Can we improve multilevel regression and poststratification (MRP) through new ways to leverage information?

A Thesis

Presented to

The Department of

Politics, International Relations, and Philosophy

Royal Holloway, University of London

For the Degree of

Doctor of Philosophy

Benjamin Lobo

01 May 2022

Declaration

I Benjamin Lobo hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Benjamin Lobo

30 September 2021

Abstract

Multilevel regression and poststratification (MRP) has become a popular and important small area estimation method in social sciences. The method enables researchers to reliably estimate public opinion in small areas such as constituencies, states, and districts. Since it was first developed, numerous studies have extended and advanced the method. This thesis asks whether we can further improve MRP estimates with alternative methodological approaches. Each chapter explores how these methodologies could be applied with MRP, and whether each improves MRP estimate accuracy. As a preface to the introduction of these alternative methods, in chapter two, the thesis asks: what is standard practice for the application of MRP in social science? I address this question through a systematic review of 86 studies which use MRP. Drawing on the collective wisdom of researchers to date, the chapter details how each of the main MRP characteristics are typically applied in practice. In chapter three, I explore whether using cross-validation lasso regression can improve variable selection for MRP applications. I explore how the method should be applied, and whether this method is an improvement on what might be considered current standard practice for variable selection. The results are somewhat mixed but show that lasso could be a useful tool. I argue that incorporating lasso into the model building process, alongside standard variable selection approaches, would represent an improvement over current MRP variable selection practice. In chapter four, I explore whether an unevenly distributed sample among small areas might be a useful strategy when applying MRP to electoral forecasting. The chapter set outs how this method could be

applied and explores whether it improves MRP. Overall, the results show the method can improve estimate accuracy in important small areas, and in turn, can improve the probability of correctly forecasting an electoral outcome. In the final chapter, I explore how we can use informative priors with MRP. I employ a two-stage prior elicitation method with MRP and apply to estimating vote choice at numerous elections. The results indicate the method can improve estimate accuracy and precision. The results also give some indication that this method could be useful for improving sub-group inference and computational efficiency. However, improvements are inconsistent across different elections, and often improvements are only significant for the smallest sample sizes.

Acknowledgements

The work that has gone into this thesis has benefitted from academic, practical, financial, and emotional support from numerous people and organisations.

I first wish to thank my wonderful supervisors, Chris Hanretty and Oliver Heath. I am especially grateful to Chris, who has provided immeasurable support and advice throughout. Without Chris, this thesis would simply not have been possible. It was also his inspirational teaching of statistics and R during my Masters that started me on my quantitative research journey. Oli has been a great source of encouragement throughout my PhD. His critique of my work, and advice on how best to shape and frame a paper has considerably improved this thesis.

I would like to thank numerous staff and students in the department who have reviewed, commented, and helped me throughout my time at Royal Holloway. I am particularly grateful to Dr Cassilde Schwartz, who has read and provided thought provoking feedback at each stage of the PhD process. Equally, feedback and academic support from both Dr Kaat Smets and Dr Sofia Collignon has helped me, and improved my work significantly. To my entire PhD cohort, I thank you for providing academic support, as well as friendships.

To my parents, none of this would have been possible had you not encouraged me from a young age to pursue whatever I found interesting. To all my family and friends, your support throughout this have been invaluable, and time spent with you has provided much needed respite from the PhD. A special thanks goes to Allan, Charlene, Felix, Jake, Jo, Laura, Lydia, May, Mona, Pete, Tom, and Zhamilya for

generously giving their time to proofread chapters.

Finally, I would like to thank both Royal Holloway and Survation for providing the financial support for this project to be completed.

Table of Contents

List of Tables

List of Figures

Dedication

For Ellie.

Chapter 1

Introduction

Recent methodological advances in public opinion research have led to the development of multilevel regression and poststratification (MRP). The method enables researchers to estimate opinion or behaviour of populations in sub-national geographic units known as small areas, such as states, districts, and constituencies. Standard public opinion polls rarely offer the opportunity to infer below a national level picture, while older methods have been shown to achieve poorer estimate accuracy than MRP (Lax and Phillips, 2009b; Warshaw and Rodden, 2012).

Since its first development (See Gelman and Little, 1997; Park et al., 2004), the method has become impressively popular (Leemann and Wasserfallen, 2017) and is regarded by some as the 'gold standard' of small area estimation (Selb and Munzert, 2011). Overall, the development of the method means researchers are better equipped to understand what the public think and how they behave.

To date, numerous studies have made significant contributions towards demonstrating the efficacy of the method and setting out how MRP can be applied in practice (See Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Buttice and Highton, 2013; Kastellec et al., 2016). Subsequently, the method has been applied across numerous fields in social science. However, although the method has gained popularity and is

increasingly used in a variety of research settings, there are still challenges to the wider application of the model (Gelman et al., 2016). This thesis seeks to contribute to the continued development of the method by assessing whether alternatives to the standard application can improve MRP estimates. In doing so it is hoped the thesis will advance our understanding of the method overall, and advance our understanding of how best to apply the method.

The motivation for this research is based on two inter-related realities of public opinion research. First, research analysing sub-national small area opinion is important if not vital for social science fields. For extensive periods in social science, research has overlooked geographic variation in opinion, instead focusing on the national-level picture (Rodden, 2010). Today however, this is no longer the case, with analysis of sub-national opinion a growing area of academic interest. There now seems to be both a broad recognition that opinion geographically varies within countries, and a recognition that studies and academic fields need to account for this.

Second, directly inferring small area opinion from surveys is simply not feasible. In nearly all applications, surveys are designed to be nationally representative and do not enable researchers to investigate below a national or regional level. This means researchers are restricted in what research questions they can address, and the extent to which they can truly analyse the nature of opinion within countries.

Both these realities are why the development of MRP has been important for social science research on public opinion. The method enables researchers to reliably estimate public opinion and behaviour within small areas. This has led to a plethora of studies applying the method to address numerous key substantive questions in social science. Importantly, the method comes with little (financial) costs to the researcher. Free to use analysis tools (R and Python) and state-of-the-art probabilistic programming tools (Stan), make the method *relatively* easy to implement. Similarly, advances in computational power mean that estimating - even highly complex - MRP

models is relatively quick and easy.

But as an emerging methodology, there is still much we do not know about the method. For example, we do not know how the method is applied in practice. Although there is often wide-spread recognition that the method has become popular and has grown in use in social science, there is no documentation which accounts for its use in studies. There is also much to learn about the best ways we can and should apply the method. There is no one-size fits all with MRP. Across different settings, the model will require tailoring according to the unique aspects of each application. For example, when estimating a variety of different small area opinion, the relevant and predictive variables will vary and researchers need to incorporate different sources of information to account for this. In order to better equip researchers with the knowledge how best to apply MRP, we need to continue the development of the method exploring how we might apply to alternative settings. Finally, we also need to continue exploring how we might improve the method, asking whether variations to the standard application might enhance our small area opinion estimates.

This thesis seeks to contribute towards our understanding of the method in both how it is applied, and how it should be applied. In this thesis, I do so from the perspective that MRP is successful at estimating small area opinion because it is highly adept at combining a variety of information sources to produce estimates. It combines information about individual person types through respondent-level survey data, information about the small areas through area-level variables, and structural information about the population through the poststratification frame.

The thesis will assess whether we can better leverage information or leverage new information for MRP applications. In the three main chapters, I assess three methodologies, exploring how we might incorporate these alternative approaches with MRP applications, and importantly, assess the degree to which each improves MRP estimates. Chapter three explores whether we can improve how we select information

(variables) for MRP. Chapter four asks whether we can use past information about the estimated opinion or behaviour to determine a sampling strategy that improves MRP estimates. The final chapter examines whether we can incorporate further information from past models through informative priors. For the remainder of this introductory chapter, I introduce MRP, discussing the method background, theory, and development, and conclude with a roadmap for the thesis ahead.

1.1 Background

Public opinion surveys

Public opinion and behaviour are core areas of interest for much of social science. What people think and how they behave are important questions for researchers interested in better understanding the human world. In order to investigate these topics, researchers make use of various forms of qualitative and quantitative methods. In quantitative studies, surveys are perhaps the most common method used. They provide researchers the opportunity to speak to a small but representative group of the population delivering a snapshot of public opinion or behaviour.

The method offers a relatively quick and inexpensive means to carry out research, with results which are - in theory at least - representative of the wider population. Such is the popularity of the method that today, opinion survey research is a large commercial industry (Gelman et al., 2016).

Unfortunately, opinion surveys are not without their problems. Indeed, there is an extensive range of methodological issues associated with them (for a detailed discussion, see Groves, 1987). The main issue that concerns this thesis is the inability of surveys to reliably capture opinion or behaviour below the national level. If a researcher is interested in the opinion or behaviour of small area populations such as states, districts or constituencies, surveys are for the most part unsuitable.

Surveys are designed to be nationally representative with the goal of enabling researchers to make claims about the population under study. In practice, researchers choose a sample size which will provide estimates of national public opinion with a margin of error of +/-3% at the 95% confidence interval. This means that a typical survey sample is too small for scholars to be able to make inferences about sub-national small area opinion (Warshaw and Rodden, 2012: 203; Buttice and Highton, 2013: 449). For example, using a sample size of 2,000 to investigate opinion in US states would result in a margin of error of +/-15%.[1] In the United Kingdom, to achieve a +/-3% margin of error for 632 constituencies, assuming a population of 100,000 in each, a sample of over 500,000 would be required.

There are examples of surveys with large enough sample sizes to enable researchers to directly infer small area opinion. However, these are mostly confined to the United States, where the number of areas is relatively small (i.e. 51 states including District of Columbia). And even in the United States, it is widely recognised these surveys are simply not suitable when investigating opinion at the sub-state level (Hersh and Nall, 2016: 292). Surveying small area populations is possible, although applications are rare, expensive, and carried out infrequently meaning that comparison between small areas is often impossible (Park et al., 2004: 375; Lax and Phillips, 2013: 1).

The need for subnational estimates

The inability of surveys to gauge sub-national opinion or behaviour is not a problem for the majority of studies. However, for some topics there is a need to be able to gauge opinion or behaviour of sub-groups in the population. For dyadic representation, researchers need reliable estimates of opinion for small area populations (for early and key studies on dyadic representation, see Miller and Stokes, 1963; Hill and Hurley, 1999; Weissberg, 1978). This is because the study analyses the extent to which

[1]This calculation is based on the assumption that respondents are evenly distributed among states.

elected representatives vote in line with the opinion of the populations that elect them (for a discussion on how MRP has been important for representation studies see Caughey and Warshaw, 2019). MRP examples include Lax and Phillips (2012), whose study demonstrated that there is a gap between the views of US voters and the way senators vote across 39 different policy areas. Lewis and Jacobsmeier (2017) use MRP to estimate dynamic state-level support for same-sex marriage. Using these MRP estimates, they showed that direct democracy institutions can improve congruence between opinion in states and elected state representatives. In the UK, Hanretty et al. (2017) show that MPs are responsive to the opinion of constituents on support for same-sex marriage and Britain's membership in the EU.

Small area estimates are similarly important for electoral forecasting. To predict an electoral outcome we require reliable and accurate estimates of voting behaviour in each small area. In parliamentary democracies, who governs is not based on which party receives the highest national level vote share, but rather which party wins a majority of parliamentary seats. To forecast the electoral outcome, we therefore need estimates which enable the prediction of the party winner in each small area rather than a national level picture. Even in presidential electoral systems such as the United States, the president is not elected by 'popular vote' but rather based on electoral college votes. To forecast the presidential election we need to be able to forecast vote share in each state. Examples of electoral forecasting with MRP include predicting the 2016 US presidential election, the UK 2017 general election, and the UK 2016 EU referendum (Kiewiet de Jonge et al., 2018; Wyatt et al., 2016; Lauderdale et al., 2020).

In other fields, limiting analysis of opinion to a national level might not inhibit research altogether. However, only analysing opinion at a national level might mean that we fail to capture idiosyncrasies that are apparent in societies. In recognition of geographic heterogeneity in opinion, numerous researchers have applied MRP to

various fields in social science, exploring both the causes and implications of variation in opinion. For example, studies have used MRP to explore geographic variation of opinion towards gender equality (Koch and Thomsen, 2017), policy mood and ideology (Enns and Koch, 2013), attitudes toward environment (Howe et al., 2015; Mildenberger et al., 2016; Howe, 2018; Kaufmann et al., 2017 ; Fowler, 2016, 2017; Eun Kim and Urpelainen, 2018), attitudes towards immigrants (Butz and Kehrberg, 2015, 2016), and migration intentions (Williams et al., 2018).

Estimating sub-national opinion or behaviour

Awareness that surveys are incapable of providing sub-national opinion or behaviour estimates is by no means recent. Researchers have long been developing methods to enable them to investigate small area opinion or behaviour. One of the simplest methods is to use a proxy of opinion or behaviour and treat the proxy as a reliable indicator of the opinion or behaviour of interest. For instance Berkman and O'Connor (1993) use state-level membership of an abortion rights group and number of Christians as indicators of state-level opinion towards abortion. Another example is that of Berry et al. (1998), who used voting records of elected representatives to construct a state-level ideology score. However, this method is considered problematic as it assumes that opinion or behaviour maps onto the proxy often without supporting empirical evidence (Pacheco, 2011; Lax and Phillips, 2013). Ultimately, the proxy measure was viewed as a sub-optimal solution, but viewed as satisfactory in the absence of an alternative method (Norrander and Wilcox, 1999).

Subsequent scholars developed disaggregation, a method which aggregates numerous polls into one large N survey. The combined survey has a large enough sample size to enable researchers to directly estimate small area opinion or behaviour. It was first developed by Erikson et al. (1994), who used the method to establish state-level public opinion on government policies. Brace et al. (2004) later used the method to

investigate how political ideology changes over time in US states.

Although an improvement on proxy measures, the need for identical survey questions across numerous polls means there is a limited number of topics which can be investigated (Kastellec et al., 2016; Buttice and Highton, 2013; Lax and Phillips, 2009a). Furthermore, because there are typically an insufficient number of polls in a single year, researchers must collate polls across numerous years. As a result, estimates are not snapshots at one given time period, but the average across time periods. In turn, this means we cannot measure temporal changes in opinion or behaviour (Kastellec et al., 2016; Howe et al., 2015). The method is also only possible in applications where the number of areas is relatively small (Warshaw and Rodden, 2012). This is why the application is mostly restricted to the United States, where studies estimate opinion or behaviour within 51 states. For applications with a larger number of small areas, the method is no longer a feasible way to estimate small area opinion or behaviour.

MRP

In response to the need for better methods to estimate small area opinion or behaviour, scholars developed multilevel regression and poststratification. MRP was first developed in the United States with the work of Gelman and Little (1997), and subsequently Park et al. (2004). Their work built upon previous research which had developed poststratification (Pool et al., 1965; Weber et al., 1972), combining it with multilevel modelling. Importantly, studies have demonstrated that MRP is better able to estimate public opinion than previous methods. When compared to disaggregation for example, MRP consistently produced more reliable and accurate estimates of small area opinion (Park et al., 2004; Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Pacheco, 2011). Even Buttice and Highton (2013), who are cautionary in their support of MRP, contend that MRP is a superior method of opinion estimation than

disaggregation.

The first application of MRP estimated opinion and behaviour at US state-level, but has since been applied to a range of small areas including US congressional and senate districts (Warshaw and Rodden, 2012), US cities and towns (Tausanovitch and Warshaw, 2013), and US counties (Kaufmann et al., 2017). Outside of the United States, MRP has been applied in German electoral districts (Selb and Munzert, 2011), Swiss Cantons (Leemann and Wasserfallen, 2016, 2017), UK constituencies (Hanretty et al., 2016, 2017), EU states (Toshkov, 2015; Kolczynska et al., 2020), EU regions (Lipps and Schraff, 2021), Canadian federal districts (Mildenberger et al., 2016), and South African electoral districts (Ornstein, 2017).

The method can also be applied to estimate variables that are not strictly opinion or behaviour. For example, studies in both the US and the UK have combined MRP with item response theory (IRT) to produce estimates of broad political sentiment (Tausanovitch and Warshaw, 2013; Hanretty et al., 2017). Outside of social science, notable applications include population health studies (See Zhang et al., 2014; Downes et al., 2018; Downes and Carlin, 2020a, 2020c, 2020b).[2]

Theory

MRP sits within the wider field of small area estimation methods (SAE). The field incorporates a wide-range of advanced methodologies which produce estimates for small area populations. Small areas in this field are unique geographic sub-national units, including states, districts and constituencies. As discussed above, direct estimation from surveys is not possible, instead this field focuses on developing model-based methods which typically use survey data to estimate means or quartiles for small area populations (for a discussion on SAE, see Rao and Molina, 2015; Pfeffermann, 2013).

[2]This thesis will focus entirely on the use of the MRP in social sciences. The method is largely unchanged between applications in social and health sciences. However, the development of the method in the two applications is largely separate in the literature. Furthermore, the use of MRP requires subject-specific tailoring, and lessons here may not be applicable to other contexts.

There are two stages of an MRP model: first, a multilevel model is fit to the data; second, the model estimates are weighted to small area population figures through the poststratification frame.

At the first stage, a multilevel model is estimated where individuals are nested within small areas. At level-1 of the multilevel model, opinion or behaviour is modelled as a function of individual-level characteristics (Lax and Phillips, 2013: 5). Individual-level variables are obtained from survey data and are typically demographic characteristics of respondents. This is not a technical requirement but necessary due to limitations on available data.[3] In most applications of MRP, all individual-level variables are estimated as random intercept terms. This means each individual-level category is drawn from a common distribution of parameter effects (Hanretty, 2019).

At level-2 of the multilevel model, the parameters for each small area are estimated. The parameters are again estimated as random intercept terms drawn from a common distribution. To help the model better identify variation between small areas, most MRP applications use additional area-level variables. The inclusion of these variables improves parameter estimation by shrinking each small area parameter towards other areas with similar characteristics (Gelman and Hill, 2007: 269; Hanretty et al., 2016: 574). Models may also include additional levels to further improve estimate accuracy. This means that individuals are nested within small areas, and these areas are in turn nested within a higher geographical unit such as a region. The additional level can improve estimate accuracy as small area estimates are shrunk towards the mean of each higher geographical unit.

One of the key reasons why MRP is particularly effective at estimating opinion or behaviour in small areas is because of partial pooling in the multilevel model. The classic regression case either fully pools respondents or does not pool respondents at all (Gelman and Hill, 2007: 254; Lax and Phillips, 2013: 11; Warshaw and Rodden, 2012:

[3]The is predominantly because of data requirements for the poststratification frame, as will be explained in further detail below.

206). In the fully pooled case, the model uses the full sample but does not account for differences between small areas. In the case of no-pooling, the model estimates each group separately using only using sample from each small area. Multilevel modelling, on the other hand, partially pools respondents across small areas. This means small area estimates are directly estimated from respondents within each small area, as well as from the wider sample. In practice, small area estimates are shrunk towards the population mean. The degree of shrinkage is decided by the model internally, with areas with fewer respondents shrunk towards the overall mean to a greater extent (Lax and Phillips, 2013: 11). Pooling is also greater when the area level variance is smaller, that is, less variation in opinion between small areas.

Although there are numerous ways to specify the model, for explanatory purposes here, I will describe a simple multilevel model with four individual-level variables (age, education, ethnicity and gender) and three area-level variables (region, past vote share and religion).[4]

In most applications the multilevel model is estimating Y, a binary variable. The model is typically a multilevel logistic regression model, estimating where $Y = 1$. It can be written as follows:

$$\begin{aligned} Pr(Y_i = 1) = logit^{-1}(\beta^{\theta} + \ \beta^{Female} \ \cdot Female_{[i]} + \ a_{s[i]}^{Area} + \ a_{j[i]}^{Age} + \\ a_{k[i]}^{Education} + \ a_{l[i]}^{Ethnicity} + \ a_{m[i]}^{Region}), \ for \ i = 1, ..., n. \end{aligned} \tag{1.1}$$

Here i indexes the individual respondent in the survey, for $i = 1, ..., n$. β^{θ} refers to the global intercept, β^{Female} is the parameter for female respondents.[5] a_s^{Area}, a_j^{Age},

[4] I use these variables because the systematic review identified them as some of the most commonly used in MRP applications.

[5] The parameter for gender (female) can also often specified as a varying intercept term.

$a_k^{Education}$, $a_l^{Ethnicity}$ and a_m^{Region} are all varying intercept terms. Where:

$$\begin{aligned} a_j^{Age} &\sim N(o,\ \sigma^2)\ for\ j = 1, ..., J \\ a_k^{Education} &\sim N(o,\ \sigma^2)\ for\ k = 1, ..., K \\ a_l^{Ethnicity} &\sim N(o,\ \sigma^2)\ for\ l = 1, ..., L \end{aligned} \tag{1.2}$$

The parameters for each category of age, education and ethnicity variables are drawn from a normal distribution with mean of 0 and some variance (σ^2). The a_s^{Area} term is estimated similarly, but is estimated as a function of the area-level variables, that is:

$$a_s^{Area} \sim N(a_{m[s]}^{Region} + \beta^{vote} \cdot vote_{[s]} + \beta^{Religion} \cdot Religion_{[s]},\ \sigma^2) \tag{1.3}$$

Here the small area parameter a_s^{Area} is estimated as a function of the Region of the small area, the past vote share at the small area-level, and the proportion of a religious group in each small area. Region is itself a modelled random intercept term, where:

$$a_m^{Region} \sim N(o,\ \sigma^2)\ for\ m = 1, ..., M \tag{1.4}$$

The method's first stage estimates opinion or behaviour for each person-type across all small areas. That is, the model produces estimates for each combination of all individual-level categories within each small area. For example, if we include age (with categories of 18-24, 25-44, 45+) and ethnicity (with categories of white, non-white), the model would produce estimates of opinion or behaviour for 18-24 white people, 18-24 non-white people and so on.

The second stage of MRP involves poststratification. The process of poststratification is principally a weighting scheme, where the estimated opinion or behaviour for each person-type are weighted according to the proportion each person-type represents in the population. The poststratification frame is constructed by accessing

census data, or similarly large individual-level survey-type data, which provides the joint-distributions for all individual-level variables included in the first level of the multilevel model. Using the above example, we need data which will enable us to determine the proportion of each small area population which is 18-24 and white, 18-24 and non-white, and so on for each combination of all individual-level variables included. The poststratification procedure is as follows:

$$Y_s^{Pred} = \frac{\sum_{c \epsilon s} N_c \, \pi_c}{\sum_{c \epsilon s} N_c} \tag{1.5}$$

Where N_c is the population count for each small area, and the π_c is the person-type estimate. The final small area estimations of opinion are thus person-type estimates weighted according to the proportion they represent in the population within each small area.

Requirements for MRP to perform well

MRP is a method which is useful to forecast opinion or behaviour in small areas. However the method may not always be applicable nor the best method for estimating opinion. Data requirements of MRP are a non-trivial restriction to the wider application. The method requires at least three separate data-types, each limited by distinct but related ways.

1. Individual-level survey data

   Individual-level variables are obtained from survey data, which capture the opinion or behaviour of interest as well as characteristics of each respondent. In most academic applications, scholars typically use free-to-access surveys. This can be a limiting factor as researchers can only estimate opinion or behaviour included in the survey. And further, can only use surveys where the necessary individual-level characteristics are available. In most applications these variables

are demographic characteristics, but this is not a technical requirement. All individual-level characteristics used are also required for the poststratification frame, where there are far greater data limitations.

2. Area-level variables

   Area-level variables are used at level-2 of the multilevel model. They are typically continuous variables and capture information about the small area or the population of the small area. For example the percentage which previously voted for a certain political party, or the percentage of the population from a certain religion. These variables are often the least restrictive, as in most applications, governments publish a variety of free-to-access statistics about small areas.

3. Poststratification frame data

   Data for the poststratification frame is the hardest to obtain. It requires the joint-distribution proportions for all individual-level categories in each small area. In the United States, this data is available from the census or the ACS, a large census-like survey. In most other countries, the required micro-level data is not accessible. This has led to researchers developing alternative methods to produce a poststratification frame (See Hanretty et al., 2016; Leemann and Wasserfallen, 2017). While these methods overcome data limitations present for the construction of a poststratification frame, they require further methodological investment from researchers.

Another requirement for MRP to perform well concerns the variation in opinion between small areas (inter-subgroup), and within each small area (intra-subgroup). MRP is best used to capture inter-subgroup variation in opinion or behaviour. Subgroups here can be small area populations or demographic subgroups. Should there be limited

variation between subgroups, the benefits of MRP are likely to be negligible. For example, at the onset of the UK coronavirus lockdown, if we estimated the proportion in each constituency which believed the pandemic was the biggest issue facing the country, we would likely have seen little difference across small areas and the benefits of MRP would be minor[6]

If there is significant intra-area group variation (that is significant variation of opinion within each small area) MRP may produce poor estimates. This is because if there is variation in opinion that we are not accounting for in the model, the model may be poor at estimating opinion. This risk can be averted - or at least minimised - by the inclusion of individual-level variables which capture the intra-group variation.

The method is also known to struggle at reliably estimating low incidence opinion (Hanretty, 2019). This means if only a small proportion of a population have a certain opinion or exhibit a certain behaviour, MRP will most likely perform poorly. For example, in electoral forecasting MRP estimates for small parties will typically be less accurate and reliable than estimates for larger parties.

Sample size

Part of the appeal of the method is that MRP can produce accurate and reliable estimates with *relatively* small sample sizes. For instance, Lax and Phillips (2009b) argued their study demonstrated MRP was able to produce accurate estimates in 50 US states with a total sample of around 1,400. Kastellec et al. (2016) echoed such findings, again arguing that a sample of around 1,400 was sufficient to estimate in 50 US states. Warshaw and Rodden (2012) extended the application by exploring how the method performed at lower geographical levels. They found sample sizes of 2,500 and 5,000 were sufficient for US congressional districts (436 small areas) and for

[6]Ipsos Mori have tracked the single biggest issue among UK public since 2010. They have shown that at start of lockdown, 85% of UK public said that the Coronavirus / pandemic was the biggest issue. See `https://www.ipsos.com/sites/default/files/ct/news/documents/2021-02/issues_index_jan21_cati_v1_public.pdf`

Senate districts (1,942 small areas), respectively. Outside the United States, Hanretty et al. (2016) demonstrated the effectiveness of the method with between 8,000-12,000 respondents for 632 small areas.

Importantly, the necessary ratio of respondents to the number of small areas is lower in applications where there are a large number of small areas (Hanretty, 2019). Put a different way, when we are estimating for a larger number of small areas, the required number of respondents per small area will be lower than if we were estimating for relatively few small areas. This is important because it means that we are able to estimate for a large number of small areas with obtainable sample sizes. Whereas for disaggregation, to estimate in a large number of small areas the method would require a sample size that is simply not feasible.

Survey non-response

MRP is a particularly useful technique in the context of the growing problems of survey data. Response rates for random probability phone surveys have continually declined in recent periods making it increasingly difficult to obtain samples. This has led to the increase in non-probability online samples, which are far easier to obtain, but can have significant selection effects. Both methods suffer from survey non-response and a lack of representative samples. MRP however, is a method that can alleviate concerns about survey non-response (Park et al., 2004: 376; Gelman et al., 2016: 5). Indeed, Wang et al. (2015) demonstrated that MRP is able to produce accurate estimates of voting behaviour even with highly unrepresentative samples. Numerous studies have since demonstrated that MRP can be highly effective at estimating reliable opinion or behaviour with non-probability samples (Kennedy and Gelman, 2020; Cerina and Duch, 2020a, 2020b). This means the method equips researchers with a viable mechanism to deal with the problems associated with modern survey samples.

Dynamic opinion and behaviour

One of the drawbacks of disaggregation was that typically the method was not able to measure change in opinion or behaviour across years. This was because the method required researchers to merge multiple surveys - often across years - into one large sample to disaggregate. MRP, on the other hand, is highly adept at capturing temporal changes in opinion or behaviour within small areas. Pacheco (2011) developed a pooled approach to use MRP to estimate dynamic opinion and behaviour at US state-level (Pacheco, 2011). She has subsequently applied this method to measure dynamic opinion towards a range of political and social issues (Pacheco, 2012, 2013, 2014; Pacheco and Maltby, 2017, 2019). More recently researchers have estimated dynamic opinion or behaviour by including the time-period as a parameter in the model. In this specification, time-periods can be included as a linear or quadratic term (See Gelman et al., 2016), or estimated with far greater flexibility through estimating the time-period with a spline (See Kolczynska et al., 2020).

Sub-group inference

MRP is also a particularly useful method when we are interested in the opinion or behaviour among subgroups in small areas. For example, Ghitza and Gelman (2013) showed how including multiple and cross-level interactions can significantly improve inference among subgroups of the population. A more recent variant of MRP - which includes structured priors - has been shown to potentially enable researchers to investigate much smaller sub-groups than was previously possible (Gao et al., 2021). In this study, structured priors enabled them to produce reliable estimates for up to 72 separate age categories (Gao et al., 2021).

Extensions of MRP

The initial studies on MRP validated the method through applications to the United States, where the necessary data to construct a poststratification frame is readily available and *relatively* easy to access. Outside of the United States, there is far more limitation on what data is available, restricting the countries where MRP could be applied. In response, Selb and Munzert (2011) developed a slight variant to MRP which did not have such strong data demands. Through an application to German electoral districts, they demonstrated the proficiency in estimating opinion with their alternative (Selb and Munzert, 2011).

Further studies have since focused on developing alternative methods to construct poststratification frames. All these alternative methods lessen the strict data requirements of MRP and enable the application to countries where the necessary census-like data is not available. Hanretty et al. (2016) used a raking procedure to construct a poststratification frame. Although this would still require large N sample data, the necessary sample size is significantly lower than for the standard MRP case. Leemann and Wasserfallen (2017) developed what they call synthetic poststratification, which can use standard survey sample sizes to synthetically construct the joint-distributions necessary. More recently, Cerina and Duch (2020b) have developed a method of constructing a poststratification frame through imputation.

Other extensions of the standard MRP model have focused on replacing the multilevel model component of MRP with alternative regularisation methods. These alternative approaches produce estimates for each person-type which are then poststratified to the population. Bisbee (2019) developed BARP, a method which uses Bayesian additive trees and poststratification. The non-parametric approach used here is argued to provide improved regularisation and can be particularly useful when modelling complex (or deep) interactions. Ornstein (2020) introduced stacked regression and

poststratification (SRP) which employs a model averaging technique.[7] Ornstein argues that this technique is beneficial with interactions - especially cross-level interactions - and consistently demonstrates improvements over the standard MRP case. The improvements in SRP are also particularly notable when estimating outcome variables that are not cultural topics. In another machine learning application, Cerina and Duch (2020b) have employed a random forest model to replace the multilevel model in MRP.

Along the same lines, some extensions have replaced the standard multilevel model with the goal of automating variable selection for MRP. In a similar manner to Ornstein (2020), Broniecki et al. (2021) used a Bayesian model averaging method with poststratification. Their method was used to improve variable selection, and combines the estimates from a range of variable selection models. They showed consistent improvements when compared to the classic MRP case across a large range of different estimated opinions. Another feature selection method combined with MRP is that of Goplerud et al. (2018), whose Sparse multilevel regression and poststratification (sMRP) utilises lassoPlus to induce sparsity acting as a feature selection process. This method was shown to be particularly useful in situations where deep interactions are modelled.

1.2 Roadmap

I begin the thesis in chapter two where I set out how the method has been applied in social science to date. This is achieved by a systematic review, where I document how each main model component is specified across 86 unique studies within social science. Across all studies, I document what might be considered the key components of MRP, including sample size, variables, small areas, topics, and time periods. The chapter

[7]The model averages estimates from lasso, k-nearest neighbour, random forest and gradient-boosting.

asks the question: is there a standard application of the method, and if so, what does this standard specification look like? The chapter seeks to identify areas where there is standard practice in the use of MRP, and explore areas where variation in the application exists. This is accompanied by some discussion on the considerations that drive variation in MRP specification. Rather than act as a document of best practice, the chapter is intended to improve our understanding of how the method is applied in social science.

In chapter three, I explore whether we can improve variable selection for MRP. Although there has been some notable and innovative work on variable selection and MRP (see Broniecki et al., 2021; Goplerud et al., 2018), this chapter seeks to explore how we can incorporate automated variable selection on all MRP variables, while maintaining the standard MRP form. This is achieved by exploring whether a variant of lasso (Group-lasso interaction-NET) can be used to simultaneously select individual-level variables, area-level variables, and interactions. Through an application to forecasting 2017 Conservative party vote share in GB constituencies, I first explore how this method is best applied to the MRP case. I subsequently assess whether this method improves MRP estimate accuracy when compared to path dependency and theory-based variable selection.[8] On the one hand, the chapter argues that the results cannot be used to support the outright use of automated variable selection in the form used here. On the other hand, the chapter contends that when used alongside other variable selection methods, incorporating lasso into the model building process represents an improvement on what is most likely current standard practice for MRP variable selection.

The fourth chapter investigates whether we can improve MRP predictions by oversampling respondents from certain small areas. In electoral forecasting, to correctly predict an electoral outcome we often require a higher degree of accuracy in certain

[8]I use the phrase 'path dependency' to refer to a situation where a researcher chooses identical variables to previous studies.

small areas known as marginals. This chapter explores whether a method which allocates a higher proportion of the sample to marginal small areas can improve MRP prediction accuracy. Through a simulation study and two real-world applications in the UK and US, I set-out how this could be applied to MRP electoral forecasting, and assess whether this method improves MRP estimates. I argue that, whilst not useful in all settings, in electoral forecasting an uneven sample distribution can improve estimate accuracy in certain small areas, which in turn can improve the probability of correctly predicting an electoral outcome. The results have direct implications for the use of MRP to forecast elections and, I argue, wider implications for the use of MRP where survey samples are typically unevenly distributed among small areas.

The fifth chapter explores how we can use informative priors in MRP. The method of MRP is increasingly estimated in a Bayesian framework which means researchers must specify priors for the models. To date, most use either weakly-informative or non-informative priors. This chapter explores whether using informative priors could be useful for MRP estimates. Specifically, through the application to electoral forecasting, I assess whether we can improve MRP by combining it with a two-stage prior elicitation method. In practice, this takes the form of imputing prior distributions from past election model posteriors. I first set out how this method could be applied, and second, assess whether this method improves estimate accuracy, estimate precision, parameter estimation, and computational efficiency. The chapter shows that informative priors can both improve and harm estimate accuracy and precision. Whether the priors improve or worsen accuracy is a result of numerous factors including the similarity between elections, the reliability of past data/models, and the sample size. The results also show the method can improve computational efficiency and could have the potential to improve subgroup inference. Overall however, given the variability in accuracy improvements, I argue the risks involved in the two-stage prior elicitation method outweigh the potential benefits.

The concluding section of this thesis looks back across each chapter, summarising the main findings and contributions from chapter three, four, and five. This is followed by a discussion of the main limitations of this thesis, potential future research areas, and finally, a note of advice for the applied MRP researcher.

Chapter 2

What is MRP standard practice?

Multilevel regression and poststratification (MRP) is a method developed to estimate public opinion of populations who live in sub-national small areas.[1] Sitting within the wider field of small area estimation methods, MRP is a model-based technique which uses survey data along with statistical modelling to produce estimates of opinion in small areas.

Since its first development, it has become popular and its use has been impressively high for a new method (Leemann and Wasserfallen, 2017: 1,003). The method's rise in popularity can broadly be attributed to two main aspects. First, it is an improvement on older alternative methods which estimate small area opinion (see Lax and Phillips, 2009b). Second, the method has enabled researchers to investigate sub-national public opinion on topics that were not previously possible (Lax and Phillips, 2009a: 371). Together, these have led many to conclude that MRP is theoretically and statistically superior to alternative methods (Fowler, 2016), and represents the 'gold standard' of sub-national small area opinion estimation (Selb and Munzert, 2011).

Following the first introduction of the method (see Gelman and Little, 1997; Park et al., 2004), numerous studies have demonstrated its proficiency at estimating small

[1] Small areas are geographically exclusive units within a country, typically below national and regional levels. These include administrative, political or census defined areas such as districts, constituencies, states, counties or equivalent.

area opinion or behaviour (see Lax and Phillips, 2009b; Warshaw and Rodden, 2012), have compared the relative contributions of different MRP components (see Hanretty et al., 2016), and developed various extensions of the standard model (see chapter 1). This has led to the continued growth in the use of MRP, with it becoming an important tool for researchers in academia and further afield. While there are studies which provide a worked example of how to apply MRP to estimate opinion or behaviour (see Hanretty, 2019; Lopez-Martin et al., 2019; Kennedy and Gabry, 2020), there is still limited understanding of how the method is typically applied by researchers.

The following chapter seeks to address this gap by reviewing how MRP has been applied across social science to date. In doing so, the chapter will explore to what extent there is a standard practice in the application of MRP, and if present, what does this standard practice look like? The chapter will also explore and identify where there is variation in the application of MRP. Differences in application are driven by either methodological or substantive decisions associated with the topic of interest. Where there is variation in the application, the chapter will provide some discussion on the methodological and substantive decisions that researchers face.

To achieve the above, the chapter carries out a systematic review of the application of MRP in social sciences, documenting how each of the MRP model components are applied. In effect, the research will draw on the collective wisdom of social science researchers, with the aim of advancing our understanding of the method's application. Below, I describe the systematic review process, followed by the presentation of descriptive statistics for each MRP model component.

2.1 Method

This study was designed to summarise how MRP models have been used in previous social science research. To achieve this, the research must first identify relevant

literature which has used MRP, and second, document how each model characteristic has been used.

Selection criteria

To select the relevant studies and provide some practical parameters, the following selection criteria were used.

1. Studies in social science

   The application of MRP is not restricted to social science, with notable applications in public health (see Zhang et al., 2014). However, this research will focus solely on the use of MRP in social science. This is because there may be significant differences in how MRP is applied across academic disciplines. If these differences are nontrivial, the standard practice documented in this chapter will no longer be applicable to the application in social science.

2. Published studies

   The analysis presented in this paper only takes into consideration published work. This decision is based on both methodological and practical considerations. From a methodological standpoint, if we are primarily concerned with identifying standard practice, only considering published and peer reviewed work seems reasonable. Although this approach could be argued to risk publication bias, because MRP is not concerned with statistical significance, I believe this risk is minor. From a practical standpoint, this approach was necessary to create a feasible research project. An initial search of both published and unpublished papers produced over 140 studies. While a larger N is often preferable, I deemed this to be an unfeasible number of papers to document.

Finding studies

Identifying relevant research in a systematic manner was one of the main challenges of this project. This is primarily because published research which uses MRP spans across numerous social science fields, lacking a uniform title, topic, or abstract. To ensure I had the best opportunity to identify all relevant studies, I used two independent approaches: keyword and citation-based search approach. The keyword approach searched for studies published between 1997-2018 in two academic study databases (Jstor and Web of Science). The search terms were as follows:

- State level public opinion
- District level public opinion
- Constituency level public opinion
- Multilevel regression

The search terms were intentionally broad to reduce type I error. Once duplicates and invalid results had been removed, the full list of potential articles from this stage totalled 1,350.

The citation based approach searched for papers which had cited two articles considered as the founding papers: Gelman and Little (1997) and Park et al. (2004). These papers are recognised as the first to publish work solely focused on developing MRP and applying it to social science. To find papers which had cited the above studies, I used Google Scholar's 'cited' function. I accessed the lists in September 2019 and found 159 and 321 results for Gelman and Little (1997) and Park et al. (2004), respectively. Once combined there was a total of 373 unique potential papers.

To identify relevant studies from the lists of potential papers, I first excluded studies which had not used MRP, or I was not able to freely obtain access to. Second, I excluded papers outside of social science (for example health studies) and papers that had not yet been published. I also excluded papers which used MRP estimates

from a previous study or studies which did not included details on model specification. The final list of papers totalled 86.[2] I show the breakdown of studies identified by each approach in table 2.1.

Table 2.1: Number of papers by identification method

Identification method	N
Key-word	13
Citation	37
Both	36
Total	86

As noted above, identifying relevant studies by paper methodology was challenging. The key-word search produced a large number of false-positives and was an inefficient method of finding relevant literature. The citation-based approach, while more efficient, was more prone to false-negatives as it relied on citation of two studies rather than paper content. Nonetheless, the overlap of studies identified by the two approaches gives confidence that together, the two strategies were able to identify a majority of studies which have applied MRP.

Documenting model characteristics

Once relevant studies had been identified, I created a data-set which documented model characteristics for each *unique* MRP model in every paper. For the purpose of this study, I classed a model as unique when at least one of the model characteristics differed from other models in the same paper. For example, if one paper estimated five models which were identical except for the opinion or behaviour estimated, each would

[2] I provide a list of the 86 studies in appendix A.1.

be included as a unique model in the data-set. This could lead to the analysis giving too much weight to studies which estimated numerous MRP models. To factor for this, all descriptive statistics only count each unique feature from each paper once. For instance, in the above example the five unique opinions or behaviours would be counted in the descriptive statistics. However, all other model characteristics would only be counted once as these were all identical cases. Although this is an imperfect solution to documenting model characteristics, I believe this is the most suitable approach. Using this method, papers which estimated numerous models were not over-represented in the descriptive statistics, but the range in unique models/characteristics was still captured. Table 2.2 shows N for papers, unique models, and unique characteristics.

Table 2.2: Number of unique papers, models and characteristics

	Unique N
Papers	86
Unique MRP models	441
Estimated opinion or behaviour topic	396
Sample size	212
Small areas	105
Individual-level variables	104
Interactions	91
Area-level variables	118
Hierarchical levels	116
Time periods	102

Bayesian models	86

Model characteristics were determined by an initial review of MRP literature. The process was iterative with numerous amendments before the final list of characteristics was confirmed. Although the list did not capture every single possible component of MRP models, it accounts for what could be considered the core characteristics of standard MRP models. Below I provide the list of characteristics, notes on coding and what will be reported for each.

Estimated opinion or behaviour topic: The topic of each model was coded according to the topic of the estimated opinion or behaviour. The code-frame for topics was adapted from the Pew Research Centre list of research topics, which provides a comprehensive list of public opinion and behavioural research areas.[3] I coded each opinion or behaviour according to the main categories in the Pew research topics. The exception is when a specific sub-category appeared more than five times. In these instances, I included the sub-category as a separate code. For instance, 'Abortion' is included in the Pew Research 'Politics & Policy' category, but I included as a separate category as more than five models estimated opinion on this topic. Where a model opinion or behaviour had clear overlap across topics, I counted both topics. The figures reported are the proportion of models which estimated opinion or behaviour on each topic.

Small areas: I report the small area as well as the country which the small area belongs to. In some cases, a model used fewer small areas than there are. As I was interested in the model and not the geographic units, I recorded the number of small

[3]The full list and inclusion criteria is provided in the appendix A.2. The Pew research topics can be accessed here: `https://www.pewresearch.org/topics-categorized/?menuItemId=725ceef3d3181b99e26f459ffd55a01a`

areas used in the model.

Sample size: The sample size was recorded as the total sample for each model. Figures were taken from the paper, or if not reported, the published survey sample size. If a study estimated temporal changes in opinion by estimating a separate model per time period, I treated the numerous models as one and report the total combined sample size.

Hierarchical levels: I report the number of hierarchical levels included in each model. For example, a three-level model would include individuals nested in states, which in turn are nested in regions. I code a model as two-levels when a researcher fails to explicitly state that they use more than two-levels.

Individual-level variables: I report the number of individual-level variables used, the combined total of categories for all variables, and the list of variables used. The combined categories are the sum of all individual-level variable categories. The final list of variables is the result of grouping analogous variables together. Where interactions are included, I count each individual variable separately. I report the proportion of unique models which use each individual-level variable.

Area-level variables: I recorded the area-level variable and the variable type (categorical or continuous). The list of variables was re-coded grouping similar variables into over-arching categories.

Interactions: I report the number of number of interactions included in MRP models and the list of interacted variables.

Time periods: I documented time period as each unit of time that estimates are produced for. For example, both one day and one year were recorded as one unit. I report the proportion of models which estimate across time periods, the average number of time periods, and the estimation method. Estimation method was recorded as either a static model for each time period, pooled sample estimates, or directly including time periods as a model parameter.

Bayesian modelling: I report the number of models which state they were estimated as a Bayesian model. Models which do not explicitly state, or it could not be directly inferred, were not recorded as Bayesian models. I also report prior specification of Bayesian MRP models.

Missing information

Broadly papers which have used this method fall into two categories: those with a methodological focus, or those which focus on addressing a substantive topic. The latter typically dedicated significantly less discussion towards the MRP model. This meant for some papers, details on the MRP model were brief or for certain characteristics not reported. Furthermore, within the literature there was not always consistency how the model characteristics were explained. This led to some difficulties in interpretation, and in some cases, it was not possible to determine how they had used the model. In these instances, I did not record model characteristics. Though somewhat unavoidable, this means the analysis will give more weight to papers which report model specification clearly.

Descriptive statistics

In this chapter I report a range of descriptive statistics for the model characteristics identified above. To report the frequency with which a characteristic is used, I report the proportion of unique models which use the characteristic. For interactions and Bayesian priors, the N I report is a percentage of models which use interactions or are Bayesian. To report the range of a given characteristic, for example sample size, I report the median, first (25%) and third (75%) quartile.[4]

2.2 Analysis

Growing use of MRP in social sciences

The application of MRP in social sciences has grown because the method enables researchers to estimate sub-national small area opinion or behaviour. The growth in the use of the method can be seen figure 2.1, which shows study by year of publication. The histogram shows the proportion of studies which were published in each year, along with a smoothed density line (red dashed-line). The figure shows there is a general trend upwards starting in 2008. The method has clearly grown in popularity and seems to have been particularly popular in 2017-2018. The figure shows a downward trend in the past two years (2019-2020), but this is most likely due to the cut-off date for identifying studies. Although the figure only includes published studies, and excludes studies outside of social sciences, the figure demonstrates the growing popularity and use of the method since its first application to social sciences in 2004.

[4]I report the median rather than the mean as the distribution for most of the characteristics is skewed.

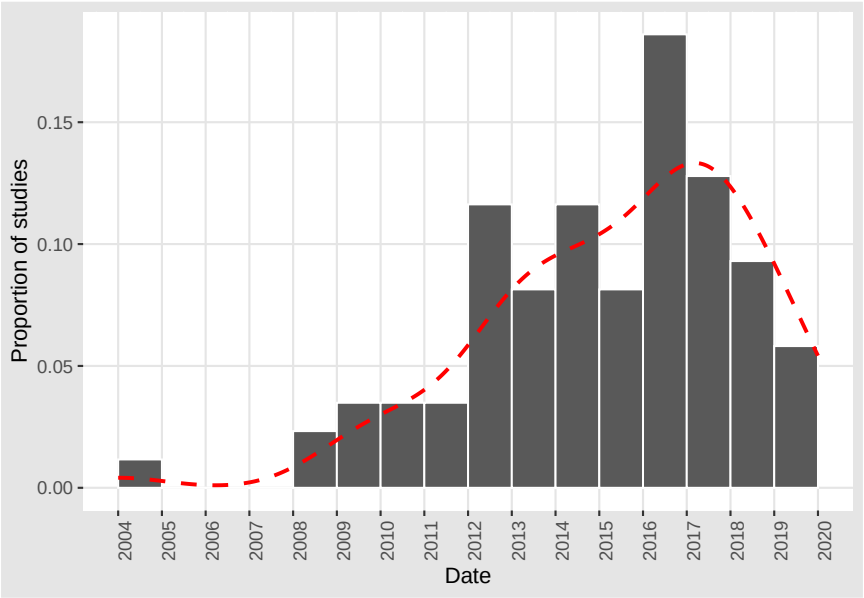

Figure 2.1: Study publication date.
Notes: showing proportion of studies by publication date.

Estimated opinion or behaviour

As set out in the introduction of this thesis, MRP was developed to produce reliable estimates of small area opinion or behaviour. This subsequently meant researchers have been able to investigate questions and topics that were previously impossible (Lax and Phillips, 2009a: 371). Yet, at present no comprehensive list of topics which MRP has been applied towards exists. Table 2.3 provides a list of opinion or behaviour estimated with MRP, grouped into overarching topics. In the left column I show the topic and the right hand column reports N, the percentage of unique models which estimated the given topic.

Table 2.3: Topic of MRP estimates

Topic	N
Gender & LGBT	17%
Economy & Work	15%

Elections & voters	10%
Politics & Policy	9%
Criminal justice	8%
Climate, Energy & Environment	7%
Health policy	6%
Immigration & Migration	6%
Abortion policy	5%
Education policy	4%
Family & Relationships	3%
Defense & National security	3%
Gun policy	2%
Race & Ethnicity	2%
Other	1%
Science	1%
Religion	1%
International Affairs	1%

Note: N = 396

From table 2.3 we can see that to date, MRP has been used across a relatively broad range of topics with no single dominant subject. Clearly, in social science there is no single standard topic which MRP can or should be applied to. Most topics estimate opinion rather than behaviour. The main exception to this is 'Elections and voters', where the models are estimating voting behaviour. Although there is variation in estimated opinion or behaviour, most topics in table 2.3 are established political cultural debates. That most studies fall under this overarching topic is most likely because of three reasons. First, MRP is argued to be adept at estimating opinion or behaviour for topics under this category, while seen as less useful for

economic topics (Buttice and Highton, 2013; Ornstein, 2020). Second, most studies use publicly available surveys, where the list of available questions are often focused on key economic and political debates. Third, these are key topics in social science and typically attract significant interest from researchers.

Variation in topics is driven by what interests the researcher, with choices of what to estimate being dependent on the research objectives of the wider study. However, methodological concerns most likely also guide researcher decisions, with researchers needing to consider the suitability of MRP to estimate their opinion or behaviour of interest. Researchers need to ensure that there is sufficient variation in opinion or behaviour within the population, including geographic variation among small areas. Furthermore, researchers need to ensure that their topic will have sufficient incidence within the population, as MRP is known to perform poorly when the opinion or behaviour has low incidence (Hanretty, 2019).

Small areas

Table 2.4 reports the small areas which the MRP models are estimating opinion or behaviour for. From left to right, I report the country of the small area, the small area itself, the number of small area units, and the percentage of unique models which estimate for the given small area. From Table 2.4 we can see most models estimate opinion in the United States, with a majority (53%) of all models estimating at state level. This is perhaps unsurprising because the method was first developed in the United States, the necessary data is more extensive and readily available, and the political system means that representation studies are particularly relevant there. Outside of the United States, MRP models have been applied in Canada, the United Kingdom, Germany, Switzerland, Denmark, Portugal, EU country level, and Taiwan. These countries are indicative of standard practice for MRP in terms of two necessary features. First, populations where there is variance in opinion by geography, and

second, countries where there is available data required for MRP.[5]

Table 2.4: Small areas

Country	Small area	Number of small areas	Models (N)
United States	States	43-51	53%
United States	Congressional Districts	435-436	13%
United Kingdom	Constituencies	632	5%
United States	Counties	1835-3143	5%
United States	Census administrative area	917-9981	4%
United States	State legislative districts	1942-4335	3%
Switzerland	Cantons	26	2%
United States	Cities (>20,000)	1600	2%
United States	Metropolitan areas	365-381	2%
Australia	Electoral districts	151	1%
Canada	Electoral districts	338	1%
Canada	Metropolitan areas	4	1%
Canada	Provinces	10	1%
Denmark	Congressional Districts	12	1%
EU	Countries	27	1%
EU	Regions	9	1%
Germany	Electoral districts	299	1%
Portugal	Administrative districts	18	1%
Taiwan	Census administrative area	23	1%
United Kingdom	Local authority	380	1%

[5]See 'Introduction to MRP' section which explains the necessary data required for MRP. Although alternative methods for the construction of a poststratification frame can alleviate such strict data requirements (see Hanretty et al., 2016; Leemann and Wasserfallen, 2017), most applications still use the standard method.

United States	Designated Marketing Area	101	1%

Note: N = 105

The number of small areas is interesting and highlights the variety of settings which MRP has been applied to. For instance, the range of small areas from 4-9,981 shows that MRP has been applied to estimating small area figures for both large (Canadian metropolitan areas) and small populations (US census administrative districts). However, most models do not estimate for such large of small populations. Indeed, if the goal is solely to estimate opinion or behaviour, the benefits of MRP will be somewhat negligent for such large populations. For such small populations, the model would need a large survey to obtain a sufficient number of respondents per small area, or if there is not sufficient coverage, many small area estimates will simply be a function of the area-level variables included. Decisions on which small area to estimate for will most likely be solely substantive, with the research question dictating what area the researcher requires opinion estimated within.

Sample size

One of the reasons that MRP was argued to be better than disaggregation was the ability to estimate small area opinion with much smaller sample sizes (see Lax and Phillips, 2009a). Indeed, early research into MRP argued that the method was able to produce reliable estimates of opinion from standard sample surveys of around 1,400 (Lax and Phillips, 2009b). Yet, despite acknowledgments that MRP enables researchers to use smaller samples, there are few guidelines on required minimum sample size. This is partly because sample size requirements for multilevel modelling are complicated. As opposed to standard regression, multilevel models have sample size requirements at multiple levels (Bell et al., 2008). In MRP, and multilevel modelling more generally, there is no consensus on minimum or optimal sample size. Often suggestions for

minimum or optimal sample size are study-specific and somewhat subjective. The below figures cannot provide a minimum nor best sample size for MRP, but advance our understanding by showing the standard range which most MRP sample sizes fall within.

Overall sample size

Table 2.5 reports the median, first, and third quartiles for total sample for all MRP models. The median figure of 7,742 is an indicator of the total sample size used by MRP models in general, while the first and third quartiles of 2,512 and 54,196, respectively, highlight the range of sample sizes used by most MRP models. The median figure highlights that typically MRP uses surveys which have a larger sample size than is conventional for nationally representative surveys.[6] The interquartile range of 51,684, shows that sample size for MRP varies significantly. This is largely because necessary sample size is somewhat dependent on the number of small areas and the number of time periods which opinion or behaviour is being estimated for. An increase in either of these typically requires a larger sample size.

Table 2.5: Total sample size

	Median	First quartile (25%)	Third quartile (75%)
All models	7742	2512	54196

Note: N = 212

Sample size and small areas

Sample size considerations are closely linked to the number of small areas used. Accordingly, the below analysis highlights how sample size varies depending on number of small areas. To explore this, I first group MRP models according to the number of

[6]Typically nationally representative survey samples are between 1,000-2,000.

small areas opinion or behaviour is estimated for. Among unique models, the range in number of small areas is from 4 to 9,981 with a median of 50. I divide MRP models according to the number of small areas, with > 51 small areas labeled as the upper-half, and models which estimate for ≤ 51 labeled as lower-half.[7]

Table 2.6 reports sample size according to the two groupings of small areas. From the median figures, it is clear the number of small areas affects sample size. Models in the upper-half (i.e. models estimating in > 51 small areas) use much larger sample sizes than those in the lower-half (i.e. models estimating ≤ 51 small areas). Upper-half models have a median sample size of 27,116 and a first and third quartile of 9,342 and 71,437, respectively. Whereas, lower-half models had a median sample size of 5,110, with a first quartile of 2,004 and third quartile of 19,766.

Table 2.6: Total sample size (Models grouped by number of small areas)

	Median	First quartile (25%)	Third quartile (75%)
Lower-half	5110	2004	19766
Upper-half	27116	9342	71437

Note: N = 212

As models which have a greater number of small areas will require more respondents to ensure adequate sample per small area, it is unsurprising that the total sample size for upper-half models is larger. However, although total sample size for upper-half models will be larger, sample size per small area will most likely be smaller. Table 2.7 reports the number of respondents per small area for all models, upper and lower-half categories. These assume an evenly distributed sample among small areas, something which is unlikely, but nonetheless are useful to better understand standard MRP

[7]Although the median number of small areas is 50, I decided to separate models by ≤ 51 and > 51. This is because models with 50 or 51 small areas are both estimating opinion or behaviour at US state-level. It did not make sense to analyse these models separately and so I increased the lower-half threshold. The two groupings therefore are unequal sizes, but serve the purpose here of exploring how sample size varies depending on number of small areas.

sample sizes.

Table 2.7: Average sample per small area

	Median	First quartile (25%)	Third quartile (75%)
All models	87	25	243
Lower-half	110	40	516
Upper-half	19	13	153

Note: N = 212

In contrast to table 2.6, which highlighted that lower-half models typically use smaller samples, table 2.7 shows that lower-half models use more respondents per small area than upper half models. For all models, MRP uses samples which have on average 87 respondents per small area. However, in the lower-half category this rises to 110 respondents, while for the upper-half, the number of respondents per small area decreases to 19. This pattern is most likely because estimating small area parameters is easier with a greater number of small areas, and therefore requires fewer respondents per small area (Theall et al., 2011). Whereas, with fewer number of small areas, the estimation of small area parameters is more difficult and therefore requires a bigger sample per small area. Furthermore, while we need adequate coverage of respondents per small area, we also need to ensure that overall we have a large enough sample to reliably estimate all parameters included in the model. This means that models with fewer small areas will typically have more respondents per small area.

Sample size and time periods

The final characteristic to account for when considering sample size is whether the model estimates opinion over a time period. These models require larger sample sizes to ensure adequate respondents within small areas and across time periods. Table 2.8 reports sample sizes controlling for time periods of a model. Once we account for time

periods, total sample size is much lower than the figures reported in either table 2.5 or 2.6. The median sample size for all models is 4,583, with a first and third quartile of 1,977 and 12,001, respectively. For lower-half category models the median sample size is 2,784 and for upper-half models the median sample size is 12,063.

Table 2.8: Total sample per time period

	Median	First quartile (25%)	Third quartile (75%)
All models	4583	1977	12001
Lower-half	2784	1607	7435
Upper-half	12063	5619	56024

Note: N = 212

Finally, table 2.9 reports respondents per small area accounting for time periods for all models, upper and lower-half small area categories. From table 2.9 we can see that the standard (median) sample size per small area is 49 for all models, 61 and 14 for lower and upper-half models, respectively. We again see a similar trend, that is, models with a greater number of small areas have fewer respondents per small area, while models with fewer small areas typically use samples with more respondents per small areas.

Table 2.9: Average sample per small area and time period

	Median	First quartile (25%)	Third quartile (75%)
All models	49	21	141
Lower-half	61	37	146
Upper-half	14	7	96

Note: N = 212

When considering sample size for MRP, researchers may refer to (Lax and Phillips, 2009b) who have argued that around 1,400 respondents are adequate to reliably estimate opinion in 49 small areas, while Warshaw and Rodden (2012) suggested that 2,500 and 5,000 samples were suitable for 436 and 1,942 small areas, respectively. However, analysis by Buttice and Highton (2013) noted that small sample sizes - such as 1,500 for 50 small areas - resulted in significant variation in MRP estimate accuracy. The results presented here are not meant to be a guide for minimum sample size but provide some results which are indicative of standard practice for the application of MRP in social science. The figures highlight how the number of small areas and time periods must be taken into account when considering sample size. Accordingly, the figures reported in table 2.8 and 2.9 - both of which take into account time period and number of small areas - perhaps best reflect the standard sample sizes used by MRP to date.

However, these tables show even when we account for number of small areas and time periods, there is still significant variation in sample sizes, as shown by the interquartile ranges. This variation will most likely be due to other practical and methodological considerations that researchers face. For example, researchers who wish to estimate a more complex model will require a larger sample size.

Hierarchical levels

One of the key principles of MRP is that opinion is modelled as a function of demographic and geographic variables hierarchically. That is, respondents are nested within small areas. To explore standard practice among MRP models, table 2.10 reports the number of hierarchical levels of MRP models. The table shows that across all models, most use three hierarchical levels. This means that individuals are nested within small areas, which are then in turn nested within a higher geographical area such as region. That the majority of models use three hierarchical level suggests that

most researchers believe small area estimates benefit from including an additional level.

For lower and upper-half small area categories, there seems to be little difference between how many hierarchical levels are used. Both have a median of three, while the small differences in first and third quartiles indicate that lower-half models sometimes use two hierarchical levels, while upper-half models sometimes use four hierarchical levels.

Table 2.10: Hierarchical levels

	Median	First quartile (25%)	Third quartile (75%)
All models	3	2	3
Lower-half	3	2	3
Upper-half	3	3	4

Note: N = 116

Individual-level variables

Individual-level variables are an important feature of MRP models. Indeed, opinion is modelled as a function of individual-level variables (Lax and Phillips, 2013: 5). Typically, these are demographic characteristics of respondents. This is not a technical requirement, but rather a function of what data is available to researchers.[8] Despite a limited set of possible variables, researchers are still faced with decisions of what to include and what to exclude. It is therefore useful to explore how many and which individual-level predictors are typically used by MRP models. This should advance our understanding of what type of variables are standard - and useful - for estimating opinion and behaviour.

[8]All individual-level predictors are required to be part of the poststratification frame. To build a poststratification frame we need joint distributions of all variables for each small area. This is restricted to a limited set of primarily demographic variables.

Number of categories

Table 2.11 reports the standard number of individual-level variables used by MRP models and the standard number of combined categories of the variables used. For example, if a MRP model used gender (Male, Female) and Ethnicity (White, Non-white), the number of variables would be two, with four categories. The first row shows that standard practice is to use four individual-level variables, as can be seen by the median and first quartile figure of four. There is a greater range in the number of categories that researchers use for the individual-level variables. Standard practice is for a combined total of 13 categories, but first and third quartiles of 8 and 16 highlight that there is significant variation in the number of categories typically used. Decisions on the number of individual-level variables will most likely be associated with sample size, with larger sample sizes enabling the use of more variables and a greater number of categories. Although, researchers may simply prefer to estimate simpler models with fewer variables, especially if they think that gains from additional individual-level variables will be marginal, or non-existent.

Table 2.11: Number of individual-level variables and categories

	Median	First quartile (25%)	Third quartile (75%)
Variables (N)	4	4	5
Categories (N)	13	8	16

Note: N = 104

Variable choice

As noted above, researchers face constraints on which individual-level variables are available for use. Nonetheless, their choices between even a limited set of individual-level predictors, represents assumptions and expectations that the given variables are predictive of the opinion of interest. Table 2.12 reports the proportion of unique

models which used each individual-level variable.

Table 2.12: Individual-level variables

Variable	N
Gender	92%
Education	89%
Age	86%
Ethnicity	70%
Past vote	17%
Poll	8%
Time period	8%
Income	6%
Marital status	6%
Employment	5%
Housing tenure	5%
Social Grade	5%
Ideology	3%
Foreign born	2%
Language	1%
Policy	1%
Religion	1%

Note: N = 104

From Table 2.12 it is clear four individual-level variables are the most common: gender, age, education and, ethnicity. Gender was used in 92% of the unique models analysed here, education 89%, age 86%, and ethnicity in 70% of models. There are some differences in how researchers use these variables, whether as categorical

or dichotomous, or whether combined with another variable, for instance gender & ethnicity. But clearly previous research has strongly relied on these demographic features as drivers of opinion across various topics. Other demographic variables include characteristics concerning a respondent's income, family or relationship status, or their housing situation. Past vote is also a widely used individual-level variable, owing to the frequency with which MRP is used to forecast elections and because vote choice is often highly predictive of other political opinions. The other non-demographic, but widely used variable, is time period. This refers to the time period which the respondent was surveyed.

The prevalence of the individual-level variables reported in table 2.12 will most likely be a combination of methodological and substantive decisions. From a methodological standpoint, researchers are limited by which variables are available in the survey and what variables they can include in the poststratification frame. From a substantive perspective, researchers should choose variables that they deem predictive of their opinion or behaviour of interest. While Buttice and Highton (2013) caution against assuming past variables will be predictive of other opinion, it is clear that past applications of MRP believe gender, education, age, and ethnicity are strongly predictive of a wide variety of opinion or behaviour.

Area-level variables

Area-level variables are an important aspect of MRP, as they are used at level-2 of the model and contribute towards the model better identifying area-level variance. The importance has been demonstrated by studies which have found that estimate accuracy is most impacted by the inclusion of area-level variables (see Buttice and Highton, 2013; Hanretty et al., 2016). It is therefore important that researchers choose relevant and predictive variables to ensure accurate estimates.

Number of variables

An important first consideration is how many area-level variables researchers use. Table 2.13 reports the median, first, and third quartiles for all models, lower and upper-half small area categories. From table 2.13 we can see for all models the median number of area-level variables is three, with most models using between two and five, as can be seen by the first and third quartiles. The number of area-level predictors is clearly related to the number of small areas. For models with fewer small areas (lower-half category), standard practice is to use two area-level variables, while for models with a larger number of small areas (upper-half category), the average is six. This difference can be explained by the difference in potential risk of overfitting the model. With fewer small areas, the number of area-level variables that can be used before the model overfits is lower than is the case for models with a greater number of small areas.

Table 2.13: Number of area-level variables

	Median	First quartile (25%)	Third quartile (75%)
All models	3	2	5
Lower-half	2	2	3
Upper-half	6	4	8

Note: N = 118

Type

Individual-level variables are always categorical as we need to poststratify by the variable categories. However, area-level variables can be either categorical or continuous. Table 2.14 reports the breakdown of area-level predictors into continuous and categorical variables. As can be seen in table 2.14, using continuous area-level variables is far more common than categorical variables. Indeed, continuous variables

account for over nearly three quarters of the area-level variables used. This is most likely because continuous variables are typically better at helping the multilevel model identify variation in small area opinion or behaviour. When researchers use categorical variables, it is mostly to account for geographical units above the small areas.

Table 2.14: Type of area-level variable

Variable Type	N
Continuous	72%
Categorical	28%

Note: N = 236

Variable choice

Table 2.15 reports area-level variables used among the MRP models under consideration here. The first column lists the variable, and the second column lists the percentage of unique models which include each area-level variable. The most common is higher geographical area, which accounts for nearly all the categorical variables identified in table 2.14. This variable is typically the region which the small area is in. It is used in MRP to encourage shrinkage of opinion or behaviour to the mean of the region. Although this is typically classed as an area-level variable, it is specified in the same way as small areas and individual-level variables - that is - as a random intercept term. The second most common variable used is party vote share. The prevalence of past vote share is unsurprising, as it is often available at small area level and can be highly predictive of political opinion or behaviour. The remaining variables are predominantly demographic, or characteristics of the population in the small areas. Some of the more commonly used variables include Religion (43%), Jobs / Employment (25%), and income of the small area population (22%). As with individual-level variables, data availability is perhaps the most significant factor in determining which variables are

used. However, at the area-level there is a much greater range in variables available, including free-to-access government statistics on small areas.

Overall, researchers tend to use variables that capture something (typically demographic) about the small area population. However, numerous studies also use variables which are specific to their estimated opinion or behaviour. For example, if interested in the opinion towards the environment, researchers might include variables about the environment in each small area. While the number and type (categorical or continuous) of area-level variables are methodological decisions, variable choice is substantive and should be based on which variables a researcher believes to be predictive of their topic.

Table 2.15: Area-level variables

Variable	N
Higher geographical area	69%
Past vote share	51%
Religion	43%
Jobs / Employment	25%
Population income	22%
Population ethnicity	14%
Urban / Rural	10%
Education	9%
Same-sex households	9%
Military veterans	8%
Political attitudes	8%
Geography	6%
Car / drivers	5%
Immigrant population	5%

Population age	4%
Population size / density	4%
Environmental	3%
Government attribute	3%
Health / disability	3%
Population language	3%
Poverty	3%
Electoral candidate	3%
Housing tenure	3%
Social grade	3%
Crime rates	2%
Family size	2%
Female population	2%
Marital status	2%
Party vote share	2%
Segregation	1%
Time period	1%
None	7%

Note: N = 118

Interactions

Researchers may also use interactions in the MRP models. Including interactions has been shown to improve prediction and inference, especially among sub-groups within each small area (Ghitza and Gelman, 2013). In MRP, researchers can interact

individual-level variables - as is the most common case - but can also include cross-level interactions between individual and area-level variables.

While the majority of MRP models do not use interactions, the use is still prevalent in MRP applications with just over one-third (36%) of unique models including interactions. For models which use interactions, standard practice is to use a single interaction, as can be seen in table 2.16, which reports the median, first and third quartiles of number of interactions. Using a single interaction is most likely due to a number of factors, including sample size and concerns about model complexity. Including one or more interactions will typically require a larger sample size. Researcher decisions on the number of interactions is therefore restricted by the sample size available. Similarly, including interactions makes the model more complex and harder to estimate. This means that researchers will often restrict the number of interactions to ensure they can reliably and efficiently estimate the model.

Table 2.16: Number of interactions included

Median	First quartile (25%)	Third quartile (75%)
1	1	2

Note: N = 33

Finally, in table 2.17 I report the interacted variables and frequency of use among interaction models. The interactions most used in MRP applications are either ethnicity and gender or age and education. These interactions are not particularly surprising, given that they include the four individual-level variables most used by MRP models. When including an interaction, it would seem that standard practice is to include one interaction and interact two individual-level variables already included in the model. Although cross-level interactions are possible to implement in MRP, only a few papers make use of them.

Table 2.17: Variables interacted

Interaction	N
Ethnicity & Gender	58%
Age & Education	42%
Age, Ethnicity & Gender	9%
Age & Ethnicity	6%
Age & Past Vote	6%
Education & Ethnicity	6%
Education & Past Vote	6%
Age & Gender	3%
Age & School	3%
Age, Education & Employment	3%
Education & Gender	3%
Ethnicity & Income	3%
Ethnicity & Past Vote	3%
Ethnicity & Region	3%
Ethnicity & State	3%
Foreign born & Gender	3%
Gender & Profession	3%
Income & Region	3%
Income & State	3%
Past Vote & Past Vote	3%
Past Vote & Time	3%

Note: N = 33

Time periods

One argument put forward for why MRP was an improvement over disaggregation was its ability to measure opinion change over time in small areas (Howe et al., 2015). The smaller sample sizes required by MRP, compared to disaggregation, meant researchers now had the opportunity to investigate temporal shifts in opinion or behaviour. Indeed, using MRP to estimate small area opinion change over time periods is popular among researchers who use MRP. Among all unique models analysed here, just over two-fifths (42%) estimated opinion or behaviour over a time period. Decisions about whether to estimate temporal changes in opinion or behaviour will solely be based on whether the research question requires it. However, greater data requirements - and model complexity depending how time periods are estimated - will most likely also play a role in the decision.

On average, the median number of time periods estimated for is 20, with first and third quartiles of 6 and 29, respectively. Time periods here could be years, days or any similar units, where respondents can be nested within. How a model estimates opinion across these given time periods also varies significantly. Broadly, there are three ways that MRP can be used to estimate opinion or behaviour over time periods. First, a separate static estimate, where there is a separate MRP model for each time period. Second, pooled-sample estimates where again there is a separate MRP model for each time period. However in this application, the sample is pooled so that survey data from the preceding and subsequent time periods are included to produce estimates for each time period. Third, time periods are directly included in the MRP model as a parameter. This method estimates one model with all data and a parameter for the time period. Table 2.18 reports the breakdown of MRP time period models by method. Including time period as a MRP parameter is the most commonly used approach, followed by the pooled-sample, while static estimates are the least common method. Which method a researcher uses will most likely be driven by methodological

decisions about the method they think will produce the best estimates.

Table 2.18: Time period estimation method

Method	N
Modelled	35%
Pooled	30%
Unique model	21%
No data	14%

Note: N = 43

Bayesian models

MRP is increasingly estimated as a Bayesian method. The benefit of Bayesian estimation for MRP is in part because it "propagates uncertainty across the modeling, and thus gives more realistic confidence intervals" (Lopez-Martin et al., 2019). Table 2.19 reports the percentage of models which are Bayesian. The majority of models are still not estimated as Bayesian models, but nonetheless, nearly a third of the models under analysis here are estimated as Bayesian models.[9]

Table 2.19: Bayesian model

Bayesian estimation	N
No	71%
Yes	29%

Note: N = 86

For Bayesian analysis, priors are an additional model component that researchers must specify. Priors are a means to incorporate knowledge that we have before

[9]These figures are indicative of whether a paper stated the model was Bayesian. Papers which did not explicitly state the model was Bayesian were assumed to not be.

estimation into the model and affect the estimation of parameters. Typically, we can categorise them as non-informative where the data is allowed to speak for itself. Weakly-informative priors where we rule out unlikely or impossible parameter values or informative priors where we directly encode strict numerical values for the expected distribution of the parameter. In table 2.20, I report the prior specification of the Bayesian MRP models. In MRP papers, many are not explicit or simply do not state the type of prior that they use in their Bayesian model. Indeed, half of the Bayesian models analysed here do not report prior specification. Among those who do explain their prior distribution, half are non-informative and half are weakly-informative priors. The debate between non-informative and weakly informative priors is solely methodological, and represents a wider debate in Bayesian analysis over which researchers should use.

Table 2.20: Prior specification

Prior	N
Non-informative	20%
Weakly-informative	20%
Other	8%
Not reported	52%

Note: N = 25

2.3 Conclusion

The motivation for this chapter was the belief that at present there is limited clear guidance on how best to use MRP. By documenting how the method has been used in social sciences to date, the chapter drew on the collective expertise of previous scholars who have used MRP. By doing this, the chapter has advanced our understanding of how

MRP is typically applied in social science. The chapter has identified standard practice when present, and identified areas where there is greater variation in the application of the method. When identifying variation, the chapter also provided discussion on the key methodological and substantive considerations that drive researcher decisions.

The analysis identified that standard practice for sample size was around 49 respondents per area. This varied depending on the number of small areas, with fewer small areas requiring more respondents per small area, whereas the larger number of small areas required fewer respondents per small area. Similarly, when we account for time periods the standard sample sizes decrease. Indicating, that while time period MRP models require larger sample sizes overall, they require fewer respondents per time period, when compared to models which estimate for a single time period.

For individual-level variables, the analysis highlighted that standard practice is to use four individual-level predictors, and between 8-16 categories altogether. The most used individual-level variables in MRP were age, gender, education, and ethnicity. These variables have been used by most models under analysis here. The prevalence is most likely due to the predictive power of these demographic characteristics and that they are often readily available at the right level and in the right format. For area-level variables, for all models the standard (median) was three, with most models using between two and five area-level variables. This also varied depending on the number of small areas, with models which have more small areas using more area-level variables. There is a broader range of area-level variables compared to individual-level variables. Nonetheless, area-level variables used typically fall into three categories, geographic, demographic and political.

Interactions are used by around a third (36%) of all papers analysed here. Typically, researchers use one interaction, and most often, researchers interact either age and education or gender and ethnicity. The analysis also identified that most models used three hierarchical levels, that is individual respondents nested in small areas, which in

turn were nested in a higher geographical unit. When investigating temporal changes in opinion in small areas, researchers differ in how they use MRP. To date, most common is to directly include the time period in the model as a parameter, followed by the survey pooling method. Most MRP models are not Bayesian, although this is a growing feature of MRP in applied work. When estimating MRP as a Bayesian method, researchers often do not state what their prior distributions are, and of those who do, researchers typically either use non-informative or weakly informative priors.

Across all the MRP components identified in this chapter, a standard practice does seem to emerge. However, for each component there is still variation. This means standard practice is not a single specification, but a range in which most of MRP models fall within. Researchers who wish to apply MRP should take note of these ranges, and use them to guide their MRP application decisions. The chapter has also provided some discussion for the main reasons for variation in application, highlighting the different methodological and substantive considerations of researchers when applying MRP. For the most part, these decisions are largely driven by methodological concerns and the limitations that a researcher faces, perhaps of most significance, are the sample size available and limitations of available data.

Chapter 3

MRP and variable selection

Multilevel regression and poststratification (MRP) is a method used to produce estimates of public opinion or behaviour in sub-national units of interest (small areas). Though the method has gained popularity and has been applied in social sciences to answer numerous substantive questions, the method is far from a panacea for all situations. Indeed, there are many instances where MRP is not a useful method (for a detailed discussion see Hanretty, 2019). As well as not being suitable in every situation, MRP cannot be uniformly applied to estimate small area opinion or behaviour.

One reason for this is the need for tailored variable selection. Researchers need to undertake preliminary research and analysis to ensure their variable selection choices are rooted in existing theory. This is a time-consuming task and, for many researchers, not a viable method to determine relevant variables. As an alternative, most applications of MRP directly replicate variables from previous applications (a method referred to here as path dependency). This method is often employed regardless of whether estimated opinion or behaviour is the same. Path dependency is presumably employed on the assumption that replicated variables will be predictive of all opinion or behaviour in social sciences. However, previous research by Buttice and Highton (2013) found significant variation in estimate accuracy when employing

a single set of variables to estimate a variety of opinion and behaviour.

Motivated by the fact that path dependency is unsatisfactory, while theory-based variable selection is not feasible in many applications, this chapter explores whether automated variable selection through 'Least Absolute Shrinkage and Selection Operator' (lasso) can offer researchers a viable alternative. Lasso regression is a regularisation method which "shrinks the coefficient estimates toward zero" (James et al., 2013).[1] For variable selection, variables with a non-zero coefficient value are selected for inclusion in the model. By combining with MRP, this would avoid simple path dependency and relieve the requirements of extensive preliminary analysis of theory.

The findings presented within this chapter suggest that when using lasso to select variables for MRP, the estimates are as, if not more, accurate than path dependency variable selection. The results also highlight that the benefits of lasso are notable for area-level variables, though less clear for individual-level variables and interactions. While it is difficult to compare the accuracy of lasso to theory based variable selection, a more conservative approach would favour theory-based variable selection. This chapter proceeds first with a brief introduction to MRP, variable selection and lasso regression. This will be followed with the theory of lasso regression, a description of the chapter method, the presentation of the results, and finally discussion of the findings.

3.1 Background

When interested in public opinion or behaviour, researchers typically use surveys to gauge a national-level picture. However, when interested in opinion or behaviour within sub-national small areas, surveys are no longer a practical method. In these instances, researchers need to make use of small area estimation techniques, such as

[1] In contrast to other regularisation methods such as ridge regression, lasso may shrink coefficients to exactly zero.

MRP. The method was first developed by Gelman and Little (1997), and subsequently applied to estimating opinion in social science by Park et al. (2004). To date, there have been numerous studies which have made significant contributions toward setting out how MRP can be used (See Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Buttice and Highton, 2013; Kastellec et al., 2016). As a result, it has now been applied by numerous researchers in social sciences across a range of mostly political topics.

MRP works in two stages. First a multilevel model is estimated where opinion or behaviour is modelled as a function of demographic-geographic variables (Tausanovitch and Warshaw, 2013: 334). For example, if the model included gender, age, and education, the first stage would produce an estimate for every female, aged 18-24, and university educated in each small area. Second, the estimates for each demographic-geographic person type are weighted to the population frequencies (Pacheco, 2011: 420). That is, the estimates are weighted according to the proportion each person type represents in the population.

MRP variables

MRP variables are one of the most important components of the model. Indeed, the model estimates opinion as a function of different individual and area-level predictors (Tausanovitch and Warshaw, 2013: 334). It is therefore imperative that these variables are well-suited to the opinion of interest, as poorly selected variables that are not predictive of the estimated opinion will result in inaccurate small area estimates. In recognising the importance of variable selection choices, scholars have emphasised the importance of researchers tailoring variable selection for their topic (Warshaw and Rodden, 2012; Buttice and Highton, 2013). As well as choosing the right variables, selecting the right interactions between variables may have significant impact on the estimates. Key work by Ghitza and Gelman (2013) has demonstrated that including 'deep-interactions' in MRP models can significantly improve estimate accuracy.

At level-1 of the multilevel model, opinion or behaviour is modelled as a function of individual-level variables. In most cases, these are demographic characteristics of respondents. In applications of MRP to date, there are four main individual-level variables used: age, gender, ethnicity, and education (see chapter 2). Whether combined or used alone, these variables are used by most MRP models, regardless of country or topic.

Researchers are faced with numerous restrictions when selecting individual-level variables. First, the variables must be included in the survey, second, the poststratification frame requires the small area joint distribution proportions for all individual-level variables included. For example, if we include gender (Male and Female) and age (18-24, 25-49, 50-64, 65+) as individual-level variables, we need the proportions of Males 18-24, Females 18-24, and so forth in each small area. Typically, this is done by directly inferring these proportions from a census or similar data frame. This restricts the application of MRP to countries where the required data is available. However, the development of alternative methods to construct a poststratification frame have enabled the wider application of MRP. For instance, Hanretty et al. (2016) constructed a frame by a raking procedure, while Leemann and Wasserfallen (2017) developed what they call 'synthetic' poststratification which allows researchers to construct a poststratification frame with relatively small N survey sizes.

At level-2 of the multilevel model, we estimate the effects of each small area. This is equivalent to estimating separate intercepts or slopes for each small area.[2] Typically, small area effects are modelled as a function of area-level variables (Kastellec et al., 2010: 772). Area-level variables are auxiliary data which capture some characteristic of each area. These could be political, demographic or geographic features of each small area. Using area-level variables is key to the MRP model better identifying small area variation (Gelman and Hill, 2007: 269). Indeed, previous research has demonstrated

[2]Most MRP applications only estimate a random intercept.

that the inclusion of area-level variables has the largest impact on estimate accuracy (Warshaw and Rodden, 2012; Buttice and Highton, 2013; Hanretty et al., 2016).

As with individual-level variables, availability of data in the correct format is the main limitation when researchers are deciding which area-level variables to use. However, there are fewer restrictions than for individual-level variables. Indeed, in many applications there are numerous free-to-access small area statistics, providing researchers with an abundance of potential variables to use.

The direct impact of both individual and area-level variables has led scholars to stress the importance of carefully selecting variables, and tailoring them towards the estimated opinion or behaviour (Buttice and Highton, 2013). Yet, chapter 2 highlighted a distinct lack of variation in variables used across MRP applications regardless of country or topic. This might partly by due to researchers' inability to ensure their variable selection will lead to accurate estimates. Validating variable selection is difficult because we rarely have *true* small area opinion figures with which to compare MRP estimates against.

One application where this is possible is electoral forecasting, where researchers can validate estimates against electoral results. This means that researchers can determine whether their selected variables are well suited. For most other topics this is not possible. Instead, it seems researchers base variable selection on previous applications of MRP which estimate a political opinion or behaviour. This is presumably employed on the basis that drivers of all political opinion or behaviour are identical or similar. However, researchers cannot be certain these variables will be well suited to their specific application of MRP. Indeed, Buttice and Highton (2013) noted significant variation in accuracy when using a single set of variables to predict a variety of opinion or behaviour.

In applications where no previous validated examples exist, the process of variable selection is more difficult. Researchers must carry out an extensive survey of the exist-

ing theory, or if no theory exists, in-depth preliminary analysis to identify predictive variables. The latter is a more onerous task than the former, but nonetheless, both represent a significant challenge to researchers who wish to apply MRP to new topics.

Although examples of researchers tailoring their models are rare, there are some. For instance, Leemann and Wasserfallen (2016) tested different combinations for each MRP model, and only presented the findings of the best model. However, this example is the exception rather than the rule. In most cases there seems to be limited consideration given to how the variables used will affect MRP estimates - beyond recognising that the variables have been demonstrated to perform well on political issues in previous studies.

Variable selection

Variable selection is a key component of any statistical modelling. Which variables we use to predict, explain, or describe is one of the most important decisions researchers face. The decisions have direct impact on the validity and reliability of statistical models and the research more broadly. Researchers thus need to carefully select variables, ensuring they include all relevant and predictive variables, exclude unnecessary and irrelevant variables, and ensure they avoid collinearity among variables. This is a time-consuming task, and with the growth in volume of available data, this task has become increasingly onerous. Furthermore, should a researcher wish to include interactions between variables, the tasks becomes more problematic as the number of features significantly increases.

Combing MRP with automated variable selection is not entirely new. Sparse MRP (of sMRP) is currently being developed by Goplerud et al. (2018). The method combines variable selection of lasso with a multilevel model (Goplerud et al., 2018: 1). Their work has demonstrated that by applying lassoPLUS priors on the coefficients,

they introduce sparsity directly into the multilevel model.[3] The estimates may then be poststratified as is done in the standard MRP application. However, unlike a normal multilevel model, where individual-level variables are treated as varying intercept terms, their model treats individual-level characteristics as fixed effects. A main argument for multilevel modelling was the benefits of partial pooling (Gelman and Hill, 2007), which was achieved by the varying intercept terms. By treating individual-level characteristics as fixed effects we lose this advantage. Proponents of sMRP would maintain that the sparsity introduced by the priors achieves the same goal as partial pooling. While it is beyond the scope of this chapter to comment in favour of either strategy, for those who wish to use varying intercepts, sMRP is not suitable.

Another recent development is work by Broniecki et al. (2021), whose method uses Ensemble Bayesian Model Averaging (EBMA), to produce estimates which are a weighted combination of five independent variable selection methods.[4] They demonstrated that EBMA performs better than theory-based selection, and better than each of the five independent selection methods, including lasso. However, their method only performs variable selection on area-level variables, and though there is good reason for this, if we are interested in variable selection of both individual and area-level variables this method is not suitable.[5]

Lasso variable selection

In regression analysis, researchers have developed automated variable selection to lessen the burden. These methods seek to identify the *best* model by determining the optimal combination of variables to use. That is, they seek to determine the best fitting model to predict an outcome variable. The growth in use and popularity of automated

[3]The lassoPLUS framework is itself a recent development by Ratkovic and Tingley (2017).

[4]The five variable selection methods they use are: best subset, lasso, PCA, gradient boosting, and support vector machine.

[5]The improvements in accuracy are greatest from the inclusion of area-level variables. They also note that the risks of overfitting are high for area-level variables, but much less problematic for individual-level variables.

variable selection may partly be attributed to the growth in computational power, which has meant researchers can sift through large numbers of potential candidate models with relative ease and speed.

Perhaps the simplest and most widely used are stepwise and best-subset methods. These methods seek to find the *best* model by selecting variables according to some predefined criterion. Stepwise works by starting with either a full or null model, then sequentially adding or subtracting variables according to whether the variable improves the model. The original, and still widely used criteria, is whether the added variable is statistically significant (See Heinze and Dunkler, 2017). The *best* model is when all variables included have a statistically significant p-value. Later alternatives have made use of AIC or BIC to determine variable inclusion (Hastie et al., 2009).

Best-subset works along a similar line of logic, the method works by searching through all possible variable combinations to find the *best* model according to a set criterion (Hastie et al., 2009). The criteria used to select variables include R-squared, AIC, BIC, log-likelihood, and Mallows C, depending on the model and researcher preference.

Both methods are widely used within social sciences and further afield. However, neither offer satisfactory solutions to variable selection. They are both criticised for their discrete process, tendencies of high variance, and inability to reduce prediction error from the null or full model (Hastie et al., 2009). Indeed, Breiman (1995) demonstrated that traditional model selection methods were not suitable as they did not lead to better prediction and did not provide stable solutions. The solutions provided by both methods typically fit well locally but perform poorly globally (Yuan and Lin, 2006).

In response to such problems, researchers have recommended least absolute shrinkage and selection operator (lasso). Lasso works be applying a penalty (known as lambda) to the coefficients, which shrinks them towards, and in some cases to zero.

For variable selection, we simply take-forward variables with non-zero coefficients. As a variable selection method, lasso has been demonstrated to improve overall prediction and model interpretability (Tibshirani, 1996). The method provides stable solutions and does not suffer from high variability that best-subset and stepwise do (Hastie et al., 2009). Although there is some dispute whether lasso is useful for parameter inference, the shrinkage induced by lasso is widely seen as an optimal solution for prediction purposes (Heinze et al., 2018).

The shrinkage (or regularisation) of lasso regression adjusts the bias-variance trade off, by reducing variance while increasing bias. Importantly, the reduction in variance can improve prediction accuracy as the risks of overfitting are restricted. This is especially important for out-of-sample prediction, where reducing variance is necessary to ensure the model can be applied to new data. The MRP case is an example of out-of-sample prediction, as the poststratification stage is cell-prediction using the multilevel model and the poststratification frame as new data. Therefore, although it is not desirable to oversimplify the model and increase bias, it is a necessary risk to reduce variance and ensure the predictions are stable and ultimately accurate.

There are alternative regularisation methods that have been proposed such as ridge or elastic-net, however, I believe that lasso is the best solution to use here. Ridge shrinks coefficients towards zero, but never to zero and therefore cannot be used for variable selection. A more recent innovation is elastic-net, which incorporates the benefits of both ridge and lasso regularisation (Zou and Hastie, 2005). Elastic-net was designed to resolve two issues associated with lasso: first, lasso has been shown to perform poorly when the number of predictor variables are greater than N; second, when there are two collinear variables, lasso selects one at random. However, I would argue that these problems are unlikely to arise in the MRP case, or at least, do not represent significant problems for the MRP application. It is highly unlikely, and I am not aware of, a MRP application where variables are greater than N. While

the collinear problem is of limited concern given that MRP is a prediction method solely concerned with estimate accuracy. Furthermore, the benefits achieved with extensions of the standard lasso - used in this chapter and explained below - outweigh the potential problems that can arise with lasso.

3.2 Theory

Lasso is an extension of standard ordinary least squares (OLS) regression. It works along the same procedure as OLS regression but introduces a penalty that shrinks (or regularises) the coefficients. In some cases, the coefficients are shrunk to zero and because of this, we also refer to lasso regression as introducing sparsity.

First, consider the standard linear case, where:

$$Y = \beta_\theta + \beta_X + \epsilon \tag{3.1}$$

Here β_θ refers to the intercept, β_X are coefficients for X, a matrix of $X_1, X_2, ..., X_p$ variables. To solve this problem, we introduce the ordinary least squares (OLS) method which aims to find values of β_θ and β_X which minimise the sum of squared error (or residual sum of squares, RSS). Put another way, we wish to find a solution that provides us with the 'line of best fit' given a set of data. The OLS solution is thus:

$$\min_{(\beta\theta,\beta X)} = \sum_{i=1}^{n}(y_i - \beta\theta - \sum_{j=1}^{p}\beta_j X_{ij})^2 \tag{3.2}$$

Where i represents the individual case, j references a unique variable, and p refers to the total number of variables. Here the OLS solution is an intercept and slope where the residuals (the difference between the actual value of y and the predicted y) are minimised.

The error in a linear regression case can be further divided into squared bias,

variance, and irreducible error (error which cannot be avoided). Variance is the amount by which the regression solution would change if the model were estimated with new data (James et al., 2013: 34). Whereas bias is the degree to which the regression model is too simple to capture the complexity of the relationship (James et al., 2013: 35). In the linear regression case, bias is often low and variance high. However, in some applications we wish to modify the bias-variance trade-off, meaning we wish to reduce variance at the cost of introducing bias. Lasso regression is one method that achieves this. This is done by introducing a penalty on the regression coefficients that shrinks them towards - and in some cases to - zero. The solution works in a similar way to OLS regression. The lasso solution is as follows:

$$\min_{(\beta\theta,\beta X)} = \sum_{i=1}^{n}(y_i - \beta\theta - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.3}$$

As can be seen above, the lasso equation is identical as the OLS (Eq3.2) except for the penalty: $\lambda\sum_{j=1}^{p}|\beta_j|$. The penalty equates to sum of absolute values of β coefficients * lambda (λ). In the lasso solution, λ induces shrinkage, with larger values of lambda inducing greater shrinkage of the coefficients. Where $\lambda = 0$, this is equivalent to the standard OLS solution and where $\lambda = \infty$, all regression coefficients are equal to zero.

For each value of lambda, we are provided with a different solution with different regression coefficient values relative to the degree of regularisation. To determine which value of lambda is most suited, we make use of k-folds cross validation. K-fold cross validation works by partitioning the data into K subsets. We take one group of K and set aside, then estimate the model using the data of the remaining groups. The model is tested on the group that was partitioned to determine model accuracy. This is repeated for each of the K groups. For instance, if K =10, we would use 9/10 groups to estimate (or train) the data, and the remaining group to test the model. In each case, the model itself is discarded, but the evaluation score is kept and used to

evaluate which model is *best*, according to whichever pre-defined evaluative criterion the researcher uses.

When applied to lasso regression, we use K-fold cross validation to determine which value of λ provides us with the lowest cross-validation error, this value is known as lambda-hat ($\hat{\lambda}$). For variable selection, we would simply take forward the variables in the $\hat{\lambda}$ model that have non-zero coefficients. However, $\hat{\lambda}$ solutions have been demonstrated to select overfitted models (Krstajic et al., 2014. 2014: 11). As an alternative, it is recommended that researchers should use $\hat{\lambda}1Std$, which represents the largest λ value within 1 cross-validation standard error from $\hat{\lambda}$ (see, Breiman, 1998; Hastie et al., 2009). While $\hat{\lambda}$ often produced an overfitted solution, $\hat{\lambda}1Std$ enforces greater regularisation leading to a more parsimonious solution.

The classic lasso was developed for the linear regression case, but this research will make use of numerous extensions beyond the linear application. First, the research will use the group-lasso to select variables. In situations where variables are a grouping of categories, it would not make sense to include some categories of the group and exclude others. For example, if we were to include regions as a group-effect, it would not make sense to only include some regions. To overcome this issue, Yuan and Lin (2006) introduced the group-lasso. The group-lasso works along similar lines as the standard lasso, however the process of selection takes place at the group level, where, if any category of a group is selected, all categories are selected and included in the model. This is especially important for selecting variables in the MRP case, as individual-level variables are included as a group of varying intercept effects.

Another extension to be utilised here is the lasso for generalized linear models. Lasso for generalized linear models works along the same procedure as the linear case but substitutes the sum of squares with the negative log-likelihood (Meier et al., 2008). It was first developed by Lokhorst (1999), and later developed to become the group lasso for logistic regression (Meier et al., 2008). The logistic group lasso can be written

as:

$$\min_{(\beta\theta,\beta X)} = l(\beta) + \lambda \sum_{g=1}^{g} s(df_g)||\beta_g||^2 \tag{3.4}$$

Here g references the g^{th} predictor from $g = 1...G$ groups of predictors, while df is the degrees of freedom for each g predictor. S is a rescaling function necessary to rescale the penalty. $l()$ is the log-likelihood function, i.e.:

$$l(\beta_\theta, \beta_X) = \prod_{g=1}^{g} p(x_i) \tag{3.5}$$

The final extension that this research will make use of is the 'Group-lasso interaction-NET' which was developed to enable researchers to find pairwise interactions via the group-lasso (Lim and Hastie, 2015). This is especially important for the purpose of this research, which aims to discover interactions for modelling with MRP. In a similar way to the group-lasso, when an interaction is selected, both unique variables are also selected for inclusion. This is important, as we know failing to include interaction variables as independent variables will most likely lead to biased parameters (Brambor et al., 2006: 68).

3.3 Data and methods

This research was designed to test how we can use lasso to select variables for MRP, and whether the variables selected lead to accurate MRP estimates. To achieve this, the research is primarily focused on answering the following:

1. How we can use lasso to select variables for MRP? And what degree of regularisation selects variables that produce the most accurate MRP estimates?

To test whether lasso is a viable variable selection method for MRP, we first need to establish how best to apply it to the MRP case. To use lasso regression to select variables, I make use of cross-validation (CV) to determine a value of λ that selects

variables with the lowest prediction error. As discussed above, $\hat{\lambda}$ is the model with the lowest CV error, but it has been argued that we should instead use the model associated with $\hat{\lambda}1Std$ (Breiman, 1998).

To test this, I first utilise CV lasso regression to select individual and area-level variables, and any associated pairwise interactions. Next, using the selected variables, I estimate the 2017 Conservative vote share within GB constituencies using MRP. Each model has a different set of variables and represents a different value of λ: from $\hat{\lambda}$ through to $\hat{\lambda}1Std$. Because there were some cases where different values of lambda produced identical solutions, and because I was interested in estimate accuracy of greater regularisation, I estimated a further three models. These models all had values of $\lambda > \hat{\lambda}1Std$. In total I planned to estimate 18 MRP models each with a unique set of variables.

2. Are lasso regression MRP estimates more accurate than path dependency or theory-based variable selection?

As identified previously, it seems evident that many applications of MRP select variables by replicating previous examples, referred to here as path dependency variable selection. The second stage seeks to analyse how lasso-MRP compares to 'off-the-shelf model specification', that is, models which replicate variable choices from previous applications of MRP. For this stage, I take forward $\hat{\lambda}1Std$ and compare with two models which replicate variables used in two previous UK studies.[6] This stage also makes a brief comparison between lasso MRP estimates and MRP estimates from Lauderdale et al. (2020), which I treat as a theory-based variable selection model.

Data

Multilevel regression and poststratification requires three data types, individual and area-level data used in the multilevel model, and data used to construct the

[6]Hanretty et al. (2016) and Hanretty et al. (2017).

poststratification frame. Individual and area-level data are subject to data availability issues, but broadly, researchers can access the necessary data with relative ease. Data required for the poststratification frame, on the other hand, is much harder to obtain. To construct, researchers need the joint-distribution proportions for each individual-level variable. The easiest way to construct a poststratification frame is to use the census or similar data. However, in most countries, there is insufficient available data to do this. Researchers have instead developed a raking procedure (Hanretty et al., 2016) or used synthetic poststratification (Leemann and Wasserfallen, 2017). To avoid this added level of complication, I make use of a poststratification frame published by Hanretty (2019).[7]

Individual-level data

Individual-level data used in this research is from the British Election Study (BES), a large survey of the British public's political opinion and behaviour. The study surveys around 30,000 respondents online at intervals each year, as well as prior and post a general election. This study makes use of the 2017 pre-election campaign survey data, wave 12 of the British Election Study. Once missing data was removed, the final sample was around 25,000.

Individual-level variables were restricted by variables included in the poststratification frame. The following variables were included: age (16-19, 20-24, 25-29, 30-44, 45-59, 60-64, 65-74, 75+), education (None, Level 1, level 2, level 3, level 4+, other), housing (own, rent), Social grade (AB, C1, C2, DE), and sex (Male, Female). I also included a variable which indexed the week of the campaign period that the survey data was collected. This was done to account for temporal changes in behaviour over the course of the campaign period.

[7]The poststratification frame (and number of categories) includes gender (2), age (8), education (6), social grade (4) and housing (2). It can be accessed here: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IPPPNU`

To produce estimates of vote share I estimated vote intention and turnout from the BES survey data. Conservative vote intention was derived from the 2017 vote intention question. This was re-coded into a dummy variable, Conservative = 1, all other, including would not vote, = 0. Turnout was also derived from the same 2017 vote intention question, intention to vote was coded = 1 and 'would not vote' = 0.

Area-level data

The area-level variables used within this research were from a data set assembled by the *Financial Times*, for their analysis of the 2017 election.[8] All variables were publicly available data and were predominantly demographic, health, economic and political variables. For variables with 5 or less missing values, I replaced missing values with figures from the preceding year, or closest possible year.[9] Variables with more than 5 cases of missing data were excluded, leaving a total of 39 area-level variables.

Using a pre-assembled set of area-level variables meant I was able to forgo an initial research step of finding a 'long-list' of potentially relevant variables. The data has most likely excluded variables which are relevant and predictive of 2017 voting behaviour. However, I believe this data set is sufficient for the purposes of testing whether lasso is able to select the most predictive and relevant variables from a set of variables.

Lasso regression

To select variables and any associated pairwise interactions, I make use of lasso regression. I first merged the individual and area-level variables into a single matrix of

[8] Accessed here: `https://github.com/ft-interactive/ge2017-dataset`

[9] For the 6 variables that had missing data, each only had a two missing cases. None of these variables were selected as stand-alone area-level variables, but were selected as interactions for some models. To check that the replacement of values was not affecting selection, I re-ran the lasso model on two subsets of the full sample (one with respondents from these areas and one without respondents from these areas). The results showed very small differences giving confidence that missingness and imputation was not affecting lasso selection.

predictor variables and estimated a hierarchical group-lasso logistic model, using Conservative vote choice as the binary dependent variable. I limited pairwise interactions to only allow interactions between variables of the same level; meaning individual-level variables could only be interacted with other individual-level variables and likewise for area-level variables.[10] The model used k-fold cross validation to select the value of lambda ($\hat{\lambda}$) with the least prediction error. I estimated the lasso models with the glinternet.cv function available through the R package glinternet (Lim and Hastie, 2013). The output included 50 values of lambda and the corresponding selected variables.[11]

Multilevel regression and Poststratification

For both stage 1 and 2, to estimate 2017 Conservative constituency vote share, I use multilevel regression and poststratification. The modelling strategy pursued here is motivated by previous studies which estimate vote choice using MRP. First, as with Selb and Munzert (2011), and Hanretty et al. (2016) I estimate vote choice as a binary outcome. Second, I follow the modelling procedure of Kiewiet de Jonge et al. (2018) by estimating turnout from the survey data and applying to vote estimates. Turnout was estimated by using the lasso and MRP strategy outlined in this chapter.[12] The one exception is that rather than post-stratify to constituency level, I keep the estimates for each row of the poststratification frame. I then apply the turnout estimates to vote intention estimates for each row of the poststratification frame. The final vote

[10]This was because I wanted to limit the complexity of the MRP models by excluding cross-level interactions.

[11]I provide a full of list of all variables and interactions selected for each lambda value in appendix B.2

[12]Because this chapter is testing whether lasso can select variables for MRP, it could be argued I should not have used this method to select variables for the turnout model. To ensure the results, and interpretation of results, were not a function of this turnout measure, I also produced results with two further turnout measures. Broadly, the interpretations do not change regardless of turnout measure. I show stage-1 results with the two alternative turnout measures in appendix B.1.

intention was thus:

$$PrFinalVote_i = Pr(T_i) * Pr(V_i) \tag{3.6}$$

Where i represents each cell-type, T refers to turnout and V to vote. Thus, to estimate the 'Final Vote' probability for any given cell-type I multiply turnout by vote share.

Except for variation in variables, all MRP models were identical. Each included campaign week as an additional random effect. All were estimated as Bayesian multilevel logistic regression models, with weakly informative student-t priors on the intercept and model coefficients (for a discussion on priors, see Gelman et al., 2008). I estimated the model with 2 Markov chains each with 1000 iterations (500 warm-up and 500 sampling). To decrease computational time, I included a QR decomposition in each model. All models were estimated using the stan_glmer function through the rstanarm package (Goodrich B, Gabry J, Ali I, Brilleman S 2020).

Following the estimation of the multilevel model, I drew 500 samples from the posterior, using the poststratification frame as a new data. I generated mean and 90% credible intervals (Low and High estimates) from the 500 posterior samples. Because the poststratification frame includes figures for the entire adult population, the final constituency estimates are Conservative vote share as a percentage of all adults, rather than of voters as is typically used in psephology.

3.4 Results

In most instances, conventional measures of goodness-of-fit would be suitable to assess the accuracy of a model. However, for MRP, the goal is to estimate opinion or behaviour of a sub-national unit. Assessing the goodness-of-fit of the multilevel model, or the prediction accuracy at the individual-level are therefore not particularly useful. It is far more suitable to assess the accuracy of estimates in each sub-national unit. In the application here, this is relatively straightforward as we can compare

the MRP estimates with actual 2017 Conservative constituency vote share. To assess the accuracy of MRP estimates I use three measures: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Correlation (Cor). MAE and RMSE are first calculated for each small area across the 500 posterior samples. I then calculate the average, and 90% credible intervals (Low, 5% and High 95%) across all constituencies to provide overall MAE and RMSE. Correlation is calculated between the mean constituency estimates and true constituency vote share.

In the first stage, I compare MRP estimates of various different λ values. Much of the emphasis here is on comparing results of various λ solutions to $\hat{\lambda}1Std$, which I treat as the baseline model. In the second stage, I take forward the baseline, and compare to two 'off-the-shelf' models and estimates from a theory model.

3.4.1 Comparing CV lasso lambda solutions

I first present the results from stage 1 of the research: how we can use lasso to select variables for MRP? In this stage I show MRP estimates for a range of λ solutions. To first give an indication of the relative size of each λ value, and the associated in-sample prediction error, in figure 3.1 I show the cross-validation (CV) error for all lambda values. In the figure there are two plots, on the left plot the x-axis shows raw lambda values and on the right plot the x-axis shows the logarithm of lambda. For both, CV error is shown on the y-axis and the two red lines show $\hat{\lambda}$ (left) and $\hat{\lambda}1Std$ (right). I estimate MRP models with selected variables from $\hat{\lambda}$ through to $\hat{\lambda}1Std$, and a further three models where $\lambda > \hat{\lambda}1Std$.

The first plot is useful to see the relative size of each lambda value, and thus the relative degree of regularisation that each λ enforces. Although there is a clear difference between $\hat{\lambda}$ and $\hat{\lambda}1Std$ values, they are relatively similar compared to the range in λ value sizes. The right-hand plot makes it easier to see the CV error of all the λ solutions estimated in this section. As we would expect, $\hat{\lambda}$ has the lowest CV

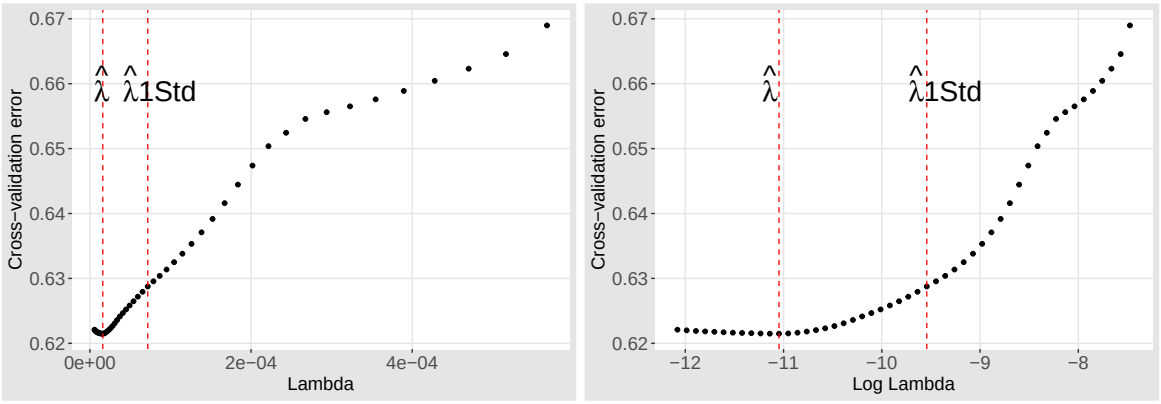

Figure 3.1: Lambda CV error
Notes: showing raw lambda on left and log lambda on right.

error, while $\hat{\lambda}1Std$ has 0.01 greater CV error - hence why selecting $\hat{\lambda}1Std$ is called the one-standard-error rule (Hastie et al., 2009: 244).

Next, I present the accuracy of the MRP models with variables selected with the different lambda values. Here I report the estimates of 16 separate MRP models, each representing a different value of lambda. I originally intended to estimate 18 models in total, however, two MRP models (12 and 13) are not included in the results because the multilevel models failed to estimate. These models are the two with the smallest λ values, $\hat{\lambda}$ and the value immediately above it. For both models, two columns (coefficients) were dropped because of rank deficiency and thus the resulting estimates cannot be analysed. The rank deficiency is most likely a result of multicollinearity among the matrix of predictor variables. Indeed, a surface-level examination of selected variables showed there are cases where one area-level variable was a direct linear combination of two other variables.[13] The degree of regularisation imposed by these two lambda values was clearly too small and the resulting associated models were overfit to the data.

Figure 3.2 reports the MAE, RMSE and correlation for each of the MRP models. I report the accuracy figures for each model, labeled by the λ index number. The larger

[13]For instance, both models include social grade variables C2, DE, and C2DE. The latter is the combined percentage of the first two, and thus, a perfect linear combination.

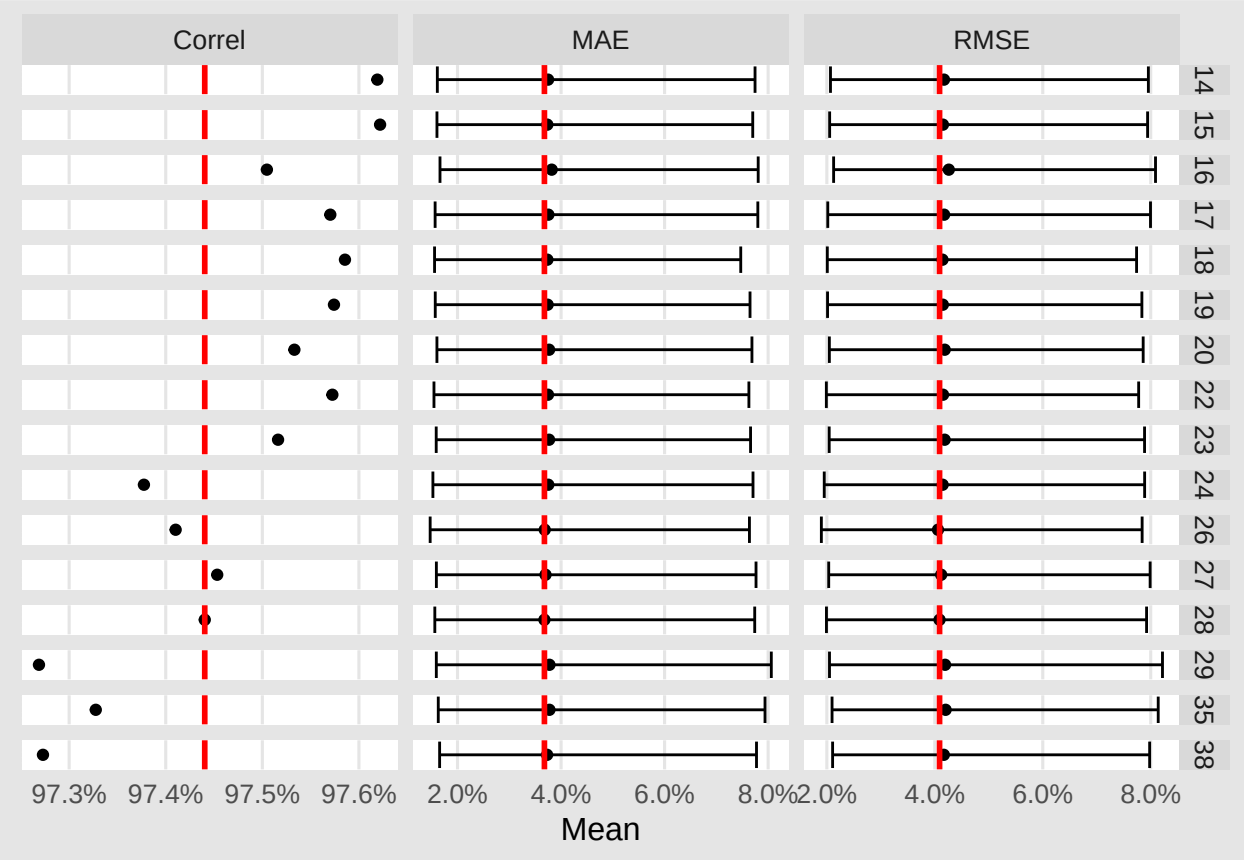

Figure 3.2: Estimate accuracy for lambda solutions
Notes: showing correlation, MAE and RMSE. Points show mean accuracy value. For MAE and RMSE lines indicate 90% credible intervals. Red lines show accuracy value for the baseline model.

index number indicates a larger λ value and therefore a higher degree of regularisation. On the left-hand side of figure 3.2 I show correlation between model estimates and actual vote share, MAE is displayed in the middle and RMSE displayed on the right. In each chart, the accuracy value is represented by the circle, with MAE and RMSE also showing lines either side of the circle indicating the 90% credible intervals. Finally, the red dashed line shows the value for the $\hat{\lambda}1Std$ (the baseline| Model 28) and should be used as a point of comparison for all other models.

Looking at correlation in figure 3.2 - the left-hand graph - we can see there is little difference between each model's correlation to actual Conservative constituency vote share. Indeed, the difference between the lowest and highest correlation is less

than 0.5%, ranging from just below 97.3% to just above 97.6%. Although differences are small, most points are to the right of the red line, indicating that most models achieve higher correlation than the baseline. Overall, a general pattern seems to be present where models with a smaller λ value (i.e. less regularisation) achieve greater correlation.

Turning to the middle chart which reports MAE of model estimates. For all models, MAE is just under 4%, with the credible intervals ranging from around 2-8%. Overall, there seems to be small differences between each model's MAE. Although, most MAE points are just to the right of the red line, which indicates that most model's MAE is marginally higher than the baseline. Finally, looking at RMSE on the right-hand side of figure 3.2, we see a similar picture to MAE. RMSE for all models is around 4%, with the upper and lower estimates ranging from around 2-8%. Again, most RMSE point estimates seem positioned just to the right of the red line indicating most model RMSE is marginally worse than the baseline.

To further explore how each model compares to the baseline model, in table 3.1, I report increase or decrease from the baseline for correlation, average MAE and average RMSE. To show +/- from the baseline, I report figures which are a percentage of the baseline value. For example, model 38 MAE is 1%, which indicates an increase by 1% of the baseline MAE value.

Looking first at correlation, each model has a 0% figure. This indicates that in each case, there is no difference between any model and the baseline model correlation. This is evident in figure 3.2, which shows less than 0.5% difference between all model correlations. In figure 3.2 to account for the 90% credible intervals, the x-axis for MAE and RMSE extends from 2-8%. This made it difficult to determine whether any model achieved better MAE or RMSE than the baseline. However, using table 3.1 we are better able to identify differences no matter how small. For MAE, it is evident no model performed better than the baseline (Model 28), although model 26

Table 3.1: Comparing lambda solution's accuracy to baseline

Model	Correl	MAE	RMSE
38	0%	1%	2%
35	0%	3%	3%
29	0%	3%	2%
27	0%	1%	1%
26	0%	0%	-1%
24	0%	2%	1%
23	0%	2%	2%
22	0%	2%	2%
20	0%	2%	2%
19	0%	2%	1%
18	0%	2%	1%
17	0%	2%	2%
16	0%	4%	4%
15	0%	2%	2%
14	0%	2%	2%

Note:
Values are % increase/decrease of the baseline value

is equally as accurate. Whereas for model 16, MAE increases by 4%. Finally, looking at RMSE, model 26 is an improvement on the baseline, with a 1% improvement for RMSE. However, all other models perform worse than the baseline, with model 16 again faring the worst with RMSE 4% higher. Although these differences are small, when assessed alongside the failure of $\hat{\lambda}$ to estimate altogether, the results seem to support the case for $\hat{\lambda}1std$.

3.4.2 Comparing lasso with path dependency variable selection

The first stage explored how best we should use CV lasso to select variables for MRP. I next compare the approach with the path dependency strategy, which replicates variable selection from past studies (referred to here as 'off-the-shelf' model specifi-

cation). Below, I compare the model variables and the estimate accuracy of three models: lasso-MRP (Baseline | $\lambda 1 \hat{Std}$)[14], off-the-shelf-A which replicates selection from Hanretty et al. (2016), and off-the-shelf-B which replicates variables from Hanretty et al. (2017).[15]

First, in table 3.2 I present the variables used in each of the three MRP models. As can be seen from table 3.2, there is little variation in individual-level variables used by all three models. This is primarily because each model used the same poststratification frame, and therefore each was restricted to the same set of individual-level variables. This has clearly made it difficult to determine the benefits of lasso when selecting individual-level variables. The omission of gender in the lasso model is the only distinguishing feature for variable selection at this level. Though small, this difference may highlight the ability of lasso to select relevant and predictive variables and exclude irrelevant individual-level variables.

The difference between models is much clearer for area-level variables. First it is notable that the lasso-MRP model selected three area-level variables, while Off-the-shelf-A and B used 11 and 15, respectively. The lasso model used two political variables and one demographic, off-the-shelf-A used mostly demographic and two geographic variables, while off-the-shelf-B used a mixture of political and demographic variables.

The differences in accuracy are most likely due to the differences in these variables. Of note, is the selection of political variables by the lasso model and off-the-shelf-B. There is extensive evidence in the literature which demonstrates political variables (and particularly past vote choice) are highly predictive of vote choice. It is therefore unsurprising that lasso-MRP and off-the-shelf-B achieved higher accuracy than off-the-

[14]In line with the recommendations of Breiman (1998), I have taken forward $\hat{\lambda} 1 Std$ as the lasso model to compare.

[15]It would have been useful to replicate models from different authors, but there are few studies which apply MRP to the UK. I am only aware of one with entirely different authors, but they do not report model variables.

Table 3.2: Model variables

Variable	Lasso-MRP	Off-the-shelf-A	Off-the-shelf-B
Individual-level	Age	Gender	Gender
	Education	Age	Age
	Housing	Education	Education
	Social grade	Housing	Housing
		Social Grade	Social Grade
Area-level	Leave vote share	Region	Region
	Conservative 2015 vote	Density	2015 Conservative vote share
	Aged 18-24	Christian population	2015 Labour vote share
		Other religion	Leave vote share
		Non-white	2015 Green vote share
		Owns house	Plaid Cymru 2015 vote
		Female population	Aged 18-24
		Average education	Aged 65+
		Married	Own house
		Private sector	Self-employed
		Median social grade	Unemployed
			Economically inactive
			White population
			Density
			Level 4 Qualifications
			Health bad

shelf-A (full accuracy results are reported below). The inclusion of political variables for these models was most likely a significant contributing factor to the improved accuracy. This is significant because lasso has enabled us to to identify political variables as the most predictive without the need for in-depth research.

The lasso method selected a mixture of political and demographic variables, but a more restricted list than either of the off-the-shelf models. However, lasso-MRP achieved similar, if not better, accuracy than the two off-the-shelf models with a much smaller set of variables. This could be seen as evidence that the method is highly efficient at selecting the *best* predictive and relevant variables, while also able to exclude variables that do not contribute to improved accuracy.

Turning to the comparison in accuracy, figure 3.3 presents three scatter plots,

one for each of the three models being compared here (lasso, off-the-shelf-A, and off-the-shelf-B). The y-axis shows the MRP estimate and the x-axis shows the actual 2017 Conservative vote share. In each individual graph, the MRP mean estimates are shown by the points, with vertical blue lines representing the 90% credible intervals. The red line shows the relationship between the estimates and true vote share (with the line angle reported on each plot), while the dashed grey line visualises a perfect linear relationship. For all the model estimates, we can see the red line is above the dashed line when Conservative vote share is small, and below the dashed line when Conservative vote share is high. This indicates that the models over-estimate Conservative vote share when true vote share is small, and under-estimate vote share when true vote share is high. However, lasso and off-the-shelf-B have a line angle of 35° whereas for off-the-Shelf-A has a 32°. This means that lasso and off-the-shelf-B estimates are closer to a linear relationship (i.e. 45°) with true vote share. Furthermore, it seems evident from the scatter plot that the 90% credible intervals for lasso and off-the-shelf-B are shorter, meaning we have greater confidence in the estimates and generally indicates better model estimation.

Comparison of the model's accuracy is further explored in figure 3.4, which shows MAE, RMSE and correlation for each model. For both MAE and RMSE, the circle is the point estimate while the lines indicate the 90% credible intervals for the accuracy measure. The plot shows that both lasso and off-the-Shelf-B have higher correlation (both over 97%), while off-the-Shelf-A has a lower correlation of just under 93%. Turning to MAE, it is evident that lasso and off-the-shelf-B perform better than off-the-Shelf-A, with nearly a 1% reduction in error. RMSE shows a similar pattern, lasso and off-the-shelf-B have RMSE of around 4%, whereas off-the-shelf-A has RMSE of over 5%.

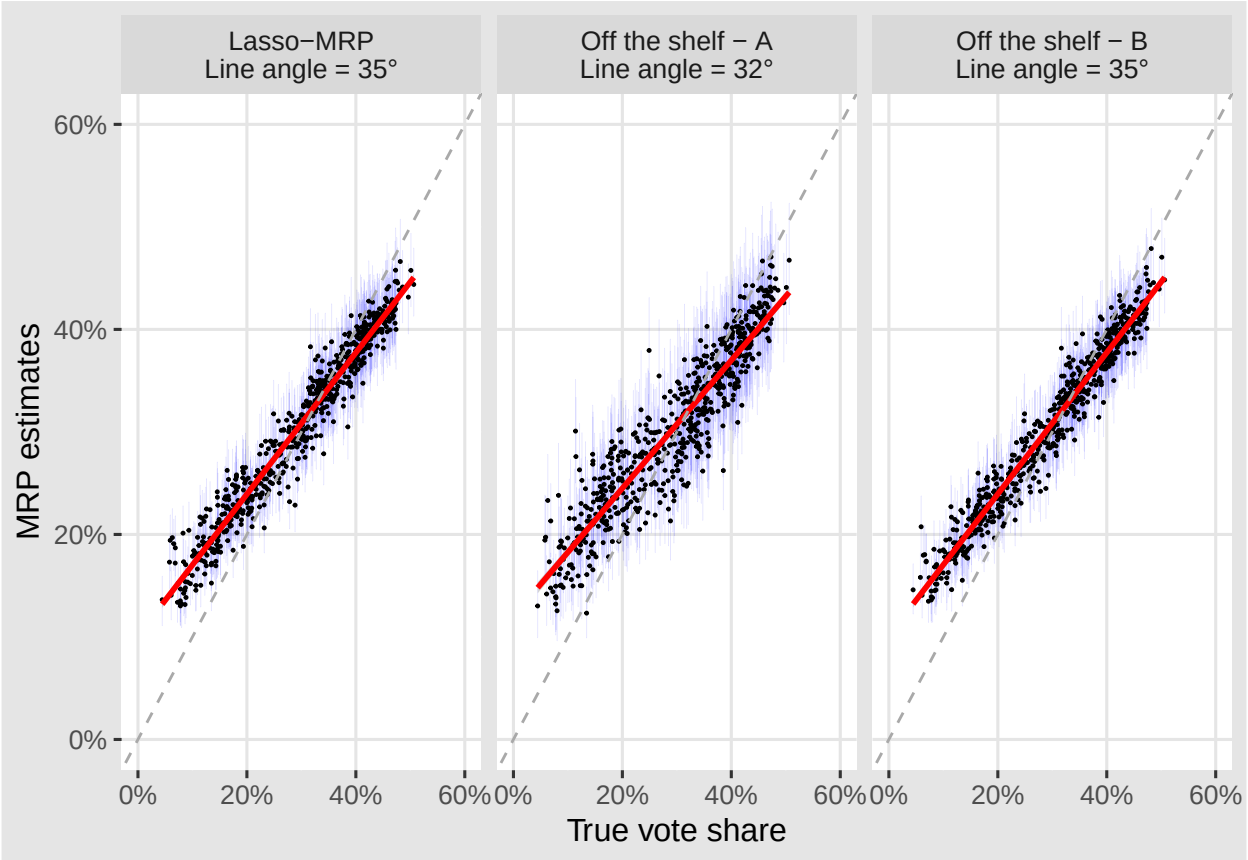

Figure 3.3: MRP estimates: lasso versus Off-the-shelf
Notes: showing estimates versus true 2017 Conservative vote share.

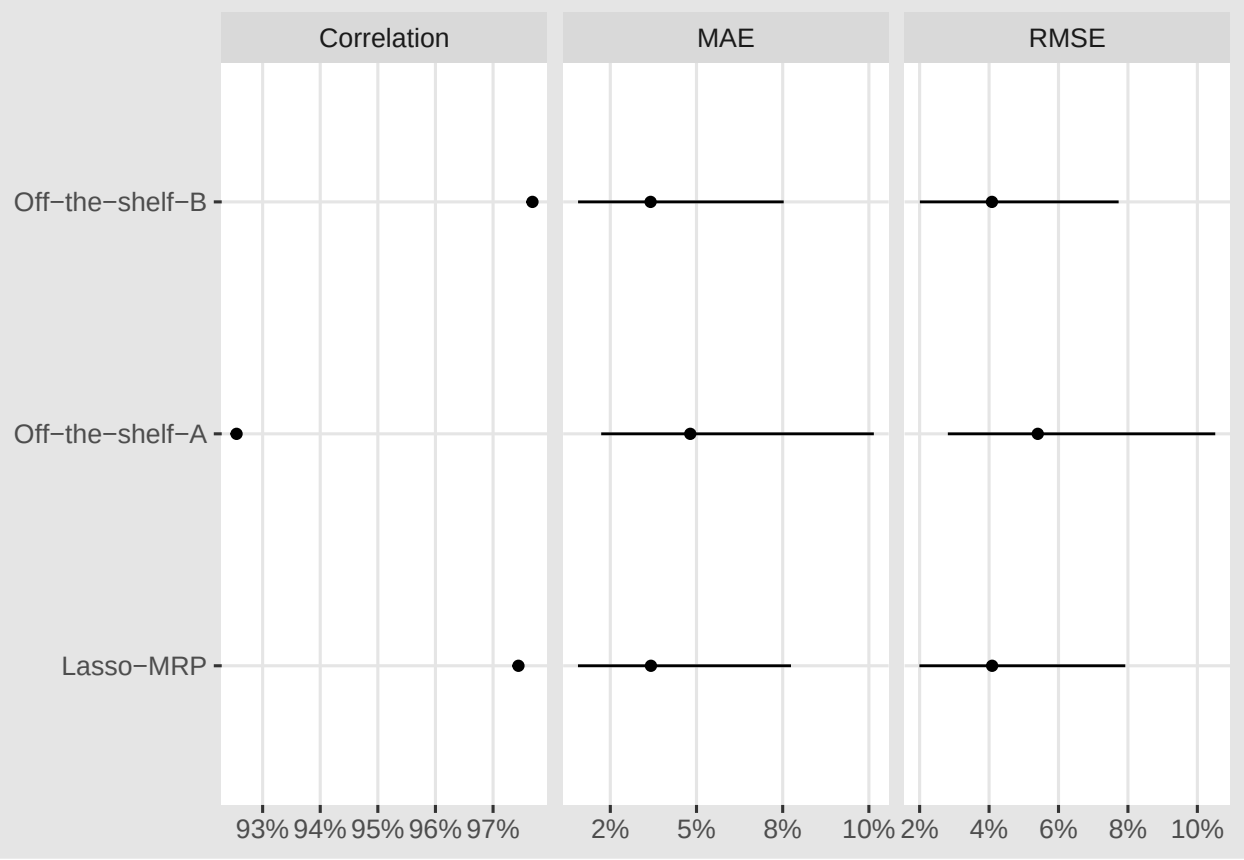

Figure 3.4: MRP accuracy: lasso versus Off-the-shelf
Notes: Correlation (left), MAE (middle), RMSE (right). MAE and RMSE points show average accuracy with lines indicating 90% credible intervals.

3.4.3 Comparing lasso with theory-based variable selection

To make some comparison between lasso and a theory-based model, I next compare the lasso estimates to those of Lauderdale et al. (2020). Their study presented results of their MRP 2017 UK general election estimates, including Conservative party vote share. While I am not aware of their exact variable selection process, the variables are not identical to those from previous studies and thus we can assume that theory and prior analysis informed their variable choices.

The estimates were originally published by YouGov before the election and correctly predicted a hung parliament.[16] In order to directly compare the two, I first convert YouGov estimates to the same format as lasso-MRP estimates (a percentage of the constituency population as opposed to percentage of voters).[17] I then calculate accuracy for these figures. To make the lasso estimates directly comparable, I calculate accuracy on mean constituency estimates, as opposed to the previous section where I calculated accuracy across iterations of the poststratification stage.

The YouGov Conservative party estimates achieved MAE of 2.5%, RMSE of 3.1% and correlation 97.8%. The baseline lasso model achieved 3.3% MAE, RMSE of 4.1% and has a correlation of 97.4%. The Lauderdale et al. (2020) estimates are a significant improvement on the lasso-MRP estimates, and would support the idea that theory driven variable selection leads to improved estimate accuracy. However, it should be noted that the comparison here is not entirely fair, as the Lauderdale et al. (2020) model uses a much larger sample, individual-level past vote and cross-level interactions. It is therefore difficult to disentangle whether the improved accuracy is because of improved variable selection choices or other beneficial characteristics.

[16] Can be accessed here: `https://yougov.co.uk/topics/politics/articles-reports/2017/05/31/how-yougov-model-2017-general-election-works`

[17] Although Lauderdale et al. (2020) provide accuracy figures in their paper, it was necessary to convert estimates to percentage of population first, in order to compare with lasso-MRP estimate accuracy.

3.5 Discussion

This research was designed to test whether researchers could make use of lasso regression to select variables for MRP. When estimating small area opinion or behaviour with MRP, individual and area-level variables have a direct impact on the accuracy of the estimates. This is because the opinion or behaviour is modelled as a function of these variables (Kastellec et al., 2010: 771). In recognition, researchers have argued that users of MRP need to carefully select variables and ensure their relevance to estimated opinion or behaviour (See Warshaw and Rodden, 2012; Buttice and Highton, 2013). Yet, at present, it seems most variable selection for MRP is based upon a path dependency approach, as opposed to driven by the existing theory. This is most likely because theory driven variable selection is a time-consuming strategy. To provide researchers with an alternative, this research tested whether using lasso regression as a preliminary stage to MRP could automate variable selection and produce accurate estimates.

As a precursor to this, the research first sought to establish how lasso could be used for applications with MRP. I used cross-validation and tested a variety of λ values to determine best practice for MRP. Breiman (1998) and Hastie et al. (2009) recommend using $\hat{\lambda}1Std$ rather than $\hat{\lambda}$. They argue the higher degree of regularisation leads to a more parsimonious model and better prediction accuracy. The results presented in this chapter show that for applications with MRP, this recommendation should be followed. Indeed, the MRP model associated with $\hat{\lambda}$ failed to estimate entirely, whereas, $\hat{\lambda}1Std$ consistently demonstrated among the best prediction accuracy. Although, among all λ solutions there were only marginal differences in accuracy.

Importantly, this research has demonstrated that lasso and MRP work to produce estimates that are equally, if not more, accurate than simple path dependency. Comparing lasso MRP estimates with two different off-the-Shelf models, lasso-MRP produced estimates equally as accurate as one model and better than the other. On the

one hand, it could be argued that this research has not demonstrated any particular benefit to using lasso for the MRP case. Rather, we have demonstrated that path dependency can lead to accurate estimates, despite previously suggesting the opposite would be true. However, by demonstrating that lasso is equally - if not more - accurate than path dependency, I would argue lasso is preferable. Lasso, at worst is equal to path dependency, and at best, improves accuracy.

Beyond accuracy, lasso may be viewed as superior to path dependency models as it produced competitively accurate estimates with a far simpler model. Although there has been a drive for complex solutions in science, econometrics, and social sciences, most scientists still contend that simpler statistical models are preferable (See Green and Armstrong, 2015; Zellner et al., 2002). As was highlighted in table 3.2, the lasso solution was far simpler than either off-the-shelf-A or B. This may lead us to conclude that lasso was more adept at determining predictive variables than simple path dependency

Directly comparing lasso to theory-based variable selection is difficult and this chapter has not been able to make an adequate comparison. Nonetheless, the limited comparison that this chapter made suggests that theory-based variable selection can lead to higher accuracy than lasso MRP models.

Overall, this research has demonstrated that lasso can be a useful tool to select variables for use with MRP. Equally, the research has shown that theory and path dependency can produce accurate estimates. The comparison between these methods throughout this chapter has created a narrative that these strategies are in opposition to each other. However, in most applied settings, the methods work best together. It therefore seems evident that best practice for researchers is to use a combination of these methods, with all incorporated into the model building process. Although this is not a fully automated process, this represents an improvement over the current standard practice for MRP model building, which largely seems to be based upon

simple path dependency.

The inclusion of lasso in the model-building process may prove particularly useful in applications where there is limited theory and no previous MRP application examples. In these instances, preliminary analysis with lasso can help guide variable selection. The lasso solution is not a panacea, and researchers will still need to find a 'long-list' of potentially predictive variables which lasso can select from. But, as has been demonstrated here, we can have confidence that lasso will select variables that are predictive of the opinion or behaviour of interest.

The results presented in this chapter have shown the benefits of lasso are particularly notable for the selection of area-level variables. This is an important finding as previous work has demonstrated that these variables have the largest impact on estimate accuracy (Hanretty et al., 2016). When compared to path dependency, lasso selected a more limited set of variables, but was still able to achieve comparable accuracy. This signifies that, unlike path dependency, lasso was able to efficiently select relevant and predictive variables, while excluding additional variables that did not contribute to improved MRP estimate accuracy.

When a researcher wishes to include interactions in a model, the potential number of features to choose from increases dramatically. In such instances, lasso may be particularly useful as it can be used to detect possible predictive interactions. However, from the results here, it is unclear whether lasso is efficient at detecting and selecting interactions. In stage one, some MRP models included interactions, but these models performed no better than models without interactions. Perhaps not permitting cross-level interactions was too restrictive, as research where these are included has demonstrated high levels of accuracy (see Lauderdale et al., 2020). Conversely, it may be the case that no interactions were particularly useful for predicting 2017 Conservative vote share, and lasso was able to identify this. Unfortunately there is no 'ground truth' to determine which, if any, interactions are predictive of voting behaviour.

This means we are ill-equipped to be able to fully assess the lasso performance on selecting interactions. Future research could focus on a simulation study to be able to properly investigate this.

Similarly, the results of using lasso to select individual-level variables are not clear. This is because this research was limited to a small set of individual-level variables due to all strategies using an identical poststratification frame. The comparison between path dependency and lasso variables highlighted that all models used near identical individual-level variables, and therefore we are not able to properly establish the benefit of lasso to select these variables. Future research may find it interesting to investigate whether using lasso with a larger set of individual-level variables -including variables beyond demography- can improve small area estimate accuracy.

A consideration that should be taken into account, but has been overlooked throughout this study, is at what point in the model building process or research overall do we use this method? In this study, I have used variable selection at a single and fixed point, using the method to select variables for the final vote share prediction model. However, in scenarios where we wish to estimate opinion or behaviour across time periods, further consideration needs to be given to where this method fits into the process. If we are forecasting vote share in small areas over the course of an election, should variable selection be undertaken at the start of a campaign period and not changed throughout the election? Or should the process be continuous, where we constantly update variable selection based on new data collection? While substantial shifts in voting patterns over the course of an election are rare, failing to account for them in the variable selection process could be problematic. Similarly, studies which apply MRP to estimating opinion over the course of a significant time-period may need to consider how we select variables that account for potential shifts in the relationship between the opinion or behaviour of interest and predictor variables.

3.6 Conclusion

This chapter set out to explore how best to use lasso variable selection for MRP, and whether the method produces accurate estimates. The chapter has demonstrated that lasso can be used efficiently to select variables for MRP and these models can, in turn, produce accurate estimates. The research has shown how we can use cross-validation lasso to select variables for applications with MRP, and shown what degree of regularisation is appropriate for MRP. In line with Breiman (1998) and Hastie et al. (2009), the results support the argument that $\hat{\lambda}1Std$ is the preferable λ value, leading to stable and accurate estimates.

Importantly, the results show that the estimates are equally, if not more, accurate than a simple path dependency approach. This chapter has not been able to adequately compare theory-based variable selection with lasso, but a conservative interpretation of the comparison made here would suggest that we should favour theory-based selection where possible. Nonetheless, this chapter has contributed to our wider understanding of MRP variable selection, showing that lasso may be a useful tool in the model-building process.

However, the results presented here have not made a significant contribution towards demonstrating that lasso is beneficial to selecting individual-level variables. This is partly because this research made use of a limited set of individual-level variables which have differed little between models. Similarly, we have not been able to make any inferences about the benefit of lasso to select interaction pairs. Few lasso models selected any interaction pairs, and as the results show, the inclusion of interactions had no impact on estimate accuracy.

For selecting area-level variables, this research has demonstrated that lasso is an efficient method, and associated MRP models produce accurate results. This is important, as work has demonstrated that inclusion of area-level variables improve estimate accuracy to the greatest extent (Warshaw and Rodden, 2012; Hanretty et

al., 2016). Furthermore, as there are risks of overfitting for area-level variables, it is particularly important that selection is considered (Broniecki et al., 2021). Although if the benefits of lasso are only notable for area-level variables, researchers may be better served by applying the autoMRP approach developed by Broniecki et al. (2021). Their research demonstrated the effectiveness of this method to select area-level variables, and outperformed lasso variable selection.

Chapter 4

Improved MRP sample distribution

Multilevel regression and poststratification (MRP) has become increasingly popular in academia and further afield. This is largely because the method enables researchers to produce reliable and accurate estimates of public opinion or behaviour in sub-national small areas. One of the main benefits of MRP is that the method requires relatively small survey sample sizes. Indeed, much early work was dedicated to investigating minimum and optimal sample sizes necessary for MRP (See Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Buttice and Highton, 2013)

However, discussion on MRP samples rarely goes beyond sample size. This is probably because most applications of the method make use of publicly available surveys, where researchers have no control over the survey or the sampling procedure. Therefore the discussion focuses solely on whether a survey can - or cannot - be utilised for MRP with a given sample size. But, in applications where researchers have input into the survey design and sampling strategy, they should also consider other characteristics which directly affect both estimates and estimate accuracy. For example, the distribution of respondents among small areas directly affects sample estimates but is almost never discussed in the MRP literature.

Again, the lack of discussion on the distribution of the sample is most likely

because typically researchers have no input into the sampling strategy. However, when researchers have control over the sample design, in certain applications they might want to consider whether oversampling respondents from certain small areas could improve their MRP estimates. In practice, the strategy would involve adjusting the sample distribution so that certain small areas receive a greater proportion of the sample. For example, in electoral forecasting we may wish to pursue a strategy where to improve prediction accuracy in certain small areas we allocate them a larger proportion of the sample. This strategy may be advantageous if predicting an electoral outcome rests on correctly predicting the outcome in certain small areas.

Accordingly, this chapter seeks to explore whether for electoral forecasting an uneven sample distribution can improve prediction accuracy in certain small areas. I address this question with a simulation study and two real-world applications. Overall, the results presented in this chapter are supportive of the strategy. The findings show that the method can improve estimate accuracy in small areas deemed important, and this in turn can improve our ability to forecast elections.

4.1 Background

MRP was developed to provide reliable estimates of opinion or behaviour for populations in small sub-national areas. Its development was necessary because traditional methods of inquiry such as surveys are not suitable when interested in these populations (See CH1: Introduction to MRP). Although alternative small area estimation methods have been utilised in previous research, these have since been demonstrated to produce unreliable and inaccurate estimates. MRP, on the other hand, has been shown to produce reliable estimates with much smaller sample sizes than would be necessary to directly infer opinion or behaviour. For instance, Lax and Phillips (2009a) argued that their study demonstrated MRP was able to produce accurate estimates in

50 US states with a total sample of around 1,400.[1] Kastellec et al. (2016) have echoed such findings, again arguing that a sample of around 1,400 was sufficient to estimate in 50 US states. Warshaw and Rodden (2012) subsequently found that a sample of 2,500 and 5,000 were suitable for US congressional districts (436 small areas) and for Senate districts (1,942 small areas), respectively. Outside of the United States, Hanretty et al. (2016) suggested a sample of between 8,000 and 12,000 respondents for 632 small areas was sufficient.

Why does MRP perform better with smaller samples?

MRP benefits from the hierarchical structure of multilevel modelling, which, by partially pooling respondents across small areas can produce reliable estimates for each small area even with small sample sizes (Selb and Munzert, 2011; Leemann and Wasserfallen, 2017: 1005). In practice this means the final estimates for each small area are derived from information across the entire sample, as well as information specific to each small area. In standard regression, analysis is typically one of either full or no pooling. The former means the sample is pooled together, and the model is blind to any variance by groupings such as small areas. The latter - no pooling - means we model opinion or behaviour separately for each small area. This is also an unsatisfactory solution as it means our analysis fails to account for the variation across the population in general and will most likely lead to poor estimates. Furthermore, in most cases sample sizes are too small to enable reliable model estimation for each small area separately.

Partial pooling is a midpoint between the two, where small area estimates "borrow strength" from the whole sample (Gelman and Hill, 2007). Area parameters are drawn for a common distribution, which means the estimation of area-level parameters are

[1] Buttice and Highton (2013) noted significant variation in accuracy with these sample sizes. Although, they still argued that MRP represented an improvement on past alternative methods such as disaggregation.

derived from information specific to each small area, as well as information from all small areas (Hanretty et al., 2016; Hanretty, 2019). In practice, area-level parameters are shrunk towards the overall sample mean after controlling for individual-level variables (Kastellec et al., 2010). The degree of shrinkage can be considered as a weighting scheme, where areas with smaller sample sizes have estimates shrunk towards the overall mean to a greater extent (Gelman and Hill, 2007: 254). While areas with a larger sample size are shrunk to a lesser extent and estimates can vary away from the overall mean to a greater extent.[2] The partial pooling is also relative to the degree of variance, with more pooling when variance between small area opinion or behaviour is small (Lax and Phillips, 2009b: 111).

Partial pooling principally manifests in two related but independent ways: the partial pooling of individual-level random effects and the partial pooling of area-level variance effects. Individual-level characteristics (typically demographic) are included in the multilevel model as varying intercept effects. This means that the entire sample informs the individual-level variable parameters regardless of which small area a respondent resides in (Selb and Munzert, 2011: 457). The second and perhaps more important way that partial pooling benefits MRP is through the estimation of area-level parameters, which are again estimated as varying intercept effects. Because the sample per area is typically small (or at least too small to directly infer parameters) the partial pooling is essential to reliably estimate parameters for each small area. Typically, area-level parameters are also modelled as a function of area-level variables. This improves the partial pooling as small area estimates are shrunk towards the mean of areas similar to each specific small area (Gelman and Hill, 2007: 269). Indeed, partial pooling alone has been shown to have limited impact on estimate accuracy (Hanretty et al., 2016). However, With the introduction of area-level variables there is consistent and significant improvement in estimate accuracy (See Warshaw and

[2]Assuming the individual small area estimate is different to the overall mean.

Rodden, 2012; Selb and Munzert, 2011; Hanretty et al., 2016).

MRP sample distribution

Discussion on MRP sample tends to focus on the total sample size, with a specific focus on the necessary minimum and optimal sample sizes. This is a natural extension of one of the main reasons MRP was developed: the inability to reliably infer small area opinion or behaviour with standard sample sizes. However, there may also be value in paying closer attention to how the sample is distributed across small areas. That is, how many respondents each small area has and how this may - or may not - affect our estimates of opinion or behaviour.

Most applications of MRP make use of publicly available survey data which means researchers do not have control over sample distribution. As a result, discussion on sample rarely goes beyond recognising that the sample distribution may impact individual small area estimates (Buttice and Highton, 2013; Muller and Schrage, 2014: 144; Park et al., 2004: 320; Lax and Phillips, 2009b: 111; Pacheco, 2011). However, if we have control over the sampling procedure, researchers should pay attention to the distribution as this has the potential to affect estimates. Indeed, in certain applications, structuring the sample distribution so that certain small areas receive a larger proportion of the sample may improve prediction accuracy. This is particularly relevant if we have finite resources and therefore a limited total sample size. In these instances, an uneven sample distribution may be the best strategy to achieve the highest accuracy possible.

Applications for an uneven distribution

Distributing the sample unevenly among small areas is most likely not preferable in all MRP applications. However, in instances where it is preferential or required to achieve a higher level of accuracy in certain small areas, an uneven sample distribution

may be useful.

The most obvious example is in electoral forecasting, where our ability to correctly predict an electoral outcome (i.e. who governs) rests on prediction accuracy within a subset of small areas, not accuracy across all small areas.

This is because in representative democracies, the electoral outcome is often decided by which party wins an overall majority of all electoral districts (i.e. small areas).[3] However, in practice the distribution of the electorate is such that most electoral districts are consistently won by the same party, while some electoral districts - known as marginals - regularly switch between parties and have lower margin of victory. For example, among the 632 GB constituencies, 70% have been won by the same party over the last four elections, with an average margin of victory of 27%. Whereas constituencies which have changed twice or more over the last four elections have an average margin of victory of 10%.

A strategy which focuses on ensuring greater accuracy in specific small areas might, therefore, improve our ability to predict an electoral outcome. By focusing on improving accuracy in marginal small areas we risk worsening accuracy in both non-marginal and across all small areas. However, if our goal is electoral prediction the benefits of this method might still outweigh these negatives.

Take for example the fake scenario in table 4.1. The table shows vote share estimates for two different models, as well as overall model MAE. Both models are estimating vote share for party X in small areas area 1 and 2. Assuming a two-party competition where a party needs >50% to win, Party X wins the vote in both small areas 1 and 2. In Model A, the predictions are within 3% of the true vote share, but in area 2 the prediction fails to correctly forecast party X as the winner. On the other hand, model B has improved accuracy in area 2, significantly worse accuracy for small area 1, but predicts the seat winner for both. Model A achieves better MAE

[3]This is major reason why MRP has been used to predict multiple elections, as the method enables us to predict vote share in each small area.

Table 4.1: Example scenario

		Model A		Model B	
	Actual Vote	Estimates	Absolute error	Estimates	Absolute error
Area 1	70%	67%	3%	60%	10%
Area 2	52%	49%	3%	51%	1%
MAE			3%		6%

overall and would be considered the better model by this metric. However, if we assess by the ability of models to predict seat winners, then Model B is preferable. Such a contrived example may never manifest in practice but illustrates how assessing a model by estimate accuracy may not lead us to the *best* model should our objective be predicting an election, rather than estimating vote share.

Multilevel modelling and uneven sample distribution

For small areas where higher levels of accuracy are deemed to be of greater importance, the benefit of an uneven sample distribution may be two-fold. First, as these small areas would receive a larger proportion of the sample, the raw data for these small areas will most likely be closer to the true value. This would mean the model should be better informed and equipped to produce accurate estimates.[4] Second, assuming the raw data is more accurate, the larger sample size will mean less shrinkage induced through partial pooling. Therefore, the resulting area-level parameters will be allowed to be distinct from the overall mean to a greater extent.

Conversely, small areas which receive a smaller proportion of the sample are at potential risk of greater inaccuracy. Furthermore, it is currently unknown how an uneven sample distribution may affect the estimation of small area variance. Area-level parameter estimation may be problematic with an uneven sample distribution because

[4]This assumes that increasing sample size will improve accuracy of the raw data. This assumption will not necessarily hold true in every case. However, across all small areas it seems a fair assumption that for the most part increasing sample size will improve accuracy.

the estimation is based on a sample over-represented by respondents from certain small areas. If voting behaviour in over-sampled small areas is distinct from other small areas, then the model will learn too much from a sub-group who are not representative of the wider population.

The same potential problem applies to the individual-level variables (which are modelled as varying intercept terms). If the unevenly distributed sample leads to the over representation of certain sub-groups, then the estimation of individual-level characteristics will be more difficult as some parameters will be estimated with low N. However, this is less likely to arise as individual-level variables typically have few categories, while the sample should be sufficiently large that the uneven sample distribution will not affect estimation of these parameters.

As already noted, in MRP applications there is somewhat limited discussion on sample characteristics, including the distribution of respondents among small areas. Where discussion is present, it is largely restricted to simple acknowledgment that areas with a larger sample size are more likely to have estimates distinct from the overall mean (see Muller and Schrage, 2014: 144; Park et al., 2004: 320). One of the few studies which has directly addressed varied sample distribution and estimate accuracy is Toshkov (2015), whose study explored potential use of MRP to estimate opinion among EU states. As part of the study, he estimated two identical models with different sample distributions. One model had an evenly distributed sample, and in the other, each country sample size was relative to its proportion of the total EU population. He found no significant differences between the two strategies, but in some cases the uneven sample distribution produced more accurate estimates (Toshkov, 2015: 457).

Research by Wang et al. (2015) investigated the use of highly unrepresentative samples to estimate vote share for the 2016 US Presidential election. Their sample was acquired by a survey of online Xbox gamers. The sample was highly unrepresentative,

comprised largely of young males. They showed that the "borrowing of strength" of MRP enabled the model to reliably estimate for all voter types, despite the sample being unevenly spread across groups. However, the study used a sample which had around 400,000 respondents. This sample size meant that groups who were proportionally underrepresented were still much larger in absolute terms than sample sizes in most MRP applications. Therefore, the results from this study might not be indicative of parameter estimation when samples are much smaller.

The limited discussion on sample size distribution in MRP is mirrored in the wider multilevel modelling literature. Here, most studies discuss and analyse sample size requirements assuming the even distribution of sample among sub-groups (Cohen, 2005; McNeish and Stapleton, 2016). Although more recently, some studies looking at multilevel modelling have begun to incorporate or directly address unevenly distributed samples. Across a variety of fields and different conditions, several studies have investigated how uneven sample distribution may - or may not -impact multilevel model parameter estimation. In each case the authors stress the importance of not interpreting the findings beyond the specific conditions they analyse. Nonetheless, the findings can inform our expectations of how MRP estimation will perform with uneven samples.

Research in the field of randomized control experiments, tested whether an uneven sample distribution negatively impacts Type I error rate, statistical power, bias, and standard errors for mixed effect logistic regression models (Heo and Leon, 2005). They argue their results indicate an uneven distribution does not negatively impact the measures they use to evaluate the models (Heo and Leon, 2005). In education research, Milliren et al. (2018) investigated whether an unbalanced sample distribution was detrimental to linear random effect models. Their results also show no significant negative impact on the ability of a model to identify true random effects with an unbalanced sample distribution. Similar conclusions were reached by Schoeneberger

(2016) for the multilevel logistic case.[5] Specifically, they found that performance is mostly affected by level-2 sample size, with logistic models requiring larger sample sizes (Schoeneberger, 2016).

Investigating data sparseness for multilevel models, Bell et al. (2008) explored how increasing the proportion of cells which had a single respondent impacted model performance for linear multilevel models. They found that the proportion of singleton cells had a limited impact on model performance when there was a sufficiently large number of small areas.[6] However, when there were few areas and a higher proportion of singletons, the accuracy of level-2 parameters were negatively impacted. Broadly, in linear and logistic multilevel models, data spareness does not negatively impact model estimation as long as there is sufficient number of small areas (>50) and sufficient sample per small area (>3-5 respondents) (Clarke and Wheaton, 2007; Clarke, 2008). However, when these conditions are not met, variance estimation can be problematic, with inaccurate estimation more pronounced for logistic applications (Clarke, 2008).

4.2 Data and Methods

4.2.1 Simulation study

The simulation component of this research involved two stages. First, creating fake data and, second, estimating MRP models for different sample sizes, number of small areas, and sample distributions.

[5] Although their study did not directly address uneven sample distribution, they incorporated uneven sample distribution into the simulation study.

[6] They evaluated model performance by looking at variance estimation, fixed effects, standard errors, and convergence.

Simulating fake data

To create "fake data" for the simulation study, I followed the procedure itemised below:

(1) Start with a "real" poststratification frame.

To create fake data that was comparable to real data for an actual population, I used a "real" poststratification frame of the GB population.[7] The poststratification frame included the joint distributions for a variety of individual characteristics for all 632 GB constituencies. This acted as the base from which I created a simulated data for a fake population.

(2) Create individual-level variables and append area-level variables.

In line with standard practice identified in chapter 2, I use 4 individual-level and 3 area-level variables. To create the individual-level variables I re-coded variables from the real poststratification frame, collapsing some categories. This provided me with 4 individual-level variables with varying categories: X1 (2 categories), X2 (3 categories), X3 (4 categories), and X4 (5 categories). Altogether, this meant that each small area had 120 different person-type or individual-level variable joint-distribution variations. To generate area-level variables, I used actual political and demographic statistics for GB constituencies.[8] The area-level variables (renamed A1, A2 and A3) were appended to the poststratification frame for each small area. To create a person-type weight for the poststratification frame, that is, the proportion each person type represents in each small area, I made use of the weight included in the original poststratification frame. As the original frame included more person types, I recalculated the weight so it was representative of the 120 person-types in the new poststratification frame.

[7] It can be accessed here: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IPPPNU`

[8] I used Conservative party 2017 vote share, 2016 EU referendum vote choice, and % 18-24. These were chosen as variables selected by lasso regression in chapter 3.

(3) Creating fake Y.

The procedure of creating Y was undertaken with the intention of imitating a binary vote choice in a two-party system where 1 indicates a vote for party A and 0 a vote for party B. To generate Y for each person type (i.e. each row of the poststratification frame), I first calculated the probability that Y = 1 for each person type. This was achieved by taking the inverse-logit of the sum of each variable multiplied by a coefficient. Thus $Pr(Y_i = 1)$ was calculated by the following:

$$Pr(Y_i = 1) \;=\; logit^{-1} \sum_{g=1}^{g} X_{g_i} \beta_g \tag{4.1}$$

Here g references the gth variable of $g = 1..., G$ variables, while X is a N*G matrix, β_g is a vector of coefficients for each variable, and i indexes an individual. To produce Y, I drew binary outcomes (0,1) from a Bernoulli distribution:

$$Pr(Y_i = 1) = p_i, \;\; Pr(Y_i = 0) = q_i = 1 - p_i \tag{4.2}$$

Where i references the ith individual in the data frame, p is the probability from equation 1 and q is the inverse of p. I directly included area-level error into the calculation of p_i, while individual-level error was introduced by the random draws.

(4) Create "master" poststratification frames for each small area N.

To create the poststratification frames for each small area N, that is, for the various number of small areas tested here (either 50, 200, 400 or 600), I randomly sampled small areas from the poststratification frame created by steps 1-3. The sampling was random without replacement and provided me with four master poststratification frames for each of the small area N: 50, 200, 400, and 600.

(5) Create fake individual-level survey data.

To generate fake survey data, I sampled person types from the master poststratification frames. The procedure was random sampling with replacement, using the person type weight as probability for selection. This procedure meant I imitated random survey sampling where more populous sub-groups are represented to a greater degree in the survey sample. For each small area N, the sampling procedure was replicated for four different average small area sample sizes (5, 10, 20 and 30) and for three different sample distribution ratios (Even, 2:1, and 3:2:1).[9] In total, the process generated, 48 data sets with four unique small area sample sizes (5, 10, 20, and 30), three different ratios (Even, 2:1, and 3:2:1), and four unique number of areas (50, 200, 400, and 600).[10]

Ratios were calculated by using the small area margin of victory (i.e. the difference between proportion of Y =1 and Y= 0). Thus, where s references each small area, the calculation was as follows:

$$Margin = abs(\sum Y_s = 1 - \sum Y_s = 0) \tag{4.3}$$

Using the margin, I divided small areas evenly into thirds or halves depending on the sample distribution ratio. There were three sample distribution ratios used: 'Even' which meant the sample was evenly spread among small areas, '2:1' ratio, where the 'most marginal' group (i.e. smallest margin of victory) received around 2/3 of the sample and the 'least marginal' group a 1/3 of the sample. Finally, the '3:2:1'"' ratio where the 'most marginal' group received 1/2 of the sample, the 'mid marginal' group received 1/3 of the sample and the 'least marginal' received 1/6 of the sample.

[9]To ensure that the sampling procedure and the resulting samples were comparable, I used the R set.seed function to ensure consistency in sampling. This meant, that for data frames where the sample per small area ratios were either '2:1' or '3:2:1', the areas with more respondents would have the same respondents as the 'Even' sample dataset, plus additional respondents. This was to ensure that increasing sample size would most likely increase accuracy of the raw data.

[10]For samples where the ratios are not even, the small area sample sizes are average sample size per small area.

Simulation modelling procedure

For the 48 variations in sample size, small area N and sample distribution ratios, I estimated the proportion of Y = 1 in each small areas using multilevel regression and poststratification (MRP). Except for the sample size, small area N and sample distributions, all models were estimated in an identical way. I first estimated Bayesian multilevel logistic models. Each mode included X2, X3 and X4 as random effects, and X1, A1, A2, and A3 as fixed effects.[11] Each model included weakly-informative student-t priors on both the intercept and the coefficients (for a discussion on prior choices, see Gelman et al., 2008). The model was thus as follows:

$$Pr(Y_i = 1) \ = logit^{-1}(\beta_\theta + \beta^{X1}_{X1[i]} + \alpha^{X2}_{j[i]} + \alpha^{X3}_{k[i]} + \alpha^{X4}_{l[i]} + \alpha^{Area}_{s[i]}), \tag{4.4}$$

Where each random intercept term ($\alpha^{X2}_j,\ \alpha^{X3}_k \alpha^{X4}_l$) is assumed to be normally distributed with a mean zero and some variance σ^2.

$$\alpha^{X2}_j \ \sim \ N(0,\ \sigma^2)\ for\ j = 1, ..., 3$$

$$\alpha^{X3}_k \ \sim \ N(0,\ \sigma^2)\ for\ k = 1, ..., 4$$

$$\alpha^{X4}_l \ \sim \ N(0,\ \sigma^2)\ for\ l = 1, ...5$$

The small area random intercept term (α^{Area}_s) is itself modelled as a function of area-level variables $A1,\ A2,\ A3$, and again, assumed to be normally distributed with some variance. Where s indexes the relevant small area, it is:

$$\alpha^{Area}_s \ \sim \ N(\beta^{A1} \cdot A1_s + \beta^{A2} \cdot A2_s + \beta^{A3} \cdot A3_s,\ \sigma^2)\ for\ s = (1, ..., S)$$

[11] X1 was an individual-level variable and could have been included as a varying intercept term, as was done with all the other individual-level variables.

For all β and the σ parameters for the random intercept terms, I include student-t priors with 5 degrees of freedom, mean 0 and standard deviation of 5. Each model was estimated with 2 chains, 500 warm-up and 500 sampling. To improve estimation and reduce computation time, the models included QR decomposition on the fixed effects. All models were estimated with Stan called through RStan (S. D. Team, 2020) and implemented with stan_glmer function of rstanarm (Goodrich et al., 2020). To estimate the proportion of individuals in each small area where Y = 1, the multilevel estimates were poststratified using each small area N poststratification frame. In practice this meant I drew 500 samples from the multilevel model posteriors using the poststratification frames as new data.

$$Y_s^{mrp} = \frac{\sum_{c \epsilon s} N_c \, \pi_c}{\sum_{c \epsilon s} N_c} \tag{4.5}$$

Here s indicates the small area, N_c is the relevant population count and π_c is the person-type cell estimate. The final small areas estimates are the mean and the 90% credible intervals for the estimates.

4.2.2 External validation

The second stage of the research - the external validation - estimated voting behaviour in the UK and US. I estimated Conservative Party vote share in UK constituencies for the 2019 election and Republican vote share at state-level for the 2016 US presidential election. These elections represent real-world manifestations of the upper and lower small area N used in the simulation study.[12] Broadly, the external validation followed a similar strategy to the simulation study, with minor modifications necessary for each applied case.

[12]UK constituencies number 632 and US states number 50.

Individual (survey) data

For the UK validation study, I made use of the British Election Study (BES) election campaign wave survey. The survey included around 30,000 respondents from all UK constituencies and included vote intention as well as respondent demographic characteristics. For the 2016 US Presidential election, individual-level data was obtained from the Cooperative Congressional Election Study (CCES). Here I use the pre-election survey data. The survey sample size was around 50,000+ respondents, with respondents from all 50 US states.

Poststratification frame

For the UK, I used a publicly available poststratification frame from Hanretty et al. (2017). This poststratification frame included all demographic individual-level variables used in the multilevel model. The frame was constructed using UK census data and a raking procedure to provide a range of individual-level variable joint-distributions at GB constituency level.[13] For the US poststratification frame, I constructed using the 2014-2018 American Community Survey (ACS) 5-year Public Use Microdata Sample. The data has over 18 million individual responses and is regarded as a reliable source for deriving demographic joint-distributions required for the poststratification frame. I calculated person type (or cell) weights by summing the number of individuals in each cell and dividing by the total population for the given state. The procedure was as follows:

$$\frac{\sum_{c \epsilon t} N_c}{\sum_{c \epsilon t} N_s} \tag{4.6}$$

Where t references the total number of person types, c indexes a unique cell in t, and s the state.

[13]For a detailed account of poststratification construction, please refer to Hanretty et al. (2016), Hanretty et al. (2017), and Hanretty (2019).

Variables

Vote intention for the Conservative Party was derived from a vote intention question which asked whether "in the upcoming election the respondent knew which party they would vote for". I excluded those who reported they would not vote and those who reported they "Did not know". US vote intention was derived from one of two questions in the survey. For those who had already voted, I used their declared vote. For those who were yet to vote, I derived vote intention from a question which asked for a respondent's preferred candidate for the 2016 upcoming US presidential election.[14] I

Table 4.2: Model variables

Individual-level	Area-level
UK	
Gender (2)	GE2017 vote
Age (8)	EU2016 vote
Education (6)	% Long-term unemployed
Campaign week (4)	% Industry manufacturing
	Population density
	Region
US	
Gender (2)	Region
Age (4)	2012 vote
Ethnicity (4)	
Education (4)	
Marital status (4)	
Campaign week (4)	

excluded those who stated they would not vote and those who did not intend to vote for either Republican or Democrat presidential candidates.

Individual and area-level variable selection was based on variables used in the MRP models of Lauderdale et al. (2020). In their paper, they estimate UK party vote share for the 2017 General Election and the US 2016 presidential election. I have not included individual-level past vote due to limitations on available data and to

[14]Candidate preference is an imperfect proxy for vote intention, but I believe is a sufficient proxy for the purposes of this study: comparing sampling strategies.

reduce model complexity/computation time, I have not included interactions used in Lauderdale et al. (2020). For both models, to account for temporal change in vote choice, I also include a varying intercept term for the week of survey fieldwork. The variables used in each case are reported in table 4.2.

Sample distribution and size

Assigning small areas into marginal groups is more complicated in a real-world setting than was the case for in the simulation study. First, in a real world setting we would not know the margin of victory prior to the election. As the best substitute, I made use of margin of victory in the previous election.[15] For the UK, I used past vote from the 2017 election, calculating difference between the Conservative and all other parties' vote share. For the US, I calculated the difference between Republican and Democratic vote share. Second, in a real-world setting, splitting small areas into groups of thirds or halves - as has been done in the simulation study - is not practical. Using this method could result in small areas included in the 'most marginal group', even though by conventional metrics they would not be considered a marginal small area. Instead, I assigned a small area into a marginal group based on whether past margin of victory is below either 5% or 10%. In table 4.3 I report the cut-off points used to assign small areas into marginal and non-marginal groups.

Table 4.3: Marginal group categories

	Most marginal	Mid marginal	Least marginal
Two groups (2:1)	<10%		≥ 10%
Three groups (3:2:1)	<5%	≥ 5% & <10%	≥ 10%

[15]There are risks by using past vote as we are assuming that small areas which were marginal in a past election will again be marginal. If this is not the case, our strategy risks greater inaccuracy in important small areas than would be achieved with an evenly spread sample.

Determining sample distribution among small areas groups is also more complicated than was the case in the simulation study. In most applications there are too few small areas in the 'most marginal' and 'mid marginal' groups, such that the ratios used in the simulation are no longer applicable from both a modelling and a practical standpoint. From a modelling perspective, the 'least marginal' group, which includes most small areas, would receive too small a proportion of the sample. Although we wish to redistribute the sample, we still need sufficient number of respondents in non-marginal small areas for reliable estimation. From a practical perspective, the necessary samples sizes for 'most' and 'mid' marginal small areas are not feasible with the surveys I use here (BES and CCES).

As an alternative, I calculated a weighted ratio for the sample distribution. To calculate this, I first create a weight that is the number of small areas in a group divided by the total number of small areas. I take the proportion of sample that each group should receive, according to the original ratios, and multiply by the weight. The weighted ratio is then rescaled so that the sum of weighted ratios equals 1. The calculation is as follows:

$$\frac{wt_g \cdot R_g}{\sum_{g=1}^{G} wt_g \cdot R_g} \tag{4.7}$$

Where g references the group of small areas, R the original ratio and wt the weight as described above. The weighted ratio thus provides a means to calculate the sample distribution that accounts for marginal group size, that is, the number of small areas in each group.

The weighted ratio determined the sample distribution, and provided the required sample per small area. To select survey respondents from the total sample, I randomly sampled respondents without replacement from each small area using the survey weight as the probability for selection. I completed the sampling procedure once for both US and UK. This provided me with a single survey sample for each of the ratios. In the US, the three samples have on average 30 respondents per small area, with the

sample distributions as follows:

- Even: 30 respondents per small area
- '2:1': 23:46 respondents per small area
- '3:2:1': 21:42:63 respondents per small area.

In the UK, the BES sample did not have adequate coverage across small areas to enable an average of 30 respondents per small area. Instead, I tested an average of 20 respondents per small area across the three distributions.[16] The distributions are as follows:

- Even: 20 respondents
- '2:1': 16:32 respondents per small area
- '3:2:1': 15:29:44 respondents per small area.

Model estimation

For each unique sample - 6 in total - I have followed the estimation strategy pursued by Hanretty et al. (2016) and Selb and Munzert (2011) where I estimate vote choice as a binary outcome .[17] In the UK application, I estimated Conservative vote share within each GB constituency (632) for the 2019 general election (Conservative = 1, all other parties = 0). In the US case, I estimated Republican vote share in each US state (50) for the 2016 presidential election. I again estimated vote choice as a binary outcome, where Republican = 1 and Democrat = 0. In both cases I excluded those who reported they would not vote and those who stated they "Don't know". In the

[16]Even with an average of 20 respondents, some small areas did not have the sufficient sample size required for a given ratio distribution. For the 'Even' and '2:1' distribution, around 99% of small areas had sufficient sample sizes. For the '3:2:1' distribution, only 83% of small areas had the necessary 44 respondents. However, all except one small area had larger sample sizes than 29, with the majority relatively close to the desired 44 respondents.

[17]An alternative would be to estimate all parties with a multilevel multinomial model, as is the case in Lauderdale et al. (2020). There are clear benefits to this strategy but it is far more computationally demanding.

US application, I also excluded those who did not intend to vote for either Republican or Democrat.

In both US and UK applications, except for different sample distributions all models were identical. In each case, vote choice was first estimated as a Bayesian multilevel logistic regression model. I estimated the models with student-t priors on all varying intercept σ and all β parameters, with 5 degrees of freedom, mean of 0 and 5 standard deviation. Each model had two chains of 1000 iterations (500 warm-up and 500 sampling). All models were estimated with Stan called through RStan (S. D. Team, 2020) and implemented with the brms package (Bürkner, 2017).[18]

To produce small area vote share, I poststratified the multilevel estimates for UK and US with their respective poststratification frames. In both applications, the poststratification frames I have used are for all adults. This means that the estimates are vote share as a proportion of all adults in each small area, rather than of voting population as would be normal in electoral forecasting.

To account for voter turnout, I have applied actual small area turnout to estimates for each election. This strategy fails to account for any sub-group turnout differentials, something which we know to be apparent. However, to be able to apply differential turnout to subgroups in the electorate we need to estimate turnout first. This is a difficult task, and one which is often harder than forecasting vote choice. For instance, Lauderdale et al. (2020) argue that their turnout estimates were consistently unsatisfactory and responsible for some of the error in their constituency and state vote share estimates. For simplicity and to avoid needing to disentangle effects of turnout and the effects of uneven sample distribution, I apply actual turnout to small

[18]I include full UK and US model notation in appendix C.2 and C.3.

area vote estimates.[19] The final estimates are thus:

$$Y_s^{Vote} = Y_s^{mrp} * Turnout \tag{4.8}$$

4.3 Results

I first present the results of the simulation study followed by the two external validation exercises. When presenting the results here, I discuss accuracy of the MRP estimates rather than the estimates themselves. However, accuracy in this chapter takes on two related but different meanings. For the most part, accuracy refers to the comparison between the estimates and actual party vote share. The measures used to represent this type of accuracy are mean absolute error (MAE), root mean squared error (RMSE) and correlation. In the validation exercises, I also refer to seat prediction accuracy, which refers to whether the estimates correctly predicted the party winner. For seat prediction accuracy, I make use of the Brier score.[20]

4.3.1 Simulation study

The simulation study included 48 separate models; for four different small area N (50, 200, 400 and 600), four different average sample sizes (5, 10, 20 and 30) and three different sample distributions (Even, 2:1 and 3:2:1).[21] Below I present MRP estimate accuracy by comparing estimates to *true* vote share generated in the "simulating fake data" procedure.

[19]To provide confidence that the turnout measure did not impact the results, I produced results with a model-based turnout measure as well. I explain estimation procedure for the model-based turnout and show results produced with this alternative turnout in appendix C.5. Importantly, the findings in this chapter do not change with the model-based turnout.

[20]The brier score is a measure used to calculate the accuracy of probabilistic predictions.

[21]For uneven distribution sample designs, the sample size refers to the average sample size.

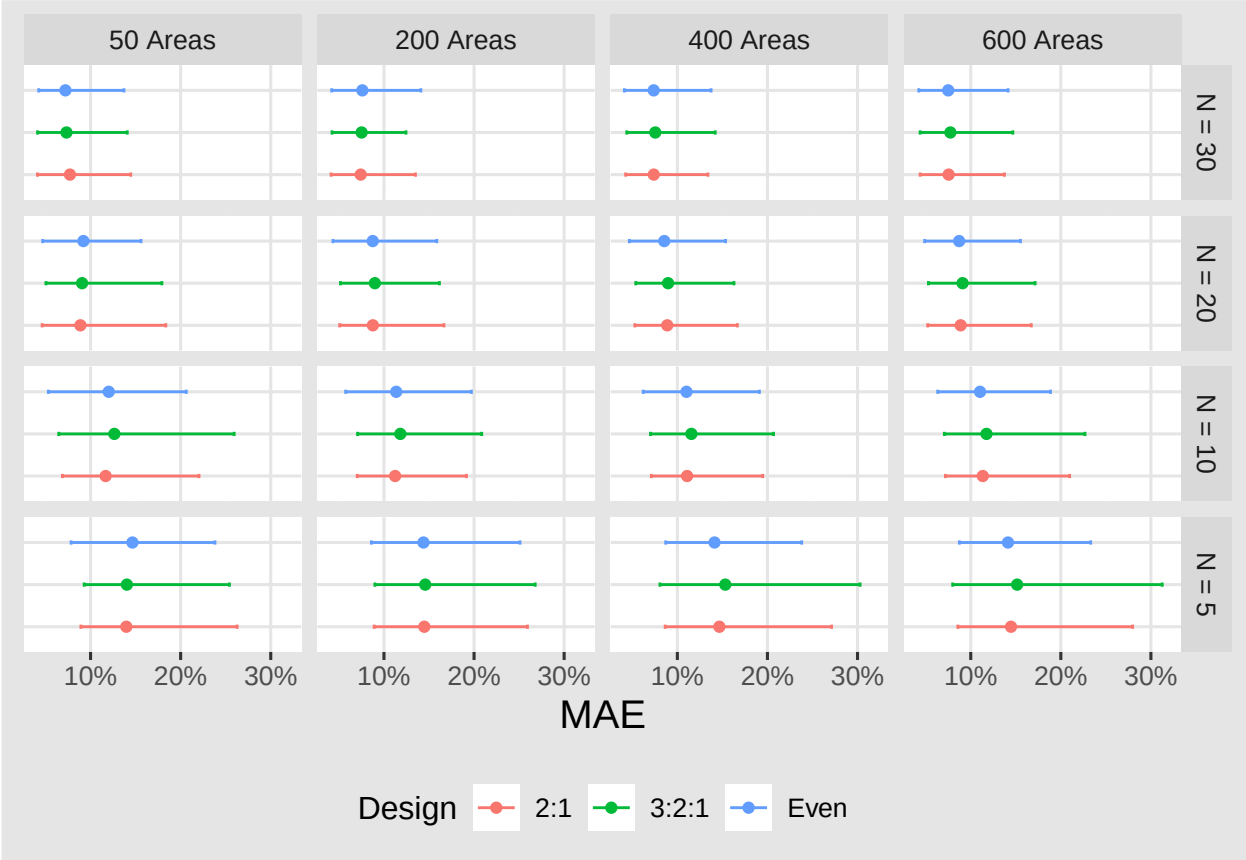

Figure 4.1: Simulation model accuracy
Notes: showing MAE for all sample sizes (rows) and small area N (columns). Points show average MAE with 90% CI shown by lines. Even sample distributions show in blue, 2:1 in red and 3:2:1 by green.

The uneven sample distribution has been implemented to improve accuracy in certain small areas. However, it is important to consider how changes in the sample distribution impact overall small area accuracy. This will allow us to understand how the uneven sample distribution affects model estimates more generally. Accordingly, figure 4.1 shows MAE of estimates for all small area N, sample size and sample distribution. Each column represents a different number of small areas, starting at 50 on the left and increasing to 600 small areas in the right-hand column. Each row shows MAE for different average sample size per small area, starting at N = 30 at the top and going down to N = 5. Finally, the colours indicate the sample distribution. Blue shows MAE for 'Even' distribution models, red for '2:1' distribution models, and

green '3:2:1' distribution models. In each plot, the point represents the average MAE across all small areas, while the lines indicate the 90% credible intervals.

Overall, the accuracy figures reported in figure 4.1 are in line with expectations informed from previous research. As we would expect, the larger samples produce more accurate estimates regardless of sample design and number of small areas. From the figure, it seems the uneven sample distribution has relatively little effect on overall estimate accuracy. In each plot, the point estimates (indicating average MAE), show little difference regardless of sample distribution. Similarly, the 90% credible intervals seem similar in length for nearly all plots. The results are largely unaffected by the number of small areas, but sample size seems to have a small effect, with '2:1' and '3:2:1' MAE widths are longer when sample sizes are smaller. Altogether, the figure highlights that the uneven sample distributions do not have a significant effect on overall MRP estimate accuracy.

As previously noted, our primary interest is to compare the prediction accuracy of certain small areas between different sample distributions. To determine sample distribution, the simulation study assigned small areas according to margin of difference (calculated as the difference between the proportion $Y = 1$ and the proportion $Y \neq 1$). This meant that, for '2:1' and '3:2:1' sample distributions, areas with a larger margin of difference received a smaller proportion of the sample and should therefore have poorer estimate accuracy. On the other hand, small areas with a smaller margin of difference received a larger proportion of the sample and should therefore have higher accuracy.

To investigate how uneven sample distributions impact estimate accuracy for marginal small areas, table 4.4 analyses results by grouping small areas according to the marginal grouping (most, mid and least marginal). These are the same groupings used to determine the sample distribution, where the most marginal group received the largest proportion of the sample. The table shows MAE and the width - calculated as

Table 4.4: Simulation small area accuracy (3:2:1 distribution)

	50 Areas		200 Areas		400 Areas		600 Areas	
Groups	MAE	Width	MAE	Width	MAE	Width	MAE	Width
Sample: 30								
Most marginal	-0.01	-0.03	-0.03	-0.04	-0.02	-0.04	-0.02	-0.03
Mid marginal	0.00	0.01	0.00	0.00	0.00	0.00	0.00	-0.01
Least marginal	0.02	0.07	0.02	0.07	0.02	0.06	0.03	0.07
Sample: 20								
Most marginal	-0.02	-0.06	-0.02	-0.04	-0.01	-0.05	-0.01	-0.05
Mid marginal	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Least marginal	0.01	0.07	0.02	0.07	0.03	0.07	0.02	0.07
Sample: 10								
Most marginal	-0.03	-0.08	-0.03	-0.06	-0.02	-0.06	-0.02	-0.06
Mid marginal	0.00	-0.02	-0.01	-0.01	0.00	-0.02	0.00	-0.01
Least marginal	0.05	0.08	0.05	0.08	0.04	0.08	0.05	0.07
Sample: 5								
Most marginal	-0.03	-0.08	-0.03	-0.09	-0.05	-0.11	-0.04	-0.11
Mid marginal	0.00	-0.01	-0.01	-0.04	-0.01	-0.07	-0.01	-0.06
Least marginal	0.03	0.04	0.05	0.04	0.10	0.03	0.09	0.02

Note: Showing +/- from 'Even' MAE and width. Areas grouped into marginal categories

the difference between the 90% credible intervals - for the '3:2:1' sample distribution.[22] I report results for the different sample sizes (across rows) and the different small area N (shown in the columns). The figures reported in the table are the increase or decrease from the 'Even' sample distribution. A minus figures indicates a lower MAE or width (i.e. an improvement), while a positive value indicates higher MAE or a larger width (i.e. poorer accuracy). From the table, we can therefore see that when estimating in 50 areas with a sample size of 30, the most marginal group MAE is 1% less than the 'Even' distribution, while the width is 3% shorter.

Importantly, the results presented in the table demonstrate the effectiveness of an uneven sample distribution. For every sample size and for every small area N, we

[22]In appendix C.1 I present the same table for the sample distribution with two groups and a ratio of '2:1'.

consistently see that for the most marginal small areas MAE decreases and widths shorten. In each case, this improvement represents a reduction in MAE of between 1-5%, while widths are between 3-11% shorter. This means that for the most important small areas, we improve accuracy as well as precision and confidence in our estimates. However, for the least marginal areas MAE increases and widths widen. MAE increases are typically around 2-5%, but for the smallest sample size go up to 10%. For the 'mid marginal' group, MAE is either identical or 1% less than the 'Even' distribution MAE.

This pattern is evident for all sample sizes but is particularly notable when average sample sizes are smaller. This is most likely because with smaller sample sizes, the increase/decrease in respondents may have a larger impact on making the underlying raw data more/less accurate than might be the case for larger sample sizes. And as a result, when the sample is unevenly distributed the differences in prediction error are starker.

4.3.2 External validation

Turning to the external validation, I next present the findings from the application of an uneven sample distribution to real-world settings: 2019 UK General Election and 2016 US presidential election. Overall the findings presented from the simulation study seem to be upheld in the real-world application.

UK

I first present the accuracy for models estimating Conservative Party vote share at the UK 2019 General Election. The election saw the Conservative Party win an overall majority, by winning 363 of the 632 Great Britain electoral districts available.[23]

[23]The UK parliament is made up for 650 MPs elected to represent UK electoral districts. However, no GB party stands in Northern Ireland, and typically modelling of UK general election does not include constituencies for Northern Ireland.

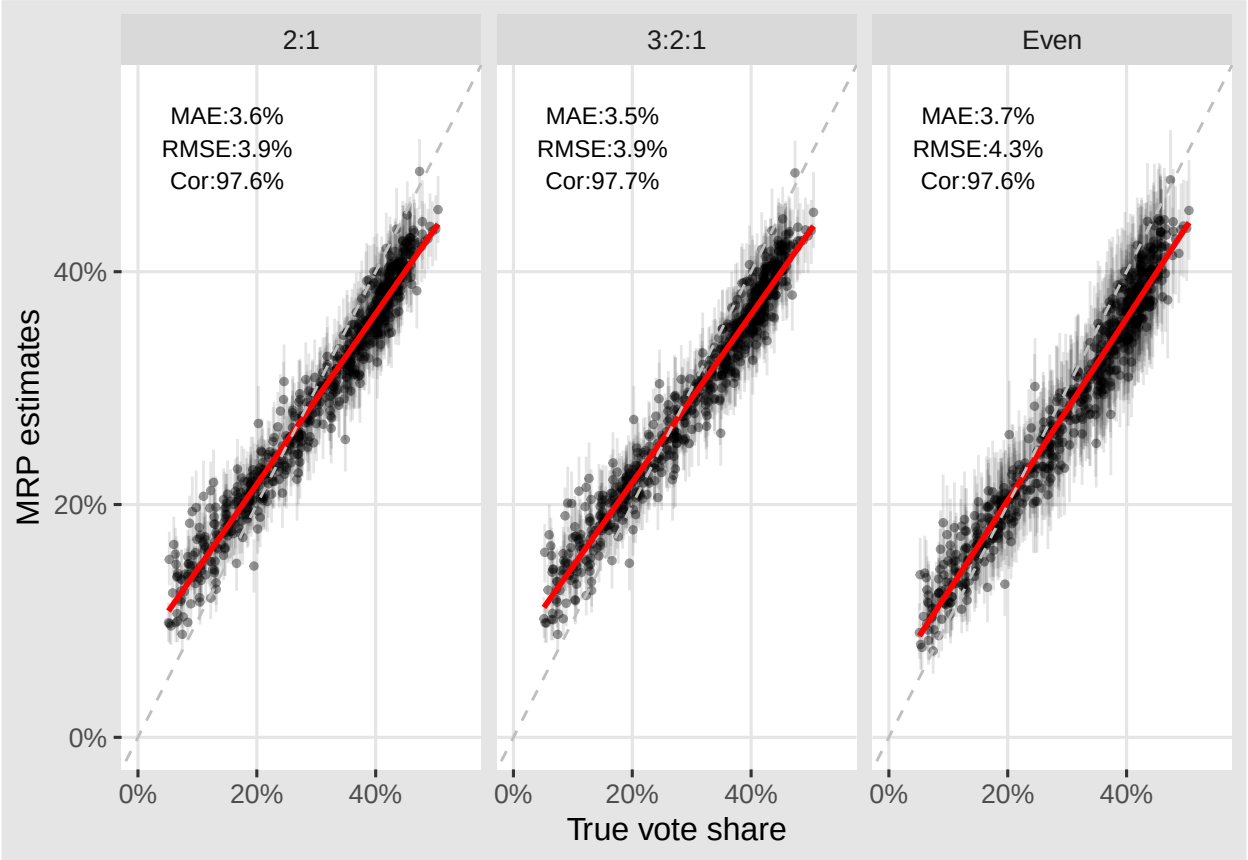

Figure 4.2: UK estimates vs. true vote share
Notes: Points show mean constituency estimate with 90% CI shown by vertical lines. Red line shows line of best fit and dashed grey line shows 45 degree line.

Figure 4.2 shows scatter plots showing estimates (on the y-axis) vs the actual 2019 Conservative vote share (on the x-axis). The figure shows three plots for the 'Even', '2:1' and '3:2:1' sample distributions. On each, the points represent the mean estimate while the vertical lines show the 90% credible intervals. The red line represents the line of best fit across all small areas, while the grey dashed line represents the line for a perfectly linear relationship. On each plot I also report the MAE, the root mean squared error (RMSE) and correlation (Cor).

From figure 4.2 we can see the '3:2:1' sample distribution achieves the lowest MAE and RMSE, followed by the '2:1' model. This is contrary to the results of the simulation study where overall accuracy showed little difference between sample distributions. However, the differences are relatively small, with all models achieving

Table 4.5: UK: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.035	0.081
Mid marginal	0.041	0.094
Most marginal	0.045	0.093
3:2:1		
Least marginal	0.036	0.057
Mid marginal	0.035	0.062
Most marginal	0.037	0.061

Table 4.6: UK: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.035	0.081
Most marginal	0.043	0.093
2:1		
Least marginal	0.035	0.057
Most marginal	0.036	0.062

a MAE within 0.2% of each other. Similarly, the difference in correlation between all three models is almost non-existent. The '3:2:1' model achieves correlation of 97.7%, while both other models achieve 97.6%. RMSE shows slightly greater differences between the uneven and 'Even' sample distributions. Two uneven samples achieve a RMSE of 3.9%, while the 'Even' sample distribution achieved RMSE of 4.3%.

Next, I examine the difference in accuracy of small area marginal groups - the same groupings used to determine sample distributions. Table 4.5 compares 'Even' and '3:2:1' distribution, showing MAE and width for the 'most', 'mid', and 'least' marginal small areas. Table 4.6 compares accuracy between 'Even' and '2:1' sample distribution model, showing MAE and width for 'most' and 'least' marginal small areas.

From table 4.5 and 4.6 we can see that the '2:1' and '3:2:1 'sample distribution models achieved greater accuracy for the most marginal small area groups when

compared to the evenly distributed sample. In both cases, the MAE of most marginal small areas was between 0.7-0.8% less than for the most marginal small areas of the 'Even' distribution model. Unlike in the simulation study, the improvements in accuracy for most marginal small areas was not at the cost of accuracy in the least marginal small areas. In the '3:2:1' distribution the least marginal small areas has 0.1% worse MAE, while for the '2:1' distribution MAE was the same as the 'Even' sample distribution. Furthermore, we can see that the estimate widths for the uneven sample distribution models are shorter than the 'Even' distribution model for all small areas regardless of which marginal grouping they were. Broadly, the results seem to indicate that in a real application, we can improve estimate accuracy in certain small areas with an uneven sample distribution. However, this does not tell us whether we have improved our ability to predict an election, as was the basis and justification for an uneven sample distribution.

To investigate whether uneven sample distributions improve seat prediction accuracy, I next show the brier score of estimates for each model. The results are presented in the table 4.7.[24] The brier scores of the '3:2:1' and '2:1' sample distribution were identical (although unrounded scores show '3:2:1' achieves the best brier score), with both showing an improvement over the 'Even' sample distribution score. The results certainly lend weight to the central argument here: an uneven sample distribution improves accuracy in marginal small areas, and this in turn can improve the probability of predicting an electoral outcome.

US

The results for the 2016 US presidential election are presented below. They broadly show similar patterns to the simulation and the UK application. The 2016 US

[24]I calculate the probability of a Conservative victory in each seat by calculating the number of times (out of the 500 posterior samples) that the Conservative estimate is greater than the largest *true* vote among all other parties.

Table 4.7: UK election prediction accuracy

Model	Brier score
Even	0.033
2:1	0.024
3:2:1	0.024

presidential election result was contrary to most predictions prior to the election, with the Republican candidate Donald Trump winning the presidency. Although the Democrats won 51% of the popular vote between the two parties, Donald Trump won the presidency by winning enough states to secure more than the 270 electoral college votes needed to win overall.[25]

Figure 4.3 presents the scatter plots showing the estimates vs the actual 2016 Republican vote share. The points show the mean estimates while the lines indicate the 90% credible intervals. The red line is the line of best fit for the estimates, while the grey dashed line is 45° line representing a perfectly linear relationship. In each scatter plot I also report the MAE, RMSE and correlation. We can see from the graphs that the model with a '3:2:1' distribution achieves highest overall accuracy, with MAE of 3.3%, RMSE of 4.2%, and correlation of 95.2%. The '2:1' distribution is the second most accurate, with MAE of 3.5% and RMSE of 4.4%, while the 'Even' sample distribution achieved MAE of 4% and RMSE of 5%. From the scatter plots we can see the 'Even' sample distribution fitted line is consistently under the 45° line, indicating that this model consistently under-predicts Republican vote share. Conversely, the two uneven sample distributions show under-prediction of Republican vote share in states where the party received low vote share. Importantly however, the models show estimates are closer to actual Republican vote share in states where the party received a greater share of the vote.

Turning to comparing accuracy between states according to marginal groups. Table

[25]This is a classic example of where MRP may be useful, as it could be used to forecast results in each small area rather than assessing electoral outcome probabilities on a national vote share.

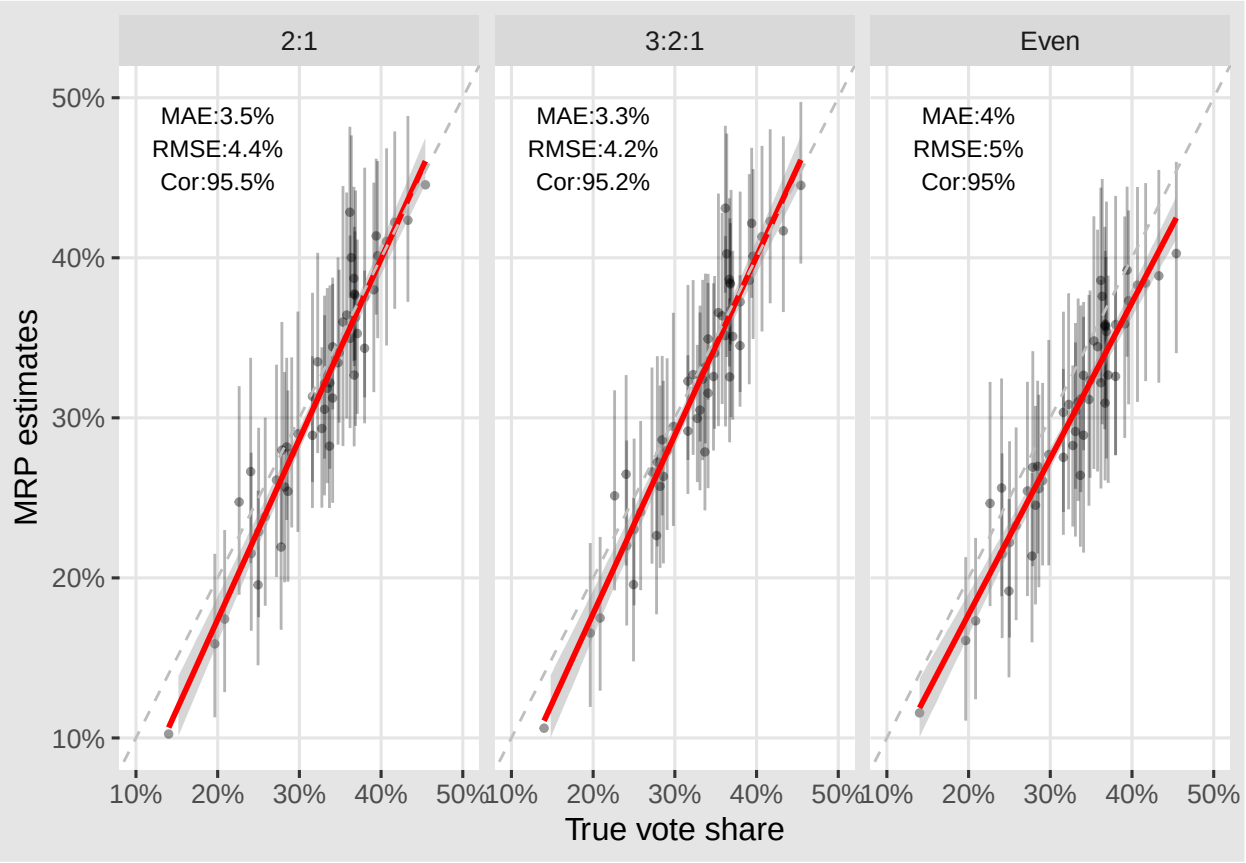

Figure 4.3: US estimates vs. true vote share

Notes: Points show mean state estimate with 90% CI shown by vertical lines. Red line shows line of best fit and dashed grey line shows 45 degree line.

Table 4.8: US: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.039	0.117
Mid marginal	0.043	0.133
Most marginal	0.040	0.130
3:2:1		
Least marginal	0.035	0.111
Mid marginal	0.036	0.122
Most marginal	0.032	0.123

Table 4.9: US: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.039	0.117
Most marginal	0.042	0.132
2:1		
Least marginal	0.033	0.104
Most marginal	0.032	0.113

4.8 shows MAE and width for 'most', 'mid' and 'least' marginal small areas, comparing 'Even' and '3:2:1' sample distribution models. We can see that the estimates for the '3:2:1' distribution are more accurate and have marginally shorter widths. At odds with results from the simulation and the UK application, here we see that the uneven sample distribution improves accuracy for all states, including those in the 'least' marginal group. This means that the '3:2:1' distribution improves estimate accuracy, even in states where the sample size is lower than if we had an evenly distributed the sample. Next, table 4.9 compares 'Even' and '2:1' distribution models, showing MAE and width for small areas separated into 'most' and 'least' marginal groups. Again, the uneven sample distribution achieves better MAE and marginally shorter widths for all small areas, including in the 'least' marginal small areas.

Finally, table 4.10 presents the brier scores for each model. The scores are identical for '3:2:1' and '2:1' sample distributions, although again the unrounded brier scores

Table 4.10: US election prediction accuracy

Model	Brier score
Even	0.074
3:2:1	0.052
2:1	0.052

show '3:2:1' as lowest. Importantly, the results once again show the uneven sample distributions improve electoral prediction, with the 'Even' sample brier score higher than both '3:2:1' and 2:1 sample distributions. Clearly, the uneven sample distribution has again improved accuracy in the small areas which are key to predicting an electoral outcome, i.e. those with a small margin of victory. And this in turn, has improved our ability to correctly predict seat winners and the overall result. The results however, contrary to the simulation and UK application, show the uneven sample distribution improves accuracy across all small areas. This means that the 'Even' sample distribution achieved poorer accuracy in small areas where the sample size was greater than was the case for both the uneven sample distributions.

The improvements in accuracy in least marginal small areas is most likely due to two reasons specific to the US application. First, the uneven sample distributions resulted in samples with an overall Republican vote share of 52-53%, compared to the 50% of the 'Even' sample distribution. Because of partial pooling, state estimates will be shrunk towards the overall vote share figure. This meant for the 'Even' distribution, estimates were shrunk towards 50%. This would have acted to the detriment of estimates in states where *true* Republican vote share is high. This can be seen in the scatter plots, which show the 'Even' model consistently under-predicting Republican vote share.

The Second, and perhaps more likely reason, is the underlying data is inaccurate. The basic assumption of an uneven sample distribution is that by increasing our sample in any given small area we are improving the probability that the raw underlying

data will be more accurate. This is probably a fair assumption, however, by no means a given. In the simulation and the UK application this has broadly been the case, but in the US example this assumption has not held. For instance, in the '3:2:1' distribution, 'least' marginal small areas had 21 respondents, 9 fewer than when the sample was distributed evenly. However, for some of the 'least' marginal small areas, the raw data is more accurate than is the case for the 'Even' distribution sample. That is, state Republican vote share is closer to *true* republican vote share. If the underlying assumption that more respondents = improved raw data accuracy is not upheld, then an uneven sample distribution will quickly become far more problematic than an 'Even' distribution. In the US application, the underlying assumption has not held, although this resulted in improved estimates for all small areas with an uneven distribution. Importantly however, this highlights that increasing sample size might not improve raw data accuracy. If this problem arises, we might make estimates far worse than would have been the case with an evenly distributed sample.

4.4 Discussion

This chapter has introduced a method to determine sample distributions and shown that this can improve MRP estimate accuracy. The chapter contributes to the wider literature on forecasting elections, and specifically forecasting elections with MRP, by demonstrating the improvements in estimate accuracy of this method. This chapter has demonstrated that by over-sampling in certain small areas, we can improve estimate accuracy in the most important small areas which in turn improve our ability to forecast elections.

One of the notable strengths of this approach is that the benefits are accompanied with *relatively* low-level risks. In the simulation study, overall estimate accuracy was largely unaffected by the sample distribution. However, the uneven sample

distributions resulted in improved estimate accuracy in the most marginal small areas. The uneven sample distribution also impacted non-marginal small areas, but for these areas it was to the detriment of estimate accuracy. In the two-applied settings, the results were more emphatic, showing the uneven sample distributions improved accuracy overall. Improvements in estimate accuracy in the most marginal small areas was not accompanied by significant decreases in accuracy in non-marginal small areas. This, I believe, is the result of the weighted ratio measure I have introduced in this chapter.[26] The weighted ratio determined sample distributions which increased marginal small area sample size significantly, but only at the cost of nominal reduction in sample size for non-marginal small areas. Altogether, the results in both simulation and real-world applications show that the benefits of this method can be significant, while the risks of poorer accuracy in non-marginal small areas seem small at worst, and non-existent at best.

In many applications of MRP, researchers use publicly available surveys where they don't have control over the sampling strategy. However, when researchers have control over the sample, this research will be of interest. This is particularly the case given that most research has finite resources and a limit on total sample size. In these circumstances, researchers should take into consideration how they wish to structure their sample so that their estimate accuracy can be targeted in the small areas which they deem most important. In an electoral setting, this manifests by over-sampling marginal small areas.

This method has demonstrated how we can get the most out of our sample for any sample size. However, the results are of particular interest to researchers with small sample sizes. The results of the simulation study show that the benefits (and

[26] In a previous version of the real-world application, I split groups evenly by past-vote share, in an identical way to the simulation study. Results were similar to those present in the chapter, but for the UK, non-marginal small area estimates had poorer accuracy. For the US, improvements in marginal small areas were not as significant as those presented above. This gives confidence that the weighted ratio is, in part at least, responsible for reducing risks involved with the uneven sample distributions. See appendix C.4 where I provide figures and tables with the previous version results.

risks) of an uneven sample distribution are most prominent for smaller sample sizes. Therefore, should a researcher only have resources that will permit a relatively small sample size, then an uneven sample distribution may be particularly beneficial.

The chapter also contributes to our wider understanding of how sample distributions can affect MRP estimate accuracy and estimate precision. Although this method will not be beneficial to all applications of MRP, the results should be of interest to all who use MRP. This is because the results highlight that the sample distribution can have significant consequences on MRP estimates. As mentioned above, most MRP applications use publicly available surveys, where samples are typically distributed unevenly in order to be representative of the wider population. This means that across small areas there will be varying degrees of accuracy and precision. This is rarely given consideration in many MRP studies, with researchers not accounting for this variation when presenting results or using the MRP estimates for further analysis.

In the chapter I have often discussed how increasing sample size will lead to improved MRP estimates because we are increasing the sample size within a small area. If this assumption does not hold true, then the estimates will certainly not benefit from improved accuracy, and rather, we will entrench inaccuracy further. This potential problem was borne out in the US application, although in this chapter this worked to the detriment of the 'Even' rather than uneven sample distributions.

One consideration that has largely been overlooked in this chapter, is how variance of opinion or behaviour within and between small areas could affect this strategy. This is a significant omission, as the degree of shrinkage through partial pooling is determined in part due to this variance. Where variance is limited, the partial pooling between small area estimates is greater. This has significant implications for the use and efficiency of an uneven sample distribution. Here I have neglected discussing variance because it is entirely beyond the control of the researcher. Nonetheless, researchers need to explore how their variable of interest varies between and within

small areas. As with multilevel modelling more widely, the uneven sample distribution should be particularly useful when there is greater variation between small areas.

An important consideration for the use of an uneven sample distribution is how we determine which small areas should receive a greater proportion of the sample. In the simulation study I used the difference between $Y = 1$ and $Y \neq 1$. However, in real-world applications, this is obviously not possible, and in the external validation I used past vote. This poses significant risks as this assumes that small areas which were previously a marginal will be a marginal in the future. In many cases there is consistency between elections, but this is not guaranteed. If previous non-marginals become marginals (and vice versa), we risk greater inaccuracy than if we had an evenly distributed sample.

In all applications in this chapter, I have assigned small areas into marginal groups based on the absolute difference between $Y = 1$ *and* $Y \neq 1$. That is, the absolute difference between vote for one party minus vote for all other parties. However, in a multiparty system, this is a problematic way to assign small areas into marginal group and risks misclassification. It would be more appropriate to estimate a multinomial model, estimating all parties at once. In this application, we would simply use the majority of the previous winning party to determine whether a small area was a marginal or not. Future work could apply the uneven sample distribution tested here to forecasting elections with an MRP multinomial model.

Similarly, future research could further advance this method by adapting the weighted ratio so that it could be applied in other electoral contests. This is necessary because in its present form, it may not be useful in all electoral systems. For instance, in the German Bundestag where voters have two votes: one for the constituency representative, and one for the proportional party make-up of parliament. The system requires accurate forecasts in each constituency to forecast the political make-up of parliament. In this application there might be electoral districts where first choice

is a marginal but the second vote is not a marginal - or vice versa. For the wider application of the weighted ratio, researchers will need to tailor it to account for the idiosyncrasies of different electoral systems such as the German Bundestag.

4.5 Conclusion

This research has tested whether an uneven sample distribution can improve MRP estimate accuracy. The sampling strategy proposed here will not be useful in many - if not most - applications of MRP. But in applications where we are more concerned with accuracy in certain small areas, this chapter has demonstrated this strategy can be useful. The most obvious application for this strategy - and the application which has been applied here - is to forecast elections, where typically a minority of small areas decide electoral outcomes.

Through a simulation study and two external validation exercises, this chapter sets out how this method can be applied, as well as demonstrating the benefits. In both the simulation study and the two applied cases in the UK and US, the research showed this strategy can improve estimate accuracy in small areas that we deem more important. In the two applied cases, the uneven sample distribution improved estimate accuracy in the most important small areas by just under 1%. In turn, these improvements translated into an enhanced ability to forecast the elections.

Chapter 5

MRP and informative priors

Multilevel regression and poststratification (MRP) is a small area estimation method that has grown in use and popularity in academic and non-academic settings. In academia, the use of the method has grown since its first introduction (See Gelman and Little, 1997), with notable applications in social sciences and population health studies. The use in social sciences extends across a range of behaviour and public opinion topics, as was documented in chapter 2. Outside of academia, the notoriety of the method and its application is largely restricted to forecasting elections, which it has been shown to be proficient at (see Lauderdale et al., 2020).

Part of the reason MRP is adept at estimating small area opinion and behaviour is because it enables researchers to combine numerous sources of information into one unified framework. For instance, we use individual-level data from surveys, incorporate information about the small areas as area-level variables, and structural information about the population through the poststratification frame. With MRP increasingly estimated as a Bayesian model, we have the opportunity to incorporate further information through the specification of informative priors.

These are an important - albeit sometimes seen as a controversial - aspect of Bayesian estimation. And although MRP is increasingly Bayesian, the literature

rarely discusses the specification of priors. Indeed, in the literature on MRP, there are only a few papers which directly address priors (see Downes et al., 2018; Gao et al., 2021). This omission might be because most see the inclusion of informative priors as unnecessary, given that we already include prior information through the use of area-level variables and the poststratification frame (Gelman, 2009). However, for electoral forecasting there is sufficient reason to believe that incorporating knowledge from previous elections could be beneficial to estimating vote share.

In representative democracies, from election-to-election there are rarely substantial shifts in the voter-type for a given political party. Variables that are predictive of vote share in one election will be predictive in the next election, and the magnitude of parameter effects will most likely be consistent. This means we can take information from the previous election and use this to help the model for the current electoral forecast. This may improve estimate accuracy, enhance the precision of estimates and improve estimation efficiency. There are, however, potential risks involved with such a modeling strategy. Perhaps the greatest risk is if the current election differs significantly from a previous election, we may do more harm than good by incorporating past information into the model.

This chapter sets out to explore how we can directly incorporate past information into current electoral forecasting models, and importantly, if these improve small area estimate accuracy. Specifically, in a series of tests I explore whether a two-stage prior elicitation strategy could be useful for MRP electoral forecasting.[1] I test the method at elections in both the US (2008 and 2012) and the UK (2017 and 2019). Broadly, the results show that informative priors, as operationalised in this study, only improve estimate accuracy for the smallest sample size. Furthermore, improvements are dependent on the election and the value of lambda used. However, informative priors do seem to improve computational efficiency, as measured by model run-time,

[1] I formally specify the models in the theory section.

and show signs that they could improve inference for sub-groups in the population.

5.1 Background

Priors

In Bayesian analysis priors are used to incorporate information that we have before we conduct any formal research or analysis (Wesel et al., 2011). They provide researchers with a means to systematically incorporate current knowledge into the model. More specifically, priors are the expected distribution of any given parameter effect or coefficient. The prior affects the posterior (i.e. the estimates) through Bayes' rule, which states:

$$P(\theta|data) = \frac{P(data|\theta) \cdot P(\theta)}{P(data)} \tag{5.1}$$

Here $P(\theta|data)$ is the posterior, $P(data|\theta)$ the likelihood , $P(\theta)$ the prior, and $P(data)$ is the probability of the data given the likelihood and prior. Thus, the prior has a direct impact on the posterior through its interaction with the likelihood. The resulting posterior distribution is a combination of the two components.[2] In practice, the impact of the prior on the posterior is by giving the model a distribution which acts as a search space for a given parameter.

Prior distributions can be of any form (for example normal, student-t, or uniform), and can range in specificity or informativeness (Schoot et al., 2021). However, because priors can have a significant impact on the posterior, and because they are not directly informed from the data, they are often viewed as the most controversial feature in Bayesian analysis. This has led to debate about how informative a prior should be. Broadly, we can categorise them into three distinct groups, non-informative, weakly-informative and informative priors (Depaoli and Schoot, 2017).

[2]In the estimation of parameters, the resulting posterior might not cross-over with the prior, if the likelihood and data are strong enough.

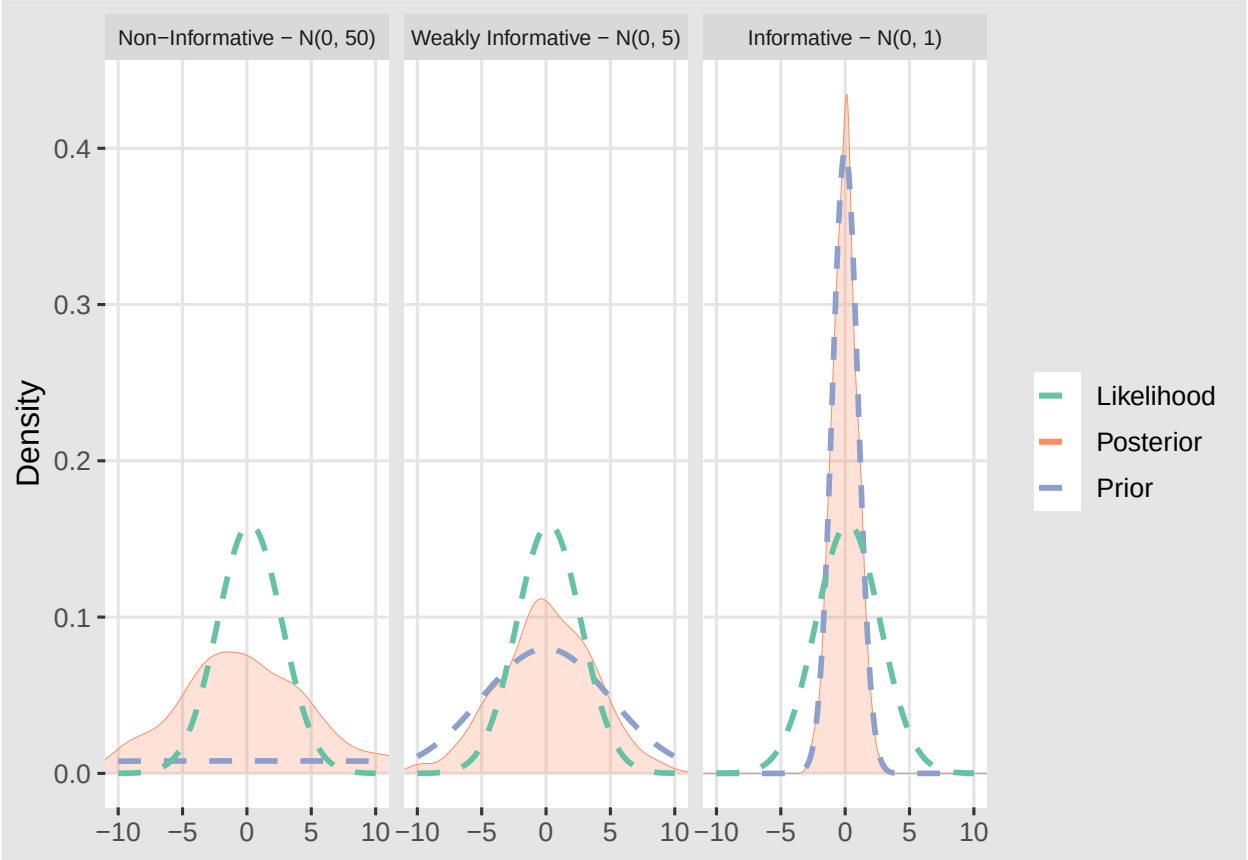

Figure 5.1: Impact of the prior
Notes: showing example plot of likelihood and posterior for Non-Informative, Weakly-Informative, and Informative priors.

In figure 5.1 I visualise how these three types of priors can affect the posterior.[3] In each plot the green dashed line is the likelihood (which remains identical in each plot), the blue dashed line is the prior, and the orange shape shows the resulting posterior distribution. As is evident in the figure - and discussed more specifically below - prior choices can have a strong and significant impact on the posterior.

Non-informative priors

Non-informative (or flat) priors are implemented with the goal of providing as little information as possible. They do not attempt to restrict the parameter estimation,

[3]Plots originally inspired by those in (Depaoli and Schoot, 2017). To produce plots, I adapt code from: `https://gist.github.com/wjakethompson/1189514071478a2ca59491f43f21afec`

and allow the data to speak entirely for itself. In the left-hand column of figure 5.1 we can see the non-informative prior is mostly flat, shown by the dashed blue line across the plot. This means that all parameter values are given equal probability of sampling for the posterior. The resulting posterior distribution is similar to the likelihood, with a peak close to around 0. However, while the data is 'allowed to speak for itself', the flat prior results in a posterior with more density in the tails of the distribution than the likelihood.

The use of non-informative priors are advocated by some, who argue that informative priors amounts to 'data falsification' (García-Pérez, 2019). This argument rests on the idea that informative priors manipulate the estimation of data, by forcing it to conform to paradigms that we set, rather than ones identified in the data. However, suggestions that priors are subjective overlook the fact that many decisions in the analysis and modeling of data is subjective, and prior specifications are no different (Gelman and Hennig, 2015). Furthermore, the 'flat' nature of non-informative prior distributions can have significant effects on the resulting posterior and this could be considered analogous to the effect of informative priors (Lemoine, 2019). Indeed, research has demonstrated that non-informative priors can act to the detriment of estimation and harm the posterior (Lenk and Orme, 2009).

Weakly-informative priors

An alternative to non-informative priors are those known as weakly-informative priors. These are more restrictive than non-informative priors, and limit the search space for parameter estimation. The priors, while not enforcing a strict distribution on the posterior, rule out impossible or highly unlikely parameter values. Weakly-informative priors are thus argued to offer a good balance between informative and non-informative priors, by ruling out unlikely outcomes and encouraging some shrinkage (Simpson et al., 2015). In figure 5.1 the middle plot shows the posterior of a weakly-informative

prior. Unlike the non-informative prior, the prior here is a normal distribution with a much smaller standard deviation. The distribution is still centered around 0, but the smaller standard deviation means there is less probability given to values further from 0. The resulting posterior distribution has a higher peak and less density in the tails and mirrors the likelihood to a greater extent than the non-informative posterior. However, the weakly-informative posterior distribution is still flatter than the likelihood with greater density in the tails.

Previous research has demonstrated that weakly-informative priors are an improvement on non-informative priors. They are argued to stabilise estimates, encoding a good degree of regularisation while still being vague enough to be used in a wide variety of contexts (Gelman et al., 2008). The benefit of weakly-informative priors is that when the data is strong, the prior has limited impact on the posterior. However, when data is weak, the weakly-informative prior will have a much greater influence on the estimation of the posterior. Some argue that weakly-informative priors can work well as a default prior, but because they do not contribute any domain specific knowledge, the goal for researchers should still be to use informative priors (Gelman et al., 2017).

Informative priors

Informative priors are where we encode strict numerical values to structure the posterior (Depaoli and Schoot, 2017). This means that the prior has a much stronger impact on the posterior, providing the model with a specific parameter search space. Such priors are a means to directly incorporate theoretical, logical, and empirical knowledge we already have (Lee and Vanpaemel, 2018). If we have such valuable information, we should incorporate it into our model, and this is no different to a variety of decisions researchers make based on past evidence (Golchi, 2019). In figure 5.1 the right-hand column shows an informative prior and the resulting posterior distribution. As can be

seen in the figure, relative to both the non and weakly-informative, the informative prior distribution is much narrower with a high peak. The prior distribution means that we place most probability within a narrow range and have a high degree of certainty about the parameter. The resulting posterior is similarly narrow, placing nearly all density in a short parameter range.

Importantly, studies have shown that when compared to non-informative priors, informative priors can improve estimation and computational efficiency (Golchi, 2019; Grzenda, 2016), and can aid with model identifiability and reduce model complexity (Lee and Vanpaemel, 2018). These improvements in estimation are often particularly noteworthy for smaller parameter effects (Zondervan-Zwijnenburg et al., 2017; Jaynes, 1985).

By ensuring that the posterior distribution is in line with our domain knowledge, informative priors can significantly aid estimation of parameters. However, some argue the risks of informative priors outweigh the potential benefits, and therefore, researchers should use weakly-informative priors instead (S. D. Team, 2020). One of the main risks associated with informative prior is that when there is a discrepancy between the prior and the likelihood, an informative prior can shift the posterior away from the likelihood (Schoot et al., 2021). This could mean that our posterior does not capture the likelihood well, and our parameter estimation is poor. In these instances, as well as poor parameter estimation, the statistical and computational benefits of informative priors are typically no longer realised.

Determining informative priors

Informative priors are often criticised for being subjective in nature. Advocates of this view contend that the process of defining priors is simply the researcher arbitrarily choosing probability distributions. However, this is often not the case, but rather an objective scientific process whereby previous knowledge is formalised into a numerical

format for the prior distribution. Skepticism towards informative priors might partly be because there is an absence of well-defined methods for specifying them. This could also be one of the main reasons why non-informative priors are often the default choice (Lee and Vanpaemel, 2018). Recognition of both of these is not new, indeed, there have been calls for more work on how to define priors for many years (see Jaynes, 1985).

More recently, these calls seem to have been heard with the growth in research on prior elicitation techniques. Prior elicitation methods involve systematically gathering information and subsequently transforming the information into a probability distribution (Depaoli and Schoot, 2017). For example, gathering expert opinion, knowledge, and judgment to define probability distributions is one growing field (O'Hagan, 2019). For a detailed review see O'Hagan (2006). Another method is the use of past studies to determine a prior distribution. This method uses past theory or empirical results, which are sourced from one or more studies and sometimes in the format of a meta-analysis (Zondervan-Zwijnenburg et al., 2017; Lee and Vanpaemel, 2018). Alternatively, researchers could carry out a pilot study, with the results used to investigate priors for the main study (for a worked example of this method, see Gelman et al., 1996). A similar approach includes splitting the data and using a sub-set for a training model. The training model posterior is in turn used to determine priors for the full model (Wesel et al., 2011).

A variant of splitting the data is the use of historic data to determine priors for the current model. The 'two-stage' solution uses the posterior of a model with historic data to define a prior distribution for the current model. This method of prior elicitation has shown promising results in the work of Yu and Abdel-Aty (2013). Their study found the 'two-stage' solution produced the best results when compared with non-informative priors alongside other prior elicitation methods (Yu and Abdel-Aty, 2013). Along similar lines, Chen and Ibrahim (2000) have developed the 'power-prior'

which includes both historic and current data in a single model, specifying coefficient priors from the joint data. The method is particularly effective because the power-prior is scaled by a parameter $a0$, which controls for the similarity between the current and historic data (for a detailed explanation, see Chen and Ibrahim, 2000; Ibrahim et al., 2015). Others, rather than focus on direct prior elicitation, emphasise the importance of prior predictive checks to test the sensitivity of priors and that they comply with domain knowledge (Gabry et al., 2019).

MRP and priors

As identified in the systematic review, the majority of MRP applications to date are not Bayesian and therefore do not incorporate or discuss priors.[4] While it is difficult to ascertain the point at which MRP became a fully Bayesian methodology, Ghitza and Gelman (2013) recognised that while their work was not yet Bayesian, they hoped future analysis would be estimated in a Bayesian probabilistic program. This seems to have been realised with more recent MRP work being estimated as Bayesian models.

However, for Bayesian MRP models there is still limited discussion on prior specification. Among Bayesian MRP models, around half do not state what priors they use in their model. The remaining half who provide details of prior specification, use either non-informative or weakly-informative priors. Overall, the use and discussion of priors in MRP applications is somewhat limited.

In the broader MRP literature, priors have been discussed by Downes et al. (2018), whose study investigated the performance of MRP to estimate health outcomes. The study also assessed the impact of three different priors (non-informative uniform, bounded uniform and weakly-informative normal). They concluded that priors had little impact on the posterior and estimates. Although they noted the more informative priors produced more precise parameter posteriors, and this was especially the case

[4] Or at least the majority do not explicitly state that they estimated a Bayesian model.

for variables with few categories (Downes et al., 2018: 1789).

More recently, the innovative work of Gao et al. (2021) demonstrates how priors could be deployed to improve MRP estimates. Their work introduced structured priors, a method which can improve MRP estimates and significantly enhance sub-group inference. Structured priors take advantage of the structures within the population to improve estimation. The standard MRP specification benefits from partial pooling, where group parameters are shrunk towards the global mean. This is particularly useful for groups with small samples and works to stabilise estimates by reducing variance. Structured priors, rather than shrinking to the global mean, shrink group estimates to the mean of groups closer and more similar to them. For example, probability of voting for a *left-wing* party often decreases with age. With structured priors, the younger age categories would be shrunk towards age categories above and below, and vice-versa for older groups. This means that estimation takes into account the known directional structure of the relationship between age and voting behaviour. The work demonstrates how using priors can significantly enhance prediction accuracy and sub-group inference. Importantly, it shows how priors can act as an additional source of information that can aid MRP estimates.

5.2 Theory

This chapter seeks to explore how informative priors can be incorporated into MRP electoral forecasting and whether such a strategy improves small area estimate accuracy. To do this, the chapter will explore forecasting multiple elections in the UK and US using a standard MRP format with informative priors and an alternative MRP format with informative priors. I explain model specification below using the UK case.[5]

[5]For explanation here I use the UK model as an example. In appendix D.1 and D.2, I show notation for the US model.

5.2.1 MRP with weakly-informative priors

In the classic case, MRP is estimated as a multilevel logistic regression model, with individual and area-level variables. The model can be written as follows:

$$Pr(Y_i = 1) = logit^{-1}(\beta^{\theta} + \beta^{Female}_{Female_i} + a^{Area}_{j[i]} + a^{Age}_{k[i]} + a^{Education}_{l[i]} + a^{Region}_{m[i]}), \;\; for\; i = 1, ..., n. \tag{5.2}$$

Where a^{Area}_j, a^{Age}_k, $a^{Education}_l$ and a^{Region}_m are varying intercept terms and β^{Female} is a fixed effect for female gender. We assume the varying intercept terms are drawn from a normal distribution with a mean 0 and some variance, which is itself a modelled parameter.

$$\begin{aligned} a^{Age}_k &\sim N(0, (\sigma^{Age})^2) \;\; for\; k = 1, ..., 8 \\ a^{Education}_l &\sim N(0, (\sigma^{Education})^2) \;\; for\; l = 1, ..., 6 \\ a^{Region}_m &\sim N(0, (\sigma^{Region})^2) \;\; for\; m = 1, ..., 11 \end{aligned} \tag{5.3}$$

The area term is modeled as a function of region, Labour vote share at previous election (lab), percent of constituency which is classed as long-term unemployed (unem), population density (dens), percentage of constituency which work in industry and manufacturing (ind) and EU referendum constituency leave vote share (leave).

$$\begin{aligned} a^{Area}_j \sim N(a^{Region}_{m[j]} + \beta^{lab} \cdot \; lab \; + \beta^{unem} \cdot \; unem \; + \beta^{dens} \cdot \; dens \; + \beta^{ind} \cdot \; ind \; + \\ \beta^{leave} \cdot \; leave, (\sigma^{Area})^2), \;\; for\; j = 1, ..., 632 \end{aligned} \tag{5.4}$$

We specify priors for the variance σ^2 and β parameters. For weakly-informative priors, this might be in the format where, variance is given a student-t prior with scale 5, mean 0 and standard deviation 5. The intercept and all β terms are given normal

priors with a mean 0 and standard deviation of 5.

$$\begin{aligned} \sigma^{Area},\ \sigma^{Region},\ \sigma^{Age},\ \sigma^{Education} &\sim Stt(5,\ 0,\ 5) \\ \beta^{\theta},\ \beta^{female},\ \beta^{pastvote},\ \beta^{unemployed},\ \beta^{density},\ \beta^{industry},\ \beta^{leave} &\sim N(0,\ 5) \end{aligned} \tag{5.5}$$

These can be considered weakly-informative as they attempt to 'rule out' effects greater than +/- 5 standard deviations from 0. Effects outside this range are highly unlikely, but the prior distribution is still vague enough to allow the 'data to speak for itself' and the likelihood to strongly influence the posterior.

5.2.2 MRP with informative priors

Informative priors enter the above model by replacing the weakly-informative priors, defined in equation 5.5. Informative priors are derived from the historic model posterior $P_H(\theta_H|X_H)$, where H denotes the historic model. I calculate the historic median ($\tilde{X}_H$) and historic standard deviation (σ_H) of the distribution of each given parameter posterior. The $\tilde{X}_H$ and σ_H are then directly imputed as priors for the current election MRP model. For example, take the a_k^{Age} term for varying age intercepts, the prior specification would be as follows:

$$\begin{aligned} a_k^{Age} &\sim N(0,\ (\sigma^{Age})^2)\ for\ m = 1, ..., 8, \\ \sigma^{Age} &\sim Stt(5,\ \tilde{X}_H,\ \sigma_H) \end{aligned} \tag{5.6}$$

As before, we assume the a_k^{Age} term is drawn from a normal distribution, with mean 0 and some standard deviation. The standard deviation is still given a student-t distribution prior with a scale of 5, but now, rather than a mean of 0 and standard deviation of 5, we give the prior $\tilde{X}_H$ and σ_H which refer to the historic parameter posterior median and standard deviation, respectively. This process is replicated for all parameters, with β terms given a normal distribution and $\tilde{X}_H$ and σ_H.

To account for differences between the historic and the current election, I multiply the standard deviation by a scaling constant, lambda (λ). For λ values > 1, this acts to widen the prior distribution, which should account for potential small shifts in a parameter effect. If for example, from one election to the next age categories have a different effect on vote choice, the wider prior distribution should better account for this shift. The intention of this can be seen as somewhat comparable to the intention of the power prior scaling constant parameter a_0 (Chen and Ibrahim, 2000; Ibrahim et al., 2015). That is, an attempt to account for variation in the similarity between historic and current data.

By providing the MRP model with informative priors, the benefits could be three-fold. First, we may improve estimate accuracy. If we have good knowledge of a given parameter effect, specifying informative priors should influence the parameter posterior towards the *true* parameter value. This in turn should ensure greater estimate accuracy. Second, we may improve the degree of certainty in the parameter estimate. That is, we may shorten the credible interval widths of the parameters and the final estimates. With informative priors, we place most probability in a narrow range of parameter values. The resulting posterior should similarly be narrow with limited density in the tails of the distribution. Third, informative priors may improve computational efficiency in model estimation.

However, as noted previously, the risks of informative priors have been suggested to outweigh the potential benefits (S. D. Team, 2020). The potential risk is that if we specify a prior that does not align with the data and the likelihood, the posterior estimation can be problematic. This means if there is limited overlap of prior and likelihood distributions, we risk shifting the posterior away from the likelihood (Schoot et al., 2021). In the application here, this problem could arise if the historic election is dissimilar to the current election, or the historic data and model are inaccurate. If either of these two manifest, our prior will not align with the likelihood and data, and

therefore, we may be providing the current model with priors that are not close to the *true* parameter distribution.[6]

5.2.3 Alternative MRP specification with informative priors

The method described above incorporates information from past elections into the current election forecast. It is possible to extend this approach to incorporate this information in a more direct and stricter fashion. To do this, each category of each variable is estimated as a separate β term (or fixed effect), rather than a group of varying intercept terms. Each β term for each variable category is then given its own specific prior, whereas in the standard specification, we give a prior for the distribution from which parameter effects are drawn from.

This approach is not possible for all varying intercept terms, but for age education and region, where the number of categories is relatively few, there is a large enough sample size to directly estimate each category effect. For the area varying intercept terms, this would not be possible because for each area we would only have sample sizes between 5-30 respondents.[7] In the standard MRP notation, a_k^{Age}, $a_l^{Education}$, a_m^{Region} are all varying intercept terms, where each category parameter is drawn from a normal distribution, which we assume to be 0-centred with some variance:

$$\begin{aligned} a_k^{Age} &\sim N(0, (\sigma^{Age})^2) \ for \ m = 1, ..., 8 \\ a_l^{Education} &\sim N(0, (\sigma^{Education})^2) \ for \ m = 1, ..., 6 \\ a_m^{Region} &\sim N(0, (\sigma^{Region})^2) \ for \ m = 1, ..., 11 \end{aligned} \tag{5.7}$$

In the alternative MRP specification, area and gender terms remain the same as in equation 5.2. However, the terms for age, education and region are changed so that

[6] We can of course check that the historic model correctly predicts past vote. However, accurate small area estimates does not guarantee that parameters are estimated well or correctly.

[7] This range in sample size per small area is indicative of the sample sizes used in this chapter. With larger sample sizes, we might have sufficient number of respondents per small area to reliably estimate small area parameters as β terms.

each category of these variables is specified as an independent β terms. The model would thus be written as follows:

$$Pr(Y_i = 1) = logit^{-1}(\beta_\theta \ + \ a_{j[i]}^{Areas} \ + \ \beta^{Female} \ \cdot \ Female \ + \ \beta^{Age} \ \cdot \ \boldsymbol{X}^{Age} \ + \\ \beta^{Education} \cdot \ \boldsymbol{X}^{Education} \ + \ \beta^{Region} \ \cdot \ \boldsymbol{X}^{Region}), \ for \ i = 1, ..., n. \tag{5.8}$$

Where β^{Age}, $\beta^{Education}$, β^{Region} are vectors of β coefficients of P categories for the variables age, education and region. $\boldsymbol{X}^{Age}, \boldsymbol{X}^{Education}$, $\boldsymbol{X}^{Region}$ are matrices of 0 and 1s, with each row indexing the ith respondent from $i = 1..., n$ and columns for each category of the given variable, where there are p categories, from $1..., P$.

For prior specification, I use an identical historic model to that used for the standard MRP with informative priors. However we now take different information from the historic model. For the standard MRP format, we use $\tilde{X}_H$ and σ_H from the varying intercepts variance parameter. For the alternative specification, I take the $\tilde{X}_H$ and σ_H for each variable category intercept. These are then imputed as the priors, such that each β parameter has a prior that is a normal distribution with mean and standard deviation from the historic posterior, $\beta \ \sim \ N(\tilde{X}_H, \ \sigma_H)$.

We can consider this method a more direct and strict incorporation of past historic information than the standard MRP format. In the standard format, we provide the model with a distribution from which a parameter is drawn from. This method, however, estimates each variable-category separately and provides specific numerical values for the expected distribution of each variable-category. This means that we bring forward more specific information (i.e. the exact effect of each category). And incorporate this into the current model in a more specific way (i.e. specifying the exact expected distribution of a given variable category effect).

This specification imposes stricter restrictions on the model parameters. This could benefit the model, as the stricter parameters are informed by the historic election model. However, this could prove problematic if there are differences between

elections, or the historic model is not good at capturing the *true* relationship between our predictor variables and voting behaviour. If either of these problems manifest, then our forecasts will be inaccurate, and inaccurate to a greater degree than the standard MRP format. This method also loses the benefit of partial pooling for all parameters which were previously estimated as varying intercept terms. Partial pooling is credited as one of the key reasons MRP is adept at forecasting sub-national opinion, and removing this could act to the detriment of MRP estimates. Partial pooling is useful because it enables the 'borrowing of strength' which means small N group parameters have less variance. However, the narrow distribution of informative priors is known to alleviate problems with variance, which might arise when estimating parameters for groups with smaller sample sizes.

5.3 Data and methods

This study seeks to test whether informative priors in two-different MRP specifications can improve small area estimate accuracy. This will be achieved by comparing these two specifications against a standard MRP model with weakly-informative priors (the baseline). The study will forecast numerous elections in both the UK and US, comparing primarily estimate accuracy, but also estimate widths, parameter coefficients, and computational efficiency. In the UK I estimate Labour vote share in constituencies for the 2017 and 2019 parliamentary elections, and in the US I estimate state-level Democrat vote share in the 2012 and 2016 presidential elections.

Data

For MRP we use both individual and area-level variables. The former captures characteristics of the respondent, while the latter captures characteristics of the small area. Individual-level data comes in the format of surveys where respondents provide

their vote choice, demographic characteristics, and the small area which the respondent resides. In the UK, I make use of the publicly available British Election Study (BES), which carries out large online-sample surveys at regular intervals, as well as before and after elections. I make use of the campaign waves (5, 12, 18), with interviews conducted throughout the month immediately prior to the election. In the US case, I use the American National Election Studies (ANES) time-series study survey data. These surveys have been carried out pre and post US presidential elections since 1948. The samples are a mixture of online and face-to-face random probability samples.

Area-level variables are also included in both the UK and US case. These are all publicly available at either UK constituency or US State-level. For the UK, I access the necessary area-level variables through the BES 'Linked data', which provides previous election results and census data at the constituency level. For the US, the only area-level variable I use is past vote, obtained through the MIT Election Data and Science Lab dataframe, 'US President 1976-2020' which provides historic voting records for US states.

Variable selection in this study is directed by the Lauderdale et al. (2020) paper on forecasting elections with MRP. The study discusses MRP electoral forecasting for numerous elections including UK 2017 and US 2016. Lauderdale et al. (2020) demonstrate the ability of MRP to forecast elections with MRP. Where possible, I replicate variable selection used in their models. The individual and area-level variables used in this study are reported in table 5.1.

Past vote in table 5.1, refers to the past Labour or Democratic vote share in the preceding election. In the UK case, as the 2016 EU referendum was after the 2015 election, I do not include 'Leave vote' in the 2015 historic model, but include in the 2017 and 2019 election models. Lauderdale et al. (2020) also use a variety of individual-level political variables, including past vote and political attenuation. To use these, researchers need to construct poststratification frames that include

Table 5.1: Individual and area-level variables

Individual-level	Area-level
UK	
Gender (2)	Past Lab vote
Age (8)	EU2016 Leave vote
Education (6)	% Long-term unemployed
Campaign week (4)	% Industry manufacturing
	Population density
	Region
US	
Gender (2)	Region
Age (4)	Past Dem vote
Ethnicity (4)	
Education (4)	
Marital status (4)	
Campaign week (4)	

Note:
Number of categories for each variable in brackets
'EU2016 Leave vote' is only included in the 2017 and 2019 models

joint-distributions of these variables.[8] For this study, this is an added complication I chose to avoid and only use demographic individual-level variables that are available in the poststratification frames I use. I also do not include interactions that are used in the Lauderdale et al. (2020) model. The introduction would increase computational demand, and for the purposes here, I think the more limited set of variables will still enable the study to sufficiently address the research question(s).

Poststratification frame

In the US case, I construct a poststratification frame with American Community Survey (ACS). The ACS regularly conducts large N sample surveys which provide a data source for reliably estimating the proportion of each demographic sub-group

[8]This could be achieved by a raking procedure (Hanretty et al., 2016), synthetically constructing the joint-distributions (Leemann and Wasserfallen, 2017) or through imputation (Cerina and Duch, 2020b)

in each US State. I use the 5-year Public Use Microdata Samples (PUMS) of the ACS, for the 2012 I use 2008-2012 file and for 2016 I use the 2012-2016 file. In the UK, data limitations mean that to construct a poststratification frame researchers need to use alternative methods (See Hanretty et al., 2016). To avoid this, I make use of a publicly available poststratification frame for all UK adults.[9] The frame was constructed using 2011 census data, which could pose problems for estimate accuracy, but should be sufficient to produce estimates for a range of models with the goal of assessing differences in accuracy.

Turnout

Turnout is notoriously hard to estimate, owing in part to over-reporting of turnout by respondents in surveys. Poor turnout estimates at state or constituency level have previously been noted as a common source of error for MRP estimates (Lauderdale et al., 2020). In order to limit this potential, I apply turnout to estimates by using actual Constituency or State level turnout from the election being forecast. This means the final estimates do not take into account differential turnout among demographic sub-groups, which could be problematic for estimate accuracy. However, I believe the problems that could arise from this are less problematic than potential error introduced by an inaccurate turnout measure, and the subsequent need to disentangle effects of informative priors from error associated with a turnout measure.[10]

Samples

In this study I test four different sample sizes (5, 10, 20 and 30 respondents per small area). To generate these, I sample respondents from the full survey sample. Because

[9]Poststratification frame can be accessed: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IPPPNU`

[10]To give confidence results presented in this chapter are not a function of using actual turnout, I also produced results with a model-based turnout measure. Details of the turnout measure and some of the results with the model-based turnout are presented in appendix D.4.

respondents are not evenly distributed among small areas in the BES and ANES, I was not able to generate samples where respondents are evenly spread. Instead, for each sample size above, I calculate what the total sample size should be, i.e. respondents per small area * number of small areas. I then sample respondents ensuring that the proportion of respondents from each small area is relative to the proportion in the full sample. I ensure that every small area has at least 1 respondent in the sample. This means the sample sizes I test, are an average of 5, 10, 20 and 30 respondents per small area. I use this sampling procedure once, to create a single dataset for each sample size for each election.

Modelling strategy

For each election the procedure of estimating vote share with informative priors involves first a model for the historic election (the preceding election) and then a model to forecast the current election. The elections to be forecast (including historic model) are as follows:

- 2017 UK election (2015 UK election)
- 2019 UK election (2017 UK election)
- 2012 US election (2008 US election)
- 2016 US election (2012 US election)

Historic model

The first stage of forecasting each election is estimating a historic model. In each case, I estimate the historic model in the standard MRP format (shown in equation 5.2). The historic model uses the full survey sample for each election, whereas the informative prior models use sub-samples identified above. I estimate the models with the weakly-informative priors identified in the theory section, that is, student-t priors, with a scale of 5, mean of 0 and standard deviation of 5 on the random-intercept

standard deviations, and normal priors with mean 0 and standard deviation of 5 on the intercept and β parameters

Specifying informative priors

To specify the informative priors for the current election, I follow the two-stage prior elicitation method. I take the historic parameter posterior distribution and calculate the median and the standard deviation of the distribution. I apply a scaling value (λ) to account for differences between the historic and current election. In practice, this is applied by multiplying the historic standard deviation σ_H or beta β_H by λ. I test three λ values here, 1, 1.5, and 2. For the alternative specification, because the demographic and regional terms are now β parameters, one category is specified as the reference for all other variable categories. To account for this in the priors, I add the reference category $\tilde{X}_H$ to the $\tilde{X}_H$ for all other categories.

Modelling estimation

To forecast vote choice I estimate Bayesian multilevel logistic regression models, where vote choice for either Labour or the Democrat Party = 1, and vote choice for other parties in the UK or Republicans in the US = 0. For the UK I exclude those who would not vote and those who 'Don't know', and for the US I exclude those who would not vote for either Democrat or Republican. In both the UK and US applications, the variables in both elections remain consistent, as identified above. However, how the models are estimated varies. With the standard application I estimate demographic and the region variable as random intercept parameters. With the alternative specification, each category of each variable is specified as a separate β parameter. Models were estimated with Rstan (S. D. Team, 2020), and called through the brms package (Bürkner, 2017). Each model had 2 chains, each with 1000 iterations (500 warm-up and 500 sampling).

The multilevel model estimates were poststratified to produce estimates for Labour or Democrat vote share in UK constituencies and US States respectively. To post-stratify I drew 500 samples from the posterior using the poststratification frame as new data. I apply turnout to each row of the poststratification frame, such that each estimated probability for voting for either Labour or Democrat have been adjusted for the associated constituency or state turnout. Because the poststratification frames I use are for total adult population, the final Labour and Democrat small area estimates are as a percentage of the total adult population, rather than the eligible voting population.

5.4 Results

For this study, accuracy refers to accuracy of estimates of Labour vote share in UK constituencies or Democratic vote share in US states. To measure this, I use mean absolute error (MAE). That is, the absolute average error of predicted vote share across all small areas. This is calculated first across iterations and next across small areas. When presenting results below, I show changes in MAE from the baseline model rather than show MAE of each model. In each case, the baseline is the standard MRP specification with weakly-informative priors, and unless otherwise stated, the baseline has an identical sample size to the model it is being compared with. A decrease from the baseline represents an improvement in accuracy as MAE has decreased, while an increase in MAE represents a worse model.

5.4.1 Alternative MRP with informative priors

I first present the results of the alternative MRP specification, where all variables except the small area, are modelled as separate β parameters. Table 5.2 shows the change from the baseline model(s) for the two UK elections, 2017 and 2019. In each

column are the results for each of the four sample sizes (average of 5, 10, 20 and 30 respondents per small area). In the rows of the table are the different λ values 1, 1.5 and 2, which in turn are grouped by election year, 2017 and 2019. In table 5.2, we can see that overall the alternative MRP specification has a small impact on estimate accuracy. With the exception of one model, for both elections, all sample sizes, and all λ values, differences from the baseline model(s) range between 0-0.3%. This means that nearly all model estimates were either the same or worse than the baseline model. The one case of improvement was for the 2019 election, a sample of 5, and

Table 5.2: Alternative MRP accuracy (UK)

	Sample			
Model	5	10	20	30
2017				
Lambda = 1	0.1%	0.2%	0.3%	0.3%
Lambda = 1.5	0%	0.1%	0.2%	0.2%
Lambda = 2	0%	0.1%	0.2%	0.2%
2019				
Lambda = 1	-0.2%	0.2%	0%	0.1%
Lambda = 1.5	0%	0.2%	0.1%	0.1%
Lambda = 2	0.1%	0.2%	0%	0%

Note: showing inf. prior MAE as +/- from baseline.

where $\lambda = 1$. The tables shows this model decreased MAE by 0.2%, which represents an improvement from the baseline model. Clearly, in this particular example, the alternative specification with informative priors works to the detriment of accuracy, albeit, to a small degree.

Turning to table 5.3, the table shows the results for the alternative MRP specification for the US results. For the US, the differences between baseline and alternative specification are much starker. In 2012, the alternative MRP specification significantly improves MRP estimates for the smallest sample size. For each λ value, MAE has decreased by 1.9%, 1.5% and 1.3%, respectively. However, as sample size increases, the

differences in MAE are largely nonexistent. When $\lambda = 1$ the models show a consistent improvement in MAE, but MAE only decreases between 0.1-0.2%. For larger sample sizes, and when $\lambda = 1.5$ *or* 2, MAE is the same or marginally worse (ranging from a 0-0.3% increase). For 2016, the results show the alternative specification is consistently

Table 5.3: Alternative MRP accuracy (US)

	Sample			
Model	5	10	20	30
2012				
Lambda = 1	-1.9%	-0.2%	-0.1%	-0.1%
Lambda = 1.5	-1.5%	0%	0.1%	0%
Lambda = 2	-1.3%	0.1%	0.3%	0%
2016				
Lambda = 1	0.4%	0.6%	1.5%	1.8%
Lambda = 1.5	0.3%	0.4%	1.2%	1.5%
Lambda = 2	0.3%	0.6%	1%	1%

Note: showing inf. prior MAE as +/- from baseline.

worse. For the smallest sample size, each model has MAE that represents an increase from the baseline, ranging from 0.3-0.4%. As sample size increases, the difference in MAE from the baseline also increases. For the largest sample size, the alternative specification with informative priors increases MAE between 1-1.8%.

As well as sample size having an impact, the results highlight how the different λ values affect the estimates differently. In 2012, the lowest λ value improves accuracy consistently and to the greatest extent. Whereas in 2016, the lowest λ value worsens accuracy to the greatest degree. However, sample size and λ values do not account for all differences between elections. This suggests that the elections themselves have an effect on how this method impacts estimate accuracy. Intuitively, this makes sense as we would expect similarity between historic and current elections to affect the extent to which informative priors improve estimate accuracy.

Altogether, the results do not suggest that the alternative specification with

informative priors is beneficial to estimate accuracy. In a very specific circumstance, and for the smallest sample size, estimate accuracy improves. But the results also show far bigger and more consistent increases in MAE, with some models producing results that are on average nearly 2% worse.

5.4.2 Standard MRP with informative priors

For the remainder of the results section I focus on the standard MRP specification with informative priors. First, in table 5.4 and 5.5 I show the increase or decrease in MAE from the baseline model. As with the tables above, results by sample size are displayed in the columns, while λ values are along the rows grouped by election year.

The UK results, in table 5.4, are somewhat mixed. Informative priors seem to have a different impact depending on the election and sample size. Looking at the first three rows, the results for 2017 show that informative priors for this election do not seem to change estimate accuracy significantly. For the smallest sample, informative priors improve accuracy with a decrease in MAE of 0.1-0.2%. For a sample of 10, MAE becomes worse, for 20 respondents the accuracy is identical to the baseline and for the largest sample size, when $\lambda = 1$ *or* 1.5 MAE increases by 0.1%, while when $\lambda = 2$ there is no difference. For 2019, the results are clearer and more supportive of the case for informative priors. The smallest sample size of 5 shows significant improvements in MAE of around 0.8-0.9%. For the average sample sizes of 10 and 20, MAE decreases by 0.1% for all λ values, and for the largest sample size MAE is identical to the baseline model.

The results for the US are once again starker than the results for the UK, with larger and more consistent improvements in MAE. For the 2012 US Presidential election, the smallest sample size shows improvements in MAE between 1.3-1.4% for the three λ values. However, for all sample sizes of 10 and above and for all λ values, MAE increased when compared to the baseline model(s), ranging from 0.1-0.4%. For

Table 5.4: Standard MRP accuracy (UK)

	Sample			
Model	5	10	20	30
2017				
Lambda = 1	-0.1%	0.2%	0%	-0.1%
Lambda = 1.5	-0.1%	0.2%	0%	-0.1%
Lambda = 2	-0.2%	0.1%	0%	0%
2019				
Lambda = 1	-0.8%	-0.1%	-0.1%	0%
Lambda = 1.5	-0.8%	-0.1%	-0.1%	0%
Lambda = 2	-0.9%	-0.1%	-0.1%	0%

Note: showing inf. prior MAE as +/- from baseline.

Table 5.5: Standard MRP accuracy (US)

	Sample			
Model	5	10	20	30
2012				
Lambda = 1	-1.4%	0.4%	0.3%	0.3%
Lambda = 1.5	-1.3%	0.2%	0.2%	0.1%
Lambda = 2	-1.3%	0.1%	0.1%	0.1%
2016				
Lambda = 1	-0.8%	-0.3%	-0.3%	-0.1%
Lambda = 1.5	-0.8%	-0.3%	-0.2%	-0.1%
Lambda = 2	-0.6%	-0.2%	-0.1%	-0.1%

Note: showing inf. prior MAE as +/- from baseline.

2016, the results show consistent improvement in MAE for all informative prior models. For the smallest sample size, the improvements in MAE are not as large as shown for 2012, but still range from 0.6-0.8%. As sample size increases, the improvements in MAE are smaller, but are still consistently an improvement from the baseline model.

Although the differences between different λ values seems to be small, there does seem to be a pattern for both elections. In 2012, with the exception of the smallest sample size, error is larger for smaller λ values. Whereas in 2016, the smaller λ values

generally result in larger improvements.

5.4.3 Do informative priors enable smaller sample sizes?

I next present analysis which explores whether informative prior models could enable the use of smaller sample sizes. That is, whether the accuracy of models with a smaller sample size and informative priors are comparable to the accuracy of models with a larger sample size and weakly-informative priors. In table 5.6 and 5.7, as with the other tables shown above, I show the increase or decrease of MAE from the baseline model. However, here the baseline model is the model with weakly-informative priors and one sample size above. For example, in table 5.6 the smallest sample size of 5, shows the informative prior model being compared with the model with weakly-informative priors and a sample size of 10.

Table 5.6: Larger sample accuracy comparison (UK)

	Sample		
Model	5	10	20
2017			
Lamba = 1	0.3%	0.6%	0%
Lamba = 1.5	0.3%	0.6%	0%
Lamba = 2	0.2%	0.5%	0%
2019			
Lamba = 1	0.3%	0.1%	0%
Lamba = 1.5	0.3%	0.1%	0%
Lamba = 2	0.2%	0.1%	0%

Note: showing inf. prior MAE as +/- from larger sample baseline.

In table 5.6 I show the results for the two UK elections. Overall, the results show that informative prior models do not allow us to produce accuracy that is comparable to models with a larger sample size and weakly-informative priors. For both elections, for sample sizes of 5 and 10 and all λ values, the estimates have greater error than models with larger sample sizes and weakly-informative priors. For the sample size of

Table 5.7: Larger sample accuracy comparison (US)

	Sample		
Model	5	10	20
2012			
Lamba = 1	0.7%	1.3%	0.6%
Lamba = 1.5	0.8%	1.1%	0.5%
Lamba = 2	0.7%	1.1%	0.4%
2016			
Lamba = 1	0.2%	0.2%	0.1%
Lamba = 1.5	0.2%	0.3%	0.1%
Lamba = 2	0.4%	0.3%	0.2%

Note: showing inf. prior MAE as +/- from larger sample baseline.

20, for both elections and all λ values the results are identical to the larger sample size and weakly-informative prior model.

The results for US, shown in table 5.7, show a similar story. The smaller sample size models with informative priors do not produce estimates that are comparable to the larger sample size models with weakly-informative priors. The poorer estimate accuracy is particularly prevalent for the 2012 election, where increase in MAE ranges from 0.4-1.3%. For the 2016 election, the increase in MAE is smaller than is apparent for 2012, with increases ranging from 0.1-0.4%.

Overall, the results presented in table 5.6 and 5.7 show that sample size has a larger impact on estimate accuracy than informative priors might. The degree, if any, to which informative priors enable better estimate accuracy is somewhat unclear. But it is clear that any gains in estimates accuracy are not sufficient to allow researchers to use smaller sampler sizes.

5.4.4 Estimate precision

The results so far show improvements in estimate accuracy are inconsistent. Significant improvements are observed for some elections, but only present for the smallest sample

sizes. Below I present analysis which explores whether informative priors improve precision in estimates. To explore this I use estimate widths, that is, the difference between the 5% and 95% credible interval. In figure 5.2 I show the range in estimate widths across all UK constituencies for the 2019 election. In the figure the points indicate the median width (also shown in the top left of each plot), the thicker lines show the 50% interval and the thin line shows the 80% interval in estimate widths. I show individual plots for each sample size (shown in the columns) and for the different priors (shown in the rows). Points and intervals to the right signify a larger width, while points and intervals to the left signify shorter widths and therefore greater precision in MRP estimates.

Overall, figure 5.2 shows that informative priors tend to improve estimate precision. This is evident in the plots which show most point and interval ranges to the left of the baseline. Furthermore, for nearly all lambda values and sample sizes, the median estimate width is smaller than the baseline median width. This is true for all sample sizes, but the differences are more pronounced for smaller sample sizes.

In figure 5.3, I show estimate widths for the US 2016 election, with sample sizes in columns and lambda values in rows. The results are somewhat less clear than the UK case. For the two smallest sample sizes, the figure clearly shows that informative priors improve precision. Both the median width reported, as well as the range in widths, show improvements from the baseline. For the two larger sample sizes, the median estimate widths are identical to the baseline. For the sample size of 20, the interval range for estimate widths seems marginally smaller than the baseline, but for the largest sample size the range in widths seems broadly similar to the baseline.

For both the US and the UK applications, differences in estimate widths seem to follow a similar pattern to the one that emerged when looking at estimate accuracy. Informative priors show improvements, but these are mostly evident for smaller sample

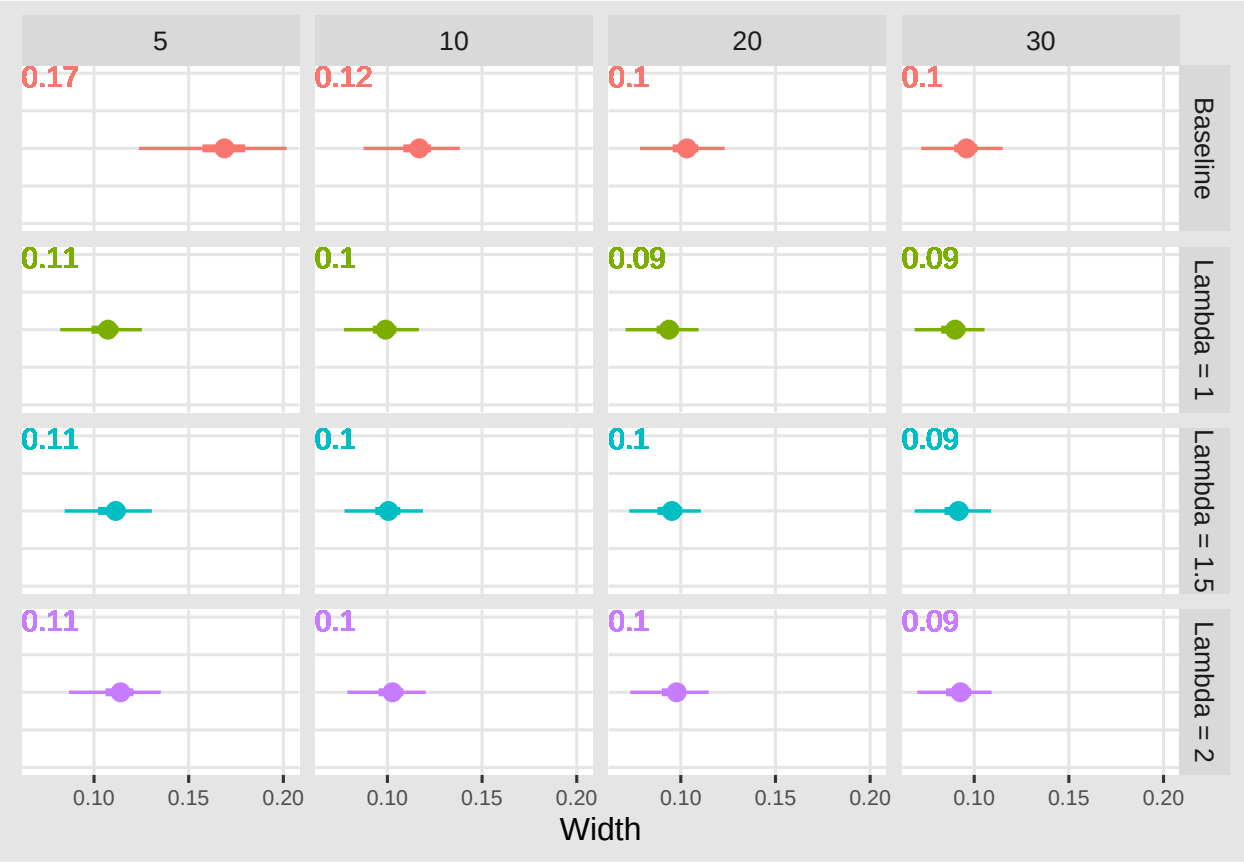

Figure 5.2: UK (2019) average widths

Notes: Showing interval range of constituency estimate widths. For clarity range restricted to 80% interval range.

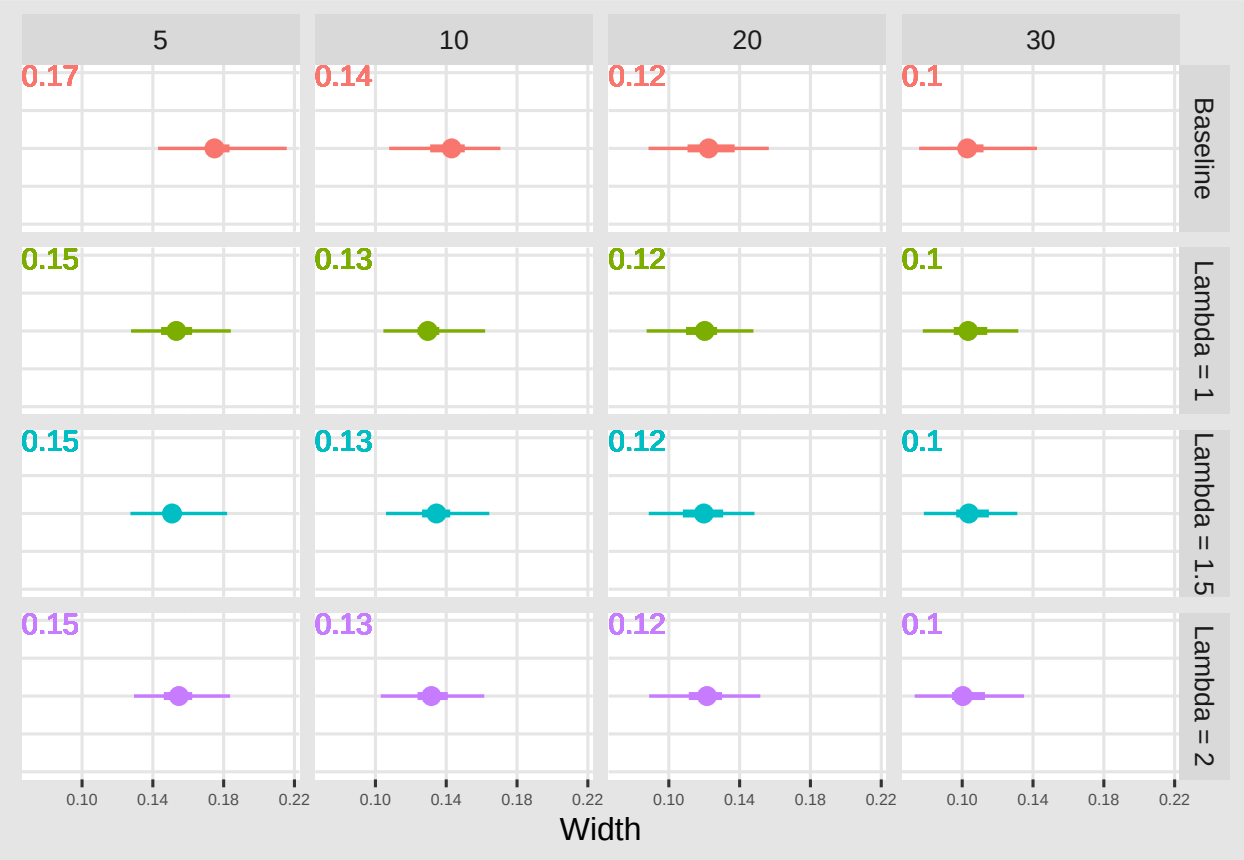

Figure 5.3: US (2016) average widths

Notes: Showing interval range of state estimate widths. For clarity range restricted to 80% interval range.

sizes, and are more prominent for the UK 2019 and US 2016 election.[11] For larger sample sizes and for UK 2017 and US 2012 election, improvements are small in real terms and there are also examples when informative priors perform worse than the baseline.

5.4.5 Parameter estimation

The above analysis has shown how small area estimates change with informative priors. I next explore where and how these changes manifest in the estimation of parameters. I use two variables to show differences between the distribution of parameters. In

[11] I show the plots for UK 2017 and US 2012 in appendix D.3.

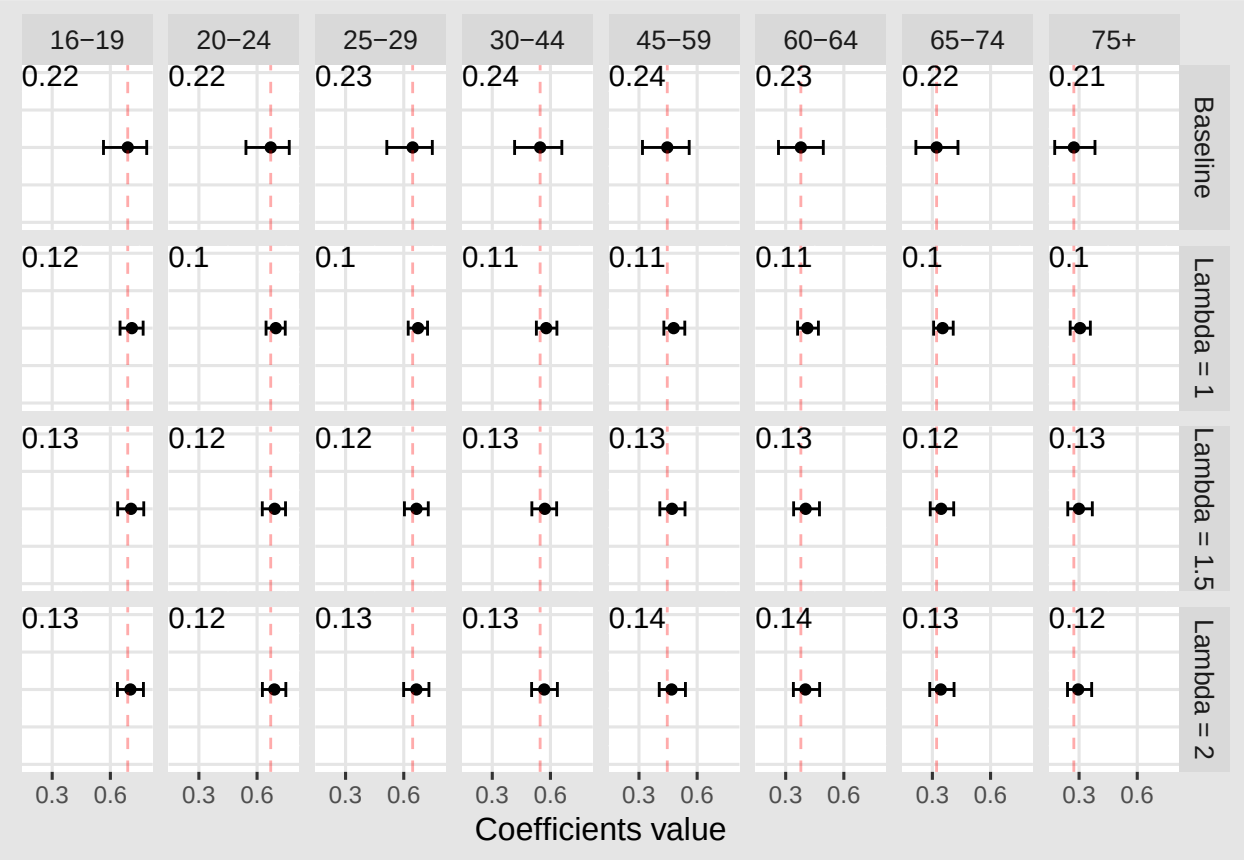

Figure 5.4: UK age coefficient plot
Notes: showing UK 2017 model age category coefficients. Values converted from logit to probability.

the UK I show age categories and in the US I show education categories. These parameters were selected because there are starker differences between the baseline and the informative prior parameters. For some parameters the differences are not as notable, while some parameters show no signs of difference from baseline models. Figure 5.4 shows the parameter coefficient of each age category for voting Labour in the UK case. The plot is for the 2017 election and for models with a sample of 30.[12]

In figure 5.4, each column is a separate age category and each row is either the baseline or an informative prior model with a λ value of 1, 1.5 or 2. In each plot the point is the median parameter value and the lines show the 90% credible interval for

[12]I selected the largest sample size to demonstrate how informative priors can impact parameter estimation despite this sample size showing little change in the resulting MRP estimates.

the parameter. In each individual plot the value in the top left corner is the width (the difference between the 5% and 95% interval). The red dashed line in each plot represents the median point for the baseline model for each given age category. Using the red dashed line, we can see that the points (i.e. the median of the parameter distributions) are consistently similar, although not identical. Where $\lambda = 1$ *or* 1.5 the parameter estimates are to the right of the red dashed line. This indicates that these models have effects which increase the probability of voting Labour to a greater extent than the Baseline. This is also true when $\lambda = 2$, but the differences seem marginally smaller. This is in line with the wider results, which have highlighted that when $\lambda = 2$ estimates are closer if not identical to the baseline. Importantly, results show how the informative priors affect the credible interval ranges. For each λ value, the informative priors decrease the width of the estimates. This highlights how informative priors could enable improved inference for sub-groups of the population (here age categories). The improvements in credible intervals are particularly notable for smaller λ values, with width increasing as λ increases.

A similar pattern is shown in figure 5.5, which shows the parameter coefficients for each education category for the US 2012 election. As with the UK case, the median parameter values are similar to the baseline, but show some small differences. For those with 'no high school diploma' and 'postgraduate' the parameters are smaller than the baseline, as shown by the median points positioned to the left of the dashed line. For all other categories, the parameters show a marginally larger effect than the baseline, shown by the median points positioned to the right of the dashed red line. In absolute terms these differences are small and likely insignificant. Again, the more significant difference is exhibited in the credible intervals of each education category. We can see that when $\lambda = 1$, the 90% credible interval is smaller by up to 10% for each education category. The improvement in inference is still significant for the two larger λ values, but to a lesser extent.

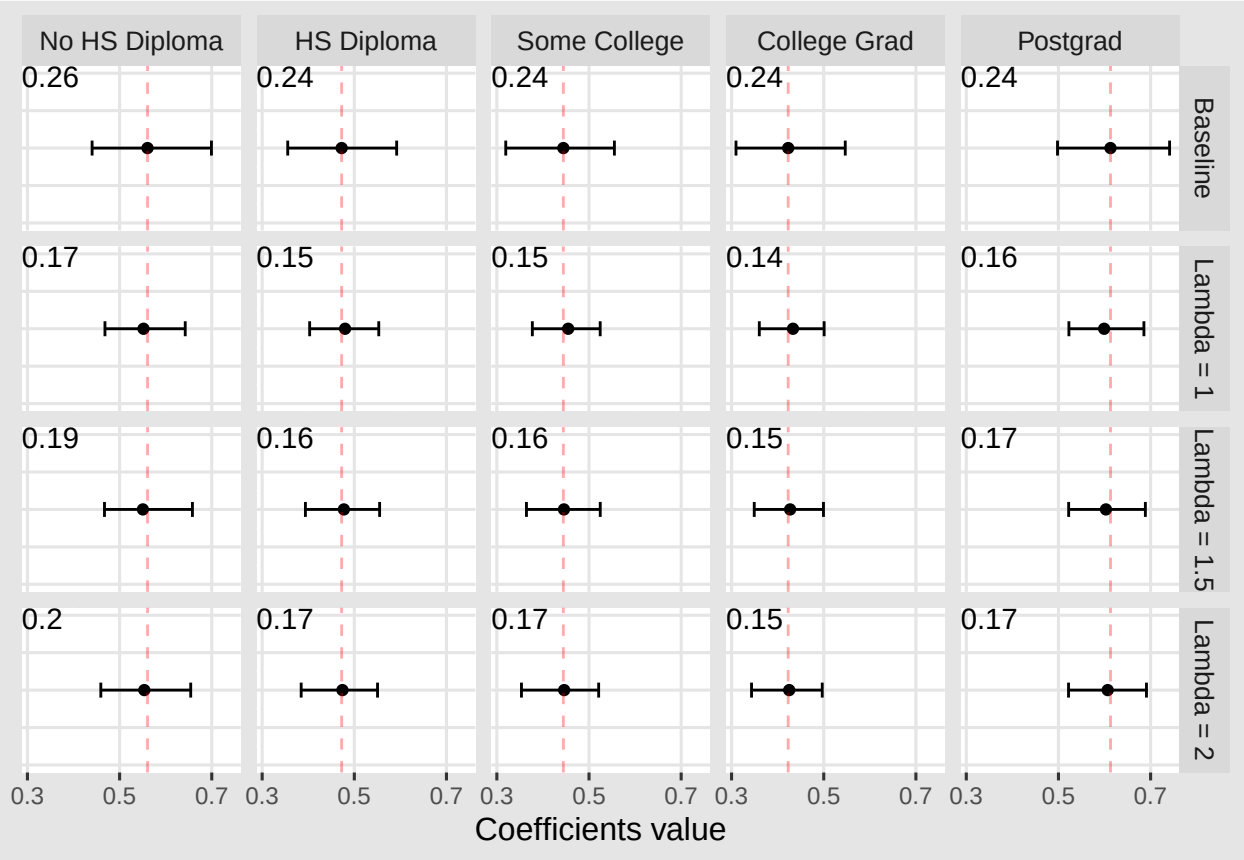

Figure 5.5: US education coefficient plot

Notes: showing US 2012 model education category coefficients. Values converted from logit to probability.

5.4.6 Computational efficiency

MRP can be computationally demanding, owing partly to large sample sizes, complex models with numerous cross-level interactions, and a large number of small areas.[13] Here, informative priors may be of use as a means to improve estimation efficiency. Informative priors can achieve this because the parameter search space is reduced, better enabling the model to estimate parameters efficiently. To explore whether informative priors have been useful here, I use model chain estimation time, including both warm-up and sampling.[14]

In table 5.8 and 5.9 I report the change from the baseline model for all MRP specifications, λ values, sample sizes, and elections. The figures reported are percentage increase or decrease from the associated baseline model. For example, in table 5.8, for the 2017 election, sample size of 5, and $\lambda = 1$, the figure of -35% represents a 35% reduction in estimation time. From the table, it is evident that informative priors consistently improve computational efficiency. However, improvements are most noticeable when $\lambda = 1$, with model run-time reduced to the greatest extent. For larger λ values, the improvements are less consistent. This is particularly the case for the two larger lambda values and the largest sample sizes, where improvements range from 4-7% reduction in model run-time, with one model increasing model run-time by 3%.

For the US, the results are somewhat similar. The smallest λ value consistently and significantly reduces model run-time. When $\lambda = 1.5$ model run-times are nearly always reduced, but to a lesser degree. For the largest λ value of 2, the results are mixed. For some models, model run-time is an increase on the baseline, while for

[13] For example, in Lauderdale et al. (2020) sample sizes are between 40,000-80,000 respondents, estimate for a minimum of 380+ small areas, and include numerous (cross-level) interactions.

[14] A previous version of this chapter reported model run-time as the difference between start and finish of running in R. This meant that compilation of the model in brms was also included in the time (i.e. the process of the model set-up). When model tun-time was measured in this way, smaller sample sizes often had longer model run-times. This is of note, because improvements in run-time could be offset by compilation times when sample sizes are small.

Table 5.8: UK computational efficiency

	Sample			
Model	5	10	20	30
2017				
Lambda = 1	-35%	-50%	-44%	-37%
Lambda = 1.5	-37%	-38%	-41%	-7%
Lambda = 2	-35%	-34%	-40%	3%
2019				
Lambda = 1	-26%	-32%	-39%	-30%
Lambda = 1.5	-16%	-29%	-34%	-4%
Lambda = 2	-11%	-25%	-14%	-7%

Note: Showing % +/- from baseline model run-time

Table 5.9: US computational efficiency

	Sample			
Model	5	10	20	30
2012				
Lambda = 1	-27%	-54%	-26%	-43%
Lambda = 1.5	-10%	-32%	-1%	-11%
Lambda = 2	1%	-33%	10%	-6%
2016				
Lambda = 1	-30%	-16%	-24%	-37%
Lambda = 1.5	-23%	-3%	1%	-15%
Lambda = 2	20%	5%	-12%	-4%

Note: Showing % +/- from baseline model run-time

others there is a reduction in time. Whether the model is an increase or decrease seems somewhat random, with no clear patter related to sample size or election.

Overall, gains in computational efficiency are greater for smaller λ values. When $\lambda = 1$ there is a consistent decrease in model run-time, ranging in the UK from 26-50% and in the US from 16-54%. For larger λ values, in the UK, there is still a reduction in model run-time, but often these are smaller and less consistent improvements. In the US, When $\lambda = 1.5$ *or* 2, there was a decrease in estimation time for most cases, but also instances when model run-time increased.

5.5 Similarity between elections

The results presented so far show inconsistencies in whether informative priors improve MRP estimates. I argue that one reason for these inconsistencies is variation between how similar the historic election is to the current election. When an election is similar to the historic election we would expect informative priors to improve estimate accuracy. Conversely, when there are greater discrepancies between the current and historic election, we would expect the informative priors to either have no effect on estimates, or worsen estimate accuracy.

In the UK case, the 2017 election was seen as a shift from the previous elections. The EU referendum took place in between the elections and has been argued to have had a separate and distinct effect (Hobolt, 2018), and caused a realignment of voters (Heath and Goodwin, 2017). This meant between the two elections, there were complicated voter flows (Mellon et al., 2018). Although there is some dispute whether voter 'realignment' was Brexit related (Jennings and Stoker, 2017), or was a return to older voter alignments (Johnston, Rossiter, Manley, et al., 2018), both these arguments are a recognition that there was a shift between 2015 and 2017.

For some, 2019 was a clear continuation of the political realignment that manifested

in the previous elections (Cutts et al., 2020; Flinders, 2020). However, others argue that 2019 was a continuation for the Conservative Party, but not for Labour (Prosser, 2021), or at least not to the same degree that was present for the Conservative Party (Curtice, 2020). Overall, the literature suggests that the 2019 UK election exhibited some similarities and also demonstrated some divergence from the previous election.

The UK results presented in this study are somewhat in line with the literature on the similarity between 2015-2017 and 2017-2019. For the alternative specification, the results show this method did not improve MAE. Neither 2015-2017 or 2017-2019 were similar enough for this strategy to be effective. This method incorporates past information in a stricter and more direct way. As a result, unless elections are very similar, we would not expect the alternative method to improve estimate accuracy. For the standard MRP specification, informative priors improved estimate accuracy for the smallest sample sizes, but only when elections shared some similarities with the previous election. That is, using 2017 information improved 2019 electoral forecasts because of greater similarity, but using 2015 information for 2017 did not improve estimate accuracy.

In the US case, Obama's second election in 2012 was seen as similar to the 2008 election, with Obama's 2008 coalition of voters maintained (Galston, 2013). In both elections, younger, non-white, and urban voters heavily favoured Obama. In the 2016 election, support for Clinton from this coalition was not maintained to the same degree. For example, non-white voters still favored Clinton over Trump, but not to the same margin they did for Obama over Romney. Some suggest that vote switching between 2012 and 2016 took place to a significant degree among white working class voters (Morgan and Lee, 2018). However, this trend has also been argued to have roots during the Obama presidency. White non-college educated voters shifted to the Republican party, with the more significant changes taking place after 2008 election (Sides et al., 2017; Weisberg, 2015).

The results presented in this study show signs of support for the literature and some divergence from the literature. The alternative specification improved estimate accuracy for the smallest sample size for 2012 but not for 2016. This is in line with the literature which argues that 2012 was a continuation from 2008, while in 2016 voting patterns changed. For the standard MRP specification, informative priors significantly improved estimate accuracy for the smallest sample sizes. This was apparent for both elections, but improvements were greater for the 2012 election. However, for larger sample sizes, informative priors worsened estimate accuracy at the 2012 election, whereas in 2016 informative priors continued to improve estimate accuracy. This is at odds with some literature, which argues that 2012 was more similar to its preceding election than was the case for 2016. Based on the literature alone, we would have expected 2012 to show consistent improvements rather than 2016.

To explore similarities between elections further, below I present analysis which seeks to quantify variation between the elections. I combine historic and current data and estimate a Bayesian multilevel logistic model for vote choice. The models are identical to the standard multilevel models estimated in this chapter, but also include a varying intercept term for the election year.[15] The year effect should capture variation between the elections controlling for the same variables used throughout this chapter. Although this is an imperfect measure, it should be a sufficient indicator of similarity. In figure 5.6 I show coefficient plots for the year varying intercept terms. The points show the median of the parameter distribution, and the lines show the credible intervals. Although the parameter values are somewhat arbitrary, a smaller value indicates greater similarity between the historic and current election. In the UK case, the 2019 election coefficient is much closer to zero (left-hand side) than the 2017 election, indicating 2017 showed greater variation from its historic election

[15]For the US case, I merge the total samples to model combined vote choice. In the UK, I used a total sample of 7,000 respondents, randomly sampled from the total survey sample and with an even split between historic and current data. The sample of 7,000 was chosen as this sample size consistently ensured coverage of all 632 small areas.

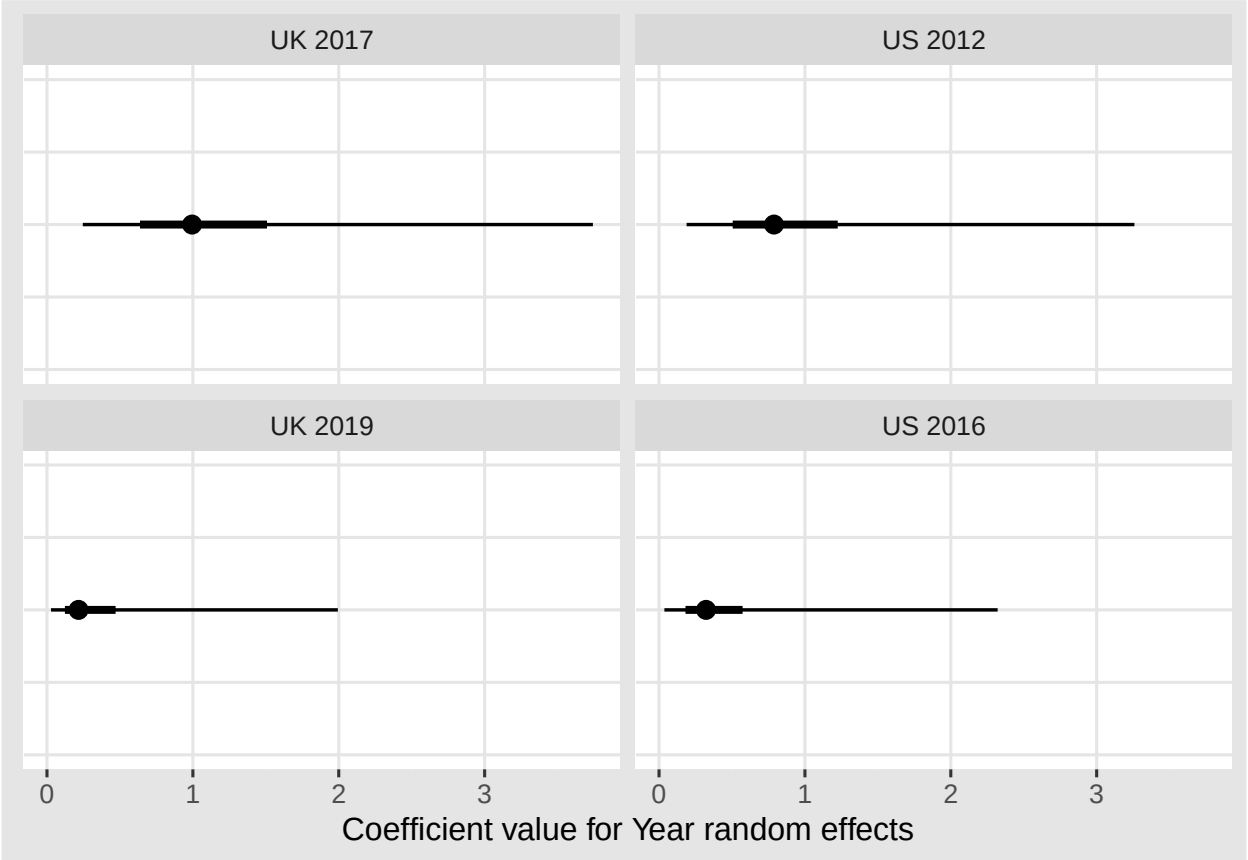

Figure 5.6: Election similarity
Notes: Models include current and historic data. Showing year varying intercept coefficient.

than was the case for 2019. This is broadly in line with both the literature and the results presented in this study. In the US case, the 2016 election shows a much smaller parameter value than the 2012 election. This demonstrates that the 2016 election was more similar to the 2012 election, than the 2012 election was to the 2008 election. This is at odds with some of the literature, but in line with the results presented in this chapter.

5.6 Discussion

This chapter explored how a two-stage prior elicitation method could be used to specify informative priors for MRP. Specifically, the chapter explored how we could

use a historic election model to derive informative priors for a current MRP electoral forecast model. The research first tested whether informative priors with an alternative MRP specification could improve estimate accuracy. Here, each variable category was modelled as a separate β parameter.[16] On the whole, the results showed the method was not beneficial to estimate accuracy and for some cases highly detrimental. I believe this is testament to the benefits of partial pooling gained through estimating demographic characteristics as varying intercepts.

For the standard MRP specification, improvements in estimate accuracy were mostly present for the smaller sample sizes. There were examples of improvements for larger sample sizes, but these were small and inconsistent. Overall, the research has shown that sample size is clearly far more important than informative priors. Even when informative priors improve estimate accuracy, the gains still do not make the estimates comparable with estimates from a model with a larger sample size and weakly-informative priors. Researchers only interested in MRP estimate accuracy, should focus on ensuring that they can achieve a sufficient sample size rather than informative priors.

Beyond estimate accuracy, the research has demonstrated that informative priors have the potential to improve precision in estimates. The results showed that estimate widths were often shorter than weakly-informative prior models. However, as with estimate accuracy, improvements were inconsistent, showing greater improvements for smaller sample sizes and for specific elections. The results, although showing informative priors could improve estimate precision, once again demonstrated that sample size seemed to have a larger affect on estimate precision than informative priors might.

Benefits of this method for estimate accuracy and precision are inconsistent. However, improvements in estimation time seem more consistent and significant.

[16]The model still estimated small area as random effects as the sample sizes used here would not permit this variable to be specified as a fixed effect.

Overall, informative priors nearly always and often significantly improve estimation times. This is especially the case for models with many small areas (i.e. the UK) and smaller λ values. Researchers will of course preference estimate accuracy over computation time, but the results show when used at the right election and properly, informative priors consistently improve computation time. The results have also shown that we can improve inference among subgroups of the population with informative priors. This is not true for all demographic variables, but there were clear examples of informative priors improving the precision of parameter estimation.

Finally, while the research tested using λ as a scaling constant to account for differences between elections, the results are inconclusive to determine which value we should use. The results indicate that the value of lambda should be linked to the similarity between the elections. Where elections are similar, a smaller λ value is more useful. On the other hand, when there is a discrepancy between elections, a smaller λ value is more detrimental to estimate accuracy. The larger λ values seem to offer less benefit to estimates, with the accuracy being similar to the weakly-informative model accuracy. This suggests that λ values of 1.5 and 2 are too large to enable the model to incorporate the benefits of informative priors. However, these values are far less risky than a value of 1, where the potential detriment to estimates is much greater. Future research could test whether this method could be improved with a different range of λ values, for example > 1, but < 1.5. These values might be better at avoiding the risks associated with $\lambda = 1$, but better able to realise the benefits of informative priors than when $\lambda = 1.5$ *or* 2.

For both the alternative and standard MRP specification, the research has shown that the potential success of such a strategy is dependent on the extent to which an election is similar to the preceding election. This, however, emphasises the major potential risk of the method. In practical applications we do not know the extent to which voting patterns will shift, and therefore the use of this method is highly risky.

Although I have demonstrated a method to determine similarity between elections (which broadly seems to align with cases where this method has worked best), I have not developed a method which can establish whether informative priors can or should be used. New and innovative work which seeks to develop a standardised measure on distance (or similarity) between elections might offer a potential solution (See Faliszewski et al., 2020, 2019).

Another potential avenue for future research could be to explore whether the method set out in this chapter could be improved with different historic information. First, research could examine whether different elections might be better suited as the historic model. For instance, it has been argued that US midterm elections, which are held two-years into a presidency, represent a 'referendum' on the incumbent President (Tufte, 1975). Using the midterm election as the historic model might provide informative priors which are closer to the *true* parameter distribution than the previous presidential election.

Second, research could explore whether informative priors could be used throughout an election campaign. In this application, a historic model would use data from the initial stages of a campaign period. Informative priors would be derived from the historic model and imputed for a model predicting vote share closer to the election date. This would negate the issue of similarity between elections, and could improve estimation for the final election forecast model. However, if there is a significant shift in opinion over the course of a campaign, this technique would be at risk of the same problems of using a dissimilar historic election.

Throughout this chapter I have discussed how we can use historic data to improve current MRP estimates. However, this overlooks the potential risk of incorporating errors from the historic model or data. If the historic model and or data is inaccurate, this method would embed these errors into the current model. We could check historic model estimate accuracy in an attempt to ensure that the model and data is accurate.

However, even when small area estimates are accurate, parameters may still have been estimated poorly. And while these errors might not harm estimates for the historic model, they may cause greater estimate inaccuracy for the current election model.

5.7 Conclusion

This study has demonstrated that informative priors can be useful in specific circumstances when forecasting elections with MRP. However, the results also demonstrate that using this method without knowledge of similarity between historic and current election could be problematic. Importantly, in this chapter I have shown that for estimate accuracy, we should be far more concerned with ensuring that we have a large enough sample. Indeed, significant accuracy improvements from informative priors were largely only evident for the smallest sample sizes.

Besides accuracy, the research has shown that informative priors may be useful to improve sub-group inference by improving the estimation of parameters. Equally, the research has shown that informative priors could aid MRP by improving computation time. In future research it would be interesting to explore whether this method is useful for much more complex models. For example if we wish to include numerous interactions (including cross-level interactions), informative priors may be particularly useful for improving efficiency in estimation.

The study has contributed to the wider Bayesian literature on prior elicitation by showing how a two-stage prior elicitation method could be applied to electoral forecasting. Furthermore, the chapter has set out how we can incorporate a scaling constant (similar to that used with power priors) and apply to the two-stage method to account for variation between historic and current data.

For the MRP literature generally, and specifically for MRP electoral forecasting, the chapter has contributed to our understanding of when this method could be employed.

This chapter has shown under what conditions this strategy could improve MRP estimates and estimation. Importantly, the chapter has shown that for most instances the risks of this strategy outweigh the potential benefits. Overall, the implications suggest that researchers should focus on sample size over informative priors. Should a researcher only have access to small sample sizes, this method might be useful. But, as has been shown, there are still risks when the historic election is dissimilar to the current election. For researchers concerned with sub-group inference, the results give some suggestion that this strategy might be useful, but more research is needed in this area.

Chapter 6

Conclusion

The contribution of MRP to social science cannot be underestimated. By equipping researchers with the means to reliably estimate opinion at a sub-national level, the method has advanced numerous academic disciplines, extending both the potential research topics and available methods to research on already established topics.

We know that opinion can vary significantly within countries, including geographically among small areas. While this thesis has not argued that all disciplines require analysis of opinion at the small area level, I maintain that failing to account for geographic variation altogether is an oversight. Accordingly, this thesis was motivated by a belief that the continued development and improvement of MRP is important and worth the attention of methodological research.

In recognising both the importance of small area estimates of opinion and the capacity of MRP to reliably estimate small area opinion, this thesis was designed with two principal aims: first, to contribute to an improved understanding of the method; and second, explore whether alternatives or extensions of the standard methodology can improve MRP estimates. Below, I briefly re-visit the main chapters of this thesis, summarising the key findings and identifying important contributions of each. This is followed by a discussion on the overarching limitations of this study and suggestions

for relevant future research. I conclude with a brief note of advice for the applied researcher.

6.1 MRP and variable selection

In chapter three, I argued that variable selection in applications of MRP needs improvement. The systematic review identified limited variation in the variables used across a *relatively* broad range of estimated opinion or behaviour. This suggests that many studies base variable selection on a path dependency approach (selecting identical variables to previous MRP studies) rather than theory based selection. Given that most applications estimate opinion or behaviour that could be deemed to have similar predictive variables, one could argue that this does not represent a significant problem. However, in line with Buttice and Highton (2013), who noted variation in estimate accuracy with identical predictor variables, I argued that we should expect more variation.

As an alternative, chapter three explored how we can use cross-validation (CV) lasso regression to select variables for use with MRP, and whether this leads to improved MRP estimate accuracy. This was achieved through the application of estimating Conservative party vote share in GB constituencies for the 2017 election. The chapter first explored what degree of regularisation (i.e. what λ value) researchers should use when using CV lasso regression to select variables for MRP. In line with recommendations of Breiman (1998) and Hastie et al. (2009), this research found that $\hat{\lambda}$ (the lambda value which minimises squared error across cross-validation) enforced too little regularisation resulting in an overfitted model. On the other hand, the model associated with $\hat{\lambda}1Std$ (largest lambda value within 1 standard error of $\hat{\lambda}$) was consistently one of the most accurate among all λ values tested. The $\hat{\lambda}1Std$ value imposed more regularisation leading to a more parsimonious model and better

out-of-sample prediction. I argued this demonstrated that for MRP variable selection applications, researchers should use $\hat{\lambda}1Std$ for CV lasso regression.

The chapter subsequently assessed the performance of lasso-MRP estimates against path dependency and theory. When compared to two path dependency models, lasso-MRP performed equally, if not better. With no discernible differences in accuracy between lasso-MRP and the best path dependency model, I argued that the lasso model represented a superior choice given that it was a more parsimonious solution (in this case fewer variables). When comparing the lasso-MRP estimates with theory driven variable selection, the theory MRP model achieved significantly better estimate accuracy. However, the research was not able to make a fair comparison between the two. The theory model used a much larger sample, included variables unavailable in the lasso-MRP model, and included numerous cross-level interactions. Nonetheless, taking a conservative approach, the chapter suggested that theory based variable selection is preferable to lasso variable selection.

The chapter argued that the results would support the use of the lasso-MRP method as a tool to be deployed in the model-building process. The systematic review seemed to indicate that many MRP applications do not use theory for variable selection. Therefore, incorporating lasso into the model building process alongside theory and path dependency would represent an improvement in MRP variable selection. Importantly, this addition comes with no financial cost to the researcher, and only relatively small computation and time costs.

This chapter has contributed to the MRP literature by continuing to explore how we can better select variables. More specifically, the chapter has contributed by setting-out how best to apply CV lasso regression to select variables for MRP. Two existing papers have explored how automated variable selection can be combined with MRP. However, Goplerud et al. (2018) method changes the MRP format, while Broniecki et al. (2021) does not offer a solution which can select individual-level,

area-level, and interactions in a unified framework. Chapter three addresses this gap by exploring how CV lasso can select all MRP variables and interactions simultaneously.

The benefits of such a method might be small when there is *good* theory to use, but when there is little or no theory to inform variable selection this method could prove particularly useful. While I argue that the results in this chapter do not support the indiscriminate use of CV lasso to select variables for MRP, the results suggest that when used alongside other methods in the model building process the method could improve applications of MRP. Overall, improving variable selection for MRP could lead to improved MRP estimate accuracy and could extend the application of MRP to new areas where we have less theory to inform variable selection.

The chapter also contributed to the wider literature on selecting λ for lasso with cross-validation curves. The 'one-standard-error' rule was proposed as a conservative approach to selecting λ values, as $\hat{\lambda}$ had been shown to select overfitted solutions (Hastie et al., 2009: 244). As an alternative, $\hat{\lambda}1Std$ selects a far more parsimonious model while also keeping within one standard error of $\hat{\lambda}$ cross-validation error. This chapter has contributed to this literature by providing further evidence that $\hat{\lambda}1Std$ is preferable for out-of-sample prediction, and that $\hat{\lambda}$ selects overfitted solutions. This speaks to the wider bias-variance trade off debate. In this example, reducing variance is preferable for prediction accuracy.

6.2 Improved MRP sample distribution

In chapter four I explored how an unevenly distributed sample can improve estimate accuracy. The premise of this chapter was that in certain applications of MRP there is greater need to ensure estimate accuracy in certain small areas. The clearest example - and the application in this chapter - is electoral forecasting. In most elections, the electoral outcome is based upon which party can win a majority of electoral contests

across all electoral districts (small areas). To predict an electoral outcome we therefore need to forecast vote choice or party winners in each electoral district.

However, in most elections small areas are divided between those where one party or candidate wins with a large majority of votes (non-marginals) and those where a party wins with a small majority (marginals). To correctly predict the electoral outcome in marginal small areas we therefore need a high level of accuracy and precision. Whereas, in non-marginal small areas we can typically afford lower accuracy without the risk of incorrectly predicting the small area party winner.

This chapter explored whether oversampling respondents from marginal small areas improved MRP electoral forecasting. This was achieved by first introducing a method to determine how the sample should be distributed among small areas, and second, by assessing whether this improved MRP estimates. Through a simulation study and two real-world applications (2017 UK and 2016 US), the chapter demonstrated that the method both improves estimate accuracy and our ability to predict elections.

In the simulation study I showed that overall accuracy was largely unaffected by the different sample distributions. However, when looking at small areas by margin of victory, the results showed that the uneven sample distributions improved accuracy and precision among small areas which received a larger proportion of the sample. On the other hand, small areas which received a smaller proportion of the sample had poorer accuracy when compared to the even sample distribution.

Improvements in marginal small area estimate accuracy also manifested in the two real-world applications. In the UK I estimated 2017 Conservative vote share and in the US 2016 Republican vote share. In the UK improvements in MAE ranged between 0.6-0.7%, while in the US improvements ranged between 0.8-1%. These improvements in estimate accuracy translated into improved ability to predict an election, with both the UK and US cases showing significant improvements in brier scores. Interestingly, in the real-world applications the improvements in estimate accuracy in marginal

small areas did not come at a cost of accuracy in non-marginal small areas. In the US, I argue this is partly a function of the data. However, across both real-world settings, I argued this is in part because of the weighted ratio I used to determine sample distributions. The weighted ratio proved an efficient way to determine a sample distribution that satisfies dual objectives: increasing the sample in marginal small areas and maintaining sufficient sample in non-marginal small areas. In the two real-world applications, the oversampling method combined with the weighted ratio improved estimate accuracy in marginal small areas, but not to the detriment of non-marginal small area accuracy.

The implications of this research are two-fold. First, for electoral forecasting both with MRP and without, this research has demonstrated how an uneven sample distribution may be useful to improve our ability to forecast an election. This is an important finding and speaks to wider literature on sampling strategies for voting behaviour. By over-sampling small areas which have small margins of victory, we can improve our ability to predict an overall electoral result. The implications of this research should be of particular interest to researchers with limited or finite resources. In these applications, researchers will have a maximum sample available, and this strategy has demonstrated how we can achieve the best possible accuracy with a given sample size.

Second, the results have implications for the wider application of MRP. In most applications of MRP, researchers make use of publicly available surveys which in most cases do not sample respondents evenly. The results presented here further demonstrate that an unevenly distributed sample will result in varying degrees of estimate accuracy across small areas. In the applied use of MRP, researchers should attempt to ensure that their results, and the interpretation of their results, better account for this variation in accuracy. This implication is especially important for studies which use MRP point estimates in a further model.

6.3 MRP and informative priors

Chapter five looked at how we could incorporate further information into the MRP model through Bayesian informative priors. MRP is increasingly estimated as a Bayesian model and this provides an opportunity for researchers to incorporate further information into the method. In this chapter, I set out how we can apply a two-stage prior elicitation method to specify informative priors with MRP. Through applications to forecasting elections in the US and UK, I showed how we can use historic election model posteriors as informative priors. Specifically, I demonstrated how to obtain the distributions for each parameter posterior and impute these as priors for the current election model. I also explored how we could incorporate a λ value into the prior specification process, with the intention of accounting for differences between the historic and current election.

This approach was applied with the standard MRP specification, and with an alternative MRP specification that enabled a more direct impact of the informative priors. For both specifications, the chapter tested whether informative priors improved estimate accuracy when compared to a standard MRP model with weakly informative priors.

The alternative specification results showed that the method did not improve estimate accuracy for the most part, and at points was actively detrimental for estimate accuracy. This, I argued, was testament to the benefits of partial pooling obtained by specifying individual-level variables as varying intercept terms. For the standard MRP specification with informative priors, the results were somewhat mixed. There were examples of the method improving estimating accuracy, but mostly showed informative priors to be of limited benefit at best, and marginally detrimental towards estimate accuracy at worst.

For the smallest sample sizes, informative priors improved estimate accuracy, but the degree of improvement varied depending on the election. For larger sample sizes,

informative priors typically either had little effect on estimates or were detrimental to accuracy. The research was able to highlight some other benefits including improved estimate precision, improvements in computational efficiency and subgroup inference. But overall, the chapter demonstrated that this strategy cannot be applied indiscriminately nor uniformly, and would require preliminary analysis to ensure the right conditions were met to achieve any benefits.

The implications of this research for the application of Bayesian MRP models are two-fold. First, the research has demonstrated that researchers should focus on sample size over informative priors. Although increasing sample size is not a cost free strategy, the research has shown that informative priors cannot produce comparable results to MRP models with larger sample sizes. More broadly, the only sample size which demonstrated somewhat consistent improvements in estimate accuracy are smaller than the standard MRP sample sizes identified in the systematic review.

Second, the research has shown under what conditions informative priors might be useful. If a researcher only has access to small sample sizes, informative priors might be an attractive strategy to improve estimate accuracy. In the application of electoral forecasting, the research demonstrates that when elections are similar, using the historical model posterior as priors for the current election can improve estimate accuracy. When there are dissimilarities between elections, this method will work to the detriment of estimate accuracy. Although I have not explored whether this holds true for other opinion or behaviour, it seems a reasonable assumption to suggest similar patterns will manifest.

For the wider Bayesian literature on priors, this chapter has contributed to exploring how historical model posteriors may be directly imputed as informative priors. This area of research is a growing field, with the two-stage prior elicitation a relatively new and emerging method. In this chapter, I provided a worked example of how this approach could be applied to an electoral forecasting context. Motivated by the

scaling constant a_0 parameter in power priors (which controls for similarity between historic and current data), the chapter also showed how the two-stage method could be adapted to account for differences between the historic and current election through the use of a λ value.

6.4 Limitations and future research

This thesis has sought to assess *new* ways to leverage information and *new* information that we can leverage. Throughout the thesis I assessed how these leveraging techniques could be applied to MRP and whether they improve MRP estimates. This was undertaken partly with the goal of contributing towards the wider application of MRP. Part of the appeal of MRP is that it can be uniformly applied across different academic disciplines, and as the systematic review identified, broadly has a standard practice in its application. However, in this thesis I have proposed and tested methodologies that could not be indiscriminately nor uniformly applied across MRP applications. This limits the impact of this research as I have not developed 'ready to use' methods for the wider application of MRP. Each additional step suggested here would most likely need to be tailored for the variety of MRP applications identified in the systematic review.

The uneven sample distribution introduced in chapter four could not be uniformly applied across electoral contests without modification. An example of this is in the German Bundestag where voters have two votes, or France, where parliamentary and presidential candidates are on the same ballot. In these situations a small area that is a marginal in one vote might not be in another. This means that researchers would need to find new methods to define marginals and determine sample distributions.

In chapter five, I show that informative priors can improve MRP estimate accuracy but only for the smallest sample sizes and only under certain conditions. Overall,

improved estimate accuracy and precision, as well as improved sub-group inference, will most likely only manifest when the historic and current election are similar. For the wider application of this method we would first have to determine a way to reliably establish electoral similarities *a priori*.

Future research could focus on ways in which these methods could be applied uniformly across a variety of applications rather than the specific applications in this thesis. For the wider application of the uneven sample distribution, one avenue would be to further develop the weighted ratio so that it could be applied universally to multiparty systems with more than one vote. This research would of course also need to assess whether these sample distributions still maintained the improvements in estimate accuracy that were reported in chapter four. For informative priors to be uniformly applied, we would need a method to assess similarity between elections. Proponents of the power prior (Chen and Ibrahim, 2000) would argue that this already exists. An extension of the basic power prior allows for the modelling of the scaling constant a_0, which controls for how influential the historical data is in the estimation of the posterior. Incorporating power priors to the MRP case might be a fruitful avenue of future research.

A related limitation is that throughout this thesis I have only applied each application to electoral forecasting. While chapter four most likely has limited application beyond electoral forecasting, chapter three and five could be applied to topics beyond voting behaviour. From a practical standpoint, voting behaviour is one of the best test applications, as we can compare estimates against known outcomes. However, by only estimating voting behaviour there are two significant limitations: first, I cannot claim that these methods and findings from this research more generally, will be applicable to other opinion or behavior; second, I cannot claim to have contributed to the wider application of the method. Although part of the premise of this research was the contribution to the extension of MRP. Without providing firm examples of

these methods contributing to the wider application of MRP, it could be argued this thesis has not yet satisfied this goal.

As well as not demonstrating application beyond electoral forecasting, the way I have estimated vote choice could be argued to restrict the extent to which I have truly tested the efficacy of these methods for electoral forecasting. Throughout this thesis I have consistently treated vote choice as a binary outcome where support for a political party = 1, and support for all other parties = 0. This decision was not substantive, but rather practical and followed the examples set out in previous MRP studies (see Selb and Munzert, 2011; Hanretty et al., 2016). On the one hand, estimating vote choice in this format was preferable to demonstrate these methods for MRP. This is because in most applications researchers are estimating opinion towards a given topic where support = 1, and opposition = 0, and thus identical to how vote choice is operationalised here. On the other hand, treating vote choice as a binary outcome in a multiparty system is most likely different to present-day MRP electoral forecasting, with researchers preferring multinomial regression which estimates all main parties in a single model.

This could be deemed problematic as this means I do not know whether various alternative applications tested in this thesis are applicable for how researchers actually forecast elections with MRP, i.e. with a multinomial MRP model. For example, in chapter three selecting variables for a multinomial model is not possible with the Group-Lasso interaction-NET I make use of. Future research would need to test alternative lasso solutions which facilitate the selection of variables with a multinomial model. In chapter four, I calculate whether a small area is a marginal by calculating the difference between the proportion in each small area where $Y = 1$ *and* $Y \neq 1$. In a multiparty system, this is a problematic way to identify marginals and could have high type I and II error rates. When treating vote choice as binary in a multiparty system, determining marginals will always be an imperfect solution. If we estimate a

multinomial MRP model for vote choice, this would negate the problem as we would define a marginal in terms of past winning candidate majority. Finally, in chapter five estimating the model as a multinomial model would not have altered the application of informative priors significantly. However, the more complex nature of a multinomial model might be a circumstance where informative priors are of greater benefit than any of the applications tested in this thesis. Again, estimating vote choice as a binary outcome has not enabled an assessment of informative priors true to how MRP is used to forecast elections.

In the introduction, I framed the thesis along the lines of how we can better leverage information for MRP applications. In each chapter, I have discussed how we can leverage information on the assumption that the information will improve MRP estimates. However, as is the case for MRP more generally, using past information in any form can be problematic. In chapter four using past information takes the form of using past vote to determine which small areas should be deemed marginal, and therefore receive a larger proportion of the sample. In chapter five, the historic model is used to derive informative priors for the current election model. In both of these applications, we are assuming that the historic information will be useful for the current MRP model. In electoral forecasting this is often a fair assumption, but an assumption nonetheless. If the information we incorporate into an MRP model is not useful, this may not only fail to improve MRP estimates, but it could also be (highly) detrimental to estimate accuracy. Overall, throughout the thesis there has been recognition that historic information could be problematic for MRP estimates, but I rarely make a significant contribution to proposing ways to alleviate these concerns.

Future research could seek to develop methods which are better able to determine whether the historic information will be useful. Of particular interest could be a new area of research which seeks to identify how we can measure the similarity (or distance) between elections (See Faliszewski et al., 2020, 2019). If we are able calculate

a standardised measure of similarity between elections, sensitivity analysis could be used to determine the distance at which a historic election is, or is not, useful.

Another important limitation concerns how I have analysed whether each proposed methods improves MRP. Assessing efficacy of information leveraging techniques proposed in this thesis has predominantly been done by focusing on estimate accuracy. As a prediction rather than inference method, this approach has largely been the most appropriate. However, there are other criteria that might have been useful to assess the alternative methods, but I have been restricted by the lack of ground truth available to make such assessments.

In chapter three, I compared lasso-MRP with a path dependency variable selection approach. In the chapter I compared accuracy and the variables selected, but was not able to compare whether either approach selected the *best* variables, as there was no ground truth on which variables were the most predictive of voting behaviour. This was particularly problematic for assessing the efficacy of the method to select interactions and individual-level variables. Without knowledge of which variables should be selected I was unable to assess whether the lasso extension I applied (Group-Lasso interaction-NET) was able to correctly select all variables and interactions. This was a significant limitation because the method was chosen partly on the basis of its capacity to select all variables and interactions in a unified framework.

In chapter five, I showed that informative priors changed some parameter values and hypothesised that these were the cause for the differences in estimate accuracy. However, without knowledge of the *true* parameter values, I could not determine whether these parameter estimates were an improvement over parameters which were estimated with weakly-informative priors.

In real-world settings, overcoming these limitations is somewhat impossible. Simulation studies might be able to provide some indication, but will not offer a comparable analysis with the same degree of complexity present in real-world applications. Ulti-

mately, assessing implications of the methods tested in this thesis by estimate accuracy is the most practical but nonetheless, poses limitations for developing an understanding of exactly how each alternative method changes MRP estimation.

For some of the methods proposed in this thesis, the benefits would most likely arise in applications with highly complex MRP specifications. However, for the most part I have consistently restricted MRP complexity in each application. In chapter three, I used a poststratification frame that only allowed for a limited set of individual-level variables and I did not allow cross-level interactions. Previous research has demonstrated that political individual-level variables and cross-level interactions can produce highly accurate MRP vote choice estimates (See Lauderdale et al., 2020). Therefore, restricting potential variables and interactions might have meant that benefits from this approach did not manifest because the *most* predictive variables and interactions were not available.

Similarly, in chapter five I explored how informative priors might improve estimate accuracy. I restricted the models to *relatively* simple specifications with no interactions altogether. However, posterior estimation is harder when including interactions and informative priors might have significantly aided posterior estimation. Overall, the simple models tested in chapter five are not representative of model complexity in real-world electoral forecasting applications. This could call into question whether I have offered a *true* test for whether informative priors improve MRP estimates and estimation.

At numerous points throughout this thesis I have stated that restricting complexity was a reasonable choice and one that did not affect satisfying research objectives. From the perspective of creating a feasible research project this is most likely true. However, from the perspective of testing whether methodologies improve accuracy, one could argue that the MRP applications in this thesis should have been closer in comparison to real MRP applications. This speaks to a wider internal vs. external validity trade-off

in methodological research. In this thesis, I have attempted to replicate how MRP is applied in practice, while also simplifying the model to ensure I could identify whether any of the proposed methodologies *truly* improve MRP estimates.

6.5 Advice for applied researchers

For the applied MRP researcher, this thesis has numerous implications. In this closing statement, I try to translate these implications into advice for applied researchers.

Before discussing direct implications from this research, I first wish to briefly discuss the implementation of MRP, the necessary software and tools, and prerequisite knowledge and skills. For applied researchers with a good foundation in statistics and programming, the implementation of all of the methods in this project are feasible. Those with limited experience in either of these should first dedicate time towards acquiring the prerequisite knowledge and skills. The application of MRP more broadly requires understanding of multilevel modelling, and more recently Bayesian statistics. Useful textbooks which will take researchers from limited statistical knowledge and coding experience, through to being comfortable with Bayesian statistics and multilevel modelling include Gelman et al. (2020); McElreath (2020), while Gelman and Hill (2007) is an excellent textbook, it assumes the reader already has some knowledge of Bayesian statistics.

Should researchers already have the prerequisite knowledge, implementation of the standard MRP form is relatively straightforward. However, those with no prior experience in applying MRP, should first refer to the numerous free-to-access online resources which provide worked examples, including code which researchers can borrow from (See Alexander, 2019; Williams, 2018; Rivers, 2018; Dunham, 2018; Mastny, 2017; Leemann and Wasserfallen, 2018; Hanretty, 2019; Kastellec et al., 2016; Lopez-Martin et al., 2019; Kennedy and Gabry, 2020).[1]

[1]These resources were largely informed from a list compiled by Josh McCrain, accessed here:

Throughout this thesis I have used R (R. C. Team, 2020), and the R package tidyverse (Wickham et al., 2019), to clean, wrangle, and visualise data. This was essential to produce the research embodied in this thesis. Researchers who have a preference for alternative programs, such as Python, could easily implement all data preparation with this software instead. However, most of the development of MRP and, to my knowledge, worked examples of MRP use R. This means that researchers who wish to implement MRP in different software will need to be able to translate R code or statistical notation into their preferred programming language.

Similarly, the MRP models in the thesis has all been implemented with the Bayesian modelling programme Stan, and called through Rstan (S. D. Team, 2020) with the aid of R packages rstanarm (Goodrich et al., 2020) and brms (Bürkner, 2017). These packages enable the specification of Stan models with standard R model-fitting functions, making implementing complex Bayesian models relatively easy and straightforward. Researchers who have a preference for Python, could implement their analysis through Pystan, the Python interface with Stan. Although there are limited Python libraries which enable researchers to forgo writing the full MRP model specification in the Stan model format.[2] Alternatively, models could be implemented in other Bayesian probabilistic programming tools such as PyMC3 (Salvatier et al., 2016), or Edward (Tran et al., 2017). Although these tools could produce different estimates, as both use different Bayesian sampling methods to Stan's default No-U-Turn Sampler (Hoffman and Gelman, 2011) Overall, for those who do not have an advanced knowledge of Bayesian statistics or advanced programming in Python, the use of Rstan (called through rstanarm or brms) is recommended for applied researchers.

Turning now to the implications of this thesis for applied researchers. First, when

`https://joshuamccrain.com/index.php/mrp-in-r/`

[2]There are some recent examples of packages designed to make implementing Stan in python easier, including pybrms (Haber 2020), the Python version of brms.

estimating opinion or behaviour with MRP, theory should drive variable selection. For instance, in electoral forecasting there is an extensive body of literature which examines what the key and main drivers of vote choice are. This body of literature now also includes numerous examples of MRP being applied to electoral forecasting. For researchers forecasting elections with MRP, they should base variable selection on both the theory and MRP examples. However, as an additional technique lasso may compliment the model building process. Importantly, it comes with relatively few costs to the researcher and may provide further confidence in variable selection.

Should a researcher wish to use CV lasso regression to select variables, researchers should make use of $\hat{\lambda}1Std$ (the largest λ value within one standard error of $\hat{\lambda}$, the value with the lowest CV error). More broadly, researchers should choose greater regularisation as this is preferential for out-of-sample prediction (i.e. in MRP the poststratification stage).

For the wider application of MRP, variable selection is often harder. This is because we cannot validate our choices by comparing previous model results with known outcomes. In these circumstances, when there are limited examples of MRP being applied to estimate the opinion, and the wider theory on which variables are predictive is limited, the incorporation of lasso into the model building process becomes more important. Using lasso as a complementary variable selection method can provide researchers with confidence that the variables selected will lead to reliable and accurate estimates. Instead of lasso, alternative solutions such as autoMRP (Broniecki et al., 2021), and sMRP (Goplerud et al., 2018) could also be viable options, should a researcher be less interested in selecting individual-level variables or maintaining the standard MRP form. Although, researchers may still prefer the lasso approach as this maintains the standard multilevel model which offers greater model intelligibility than alternative MRP regularisation approaches (Gao et al., 2021).

The oversampling strategy proposed in chapter four seems to show promising signs

for electoral forecasting. When researchers have control over sampling design, they should follow the method set out in chapter four. The method has been shown to improve accuracy in the most important small areas, while having no detrimental affect on overall accuracy. The weighted ratio might need adapting for certain applications, but it is an efficient way to determine sample distributions and should be used when applying this method. Overall, the method will improve prediction accuracy in important small areas, which in turn can improve the probability of predicting an election.

For the wider application of MRP, this method could be useful where greater accuracy is needed for a subset of small areas which can be identified in advance. However, when accuracy is equally important for all small areas, this method will rarely be suitable. Despite the real-world examples showing almost no risks involved in applying the method, the simulation demonstrated there was potential the method could be detrimental to accuracy in certain small areas. Taking a conservative approach, this method should only be applied when two criteria are met: first, improving accuracy in certain small areas is a justified necessity; second, there is a high degree of confidence that poorer accuracy in certain small areas will not impact the overall prediction goals.

Although it is not advised this method is applied to all MRP applications, the results still have implications for the wider use of MRP. The chapter provided further evidence that unevenly distributed samples result in differing degrees of accuracy across small areas. When researchers use surveys which distribute the sample unevenly among small areas, the discussion and interpretation of the MRP results should give greater consideration to these differences. This is especially important when using MRP results in further analysis. In these applications, analysis using MRP results should account for the variation in estimate accuracy among small areas.

Informative priors could be an attractive feature to employ when forecasting elections. However, the results presented in this thesis should be seen as a cautionary

tale how the indiscriminate use could be problematic. Should a researcher wish to use informative priors for MRP electoral forecasting, the advice is twofold: first, researchers should focus on ensuring they can achieve a sufficient sample size, as sample size is clearly more important than prior type; second, researchers need to ensure that there are similarities between elections, which at present we do not have the methods to achieve. If researchers are interested in improving computation time, or interested in sub-group inference of voting patterns, informative priors could be useful. However, at present there is not sufficient evidence to warrant the use of this method to achieve these goals.

Informative priors deployed in the two-stage procedure should rarely be used for wider applications.[3] This is mainly because there are most likely few applications where there is good enough past information that would warrant the use of informative priors as used in this chapter. Furthermore, the gains in accuracy seem to be small given the potential risks of the two-stage method. As with electoral forecasting, researchers should focus on sample size over informative priors. Overall, I struggle to see an application beyond electoral forecasting where we can have confidence the two-stage method, identical to the one set-out in chapter five, would improve MRP estimates. This is not to say informative priors more generally will not improve estimates, but should not be used as operationalised in this chapter.

More broadly, in line with previous literature on MRP, this research has continued to show the importance of sample size. In the simulation study in chapter four, accuracy improvements were greater for smaller sample sizes, and for chapter five, significant improvements were mostly only present for the smallest sample size. This emphasises both the impact and importance that sample size has on MRP accuracy, and suggests researchers should always first ensure they have a sufficient sample size for their MRP application. If researchers are limited by available sample size, then

[3]For the explanation of informative and weakly-informative priors, refer back to 5.1, and to see details on the two-stage prior elicitation method refer back to 5.2.

some of the leveraging techniques proposed in this thesis could be useful. This lesson has truth beyond this thesis, where alternative methodologies proposed in the MRP literature tend to show particular improvements when sample sizes are small.

Applied researchers may also improve their models by paying closer attention to what their goals of applying MRP are, and what metrics they should use to assess the accuracy of their MRP application. The uneven sample distribution was implemented because I considered that predicting an election was more important than overall estimate accuracy. Equally, we might choose to use informative priors because we are more interested in sub-group inference or estimate precision. Overall, only considering estimate accuracy might not lead to the best methodological decisions. In the applied us of MRP, researchers should give consideration to what the overall goals of their MRP application is, and how these might affect their modelling decisions.

Similarly, researchers must ensure they tailor their MRP model according to their specific application. The need to tailor MRP models for each application is a theme that has run throughout this thesis. In the introduction, I stated there is no one-size fits all for MRP and this has broadly rung true throughout each chapter. Automated variable selection was an attempt to tailor variables choices for MRP applications. The uneven sample distribution is a tailored approach of MRP based on the need for higher accuracy in certain small areas. Equally, informative priors are a way to tailor the information we provide the MRP model. Clearly tailoring the model is the basis for the successful application of MRP, and researchers should be advised to embrace this practice.

Appendix A

Chapter 2

In chapter 2 I presented the results from the systematic review which documented standard practice for the application of MRP in social science studies. In appendix A.1 I list all studies included in the systematic review. In appendix A.2 I provide the list of topics and inclusion criteria. This list was originally adapted from the Pew Research Centre list of research topics.

A.1 Systematic review studies

Berkman MB, Plutzer E (2011). "Local Autonomy versus State Constraints: Balancing Evolution and Creationism in U.S. High Schools." *Publius: The Journal of Federalism*, *41*(4), 610-635.

Binderkrantz AS, Nielsen MK, Pedersen HH, Tromborg MW (2020). "Pre-parliamentary party career and political representation." *West European Politics*, *43*(6), 1315-1338.

Bishin BG, Smith CA (2013). "When Do Legislators Defy Popular Sovereignty? Testing Theories of Minority Representation Using DOMA." *Political Research Quarterly*, *66*(4), 794-803.

Bromley-Trujillo R, Poe J (2020). "The importance of salience: public opinion

and state policy action on climate change." *Journal of Public Policy*, *40*(2), 280-304.

Broockman DE, Skovron C (2018). "Bias in Perceptions of Public Opinion among Political Elites." *American Political Science Review*, *112*(3), 542-563.

Buttice MK, Highton B (2013). "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis*, *21*(4), 449-467.

Butz AM, Kehrberg JE (2015). "Social Distrust and Immigrant Access to Welfare Programs in the American States: Social Distrust and Immigrant Access." *Politics & Policy*, *43*(2), 256-286.

Butz AM, Kehrberg JE (2016). "Estimating anti-immigrant sentiment for the American states using multi-level modeling and post-stratification, 2004-2008." *Research & Politics*, *3*(2), 1-7.

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

Canes-Wrone B, Clark TS, Semet A (2018). "Judicial Elections, Public Opinion, and Decisions on Lower-Salience Issues." *Journal of Empirical Legal Studies*, *15*(4), 672-707.

Carson A, Ratcliff S, Dufresne Y (2018). "Public opinion and policy responsiveness: the case of same-sex marriage in Australia." *Australian Journal of Political Science*, *53*(1), 3-23.

Cohen JE, Rottinghaus B (2018). "Constituent Approval, Electoral Marginality, and Congressional Support for the President: Approval, Support, and Marginality." *Presidential Studies Quarterly*, *48*(2), 202-224.

Enns PK, Koch J (2013). "Public Opinion in the U.S. States: 1956 to 2010." *State Politics & Policy Quarterly*, *13*(3), 349-372.

Enns PK, Lagodny J, Schuldt JP (2017). "Understanding the 2016 US Presidential Polls: The Importance of Hidden Trump Supporters." *Statistics, Politics and Policy*,

8(1).

Eun Kim S, Urpelainen J (2018). "Environmental public opinion in U.S. states, 1973-2012." *Environmental Politics*, *27*(1), 89-114.

Fairbrother M, Martin IW (2013). "Does inequality erode social trust? Results from multilevel models of US states and counties." *Social Science Research*, *42*(2), 347-360.

Figueiredo J, Campos P (2013). "Estimation of Underrepresented Strata in Preelection Polls: A Comparative Study." In Oliveira PE, da GraÃ§a Temido M, Henriques C, Vichi M (eds.), *Recent Developments in Modeling and Applications in Statistics*, 47-57. Springer Berlin Heidelberg, Berlin, Heidelberg. doi: 10.1007/978-3-642-32419-2_6 (URL: `https://doi.org/10.1007/978-3-642-32419-2_6`).

Flavin P, Franko WW (2017). "Government's Unequal Attentiveness to Citizens' Political Priorities: Government's Unequal Attentiveness to Citizens' Political Priorities." *Policy Studies Journal*, *45*(4), 659-687.

Flores AR, Herman JL, Mallory C (2015). "Transgender inclusion in state nondiscrimination policies: The democratic deficit and political powerlessness." *Research & Politics*, *2*(4), 205316801561224.

Fowler L (2016). "The states of public opinion on the environment." *Environmental Politics*, *25*(2), 315-337.

Fowler L (2017). "Tracking state trends in environmental public opinion." *The Social Science Journal*, *54*(3), 287-294.

Gelman A, Lee D, Ghitza Y (2010). "Public Opinion on Health Care Reform." *The Forum*, *8*(1), 1-14.

Gerber AS, Huber GA, Doherty D, Dowling CM (2016). "Why People Vote: Estimating the Social Returns to Voting." *British Journal of Political Science*, *46*(2), 241-264.

Ghitza Y, Gelman A (2013). "Deep Interactions with MRP: Election Turnout and

Voting Patterns Among Small Electoral Subgroups: DEEP INTERACTIONS WITH MRP." *American Journal of Political Science*, *57*(3), 762-776.

Gibson JL, Claassen C (2016). "Macro-Tolerance and Protest: Does a Culture of Political Intolerance Dampen Dissent?" *SSRN Electronic Journal*.

Grogan CM, Park S((2017). "The Racial Divide in State Medicaid Expansions." *Journal of Health Politics, Policy and Law*, *42*(3), 539-572.

Hanretty C (2019). "An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion." *Political Studies Review*, 147892991986477.

Hanretty C, Lauderdale BE, Vivyan N (2016). "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods*, *6*(3), 571-591.

Hanretty C, Lauderdale BE, Vivyan N (2017). "Dyadic Representation in a Westminster System: Dyadic Representation in a Westminster System." *Legislative Studies Quarterly*, *42*(2), 235-267.

Hansen ER, Treul SA (2015). "The Symbolic and Substantive Representation of LGB Americans in the US House." *The Journal of Politics*, *77*(4), 955-967.

Hare C, Monogan JE (2020). "The democratic deficit on salient issues: immigration and healthcare in the states." *Journal of Public Policy*, *40*(1), 116-143.

Hawley G (2013). "Issue Voting and Immigration: Do Restrictionist Policies Cost Congressional Republicans Votes?: Issue Voting and Immigration." *Social Science Quarterly*, *94*(5), 1185-1206.

Hertel-Fernandez A (2014). "Who Passes Business's"Model Bills"? Policy Capacity and Corporate Influence in U.S. State Politics." *Perspectives on Politics*, *12*(3), 582-602.

Hill SJ (2015). "Institution of Nomination and the Policy Ideology of Primary Electorates." *Quarterly Journal of Political Science*, *10*(4), 461-487.

Houston D (2019). "Schoolhouse Democracy: Public Opinion and Education Spending in the States." *Educational Researcher*, *48*(7), 438-451.

Howe PD (2018). "Modeling Geographic Variation in Household Disaster Preparedness across U.S. States and Metropolitan Areas." *The Professional Geographer*, *70*(3), 491-503.

Howe PD, Marlon JR, Wang X, Leiserowitz A (2019). "Public perceptions of the health risks of extreme heat across US states, counties, and neighborhoods." *Proceedings of the National Academy of Sciences*, *116*(14), 6743-6748.

Howe PD, Mildenberger M, Marlon JR, Leiserowitz A (2015). "Geographic variation in opinions on climate change at state and local scales in the USA." *Nature Climate Change*, *5*(6), 596-603.

Jaeger WP, Lyons J, Wolak J (2017). "Political Knowledge and Policy Representation in the States." *American Politics Research*, *45*(6), 907-938.

Jeong G (2013). "Congressional Politics of U.S. Immigration Reforms: Legislative Outcomes Under Multidimensional Negotiations." *Political Research Quarterly*, *66*(3), 600-614.

Kastellec JP (2018). "How Courts Structure State-Level Representation." *State Politics & Policy Quarterly*, *18*(1), 27-60.

Kastellec JP, Lax JR, Malecki M, Phillips JH (2015). "Polarizing the Electoral Connection: Partisan Representation in Supreme Court Confirmation Politics." *The Journal of Politics*, *77*(3), 787-804.

Kastellec JP, Lax JR, Phillips JH (2010). "Public Opinion and Senate Confirmation of Supreme Court Nominees." *The Journal of Politics*, *72*(3), 767-784.

Kaufmann RK, Mann ML, Gopal S, Liederman JA, Howe PD, Pretis F, Tang X, Gilmore M (2017). "Spatial heterogeneity of climate change as an experiential basis for skepticism." *Proceedings of the National Academy of Sciences*, *114*(1), 67-71.

Kehrberg JE (2017). "The Mediating Effect of Authoritarianism on Immigrant

Access to TANF: A State-Level Analysis: AUTHORITARIANISM AND TANF." *Political Science Quarterly*, *132*(2), 291-311.

Koch J, Thomsen DM (2017). "Gender Equality Mood across States and over Time." *State Politics & Policy Quarterly*, *17*(4), 351-360.

Kogan V (2017). "Administrative Centralization and Bureaucratic Responsiveness: Evidence from the Food Stamp Program." *Journal of Public Administration Research and Theory*, *27*(4), 629-646.

Konitzer T, Rothschild D, Hill S, Wilbur KC (2019). "Using Big Data and Algorithms to Determine the Effect of Geographically Targeted Advertising on Vote Intention: Evidence From the 2012 U.S. Presidential Election." *Political Communication*, *36*(1), 1-16.

Krimmel K (2019). "Rights by Fortune or Fight? Reexamining the Addition of Sex to Title VII of the Civil Rights Act." *Legislative Studies Quarterly*, *44*(2), 271-306.

Krimmel K, Lax JR, Phillips JH (2016). "Gay Rights in Congress: Public Opinion and (Mis)representation." *Public Opinion Quarterly*, *80*(4), 888-913.

Krimmel K, Rader K (2017). "The Federal Spending Paradox: Economic Self-Interest and Symbolic Racism in Contemporary Fiscal Politics." *American Politics Research*, *45*(5), 727-754.

Lauderdale BE, 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.

Lax JR, Phillips JH (2013). "How Should We Estimate Sub-National Opinion Using MRP? Preliminary Findings and Recommendations." *Unpublished manuscript*.

Lax JR, Phillips JH (2012). "The Democratic Deficit in the States." *American Journal of Political Science*, *56*(1), 148-166.

Lax JR, Phillips JH (2009). "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review*, *103*(3), 367-386.

Ledford C (2018). "Symbolic Racism, Institutional Bias, and Welfare Drug Testing Legislation: Racial Biases Matter: Racial Biases Matter." *Policy Studies Journal*, *46*(3), 510-530.

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.

Leemann L, Wasserfallen F (2016). "The Democratic Effect of Direct Democracy." *American Political Science Review*, *110*(4), 750-762.

Lei R, Gelman A, Ghitza Y (2017). "The 2008 Election: A Preregistered Replication Analysis." *Statistics and Public Policy*, *4*(1), 1-8.

Lewis DC, Jacobsmeier ML (2017). "Evaluating Policy Representation with Dynamic MRP Estimates: Direct Democracy and Same-Sex Relationship Policies in the United States." *State Politics & Policy Quarterly*, *17*(4), 441-464.

Lewis DC, Wood FS, Jacobsmeier ML (2014). "Public Opinion and Judicial Behavior in Direct Democracy Systems: Gay Rights in the American States." *State Politics & Policy Quarterly*, *14*(4), 367-388.

Magidin de Kramer R, Tighe E, Saxe L, Parmer D (2018). "Assessing the Validity of Data Synthesis Methods to Estimate Religious Populations: VALIDITY OF RELIGIOUS POPULATION ESTIMATES." *Journal for the Scientific Study of Religion*, *57*(2), 206-220.

Martin DC, Newman BJ (2015). "Measuring Aggregate Social Capital Using Census Response Rates." *American Politics Research*, *43*(4), 625-642.

Milazzo C, Townsley J (2020). "Conceived in Harlesden: Candidate-Centred Campaigning in British General Elections." *Parliamentary Affairs*, *73*(1), 127-146.

Mildenberger M, Howe P, Lachapelle E, Stokes L, Marlon J, Gravelle T (2016). "The Distribution of Climate Change Public Opinion in Canada." *PLOS ONE*, *11*(8), e0159774.

Minozzi W (2014). "Conditions for Dialogue and Dominance in Political Campaigns." *Political Communication*, *31*(1), 73-93.

Muller C, Schrage D (2014). "Mass Imprisonment and Trust in the Law." *The ANNALS of the American Academy of Political and Social Science*, *651*(1), 139-158.

Pacheco J (2011). "Using National Surveys to Measure Dynamic U.S. State Public Opinion: A Guideline for Scholars and an Application." *State Politics & Policy Quarterly*, *11*(4), 415-439.

Pacheco J (2012). "The Social Contagion Model: Exploring the Role of Public Opinion on the Diffusion of Antismoking Legislation across the American States." *The Journal of Politics*, *74*(1), 187-202.

Pacheco J (2013). "The Thermostatic Model of Responsiveness in the American States." *State Politics & Policy Quarterly*, *13*(3), 306-332.

Pacheco J (2014). "Measuring and Evaluating Changes in State Opinion Across Eight Issues." *American Politics Research*, *42*(6), 986-1009.

Pacheco J, Maltby E (2017). "The Role of Public Opinion - Does It Influence the Diffusion of ACA Decisions?" *Journal of Health Politics, Policy and Law*, *42*(2), 309-340.

Pacheco J, Maltby E (2019). "Trends in State-Level Opinions toward the Affordable Care Act." *Journal of Health Politics, Policy and Law*, *44*(5), 737-764.

Park DK, Gelman A, Bafumi J (2004). "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis*, *12*(4), 375-385.

Riddle T, Sinclair S (2019). "Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias." *Proceedings of the National Academy of Sciences*, *116*(17), 8255-8260.

Selb P, Munzert S (2011). "Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information." *Political Analysis*, *19*(4),

455-470.

Skulley C, Silva A, Lang MJ, Collingwood L, Bishin BG (2018). "Majority rule vs. minority rights: immigrant representation despite public opposition on the 1986 immigration reform and control act." *Politics, Groups, and Identities*, *6*(4), 593-611.

Tausanovitch C, Warshaw C (2014). "Representation in Municipal Government." *American Political Science Review*, *108*(3), 605-641.

Tausanovitch C, Warshaw C (2013). "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics*, *75*(2), 330-342.

Tighe E, Livert D, Barnett M, Saxe L (2010). "Cross-Survey Analysis to Estimate Low-Incidence Religious Groups." *Sociological Methods & Research*, *39*(1), 56-82.

Toshkov D (2015). "Exploring the Performance of Multilevel Modeling and Post-stratification with Eurobarometer Data." *Political Analysis*, *23*(3), 455-460.

Tsai C (2015). "Measuring Public Opinion toward Social Welfare in Taiwan." *The Taiwanese Political Science Review*, *19*(1), 242-278.

Vandeweerdt C, Kerremans B, Cohn A (2016). "Climate voting in the US Congress: the power of public concern." *Environmental Politics*, *25*(2), 268-288.

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.

Williams AM, Jephcote C, Janta H, Li G (2018). "The migration intentions of young adults in Europe: A comparative, multilevel analysis." *Population, Space and Place*, *24*(1), e2123.

A.2 List of topics

Table A.1: Topics list

Topic	Notes
Abortion policy	Originally part of Politics and Policy category
Climate, Energy & Environment	Originally part of Politics and Policy, Science and International Affairs categories
Criminal justice	Originally part of Politics and Policy category, includes death penalty
Defense & National security	Originally part of Politics and Policy, International Affairs and Immigration and Migration
Economy & Work	Included Economic Conditions, Income, Wealth & Poverty, Business & Workplace, Economic Policy Global Trade, Economic Systems, Personal Finances
Education policy	Originally part of Politics and Policy, Science, Internet and Technology and Other topics
Elections & voters	Originally part of Politics and Policy category
Family & Relationships	Included Household Structure & Family Roles, Family Caregiving, Marriage & Divorce Parenthood Romance & Dating, Friendships
Gender & LGBT	Included LGBT Attitudes & Experiences, Economics, Gender & Work/ Education/ Politics Gender & Leadership/ Religion, Gender Equality & Discrimination, Gender Roles
Health policy	Politics and Policy, Science and International Affairs categories
Immigration & Migration	Topics included Immigration Attitudes, Immigrant Populations, Immigration & Economy Integration & Identity, Legal Immigration, Refugees & Asylum Seekers Unauthorized Immigration, Border Security & Enforcement, Technology & Immigration
International Affairs	Included World Leaders World Elections, Global Image of Countries, Organizations Alliances & Treaties, Global Balance of Power, Bilateral Relations, War & International Conflict Global Economy & Trade, Global Tech & Cybersecurity, Human Rights, International Political Values
Politics & Policy	Included Trust, Facts & Democracy, Political Parties & Polarization, Politics & Media, Leaders Political Ideals & Systems, Political & Civic Engagement, Issue Priorities Political Discourse, Protests & Uprisings, Generations, Age & Politics
Race & Ethnicity	Included Racial Bias & Discrimination, Race, Ethnicity & Politics, Race Relations Racial Intermarriage, Racial & Ethnic Identity, Racial & Ethnic Shifts, Ethnic groups
Religion	Included Beliefs & Practices, Religion & Social Values, Religious Freedom & Restrictions Religion & Government, Religion & Politics, Interreligious Relations, Non-Religion & Secularism Religion & Science, Religious Demographics, Religious Identity & Affiliation Religious Leaders & Institutions, Religious Knowledge & Education
Science	Included Trust in Science, STEM Education & Workforce, Science Funding Religion & Science, Biotech, Evolution, Food Science, Gene Editing, Space, Human Enhancement
Gun policy	Added
Other	Any topic which did not clearly fit any category.

Appendix B

Chapter 3

In chapter 3 to produce MRP estimates, I used a turnout measure which was estimated using variables selected with the lasso approach. To give confidence that the results - and interpretation of results - were not a function of the turnout measure, I produced MRP estimates with two alternative turnout measures. I provide details of the turnout measures and show alternative results for all λ solutions in B.1.

In B.2 I provide a list of all variables available for selection in the lasso model. The table also provides details on variable selection for different lambda solutions, showing which were selected as a stand-alone variable, which were selected as an interaction variable, and which were selected as both an interaction and as a stand-alone variable.

B.1 Alternative turnout

For the first alternative turnout measure, I estimated by replicating variable selection from previous studies. Turnout was estimated using the same data used in chapter 3 - BES wave 12. The estimation procedure was based on a turnout model from Lauderdale et al. (2020), with variable selection also informed by the meta-analysis of Stockemer (2017). The model included age, education, region, and constituency as varying intercept effects, as well as area-level variables of past turnout and past

seat-winner majority. The second turnout measure was actual 2017 constituency-level turnout. Using actual turnout meant that estimates did not account for different turnout among sub-groups. However, this measure provided a check that results were not a function of a potentially poorly estimated turnout measure.

All turnout measures, including the one used in chapter 3, were imperfect. The two model-based turnout measures produced similar MRP estimates, but both over estimated turnout. However, the estimates produced with these turnout measures were much more accurate than the estimates produced with actual 2017 turnout. This was a result of the way I estimated vote choice - where Conservative = 1 and all other (including 'would not vote') = 0. MRP estimates which used actual turnout consistently under-predicted vote share across all constituencies. Whereas the turnout measures derived from the BES data produced estimates that consistently over predicted turnout, but this worked in favour of constituency vote share estimate accuracy.

Importantly, although MRP estimates are different with the alternative turnout measures, the interpretation of the results does not change. The results still showed more regularisation produced better MRP estimate accuracy, with marginally poorer correlation. Furthermore, both also showed that $\hat{\lambda}1tsd$ was consistently among the models with the highest accuracy. This gives confidence that the results presented in the chapter were not a product of the lasso turnout measure. In table B.1 and B.2 I show increase/decrease from the baseline for MAE, RMSE, and correlation for all λ values. The values reported are a percentage of the baseline ($\hat{\lambda}1tsd$) value. These tables are identical to table 3.1 reported in chapter 3.

Table B.1: Comparing lambda solution's accuracy to baseline

Model	Correl	MAE	RMSE
38	0%	1%	1%

35	0%	1%	1%
29	0%	1%	1%
27	0%	0%	1%
26	0%	0%	-1%
24	0%	1%	1%
23	0%	1%	1%
22	0%	0%	0%
20	0%	1%	1%
19	0%	1%	1%
18	0%	1%	2%
17	0%	3%	3%
16	0%	3%	4%
15	0%	1%	1%
14	0%	3%	3%

Note:

Replication turnout measure

Values are +/- % of the baseline value

Table B.2: Comparing lambda solution's accuracy to baseline

Model	Correl	MAE	RMSE
38	0%	2%	2%
35	0%	-2%	-2%
29	0%	-2%	-2%
27	0%	-1%	-1%
26	0%	-1%	-1%

24	0%	0%	0%
23	0%	0%	0%
22	0%	0%	0%
20	0%	1%	1%
19	0%	1%	1%
18	0%	1%	1%
17	0%	2%	2%
16	0%	2%	2%
15	0%	1%	1%
14	0%	1%	1%

Note:

Actual turnout measure

Values are +/- % of the baseline value

B.2 Lasso variables and interactions

Table B.3: Lasso variables and interactions

Variable	38	35	29	28	27	26	24	23	22	20	19	18	17	16	15	14	13	12
Individual-level																		
Gender					*	*	*	*	*	*	*	*	*	*	*	*	*	*
Age	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+
Education		-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	*	*
Housing	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Social grade				-	-	-	-	-	-	+	+	+	+	+	+	+	+	+
Area-level																		
Region															-	-	-	-
hanrettyLeave	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Con_2015	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+
Green_2015									-	-	-	-	-	+	+	+	+	+
Lab_2015														*	*	*	*	*
LD_2015																		*
Plaid_2015													-	-	-	-	-	
SNP_2015																		*
UKIP_PC_2015																	*	*
Age_18to24			-	-	-	-	-	-	-	-	-	+	+	+	*	*	*	*
Age_65plus														-	-	-	-	
Born_Britain																		*
Born_restOfEurope															*	*	*	*
Born_restOfWorld																		
Demos_nationalID																*	*	*
demos_nonWhiteUk														*	*	*	*	*
Edu_degree																		
Edu_noQuals																		*
Econ_seg_AB																		
Econ_seg_C1																	*	*
Econ_seg_C2																	*	*
Econ_seg_DE													*	*	*	*	*	*
Econ_seg_ABC1																	*	*
Econ_seg_C2DE																	*	*
Econ_activityRate								*	*	*	*	*	*	*	*	*	*	*
Econ_employmentRate						*	*	*	*	*	*	*	*	*	*	*	*	*
Econ_managers														*	*	*	*	*
Econ_professionals																		*
Econ_tech_3																		
Econ_administrative															*	*	*	*
Econ_skilledTrades																		*
Econ_careLeisure													*	*	*	*	*	*
Econ_routine																		*
BadHealth																		
Health_activitiesLimitedALot																		
Health_activiesLimitedAny																		
Turnout_2015															*	*	*	*
Pop_density																		
Edu_students_termTime																*	*	*
Econ_MedianHousePrice																		
Individual interactions N	0	0	0	0	3	2	3	4	4	6	7	7	7	7	7	7	7	7
Area interactions N	0	0	0	0	0	3	3	2	2	2	2	3	4	7	30	32	38	23

Note:

Model 13 and 12 failed to estimate due to rank deficiency

\- = Variable

* = Interaction

\+ = Both

Appendix C

Chapter 4

This appendix provides supplementary materials for chapter 4. In the results for the simulation study I showed estimate accuracy for marginal small areas, comparing the '3:2:1' and 'Even' sample distributions. In C.1 I provide results from the simulation study comparing the '2:1' and 'Even' distributions for the 'most' and 'least' marginal small areas.

The chapter included two real-world applications with uneven sample distributions. I provide model specification for the UK in C.2 and for the US model in C.3. I show the results from a previous version of the real-world application, where small areas were evenly divided into halves and thirds in C.4. Finally, in C.5 I present results of the two real-world applications which use model-based turnout measures rather than actual turnout used in the chapter.

C.1 Simulation results for 2:1 ratio

In chapter 4 I showed small area accuracy for the simulation study. I reported MAE and widths for small areas grouped into 'Most', 'Mid' and 'Least' marginal small areas, for all sample sizes (5, 10, 20 and 30), and small area N (50, 200, 400 and 600). In the chapter, table 4.4 showed results comparing accuracy between 'Even' and

'3:2:1' sample distributions. Below, in table C.1 I show the same table but comparing 'Even' and '2:1' sample distributions with small areas grouped into 'Most' and 'Least' marginal small areas.

Table C.1: Simulation small area accuracy (2:1 distribution)

	50 Areas		200 Areas		400 Areas		600 Areas	
Groups	MAE	Width	MAE	Width	MAE	Width	MAE	Width
Sample: 30								
Most marginal	-0.01	-0.02	-0.01	-0.03	-0.01	-0.02	-0.01	-0.02
Least marginal	0.02	0.04	0.01	0.04	0.01	0.04	0.01	0.04
Sample: 20								
Most marginal	-0.02	-0.05	-0.01	-0.03	-0.01	-0.03	-0.01	-0.04
Least marginal	0.02	0.04	0.01	0.05	0.01	0.04	0.02	0.04
Sample: 10								
Most marginal	-0.02	-0.05	-0.02	-0.04	-0.02	-0.04	-0.02	-0.04
Least marginal	0.01	0.04	0.02	0.04	0.02	0.04	0.01	0.04
Sample: 5								
Most marginal	-0.04	-0.07	-0.02	-0.08	-0.03	-0.08	-0.03	-0.08
Least marginal	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.02

Note: Showing MAE and widths for 2:1. Areas grouped into marginal categories

C.2 UK model specification

$$
\begin{aligned}
Pr(Y_i = 1) = logit^{-1}(\beta_\theta + Female \cdot \beta^{Female} + a^{Constit}_{j[i]} + \\
a^{Age}_{k[i]} + a^{Education}_{l[i]} + a^{Region}_{m[i]} + a^{Week}_{p[i]})
\end{aligned}
$$

$$
\begin{aligned}
a^{Age}_k \sim N(0, (\sigma^{Age})^2) \; for \; k = 1, ..., 8 \\
a^{Education}_l \sim N(0, (\sigma^{Education})^2) \; for \; l = 1, ..., 6 \\
a^{Region}_m \sim N(0, (\sigma^{Region})^2) \; for \; m = 1, ..., 11 \\
a^{Week}_p \sim N(0, (\sigma^{Region})^2) \; for \; p = 1, ..., 4 \\
a^{Constit}_j \sim N(a^{Region}_{m[j]} + \beta^{Con2017} \cdot Con2017 + \beta^{unem} \cdot unem + \beta^{dens} \cdot dens + \\
beta^{ind} \cdot ind + \beta^{leave} \cdot leave, (\sigma^{Constit})^2), \\
for \; j = 1, ..., 632
\end{aligned}
\tag{C.1}
$$

C.3 US Model specification

$$Pr(Y_i = 1) = logit^{-1}(\beta_\theta + Female \cdot \beta^{Female} + a^{State}_{j[i]} + a^{Age}_{k[i]} + a^{Education}_{l[i]} + a^{Region}_{m[i]} + a^{Week}_{p[i]} + a^{Ethnicity}_{q[i]} + a^{Marital}_{r[i]})$$

$$a^{Age}_k \sim N(0, (\sigma^{Age})^2) \; for \; k = 1, ..., 4$$

$$a^{Education}_l \sim N(0, (\sigma^{Education})^2) \; for \; l = 1, ..., 4$$

$$a^{Region}_m \sim N(0, (\sigma^{Region})^2) \; for \; m = 1, ..., 4$$

$$a^{Week}_p \sim N(0, (\sigma^{Week})^2) \; for \; p = 1, ..., 4$$

$$a^{Ethnicity}_q \sim N(0, (\sigma^{Ethnicity})^2) \; for \; q = 1, ..., 4$$

$$a^{Marital}_r \sim N(0, (\sigma^{Marital})^2) \; for \; r = 1, ..., 4$$

$$a^{State}_j \sim N(a^{Region}_{m[j]} + \beta^{Rep2012} \cdot Rep2012, (\sigma^{State})^2),$$

$$for \; j = 1, ..., 51$$

(C.2)

C.4 External validation - Alternative sample distribution

In chapter 4 I introduced the weighted ratio, which I argued was an efficient method to distribute the sample among small areas. It significantly increased sample size in marginal small areas while nominally decreasing sample size in non-marginal small areas. By minimising the decrease in sample size for non-marginal small areas, the accuracy in these small areas was largely unchanged. In a previous version I had divided small areas according to the procedure from the simulation study. In this version, the sample size for non-marginal small areas was much smaller, and this

resulted in decreased accuracy in the least marginal small areas. Because of this, in the chapter I argued the weighted ratio resulted in improved accuracy in marginal small areas, while having little effect on least marginal small areas accuracy.

Below I present the previous version results, where small areas were divided into equal groups of halves or thirds. In figure C.1 and C.2 I show estimates versus actual results for UK and US, respectively. In table C.2 and C.3 I show UK MAE results for small areas grouped according to marginal groupings. In C.4 and C.5 I show US MAE of small areas grouped according to marginal groups.

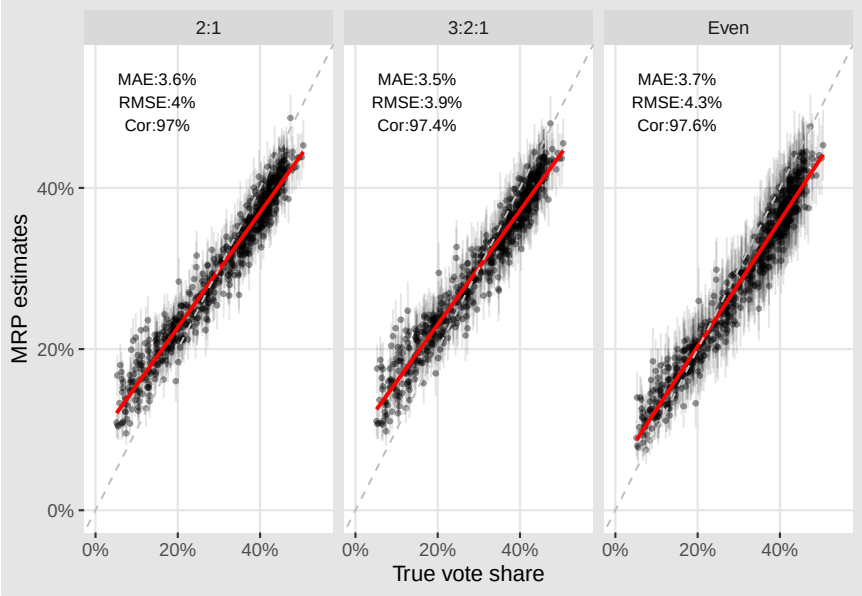

Figure C.1: UK Constituency estimates vs. true vote share

Table C.2: UK: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.029	0.067
Mid marginal	0.040	0.088
Most marginal	0.043	0.092
3:2:1		
Least marginal	0.047	0.057
Mid marginal	0.030	0.065
Most marginal	0.033	0.068

Table C.3: UK: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Most marginal	0.043	0.092
Least marginal	0.031	0.073
2:1		
Most marginal	0.033	0.061
Least marginal	0.038	0.053

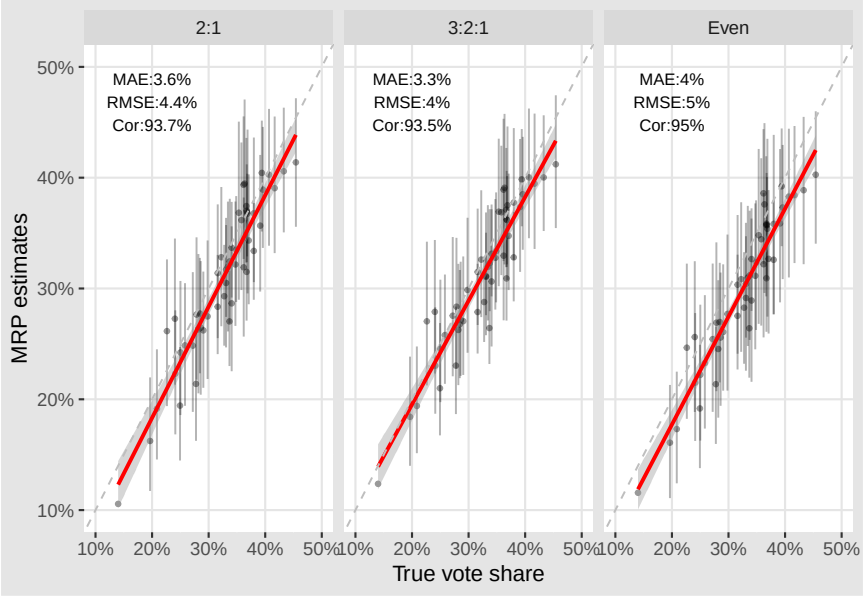

Figure C.2: US State estimates vs. true vote share

Table C.4: US: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.035	0.112
Mid marginal	0.042	0.122
Most marginal	0.043	0.131
2:1		
Least marginal	0.033	0.107
Mid marginal	0.037	0.114
Most marginal	0.038	0.118

Table C.5: US: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.037	0.114
Most marginal	0.043	0.129
2:1		
Least marginal	0.031	0.102
Most marginal	0.034	0.109

C.5 Model-based turnout results

In chapter 4 the MRP estimates were produced using actual UK constituency or US state turnout. I chose to use actual turnout because reliably estimating turnout is notoriously difficult. I argued poor turnout estimates risked blurring the correct interpretation of the results. However, it could be argued that using actual turnout, and not accounting for differential turnout among sub-groups, faces the same risk. To check chapter 4 results were not a function of using real turnout, I also produced estimates which used model-based turnout measures. These turnout measures, although somewhat unsatisfactory, allow the MRP estimates to better account for differential turnout among sub-groups. Importantly, when using the model-based turnout measures, both results and interpretation do not significantly differ. Below, I describe the methods followed to estimate turnout for the UK and US. In figure C.3, table C.6 and C.7 I show UK results with the model-based turnout. In figure C.4, table C.8 and C.9 I show US results with model-based turnout.

UK: To estimate turnout for the UK election I used the 2015 and 2017 BES validated vote data. The face-to-face survey collects recalled turnout and validates against the UK voter register. I broadly follow the procedure discussed in appendix B, which is primarily based on the method set out in Lauderdale et al. (2020). I estimated a multilevel logistic model, where Turnout = 1. I included region, age, and education as varying intercept terms. Gender, 2017 constituency turnout and 2017 constituency

majority were included as fixed effect terms. I draw 500 samples from the posterior and calculate the median turnout for each poststratification frame row. I subsequently apply turnout to vote choice estimates at each row of the poststratification frame.

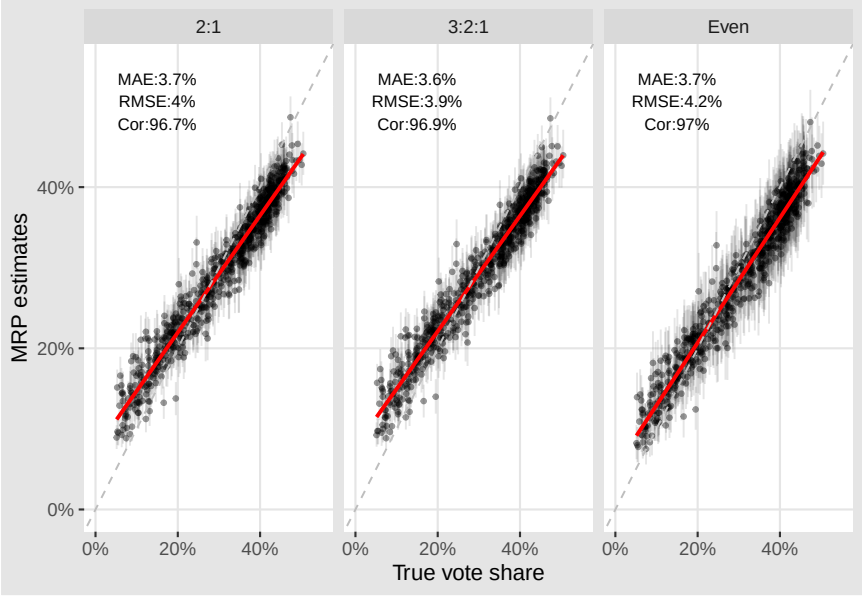

Figure C.3: UK Constituency estimates vs. true vote share

Table C.6: UK: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.036	0.079
Mid marginal	0.040	0.090
Most marginal	0.043	0.090
3:2:1		
Least marginal	0.038	0.055
Mid marginal	0.035	0.059
Most marginal	0.036	0.059

US: To estimate turnout for the US election I used the 2012 CPS data, a large N survey which captures voter turnout immediately proceeding an election. I broadly follow Lauderdale et al. (2020) for variable selection, although the estimation strategy differs. I estimated a multilevel logistic model, where Turnout = 1. I used state, age education, ethnicity, marital status as varying intercept terms, with gender, 2012

Table C.7: UK: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.036	0.079
Most marginal	0.042	0.090
2:1		
Least marginal	0.037	0.055
Most marginal	0.035	0.060

state-level turnout and 2012 state-level majority as fixed effect terms. I draw 500 samples from the posterior and calculate the median turnout for each row of the poststratification frame. However, because the CPS survey has relatively large rates of turnout over-reporting, I employ an adjustment method somewhat similar to that used by Ghitza and Gelman (2013). To calculate the adjustment factor, I divided actual state-level turnout by the unweighted state-level turnout estimate. The state-level adjustment factor was subsequently applied to turnout for each poststratification row. The adjustment was necessary because the unweighted estimates was consistently high. The weighting procedure reduced bias introduced from CPS turnout over-reporting, while maintaining the relative turnout levels between subgroups.

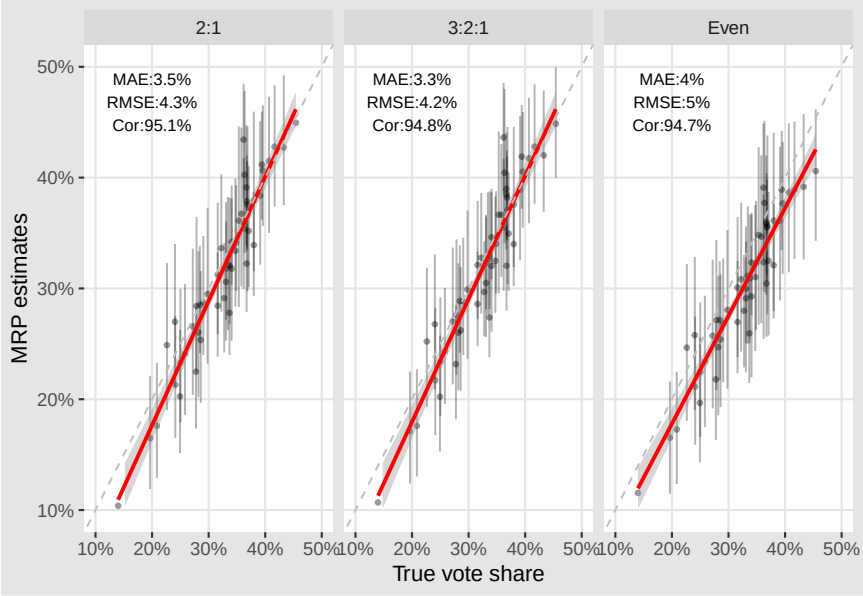

Figure C.4: US State estimates vs. true vote share

Table C.8: US: Even and 3:2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.039	0.117
Mid marginal	0.043	0.133
Most marginal	0.040	0.130
2:1		
Least marginal	0.035	0.110
Mid marginal	0.036	0.122
Most marginal	0.031	0.122

Table C.9: US: Even and 2:1 accuracy comparison

Groups	MAE	Width
Even		
Least marginal	0.039	0.117
Most marginal	0.042	0.132
2:1		
Least marginal	0.034	0.104
Most marginal	0.032	0.112

Appendix D

Chapter 5

In chapter 5, I provided model specification for both the standard and alternative MRP format. In the chapter I used the UK case to explain model specification. In D.1 I provide the full US standard MRP model, and in D.2 I provide the US alternative MRP specification.

In the results of the chapter, I showed variance in estimate widths for the UK 2019 and US 2016 models. In D.3 I show results for the UK 2017 and US 2012 models. Finally, in D.4 I provide results for the standard MRP specification with a model-based turnout measure rather than actual turnout used in the chapter.

D.1 US model specification (Standard)

$$Pr(Y_i = 1) = logit^{-1}(\beta_\theta + Female \cdot \beta^{Female} + a^{Area}_{j[i]} + a^{Age}_{k[i]} + a^{Education}_{l[i]} + a^{Region}_{m[i]} + a^{Week}_{p[i]} + a^{Ethnicity}_{q[i]} + a^{Marital}_{r[i]}), \ for\ i = 1, ..., n.$$

$$\begin{aligned} a^{Age}_k &\sim N(0, (\sigma^{Age})^2) \ for\ k = 1, ..., 4 \\ a^{Education}_l &\sim N(0, (\sigma^{Education})^2) \ for\ l = 1, ..., 4 \\ a^{Region}_m &\sim N(0, (\sigma^{Region})^2) \ for\ m = 1, ..., 4 \\ a^{Week}_p &\sim N(0, (\sigma^{Week})^2) \ for\ p = 1, ..., 4 \\ a^{Ethnicity}_q &\sim N(0, (\sigma^{Ethnicity})^2) \ for\ q = 1, ..., 4 \\ a^{Marital}_r &\sim N(0, (\sigma^{Marital})^2) \ for\ r = 1, ..., 4 \\ a^{Area}_j &\sim N(a^{Region}_{m[j]} + \beta^{Past-Dem} \cdot Past - Dem, (\sigma^{Area})^2), \\ & \qquad for\ j = 1, ..., 51 \end{aligned} \tag{D.1}$$

D.2 US model specification (Alternative)

$$\begin{aligned} Pr(Y_i = 1) = logit^{-1}(\beta_\theta + a^{Area}_{j[i]} + \beta^{Female} \cdot Female + \beta^{Age} \cdot \boldsymbol{X}^{Age} + \\ \beta^{Education} \cdot \boldsymbol{X}^{Education} + \beta^{Region} \cdot \boldsymbol{X}^{Region} + \\ \beta^{Ethnicity} \cdot \boldsymbol{X}^{Ethnicity} + \beta^{Marital} \cdot \boldsymbol{X}^{Marital}, \\ for\ i = 1, ..., n. \end{aligned} \tag{D.2}$$

$$a^{Area}_j \sim N(a^{Region}_{m[j]} + \beta^{Past-Dem} \cdot Past - Dem, (\sigma^{Area})^2), \quad for\ j = 1, ..., 51$$

D.3 Estimate precision

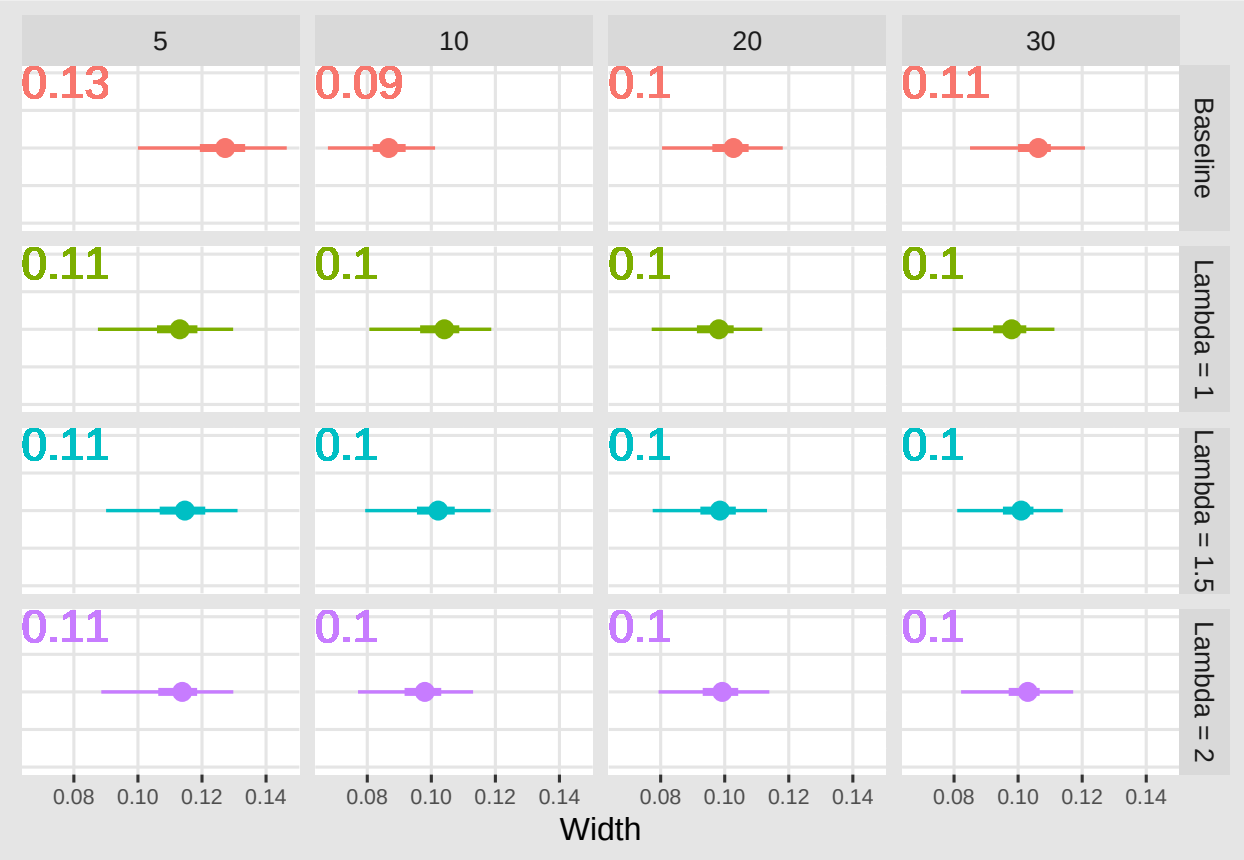

Figure D.1: UK (2017) average widths

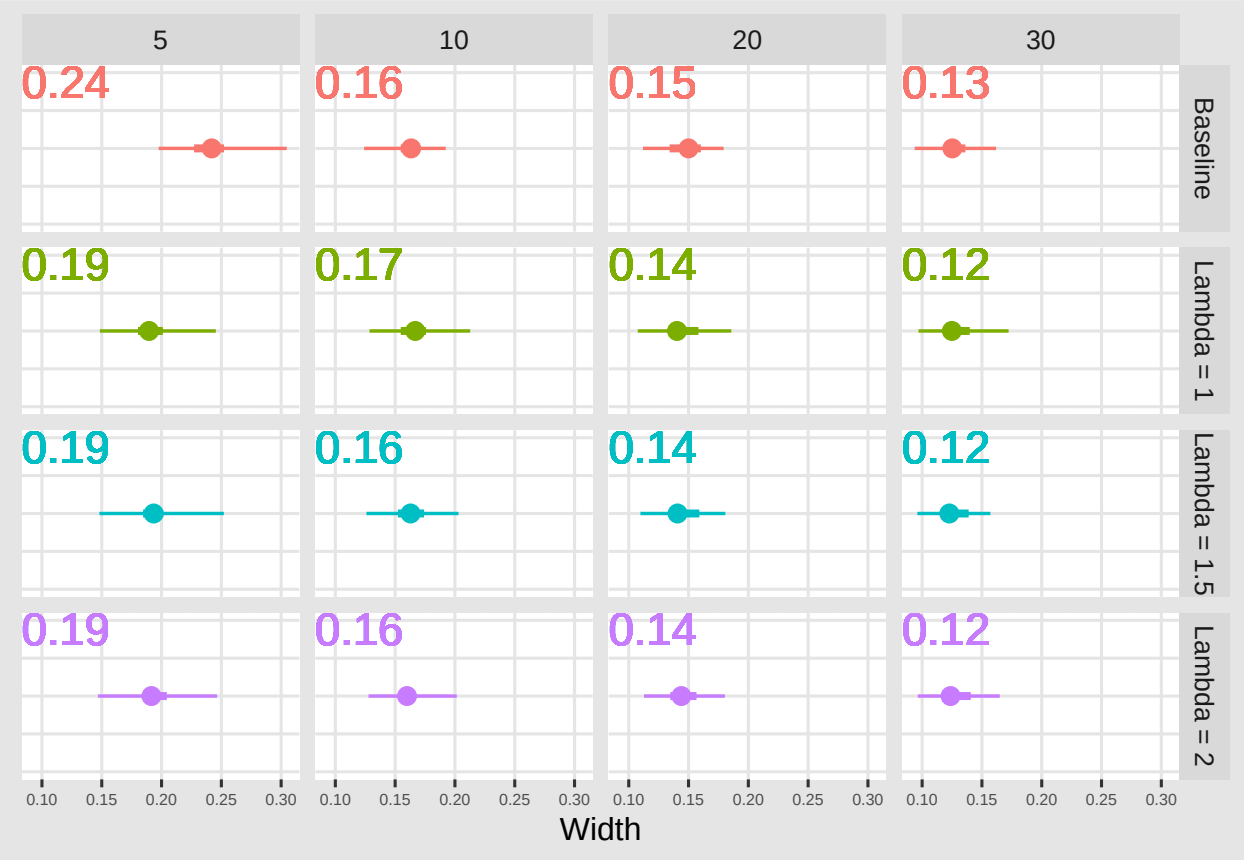

Figure D.2: US (2012) average widths

D.4 Model-based turnout results

In chapter 5 I used real turnout to produce estimates for both UK and US elections. As discussed in the chapter, this meant that the results did not take into account deferential turnout among subgroups. In order to demonstrate that using actual turnout did not affect the results nor the interpretation of the results, I produced estimates for the same elections, but use a model-based turnout measure. Below, I describe the methods used to estimate the model-based turnout measures, and subsequently show UK and US results with the alternative turnout measure.

For the UK, I used the same 2017 turnout measure estimated for chapter 4 and described in appendix C.5. For the 2019 election, I followed an identical estimation procedure to that used to estimate 2017. I used the 2015 and 2017 BES validated

vote surveys, and estimated a multilevel logistic model, where Turnout = 1, and did not vote = 0. I include varying intercept terms for age, education, and region, while 2017 turnout and 2017 winning party majority are included as area-level variables. The model estimates were poststratified to each row of the poststratification frame, taking the median of 500 posterior draws.

For the US, I estimated turnout for both the 2012 and 2016 elections, replicating the method used and explained in C.5. I use the CPS survey (2012 and 2016), model turnout with a Bayesian multilevel logistic regression model with age, education, state, ethnicity, marital status as varying intercept terms. Gender, past state turnout and past state-level majority are all included as β terms. I adjust the turnout estimates according to known state-level turnout, the same procedure explained in C.5. Finally, I poststratify the estimates to each row of the poststratification frame, taking the median of 500 posterior draws. Turnout is applied to vote choice estimates at each row of the poststratification frame. Overall, each measure over-estimated turnout, and produced estimates equal or worse than those presented in the chapter. However, unlike the turnout used in the chapter, the measures were able to capture differential turnout among sub-groups. Importantly, the interpretation of the results do not change when the model-based turnout measure is used. Below I show results identical to tables 5.4 and 5.5 presented in chapter 5. The tables show the increase or decrease in MAE from the baseline for all sample sizes and for all elections.

Table D.1: Standard MRP accuracy (UK)

	Sample			
Model	5	10	20	30
2017				
Lambda = 1	-0.2%	0.1%	0%	-0.1%
Lambda = 1.5	-0.2%	0.1%	0%	-0.1%
Lambda = 2	-0.2%	0.1%	0%	0%
2019				
Lambda = 1	-0.6%	0%	0%	0%
Lambda = 1.5	-0.6%	-0.1%	-0.1%	0%
Lambda = 2	-0.6%	-0.1%	0%	0%

Note: showing inf. prior MAE as +/- from baseline.

Table D.2: Standard MRP accuracy (US)

	Sample			
Model	5	10	20	30
2012				
Lambda = 1	-1.3%	0.4%	0.2%	0.2%
Lambda = 1.5	-1.2%	0.2%	0.1%	0.1%
Lambda = 2	-1.3%	0.1%	0%	0.1%
2016				
Lambda = 1	-0.9%	-0.3%	-0.3%	-0.1%
Lambda = 1.5	-0.8%	-0.3%	-0.2%	-0.1%
Lambda = 2	-0.7%	-0.3%	-0.1%	0%

Note: showing inf. prior MAE as +/- from baseline.

References

Alexander, R. (2019) *Getting started with MRP*. [online]. Available from: `https://rohanalexander.com/posts/2019-12-04-getting_started_with_mrp/`.

Bell, B. A. et al. (2008) Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM proceedings, section on survey research methods*. 1122–1129.

Bell, B. A. et al. (2014) How Low Can You Go?: An Investigation of the Influence of Sample Size and Model Complexity on Point and Interval Estimates in Two-Level Linear Models. *Methodology*. 10 (1), 1–11.

Berkman, M. B. & O'Connor, R. E. (1993) Do Women Legislators Matter?: Female Legislators and State Abortion Policy. *American Politics Quarterly*. 21 (1), 102–124.

Berry, W. D. et al. (1998) Measuring Citizen and Government Ideology in the American States, 1960-93. *American Journal of Political Science*. 42 (1), 327.

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

Brace, P. et al. (2004) Does State Political Ideology Change over Time? *Political Research Quarterly*. 57 (4), 529–540.

Brambor, T. et al. (2006) Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*. 14 (1), 63–82.

Breiman, L. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics*. 37 (4), 373–384.

Breiman, L. (ed.) (1998) *Classification and regression trees*. 1. CRC Press repr. Boca Raton, Fla.: Chapman & Hall/CRC.

Breukelen, G. J. P. van et al. (2007) Relative efficiency of unequalversus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*. 26 (13), 2589–2603.

Broniecki, P. et al. (2021) Improved Multilevel Regression with Post-Stratification Through Machine Learning (autoMrP). *The Journal of Politics*. 714777.

Buttice, M. K. & Highton, B. (2013) How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? *Political Analysis*. 21 (4), 449–467.

Butz, A. M. & Kehrberg, J. E. (2016) Estimating anti-immigrant sentiment for the American states using multi-level modeling and post-stratification, 2004-2008. *Research & Politics*. 3 (2), 1–7.

Butz, A. M. & Kehrberg, J. E. (2015) Social Distrust and Immigrant Access to Welfare Programs in the American States: Social Distrust and Immigrant Access. *Politics & Policy*. 43 (2), 256–286.

Bürkner, P.-C. (2017) **Brms** : An *r* Package for Bayesian Multilevel Models Using *stan*. *Journal of Statistical Software*. 80 (1),.

Candel, M. J. J. M. et al. (2007) Optimality of Equal vs. Unequal Cluster Sizes in Multilevel Intervention Studies: A Monte Carlo Study for Small Sample Sizes. *Communications in Statistics - Simulation and Computation*. 37 (1), 222–239.

Caughey, D. & Warshaw, C. (2019) Public Opinion in Subnational Politics. *The Journal of Politics*. 81 (1), 352–363.

Cerina, R. & Duch, R. (2020a) Measuring public opinion via digital footprints. *International Journal of Forecasting*. 36 (3), 987–1002.

Cerina, R. & Duch, R. (2020b) Polling India via Regression and Post-Stratification of Non-Probability Online Samples. *Submitted*.

Chen, M.-H. & Ibrahim, J. G. (2000) Power prior distributions for regression models. *Statistical Science*. 15 (1),.

Clarke, P. (2008) When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*. 62 (8), 752–758.

Clarke, P. & Wheaton, B. (2007) Addressing Data Sparseness in Contextual Population Research: Using Cluster Analysis to Create Synthetic Neighborhoods. *Sociological Methods & Research*. 35 (3), 311–351.

Cohen, M. P. (2005) Sample Size Considerations for Multilevel Surveys. *International Statistical Review*. 73 (3), 279–287.

Consonni, G. et al. (2018) Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*. 13 (2), 627–679.

Curtice, J. (2020) Brave New World: Understanding the 2019 General Election. *Political Insight*. 11 (1), 8–12.

Cutts, D. et al. (2020) Brexit, the 2019 General Election and the Realignment of British Politics. *The Political Quarterly*. 91 (1), 7–23.

Depaoli, S. et al. (2020) The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. *Frontiers in Psychology*. 11608045.

Depaoli, S. & Schoot, R. van de (2017) Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*. 22 (2), 240–261.

Desboulets, L. (2018) A Review on Variable Selection in Regression Analysis. *Econometrics*. 6 (4), 45.

Downes, M. et al. (2018) Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples. *American Journal of Epidemiology*. 187 (8), 1780–1790.

Downes, M. & Carlin, J. B. (2020a) 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.

Downes, M. & Carlin, J. B. (2020b) Multilevel regression and poststratification for estimating population quantities from large health studies: A simulation study based on US population structure. *Journal of Epidemiology and Community Health*. jech-2020-214346.

Downes, M. & Carlin, J. B. (2020c) Multilevel Regression and Poststratification Versus Survey Sample Weighting for Estimating Population Quantities in Large Population Health Studies. *American Journal of Epidemiology*. 189 (7), 717–725.

Dunham, J. (2018) *Dgo: Dynamic Estimation of Group-Level Opinion*. [online]. Available from: `https://github.com/jamesdunham/dgo`.

Enns, P. K. & Koch, J. (2013) Public Opinion in the U.S. States: 1956 to 2010. *State Politics & Policy Quarterly*. 13 (3), 349–372.

Erikson, R. S. et al. (1994) *Statehouse Democracy Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.

Eun Kim, S. & Urpelainen, J. (2018) Environmental public opinion in U.S. States, 1973-2012. *Environmental Politics*. 27 (1), 89–114.

Faliszewski, P. et al. (2019) How Similar Are Two Elections? *Proceedings of the AAAI Conference on Artificial Intelligence*. 331909–1916.

Faliszewski, P. et al. (2020) 'Isomorphic Distances Among Elections', in Henning Fernau (ed.) *Computer Science – Theory and Applications*. [Online]. Cham: Springer International Publishing. pp. 64–78.

Flinders, M. (2020) Not a Brexit Election? Pessimism, Promises and Populism "UK-Style". *Parliamentary Affairs*. 73 (Supplement_1), 225–242.

Fowler, L. (2016) The states of public opinion on the environment. *Environmental Politics*. 25 (2), 315–337.

Fowler, L. (2017) Tracking state trends in environmental public opinion. *The Social Science Journal*. 54 (3), 287–294.

Friedman, J. et al. (2010) A note on the group lasso and a sparse group lasso. *arXiv:1001.0736 [math, stat]*.

Gabry, J. et al. (2019) Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 182 (2), 389–402.

Galston, W. (2013) *The 2012 Election: What Happened, What Changed, What it Means*.

Gao, Y. et al. (2021) Improving Multilevel Regression and Poststratification with Structured Priors. *Bayesian Analysis*. -1 (-1),.

García-Pérez, M. (2019) Bayesian Estimation with Informative Priors is Indistinguishable from Data Falsification. *The Spanish Journal of Psychology*. 22 (45),.

Gelman, A. et al. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*. 2 (4), 1360–1383.

Gelman, A. (2014) How Bayesian Analysis Cracked the Red-State, Blue-State Problem. *Statistical Science*. 29 (1),.

Gelman, A. (2004) Parameterization and Bayesian Modeling. *Journal of the American Statistical Association*. 99 (466), 537–545.

Gelman, A. et al. (1996) Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association*. 91 (436), 1400–1412.

Gelman, A. (2009) Prior distributions for Bayesian data analysis in political science. *Unpublished manuscript*.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. 1 (3), 515–534.

Gelman, A. et al. (2010) Public Opinion on Health Care Reform. *The Forum*. 8 (1), 1–14.

Gelman, A. et al. (2020) *Regression and Other Stories*. 1st edition. [Online]. Cambridge University Press.

Gelman, A. (2007) Struggles with Survey Weighting and Regression Modeling. *Statistical Science*. 22 (2),.

Gelman, A. et al. (2017) The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*. 19 (10), 555.

Gelman, A. et al. (2016) Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion. *Unpublished manuscript*.

Gelman, A. & Hennig, C. (2015) Beyond subjective and objective in statistics. *arXiv:1508.05453 [stat]*.

Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*. Analytical methods for social research. Cambridge ; New York: Cambridge University Press.

Gelman, A. & Little, T. (1997) Poststratification Into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*. 23 (2), 127–135.

George, E. I. (2000) The Variable Selection Problem. *Journal of the American Statistical Association*. 95 (452), 1304–1308.

Ghitza, Y. & Gelman, A. (2013) Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups: DEEP INTERACTIONS WITH MRP. *American Journal of Political Science*. 57 (3), 762–776.

Golchi, S. (2019) Informative priors in Bayesian inference and computation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 12 (2), 45–55.

Goodrich, B. et al. (2020) Rstanarm: Bayesian applied regression modeling via Stan. *R package version 2.21.1.* [online]. Available from: `https://mc-stan.org/rstanarm`.

Goodwin, M. & Heath, O. (2019) *Low-income voters in UK general elections, 1987-2017.*

Goplerud, M. et al. (2018) Sparse Multilevel Regression (and Poststratification (sMRP)). *Working paper.*

Green, K. C. & Armstrong, J. S. (2015) Simple versus complex forecasting: The evidence. *Journal of Business Research.* 68 (8), 1678–1685.

Grogan, C. M. & Park, S. (Ethan). (2017) The Racial Divide in State Medicaid Expansions. *Journal of Health Politics, Policy and Law.* 42 (3), 539–572.

Groves, R. M. (1987) Research on Survey Data Quality. *Public Opinion Quarterly.* 51 (part 2: Supplement: 50th Anniversary Issue), S156.

Grzenda, W. (2016) Informative Versus Non-Informative Prior Distributions and Their Impact on the Accuracy of Bayesian Inference. *Statistics in Transition.* 17 (4), 763–780.

Hanretty, C. (2019) An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion. *Political Studies Review.* 147892991986477.

Hanretty, C. et al. (2016) Comparing Strategies for Estimating Constituency Opinion from National Survey Samples. *Political Science Research and Methods.* 6 (3), 571–591.

Hanretty, C. et al. (2017) Dyadic Representation in a Westminster System: Dyadic Representation in a Westminster System. *Legislative Studies Quarterly.* 42 (2), 235–267.

Hastie, T. et al. (2009) *The elements of statistical learning: Data mining, inference, and prediction.* Springer series in statistics. 2nd ed. New York, NY: Springer.

Heath, O. & Goodwin, M. (2017) The 2017 General Election, Brexit and the Return to Two-Party Politics: An Aggregate-Level Analysis of the Result. *The Political Quarterly*. 88 (3), 345–358.

Heinze, G. et al. (2018) Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*. 60 (3), 431–449.

Heinze, G. & Dunkler, D. (2017) Five myths about variable selection. *Transplant International*. 30 (1), 6–10.

Henderson, D. & Denison, D. (1989) Stepwise Regression in Social and Psychological Research. *Psychological Reports*. 64251–257.

Heo, M. & Leon, A. C. (2005) Performance of a Mixed Effects Logistic Regression Model for Binary Outcomes With Unequal Cluster Size. *Journal of Biopharmaceutical Statistics*. 15 (3), 513–526.

Hersh, E. D. & Nall, C. (2016) The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records: RACE AND INCOME-BASED VOTING. *American Journal of Political Science*. 60 (2), 289–303.

Hill, K. Q. & Hurley, P. A. (1999) Dyadic Representation Reappraised. *American Journal of Political Science*. 43 (1), 109.

Hindman, M. (2015) Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The Annals of the American Academy of Political and Social Science*. 65948–62.

Hobolt, S. B. (2018) Brexit and the 2017 UK General Election: Brexit and the 2017 UK General Election. *JCMS: Journal of Common Market Studies*. 5639–50.

Hoffman, M. D. & Gelman, A. (2011) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv*.

Howe, P. D. et al. (2015) Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change*. 5 (6), 596–603.

Howe, P. D. (2018) Modeling Geographic Variation in Household Disaster Preparedness across U.S. States and Metropolitan Areas. *The Professional Geographer*. 70 (3), 491–503.

Hox, J. J. (2010) *Multilevel analysis: Techniques and applications*. Second edition. New York, NY: Routledge.

Hurvich, C. M. & Tsai, C.-L. (1990) The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*. 44 (3), 214.

Ibrahim, J. G. et al. (2015) The power prior: Theory and applications. *Statistics in Medicine*. 34 (28), 3724–3749.

James, G. et al. (eds.) (2013) OCLC: ocn828488009. *An introduction to statistical learning: With applications in R*. Springer texts in statistics 103. New York: Springer.

Jaynes, E. T. (1985) Highly Informative Priors. *Bayesian Statistics*. 2329–360.

Jennings, W. & Stoker, G. (2017) Tilting Towards the Cosmopolitan Axis? Political Change in England and the 2017 General Election. *The Political Quarterly*. 88 (3), 359–369.

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

Johnston, R., Rossiter, D., Manley, D., et al. (2018) Coming full circle: The 2017 UK general election and the changing electoral map. *The Geographical Journal*. 184 (1), 100–108.

Johnston, R., Rossiter, D., Hartman, T., et al. (2018) Exploring constituency-level estimates for the 2017 British general election. *International Journal of Market Research*. 60 (5), 463–483.

Jones, K. et al. (1992) People, Places and Regions: Exploring the Use of Multi-Level Modelling in the Analysis ofElectoral Data. *British Journal of Political Science*. 22 (3), 343–380.

Kastellec, J. P. et al. (2016) Estimating State Public Opinion with Multi-level Regression and Poststratification using R. *Working paper*.

Kastellec, J. P. et al. (2010) Public Opinion and Senate Confirmation of Supreme Court Nominees. *The Journal of Politics*. 72 (3), 767–784.

Kaufmann, R. K. et al. (2017) Spatial heterogeneity of climate change as an experiential basis for skepticism. *Proceedings of the National Academy of Sciences*. 114 (1), 67–71.

Kennedy, L. & Gabry, J. (2020) *MRP with rstanarm*.

Kennedy, L. & Gelman, A. (2020) Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *arXiv*.

Kiewiet de Jonge, C. P. et al. (2018) Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP). *Public Opinion Quarterly*. 82 (3), 419–446.

Koch, J. & Thomsen, D. M. (2017) Gender Equality Mood across States and over Time. *State Politics & Policy Quarterly*. 17 (4), 351–360.

Kolczynska, M. et al. (2020) *Modeling public opinion over time and space: Trust in state institutions in Europe, 1989-2019*.

Kreft, I. G. (1996) Are Multilevel Techniques Necessary?: An overview, including Simulation Studies. *Unpublished manuscript*.

Krimmel, K. et al. (2011) Public Opinion and Gay Rights: Do Members of Congress Follow Their Constituents' Preferences? *Unpublished*.

Krstajic, D. et al. (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*. 6 (1), 10.

Lauderdale, B. E. et al. (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.

Lax, J. R. & Phillips, J. H. (2009a) Gay Rights in the States: Public Opinion and Policy Responsiveness. *American Political Science Review*. 103 (3), 367–386.

Lax, J. R. & Phillips, J. H. (2009b) 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. *Unpublished manuscript*.

Lax, J. R. & Phillips, J. H. (2012) The Democratic Deficit in the States. *American Journal of Political Science*. 56 (1), 148–166.

Lee, M. D. & Vanpaemel, W. (2018) Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*. 25 (1), 114–127.

Leemann, L. & Wasserfallen, F. (2017) Extending the Use and Prediction Precision of Subnational Public Opinion Estimation: Extending use and Precision of MRP. *American Journal of Political Science*. 61 (4), 1003–1022.

Leemann, L. & Wasserfallen, F. (2018) *MrP Illustration*. [online]. Available from: `https://github.com/lleemann/MrP_chapter/blob/master/MrP_Illsutration.pdf`.

Leemann, L. & Wasserfallen, F. (2016) The Democratic Effect of Direct Democracy. *American Political Science Review*. 110 (4), 750–762.

Lei, R. et al. (2017) The 2008 Election: A Preregistered Replication Analysis. *Statistics and Public Policy*. 4 (1), 1–8.

Lemoine, N. P. (2019) Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*. 128 (7), 912–928.

Lenk, P. & Orme, B. (2009) The Value of Informative Priors in Bayesian Inference with Sparse Data. *Journal of Marketing Research*. 46 (6), 832–845.

Lewis, D. C. & Jacobsmeier, M. L. (2017) Evaluating Policy Representation with Dynamic MRP Estimates: Direct Democracy and Same-Sex Relationship Policies in the United States. *State Politics & Policy Quarterly*. 17 (4), 441–464.

Lim, M. & Hastie, T. (2013) Learning interactions through hierarchical group-lasso regularization. *arXiv*.

Lim, M. & Hastie, T. (2015) Learning Interactions via Hierarchical Group-Lasso Regularization. *Journal of Computational and Graphical Statistics*. 24 (3), 627–654.

Lipps, J. & Schraff, D. (2021) Estimating subnational preferences across the European Union. *Political Science Research and Methods*. 9 (1), 197–205.

Lokhorst, J. (1999) The lasso and generalized linear models. *Honors Project*.

Lopez-Martin, J. et al. (2019) *Multilevel Regression and Poststratification Case Studies*.

Maas, C. J. M. & Hox, J. J. (2005) Sufficient Sample Sizes for Multilevel Modeling. *Methodology*. 1 (3), 86–92.

Mastny, T. (2017) *MRP Using brms and tidybayes*. [online]. Available from: `https://timmastny.rbind.io/blog/multilevel-mrp-tidybayes-brms-stan/`.

McElreath, R. (2020) *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC texts in statistical science. 2nd edition. Boca Raton: Taylor; Francis, CRC Press.

McNeish, D. M. & Stapleton, L. M. (2016) The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*. 28 (2), 295–314.

Meier, L. et al. (2008) The group lasso for logistic regression: Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 70 (1), 53–71.

Mellon, J. et al. (2018) Brexit or Corbyn? Campaign and Inter-Election Vote Switching in the 2017 UK General Election. *Parliamentary Affairs*. 71 (4), 719–737.

Mildenberger, M. et al. (2016) The Distribution of Climate Change Public Opinion in Canada Henrik Österblom (ed.). *PLOS ONE*. 11 (8), e0159774.

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

Milliren, C. E. et al. (2018) Does an uneven sample size distribution across settings matter in cross-classified multilevel modeling? Results of a simulation study. *Health & Place*. 52121–126.

Morgan, S. & Lee, J. (2018) Trump Voters and the White Working Class. *Sociological Science*. 5234–245.

Muller, C. & Schrage, D. (2014) Mass Imprisonment and Trust in the Law. *The ANNALS of the American Academy of Political and Social Science*. 651 (1), 139–158.

Norrander, B. & Wilcox, C. (1999) Public Opinion and Policymaking in the States: The Case of Post-Roe Abortion Policy. *Policy Studies Journal*. 27 (4), 707–722.

O'Hagan, A. (2019) Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*. 73 (sup1), 69–81.

O'Hagan, A. (ed.) (2006) *Uncertain judgements: Eliciting experts' probabilities*. Statistics in practice. London ; Hoboken, NJ: John Wiley & Sons.

Oravecz, Z. et al. (2016) 'Sequential Bayesian updating for big data', in Michael Jones (ed.) *Big Data in Cognitive Science*. Psychology Press.

Ornstein, J. (2017) Subnational Public Opinion Estimation Using MrsP. *Unpublished manuscript*.

Ornstein, J. T. (2020) Stacked Regression and Poststratification. *Political Analysis*. 28 (2), 293–301.

Osborne, M. (2000) A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*. 20 (3), 389–403.

Pacheco, J. (2014) Measuring and Evaluating Changes in State Opinion Across Eight Issues. *American Politics Research*. 42 (6), 986–1009.

Pacheco, J. (2012) The Social Contagion Model: Exploring the Role of Public Opinion on the Diffusion of Antismoking Legislation across the American States. *The Journal of Politics*. 74 (1), 187–202.

Pacheco, J. (2013) The Thermostatic Model of Responsiveness in the American States. *State Politics & Policy Quarterly*. 13 (3), 306–332.

Pacheco, J. (2011) Using National Surveys to Measure Dynamic U.S. State Public Opinion: A Guideline for Scholars and an Application. *State Politics & Policy Quarterly*. 11 (4), 415–439.

Pacheco, J. & Maltby, E. (2017) The Role of Public Opinion - Does It Influence the Diffusion of ACA Decisions? *Journal of Health Politics, Policy and Law*. 42 (2), 309–340.

Pacheco, J. & Maltby, E. (2019) Trends in State-Level Opinions toward the Affordable Care Act. *Journal of Health Politics, Policy and Law*. 44 (5), 737–764.

Park, D. K. et al. (2004) Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*. 12 (4), 375–385.

Pfeffermann, D. (2013) New Important Developments in Small Area Estimation. *Statistical Science*. 28 (1),.

Pool, I. de S. et al. (1965) *Candidates, Issues, and Strategies A Computer Simulation of the 1960 Presidential Election*. Cambridge: MA: MIT Press.

Prosser, C. (2021) The end of the EU affair: The UK general election of 2019. *West European Politics*. 44 (2), 450–461.

Rao, J. N. K. & Molina, I. (2015) *Small Area Estimation: Rao/Small Area Estimation*. [Online]. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Ratkovic, M. & Tingley, D. (2017) Sparse Estimation and Uncertainty with Application to Subgroup Analysis. *Political Analysis*. 25 (1), 1–40.

Ratner, B. (2010) Variable selection methods in regression: Ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing*.

18 (1), 65–75.

Rivers, D. (2018) *AAPOR Short Course on MRP*. [online]. Available from: `https://github.com/rdrivers/mrp-aapor`.

Rodden, J. (2010) The Geographic Distribution of Political Preferences. *Annual Review of Political Science*. 13321–340.

Salvatier, J. et al. (2016) Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*. 2e55.

Schoeneberger, J. A. (2016) The Impact of Sample Size and Other Factors When Estimating Multilevel Logistic Models. *The Journal of Experimental Education*. 84 (2), 373–397.

Schoot, R. van de et al. (2021) Bayesian statistics and modelling. *Nature Reviews Methods Primers*. 1 (1), 1.

Selb, P. & Munzert, S. (2011) Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information. *Political Analysis*. 19 (4), 455–470.

Sides, J. et al. (2017) How Trump Lost and Won. *Journal of Democracy*. 28 (2), 34–44.

Simon, N. et al. (2013) A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*. 22 (2), 231–245.

Simpson, D. P. et al. (2015) Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv:1403.4630 [stat]*.

Singh, M. (2016) 'The failure of a generation: The polling debacle of 2015.', in Philip Cowley & Robert Anthony Ford (eds.) *More sex, lies & the ballot box: Another 50 things you need to know about elections*. London: Biteback Publishing.

Snijders, T. A. B. & Bosker, R. J. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Second edition. London: Sage.

Steyerberg, E. (1999) Stepwise Selection in Small Data Sets A Simulation Study of Bias in Logistic Regression Analysis. *Journal of Clinical Epidemiology*. 52 (10),

935–942.

Stockemer, D. (2017) What Affects Voter Turnout? A Review Article/Meta-Analysis of Aggregate Research. *Government and Opposition*. 52 (4), 698–722.

Tausanovitch, C. & Warshaw, C. (2013) Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities. *The Journal of Politics*. 75 (2), 330–342.

Tausanovitch, C. & Warshaw, C. (2014) Representation in Municipal Government. *American Political Science Review*. 108 (3), 605–641.

Team, R. C. (2020) *R: A Language and Environment for Statistical Computing*. [online]. Available from: `https://www.R-project.org/`.

Team, S. D. (2020) *RStan: The R interface to Stan*. [online]. Available from: `http://mc-stan.org/`.

Theall, K. P. et al. (2011) Impact of small group size on neighbourhood influences in multilevel models. *Journal of Epidemiology & Community Health*. 65 (8), 688–695.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* 58 (1), 267–288.

Tibshirani, R. (2011) Regression Shrinkage and Selection via the Lasso: A Retrospective. *Journal of the Royal Statistical Society.* 73 (3), 273–282.

Toshkov, D. (2015) Exploring the Performance of Multilevel Modeling and Poststratification with Eurobarometer Data. *Political Analysis*. 23 (3), 455–460.

Tran, D. et al. (2017) Edward: A library for probabilistic modeling, inference, and criticism. *arXiv*.

Tudor, J. & Wall, M. (2021) A moving target? An analysis of the impact of electoral context on polling error variation in both British and international general elections. *Journal of Elections, Public Opinion and Parties*. 1–23.

Tufte, E. R. (1975) Determinants of the Outcomes of Midterm Congressional Elections. *American Political Science Review*. 69 (3), 812–826.

Vaccari, C. et al. (2020) The United Kingdom 2017 election: Polarisation in a split issue space. *West European Politics*. 43 (3), 587–609.

Wang, W. et al. (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.

Weber, R. et al. (1972) Computer Simulation of State Electorates. *The Public Opinion Quarterly*. 36 (4), 549–565.

Weisberg, H. F. (2015) The decline in the white vote for Barack Obama in 2012: Racial attitudes or the economy? *Electoral Studies*. 40449–459.

Weissberg, R. (1978) Collective vs. Dyadic Representation in Congress. *American Political Science Review*. 72 (2), 535–547.

Wesel, F. V. et al. (2011) Choosing Priors for Constrained Analysis of Variance: Methods Based on Training Data: Choosing priors for constrained ANOVA. *Scandinavian Journal of Statistics*. 38 (4), 666–690.

Wickham, H. et al. (2019) Welcome to the Tidyverse. *Journal of Open Source Software*. 4 (43), 1686.

Williams, A. M. et al. (2018) The migration intentions of young adults in Europe: A comparative, multilevel analysis. *Population, Space and Place*. 24 (1), e2123.

Williams, R. (2018) *Multilevel Regression with Poststratification*. [online]. Available from: `https://jayrobwilliams.com/files/html/teaching-materials/MRP#`.

Wyatt, J. et al. (2016) Estimating General Election Support for President Using Multilevel Regression & Poststratification (MRP). *Morning consult*.

Yu, R. & Abdel-Aty, M. (2013) Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis & Prevention*. 5651–58.

Yuan, M. & Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 68 (1), 49–67.

Zellner, A. et al. (eds.) (2002) *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. 1st edition. [Online]. Cambridge University Press.

Zhang, X. et al. (2014) Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*. 179 (8), 1025–1033.

Zondervan-Zwijnenburg, M. et al. (2017) Where Do Priors Come From? Applying Guidelines to Construct Informative Priors in Small Sample Research. *Research in Human Development*. 14 (4), 305–320.

Zou, H. et al. (2007) On the 'degrees of freedom' of the lasso. *The Annals of Statistics*. 35 (5),.

Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*. 101 (476), 1418–1429.

Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 67 (2), 301–320.