Small area estimates of public opinion: model-assisted post-stratification of data from voter advice applications

Simon Jackman, Shaun Ratcliff and Luke Mansillo

United States Studies Centre, University of Sydney

04 January 2019

Abstract

Small-area estimates of public opinion are vital for studies of representation, but are constrained by the costs of collecting sufficiently large surveys. We use model-assisted post-stratification procedures to repurpose data from a voter advice application (VAA) fielded during the 2016 Australian Federal election campaign. VAAs are typically run with media partners, and provide massive samples relative to commercial and academic surveys (in this case, nearly 800,000 respondents). However, considerable bias is generated from self-selection. Our procedure uses Bayesian classification trees to form predictive models of survey responses, which we project onto post-stratification frames for each of Australia's 150 House of Representatives electoral divisions. We demonstrate the utility of these data and our methodology for district-level estimates using a unique opportunity, a 2017 plebiscite on same-sex marriage, calculating small-area estimates that would have been prohibitively expensive to obtain with conventional surveys.

Small-area estimates of public opinion

Representation is a core concern of democratic theory and practice (Key 1961). For both normative (Dahl 1971) and practical reasons (Downs 1957), elected officials are presumed to be responsive to the preferences of their constituents. Understanding the mapping between mass preferences and legislative behavior — and public policy — is thus a central and enduring research program of political science.

Let a legislators' revealed preferences based on roll call voting (an ideal point or voting score) be $\xi_i \in \mathbb{R}$ (see Clinton, Jackman, and Rivers 2004; Poole and Rosenthal 2007). Many studies of representation focus on the relationship $\xi_i = f(x_i)$, where x_i is a measure of the preferences of citizens in legislative district i. For instance, many parliamentary systems have few departures from party-line voting in legislatures, with $\xi_i \approx \xi_{p(i)}$ where $p(i)$ is the party of legislator i. In particular, in a strong two-party Westminister system, we may have $\xi_i \in \{-c, c\}$ (i.e., just two unique ideal points, one for each party, with no meanignful within-party variation among legislators), as party discipline "oversmooths" f with respect to district preferences. In the United States, we might also question whether a combination of primary voting, gerrymandering and polarisation result in more extreme politicians being elected to Congress. In either case, are there electoral sanctions (Canes-Wrong, Brady, and Cogan 2002) for $\xi_i \neq f(x_i)$?

To answer this question we require estimates of x_i, the preferences of citizens at the level of legislative districts. These district-level estimates of public opinion — what statisticians call small-area estimates — are seldom produced in contexts other than the United States. Cost is the typical impediment. Miller and Stokes (1963) neatly articulated the nature of this problem: a public opinion survey with national coverage — and a conventional sample size — has very few respondents per legislative district. This makes studies of representation at the level of individual legislative districts difficult, with researchers seldom possessing the resources for the large samples needed to generate useful small area estimates.

Even in the United States, applications have largely focused on states, save for some prominent exceptions producing estimates at the level of counties or Congressional districts (e.g., Tausanovitch and Warshaw 2013, 2014; Warshaw and Rodden 2012). This kind of work has therefore often been limited to researchers with access to very large (and often expensive) surveys, such as the CCES family of studies (Vavreck and Rivers 2008).[1]

An example of this problem can be observed with the Australian Election Study (AES). In 2016, this had an N of 2818. As can be seen in Figure 1, with 150 electoral divisions represented in the House of Representatives, this is an average of 18.8 respondents per district. This is not a criticism of the AES. It is not designed for small-area estimates at the level of electoral divisions. Rather, it is built to provide canonical time series data. This is a function it serves well. If we wish to examine public opinion and representation at the level of legislative districts in the Australian context, though, we need a different data set.

Opt-in, online surveys are a potential solution to this dilemma. These are capable of de-

[1] While the CCES is relatively affordable for individual researchers who purchase modules, a very large sample needs to be collected overall to obtain a reasonable number of observations in individual congressional districts. This limits this strategy to countries with a critical mass of political scientists with budgets to fund these kinds of studies.

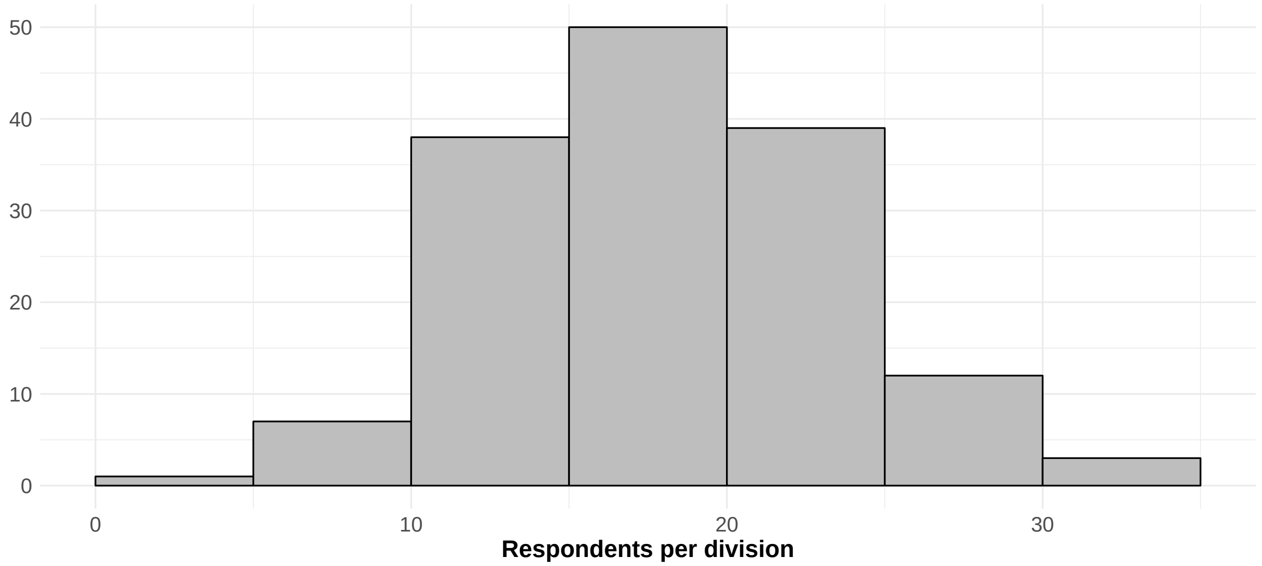

Figure 1: Distribution of number of observations per division,2016 Australian Election Study.

livering massive samples at low cost. The CCES surveys can be seen as a particular (high quality) subset of this type. Voter advice applications (VAAs) are another. These are an intriguing form of this type of survey, offering a benefit to respondents by providing them with an estimate of how close their political preferences are to the policies adopted by parties and candidates. These estimates rely on responses to many issue self-placement items. VAAs also typically measure other political and demographic attributes of respondents. In addition, they are often administered with a prominent media partner, who shoulder most of the logistical costs (for further discussion on VAAs see Garzia and Marschall 2016; Linden and Dufresne 2017). VAAs thus offer the potential for massive quantities of data spanning a host of politically important issues, at extremely low cost to social science end-users.

But low cost data is often of low quality. Unsurprisingly, data from VAAs have considerable bias, generated by (a) the recruitment mechanism typically employed — publicity via the media partner's outlets or social media channels — meaning individuals not exposed to these outlets or channels have a very small probability of participating, inducing coverage error; (b) the exclusive use of on-line, self-completion survey mode, which may exclude would-be participants without internet access; and (c) self-selection. Yet, if the resulting biases can be ameliorated, VAA data might provide significant research opportunities.

To answer this question we employ model-assisted, post-stratification procedures, attempting to remove the bias in data from the Vote Compass VAA fielded in the weeks prior to the 2016 Australian Federal election. We build post-stratification frames of Census characteristics, augmented by 2016 vote intention, for each of Australia's 150 House of Representatives divisions. We fit classification trees individually for survey responses in each division using Bayesian methods, with the post-stratification frame for divisions run through the trees to produce district-level estimates of public opinion.

We take advantage of a unique opportunity to validate our estimates against a known, post-election outcome: the 2017 plebiscite (or postal survey) administered on same-sex marriage in Australia by the Australian Bureau of Statistics. We demonstrate the utility of these data and our methodology for district-level estimates, which would have been prohibitively expensive to obtain with conventional surveys, but are now available given VAAs and methods to remove much of their inherent biases.

We show that much of the remaining error in our estimates after modelling and post-stratification appears to be driven by two factors: the under-representation of cultural diversity in the electorate, which was possibly exacerbated by low turnout in some of these diverse districts. Our findings also highlight two important considerations for the design of VAAs when administered to diverse populations: the need for a multi-lingual questionnaire and the inclusion of demographic information on language(s) spoken by respondents.

Using VAAs and model-assisted procedures to understand public opinion

Model-assisted procedures with post-stratification (such as MRP) have already been applied to answer questions about public opinion and democratic representation in the United States (Lax and Phillips 2009a, 2012; Gelman and Lee 2010; Warshaw and Rodden 2012; Canes-Wrone, Clark, and Kelly 2014; Kastellec et al. 2015; Shirley and Gelman 2015). However, due to the large samples required for small-group and area estimation, there are few examples from other democracies. Hanretty, Lauderdale, and Vivyan (2016) used MRP to estimate public opinion at the constituency-level in the United Kingdom. Selb and Munzert (2011) used multilevel models (but not MRP) to estimate levels of party support by legislative district in Germany. The existing literature does not expand much beyond these examples.

Much of the early literature on VAAs concentrated on issues specifically concerning the tools themselves, rather than potential uses, including ethical concerns (Alvarez et al. 2014), and technical issues (Linden and Dufresne 2017). Later literature expanded this to examine how accurately data collected through VAAs could predict election outcomes (Johnston 2017), and if these data could be used to map different dimensions of voters' issue preferences (Wheatley et al. 2014), but to date they have not been used for what we believe to be one of their most promising purposes: to produce estimates of public opinion for small groups and areas, such as legislative districts.

Data, methods and accuracy

We adopt a modelling strategy to reduce the effects of selection bias inherent in opt-in surveys. This includes post-stratifying the modelled estimates of public opinion on citizens' demographic characteristics. The Vote Compass VAA asked respondents their vote intention at the 2016 Australian federal election. We exploit responses to this question to improve our inferences, balancing the data against election returns and removing sources of bias correlated with partisanship that can not be addressed by balancing with respect to small set of available demographic variables.

The raw data: massive, but biased

The 2016 Australian federal election Vote Compass VAA (Vox Pop Labs 2016) was developed by a team of data scientists from Vox Pop Labs and scholars from the Universities of Melbourne and Sydney. The VAA was hosted on the website of the Australian Broadcasting Corporation, who aggressively promoted the application over the course of the election campaign. The VAA included a set of questions on respondent's demographics (age, birthplace, education, household income, religious affiliation, geographic location), vote intention and issue preferences. Over the 56 days of the formal campaign, Vote Compass collected data on 1,178,398 Australian voters, with data for nearly 800,000 available after removing observations with missing information, and duplicated responses from the same IP address (with this number varying slightly by the outcome being estimated). As can be seen in Figure 2, this leads to a minimum of approximately 1,000 respondents per division (in those with the fewest respondents) to nearly 20,000 (those with the largest).

Despite the massive sample size, Figures 3 and 4 indicate there is significant demographic and political bias in the sample. Respondents aged 65 and over were under-sampled by half. Younger voters were over-represented. University educated voters were over-represented by a factor of three, while those who had not finished high school were, despite being the largest educational cohort in the population, the smallest in our sample. There was a disproportionate number of men in the Vote Compass sample, and too few women. Respondents with household incomes in the lowest quintile were slightly over-represented and those who did not state their incomes under-sampled. Compared with the Census, there was an insufficient count of citizens identifying with Christian religions in the Vote Compass data, while

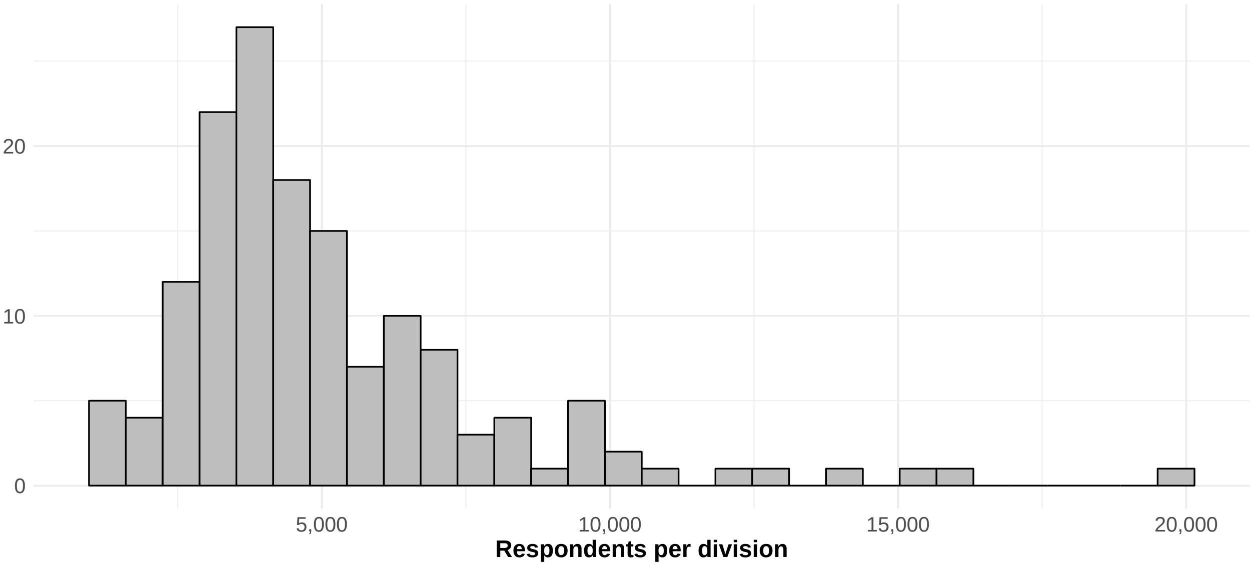

Figure 2: Distribution in number of observations per division in the 2016 Australian Vote Compass VAA

those with no religion were over-sampled by nearly a factor of two. Vote Compass also had a 40 per cent over-sample of citizens living in inner-city electorates, and double the true proportion of Green voters (comparing survey responses with 2016 House of Representatives election results). Rural voters were under-represented and those living in the inner-city over-represented.

Supporters of the centre-right Coalition parties at both the 2013 and 2016 elections were under-sampled by nearly 50 per cent in our data, as were those saying they supported other parties and candidates, or did not intend to vote at the 2016 election. Centre-left Labor voters were slightly over-represented, and supporters of the progressive-left Greens were heavily over-represented.

These biases in the data lead to poor quality estimates of public opinion at the level of electoral division, as can be seen in Figure **??**. This shows the difference between the level of support for each political party in the raw 2016 Vote Compass data and observed vote share in Australia's 150 electoral divisions at the 2016 election. Clearly we need to mitigate some of this bias if we are to obtain reasonable inferences.

Correcting for self-selection

Our model-assisted approach for rehabilitating the Vote Compass data proceeds in a number of stages. Our goal is to produce small-area estimates of the proportion of the adult, citizen

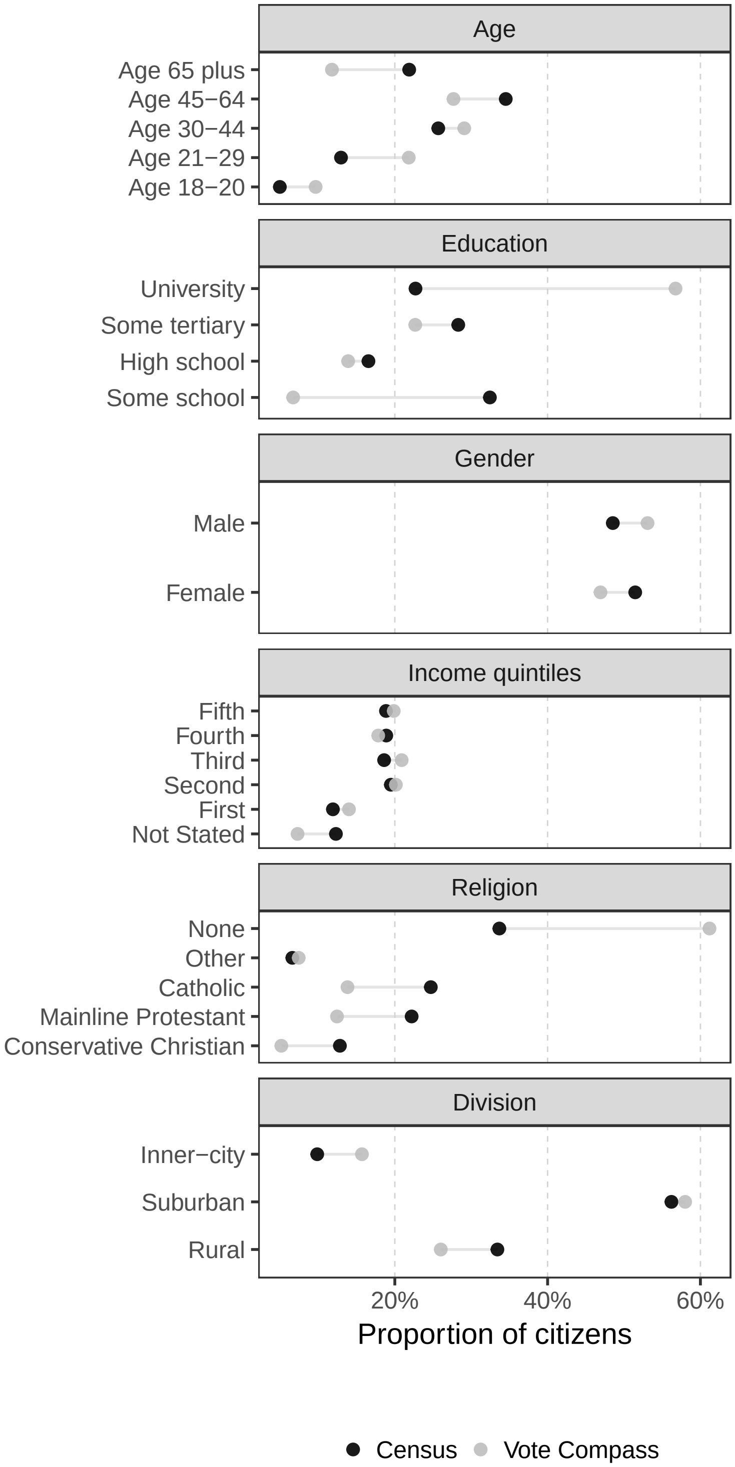

Figure 3: Differences between the demographic composition of the Vote Compass sample and the Australian adult, citizen population (according to the 2016 Australian Census).

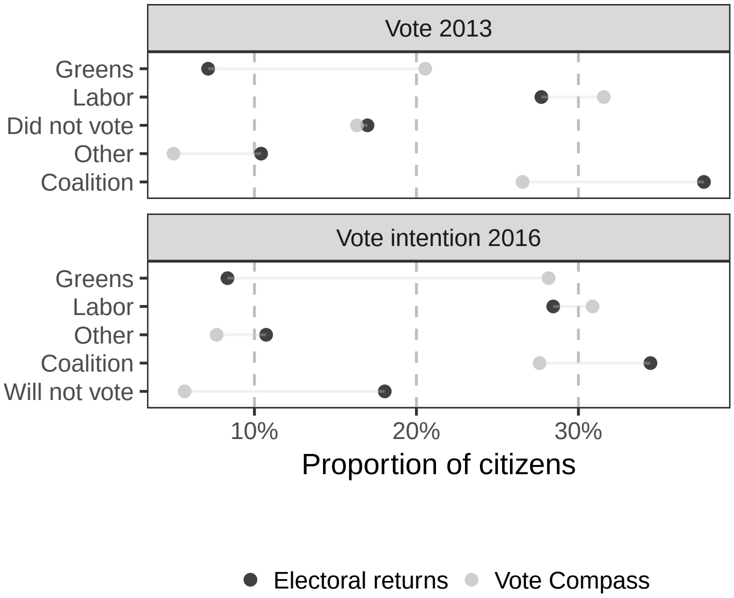

Figure 4: Differences in the voting intentions of the Vote Compass sample and election results recorded by the Australian Electoral Commission.

population with some attribute y, typically at the level of Australia's 150 electoral divisions (geographic districts represented by seats in Australia's House of Representatives). Let y be a variable of interest measured on the survey, but not available in the 2016 Census: e.g., preferences on reducing the federal deficit or same-sex marriage. We use machine-learning techniques to fit a predictive model for y as a function of demographic predictors X, which are available in both the Vote Compass data and measured in the Census.

We extract a data set of cells from the Australian Bureau of Statistics (ABS) Tablebuilder tool to build our post-stratification frame, enabling the cross-classification of X by division. For each division there are 1,100 cells, comprised of the cross-sectioning of:

- Age (18-20, 21-29, 30-44, 45-64, and 65 years and older).
- Gender (male, female).
- Education (some school, high school, a trade qualification or diploma, and a bachelor degree or higher).
- Household income (quintiles, and not stated).
- Religion (Mainline Protestants, Conservative Protestants, Catholic, other religions and no religion).

In our models, outcomes are predicted as a function of these demographic characteristics.

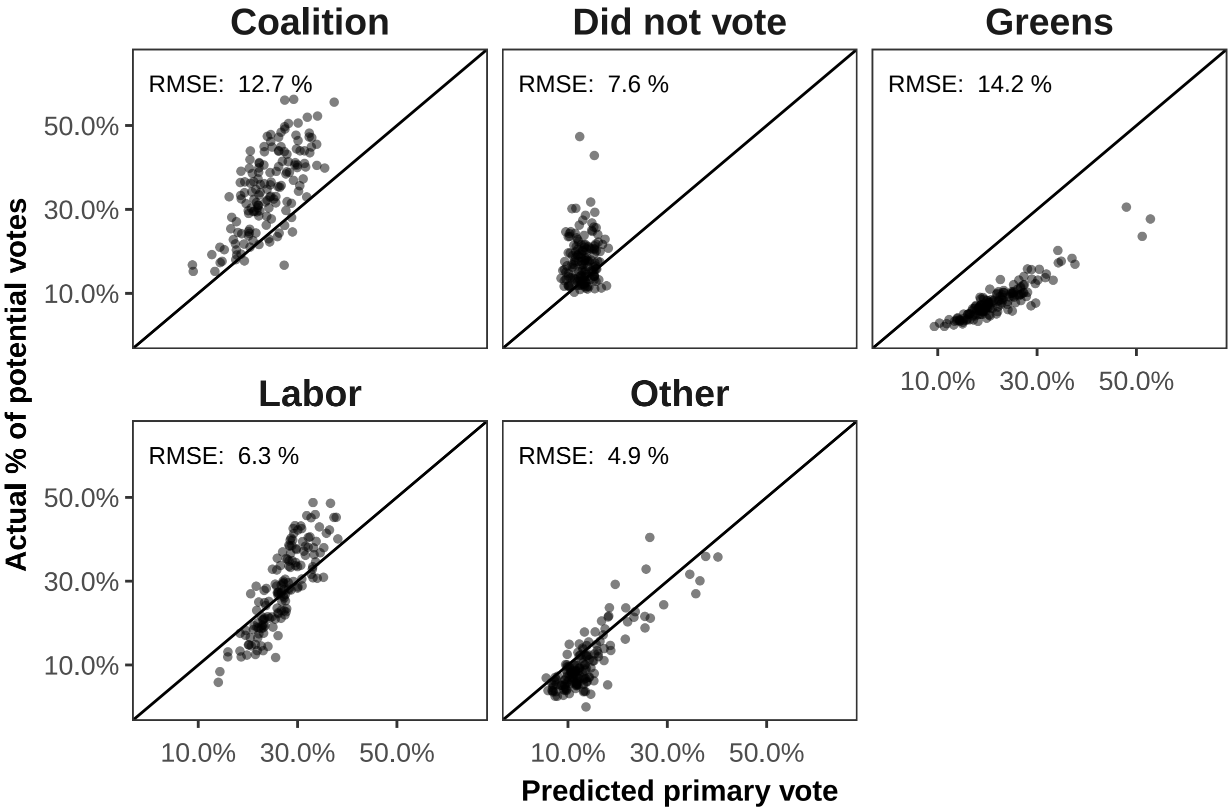

Figure 5: Comparison of the level of support for each political party in the raw 2016 Vote Compass data and observed vote share in Australia's 150 electoral divisions at the 2016 election. Each point represents a seperate division. The diagonal curve is the line of fit.

Ordinal or nominal survey responses are considered a a set of mutually exclusive and exhaustive binary indicators: i.e., $y_{ij} = 1$ if respondent i offers response j on survey item z, and otherwise $y_{ij} = 0$, where j indexes J possible responses. For each binary $y_{\cdot j}$ we fit a tree-based Bayesian classifier (Kapelner and Bleich 2013), using demographic characteristics X as predictors. Each of the J binary classification models are fit separately to the data from each of the 150 electoral divisions.

The Bayesian tree-based modeling approach regularises the models fit to the survey data, balances model fit against model complexity, the latter growing via the inclusion of more demographic variables and/or interactions between them. With draws from the posterior distribution over the classification tree, we generate probabilities of each of the J responses in each cell of the post-stratification frames.[2]

We treat these cells as a data set with which to predict y, using the model derived from the Vote Compass data, for each division. For a binary outcome y, we predict θ_c, the probability that person i in the corresponding Census cell c has attribute $y = 1$.

Each cell is assigned the relevant population frequency N_c, calculated by multiplying the probability of y for each cell with the population count from the Census. Summing over cells and dividing by the total cell count gives us an estimate for the proportion of citizens within a division with attribute $y = 1$. This procedure generalizes easily to the case of a J-class attribute y, by normalising the probabilities from the J (mutually exclusive and exhaustive) binary classifiers to sum to one in each cell. Accordingly we can measure citizens' preferences on different policy issues in all 150 electoral divisions represented in the Australian parliament.

We also add a non-Census variables to the post-stratification frames: 2016 House of Representatives vote. To do this we (a) use the modeling and prediction procedure described above for each of these survey responses, but with 2016 vote as y, followed by (b) a round of raking such that the marginal distribution of 2016 vote on the augmented frames matches actual election results, and preserves the cell counts on the original post-stratification frame.

[2] Slightly different codings of religion were employed in a small set of electoral divisions where different specifications for our model were found to provide a better fit to the data. This usually involved dropping variables when they provided little additional information. A slightly larger frame was also used in some divisions where a sixth religion category was included for voters identifying as Hindu and Muslim. This slightly alters the number of cells in each frame, and the number of citizens in each cell.

Contrast with other methods

This model-assisted procedure does have some similarities with conventional weighting procedures. Raking or rim-weighting is perhaps the most frequently used post-stratification procedure in commercial and academic polling. For a series of weighting variables $W_k, k = 1, \ldots, K$, an iterative proportional scaling algorithm produces weights for the survey data such that for each W_k, the weighted marginal distribution of our survey data, closely matches the marginal distribution of W_k in the target population. Rim-weighting suffers from a familiar trade-off: as the number of weighting variables K increases, sample cell counts become small and corresponding cell weights tend to get very large, making inference extremely sensitive to those cases and inflating the variance of the resulting estimates. In practice, K must be kept small. Typically, these weights extend to no more than three or four variables, and cells are often arbitrarily "trimmed".

Moreover, while rim-weighting ensures reasonable matches between sample data and population targets for each W_k, there is no guarantee that the sample estimate of the joint distribution of W is a good match to the population joint distribution of those variables. Examples familiar to us include weights that do a good job of balancing survey data with respect to age and vote choice separately, but do a poor job of recovering the distribution of vote conditional on age; a sign that the weighting procedure is not recovering the joint distribution of the weighting variables. In turn, this can lead to error when making inferences about y, to the extent y is a function of the W.

Valid inferences for y will result from this model based procedure if the covariates we employ, X, induce ignorability: the property that conditional on X, y is independent of survey response or non-response. Non-response bias arises when survey respondents differ from non-responders with respect to y. Ignorability is the property that we possess covariates X such that also knowing if someone responded or not would not improve predictions of y. If ignorability holds, then the model for y given X fitted using the survey data generates unbiased population estimates of y, predicted using the joint population distribution of X provided by the Census.

As with rim-weighting, model-assisted post-stratification faces a 'curse of dimensionality'. As the number of variables entering a predictive model for y increases, the Census data for a given division must be partitioned ever finer. When the N within a given cell drops below a threshold, the ABS Tablebuilder tool introduces random noise to the counts as a privacy protection measure. Therefore, when specifying our models we engage in a trade-off between the additional noise created by small-N cells and adopting those predictors that reduce the bias inherent in our data and provide useful information for predicting the outcomes in which we are interested.

Our application of this procedure addresses some important theoretical and practical issues. Without observing non-responders, ignorability conditional on X is typically a maintained, untestable assumption. Ignorability would fail to hold if even conditional on measured characteristics X, survey responders and non-responders differed with respect to y. Empirical confirmation of an absence of ignorability would be observing bias in an estimate of y based on procedures utilizing X variables thought to be sufficient to induce ignorability. We report a test of this in the following section.

Validation with the same sex-marriage postal survey

On 15 November 2017 — slightly more than a year after the election during which our data were collected — it was announced that 61.6 per cent of Australian voters had 'voted' to change the law to legalise same-sex marriage. This 'plebiscite' (officially, the Australian Marriage Law Postal Survey) was a compromise solution to resolve internal party tensions within the governing centre-right Coalition parties, and was ostensibly used to guide the actions of legislators in Australia's federal parliament, which subsequently legislated the legal recognition of same-sex marriage. This plebiscite was run by the Australian Bureau of Statistics between 12 September and 7 November, 2017, was conducted by mail, and had a participation rate of 79.5 per cent.

We compare division-level results from the plebiscite to estimates of voters' attitudes towards same-sex marriage at the 2016 Australian election derived from Vote Compass data collected during the campaign, and the model-assisated post-stratified estimates. Figure 6 displays the model-assisted, post-stratified VAA estimates (horizonal axis) and the plebiscite results (vertical axis).

Although it is unlikely public opinion changed substantially on this issue, there was more than a year and a heated campaign campaign between the collection of our data and the vote. Attitudes may have shifted in some divisions. Additionally, the question asked in the Vote Compass VAA was not the same as that asked in the plebiscite. In 2016, Vote Compass asked Australian voters whether they agreed 'marriage should only be between a man and a woman'. The 2017 vote asked 'Should the law be changed to allow same-sex couples to marry?' Additionally, we find that attitudes towards this issue have a weaker relationship to partisan choice (one of our stronger predictors) than many others. Accordingly, we conjecture that given (1) differences in question wording; (2) the time interval between the 2016 VAA and the 2017 plebiscite, then a linear transformation of the model-assisted post-stratified Vote Compass estimates should be approximately unbiased with respect to the plebiscite results. This is represented in Figure 6 by a red curve for least squares, linear regression, and a blue curve which corresponds to an outlier resistant regression.

Visual inspection of the data indicates that most of the data are consistent with a linear regression relationship between the two variables, of the sort produced by the outlier resistant regression. The RMSE for our division-level predictions was 6.49 per cent, with a mean absolute error of 5.6 per cent. The RMSE for the residuals on the linear regression line was 5.81 per cent, with a mean absolute error of 4.3 per cent. Our estimates were within five per cent of the regression line for 130 of the 150 divisions represented in the Australian House of Representatives (or 87 per cent of districts), and within 7.5 per cent for 149 divisions (99 per cent of divisions).

A reasonably small number of large depatures from linear regression appear to contaminate the data in divisions where the model-assisted post-stratified VAA estimates vastly overestimate levels of support for same-sex marriage obtained in the plebiscite. Further analysis indicates that divisions with high levels of non-English speaking households produce these large mismatches shown in Figure 6. This is demonstrated in Figure **??**, which highlights how the performance of our model-assisted, post-stratified estimates declines as divisions are included by their proportion of individuals that speak a language other than English at home.

These divisions, highlighted in Figure 8, were also characterised by relatively low levels of participation in both the Vote Compass VAA and the same-sex marriage plebiscite. We conjecture that our procedure “overcorrects” in these divisions, in the sense that we post-stratify to the entire adult, citizen population of electorates, not the non-random subset that participated in plebiscite.

Post-stratifying on 2016 vote does not improve the estimates of support for same-sex marriage in these districts. Within them (all centre-left Labor strongholds), 2016 vote was only weakly related to both Vote Compass participation, preferences on same-sex marriage and participation in the same-sex marriage plebiscite. As we show in Figure 8(b), divisions with high rates of non-voting (low turnout) on average had greater error in our models. This relationship was largely driven by the ten divisions in which more than 50 per cent of the population spoke a language other than English at home (shaded light blue in these plots), which as Figure 8(c) shows, on average had higher rates of not-voting (24 per cent versus 20 per cent for other divisions).

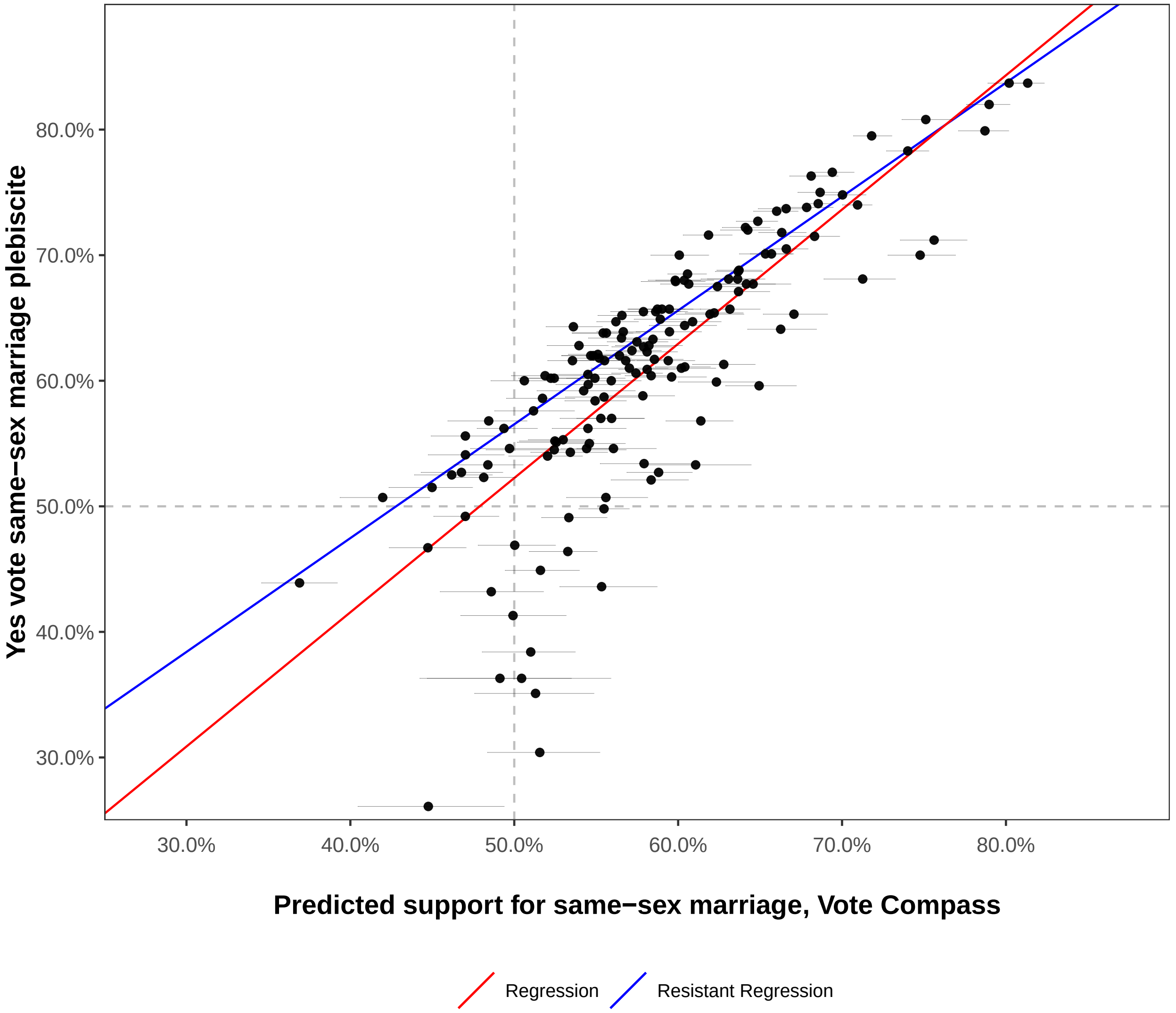

Figure 6: Comparison of estimated and observed support for same-sex marriage in Australia's 150 Commonwealth electoral divisions. Each point shows support for same-sex marriage, estimated from the Vote Compass question on whether marriage should limited to between a man and a woman (negative responses were coded as support for same-sex marriage), and the actual level of support in the marriage plebiscite. The blue line is a resistant regression estimator (minimising a weighted sum of residuals, using Tukey's biweight) between our Vote Compass estimate and the plebscite result. The red line is ordinary least squares regression.

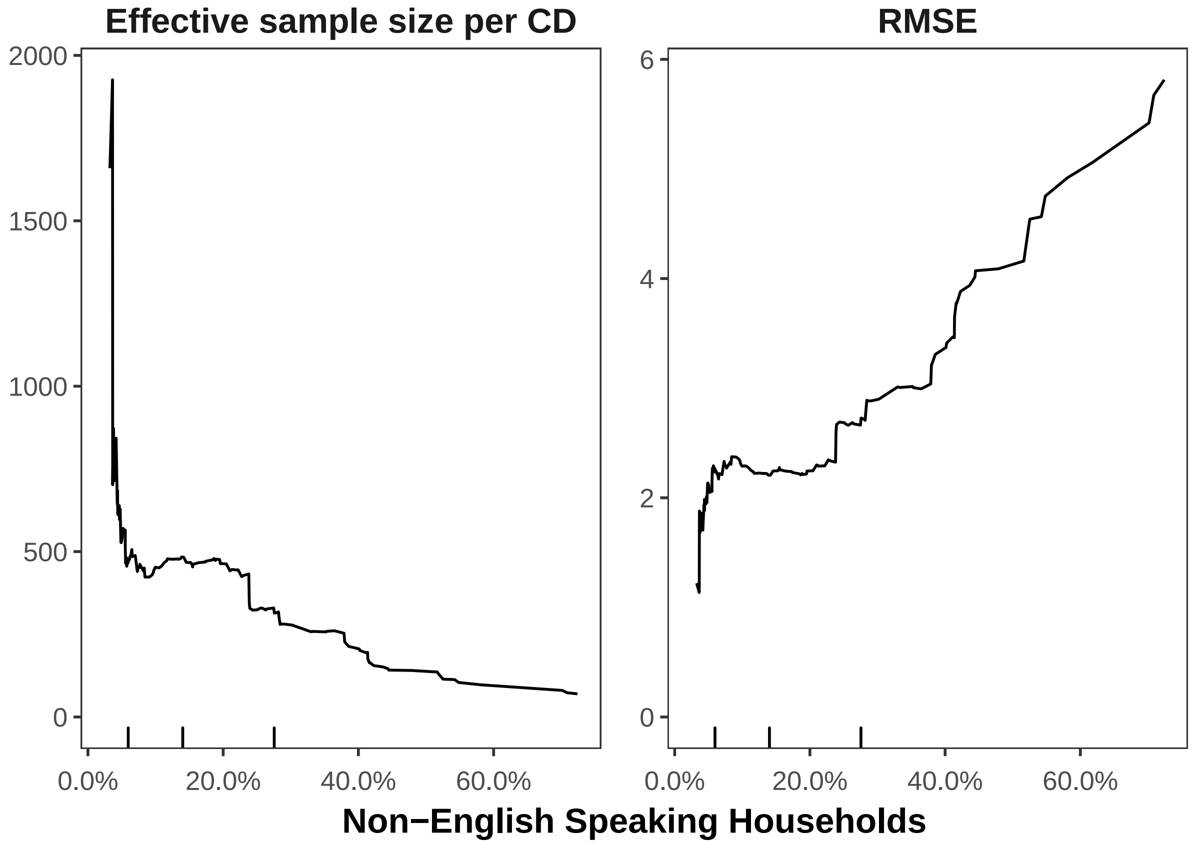

Figure 7: Performance of model-assisted post-stratified estimates for support for same-sex marriage, as divisions are included by their proportion of individuals that speak a language other than English at home. As divisions with more non-English speaking households are added to the analysis, predictive accuracy degrades, evidenced by the decline in effective sample size per electorate (left panel) and root mean square error (right panel). Tick marks on the horizontal axis indicate 25th, 50th and 75th percentiles of non-English speaking households across Australia's 150 divisions. The long tail of high-non-English districts to the right of the 75th percentile produces more than a doubling of RMSE.

This highlights two important considerations for the design of future VAA tools: the need for a multi-lingual questionnaire and the inclusion of a demographic question asking either whether English was the participant's first language, or if they spoke another language at home (to allow for post-stratification).

Despite the complexities of these particular divisions, the effective sample size of our estimates overall was 69.89 per division (using the linear regression line as the baseline for comparison). Comparatively, division-level polling for the Labor vote share at the 2016 Australian federal election, with average samples of 626 per poll, had an effective sample size of was 92 (Jackman and Mansillo 2018). This indicates that to replicate our results with data collected through classical polling methods with these accuracy rates, instead of VAA-derived data, we would need to have samples of approximately 382 respondents per division, or a survey of 57 thousand respondents in total. Considering these data include questions on vote intention, demographics and attitudes towards 30 policy issues, in either case this would have been a multi-million dollar study. Inclusion of a question on language, a multilingual VAA, or both, would likely have further increased accuracy and the effective sample size.

##Comparison with standard survey data

As a final test, we fit a similar model to a survey with a standard sample size, collected using classical methods. We use the gold standard for Australian federal election surveys: the Australian Election Study. In 2016, respondents were contacted via mail, and could complete the survey either by hand-completing the copy recieved in post, or through an online survey. The response rate was 22.5 per cent, with an *N* of 2818. When we hope to study public opinion or voter behaviour for small groups or in small areas (such as legislative districts) using data of these nature, there are significant limitations. One of these is obvious. The average sample per electoral division is small. Just 18.8 respondents. The division with the largest number of respondents had 35. The division with the smallest only 2.

Regardless, we want to use this key resource for political science research in Australia as comparison, to understand whether VAA's like Vote Compass provide any utility compared to the otherwise best available survey data for researchers in most countries (unlike the US, huge surveys using random sampling such as the Cooperative Congressional Election Study are not usually conducted in Australia and most other electoral democracies).

We conduct this comparison by fitting a multilevel regression model with post-stratification (MRP; Gelman and Little 1997; Park, Gelman, and Bafumi 2004; Lax and Phillips 2009b; Warshaw and Rodden 2012; Buttice and Highton 2013; Wang et al. 2015; Hanretty, Laud-

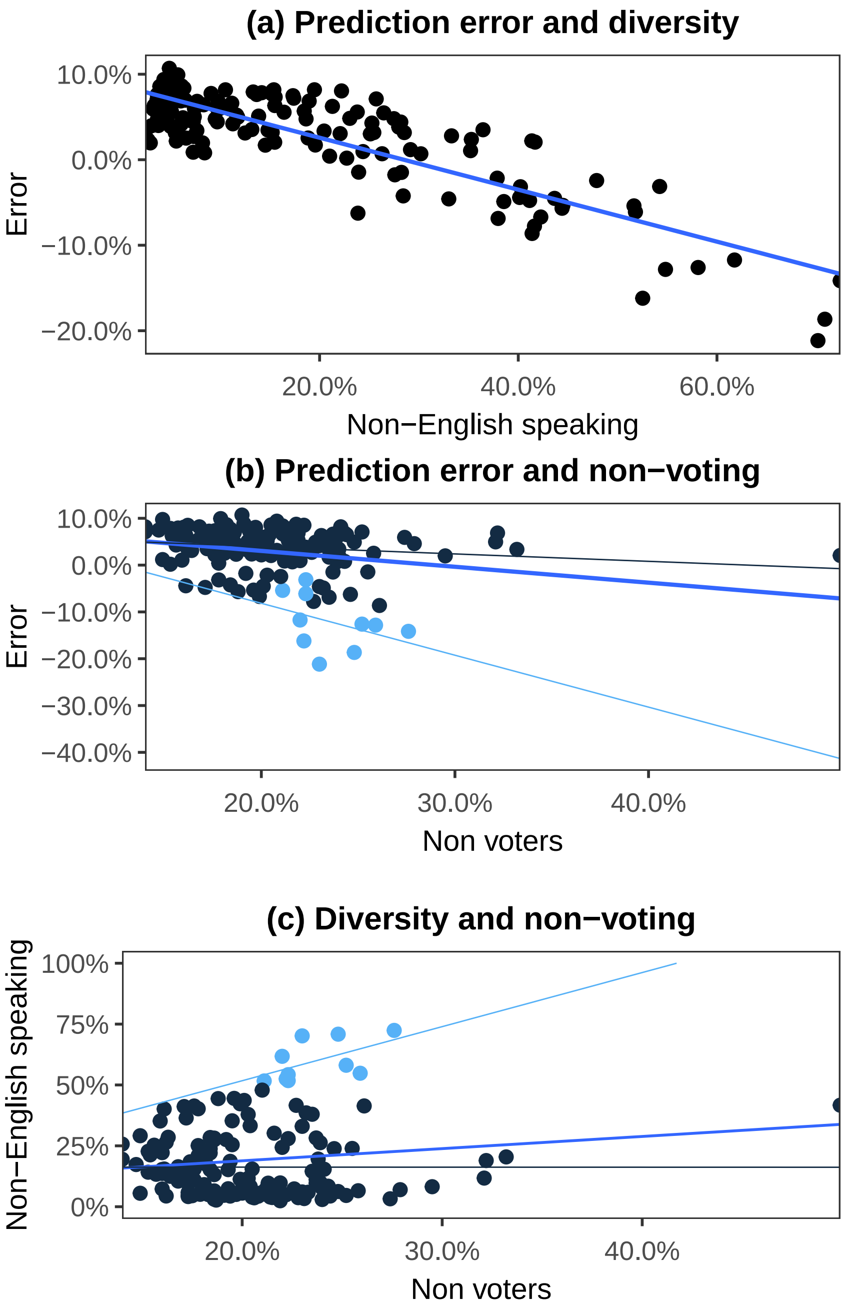

Figure 8: Comparison between the predictive error in our same-sex marriage model using the Vote Compass data with electorate diversity and non-voting in the same-sex marriage poll (shown on next page). This shows (a) how the error in our estimates on average increases with language diversity, (b) that there is also an association between error and turnout in the same-sex marriage vote, and (c) how diversity and turnout are also associated. Points and the lighter curves in (b) and (c) are also coloured, with those shaded lighter blue indicating our ten outlier divisions with non-English speaking populations greater than 50 per cent.

erdale, and Vivyan 2016) to the 2016 AES data. We use a specification identical to the models fit to the 2016 Vote Compass data, and also include division-level predictors for the two-party vote share of the centre-right Coalition parties and the proportion of the population in same-sex relationship (as recorded by the 2016 Australian Census). This helps improve the fit of these models, and is the type of specification we would use to obtain reasonable division-level inferences from these data.

As a comparison between Figures 6 and 9 shows, the model we fit to the AES data underperforms the models fit to Vote Compass. The average effective sample size of the AES model is 11.89 per division. For the Vote Compass models it is 56.1. If we use the linear regression line as the baseline instead, the gap between the models narrows, but not completely; with an average effective sample of 41.88 per division for the AES model and 69.89 for the Vote Compass models. That is, the comparison with the regression helps the AES estimates more, despite the question asked in this survey being closer to the plebiscite question.

Discussion and conclusion

With the use of model-assisted procedures, opt-in online VAAs can provide previously underutilized data sources for the study of public opinion in small geographic areas and for small population groups. These platforms provide a relatively cheap and truly massive sample size without sacrificing a large range of policy questions. This is done through the inclusion of a media partner that provides significant in-kind and financial support, and by giving respondents a payoff for completing the survey — showing them where their issue preferences sit compared to the major political parties — supplying an incentive for tens or hundreds of thousands of voters to finish a relatively long survey in its entirety at low cost.

We outline how a model-driven approach combined with post-stratification procedures using Census data and election returns can turn these heavily biased data into a tool for small group and area estimates.

This approach offers access to data that — when managed correctly — is roughly the equivalent of a classical survey with approximately 57 thousand respondents. Considering these data include questions on vote intention, demographics and attitudes towards 30 policy issues, this would likely have cost more than $1 million to collect through conventional means.

This is not to suggest these data are a replacement to high quality classical surveys where these are available. Rather, they complement them. They provides a useful tool for scholars focusing on patterns in public opinion and voter behaviour across many small electoral areas

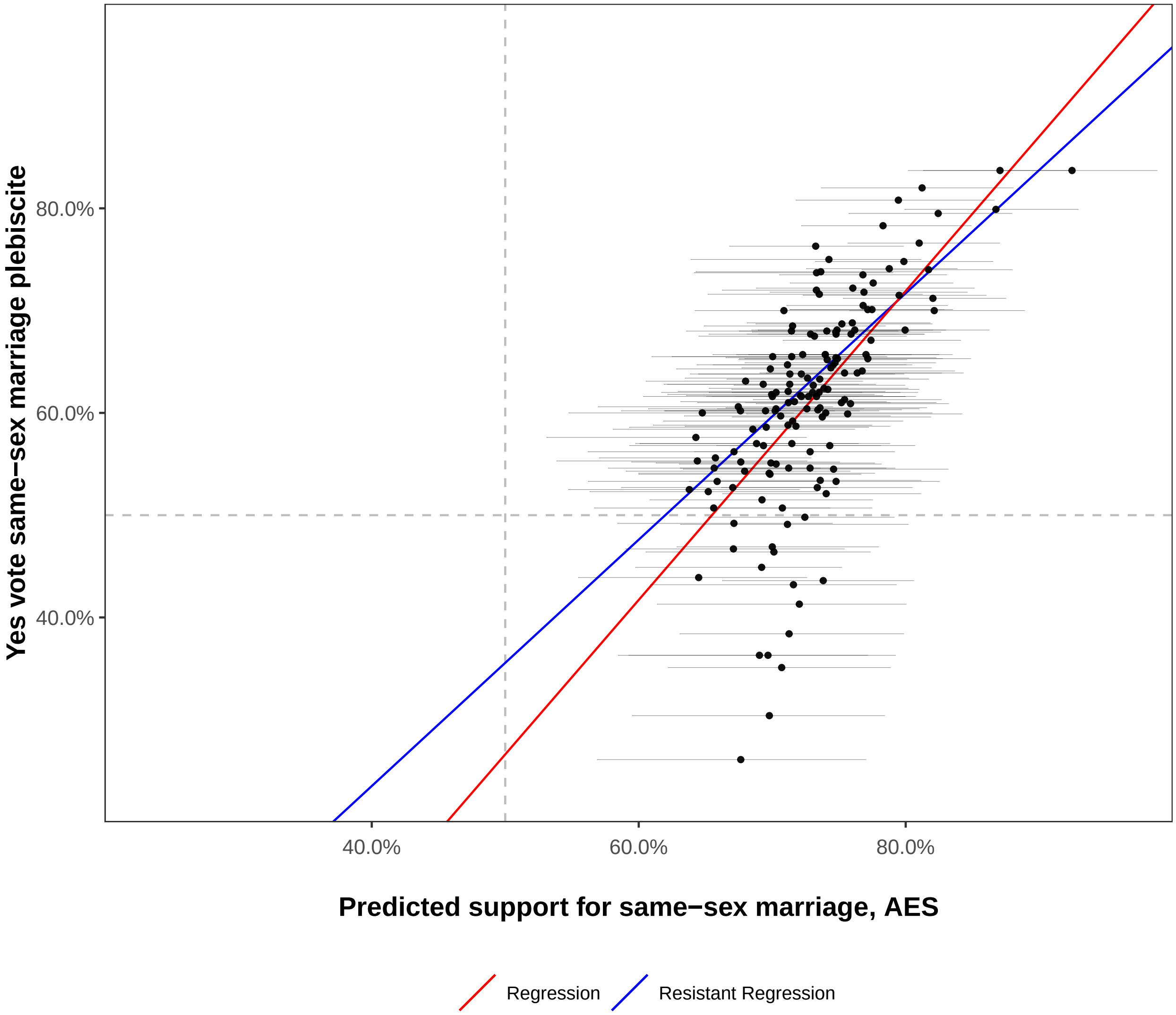

Figure 9: Comparison of estimated and observed support for same-sex marriage in Australia's 150 Commonwealth electoral divisions. Each point shows support for same-sex marriage, estimated from the Australian Election question on whether same sex couples should be given the same rights to marry as couples consisting of a man and a woman? (positive responses were coded as support for same-sex marriage), and the actual level of support in the marriage plebiscite. The blue line is a resistant regression estimator (minimising a weighted sum of residuals, using Tukey's biweight) between our Vote Compass estimate and the plebscite result. The red line is ordinary least squares regression.

and population groups where the resources are not available to survey tens of thousands of respondents using traditional methods.

In addition to providing division-level estimates, these models provide a joint distribution at the level of individual cells in our post-stratification frame. This provides the opportuntity to examine any combination of population subgroups included in the survey and post-stratification frame. As VAAs have the scope to ask participants about a wide range of issues, this creates the potential to examine spatial and demographic patterns in public opinion across many issues. A level of analysis previously difficult or impossible to achieve in almost all representative democracies where the collection of large-N data has been rare.

Besides highlighting the utility of these data, we also document two important considerations for the design of future VAA tools: the need for a multi-lingual questionnaire and the inclusion of this demographic information to ensure the diversity of the population is properly captured in estimates of public opinion.

This is an initial step in a larger research agenda. The size of this dataset, combined with our methodology for correcting for bias, provides unique opportunities to make inferences on representation and the congruence of party policies and citizens' issue preferences. Additionally, these data are not restricted to Australia. Vote Compass VAAs have also been run in elections in other Anglo-American democracies, with large samples collected in the US, UK, Canada and New Zealand and other countries (Linden and Dufresne 2017). Other voter advice tools have also been run in Australia, Europe and the US. When used with the appropriate estimation techniques, these datasets offer a range of opportunities to conduct comparative estimates of public opinion and the function of representative democracy at a level of detail greater than has previously been possible.

References

Alvarez, R. M., I. Levin, P. Mair, and A. Trechsel. 2014. "Party Preferences in the Digital Age: The Impact of Voting Advice Applications." *Party Politics* 20 (2): 227–36.

Buttice, M. K., and B. Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21 (4): 449–67.

Canes-Wrone, B., T. Clark, and J. Kelly. 2014. "Judicial Selection and Death Penalty Decisions." *American Political Science Review* 108 (1): 23–39.

Canes-Wrong, B., D.W. Brady, and J.F. Cogan. 2002. "Out of Step, Out of Office: Electoral

Accountability and House Members' Voting." *American Political Science Review* 96 (1): 127–40.

Clinton, J.D., S. Jackman, and D. Rivers. 2004. ""The Most Liberal Senator"? Analyzing and Interpreting Congressional Roll Calls." *PS: Political Science and Politics* 37 (4): 805–11.

Dahl, R.A. 1971. *Polyarchy: Participation and Opposition.* New Haven: Yale University Press.

Downs, A. 1957. *An Economic Theory of Democracy.* New York: Harper.

Garzia, D., and S. Marschall. 2016. "Research on Voting Advice Applications: State of the Art and Future Directions." *Policy & Internet* 8 (3): 376?390.

Gelman, A., and Y. and Lee D. Ghitza. 2010. "Public Opinion on Health Care Reform." *American Journal of Political Science* 8 (1).

Gelman, A., and T. C. Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23 (2): 127–35.

Hanretty, C., Lauderdale B. E., and N. Vivyan. 2016. "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods* 27 (1): 1–21.

Jackman, S., and L. Mansillo. 2018. "The Campaign That Wasn't: Tracking Public Opinion over the 44th Parliament and the 2016 Election Campaign." In *Double Disillusion: The 2016 Australian Federal Election*, edited by A. Gauja, P. Chen, J. Curtin, and J Pietsch, 133–58. Canberra: ANU Press.

Johnston, R. 2017. "Vote Compass in British Columbia: Insights from and About Published Polls." *Journal of Elections, Public Opinion and Parties* 27 (1): 97–109.

Kapelner, A., and J. Bleich. 2013. "BartMachine: Machine Learning with Bayesian Additive Regression Trees."

Kastellec, J.P., J.R. Lax, M. Malecki, and J.H. Phillips. 2015. "Polarizing the Electoral Connection: Partisan Representation in Supreme Court Confirmation Politics." *The Journal of Politics* 77 (3): 787–804.

Key, V. O. 1961. *Public Opinion and American Democracy.* New York: Knopf.

Lax, J.R., and J.H. Phillips. 2009a. "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103 (3): 367–86.

———. 2009b. “How Should We Estimate Public Opinion in the States? American Journal of Political Science.” *American Journal of Political Science* 53 (1): 107?121.

———. 2012. “The Democratic Deficit in the States.” *American Journal of Political Science* 56 (1): 148–66.

Linden, C. van der, and Y. Dufresne. 2017. “The Curse of Dimensionality in Voting Advice Applications: Reliability and Validity in Algorithm Design.” *Journal of Elections, Public Opinion and Parties* 27 (1): 9–30.

Miller, W.E., and D.E. Stokes. 1963. “Constituency Influence in Congress.” *American Political Science Review* 57 (1): 45–56.

Park, D. K., A. Gelman, and J. Bafumi. 2004. “Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls.” *Political Analysis* 12 (4): 375–85.

Poole, K.T., and H. Rosenthal. 2007. *Ideology and Congress: A Political Economic History of Roll Call Voting*. New Brunswick, NJ: Transaction Publishers.

Selb, P., and S. Munzert. 2011. “Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information.” *Political Analysis* 19 (4): 455–70.

Shirley, K.E., and A. Gelman. 2015. “Hierarchical Models for Estimating State and Demographic Trends in Us Death Penalty Public Opinion.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178 (1): 1–28.

Tausanovitch, C., and C. Warshaw. 2013. “Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities.” *The Journal of Politics* 75 (2): 330–42.

———. 2014. “Representation in Municipal Government.” *American Political Science Review* 108 (3): 605–41.

Vavreck, L., and D. Rivers. 2008. “The 2006 Cooperative Congressional Election Study.” *Journal of Elections, Public Opinion and Parties* 4: 355–66.

Vox Pop Labs. 2016. “Vote Compass Methodology.” http://voxpoplabs.com/votecompass/methodology.pdf.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. “Forecasting Elections with Non-Representative Polls. International Journal of Forecasting.” *Survey Methodology* 31 (3): 980–91.

Warshaw, C., and J. Rodden. 2012. “How Should We Measure District-Level Public Opinion on Individual Issues?” *The Journal of Politics* 74 (1): 203–19.

Wheatley, J., C. Carman, F. Mendez, and J. Mitchell. 2014. “The Dimensionality of the Scottish Political Space: Results from an Experiment on the 2011 Holyrood Elections.” *Party Politics* 20 (6): 864–78.