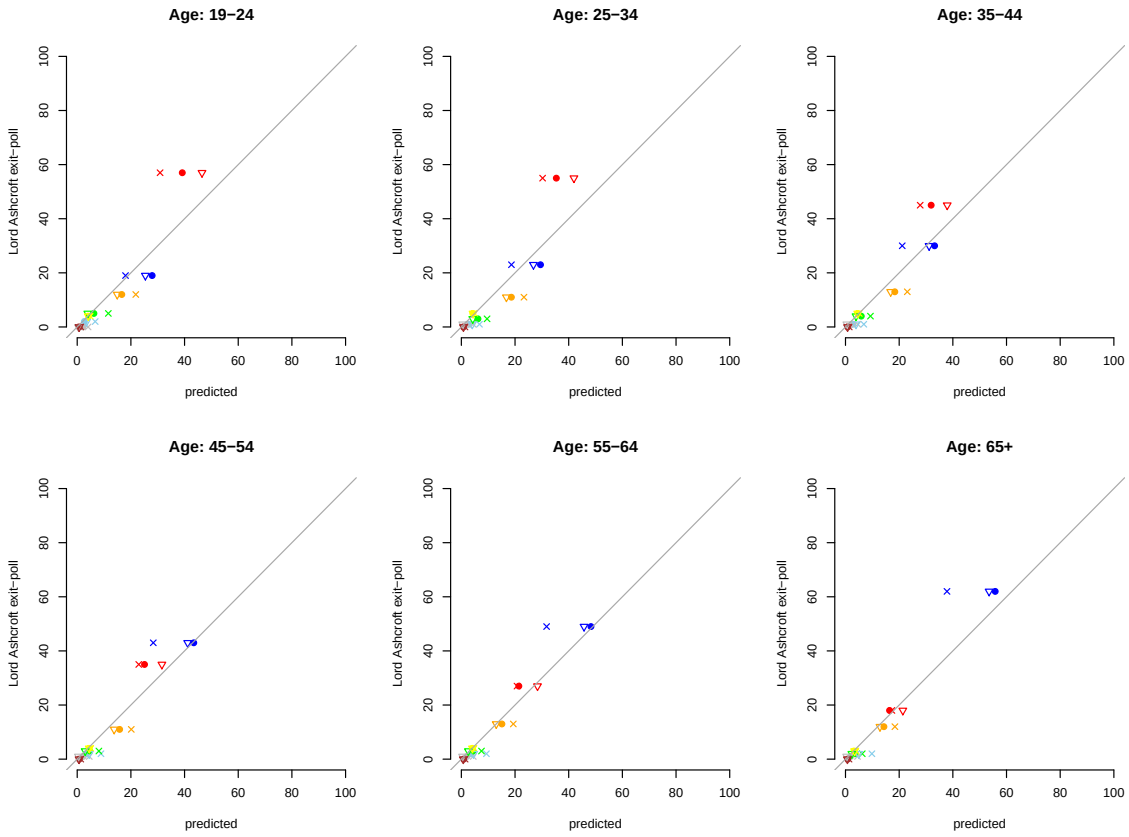


Figure A17: Vote-choice prediction-accuracy by age groups.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

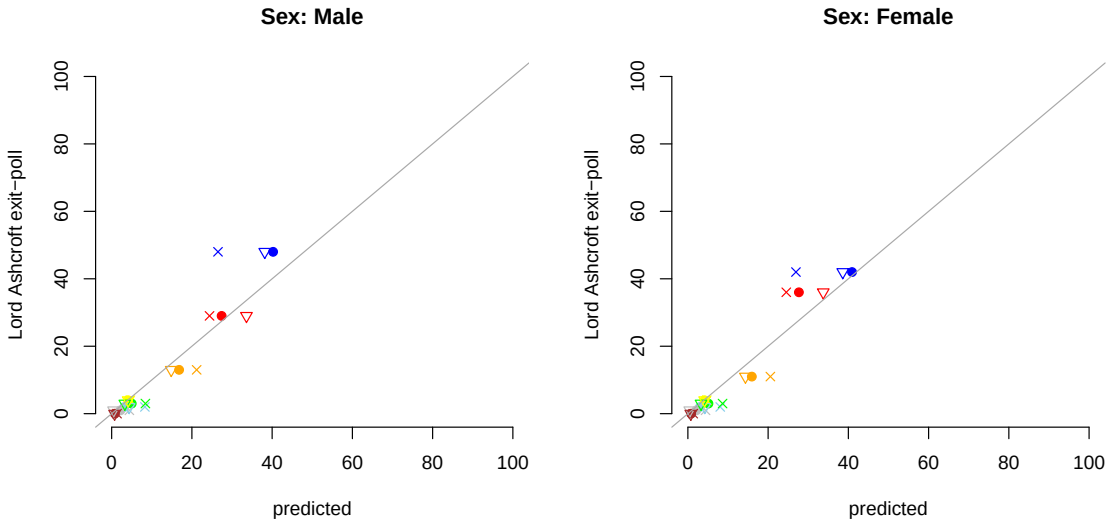


Figure A18: Vote-choice prediction-accuracy by sex.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

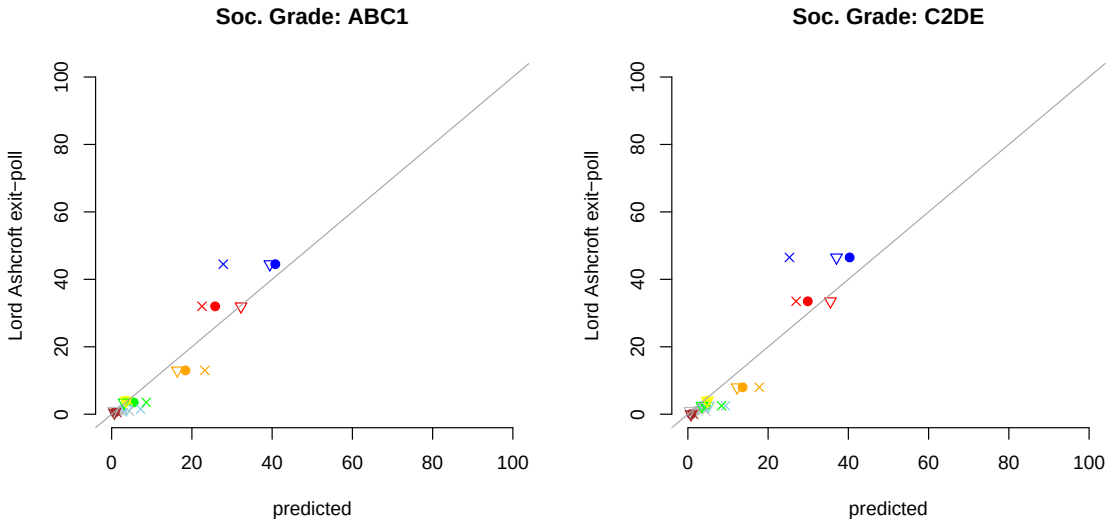


Figure A19: Vote-choice prediction-accuracy by social grade.

Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;

Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

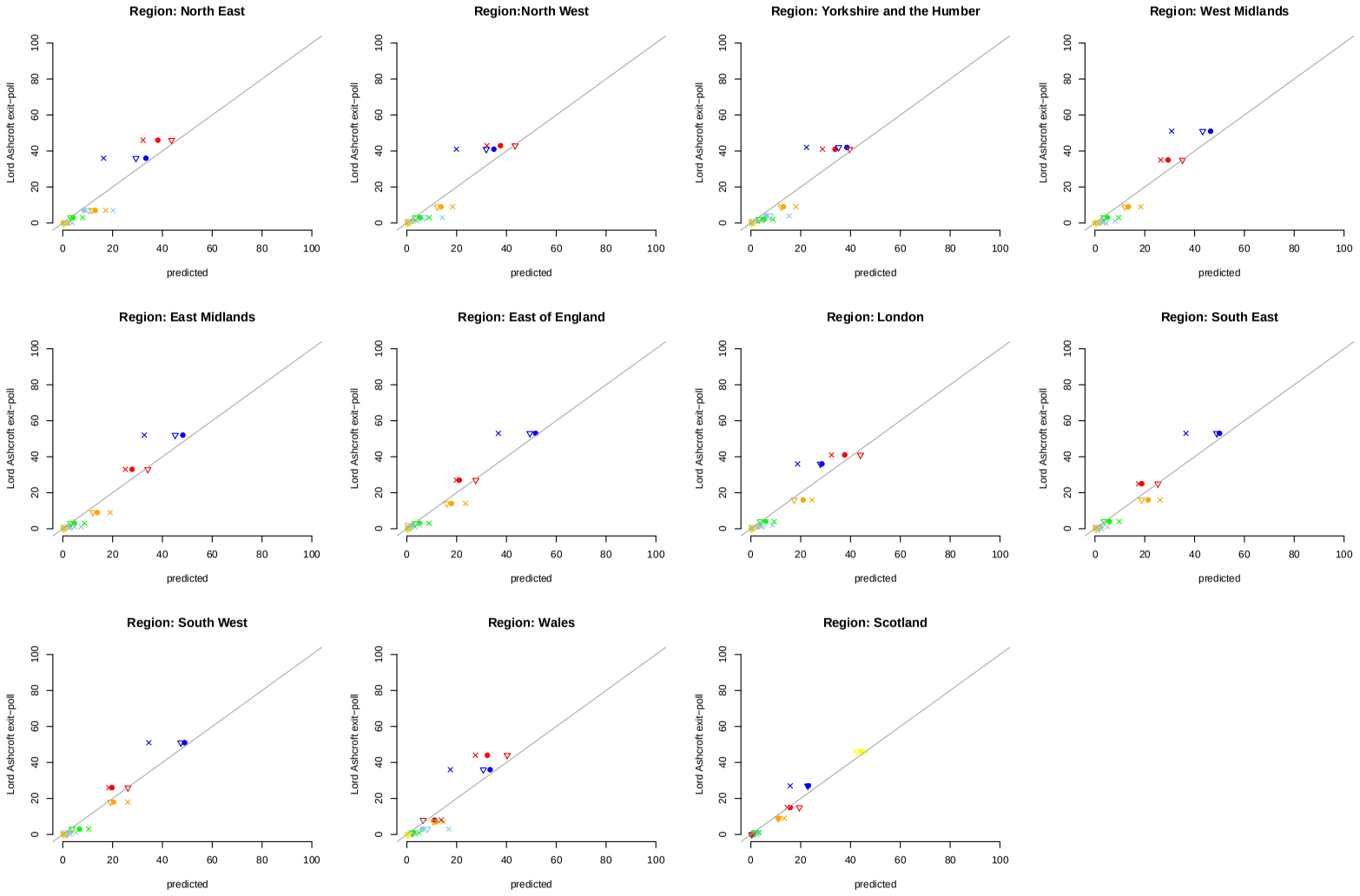


Figure A20: Vote-choice prediction-accuracy by region.
Symbols → • BES 17 & 18; × BES 16; ∇ Synthetic;
Colours → • Conservative; • Labour; • Liberal Democrat; • Brexit Party; • Green Party;
• Plaid Cymru; • SNP; • Other;

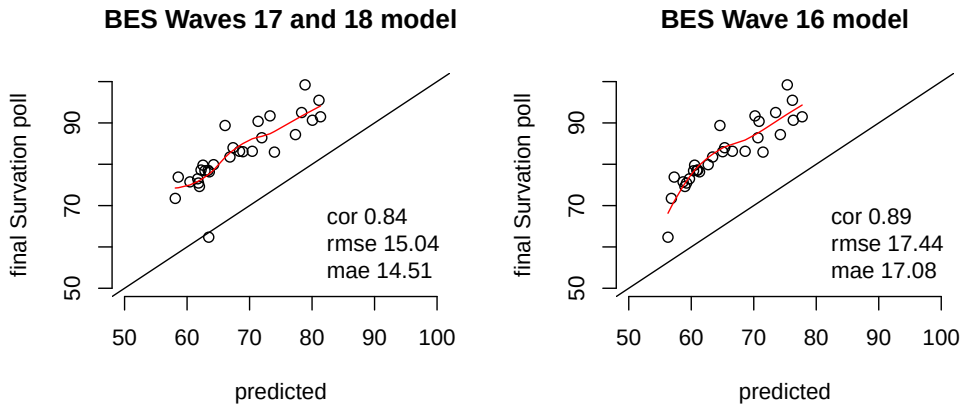


Figure A21: Turnout prediction comparison between the models outlined in this paper and the last Survation poll.

Chapter V: Rejoinder

At the end of a four-year investigation into methodological innovations to increase the power of unrepresentative samples in public opinion research, it may be wise to ask what has been learned. There are three broad themes that have been explored, which can be broadly summarized as i) convenience samples; ii) synthetic populations; iii) probability learners.

Convenience Samples. We have explored a number of convenience samples in this thesis: in 'Measuring Public Opinion via Digital Footprints' - henceforth 'paper 1' - we used a quota sampling strategy to collect reactions of Facebook users to political posts during the 2018 Texas mid-terms; in 'Polling India via Regression and Post-Stratification of Non-Probability Samples' - henceforth 'paper 2' - we surveyed Indian workers on the Amazon Mechanical Turk platform, and Indian Facebook users, in the run-up to the 2019 Lok Sabha election; and in 'Learning Voting Behaviour from Synthetic Surveys' - henceforth 'paper 3' - we explored multiple waves of the YouGov / British Election Study online subject pool in the run-up to the 2019 General Election in the UK.

Facebook and Twitter digital traces, which are examples of the data we collected in paper 1, are a-priori attractive due to two main reasons: their size / frequency, and its non-obtrusive nature. The former advertises a more complete picture about a potential voter than one could afford with a survey, whilst the latter speaks to the more 'genuine' responses we can observe when respondents don't know they are taking part in research. In our work we have found both reasons to be optimistic about the future potential of this data, as well as reasons to be more skeptical. Starting with the positives, the technological innovation brought about by `fastLink` (Enamorado et al. 2018) enabled us to match over 50% of our digital sample to voter-registration file records in Texas speaks to the high coverage that social media data has in the US; the high correlation individual-level partisan behaviour available on the voter registration files and digital traces suggests that there is correspondence between real-life and online behaviour; the reasonably good aggregate electoral predictions suggest there is signal in the digital noise. The downside of this data is that, in absence of voter-registration files, it becomes hard to extract relevant predictive features about the online users. Advances in this area are forthcoming, with the recent M3 project (Wang et al. 2019)

paving the way for large-scale demographic inference on Twitter, though error rates on the demographic predictions are still too high to compete with traditional surveying: imagine having between 30% and 70% of your survey respondents lie about their age; these are the equivalent range of successful classification rates across EU countries for age categories in the Wang et al. paper. Further difficulties with digital-trace data come with measurement: is a 'like', 'love' or an 'explicitly positive comment' on Facebook truly equivalent to a voting intention in a pre-election poll? we concluded this was the case due to the high individual-level correlation between digital traces and partisan behaviour, though more research needs to explore how fine-grained this parallel is: who are the people whose real-life behaviour does not match their online behaviour? are they common enough to represent a concern for digital social research? and can we identify them? - these are areas for future research.

With respect to the convenience samples explored in the last two papers, a number of interesting findings are worth highlighting: in India, it was surprising to see relatively stark differences in composition between Mechanical Turk workers and Facebook Users, with social status and class clearly playing a role in selection effects. Our findings about demographic and geographic composition of the Indian Mechanical Turks pool largely support existing literature (Boas et al. 2020, Ipeirotis 2010), whilst we note that Facebook-recruited users in India using quota sampling do not form a representative sample, which contrasts with tentative findings from the US (Zhang et al. 2020), where quota sampling of Facebook users was able to reproduce the population of interest. Overall, our findings support recent work on the importance of surveying multiple digital to make more robust inference (Duch et al. 2020). Whilst our effort in India proved the importance of observing respondents in multiple location, our work in paper 3 on the UK showed the importance of recency in the underlying survey data. Outdated samples cannot be easily corrected, and it appears to be vastly more important for a sample to be up-to-date than it be representative of the population. Finally, it is worth noting that in all our convenience online samples, turnout over-sampling and over-reporting is wide-spread (Kuhn & Vivyan 2018); this phenomena is well documented in the British Election Study online panel - especially when it comes to turnout of minority groups (Mellon & Prosser 2017) - but is a more general problem plaguing both novel and traditional polling (Wells 2019). Online selection exacerbates difficulties with turnout

predictions by providing incentives for the highly politically engaged to self-select into surveys.

Synthetic Populations. A second area of study in this thesis has been the joint distribution of populations of interest - typically approximated to a degree by a stratification frame. We have explored two kinds: 'pre-packaged' frames coming from the census or Voter Registration files; 'synthetic' frames, build by merging information from multiple sources. Paper 1 uses the Texas Voter Registration file to produce an extensive frame; we did not need to augment this frame, as enough political and demographic variables were available from the outset. Nevertheless, this frame too needed some modeling in order to be utilized: multiple imputation via `missRanger` (Stekhoven & Bühlmann 2012, Wright & Ziegler 2015, Wright et al. 2017, Mayer 2018) was used to complete the Voter Registration file prior to aggregation at the area-level; this has to be done as many important variables, such as education or even Age and Gender, depending on the state in question, do not have full coverage. With respect to synthetic frames, we found the Data Integration framework proposed Carrig et al. (Carrig et al. 2015) attractive for merging multiple data sources and generating deep stratification frames. Future work should directly compare this approach with MrsP (Leemann & Wasserfallen 2017) to ascertain which method should become industry-standard in model-based pre-eleciton opinion polling. We believe there are theoretical reasons why one would prefer Data Integration, which we have outlined in papers 3 and 3, though it does come at significant computational costs and extra complexity, which may end up being unnecessary for your 'run of the mill' poll. We further advocate for the use of IPF (Deming & Stephan 1940) as a tool to further refine joint frames obtained via the Data Integration procedure. We are by no means the first to use raking to further refine our frames at the area-level (Lauderdale et al. 2020, Hanretty 2019, Leemann & Wasserfallen 2017), though we are perhaps the first, within the regression and post-stratification literature, to point to the potential for IPF to be used as a real-time ecological data integration tool: by referencing some of the attractive properties of the algorithm when applied to questions of ecological inference (Johnston et al. 1982, Johnston & Pattie 2000) we open the door for real-time updating of population micro-data via IPF, deriving joint distribution targets from high-frequency data sources, such as public opinion polls. Overall, we have

learned that post-stratification weights derived from a full population joint distribution are amenable to modeling, especially thanks to advances in computational tools which enable large-scale operations at the population level.

Probability Learners. Our final set of contributions belongs to the realm of machine-learning applications, and in particular tree-based learners, to the study and prediction of probability quantities. The thesis shows a reasonable degree of progression over the three papers on this front: in paper 1 we use default settings of a probability forest (Malley et al. 2012) to predict cell-level behaviour; in paper 2 we learned to better tune these hyper-parameters according to recent advances in the literature (Olson & Wyner 2018), and obtained very encouraging results; in paper 3 we move from a random forest approach to tree-boosting (Chen & Guestrin 2016), and propose novel ways to tune hyper-parameters for the purpose of estimating electoral probabilities. A weakness of probability machines is that, unlike regression machines (Athey & Wager 2019, Athey et al. 2019), they lack asymptotic theory to describe predictive distributions. The beta-binomial model for probability-forests proposed in paper 2 seems to be a promising avenue for further research: the predictions of two very different quantities - turnout and vote-chocie - largely fit the beta distribution, with minor transgressions. This suggests that, not unlike regression-forest asymptotics, for the beta distribution to hold as the true underlying predictive distribution, some restrictions on the tree-growing mechanisms have to be assumed. This should be the subject of further work. Overall, our three papers highlight the importance of hyper-parameter tuning - whether theoretically or empirically motivated, for the success of probability machines. A final note belongs to the use of tree-based learners specifically for regression and post-stratification: we have found in our research, via trial and error, that the most important parameters to tune for area-level predictions are `mtry` and `max.depth`: making area level predictions requires trees to 'follow the signal' for quite a number of steps, and to try as many candidate variables as possible, in order to capture the relevant deep and cross-level interactions (Lauderdale & Blumenau 2019); these empirical findings point to bagged trees (Hastie et al. 2009) as the more theoretically reliable algorithm in attempting to predict area-level quantities via regression and post-stratification- though note boosted treed are able to converge to this model if it is supported by the underlying tuning-data,

and may be preferable as a result.

Final Thoughts. In this thesis I have explored ways to make useful inference from unrepresentative samples. The result of this work can be though of as a workflow, characterized by the following steps: i) *micro-data collection*: I have shown where one can collect cheap and unrepresentative samples in the run-up to an election; ii) *building stratification frames*: I have discussed and evaluated data sources to use in approximating the underlying joint distribution of a population of interest; I have further presented methods to model the frame and extend it to include variables fro auxiliary sources, as well as retaining ecological representation; iii) *updating micro-data and stratification frames*: I have shown how to update frames and micro-data using targets derived from demographic tables of publicly available opinion polls, using IPF; iv) *learning*: I have provided a framework for choosing optimal learners, which perform automatic variable selection and are orders of magnitude faster than the state of the art; we have further shown how to tune said learners, from theoretical and empirical standpoints; v) *modeling uncertainty*: I have provided practical ways to account for uncertainty in predictions of probability machines.

The contributions presented in this thesis are building toward a global artificial intelligence tool for automated polling - where the global public will no longer have to rely on the discretion of the modelers and the interest of the pollsters to obtain real-time, fine grained estimates of their community's attitudes and opinion. Though my approach does not do away with modeling decisions entirely, it is conscious of minimizing the number of pivotal decisions, and provides a principle guide as to how to make these. I believe all the pieces for the conception of this artificial intelligence lie within the work thus far presented, and dedicate my academic career to bringing this tool to life.

References

Athey, S., Tibshirani, J., Wager, S. et al. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178.

Athey, S. & Wager, S. (2019), 'Estimating treatment effects with causal forests: An application', *arXiv:1902.07409* .

Boas, T. C., Christenson, D. P. & Glick, D. M. (2020), 'Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics', *Political Science Research and Methods* **8**(2), 232–250.

Carrig, M. M., Manrique-Vallier, D., Ranby, K. W., Reiter, J. P. & Hoyle, R. H. (2015), 'A nonparametric, multiple imputation-based method for the retrospective integration of data sets', *Multivariate behavioral research* **50**(4), 383–397.

Chen, T. & Guestrin, C. (2016), `xgboost`: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics* **11**(4), 427–444.

Duch, R., Laroze, D., Robinson, T. & Beramendi, P. (2020), 'Multi-modes for detecting experimental measurement error', *Political Analysis* **28**(2), 263–283.

Enamorado, T., Fifield, B. & Imai, K. (2018), 'Using a probabilistic model to assist merging of large-scale administrative records', *Social Science Research Network (SSRN): 3214172* .

Hanretty, C. (2019), '2019 General Election MRP predictions', `https://www.survation.com/2019-general-election-mrp-predictions-survation-and-dr-chris-hanretty/`.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Ipeirotis, P. G. (2010), 'Demographics of mechanical turk'.

Johnston, R., Hay, A. & Taylor, P. (1982), 'Estimating the sources of spatial change in election results: a multiproportional matrix approach', *Environment and Planning A* **14**(7), 951–961.

Johnston, R. & Pattie, C. (2000), 'Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting', *Political Analysis* **8**(4), 333–345.

Kuhn, P. M. & Vivyan, N. (2018), 'Reducing turnout misreporting in online surveys', *Public Opinion Quarterly* **82**(2), 300–321.

Lauderdale, B. E., Bailey, D., Blumenau, J. & Rivers, D. (2020), 'Model-based pre-election polling for national and sub-national outcomes in the US and UK', *International Journal of Forecasting* **36**(2), 399–413.

Lauderdale, B. E. & Blumenau, J. (2019), Constructing and assessing seat level estimates. Reading the 2019 Election Polls: Event by the London School of Economics, Department of Methodology, and the British Polling Council.

Leemann, L. & Wasserfallen, F. (2017), 'Extending the use and prediction precision of subnational public opinion estimation', *American journal of political science* **61**(4), 1003–1022.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. & Ziegler, A. (2012), 'Probability machines', *Methods of information in medicine* **51**(01), 74–81.

Mayer, M. (2018), 'Using `missRanger`', `https://cran.r-project.org/web/packages/missRanger/vignettes/vignette_missRanger.html`.

Mellon, J. & Prosser, C. (2017), 'Missing nonvoters and misweighted samples: explaining the 2015 great british polling miss', *Public Opinion Quarterly* **81**(3), 661–687.

Olson, M. A. & Wyner, A. J. (2018), 'Making sense of random forest probabilities: a kernel perspective', *arXiv:1812.05792* .

Stekhoven, D. J. & Bühlmann, P. (2012), '`missForest` — non-parametric missing value imputation for mixed-type data', *Bioinformatics* **28**(1), 112–118.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flock, F. & Jurgens, D. (2019), Demographic inference and representative population estimates from multilingual social media data, *in* 'The World Wide Web Conference', pp. 2056–2067.

Wells, A. (2019), 'Likelihood to Vote', `https://ukpollingreport.co.uk/faq-turnout`.

Wright, M. N., Wager, S., Probst, P. & Wright, M. M. N. (2017), 'Package `ranger`'.

Wright, M. N. & Ziegler, A. (2015), '`ranger`: A fast implementation of random forests for high dimensional data in C++ and R', *arXiv:1508.04409* .

Zhang, B., Mildenberger, M., Howe, P. D., Marlon, J., Rosenthal, S. A. & Leiserowitz, A. (2020), 'Quota sampling using Facebook advertisements', *Political Science Research and Methods* **8**(3), 558–564.