Thesis title

A thesis submitted for the degree of

Master of Business Analytics

by

Dewi Lestari Amaliah

31251587

Department of Econometrics and Business Statistics

Monash University

Australia

October 2021

Contents

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Acknowledgements

I would like to thank my pet goldfish for …

Chapter 1

Introduction

Multilevel regression and poststratification, henceforth referred to as MRP, is a model-based approach to estimate population. In short, MRP incorporates a multilevel regression technique to predict the outcome of interest using a survey data. This prediction is then poststratified using the population size from a larger survey or census to get the population estimates.

According to Lopez-Martin, Phillips, and Gelman (2021), MRP is widely applied to do small area estimation in the absence of a subnational survey. In this case, MRP is able to reliably estimate small area statistics, i.e., small geographic areas, such as state or county estimates. MRP also allows the demographic-wise estimation, such as estimation by gender, age group, and education. Hence, it is useful, especially because small area statistics are often essential in policy-making, but conducting a subnational survey is often expensive and difficult. Additionally, MRP is also often applied to adjust the estimation from the non-representative survey as a result of the difficulties in recruiting representative survey respondents (Lopez-Martin, Phillips, and Gelman, 2021).

The standard method to communicate and validate the MRP estimates, such as their accuracy, is by using graphics. Indeed, statistical graphics are deemed a powerful tool to communicate quantitative information and analyse data (Cleveland, 1985); (Chambers, 1983). Wickham, Cook, and Hofmann (2015) state that statistical visualisation, particularly model visualisation, is imperative as it helps us to understand the model better, for

example, how the model changes as its parameters change or how the parameters change as the data change. They also mention that model visualisation is important to show the model's goodness of fit and whether it is good for some regions only and worse in other regions, or it is uniformly good.

While visualisation is common to communicate and diagnose MRP models, only a few discussions and studies about it, two of the studies that nearly intersect with this topic were conducted by Makela, Si, and Gelman (2017) and Schneider and Jacoby (2017). However, Makela, Si, and Gelman (2017) only focus on a graphical method for discovery and communication purposes of polling results. Besides, the MRP visualisations that they display as examples are isolated on Gelman's previous papers only. Meanwhile, the latter study by Schneider and Jacoby (2017) only focuses on how the graphics in public opinion research should be displayed. Therefore, this study tries to fill the gap by discussing the current practices of MRP visualisations generally, not only in public opinion and polling estimates scopes. It also aims to explore the possible improvements of those current practices.

Explicitly, the objectives of this study are:

1. Discuss the current practice of visualisation of MRP models.
2. Understand the implication of existing visualisation choices with real-world data.
3. Explore possible improvements of the current practice of MRP visualisation.

The first objectives will be reached by doing a systematic literature review, while the second and the third goals will be demonstrated through a case study.

1.1 MRP Overview

MRP is essentially conducted with two stages, i.e., multilevel/hierarchical regression modeling and poststratification. The idea is to combine model-based estimation commonly used in small area estimation with poststratification, which is considered the general framework as a weighting scheme in survey analysis (Gelman and Little, 1997). They

further argue that using multilevel regression estimates for poststratification allows the estimation for many more categories to gain more detailed population information.

Formally, let K be the number of categorical variables in the population and the k_{th} has J_k categories/levels, the population can be then expressed as $J = \prod_{k=1}^{K} J_k$ cells. For every cell, there is a known population size N_j. If the variable in the population is not in categorical form, then it should be categorised first. Next, suppose that the outcome of interest is a binary variable. MRP procedure is summarised in two stages as follows (Gao et al., 2021):

1. **Multilevel regression stage**. Multilevel regression is fitted to get estimated population averages θ_j for every cell $j \in \{1, ..., J\}$. The multilevel logistic regression has a set of random effects $\alpha^k_{m[j]}$ for each categorical covariate k. These random effects have the effect of pooling each θ_j partially towards $X_j\beta$, the j_{th} is the row of $\mathbf{X}$, the design matrix corresponding to θ_j, while β is the fixed effect of the model. Suppose that n is the number of individual observations in the survey data, the form of multilevel regression could be written as follows:

$$\begin{aligned} &Pr(y_i = 1) = logit^{-1}\left(X_i\beta + \sum_{k=1}^{K} \alpha^k_m[i]\right), for\ i = 1, ..., n, \\ &\alpha^k_m \sim N(0, \sigma^2_k), for\ m = 1, ..., M_k \end{aligned} \tag{1.1}$$

2. **Poststratification stage**. The probabilities of the outcome from the previous stage, θ_j, is then poststratified using the known population size N_j of each cell j to get the estimates at the subpopulation level. This stage corrects the nonresponse in the population by utilizing the known size of every cell j relative to the total population size $N = \sum_{j=1}^{J} N_j$. In other words, the estimates is a weighted average of θ_j with N_j as the weight. Suppose that S is the subpopulation which is the combination of categories in the poststratification matrix, the MRP estimates could be expressed as:

$$\theta_S = \frac{\sum_{j \in S} N_j \theta_j}{\sum_{j \in S} N_j} \tag{1.2}$$

1.2 Thesis Structure

This thesis is structured as follows:

- Chapter 2 is a systematic literature review. This chapter discuss the review of current practice in MRP visualisations in various studies.
- Chapter 3 is a case study of MRP visualisations. This chapter aims to demonstrate the MRP application in the case of U.S. presidential voting result estimation. This chapter also demonstrates how the current practice of MRP visualisation could be improved.
- The final chapter, Chapter 4, summarises the findings and concludes the contribution of this study and possible future works.

Chapter 2

Systematic Literature Review

This study is performed using a systematic review method. This method collects empirical evidence explicitly and systematically using pre-specified eligibility criteria to answer a specific research question (Green et al., 2008). Systematic literature reviews also enable the process of finding the gap in a field of science, such as understanding what has been done and what needs to be done (Linnenluecke, Marrone, and Singh, 2020). Hence, in this case, systematic literature review could assist us to understand the common practice in MRP visualisations so that we can explore how to improve.

According to Brown University Library (2021), the key criteria of the systematic literature review are: *"a clearly defined question with inclusion & exclusion criteria; rigorous & systematic search of the literature; critical appraisal of included studies; data extraction and management; analysis & interpretation of results; and report for publication."* Hence, to conform with these criteria, this study incorporates the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)'s checklist and flow diagram. The following subsections discuss the steps conducted following these criteria.

2.1 Literature Identification

MRP is applied in various scientific fields, ranging from social and political science to public health. Therefore, to identify relevant literature, this study refers to research databases instead of field-specific journals. Those databases are JSTOR, EBSCO, and

PubMed. The first two databases are chosen due to their broad range of field coverage, while the latter is chosen since MRP is sometimes also applied in the health and medical fields. These databases were also chosen to represent the heterogeneity of the field, which is one of the important factors in a systematic literature review (Schweizer and Nair, 2017).

From these databases we identify relevant articles using the combination of several search terms. Generally the search terms include the term "multilevel regression", "post-stratification", "poststratification", and "multilevel model". Our target literature is articles that are written in English. We exclude all of the publications before 1997 since this was the first proposal date for MRP. Initially we included only the title/abstract when searching these databases. However, using this method limits the set of potential articles to only include those with the search term in the abstract/title. To rectify this, we also include a search with "all field" in the search criteria. Note that for EBSCO, we directly apply the search for all fields. The detailed literature identification is shown in Table 2.1.

The total number of articles from this search criteria are 327. Next, we utilize the literature manager, EndNote X9, to manage these articles and to find duplicate articles. After removing those duplicate articles, we have 212 articles to be screened in the next stage.

Table 2.1: *Detail of literature identification*

Database	Search Terms	Search Field	Inclusion	Exclusion	Number Returned
JSTOR	(multilevel regression and poststratification) OR (“post-stratification”)	Abstract	Article, content I can access, English	anything before 1997	44
JSTOR	(("multilevel regression" AND ("post-stratification" OR Poststratification)) OR ("multilevel model" AND ("post-stratification" OR Poststratification)))	All field	Article, English	anything before 1997	142
EBSCO	"multilevel regression with post-stratification" OR "multilevel regression with poststratification" OR "multilevel regression and Poststratification" OR "multilevel regression and Post-stratification"	All field	Academic (Peer-Reviewed) Journals, English	anything before 1997	42
EBSCO	(multilevel regression AND post-stratification) OR (multilevel model AND post-stratification) OR (multilevel regression AND poststratification) OR (multilevel model AND poststratification)	All field	Academic (Peer-Reviewed) Journals, English	anything before 1997	45
PubMed	"multilevel regression with post-stratification" OR "multilevel regression with poststratification" OR "multilevel regression and Poststratification" OR "multilevel regression and Post-stratification"	Title/Abstract	Article, English	anything before 1997	26
PubMed	(multilevel regression AND post-stratification) OR (multilevel model AND post-stratification) OR (multilevel regression AND poststratification) OR (multilevel model AND poststratification)	All field	Article, English	anything before 1997	28

2.2 Screening and Eligibility Criteria

We screen all of the articles based on predetermined criteria. We find that 3 articles are apparently not research papers. This results in 209 abstracts to be screened. To screen efficiently, we use two stages. The first stage is a review of abstracts, the second a full manuscript review.

2.2.1 Stage 1: Review of abstracts

In the first stage DA and LK independently review all article abstracts with the following eligibility criteria:

1. The abstract should mention analysis of data or creation of simulation data.
2. The abstract should mention the use of MRP or multilevel models to make population estimates or the use of other regression models (BART, spatial, stacking, trees) to make population estimates.

During the screening, DA and LK agreed that 61 articles meet the eligibility criteria listed above, while 104 articles do not meet the criteria. The two reviewers disagreed on 44 articles. Accordingly, DA and LK skim the full manuscript to decide whether the paper could be included in the next stage or not. As the result, an additional 22 more articles are moved to stage 2, making a total of 83.

2.2.2 Stage 2: Full manuscript review

DA reviews the full manuscript on 83 articles based on a second set of criteria. The aim of this stage is to get the list of the final articles that would be included in the study. We set the criteria of inclusion as follow:

1. It should apply MRP as its method.
2. It should contain at least one plot relate to MRP findings.

During this stage, we exclude 4 articles as they do not meet the first criteria. Further, 7 articles are excluded as they do not meet the second criteria. Also, an article is not included

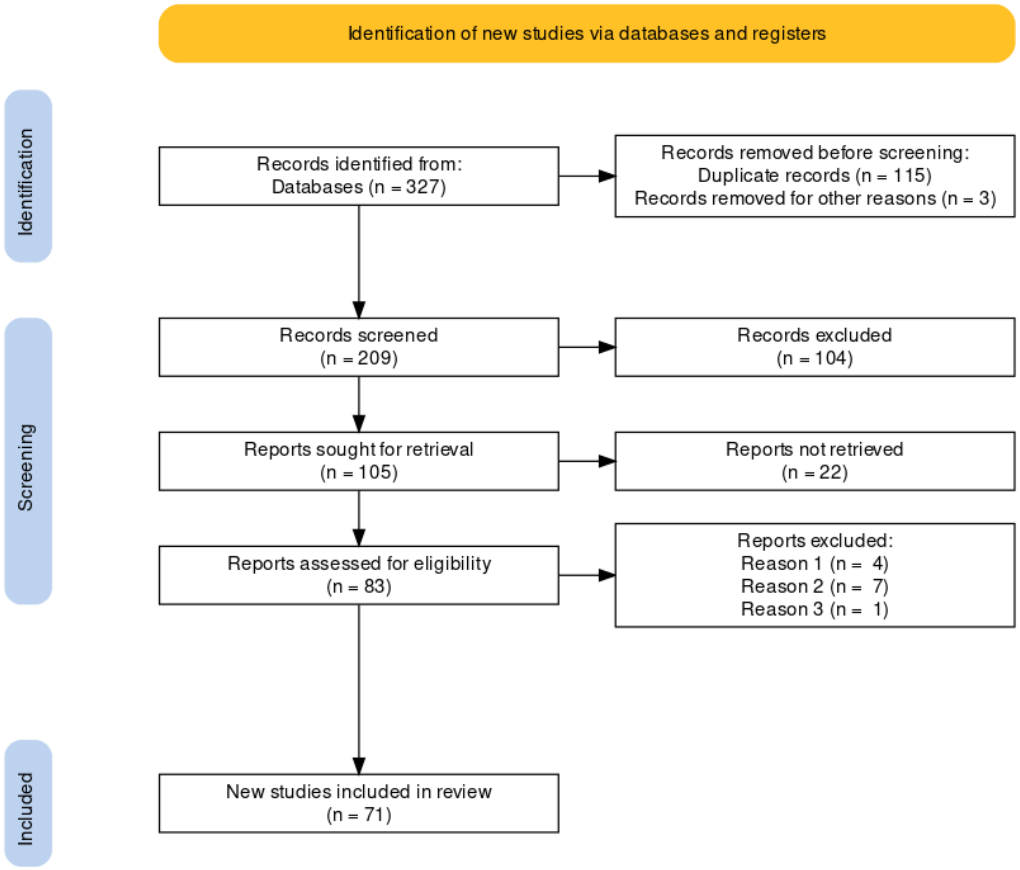

Figure 2.1: *PRISMA flow chart of this systematic literature review.*

because it is a duplicate that was not detected automatically by Endnote X9. Finally, we have 71 articles to be reviewed in the next stage. Figure 2.1 displays the PRISMA flow chart of this study. This figure is generated using `PRISMA2020` (Haddaway, Pritchard, and McGuinness, 2021).

2.3 Data Extraction and Analysis

We focus the data extraction on the MRP-related plot. We manually create a metadata for each plot (included in the supplementary material). We will use this metadata to analyse the current reporting practices with MRP. This metadata will also ensure the reproducibility of the analysis and to maintain the transparency of the systematic literature review process.

We code the plots according to their type, i.e., communication (coded to 0) and diagnostic plot (coded to 1). For diagnostic plots, we examine whether the plots compare MRP with other estimates, which are:

1. Raw (direct estimates or direct disaggregation);
2. Ground truth;
3. Weighted estimates;
4. Estimates from other MRP models, for example, a paper build several MRP models from various simulation scenarios or using different covariates;
5. Estimates from another study/survey;
6. Estimates from another method, for example comparing MRP with Bayesian Additive Tress with Post-Stratification(BARP).

Plots that show a comparison of MRP with the above list would be coded to 1, otherwise coded to 0. Diagnostic plots also categorised based on how they compare the performance of MRP. The five observed criteria are:

1. Bias;
2. Mean Absolute Error (MAE);
3. Mean Square Error (MSE)/ Relative Mean Square Error (RMSE);
4. Standard Error (SE);
5. Correlation.

Each plot is assessed based on the use of the performance metric. For each metric is scored based on whether it is used (coded 1) or not (coded 0).

We also review other features of the plot using the grammar in `ggplot2` (Wickham, 2016) as a framework. The common grammar used in practice allows us to understand to what extend MRP models are effectively visualised. It is worth noting that there is no specific convention or well-documented recommendation on how data should be visualised as building a graph more often involves choice or preference (Midway, 2020). For example, there is no specific convention on which variable should be put on the x and y-axis in a

scatter plot, even though it has been common knowledge to put the response variable on the y-axis and the explanatory variable on the x-axis. Hence, grammar assists us in evaluating well-formed graphics (Wickham, 2010). In addition, Vanderplas, Cook, and Hofmann (2020) mention that classifying and comparing graphs according to their grammar is more robust and more elegant.

Accordingly, we examine the facet, geom, axis, color, and shape. For reproducibility, the metadata also contains the article's author/s, publication year, title, and corresponding figure number as it appeared in the article. After the extraction, we analyze the data using graphical visualization with `ggplot2` (Wickham, 2016). The result will be discussed in the following subsection.

2.4 Common practices in MRP visualisations

In this study, graphics are classified into two types, i.e., communication and diagnostic plots. A plot is classified as a communication plot if the plot's goal is solely to convey the MRP result. A diagnostic plot is used to understand the MRP estimate, and typically displays the MRP estimation by showing the performance metrics or compares it with other estimation methods. From 71 articles, we extract the data of 243 plots. 47.33 % of these plots are diagnostics plots, while the remaining are communication plots.

2.4.1 Performance metrics used in MRP

According to Botchkarev (2019), performance metrics is *"a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted"* RMSE and MAE are among the most common methods used in many studies (Botchkarev, 2019). However, Willmott and Matsuura (2005) states that RMSE should not be reported in any studies since it could be multi-interpreted because it does not describe average error alone and MAE is more appropriate metric. This argument is denied by Chai and Draxler (2014) who argue that RMSE is not ambiguous and better than MAE if the distribution of model's error is normal. Accordingly, there is no single metric that fits all methods (Chai and Draxler, 2014).

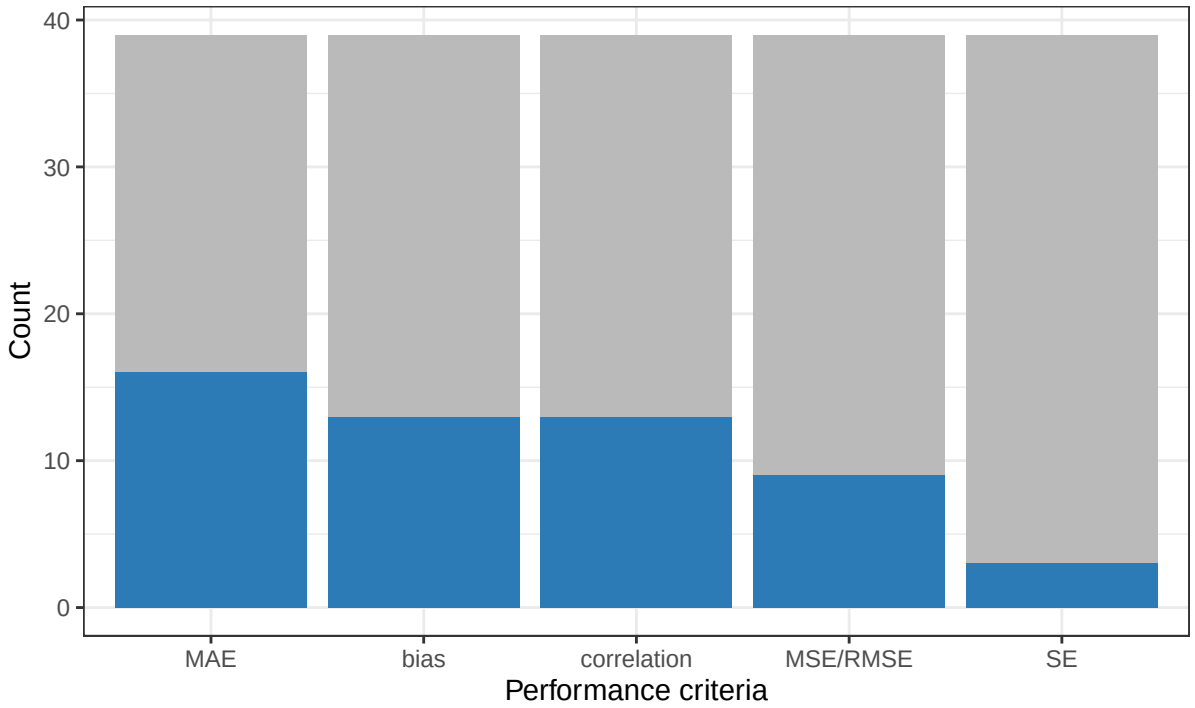

Figure 2.2: *We observed five performance metrics used: Mean Absolute Error (MAE), bias, correlation, Mean Square Error/Root Mean Square Error (MSE/RMSE), and Standard Error (SE). The blue shade represents the number of articles that show performace metrics in plot, while the grey shade represents the number of articles that show performance of MRP but did not use the corresponding metrics.*

In this study, we find that there are 39 plots out of 115 diagnostic plots (about 34%) that display performance measures. As seen in Figure 2.2, we find that MAE is the most widely used performance metric in MRP visualisations. Bias, which is interpreted similarly to MAE, is also widely used. Meanwhile, the squared error measures, which are MSE/RMSE and standard error, are only used in a few plots. It is interesting that correlation, which is not a common metric for performance, is more widely used than square error metrics.

Most of these metrics only refer to point estimates, i.e., the distance between the predicted value and the actual values. Also, these metrics mainly measure bias. However, MRP is a model in which bias-variance is applied. Therefore, other measures are also needed that reflect the degree of uncertainty and variations in the predicted value. Measures such as length of confidence or credibility interval can be used, in which the narrower the value, the more precise the estimates.

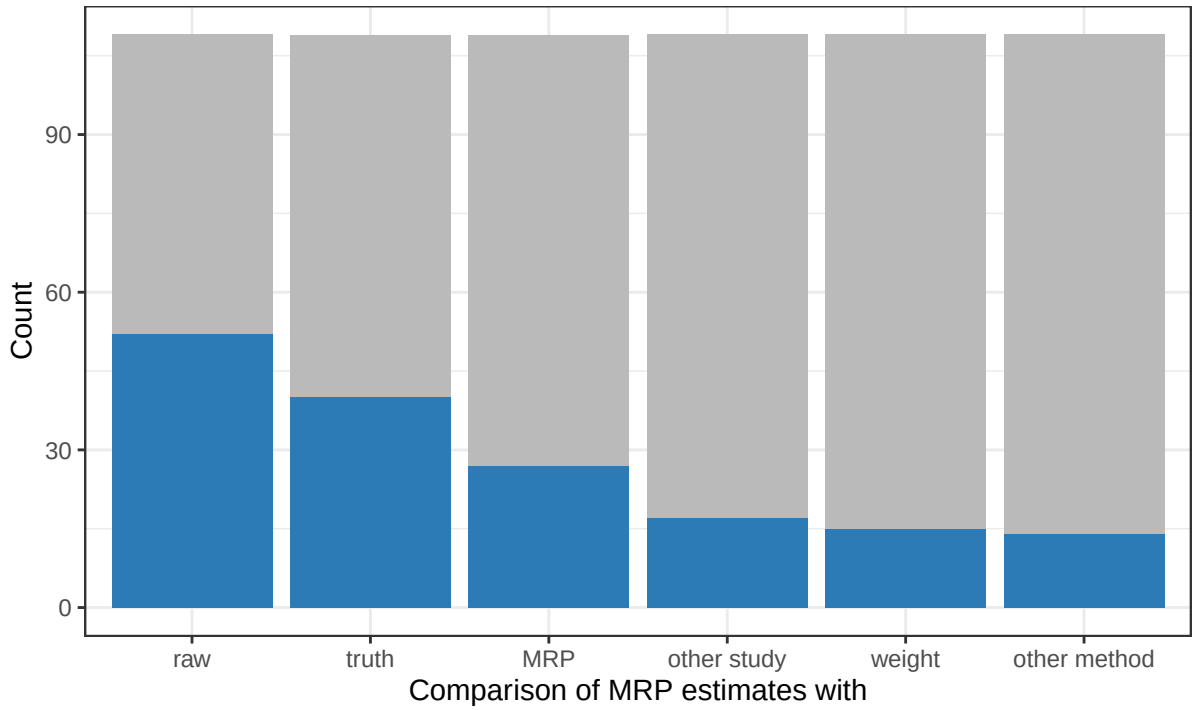

Figure 2.3: *Estimates that are compared with MRP. The blue shade represents the number of articles that compare MRP estimates with the result of other estimation methods, while the grey shade represents the number of articles that also showed comparison of MRP but did not compare to this particular estimate. Note that the blue bars do not sum to the count because some plots compared to multiple alternative estimates.*

2.4.2 Common comparisions with MRP

The goal of MRP is to make a population estimate. The method aims to adjust an unrepresentative survey to obtain accurate population and sub-population estimates. Where possible MRP is usually compared with a true value. This is generally only possible in political science applications where an election provides this true estimate. To understand how MRP improves estimates from an unrepresentative survey when compared with no adjustment, MRP estimates are usually compared with direct estimates (raw). Similarly, to understand the improvement of estimates when compared with more traditional methods, MRP is often compared with weighted estimates.

This study finds that from 115 diagnostic plots, 109 (about 95%) compare MRP estimates with estimates from other methods. Figure 2.3 shows the distribution of alternative estimates. MRP estimates are mostly compared to direct estimates and the ground truth. Some studies also compare estimates from several MRP models (usually with different

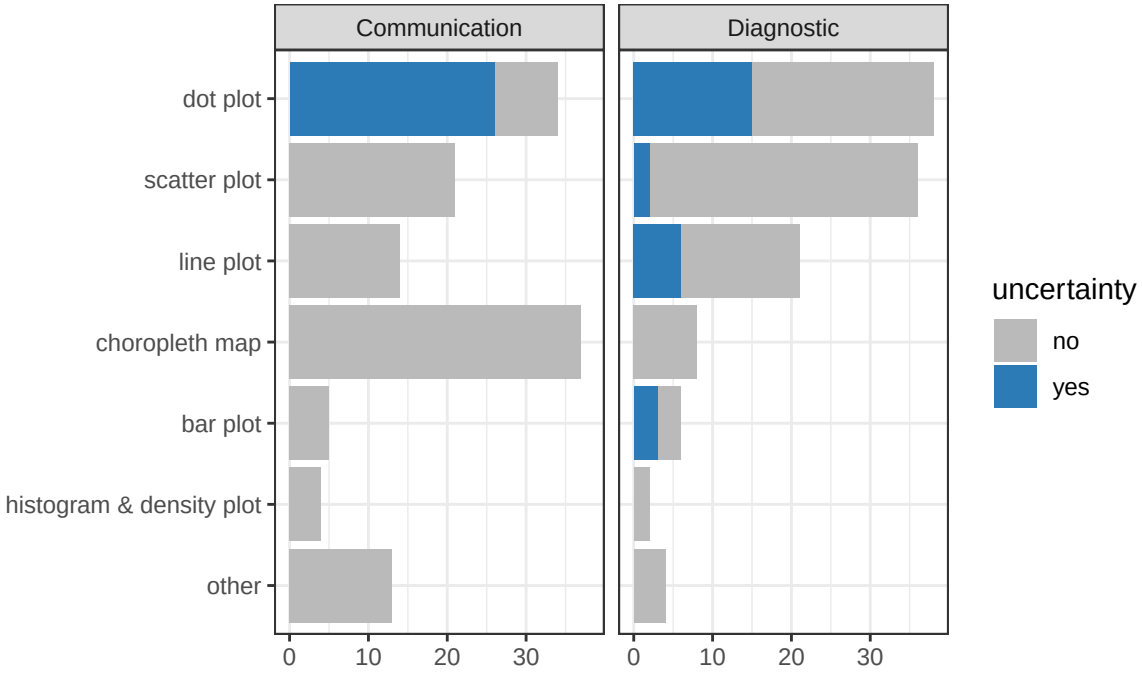

Figure 2.4: *Common plot types used in MRP visualisations. The blue shade display the number of plots that showed uncertainty, while the grey shade display the number of plots that did not show uncertainty. Both communication and diagnostics plots rarely displayed uncertainty.*

model specifications). There are not many plots showing the comparison between MRP estimates and weighted estimates.

2.4.3 Common grammar in MRP visualisations

Plot type

Plot type, referred to as `geom` in the grammar of graphics, represents the shape and features displayed in the graph. Figure 2.4 suggests that communication and diagnostic plots have a different pattern in which plot types are used. Communicating MRP estimates are mostly done using a choropleth map as MRP is often used for small area estimation. For diagnostic purposes, dot plots are mostly used to compare more than two estimation methods or to show some performance metrics.

Notice that Figure 2.4 also displays the use of uncertainty in MRP model visualisations. According to Midway (2020), displaying uncertainty in the statistical graphs is essential as the absence of this measure would produce a misleading interpretation and hinder

some statistical messages. However, he further states that uncertainty is often neglected in data visualisation. This is what we find in this study - uncertainty is not often seen in the plots. This is possibly because many of the application areas are more familiar with official statistics. In official statistics uncertainty is often unreported because results that are not sufficiently precise are not reported.

Values put in x and y-axis

The main component of a data visualisation is the axis. x and y-axis represents what value/data are exactly displayed in the graph. In MRP visualisations (Figure 2.5), estimates, small area, actual value (truth), and time are amoung the values that are displayed in the plot. We can also see that the constructs represented by the x and y-axis are more varied in diagnostic plots. It is worth noting that there are no strict rules on values to put in x and y-axis. However, it is a common that the the fixed value is represented by the x-axis, while the random variable is represented in the y-axis. We do not see this in our results as we find estimates and truth are plotted on the x and y axes interchangeably. Another common rule of thumb is that time is almost always represented on the x-axis, which is supported by the findings of our study.

Facet

Paneling or faceting is considered as to one of the effective visualisation techniques to compare the same variables by its grouping factor (Midway, 2020). We find in our results that faceting is a common practice in MRP visualisations. Figure 2.6 shows that faceting the plots by small area that is being estimated is the most common, followed by case. Small area refers to the levels of the predictors in the MRP model, for example, state, county, and religion. In several plots, small area could be referred to another variable that is associated with the MRP estimates, but is not included in the model, such as the association between health literacy and the opinion on a health-related bill. Health literacy is a variable that is not included in fitting the MRP model, while the latter is the MRP estimates. Further, case is referred to the outcome predicted with MRP.

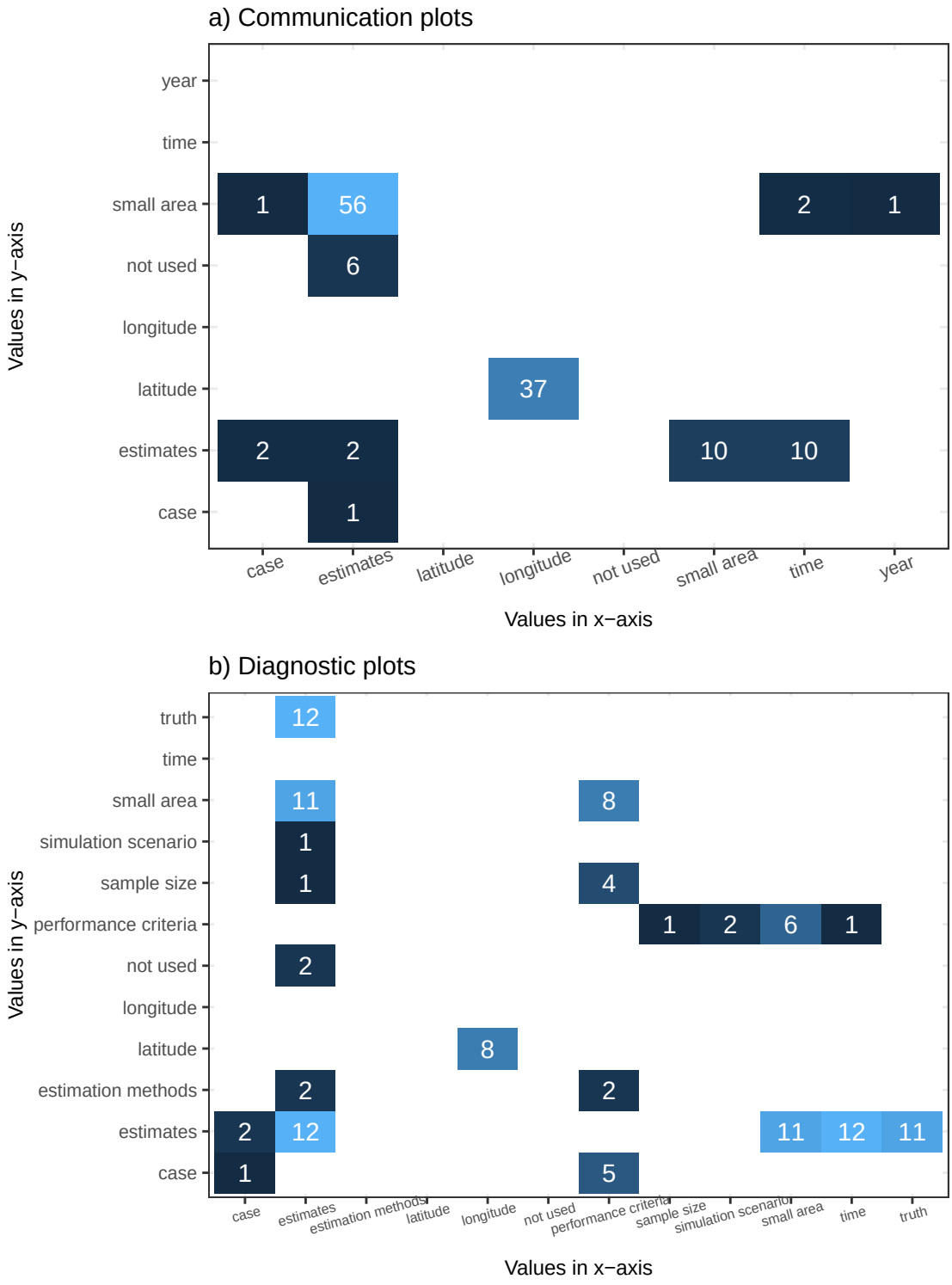

Figure 2.5: *Common values put in plots' axis. Axis in diagnostic plots more varied compared to communication plot.*

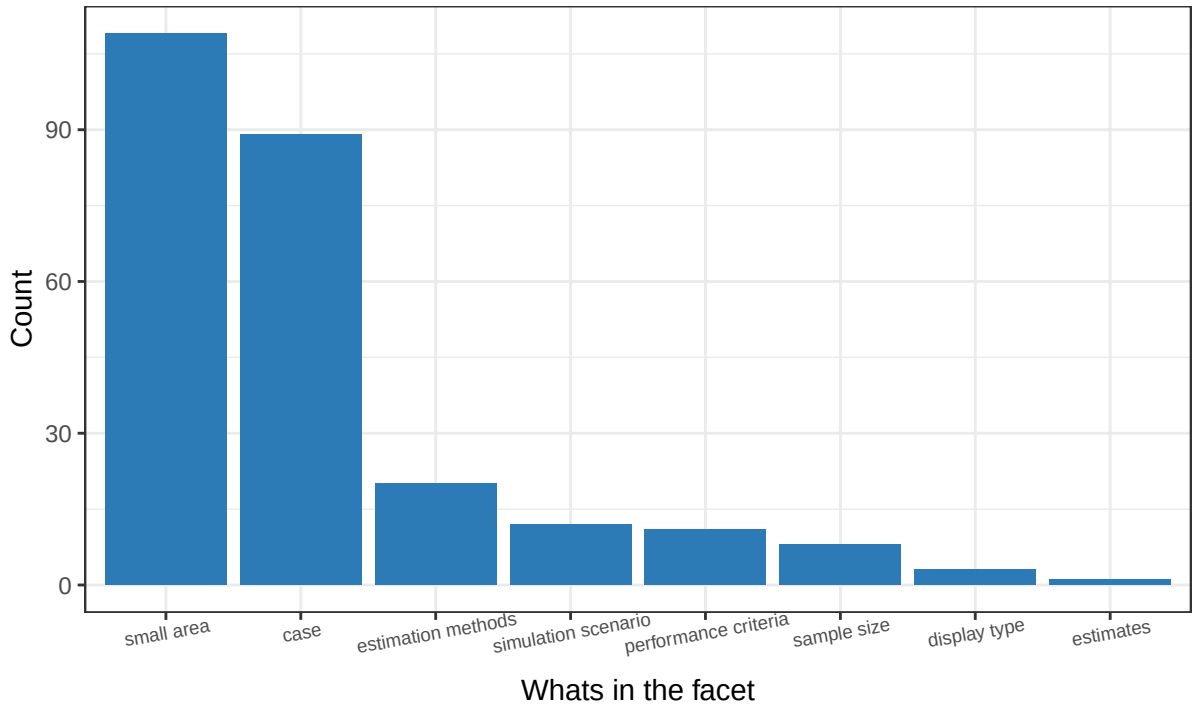

Figure 2.6: *The facetting variable in MRP visualisations*

Other features used

Besides the features explained previously, color and shape are also the components of grammar of graphics. According to a large experimental study on visualisations, color is a memorable feature of a graph (Midway, 2020). Further, Few (2008) states that the aim of color in data visualisations are to highlight particular data, to group items, and to encode quantitative values. In addition, color is sometimes displayed along with shapes to distinguish more features.

We find, as shown in Figure 2.7, that both communication and diagnostic plots incorporate color only about half the time. Shape is used less often. When there is only one feature to be displayed, for example, estimation methods, people tend to choose to use color first, rather than shape. This is seen in Figure 2.7 as after incorporating color to distinct estimates, small area, estimation methods, and performance criteria, people tend to not use shape anymore.

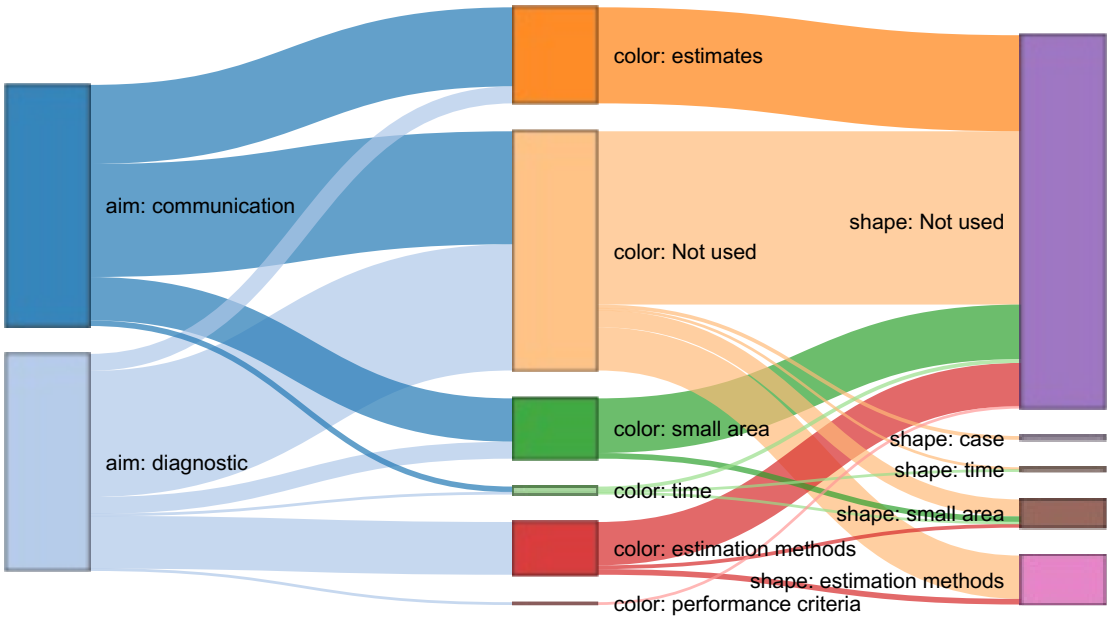

Figure 2.7: *Color and shape commonly used in MRP visualisations. Both communication and diagnostic plots rarely use color and shape features.*

Chapter 3

Case Study: Application of MRP in Presidential Voting Estimation

The majority of MRP applications are used in the context of estimating public opinion in the social and political sciences, although, in recent development, MRP is also conducted in various fields, for example, health and environmental studies. When first introduced by Gelman and Little (1997), MRP was applied to generate state estimation of the 1988 U.S. presidential election. Additionally, various subsequent studies also made presidential voting the case of interest. We recorded at least seven articles (Gelman (2014); Ghitza and Gelman (2013); Kiewiet de Jonge, Langer, and Sinozich (2018); Lauderdale et al. (2020); Lei, Gelman, and Ghitza (2017); Park, Gelman, and Bafumi (2004); Wang et al. (2015)) included in the Systematic Literature Review in Chapter 2 have applied MRP in the case of the presidential election. Therefore, in this chapter, we will also apply MRP to estimate the 2016 U.S. presidential voting outcome, especially the probability of Trump votes. This also allows us to compare MRP estimates with the actual value of the Trump votes that are already available. In this case study, we use the Cooperative Congressional Election Study (CCES) 2016 data (Ansolabehere and Schaffner, 2017) as the survey data and the American Community Survey data 2015-2017 (U.S. Census Bureau, 2021c) as the population/ poststratification data.

3.1 Data

3.1.1 Cooperative Congressional Election Study (CCES) 2016

CCES is an annual survey aims to capture Americans' view on Congress, their voting behavior and experience with regards to political geography, social, and demographic context (Ansolabehere and Schaffner, 2017). In 2016, CCES covers 64,600 samples spread over 51 states. Accordingly, Ansolabehere and Schaffner (2017) mention that the data is precise enough to measure the distribution of voters' preference in most states. In addition, beyond its reliable sample size, CCES is regarded to be desirable dataset as it measures vote preference before and after in two waves so that it is even more reliable compared to generic question (Kuriwaki, 2021b).

To fit MRP models, we use several variables from this survey. To obtain the data from the CCES website, we utilize an R package, `ccesMRPprep` (Kuriwaki, 2021a). One of the advantages of using this package is that the data has been pre-processed in particular for MRP purposes, in this case, we use the `ccc_std_demographics` function. Also, the variable names are already recoded so it has a more interpretable names. The code to get the data is available in the supplementary materials of this thesis.

The outcome variables, which will be explained in detail in Section 3.2, are the vote preference/intention (`CC16_364c`), candidate voted (`CC16_410a`), and party identity (`pid3` including leaners, i.e, coded as Independents in pid3 who expressed leaning towards a party in `pid7`). Table 3.1 shows the distribution of answers in those three variables. In `ccesMRPprep`, these variables have been renamed so that they are equivalent to `intent_pres_16`, `voted_pres_16`, and `pid3_leaner`, respectively. It is worth noting that the MRP models we would like to build use binary responses into yes/no in terms of whether the respondents vote for Trump/Republican or not.

Further, the geography and demographic variables used as covariates in the models are `state`, `age`, `gender`, `education`, and `race`. Table 3.2 shows the distribution of categories/levels of `age`, `gender`, `education`, and `race`. Initially, age is a discrete variable but in this case it is categorised. Also, `education` and `race` has more levels in the original

Table 3.1: *Percentage of each answer in CCES 2016. This question will be the MRP models outcome in this case study. Since the model outcome is binary, these answer will be converted to be yes/no in the context of vote for Trump/Republican.*

Candidate voted	percentage
Hilary Clinton	34.27
Donald Trump	29.03
Other / Someone Else	6.26
Did Not Vote	0.13
Not Sure / Don't Recall	0.35
NA	29.97

Candidate will be voted	percentage
Donald Trump (Republican)	29.76
Hillary Clinton (Democrat)	42.57
Gary Johnson (Libertarian)	4.87
Jill Stein (Green)	2.17
Other	2.91
I Won't Vote in this Election	5.13
I'm Not Sure	10.12
NA	2.47

Party identity including leaners	percentage
Democrat (Including Leaners)	48.20
Republican (Including Leaners)	32.27
Independent (Excluding Leaners)	16.24
Not Sure	3.20
NA	0.08

Table 3.2: *The response of covariates. Note that this response has been categorised into certain levels that are reflected in these tables.*

Gender	percentage
Male	45.71
Female	54.29

Race	percentage
White	69.44
Black	12.00
Hispanic	10.59
Asian	3.53
Native American	0.81
All Other	3.63

Age	percentage
18 to 24 years	8.30
25 to 34 years	19.62
35 to 44 years	15.75
45 to 64 years	38.36
65 years and over	17.98

Education	percentage
HS or Less	28.41
Some College	35.38
4-Year	23.04
Post-Grad	13.17

data but it is collapsed here to obtain fewer levels. The proportion of people answered the survey based on the state is displayed in the appendix of this thesis.

3.1.2 American Community Survey (ACS) 2015-2017

In this study, we use the ACS 2015-2017 data as the poststratification data. The ACS provides annual-basis information about jobs and occupations, demographic and citizenship, educational attainment, homeownership, and other topics (U.S. Census Bureau, 2021a). Further, the ACS uses monthly probabilistic samples to produce the annual estimates. It could be understood that the ACS is desirable data to represent the U.S. population since the coverage rate for the 2015-2017 ACS is 92.4%, 91.9%, 91.6%, respectively (U.S. Census Bureau, 2021b). It is also implied by Gao et al. (2021) that the ACS is the most accurate representation of the U.S. population every year. However, it is also worth noting that the ACS coverage do not necessarily match the CCES sample, and therefore, bias might always be presented.

To fit MRP models, we need the individual data of the ACS instead of the aggregated statistics. Hence, we use the 1-year Public Use Microdata Sample (PUMS), which carries the information/records of individual people on a yearly basis. The 1-year PUMS data reflects approximately one percent of the U.S. population (U.S. Census Bureau, 2016). Therefore, in this study, we use three years periods of the ACS 1-Year PUMS from 2015-2017 instead of 2016 only to get a better and more stable representation of the American population. Every individual in the data has a weight (`PWGTP`). Since we use three years period, this weight is then divided by 3.

The data is publicly available on the U.S. Census Bureau website. We downloaded the data in a .csv format (csv_pus.zip) year by year (2015-2019) through access on FTP site. After that, we did a data pre-processing to bind the three years of the PUMS data. We only use some variables in this data for the MRP-purposes, i.e., unique identifier of the person (`SERIALNO`), state (`ST`), weight (`PWGTP`), education (`SCHL`), sex (`SEX`), race (`RAC1P`), Hispanic origin (`HISP`), and age (`AGEP`). We also did a data munging to recode and collapse some categories in these variables. Note that the `RAC1P` did not record for Hispanic ethnicity. Hence, we introduce a new category here, Hispanic, identified if the person answers other than "1" in the `HISP` variable. Table 3.3 shows the categorised response of the variables obtained from the ACS, i.e., age, race and ethnicity, and education. Also, notice that we

Table 3.3: *The response categories of post-stratification data.*

Sex	percentage
Male	48.9
Female	51.1

Race and ethnicity	percentage
White alone	67.00
Black or African American alone	9.89
Hispanic	14.41
Asian alone	5.16
American Indian alone	0.80
Native Hawaiian and Other Pacific Islander alone	0.15
American Indian and Alaska Native tribes	0.08
Alaska Native alone	0.07
Some Other Race alone	0.19
Two or More Races	2.24

Age	percentage
Less than 18 years	20.73
18-24	8.73
25-34	11.92
35-44	11.62
45-54	13.45
55-64	14.64
65-74	10.95
75-89	7.00
90 years and over	0.95

Education	percentage
No high school	27.03
Regular high school diploma	18.82
Some college	21.25
Associate's degree	6.39
Bachelor's degree	14.50
Post-graduate	8.99
NA	3.03

get some NA values in education. This is actually the education level of under-school-age respondents. In fact, we will omit the "Less than 18 years" age group in the MRP models as it is not included in the survey data (CCES). Hence, the NAs in education response will be eventually omitted as well. The detailed code of the data pre-processing is available in the supplementary materials of this thesis.

3.2 Model Specifications

In Chapter 2, we found that the diagnostic plots shown in many articles compare MRP estimates with other estimates, one of which compares several MRP estimates with different specifications. Therefore, in this case study, we build five different MRP models as follows.

Table 3.4: *The distribution of answer in the outcome (vote). It will be the outcome in three models, i.e., baseline model, model with education as additional covariate, and model with more categories in race. We observe a reasonably large percentage of NA.*

Candidate voted	percentage
no	41.00
yes	29.03
NA	29.97

Baseline model

We begin the model fitting with the baseline model. In this model, we set the binary outcome as whether the respondents vote for Trump or not in the 2016 election. Therefore, we transform the response of `voted_pres_16` into a binary variable called `vote` ,i.e, if the value of `voted_pres_16` is "Donald Trump", then `vote` variable coded to "yes", otherwise "no". The NA values in `voted_pres_16` will stay as NA in the new `vote` variable. Hence, the distribution of the baseline model's outcome variable is displayed in Table 3.4.

The covariates used are `age`, `gender`, `state`, and the re-categorised/collapsed `race` variable into fewer levels. As seen in Table 3.2, `race` has 6 categories, i.e., `White`, `Black`, `Hispanic`, `Asian`, `Native American`, and `All Other`. In the baseline model, we collapsed the `Native American` and `All Other` into `Other`. Meanwhile, the levels of `age`, `gender`, `state` stay the same in the levels displayed in Table 3.2. The baseline model equation is:

$$\Pr(vote_{j[i]} = 1) = logit^{-1}\left(\beta_0 + \alpha^{age}_{m[i]} + \alpha^{gender}_{m[i]} + \alpha^{state}_{m[i]} + \alpha^{collapsed\ race}_{m[i]}\right), for\ i = 1, ..., n, \tag{3.1}$$

and $vote_{j[i]}$ is the binary outcome (1 = yes, 0 = no) for individual i in poststratification cell j. β_0 is the intercept. $\alpha^{age}_{m[i]}$, $\alpha^{gender}_{m[i]}$, $\alpha^{state}_{m[i]}$, and $\alpha^{collapsed\ race}_{m[i]}$ are the random effects for `age`, `gender`, `state`, and `collapsed race`, respectively. The subscript in each coefficient represents the category of the $i-th$ respondent, such as, $\alpha^{collapsed\ race}_{m[i]}$ takes value from $\{\alpha^{collapsed\ race}_{White}, \alpha^{collapsed\ race}_{Black}, \alpha^{collapsed\ race}_{Hispanic}, \alpha^{collapsed\ race}_{Asian}$, and $\alpha^{collapsed\ race}_{Other}\}$. Each random effect has an independent prior distribution, such as, $\alpha^{collapsed\ race}_{m} \sim N(0, \sigma^2_{collapsed\ race})$ and

$\alpha_m^{collapsed\ race} \sim t(3, 0, 2.5)$. Here, we use the default prior because we only want to compare models for visualisation purpose instead of looking for the best model for estimation.

Model with `education` as additional covariate

Next, we create a bigger model by adding `education` as additional covariate to the baseline model. The levels of `education` is also displayed in Table 3.2. Hence, the model specification is:

$$\Pr(vote_{j[i]} = 1) = logit^{-1}\left(\beta_0 + \alpha_{m[i]}^{age} + \alpha_{m[i]}^{gender} + \alpha_{m[i]}^{state} + \alpha_{m[i]}^{collapsed\ race} + \alpha_{m[i]}^{education}\right), \quad for\ i = 1, ..., n. \tag{3.2}$$

Model with original race categories

This model is essentially the same with baseline model, except that there are more race categories, which are `White`, `Black`, `Hispanic`, `Asian`, `Native American`, and `All Other`. The model equation is:

$$\Pr(vote_{j[i]} = 1) = logit^{-1}\left(\beta_0 + \alpha_{m[i]}^{age} + \alpha_{m[i]}^{gender} + \alpha_{m[i]}^{state} + \alpha_{m[i]}^{original\ race}\right), for\ i = 1, ..., n. \tag{3.3}$$

Model with different outcomes

Vote intention/preference

This model mimicks the model in Equation (3.2), except that we have different outcome or response variable. The response here is whether the respondent intent to vote for Trump (`yes`) or not (`no`). It is transformed from `intent_pres_16` variable in the CCES data to a new variable called `intent`. If the value of `intent_pres_16` is "Donald Trump (Republican)", then `intent` variable coded to "yes", otherwise "no". The NA values in

Table 3.5: *The distribution of answer in the outcome (intent).*

Candidate will be voted	percentage
no	67.76
yes	29.76
NA	2.47

Table 3.6: *The distribution of answer in the outcome (party).*

Party identity	percentage
not Republican	67.65
Republican	32.27
NA	0.08

`intent_pres_16` will stay as NA in the new `intent` variable. The distribution of observed "no", "yes", and NA in this variable is shown in Table 3.5.

The model is specified as follows:

$$\Pr(intent_{j[i]} = 1) = logit^{-1}\left(\beta_0 + \alpha^{age}_{m[i]} + \alpha^{gender}_{m[i]} + \alpha^{state}_{m[i]} + \alpha^{collapsed\ race}_{m[i]} + \alpha^{education}_{m[i]}\right), \quad for\ i = 1, ..., n. \tag{3.4}$$

Party identity

Beside vote intention, another outcome is the party identity in terms of whether the respondents identify themselves as Republican or not. This variable is derived from `pid3_leaner` variable and referred as `party`. If the value of `pid3_leaner` is "Republican (Including Leaners)", then `party` variable coded to "Republican", otherwise "not Republican". The NA values in `pid3_leaner` will stay as NA in the new `party` variable. The distribution of this outcome variable is displayed in Table 3.6.

The specification of covariates is also the same with model in Equation (3.2).

$$\Pr(party_{j[i]} = 1) = logit^{-1}\left(\beta_0 + \alpha^{age}_{m[i]} + \alpha^{gender}_{m[i]} + \alpha^{state}_{m[i]} + \alpha^{collapsed\ race}_{m[i]} + \alpha^{education}_{m[i]}\right), \quad for\ i = 1, ..., n. \tag{3.5}$$

The estimates from the multilevel model is then used for the second stage of MRP, which is poststratification. As the explanation in Section 1.1, poststratification is essentially taking the weight average of the cell-wise posterior estimates with the size of each cell in the population table as the weight (Gao et al., 2021). For example, the poststratification estimates of Black Men attained High School or less (HS or Less) in Alabama which is categorised in 45 to 64 years old group age who voted for Trump in the 2016 presidential election is:

$$\theta_S = \frac{\sum_{j \in S} N_j \theta_j}{\sum_{j \in S} N_j}, \tag{3.6}$$

where θ_S corresponds to the proportion of 45 to 64 years old of Black Men attained High School or less (HS or Less) in Alabama who respond to "yes" in the `vote` variable and N_j and θ_j are the size of cell corresponds to this sub-population category in the poststratification table and the posterior estimates of this sub-population category, respectively.

3.3 Model Preparation and Fitting

The MRP models require a synchronous survey and population data. Hence, to achieve this, we need to map the survey data to the population data. In this study, the model preparation and survey-population data mapping conducted with an R package, `mrpkit` (Kennedy, Gabry, Amaliah, Alexander, 2021). This package allows the transparent and reproducible workflow to build MRP model, from the data mapping until the prediction stage, including the model specification setting. The detailed code to build the MRP models is available in the supplementary materials of this thesis.

After mapping the survey and population data, we can obtain the poststratification table displayed in Table 3.7

Next, we implement Bayesian multilevel model using `brms` (Bürkner, 2018) to fit the model and obtain the posterior distributions of the parameters. `brms` itself incorporates cmdstanr (Gabry and Češnovar, 2021) as the backend. The samples of posterior distribution are

Table 3.7: *First five rows of the post-stratification table*

age	state	gender	collapsed_re	original_re	education	N_j
18 to 24 years	Alabama	Male	White	White	HS or Less	63982.0000
18 to 24 years	Alabama	Male	White	White	Some College	67957.0000
18 to 24 years	Alabama	Male	White	White	4-Year	8851.6667
18 to 24 years	Alabama	Male	White	White	Post-Grad	320.3333
18 to 24 years	Alabama	Male	Black	Black	HS or Less	40443.3333

generated in 4 chains with 1000 iterations in each chain. Since this task is computationally heavy and time-consuming, we conduct it using Monash's High Performance Cluster (HPC) infrastructure.

3.4 Results and Discussion

Chapter 4

Conclusion

Appendix A

Appendix

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Ansolabehere, S and BF Schaffner (2017). *CCES Common Content, 2016*. Version V4. https://doi.org/10.7910/DVN/GDF6Z0.

Botchkarev, A (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. eng. *Interdisciplinary journal of information, knowledge, and management* **14**, 45–76.

Brown University Library (2021). *Scientific Literature Review Resources and Services*. https://libguides.brown.edu/Reviews/types.

Bürkner, PC (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* **10**(1), 395–411.

Chai, T and RR Draxler (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. eng. *Geoscientific model development* **7**(3), 1247–1250.

Chambers, JM (1983). *Graphical methods for data analysis*. eng. The Wadsworth statistics/probability series. Belmont, Calif. : Boston: Wadsworth International Group ; Duxbury Press.

Cleveland, WS (1985). *The elements of graphing data*. eng. Monterey, Calif.: Wadsworth Advanced Books and Software.

Few, S (2008). *Practical rules for using color in charts - GitHub Pages*. https://nbisweden.github.io/Rcourse/files/rules_for_using_color.pdf.

Gabry, J and R Češnovar (2021). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

Gao, Y, L Kennedy, D Simpson, and A Gelman (2021). Improving Multilevel Regression and Poststratification with Structured Priors. eng. *Bayesian analysis* **1**(1).

Gelman, A (2014). How Bayesian Analysis Cracked the Red-State, Blue-State Problem. eng. *Statistical science* **29**(1), 26–35.

Gelman, A and TC Little (1997). *Poststratification Into Many Categories Using Hierarchical Logistic Regression*.

Ghitza, Y and A Gelman (2013). Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups: DEEP INTERACTIONS WITH MRP. eng. *American journal of political science* **57**(3), 762–776.

Green, S, JP Higgins, P Alderson, M Clarke, CD Mulrow, and AD Oxman (2008). "Introduction". In: *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd. Chap. 1, pp. 1–9. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470712184.ch1. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470712184.ch1.

Haddaway, NR, CC Pritchard, and LA McGuinness (2021). *PRISMA2020: R package and ShinyApp for producing PRISMA 2020 compliant flow diagrams (Version 0.0.2)*.

Kennedy, Gabry, Amaliah, Alexander (2021). *mrpkit: Multilevel Regression with Post-Stratification*. R package version 0.1.0.

Kiewiet de Jonge, CP, G Langer, and S Sinozich (2018). Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP). eng. *Public opinion quarterly* **82**(3), 419–446.

Kuriwaki, S (2021a). *ccesMRPprep: Functions and Data to Prepare CCES data for MRP*. R package version 0.1.8.900. https://www.github.com/kuriwaki/ccesMRPprep.

Kuriwaki, S (2021b). "The Swing Voter Paradox: Electoral Politics in a Nationalized Era". PhD thesis. Cambridge MA.

Lauderdale, BE, D Bailey, J Blumenau, and D Rivers (2020). Model-based pre-election polling for national and sub-national outcomes in the US and UK. eng. *International journal of forecasting* **36**(2), 399–413.

Lei, R, A Gelman, and Y Ghitza (2017). The 2008 Election: A Preregistered Replication Analysis. eng. *Statistics and Public Policy* **4**(1), 1–8.

Linnenluecke, MK, M Marrone, and AK Singh (2020). Conducting systematic literature reviews and bibliometric analyses. eng. *Australian journal of management* **45**(2), 175–194.

Lopez-Martin, J, JH Phillips, and A Gelman (2021). *Multilevel Regression and Poststratification Case Studies*. https://bookdown.org/jl5522/MRP-case-studies/.

Makela, S, Y Si, and A Gelman (2017). "Graphical Visualization of Polling Results". In: *The Oxford Handbook on Polling and Polling Methods*. Ed. by L Atkeson and M Alvarez. Oxford UK: Oxford University Press.

Midway, SR (2020). Principles of Effective Data Visualization. *Patterns* **1**(9), 100141.

Park, DK, A Gelman, and J Bafumi (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. eng. *Political analysis* **12**(4), 375–385.

Schneider, SK and WG Jacoby (2017). "Graphical Displays for Public Opinion Research". In: *The Oxford Handbook on Polling and Polling Methods*. Ed. by L Atkeson and M Alvarez. Oxford UK: Oxford University Press.

Schweizer, ML and R Nair (2017). A practical guide to systematic literature reviews and meta-analyses in infection prevention: Planning, challenges, and execution. eng. *American journal of infection control* **45**(11), 1292–1294.

U.S. Census Bureau (2016). *American Community Survey 2015: ACS 1-Year PUMS Files*. https://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2015_PUMS_README.pdf.

U.S. Census Bureau (2021a). *About the American Community Survey*. https://www.census.gov/programs-surveys/acs/about.html.

U.S. Census Bureau (2021b). *American Community Survey: Sample Size and Data Quality*. https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/.

U.S. Census Bureau (2021c). *The American Community Survey Public Use Microdata Sample, 2015-2017*. https://www.census.gov/programs-surveys/acs/microdata/.

Vanderplas, S, D Cook, and H Hofmann (2020). Testing Statistical Charts: What Makes a Good Graph? *Annual Review of Statistics and Its Application* **7**(1), 61–88.

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

Wickham, H (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics* **19**(1), 3–28.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H, D Cook, and H Hofmann (2015). Visualizing statistical models: Removing the blindfold. eng. *Statistical analysis and data mining* **8**(4), 203–225.

Willmott, C and K Matsuura (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. eng. *Climate research* **30**(1), 79–82.