Using Aggregated Demographic Data To Inform Electoral Boundary Redistributions: 2010 South Australian Election

Casey Briggs

Thesis submitted for the degree of

Master of Philosophy

in

Statistics and Applied Mathematics

at

The University of Adelaide

School of Mathematical Sciences

April 2015

Contents

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: . . . . DATE: . . . .

Acknowledgements

To my supervisors, Professor Nigel Bean, Dr Jonathan Tuke, and Professor Clement Macintyre: this thesis wouldn't exist without your support. Thank you for persisting with me and being so generous with your time and expertise. Thanks also to my former supervisor, Dr David Green.

Thank you to David Gully, Deputy Electoral Commissioner at the Electoral Commission of South Australia, for supplying data for this thesis, and responding helpfully to my questions even in the midst of busy election periods. My appreciation also to everyone that read drafts and provided feedback, in particular Sam and Stella who read the complete draft.

I want to thank my family, and in particular my parents, for supporting me through my entire education. There is no way I would be in the position I am now without your encouragement and love, and I'm deeply grateful to you for the opportunities you've given me.

Thanks to all the staff and students of the School of Mathematical Sciences, of whom there are far too many to name, for creating a great place to work since I commenced as an undergraduate in 2008.

Abstract

Electoral boundaries in South Australia are currently a contentious issue in politics, with allegations that the current boundaries are unfair. South Australia has fairness provisions that are unique in Australia governing the boundaries of electoral districts. However, in three of the last six state elections, the objective of fairness as characterised by these provisions has not been met.

Boundaries are drawn by the independent Electoral Districts Boundaries Commission, and are revised after every general election in South Australia. The Commission's method uses estimates for the voting behaviours in small areas to inform the decisions about boundary changes.

The objective of this thesis is to develop an alternative method for calculating these estimates, and test the credibility of the resultant estimates from our new method.

We develop a series of gradually refined regression models that use demographic data in South Australia to predict voting behaviour. The demographic data is sourced from the periodical Census of Population and Housing. In this research we also test the proposition that income, education level, and the language people speak at home are significant factors in their voting behaviour, at an aggregated group level.

We contend that the predictions calculated under the preferred model in this thesis are credible, and that the techniques used warrant further exploration.

Chapter 1

Introduction

In this chapter we give a brief description of the research problem and its motivation, and give an outline for this thesis.

1.1 Motivation

Electoral district boundaries in South Australia are reviewed, and redrawn if necessary, after every state election. These redistributions are conducted by a body independent of the government, called the Electoral Districts Boundaries Commission (EDBC).

The EDBC is required to ensure that electoral boundaries conform with a notion of fairness contained in the Constitution. In general terms this means that the boundaries should ensure that the party that receives the majority of the votes (after the distribution of preferences) at an election should be able to form government.

Since this fairness requirement came into effect in 1991, there have been six South Australian elections, and in three of these elections the party that received a majority of the votes (in all cases, the Liberal Party) was *not* able to form government. This indicates that either this characterisation of fairness is unworkable in practice, or that more information and advanced techniques are required to actually implement it.

One key part of the EDBC's method of redistribution involves calculating estimates for the strength of support for each major party in small areas of geography called 'collection districts'. There are more than 3000 collection districts in South Australia. These estimates are then used to make decisions about which collection districts to move between electoral districts.

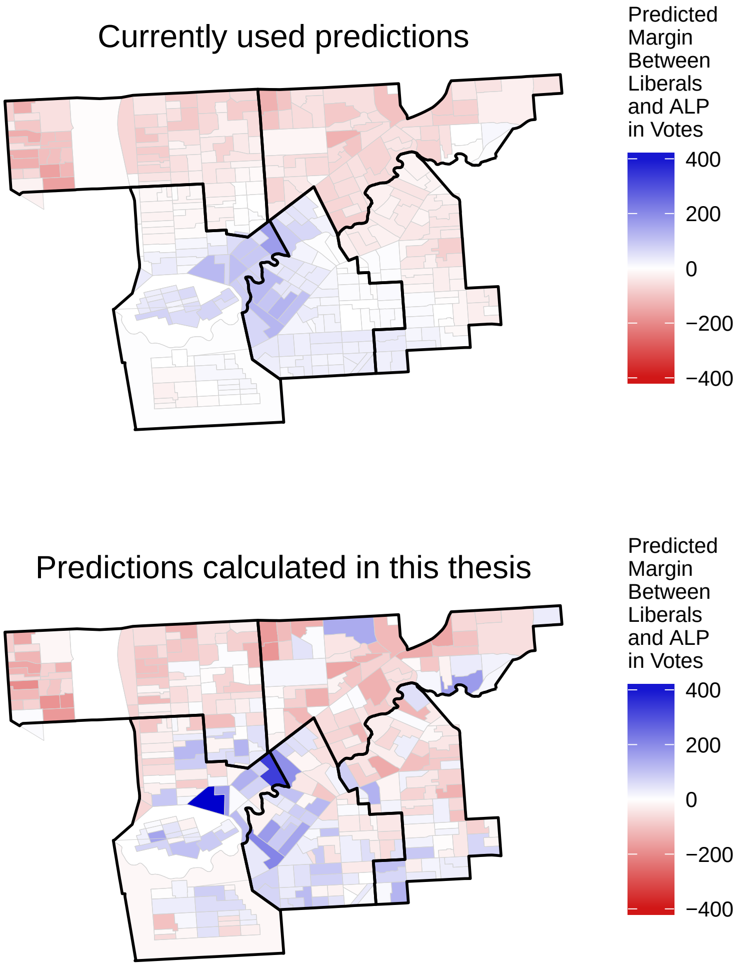
This thesis is chiefly concerned with the calculation of these estimates. We develop new methods of calculating them using new information in an attempt to improve the estimates, and hence improve the information available to the EDBC. Figure 1.1.1 gives a sample of the estimates we calculate, with the estimates that were actually used by the EDBC presented in the top of the figure, and the estimates calculated from the final model developed in this thesis presented in the lower figure.

The new information we use is data about the demographics of each collection and electoral district, sourced from the periodical Census of Population and Housing (hereafter referred to as 'the Census') conducted by the Australian Bureau of Statistics (ABS) [1]. We use data from the 2006 Census, along with election returns from the 2010 state election. Principal Component Analysis techniques are used to explore and visualise the predictor datasets.

There is a long understood link between a person's demographics and their voting behaviour (the most widely investigated characteristic being social class) and we hope this link can be exploited to give more useful estimates.

Aside from applications to electoral redistribution, the work in this thesis contributes

[1] `http://abs.gov.au/census`



Currently used predictions

Predictions calculated in this thesis

Figure 1.1.1: Estimates for the margin (in votes) between the Liberal Party and the Australian Labor Party in each collection district in the five electoral districts of Adelaide, Enfield, Hartley, Norwood, and Torrens. The estimates that were actually used by the EDBC in the redistribution following the 2010 state election are shown at the top, and estimates calculated from the model developed in this thesis are shown at the bottom.

to the literature on how a person's demographic factors effect their likelihood of voting for each major party. There is a limited literature base on this in an Australian context, and research from other countries is not automatically transferable to Australia because different countries have different electoral systems and political cultures.

This research may also be of interest to political parties, organisations that conduct polling, political scientists, and journalists.

1.2 Background

The datasets used in this thesis have been sourced from the Electoral Commission of South Australia (ECSA), the body responsible for conducting all state and local government elections in South Australia, and the ABS.

All people on the electoral roll are enrolled in one electoral district (also known as an electorate or seat), and each electoral district elects one member to the House of Assembly of the state parliament.

All votes cast in a state election are categorised as either 'ordinary' or 'declaration'. Ordinary votes are the result of a person visiting one of the polling places in their electoral district on election day. These polling places are typically, but not always, in schools, churches, and community centres. Ordinary votes are counted and reported in their polling place.

Declaration votes are the result of a person casting their vote in some other way, including by pre-poll, postal vote and by voting in a polling place *outside* their electoral district. All the declaration votes in each electoral district are treated as being cast in a declaration polling place for that electoral district.

The results from each election are reported by ECSA by polling place, and also by electoral district.

The demographic information used in this thesis is generated using an online tool provided by the ABS. We can obtain demographic information at both the electoral district and collection district levels. Demographic data is not available at polling place level as geographically they are just coordinates, not regions with a defined catchment.

This research is interesting from a statistical viewpoint partly because our electoral results and demographic predictor datasets do not align spatially.

The statistical models used in this thesis are all logistic regression models, with the response variables taking either a binomial or multinomial distribution. These models work by fitting a linear combination of predictors, with coefficients chosen that result in predictions that most closely fit the actual election results.

1.3 Thesis Outline

In Chapter 2 we review South Australia's electoral system and significant developments in its history. In particular, we review major debates and changes that have been made throughout South Australia's history to improve fairness (or perceived fairness) in the electoral system.

We also review the methodology currently used by the EDBC when it conducts electoral redistributions, including the method that this thesis aims to improve on. We review the way in which collection districts are defined and their characteristics.

Finally, we review the literature for research into the effect of demographics on voting behaviour. We find limited literature in an Australian context.

In Chapter 3 we identify from the literature a set of demographic factors to use as the predictors for this study, and perform an exploratory analysis of this data. We choose a set of 17 predictors, grouped into 4 predictor categories, and introduce a set of aggregate statistics, to permit simple rankings of collection and electoral districts for each predictor category.

We then develop a way of visualising each collection and electoral district predictor, and use this to understand the distributions of the values for the predictors, and how the distribution between predictors in a category changes as the value of the aggregate predictor changes.

We identify two characteristics of the data that impact on our models: the significantly larger spread in the predictors for collection districts than for electoral districts, and the highly correlated nature of the predictors.

Finally, we perform Principal Component Analysis on the predictors to reduce the correlation between the predictors, and investigate the factors that are most important in determining the variability in the dataset.

In Chapter 4 we begin to investigate the central question of this thesis, with a series of models which fit the predictors at electoral district level to the election results.

We fit three multinomial logistic regression models using the 17 raw predictors, the full set of principal components, and a restricted set of components.

Because we seek to calculate predictions for more than 3000 collection districts, we develop some tools to assess the validity of each model. We choose three electoral districts as case study districts and use these to investigate the usefulness of the models.

In Chapter 5 we introduce the voter location data, a dataset containing the number of people that vote in each polling place, from each collection district. We explore, verify, and clean the data.

The voter location data allows us to construct estimates for the predictors at polling place level. Using these estimates we fit the multinomial logistic regression models again, with the predictors at polling place level instead.

We then compare all of the multinomial logistic regression models through a number of summary statistics measuring how close predictions under each model are to the real election results.

This analysis leads us to a preferred model, which involves using the 17 raw predictors, but at polling place level. In all future chapters we only use this set of predictors to perform modelling.

In Chapter 6 we take a different approach and use both the predictors *and* spatial information. We know that there is a spatial nature to the system, and that local issues and candidates do play a part in voting decisions.

As the demographic predictors in this spatially-aware model we use the values of the predictors, relative to a measure of the average value of the predictors in the area. As well as this we use the estimate for the result in each collection district currently used by the EDBC in reviews of boundaries as an offset term. This estimate is a weighted average of the results at the polling places that people in each collection district voted at.

This construction results in a nested set of models, at each of the electoral district, polling place, and collection district levels.

Because this model is more complicated, we shift to fitting a logistic regression model by removing the informal votes from the system. We fit the model at both electoral district and polling place level.

This new approach gives a much better fit than the previous models that used only predictor data, and we adopt the spatially-aware model it as our preferred model.

In Chapter 7 we compare the predictions under our final preferred model to the predictions that were used by the EDBC in the redistribution following the 2010 election. While further research is required to establish the improved accuracy of our predictions, we argue that they are credible and overcome some clear shortcomings in the EDBC predictions. We therefore argue that these methods deserve further attention.

In the Conclusion we summarise the findings of this thesis, and suggest directions for further research.

Chapter 2

Literature Review and Background

In this chapter we explore the background and motivation for this research. We discuss the South Australian electoral system and its history, review the literature relevant to this thesis, and the key goal: the development of a new method for predicting the number of votes for each major party in each collection district in South Australia.

2.1 South Australia's electoral system

South Australia currently has two houses of parliament, similarly to most other parliaments in Australia. The houses are called the House of Assembly (often referred to as the lower house) and the Legislative Council (the upper house). This thesis is chiefly concerned with the process of forming government, and the Legislative Council does not play a role in this, so it will not be considered at all.

In order to form a majority government, a party (or group of members) must hold at least half of the seats in the House of Assembly.

There are 47 electoral districts in South Australia, and each of these elects a single member to the House of Assembly. The electoral districts form a partition of the state so that every person is represented by a single member of the House of Assembly. Every electoral district must also contain the same number of voters, within a 10% tolerance.

Elections for all 47 members of the House of Assembly are held concurrently every four years, on the third Saturday in March. If a vacancy arises in an electoral district in between general elections, a by-election is held to elect a replacement. Elections are conducted by the Electoral Commission of South Australia (ECSA) and voting is compulsory for all adult citizens (subject to some formal exclusions, including those of 'unsound mind')[1].

The system of voting used is a preferential voting system called the 'Alternative Vote'. Under this system each voter ranks the candidates in the election on their ballot in their order of preference. To be counted as *formal*, a vote must have all of the candidates ranked with consecutive numbers[2]. A ballot that has been marked in some way but is not formal according to the rules of the election is called *informal*.

A candidate is elected only when they have received an absolute majority (50% +1) of votes. If after totalling all the voters' first ranked candidates, no person has received an absolute majority, the candidate that has the least number of votes is eliminated and all of their votes are redistributed at full value to the candidate that

[1] Actually what is compulsory are the 'formalities' of voting: attending a polling place, having your name checked off, accepting ballot papers, and depositing them in the ballot box. This means that voters are permitted to submit blank ballots (there would be no way to police voters doing this if it were not permitted anyway, due to the nature of the secret ballot).

[2] There are circumstances under which a ballot is only partially completed but is still counted as formal — in all these cases a complete ordering of all candidates is inferred and used on the basis of a registered ticket submitted by a political party.

was next preferenced on the ballot.

This process continues until a single candidate has secured an absolute majority of the votes. If there are more than two candidates in the count when a person attains this threshold, ECSA continues to count and exclude in the same fashion until there are only two candidates left. The number of votes the final two candidates hold is recorded and called the *two-candidate preferred* count.

Majoritarian systems like this tend to be dominated by two major parties, and this tendency is known in the political science literature as Duverger's Law, named after Maurice Duverger who wrote:

> *The simple majority single-ballot system favors the two party system*[3]. Of all the hypotheses that have been defined in this book, this approaches the most nearly perhaps to a true sociological law. [15, p217]

Note that this is not to say that the party system existing in South Australia is a consequence of the electoral system. Rather, the two systems exists mutually, where the electoral system both influences the party system and is itself created by parties that exist in competition between each other [8].

In South Australia, the two parties that dominate the lower house are the Australian Labor Party (often abbreviated to Labor Party or ALP) and the Liberal Party (in this thesis we will often use the abbreviation LIB in mathematical notation to mean Liberal Party). All South Australian Premiers in over a century have come from one of these parties or their predecessors. Numerous other parties exist but, in the current environment, no commentator regards any of them to have a plausible prospect of forming a government.

Independent members can also be elected to the House of Assembly, this is an

[3] Emphasis in original.

uncommon but regular occurrence. There are currently two independents in the lower house — Geoff Brock representing the electoral district of Frome, and Martin Hamilton-Smith representing Waite (however Hamilton-Smith was elected with the Liberal Party and later resigned his membership of the party). The 2010 election saw three independent members re-elected to the House of Assembly.

The rules of formality have the consequence that regardless of who the voters' most preferred candidate is, all formal ballots can be examined to determine whether the voter preferred at one point or another in their ranking the Labor Party over the Liberal Party, or vice-versa.

In each election the ECSA examine the formal votes in this way to determine the *two-party preferred* vote, which gives the support in every district for each of the two major parties when pitted against only the other major party. In most electoral districts the two-party preferred and two-candidate preferred votes are the same, but they differ where a candidate that does not come from one of the two major parties finishes first or second in an election. ECSA have been required to produce a two-candidate preferred result for each election since amendments were made to the Electoral Act in 1976 [2].

In 2010 the two-candidate and two-party preferred results were different in four electoral districts: Chaffey, Fisher, Frome, and Mount Gambier.

For this entire thesis we only use the two-party preferred results, and treat the system as having only two parties, ignoring all minor parties.

2.2 South Australian electoral history

In this section we review the development of the current electoral system and rules governing electoral boundaries, with a focus on changes that were made with an intention to improve fairness in the system. This section is *not* an exhaustive review of the history of South Australia's electoral system. While we are chiefly concerned with the current situation, it is useful to understand the historical background behind fairness provisions currently in place.

It goes without saying that we seek an electoral system that is perceived by all stakeholders to be 'fair'. Rather than trying to comprehensively describe the criteria by which a system is regarded as fair, we consider two of the common characteristics of systems that are often regarded as being unfair - *malapportionment* and *the geographic concentration of partisan support.* Newton-Farrelly [29] provides a history of the ways both of these have manifested themselves in South Australia.

The majoritarian system used in the South Australian House of Assembly is not concerned with the fair representation of minor parties and independents. Most minor parties have low support that is dispersed across the whole of the state and this is not strong enough in any individual electoral district to win them a seat. When we consider fairness under this system we refer specifically to fairness in the formation of government.

Malapportionment

The most contentious aspect of the electoral system in the history of the parliament has been that of apportionment and electoral boundaries. A commonly accepted principle in modern electoral systems is that of 'one vote, one value'. That is, that every enfranchised person has a vote of equal influence. This principle was in fact *not* adhered to for over a century following the formation of the parliament in 1856[4].

Newton-Farrelly [29] writes that the first distribution of the House of Assembly

[4] The parliament was first elected and sat in 1857, but South Australia's constitution was ratified in 1856 by the British parliament.

"allocated twice as many seats to the rural areas as to the state capital, Adelaide, even though the balance of population was by far the other way". In other words, there were far fewer voters in rural seats, meaning that the relative influence of each rural voter over the makeup of the lower house was much higher than that of voters in Adelaide.

This malapportionment was intentional. There were two main competing arguments regarding what the primary consideration should be when constructing electoral boundaries. One supported the principle of 'one vote, one value' and argued that all electoral districts should contain the same number of votes, while the other argued that a stronger voice should be given to rural voters because of their special needs and contribution to the state economy. The latter argument was successful in implementing their proposal for over a century, and malapportionment was a feature of the South Australian parliament until 1975.

At first, the malapportionment in favour of rural areas was informal, with proposals made by *ad hoc* commissions and approved by the parliament. This was formalised in a 1936 Act which created a system with 39 single member districts, preserving a 2:1 ratio in favour of the rural areas. Dean Jaensch [20] describes the resulting electoral district sizes:

> The 1936 Bill adopted the procedure used by the Commonwealth government of allowing ±20 per cent margin of the arithmetic quota arrived at for the formation of single member districts. The quota for the metropolitan area was 15,665 and for the extra metropolitan districts 5,718, a ratio of almost 3:1 in favour of the latter. [20, p83]

Malapportionment of this scale means it is simple to construct situations where one party receives a majority of votes when aggregated across the state, but wins a minority of the seats in the House of Assembly (and thus is unable to form government).

In the same paper Jaensch considers the malapportionment in more detail and demonstrates that there is a case to be made that the South Australian government from 1933-1965, under the leadership of Liberal and Country League[5] Premier Thomas Playford, lasted for so long because of a "gerrymandered electoral system". By "gerrymandered", we mean a system in which electoral district boundaries have been drawn in such a way to create or maintain an unfairness to a particular party or parties.

Jaensch calculates a notional two-party preferred count for each election in this time period by redistributing preferences from each election. For districts that were uncontested by either of the major parties, results from the nearest federal election were used.

This methodology results in the indication that "the ALP should, on the basis of electoral support, have formed a government in six of the eight elections" [20, p85].

The paper goes on to discuss whether this was an intentional gerrymander by the Liberal Country League, or alternatively, a "silent gerrymander"[6]. The discussion focusses on the malapportionment at the time that resulted in rural voters being over-represented in the parliament, and whether this was deliberately maintained for political purposes (to win elections) or for moral purposes (a rejection of the "one vote one value" principle in favour of another principle).

After some preliminary reforms in 1969, the rural malapportionment came to an end in 1975, when it became a requirement that the principle of "one vote, one value" be applied to all redistributions. The current requirements are that each electoral district have the same quota, and all districts must contain this number of voters, within a tolerance of 10%. In practice the Electoral Districts Boundaries

[5] Equivalent to the Liberal Party today.

[6] The maintenance of rather than the creation of an unfair advantage.

Commission (see Section 2.3) aim to have the number of electors[7] in each electoral district within 3.5% of the quota at the time of a future election [29, p473].

Geographic concentration of partisan support

Another perceived cause of unfairness related to the drawing of electoral boundaries is geographic concentration of partisan support. Put simply, this refers to situations in which high concentrations of support for one party are contained in one electoral district — meaning the district is safely held by that party, but a lot of those votes are surplus to need, or 'wasted' for that party, as they cannot be counted in another electoral district.

A party having strong concentrations of support in small areas can distort the statewide result, by increasing the proportion of votes achieved by that party across the state, while not affecting the makeup of the lower house.

Jaensch gives an example of how the electoral boundary rules in 1936 "virtually made it necessary that the ALP voters outside the metropolitan area be concentrated in few districts" [20, p91].

This was because six large country towns (Gawler, Mount Gambier, Murray Bridge-Tailem Bend-Mannum, Port Pirie, Port Augusta-Whyalla and Wallaroo-Kadina-Moonta) were areas of very strong support for the ALP, and were large enough to form the bulk of their electoral districts. This concentrated rural ALP voters into very safe ALP seats containing these towns.

A more contemporary example of this differential concentration can be seen in the results of the 2014 state election. In this election the ALP won 23 seats, and 8 of these were with a two-party preferred result above 60% for the ALP. The Liberal

[7] The number of people on the electoral roll.

Party 'won'[8] 24 seats and of these, 16 gave a two-party preferred result above 60% for the Liberal Party. This indicates that the Liberal Party had more votes 'wasted' in safe seats than the ALP.

2.3 Electoral Districts Boundaries Commission

Initially, redistributions of electoral boundaries in South Australia were conducted and approved by the parliament (through the use of *ad hoc* committees). From 1882 *ad hoc* Electoral Commissions were created to propose redistributions, but their recommendations were still approved by the parliament.

The fact that redistributions were a decision of the parliament meant that proposals were often not adopted. In a research report to the South Australian State Electoral Office in 2002, Jaensch [21] summarises the history of redistributions, and writes that:

> In the colonial period, following the 1857 election, there were 14 attempts to introduce a full redistribution of the electorates: three were successful — 1861, 1872 and 1882. Major redistributions were also carried out in 1901, 1913, 1929, 1937, 1955, and 1969...
>
> In 1932, the government considered various proposals for redistribution, but did not proceed. In 1962, an Electoral Districts Redivision Act was carried through the parliament, and an Electoral Commission reported with a full redistribution. However, the Government failed to achieve the necessary constitutional majority in the House of Assembly and the redistribution Bill lapsed. [21, p154]

[8] They actually only won 22 seats, but Fisher and Frome were won by independents. On two-party preferred terms in these two seats, the Liberal Party received more votes than the ALP.

In 1929 an Electoral Commission was created to perform a redivision of the electorates, and this Commission was given the powers of a Royal Commission. Aside from the quotas discussed earlier, the Commission had a number of other criteria (all of which were secondary to the quotas in each district):

> The Commission was charged to also consider —
>
> **(a)** community or diversity of interest;
>
> **(b)** means of communication;
>
> **(c)** physical features; and
>
> **(d)** boundaries of existing Assembly districts and subdivisions. [21, p173]

In 1975 a permanent body was created to conduct redistributions, called the Electoral Districts Boundaries Commission (EDBC). This body was given the authority to declare boundaries without parliamentary approval. Any appeals against the decisions of the EDBC would have to be taken to the Supreme Court.

The EDBC currently consists of a senior judge of the Supreme Court (appointed by the Chief Justice), the Electoral Commissioner and the Surveyor-General.

The 1989 state election resulted in the ALP forming minority government with the support of two independent Labor MPs, despite the fact that the Liberal Party received 52% of the two-party preferred vote across the state. This result led to a debate on both growing malapportionment, and a lack of 'fairness' in the result that the Liberal Party could not form government despite receiving more votes than the ALP.

The Fairness Clause

Since 1975, the Liberal Party had been demanding a principle that a party receiving more than 50% of the vote should be able to form a majority government. This was not a principle that the ALP agreed with at first, with Labor Premier Don Dunstan citing concerns about bringing party politics into redistributions:

> Electoral Commissioners should not draw boundaries according to the political points of view of the electors. That is just what they ought not to be doing, because, if they do that, they will introduce Party politics into their consideration of electoral boundaries. [5]

Labor's position changed in the subsequent years, and in 1991, amendments to the Constitution were made by referendum to require redistributions to occur after every general election, and also to insert a 'fairness clause' (Section 2.3.1).

2.3.1 Current rules for redistributions

Aside from the requirement that all electoral distributions contain the same number of electors (within a 10% tolerance), the requirements for fairness and other criteria in drawing electoral boundaries are contained in Section 83 of the *Constitution Act*:

> **(1)** In making an electoral redistribution the Commission must ensure, as far as practicable, that the electoral redistribution is fair to prospective candidates and groups of candidates so that, if candidates of a particular group attract more than 50 per cent of the popular vote (determined by aggregating votes cast throughout the State and allocating preferences to the necessary extent), they will be elected in sufficient numbers to enable a government to be formed.

> **(2)** In making an electoral redistribution, the Commission must have regard, as far as practicable, to
>
> > **(a)** the desirability of making the electoral redistribution so as to reflect communities of interest of an economic, social, regional or other kind;
> >
> > **(b)** the population of each proposed electoral district;
> >
> > **(c)** the topography of areas within which new electoral boundaries will be drawn;
> >
> > **(d)** the feasibility of communication between electors affected by the redistribution and their parliamentary representative in the House of Assembly;
> >
> > **(e)** the nature of substantial demographic changes that the Commission considers likely to take place in proposed electoral districts between the conclusion of its present proceedings and the date of the expiry of the present term of the House of Assembly,
>
> and may have regard to any other matters it thinks relevant.
>
> **(3)** For the purposes of this section a reference to a group of candidates includes not only candidates endorsed by the same political party but also candidates whose political stance is such that there is reason to believe that they would, if elected in sufficient numbers, be prepared to act in concert to form or support a government. [1]

Clause (1) above is referred to as the 'fairness clause'. South Australia is the only state in Australia with this fairness clause, and there is no equivalent provision for the Federal Parliament. All of these electoral jurisdictions have a similar body to the EDBC, however each operates according to different rules.

Table 2.4.1: Two-party preferred result across all of South Australia for all state elections held since 1991, along with the party that actually formed government following the election. The bold lines are those elections at which one party received a majority of the two-party preferred vote but could not form government.

Year	ALP (%)	Liberal Party (%)	Government formed by
1993	39.0	61.0	Liberal Party
1997	48.5	51.5	Liberal Party
2002	**49.1**	**50.9**	**ALP**
2006	56.8	43.2	ALP
2010	**48.4**	**51.6**	**ALP**
2014	**47.0**	**53.0**	**ALP**

2.4 The fairness clause in operation

To examine the effectiveness of the fairness clause, we look at the results of elections since it was enacted in 1991. Table 2.4.1 shows the two-party preferred results for all state elections held since 1991.

The table shows that of the six elections held since the adoption of the fairness clause, there were three in which the party that received less than half of the two-party preferred vote was able to form government. In every case, the ALP was the beneficiary rather than the Liberal Party.

The fact that so often the goal of the fairness clause has not been met so often indicates that either the methodology used by the EDBC is insufficient, our characterisation of fairness is too difficult to implement in the first place, or that major boundary changes are required to meet the requirements of the fairness clause.

It is worth nothing that despite South Australia being the only jurisdiction with a fairness clause like this, it is also seemingly the state in which the 'wrong' party is able to form government most often. This could suggest that major boundary changes are needed, but the EDBC are hesitant to make major changes which would result in very large numbers of voters being moved to new electoral districts.

Many commentators argue that the characterisation of fairness is too hard to implement, including Jaensch [22] who in *The Advertiser* in March 2014 argued that the parliament should look into other systems to ensure fairness:

> I make no criticism of the commission. Its members are a senior Justice, the Electoral Commissioner and the Surveyor-General. If any people can be trusted to be members of a permanent and independent statutory body, they can [...] The fault lies in the politicians in the Parliament who set up an objectove [*sic*] that is simply not attainable.
>
> [The process] assumes the voters in the previous election will cast their votes in the same direction in the next election. If they do, the fairness clause will actually work as it is designed to do. But many voters do not act that way. They react to different political and personal circumstances, and change their votes. [22]

Additionally, William Bowe [10] writing in *Crikey!* says that:

> The difficulty is that electoral boundaries are an extremely blunt instrument for ensuring the majority party wins office, and they can only promise to deliver if the overall swing is precisely uniform. The last election in 2010 made a mockery of that assumption, with Labor copping huge but ultimately harmless double-digit swings throughout its Adelaide heartland, while fighting brilliantly successful rearguard actions where it mattered most.

2.4.1 Methodology of the EDBC

Knowing that an election result is unfair and fixing the unfairness are two entirely different jobs for the commission. It is more difficult to measure the size of disadvantage and fix it than it is to identify its existence in the first place. In fact, in 1976 the Commission was asked to consider a fairness clause and rejected it, saying that:

> [w]e are not satisfied, after a full consideration of the evidence presented to us, that there is any reliable method of forecasting how electors will vote next time...
>
> predicting voting patterns in future elections seems to us, with respect, to involve an interpretation of incomplete statistical data, a series of assumptions as to uncounted preferences votes, and a measure of oneiromancy[9] [3, p12-13].

Despite this, following the introduction of the fairness clause the EDBC has developed a relatively consistent approach to their work, and it is described in their reports. An example of this is the 2007 report immediately preceding the 2010 election, which gives the consideration behind the boundaries we consider in later chapters [4].

The EDBC has chosen to regard the expression "the popular vote" as referenced in Section 83(1) of the *Constitution Act* (see Section 2.3.1) to be "equivalent to the two-party preferred vote calculated on a State-wide basis" [4, p6]. Section 83(3) causes some complications, as the Commission need to make a decision about whether or not independent candidates have a political stance that would lead them to support one party or another to form government. In the case of the 2007 report the EDBC

[9] oneiromancy, n. Prediction of the future by the interpretation of dreams. Oxford English Dictionary [38].

made the decision to use the two-party preferred count in all electoral districts to determine the "popular vote" [4, p12-13].

Broadly speaking, the Commission first adjusts the boundaries to meet the condition that all electoral districts have the same number of electors within a 10% tolerance (contained in section 77 of the *Constitution Act*). It then tests proposed boundaries against a uniform statewide swing that gives a 50%+1 result, and verifies if the desired outcome transpires. If it does not, then further adjustments are made until they do [4, p14-15].

The Commission notes that s77 of the *Constitution Act* requires the ±10% tolerance be adhered to, and s83(1) requires fairness (as far as practicable), but s83(2) only requires the Commission to 'have regard to' the other criteria. This means that the fairness and equity rules for redistributions are the most important criteria for the Commission to satisfy, at the expense of other guidelines. Newton-Farrelly [29] describes an example of this issue from the 1991 redistribution:

> It moved Kangaroo Island from its traditional electorate which covered Fleurieu Peninsula (the closest mainland area to the island) to the electorate that covered the Eyre Peninsula (the furthest peninsula from the island). On the island the move was seen as a breach of all of the commissions obligations, particularly that of respecting communities of interest.
>
> [...] the commission was resolute, citing the fact that in the changes [...] the wording of the act had been changed to downgrade the importance of the community of interest criterion. A more recent redistribution has returned Kangaroo Island to its original seat, but the commission has made its point: geographic districts - even suburbs - will be split where required by fairness or equity. [29, p476]

The EDBC move small areas between electoral districts when the conduct redistributions. In order to test fairness, the Commission must therefore have an estimate for the support for each apart in small geographical areas. The unit of geography used for this purpose is the Census Collection District, or simply 'collection district'. See Section 2.5 for a full description of collection districts.

Newton-Farrelly describes the methodology used by the Commission to calculate estimates for the two-party preferred vote in each collection district. At first (in 1991 and 1994), the commission simply assumed that the electors voted at their nearest polling place and so took the nearest polling place results as the results for each collection district. This is a demonstrably false assumption.

After the first two redistributions under the new rules, a more sophisticated model was developed. Every voter on the electoral roll is linked to the collection district that they reside in. After each election, the electoral rolls used at each polling place are electronically scanned to see which one each elector had attended.

From this, an estimate for each collection district is calculated as the weighted value of the two-party preferred vote at each of the polling places attended by the voters in that collection district. This value is now used by the EDBC when conducting redistributions.

A discussion of the method for redistribution after the Commission has an understanding of the geographic distribution of partisan support is beyond the scope of this project. We focus instead on gaining an understanding and developing a more accurate and precise method for estimating the distribution of partisan support, using additional sources of data.

Once the EDBC has finalised any boundary adjustments, it then calculates a notional margin for each electoral division. However, these calculations have been contested by commentators, including Green [19], who, ahead of the 2014 state election said that:

> The EDBC has estimated the margin for Light falls from 5.4% at the 2010 election to 4.2% on the new boundaries. My estimate weighting the declaration vote produces an estimated margin of 2.8%, the value the ABC will be using [...]
>
> I'm not the only person to have noted the discrepancy. Similar calculations carried out independently for the South Australian Parliamentary Library are very close to mine and differ from the EDBC's estimates in a similar manner [...]
>
> On the EDBC pendulum, Labor would be reduced to a minority by the loss of two seats on a swing of 1.7. On my pendulum, it is three seats on a swing of 0.6%. On the EDBC pendulum Labor has five seats on margins under 3%, on my pendulum seven.

Examining the method that the EDBC uses to calculate these notional margins for electoral district is beyond the scope of this thesis.

The key question we investigate in this project is the accuracy of the estimates of partisan support in collection districts, and whether they can be improved by incorporating other sources of data, namely socio-economic data obtained from the Census. This could assist the EDBC in performing future redistributions. In addition we demonstrate the legitimacy of the link between demographic predictors and voting behaviour.

2.5 Collection Districts

The collection district, formally called the Census Collection District, is the second smallest spatial unit and the basic building block in the Australian Standard Geographical Classification. The smallest unit is called the Mesh Block but output data is not available at this level. Every person in the Census is enumerated in a single collection district.

The idea is that a collection district covers an area that a single census collector can cover (delivering and collecting forms). In 2006 there were around 225 dwellings in each collection district in the metropolitan area, with fewer in rural areas. Collection districts are redefined for each Census and so are not comparable between Censuses (although the ABS does aim to maintain comparability as much as possible).

The set of collection districts partitions the country. Individual collection districts do not cross local government boundaries but they can cross state electoral boundaries. Beyond this, the general principles for constructing collection districts are that they should follow identifiable permanent features where possible, such as the centre of roads and rivers, and conform where possible to existing suburb boundaries.

As well as the ordinary collection districts there are special districts, including for those who are off-shore, on a body of water (such as on a cargo vessel or passenger liner), and travelling overnight.

Collection districts are the smallest unit of geography in this thesis, and the goal of this research is to estimate the voting behaviour of the group of voters in each collection district.

2.6 Factors that influence voting intention

We plan to use Census data to inform our estimates of the partisan support in each collection district. In this section we review the existing literature around voting behaviour and what is already known about factors that influence it.

The focus is on literature that takes a statistical approach to infer the factors that influence how people vote. To our knowledge there is a very limited literature base on the topic, with much room for further exploration.

We restrict our literature review to just the Australian context as different countries have very different political cultures, and conclusions drawn in other countries do not necessarily hold in Australia. Additionally, most other countries, including the United States and United Kingdom, have voluntary voting systems and so the question is not just about which party they vote for, but whether they even vote at all.

Most of the Australian literature on this topic is of a non-statistical nature, but there are some useful studies.

One large dataset that can be used to analyse voting intention is the series of *Australian Election Study* (AES) surveys. These surveys have been conducted after every federal election since 1987, and ask a nationally representative sample of voters questions about their engagement with the election campaign, their party preference, their opinions on political leaders and election issues, their general political and social views, and their demographics. The datasets for each survey are available from the Australian National University [6].

In Chapter 15 of the book *Government, politics, power and policy in Australia*, Haydon Manning [41] uses AES data to study the loyalty of various types of voter to political parties. In particular, he looks at the influence of a person's social class on their voting habits, and the decline in this 'class voting' in recent decades.

Indeed, the way the major parties formed in Australia presupposes a divide in the political preference of different social classes, with Manning observing:

> The Australian Labor Party was originally formed by trade unions in the 1890s with the explicit aim of protecting wage-earners considered to

> be members of the ‘working classes’. Today’s Liberal Party was reborn in 1944 ... as a party with a particular attraction for the ‘middle classes’ [41, p274].

Manning uses AES 2007 data as evidence that wealthier voters (measured solely by household income) are more likely to vote for the Liberal Party and less wealthy voters more likely to vote for the ALP. However, this division is weakening over time. In particular, “occupational class seems to have a weaker association with voting behaviour than it did in the past” [41, p285].

Clive Bean and Ian McAllister [7] use AES 2007 data and perform a multivariate analysis to estimate the net effect of a number of factors on the vote, including a number of socio-demographic variables, campaign issues, and the party leaders themselves. They claim that no demographic variables were important in voting at the federal election in 2007. The argue:

> Of all the variables considered throughout the chapter, the only ones that had statistically significant effects on the vote in the 2007 election, net of all the other variables in the equation, were party identification, the issues of taxation, industrial relations and health, and the evaluations of Howard and Rudd. None of the social background variables had any direct impact and nor did most of the campaign issues [7].

Party identification is by far the strongest driver of voting intention in this study. However, it may be that party identification itself can be predicted using a broader range of predictors. This could equally apply to variables regarding attitudes on issues like industrial relations and tax.

Bean and McAllister’s findings are disputed by Goot and Watson [18], who use AES data from the 1996 to 2004 federal elections to identify factors responsible for the electoral success of Prime Minister John Howard and the federal coalition of the Liberal and National parties. They claim that socio-economic demographics provide a partial explanation:

> We document the shift to the Coalition, net of other factors, among older respondents, those with little education, and Catholics. And we show the shift to Labor among respondents from non-English speaking backgrounds [18, p253].

These conclusions come from comparing AES data from 1987-93 (elections won by the Labor Party) with the Howard years (1996-2007). They also identify individual policy issues that mattered most in the shift from the Labor Party to the Coalition.

Goot and Watson describe the AES as indispensable yet frustrating, as the dataset does not allow them to test all theories for the Coalition's electoral success.

Of the small literature base, there is very little research into voting behaviour in South Australia. This contributes to making this thesis novel, as we test this proposition that there is a relationship between demographics and voting behaviour.

2.7 Discussion

In this chapter we have reviewed the current South Australian electoral system and the development of rules governing fairness over time. We have considered the EDBC and its methodology for meeting the requirements of the fairness clause in the Constitution. Finally we have considered the literature base informing our current understanding of demographic factors that influence a person's voting decisions, and found limited literature.

The EDBC uses predictions for the two-party preferred vote in each collection district in their methodology of drawing electoral boundaries. These predictions are calculated using only the returns from the election and assume the homogeneity of all voters attending each polling place. Their methodology also assumes that the voters' support for each party does not change from one election to the next.

In the next chapter we introduce and explore both the 2010 election results data from ECSA and demographic data from the 2006 Census, to see how it could better inform these predictions.

This research is significantly different to all that has been found in the literature. This study will consider socio-economic variables as predictors for voting intention over much smaller geographical regions than have been considered in the literature.

The research will also consider elections from a South Australian political context, rather than in a federal context.

Chapter 3

Data

Before performing any modelling, we first investigate our datasets. In this chapter we identify and explore the predictors that will be used in the analyses in this thesis.

We first develop some basic notation for referring to the data, and then investigate the structure of the predictors through visualisations, various aggregate statistics, and principal component analysis techniques.

3.1 Raw Data

In this thesis we only use results data from the 2010 South Australian election, and demographic data from the 2006 Census of Population and Housing. This is because the 2010 election was the most recent election at the commencement of this research, and the collection districts used in the ECSA reports are those defined in the 2006 Census.

We cannot work with the 2010 election results and the 2011 Census, as different sets of collection districts are used in these two datasets.

Obviously the 2006 Census data will not precisely reflect the demographics of collection districts at the time of the 2010 election, but it is the most accurate and complete dataset that exists.

ECSA Data

These datasets contain full information and results from the 2010 state election, including all voting data that is available to the EDBC. These include electoral district records, polling place records, candidate records separated by polling place, and candidate records aggregated across polling places.

We also have access to election results for the Legislative Council, but that does not fall within the scope of this project. This data has been provided by the Deputy Electoral Commissioner, in a set of comma delimited text files.

For each district, we have

- The total number of formal votes cast in the district;
- The total number of informal votes cast in the district;
- The total number of formal declaration[1] votes in the district;
- The total number of informal declaration votes in the district;
- The total number of formal first preference votes for each candidate in the district;
- The two-candidate preferred vote in the district;

[1] Recall that a declaration vote is a valid vote that is not cast in a polling booth in the appropriate electoral district on the day of an election. These include votes that are cast before the day of the election, postal votes, votes cast by citizens internationally, and votes cast by people that have attended a polling place in an electoral district that they do not reside in.

- The two-party preferred vote in the district;
- The total number of declaration votes for each candidate in the district;
- The two-candidate preferred vote for the declaration votes in the district; and
- The two-party preferred vote for the declaration votes in the district.

For each polling place, we have

- The total number of formal votes cast in the polling place.
- The total number of informal votes cast in the polling place.
- The total number of votes cast in the polling place.
- The total number of formal first preference votes for each candidate in the polling place.
- The two-candidate preferred vote in the polling place.
- The two-party preferred vote in the polling place.

We also have data giving an indication as to where electors in each collection district vote, in line with the approach taken by the EDBC in redistributions. For each collection district, we have a record telling us the number of people that voted in each polling place.

ABS Data

The data available from the ABS includes full results from the 2006 Census, aggregated over geographical areas of varying sizes, including the smallest unit of collection district.

Data is available for individuals or households, and include characteristics such as ethnicity and language, employment and income, children, family and relationships, and education (to name a very small number of the available covariates). All of the covariates are categorical variables, often with a number of levels (in the case of variables such as income).

We are able to generate tables giving the number of people in each collection district in each category.

This data is already available to us, thanks to an agreement between the ABS and Universities Australia. Using a web tool called TableBuilder we are able to construct and download tables containing full data, in various forms.

In order to preserve the confidentiality of Census records and ensure that information likely to enable identification of individuals is not published, the ABS introduce a small amount of random error to the dataset. The full detail of this introduced random error is not publicly available, but the Census dictionary [33, p201] describes the method in general terms:

> A technique has been developed to randomly adjust cell values. Random adjustment of the data is considered to be the most satisfactory technique for avoiding the release of identifiable Census data. When the technique is applied, all cells are slightly adjusted to prevent any identifiable data being exposed. These adjustments result in small introduced random errors. However the information value of the table as a whole is not impaired.

Proportions calculated on very small collection districts could be significantly impacted by the introduced random error. We have no option but to ignore the introduced random error, noting that it only impacts in collection districts with a very small number of votes (and hence are relatively unimportant when it comes to redrawing electoral boundaries).

3.2 Electoral Data

3.2.1 Electoral District Results

The electoral district data is taken from the full set of results and is stored in a CSV. A sample of the electoral district results appears in Table 3.2.1. Each row of the table represents an electoral district, and for each electoral district we have the two-party preferred vote for each party, and the number of informal votes. There are two equivalent identifiers for each electoral district, its name and its ID (a unique number assigned by the ECSA between 601 and 647).

Table 3.2.1: Sample of the two-party preferred results (including informal votes) from the 2010 South Australian election, by electoral district.

Electoral District	District ID	ALP	Lib	Informal
Adelaide	601	9211	10909	787
Ashford	602	11625	9588	664
Bragg	603	6148	15297	416
Bright	604	10777	10610	641
⋮	⋮	⋮	⋮	⋮
Wright	647	12125	10073	711

3.2.2 Polling Place Results

The polling place data is also separated from the full set of results and stored in a CSV. A sample of the polling place results used is shown in Table 3.2.2. Each row of the table represents a polling place. Every polling place has a unique ID, and we have the total number of votes cast in each polling place as well as the breakdown of the votes between the ALP, Liberal Party, and informal votes.

Table 3.2.2: Sample of the two-party preferred results (including informal votes) from the 2010 South Australian election, by polling place .

ElectoralDistrict	PollingPlace	ALP	LIB	INF	VotesInPollingPlace
601	5	445	483	30	958
601	6	197	204	12	413
601	7	792	708	60	1560
601	8	371	221	20	612
601	10	335	537	29	901
601	11	888	1441	83	2412

All of the election results data is scrutinised very closely by stakeholders in the election, and so it is regarded as clean and consistent.

3.2.3 Voter Location Data

This dataset, supplied by ECSA, tells us the distribution of voters in each collection district across each polling place in the relevant electoral district, including declaration voters. This data requires verification.

We explore and clean the voter location dataset later, in Section 5.1.

3.3 Census Data

We obtain data for the predictors aggregated by electoral district and collection district directly from the ABS TableBuilder [35]. This data is clean and consistent and no data verification is required. Based on the findings of the literature review in the previous chapter, and consultation with Professor Clement Macintyre from The University of Adelaide [26], we choose to initially use predictors from four categories: weekly household income; language spoken at home; school education; and non-school (tertiary) education. This sample of four categories represents a set of socio-economic factors that may influence voting activity. As part of our modelling we test this hypothesis.

Within each predictor category we have a number of aggregated predictors, as indicated in Table 3.3.1. The covariates used are aggregated from the data and again chosen based on expert opinion [26].

Each predictor category has an aggregate predictor noted as ‘Other’. This contains people who for whatever reason did not answer the question or for whom the question was not applicable.

The Household Income predictor is calculated by the ABS from individual income data provided by respondents. Each individual aged over 15 in a household reports their income before tax, superannuation contributions, health insurance, amounts salary sacrificed, or any other automatic deductions [34]. These incomes are collected in ranges.

The individual income ranges are then converted to the median dollar amounts for those ranges using data obtained through other ABS surveys, and household income is calculated by summing each of the individual incomes in the household.

If any person in the household aged 15 years or over does not report their income

Table 3.3.1: Predictors used in analysis with labels, by predictor category.

Predictor Category	Aggregated Predictor	Label
Household Income (weekly)	Negative or Nil	≤ 0
	\$1 - \$499	\$1 - \$499
	\$500 - \$999	\$500 - \$999
	\$1000 - \$1399	\$1000 - \$1399
	\$1400 - \$1999	\$1400 - \$1999
	\$2000 - \$2499	\$2000 - \$2499
	\$2500 - \$2999	\$2500 - \$2999
	\$3000 - \$3499	\$3000 - \$3499
	\$3500 - \$3999	\$3500 - \$3999
	More than \$4000	≥ \$4000
	Other	N/A
Language Spoken at Home	English	English
	Not English	Not English
	Other	N/A
School Education	Completed Year 12	Year 12
	Did Not Complete Year 12	< Year 12
	Other	N/A
Non School Education	Certificate Level	CL
	Bachelor Degree or Advanced Diploma	BDAD
	Postgraduate Degree or Graduate Diploma	PDGD
	Other	N/A

then the household is counted as ‘Partial Income Stated’ [33] and is included in the ‘Other’ category for our purposes.

The school education data is calculated from the responses to the question ‘What is the highest year of primary or secondary school the person has completed?’ All responses of lower than Year 12 level are grouped into one predictor.

The non-school education data is calculated from the responses to the question ‘What is the level of the highest qualification the person has completed?’ Answers to this question are written in by the respondent and then interpreted by the ABS and sorted into groups. A person is included in the data under the ‘Other’ category for our purposes if they:

- have not completed an education qualification;
- are still studying for a first qualification;
- have a qualification outside the scope of the question; or
- are aged under 15 years.

The language predictor category comes from the question ‘Does the person speak a language other than English at home?’ If they answer ’yes’, respondents are asked to specify the language they speak at home most often. We group all of these non-English responses into one predictor. Auslan and other sign languages are treated as languages other than English for our purposes. Full information about each predictor category is available in the 2006 Census Dictionary [33].

We expect there to be strong correlation structure between these variables, particularly between the household income predictors and the education predictors. Many authors, including White [40] and Filmer [16] have explored these relationships using data from around the world.

The ABS provide data for each predictor category in tabular form, and reports the number of people (or households) in each predictor category from that region. A sample of the non-school education data is given in Table 3.3.2 (shown for collection districts, the structure is identical for electoral districts).

Table 3.3.2: Sample of the ABS data for the Non-School Education predictor category, from the 2006 Census of Population and Housing, by collection district.

Collection District	Other	BDAD	PDGD	CL
⋮	⋮	⋮	⋮	⋮
4121513	175	61	18	19
4121601	246	112	42	46
4121602	537	184	40	73
4121603	314	136	39	56
4121604	421	178	49	54
⋮	⋮	⋮	⋮	⋮

It is worth noting that the covariates in our model are not factors in the typical statistical sense, as the predictors are all interrelated. Here we are not working with unit vectors.

In a typical situation, every unit record would be in exactly one factor in each category. Here though, each unit record is a collection district, and each predictor is a proportion of the collection district in that aggregated category.

We write x_{im} to be the value of predictor m in region i. Note that if the x_{im} are represented as the number of people in each predictor, then if there were two districts with identical proportions for the predictors, but one had more voters, then that district will have a greater influence on the predictions. So as to avoid having larger districts dominate over smaller ones, we convert the predictor data into proportions within each category.

Thus, within each predictor category, the predictors must sum to one. Therefore, we remove the ‘Other’ predictor from each category without losing any information. This is necessary when we perform the model fit as we require the columns of our model matrix to be algebraically linearly independent. We do not remove any further predictors, as doing so would remove information (as there is demographic information in the ’Other’ predictor). We are therefore left with $M = 17$ predictors for this analysis. The models in this thesis will all be of the general form

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{im}.$$

3.3.1 The ‘Other’ Category

We consider the distribution of the ‘Other’ predictors. The means and variances of these predictors are contained in Table 3.3.3.

Table 3.3.3: Summary statistics on ‘Other’ predictors

Predictor category of ‘Other’ predictor	Mean	Variance
Language Spoken At Home	0.04650351	0.002615525
Household Income	0.2237498	0.01260773
School Education	0.2542419	0.003789013
Non-school Education	0.7035851	0.008816181

The variances of the values of these predictors are all quite small (all are at least one order of magnitude smaller than the corresponding means), so we conclude that the spreads of these predictors are all quite small.

The response for ‘Other’ Language Spoken At Home is very low because it is only made up of those who did not answer the relevant question on the Census form. That is, the non-response rate for this question is approximately 4.65%.

For the Household Income category, if any person over the age of 15 living in a household did not state their income then the entire household is recorded as 'Partial Income Stated' and will then be counted in our 'Other' category. The category also contains non-private dwellings (including but not limited to hotels, motels, psychiatric hospitals, staff quarters, prisons, boarding schools, and ski lodges), unoccupied private dwellings (including vacant houses and holiday homes), and other non classifiable households.

For the School Education category, the 'Other' predictor contains all those people that did not answer the relevant question and also all people that are under the age of 15. Approximately 18% of the population of South Australia in 2011 [36] were under the age of 15. so these people make up the bulk of the 'Other' category.

For the Non-school Education category, the 'Other' predictor is much higher than for the other categories. This is because it includes all people who did not answer the question, have not completed an education qualification, are still studying for a first qualification, have a qualification outside the scope of the question, or are aged under 15 years.

3.4 Exploring Collection Districts

3.4.1 Distributions of predictors

We first visually inspect the distributions within each predictor category. Consider the plot in Figure 3.4.1, showing the spread of household incomes in a particular collection district. The horizontal axis shows each of the predictors within the household income category, and the vertical axis shows the proportion of the collection district that falls into each predictor.

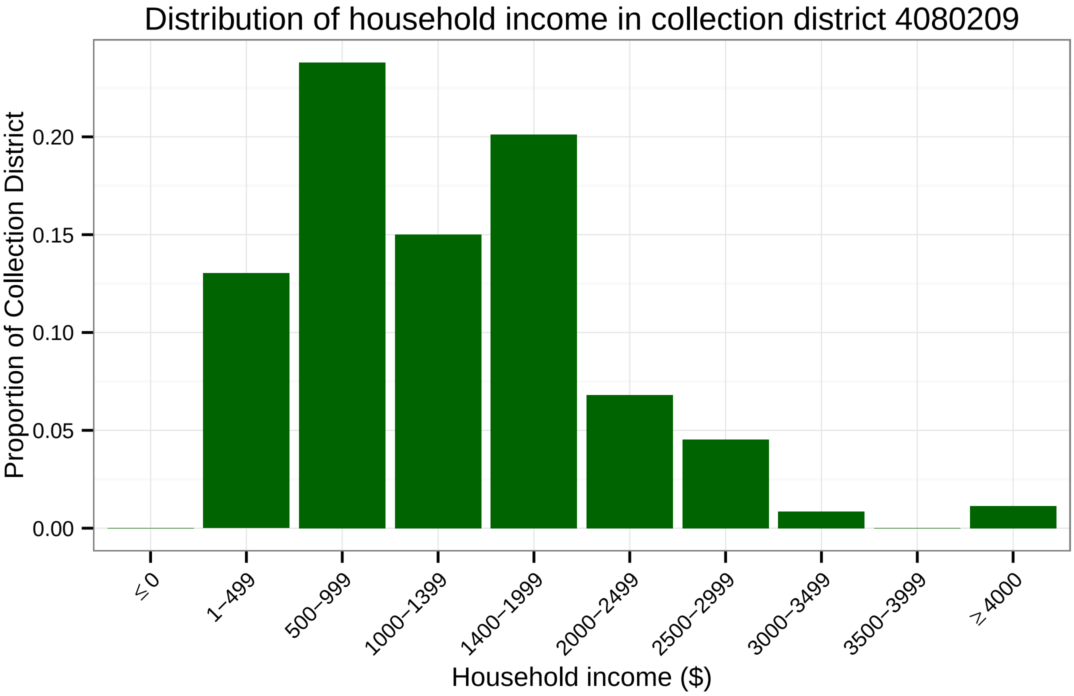


Figure 3.4.1: Distribution of household income in collection district 4071004. Labels for predictors are as defined in Table 3.3.1.

It is impractical for us to consider one of these plots for every collection district, so we wish to find a method of visualising all of the collection districts at once.

3.4.2 Aggregate Statistics

We calculate a number of aggregate statistics on each collection district to assess the data. These are described in Table 3.4.1. This provides a rudimentary way to compare the demographics of different regions.

These aggregate statistics should not be regarded as averages for each predictor category, but they give an idea of relative location of each of the collection districts within each predictor category.

[2] This is a measure of location, and *not* the arithmetic mean.

Table 3.4.1: Definitions of aggregate statistics calculated on Census data, to compare relative locations within predictor categories.

Predictor category	Aggregate statistic
Household Income	Weighted average of the lower bounds of the household incomes[2].
Language spoken at home	Proportion of people in CD that speak a language other than English at home.
School education	Proportion of people in CD that have completed Year 12 (or equivalent).
Non-school education	Weighted average of the completion of non school (tertiary) education in CD, where we code Certificate Level as 0, Bachelor Degree or Advanced Diploma as 1, and Postgraduate Degree or Graduate Diploma as 2.

The plots in Figures 3.4.2 to 3.4.5 show the distribution of the collection districts across one of the predictor categories. For a given category, the horizontal axis displays the predictors within the category. The proportion of each collection district in each predictor is indicated on the plot. The lines on the plot serve no purpose other than to connect data from the same collection districts for visualisation purposes. The lines are coloured by the relevant aggregate statistic in Table 3.4.1, as indicated in the associated legend.

Figure 3.4.2 shows the distribution of household income within collection districts. It can be seen that no collection district has more than a quarter of its households in the highest category, and the most dramatic changes in the makeup of collection districts are in the number of low income households they contain. The highest

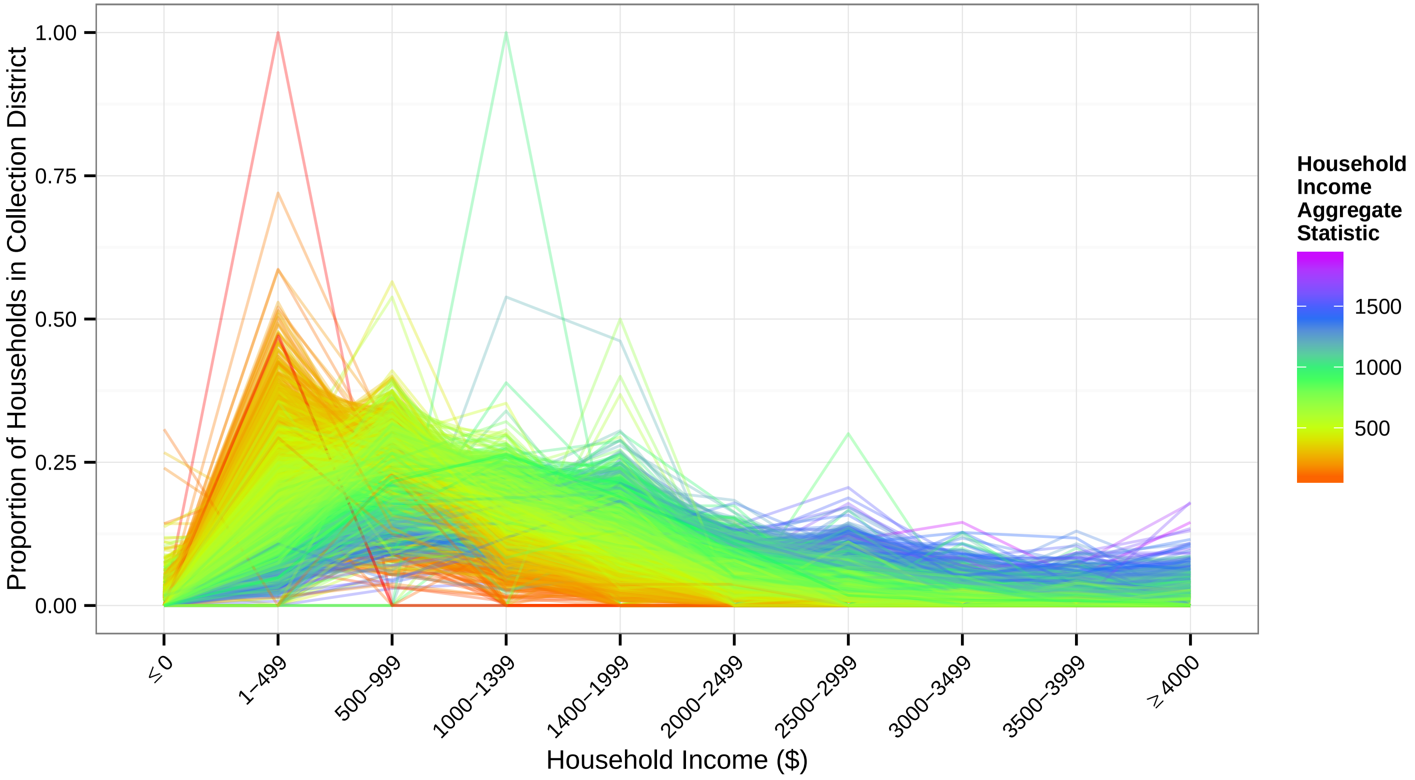


Figure 3.4.2: All collection districts coloured by the Household Income Aggregate Statistic. The sum of the individual proportions for each collection district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect collection districts and are only shown for visualisation purposes.

Household Income Aggregate Statistic of all the collection districts is \$1941.984, corresponding to an average annual household income of *at least* \$100,000 in that collection district.

It indicates that the bulk of the households in most collection districts have a weekly income of less than \$2000, regardless of the average income of the collection district.

The first abnormally high spike, coloured red, belongs to collection district 4022105. There are eight households recorded in that collection district in the Census, and they all reported an income of between \$1 and \$499 per week. The collection district is in the regional electoral district of Flinders and is located to the west of Ceduna.

The second high spike, coloured turquoise, represents collection district 4091507. There are only five households recorded in this collection district, and they all reported an income of between \$1000 and \$1399 per week. The collection district is in the metropolitan electoral district of Mitchell in the suburb of Oaklands Park. Most of the area in the collection district is occupied by the Westfield Marion shopping centre and the SA Aquatic and Leisure Centre, hence the small residential population.

Figure 3.4.3 shows the distribution of school education within collection districts. It shows that no collection district in South Australia have a Year 12 completion rate of higher than 80%, and most collection districts have a completion rate of less than 50%.

There are a small number of collection districts with abnormally low values for both predictors. The two lowest are collection districts 4120901 and 4011206.

Collection district 4120901 is located in the electoral district of Adelaide, and occupies the North-West corner of the CBD (North and West Terrace form part of the boundary for the district). Collection district 4011206 is in the electoral district of Stuart, in the vicinity of Port Augusta.

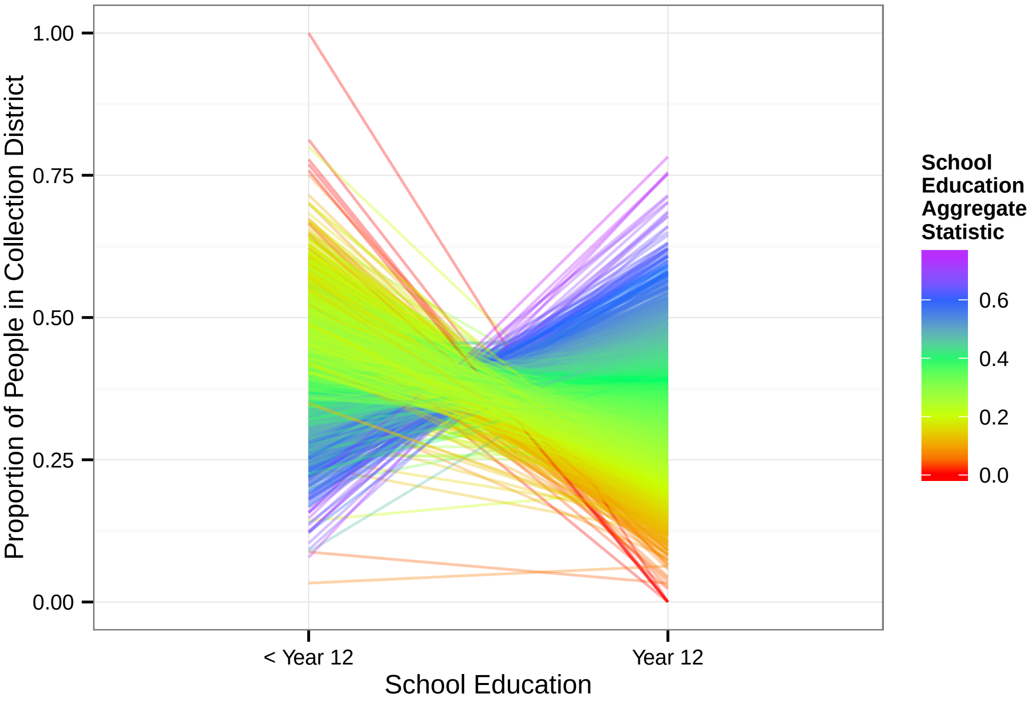


Figure 3.4.3: All collection cistricts coloured by the School Education Aggregate Statistic. The sum of the individual proportions for each collection district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect collection districts and are only shown for visualisation purposes.

Both are very small collection districts in terms of votes cast, with 13 and 31 votes in them respectively, but they still contain hundreds of residents. The districts had an abnormally high non-response rate for the question of school education. In fact, the two collection districts had extremely low response rates for all of the predictor categories we considered. It is unclear exactly why this is the case.

Figure 3.4.4 shows the distribution of higher education within collection districts. It shows that in every collection district, no more than 20% of people have completed a postgraduate qualification.

There is an abnormally high spike on the Bachelor Degree or Advanced Diploma predictor, with one collection district showing a proportion of 0.5. This is collection

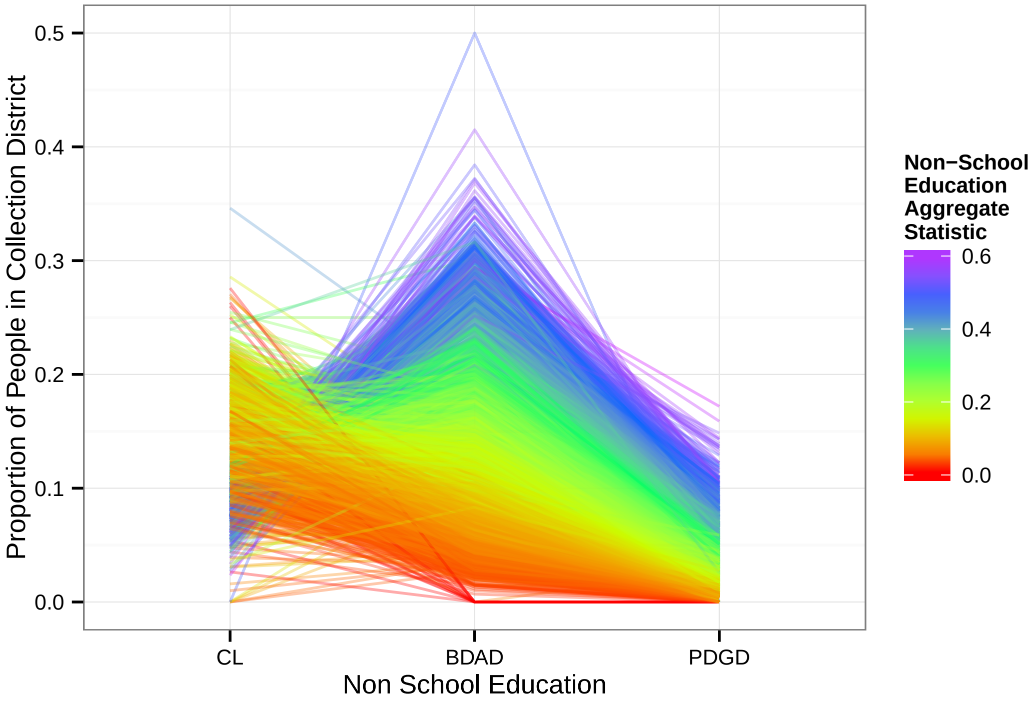


Figure 3.4.4: All collection districts coloured by the Non-School Education Aggregate Statistic. The sum of the individual proportions for each collection district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect collection districts and are only shown for visualisation purposes.

district 4091507 in the electoral district of Mitchell. This is the same collection district referred to earlier, which contains the Westfield Marion shopping centre.

There are also a number of collection districts which record zero people with a Certificate Level qualification. There are six of these in total, all of which also record zero people with a Postgraduate Level qualification, and most of which record very few people with a Bachelor Level qualification. None of these six collection districts recorded more than ten votes.

Finally, Figure 3.4.5 shows the distribution of language spoken at home within collection districts. It shows that around 20% of collection districts are populated mostly by people that speak languages other than English at home.

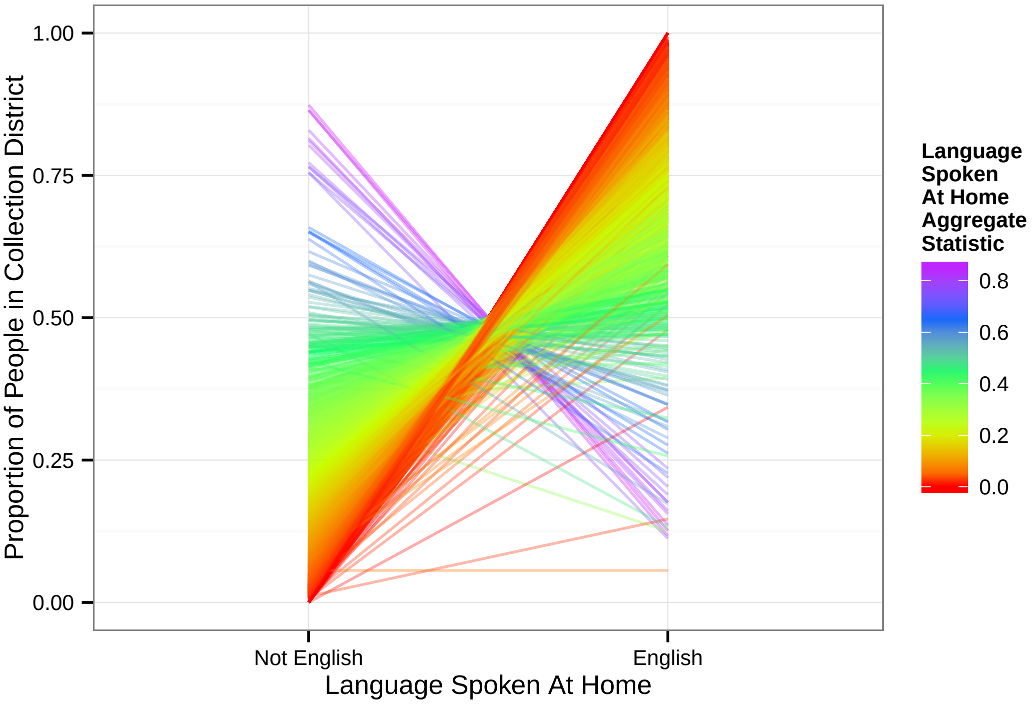


Figure 3.4.5: All collection districts coloured by the Language Spoken At Home Aggregate Statistic. The sum of the individual proportions for each collection district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect collection districts and are only shown for visualisation purposes.

There is one collection district with a very low response rate, shown as a near-horizontal red line at the bottom of the figure. This is collection district 4120901 located in the electoral district of Adelaide, highlighted earlier as a collection district with low response rate. Thirteen votes were cast in this electoral district.

3.5 Visualising Electoral Districts

We calculate the same aggregate statistics for electoral districts as in collection districts (shown in Table 3.4.1) and produce similar plots of the distribution of predictor categories within electoral districts.

Again, the horizontal axis displays the predictors within the category, while the vertical axis shows the proportion of each electoral district in each aggregate predictor. The lines connecting points on the plot are purely for visualisation purposes.

The results of this analysis are shown in Figures 3.5.1 to 3.5.3.

Broadly speaking, these figures show similar patterns as the corresponding figures showing collection districts, but with less variation between electoral districts.

There are also no clear outliers, unlike when we considered the collection districts. This is unsurprising since we are looking at aggregated collection districts now and the outliers have been ‘smoothed’ out.

Figure 3.5.1 shows that most of the variation in income between electoral districts is in the lower income households, with considerable variance in the categories \$1-499 and \$500-999 per week. No electoral districts have high proportions of households with very high incomes.

Within electoral districts with a low Household Income Aggregate Statistic (the ones coloured red in Figure 3.5.1) the distribution of household incomes looks roughly unimodal. As the Household Income Aggregate Statistic increases (toward the blue coloured electoral districts), three smaller peaks begin to emerge in the distributions.

Figure 3.5.2 shows that the proportion of people in an electoral district that have not completed Year 12 is always at least 25%, regardless of the electoral district being examined.

The aggregate statistic for non-school education is also shown in Figure 3.5.2, and it can be seen that no electoral district has a high number of people with Postgraduate Degrees or Graduate Diplomas. Again, most of the variation in higher education can be seen in the Bachelor Degree or Advanced Diploma, with proportions within electoral districts ranging between 0.049 and 0.28.

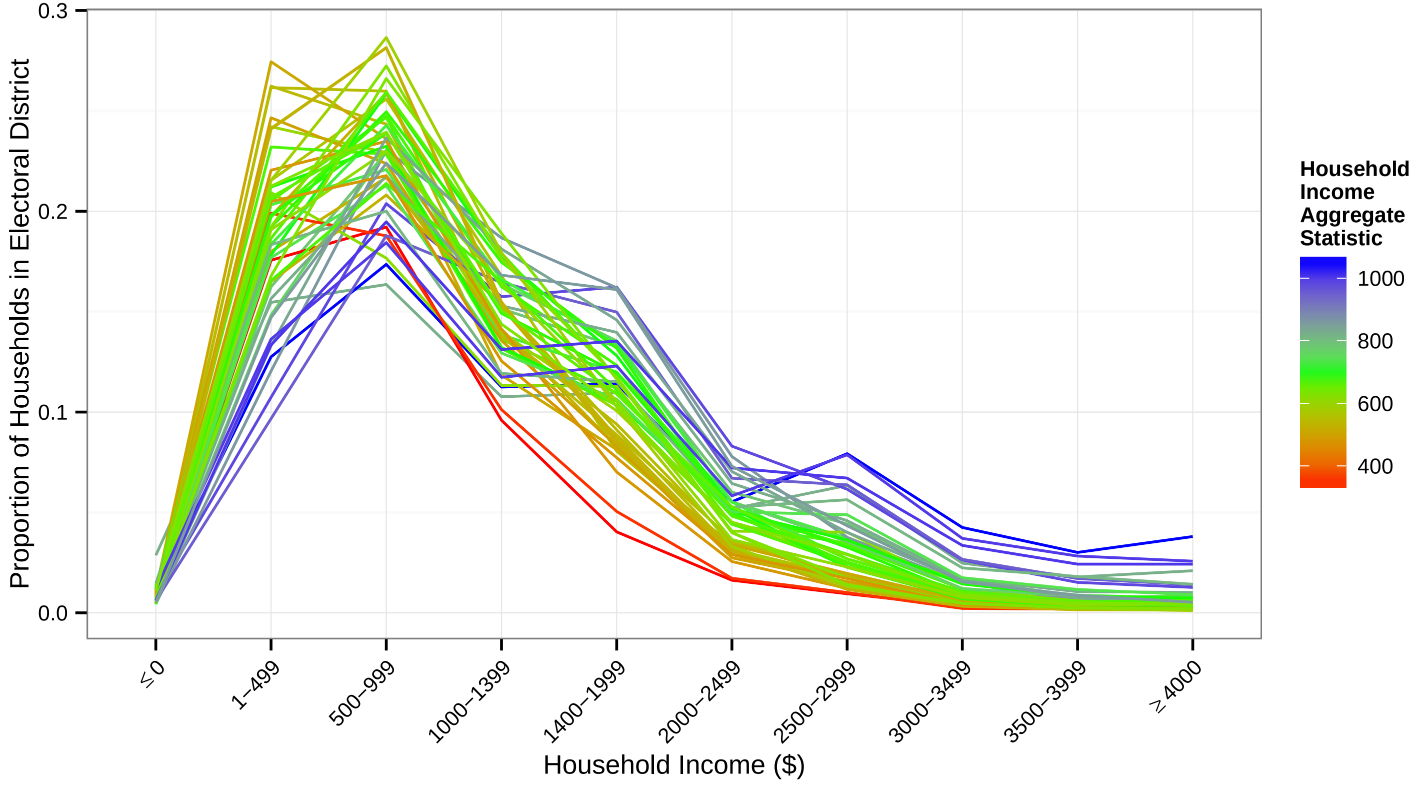


Figure 3.5.1: All electoral districts coloured by the Household Income Aggregate Statistic. The sum of the individual proportions for each electoral district is not one, due to the existence of the ‘Other’ predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect electoral districts and are only shown for visualisation purposes.

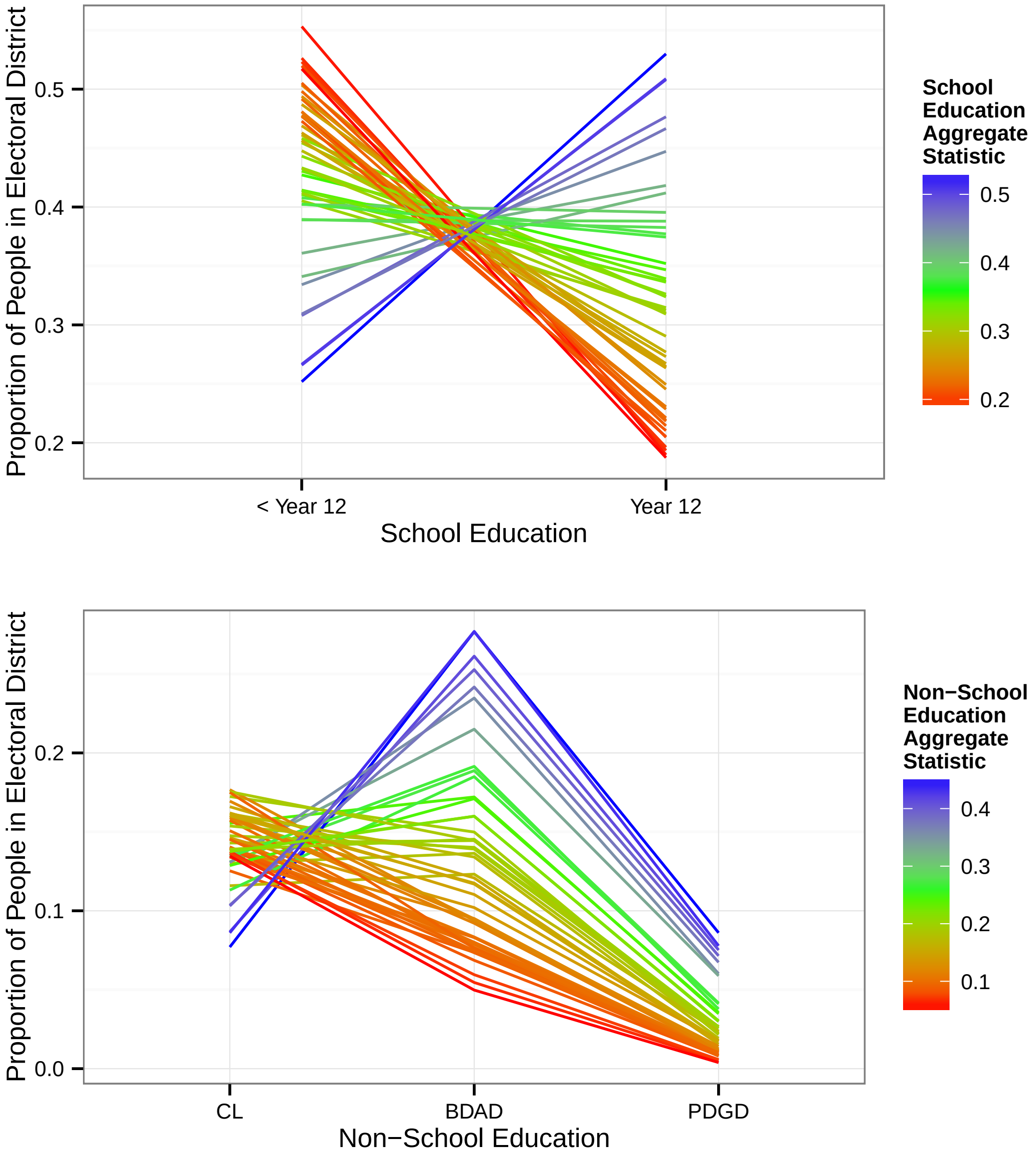


Figure 3.5.2: All electoral districts coloured by the School Education Aggregate Statistic (top) and the Non-School Education Aggregate Statistic (bottom). The sum of the individual proportions for each electoral district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect electoral districts and are only shown for visualisation purposes.

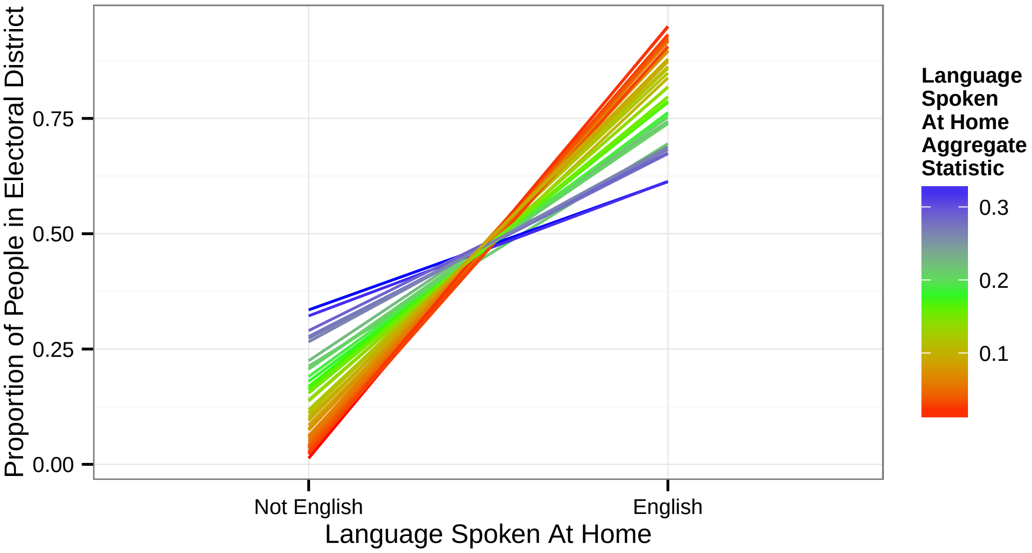


Figure 3.5.3: All electoral districts coloured by the Language Spoken At Home Aggregate Statistic. The sum of the individual proportions for each electoral district is not one, due to the existence of the 'Other' predictor for the predictor category. Labels for predictors are as defined in Table 3.3.1. Lines on the plot connect electoral districts and are only shown for visualisation purposes.

Figure 3.5.3 shows that all electoral districts have a high proportion of people that speak English at home. However, there are six electoral districts in which at least a quarter of people do not speak English at home.

The electoral district with the highest proportion of people that do not speak English at home is Croydon (proportion of 0.33). Croydon is an electoral district in the metropolitan area, in the Western suburbs of Adelaide. The electoral district with the highest proportion of people that speak English at home is Goyder (proportion of 0.95). Goyder is a rural electoral district on the Yorke Peninsula.

3.6 Extrapolation

Later in this thesis we will apply the results of a multinomial regression performed on 47 electoral districts to over 3000 collection districts. Before doing this we consider the distribution of the aggregate statistics for both collection district and electoral districts to see the degree of extrapolation this represents.

We have already noted that the variances between our predictors for collection districts is greater than those between our predictors for electoral district. Figure 3.6.1 illustrates this using the predictor 'Speaks English at Home'.

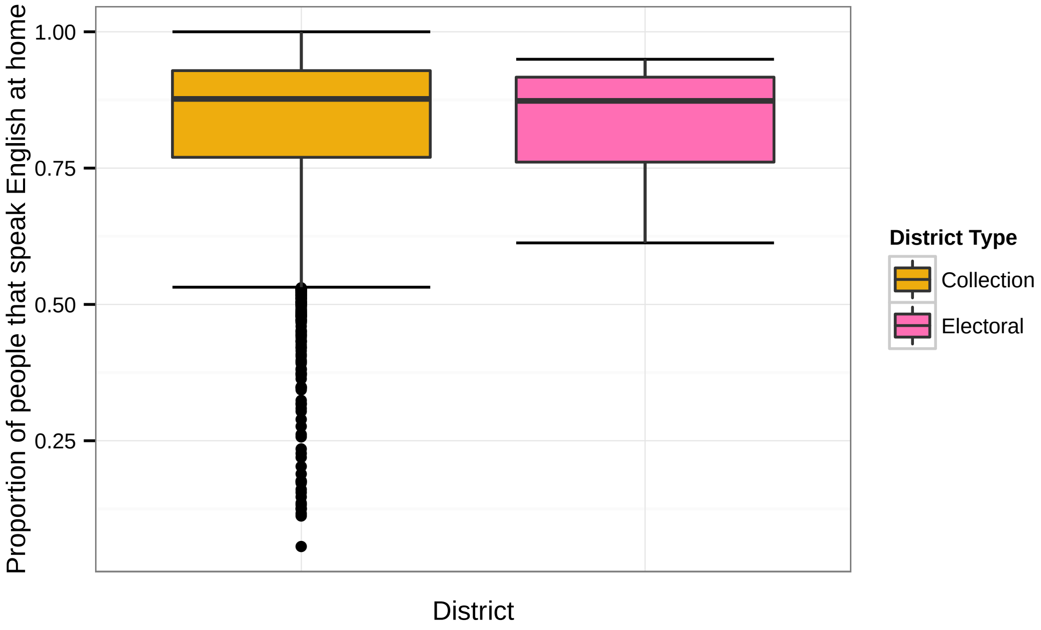


Figure 3.6.1: Boxplots of proportion of people that speak English at home for collection and electoral districts.

Although the medians and interquartile ranges for the collection and electoral districts appear comparable, the range of proportions is much larger for collection districts than it is for electoral districts. This suggests that we may have issues extrapolating results calculated for electoral districts to collection districts.

3.7 Correlation

As discussed in Section 3.3, we believe there could be a significant correlation structure in our predictors. To investigate this we calculate the covariances between each of the predictors for collection districts, and visualise them in Figure 3.7.1.

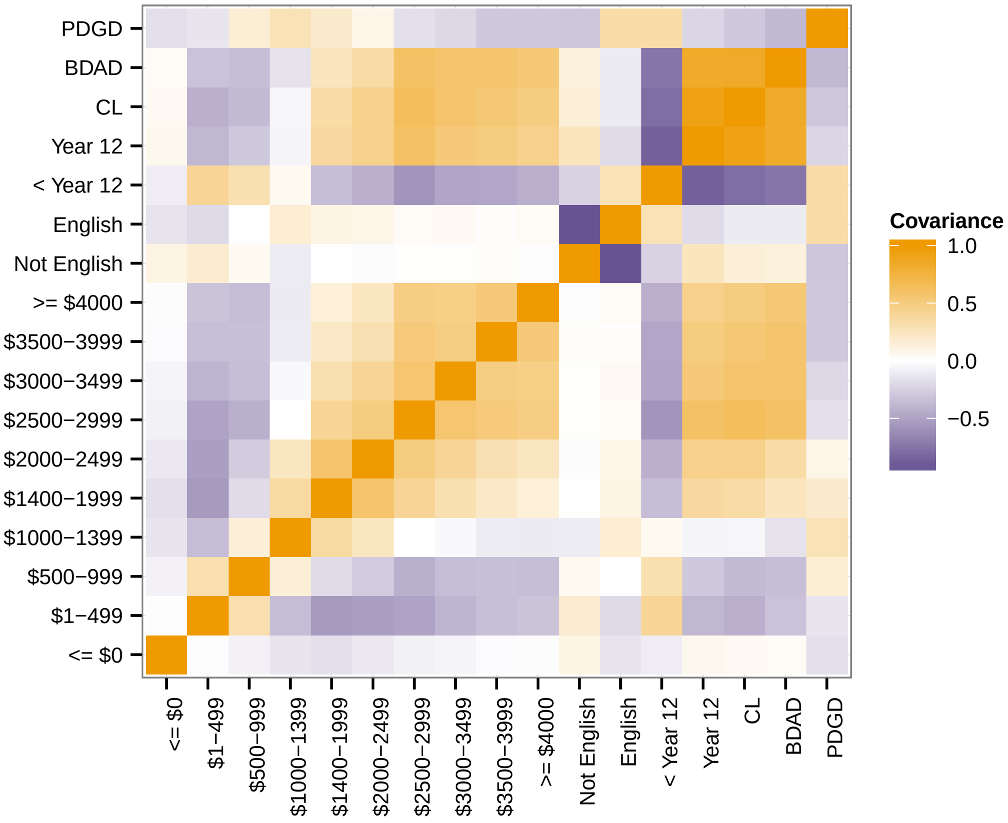


Figure 3.7.1: Heatmap containing the covariances between predictors.

As can be seen in the heatmap, there are a number of standout blocks along the leading diagonal. Unsurprisingly, there are extremely strong correlations between the two Language Spoken At Home predictors, and also between the two School Education predictors. There is also a weak positive correlation between the higher levels of household income.

Additionally, there appear to be strong correlations between the School Education and Non-School Education predictors. A person who has completed a Bachelor's Degree or a Certificate Level qualification is much more likely to have completed Year 12 (and much less likely to not have completed Year 12). Strangely, there appears to not be any strong relationship between the Postgraduate Qualification predictors and any of the other education predictors.

The higher levels of household income are weakly positively correlated with the number of people that have completed a Bachelor's Degree or Certificate Level qualification, and with the number of people that have completed Year 12. The same predictors are weakly negatively correlated with the number of people that have not completed Year 12.

Interestingly, there is almost no correlation between the language people speak at home and any other predictor.

While the household income predictors show weaker correlations than the other predictors, we can still see a positive correlation between the higher income predictors, and a weak negative correlation between higher income predictors and lower income predictors. Unsurprisingly, each of the income predictors is most strongly positively correlated with the income predictors nearest to them (for example, the predictor $1000-1399 is most strongly correlated with the predictor $1400-1999).

3.8 Principal Component Analysis

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of data and analysing the most important factors that determine the variability within a data set. By the definition of the algorithm, the technique also reduces correlation in the data.

The method works by finding an orthogonal set of linear combinations of our predictors. That is, we express the data using a new basis. The first basis vector chosen is in the direction of the most variance in the data set. We then choose a basis vector orthogonal to the first in the direction of maximum variance given that constraint, and so on choosing basis vectors orthogonal to all previous ones until we have a full basis.

A full description of the technique can be found in Jolliffe [23].

3.8.1 PCA applied

In our first model, we will fit a multinomial regression model to the 47 electoral districts and then using the results to predict values in each collection district for each party.

Given that each electoral district is orders of magnitude larger than the typical collection district, and that the demographics of each electoral district are aggregated from the demographics of the collection districts within it, we expect collection districts to be considerably more variable than electoral districts.

We can test this theory and gain some other useful insights by performing PCA on the data.

PCA on Electoral Districts (PCA Model 1)

We first perform PCA on the predictors for the 47 electoral districts and call this PCA Model 1. Table 3.8.1 shows the proportion of variance explained by the first six components of PCA Model 1. It indicates that over 92% of the variance in the data is explained by only the first two components.

Table 3.8.1: Variance in dataset explained by first six components of PCA Model 1, on electoral districts.

	Proportion of Variance	Cumulative Proportion
Component 1	0.628	0.628
Component 2	0.298	0.926
Component 3	0.0386	0.965
Component 4	0.0163	0.981
Component 5	0.00772	0.989
Component 6	0.00463	0.994
⋮	⋮	⋮

Since 92% of the variance in the data can be explained by the first two components, we are justified in visualising the PCA model using the first two scores only. We plot the first two scores for the 47 electoral districts in Figure 3.8.1. We also calculate the convex hull over the electoral districts and display this in the figure.

We see that many of the electoral districts are clumped together, indicating that they have a fairly similar makeup.

To gain an understanding of how much variance there is between the collection districts compared to the electoral districts, we also extrapolate and calculate the first two scores[3] for each collection district using PCA Model 1. We can then visualise them in the same way, also shown in Figure 3.8.1.

Figure 3.8.1 shows that there is considerably more variance between collection districts than between electoral districts. There are a large number of collection districts that look very different to any electoral district, which indicates that it may be problematic to extrapolate modelling results based on electoral districts to collection districts.

[3] Essentially taking the projection of each district onto the first two principal components.

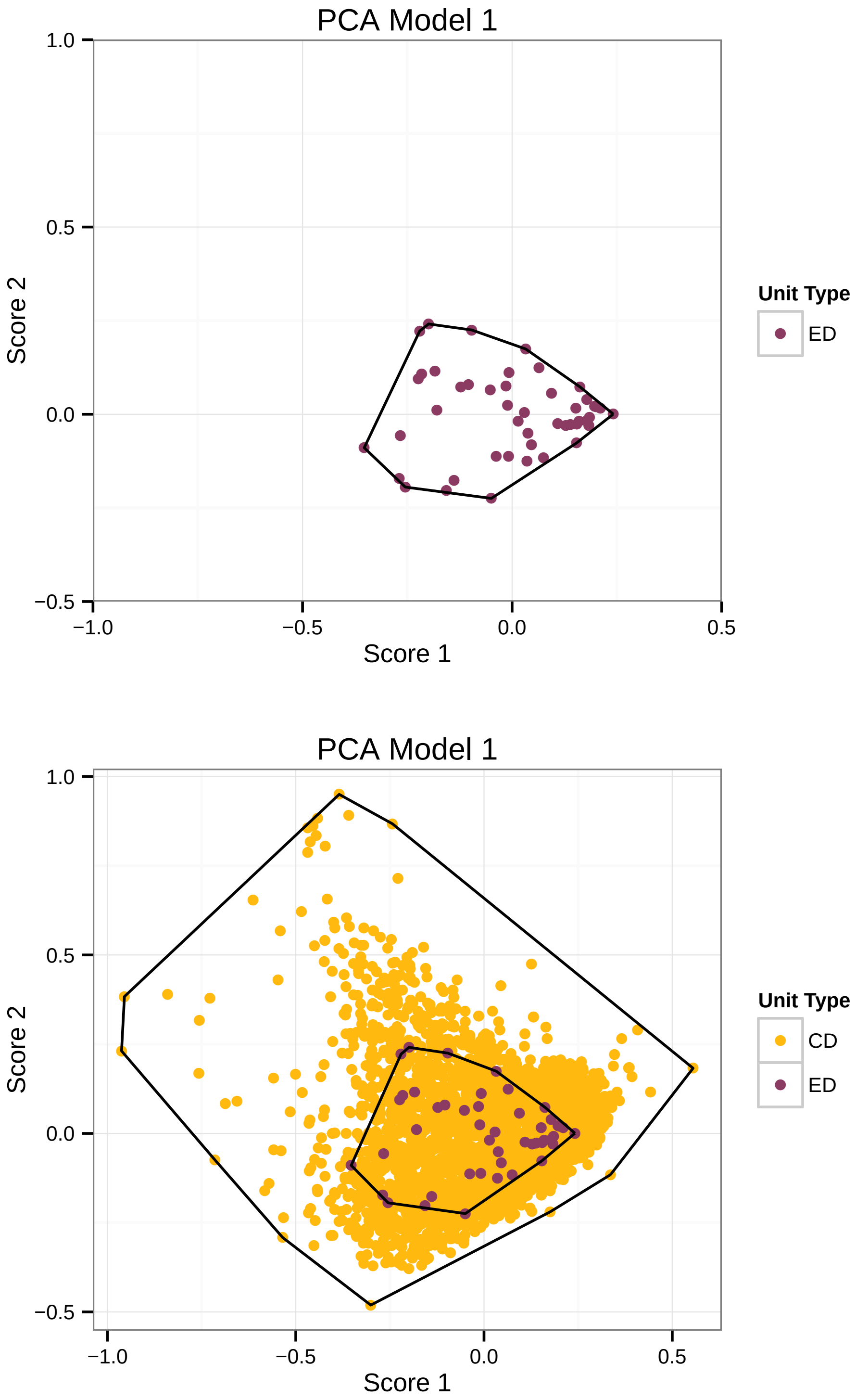


Figure 3.8.1: Plot of the first two scores for the electoral districts under PCA Model 1, with the convex hull over the electoral districts (top), and the first two scores for the electoral districts and the collection districts under PCA Model 1, with convex hulls over the sets of electoral districts and collection districts (bottom).

PCA on Collection Districts (PCA Model 2)

We also perform PCA on the set of predictors for collection districts, and call this PCA Model 2. Again, we consider the variance that is explained by each component, showed in Table 3.8.2. This table shows that there is considerably more variance between collection districts, as the first two principal components only explain approximately 76% of the variance. This suggests that the collection districts are less strongly correlated than the electoral districts.

Table 3.8.2: Variance in dataset explained by first six components of PCA Model 2, on collection districts.

	Proportion of Variance	Cumulative Proportion
Component 1	0.432	0.432
Component 2	0.324	0.757
Component 3	0.0818	0.839
Component 4	0.0458	0.884
Component 5	0.0298	0.914
Component 6	0.0219	0.936
⋮	⋮	⋮

Again, we plot the first two scores of all the districts (both collection and electoral), along with convex hulls for each type of district. For visualisation purposes, we choose to reverse the sign of the first score. This merely reorients the plot in a way that allows for simpler comparisons between PCA Models 1 and 2. The plot is shown in Figure 3.8.2.

The plots for Model 1 and Model 2 are very similar, indicating that the there are strong similarities in the structures between collection and electoral districts.

Most strikingly, the convex hulls for Model 1 and Model 2 look to have very similar

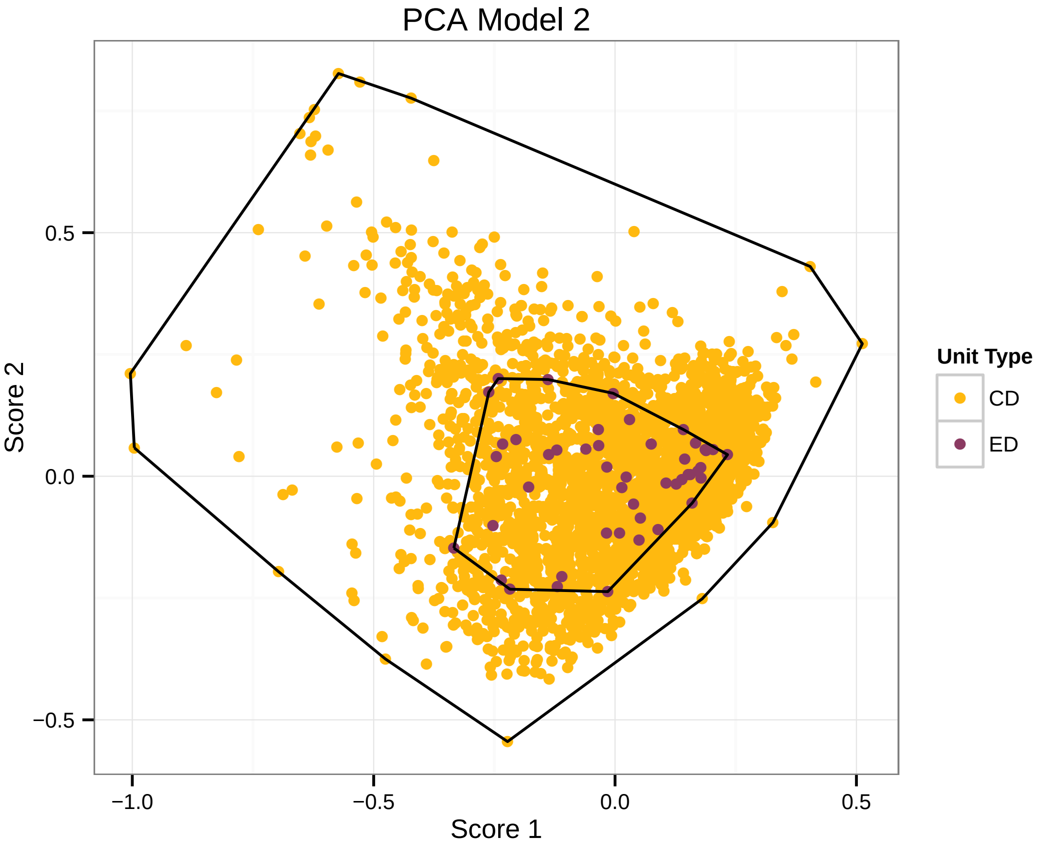


Figure 3.8.2: Plot of the first two scores for the electoral districts and the collection districts under PCA Model 2, with convex hulls over the sets of electoral districts and collection districts.

shapes and positions. To examine this further we overlay just the convex hulls for each model in one plot, shown in Figure 3.8.3.

This plot demonstrates that the structure of the collection district data and electoral district data is similar. The convex hulls have very similar shapes and locations.

Model Selection

Initially, we fit our model to the 47 electoral districts, so it makes sense that as we continue to explore the correlations and consider reducing the number of dimensions

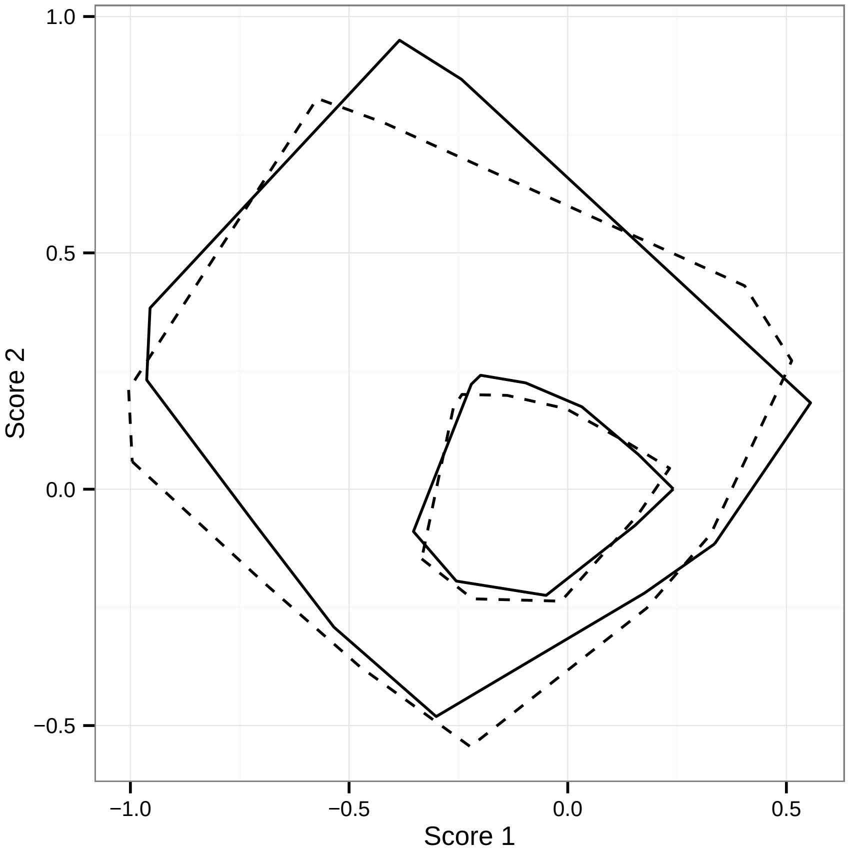


Figure 3.8.3: Convex hulls for PCA Model 1 and PCA Model 2 overlaid for visual comparison. The convex hulls for PCA Model 1 are shown with a solid line, and the convex hulls for PCA Model 2 are shown with a dashed line.

of predictors, we use PCA Model 1.

Once we are performing a fit based on the collection districts, we shall switch to using PCA Model 2 as a basis for working with and exploring the data.

3.8.2 Clusters

We are interested to know if the collection districts within each electoral district cluster together when viewed under PCA Model 1. We consider a small set of electoral districts and plot the scores of the collection districts within them to see if this is true.

We choose to use the districts of Adelaide, Chaffey, Cheltenham, Heysen, Newland, and Waite, and display these in Figure 3.8.4. The collection districts are coloured according to the electoral district in which they are located, and the electoral districts are also shown with shaded diamonds.

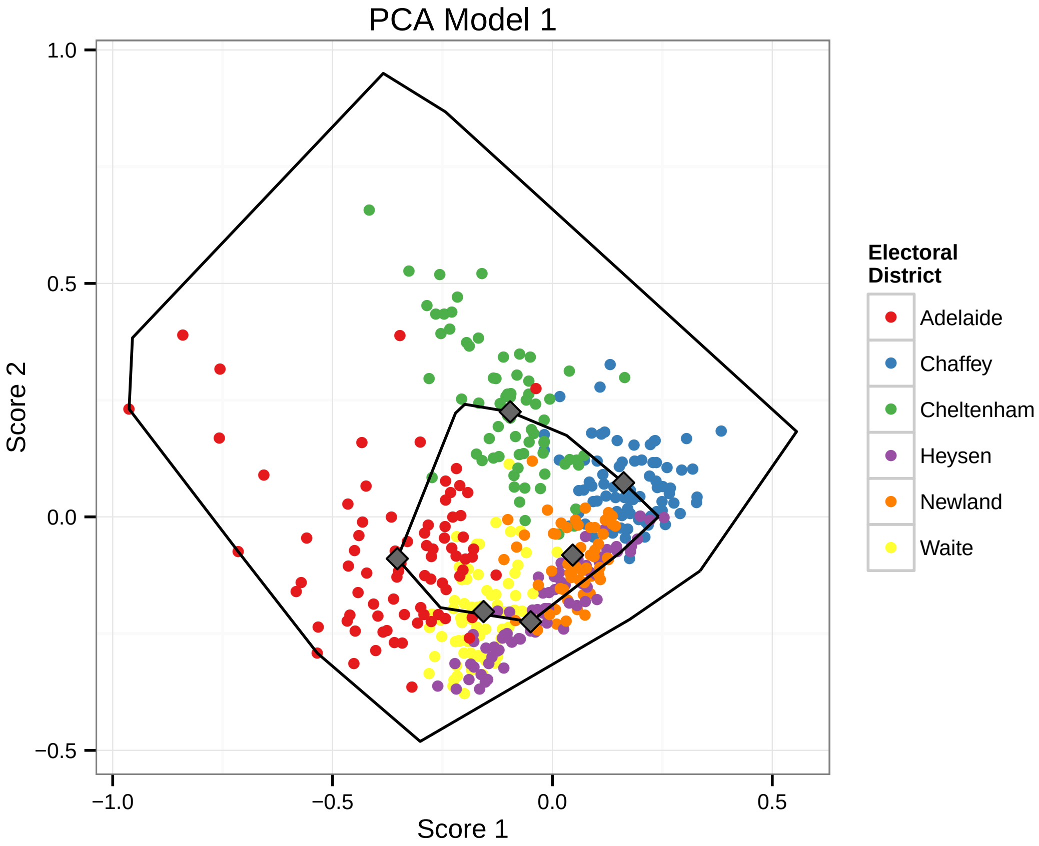


Figure 3.8.4: Subset of collection districts in South Australia, coloured by the electoral district in which they are located. The six shaded diamonds represent the position of the six electoral districts being considered.

As can be seen from the figure, all of these six electoral districts happen to sit on or near the convex hull. Each electoral district is roughly centred amongst the collection districts that sit inside it, and there is clear clustering in the collection districts. Many of the most extreme collection districts belong in only a small number of electoral districts.

A similar pattern can be seen with collection districts clustered around their electoral district using PCA Model 2 (figure not shown).

3.8.3 Structure within PCA

In order to further understand the structure and correlations between predictors, we now combine the PCA structure with the aggregate statistic information into one visualisation. We look for any patterns that give us a way of reducing the number of dimensions in the model fit.

Using PCA Model 1, we plot the first two scores of each collection district and colour them by the aggregate statistics. The results are shown in Figures 3.8.5 and 3.8.6.

In each of these figures we do not plot the electoral districts, but we do show the convex hull as a point of reference.

Figure 3.8.5 shows the Household Income Aggregate Statistic, and indicates a relatively weak pattern of increasing income from the top right corner of the plot toward the collection districts in the bottom left.

It also shows that most of the collection districts with the highest household incomes are clustered together, which, combined with the evidence in the previous sections, indicates that the majority of these collection districts are in a small number of electoral districts.

Figures 3.8.5 and 3.8.6 show that there is a much stronger pattern in the education predictor categories. The districts that are more highly educated are on the bottom left of the plots, and the broad direction in which the education level varies is the

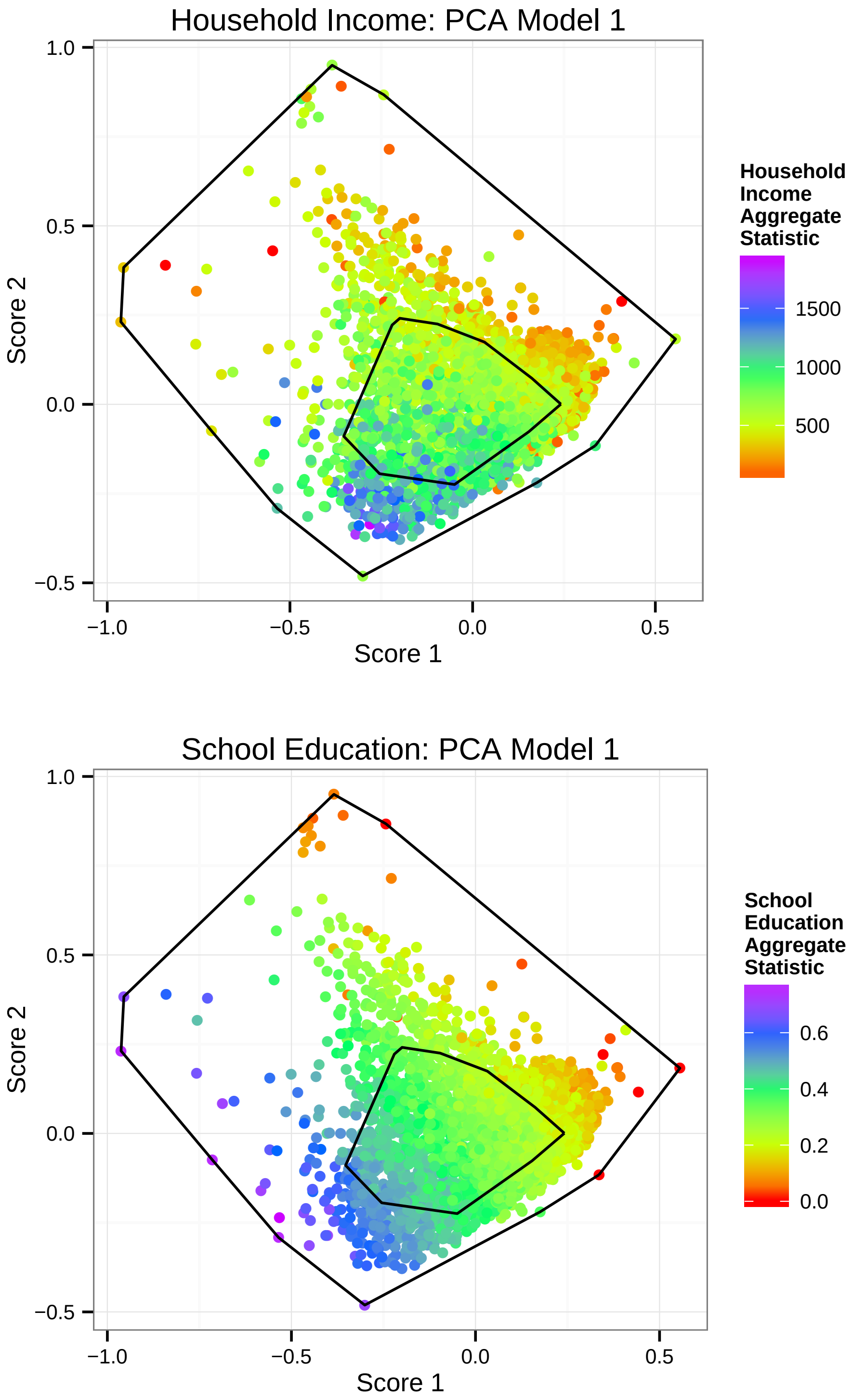


Figure 3.8.5: Plot of the first two components of each collection district under PCA Model 1. Each collection district is coloured according to its Household Income Aggregate Statistic (top) and its School Education Aggregate Statistic (bottom).

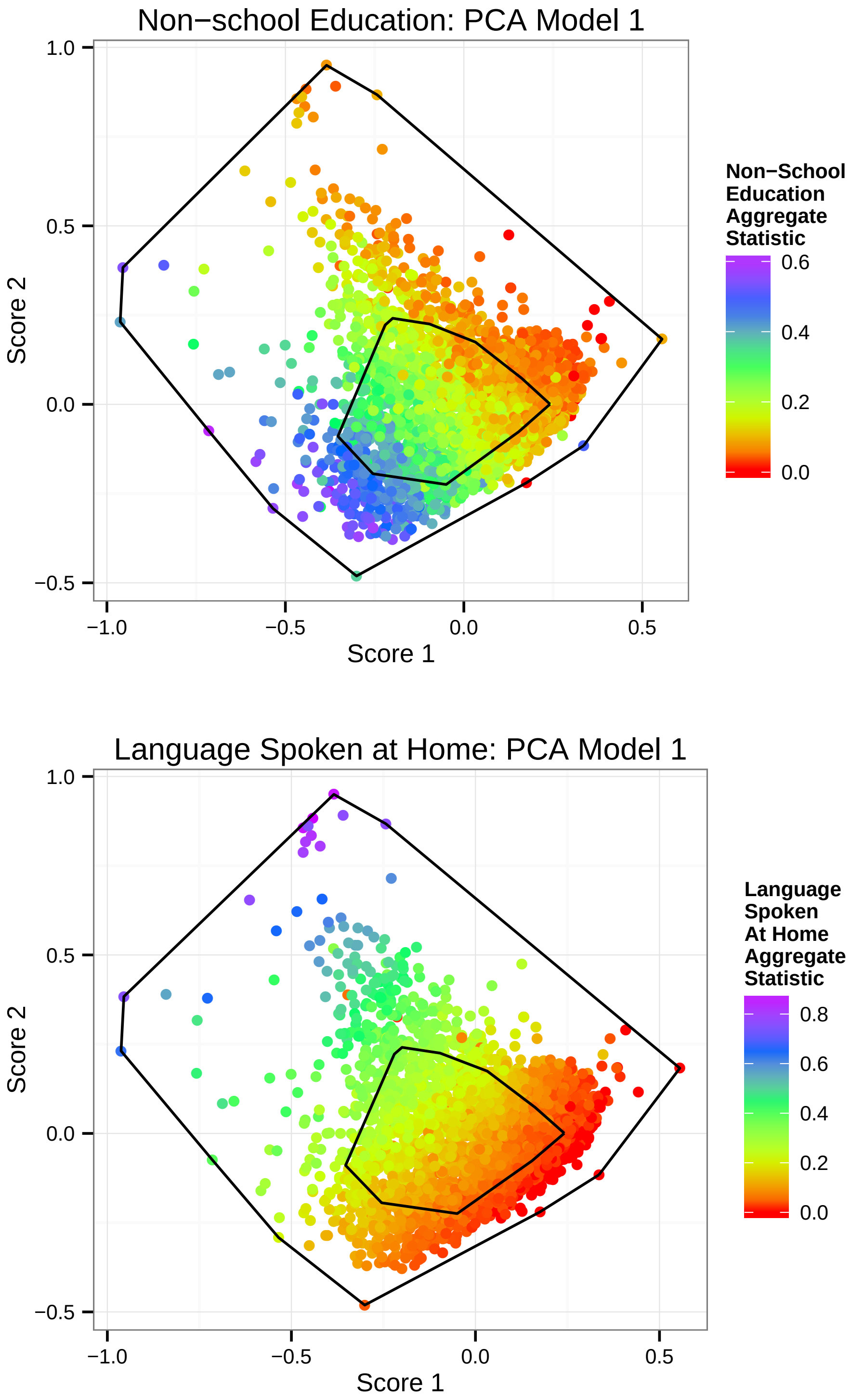


Figure 3.8.6: Plot of the first two components of each collection district under PCA Model 1. Each collection district is coloured according to its Non-School Education Aggregate Statistic (top) and its Language Spoken at Home Aggregate Statistic (bottom).

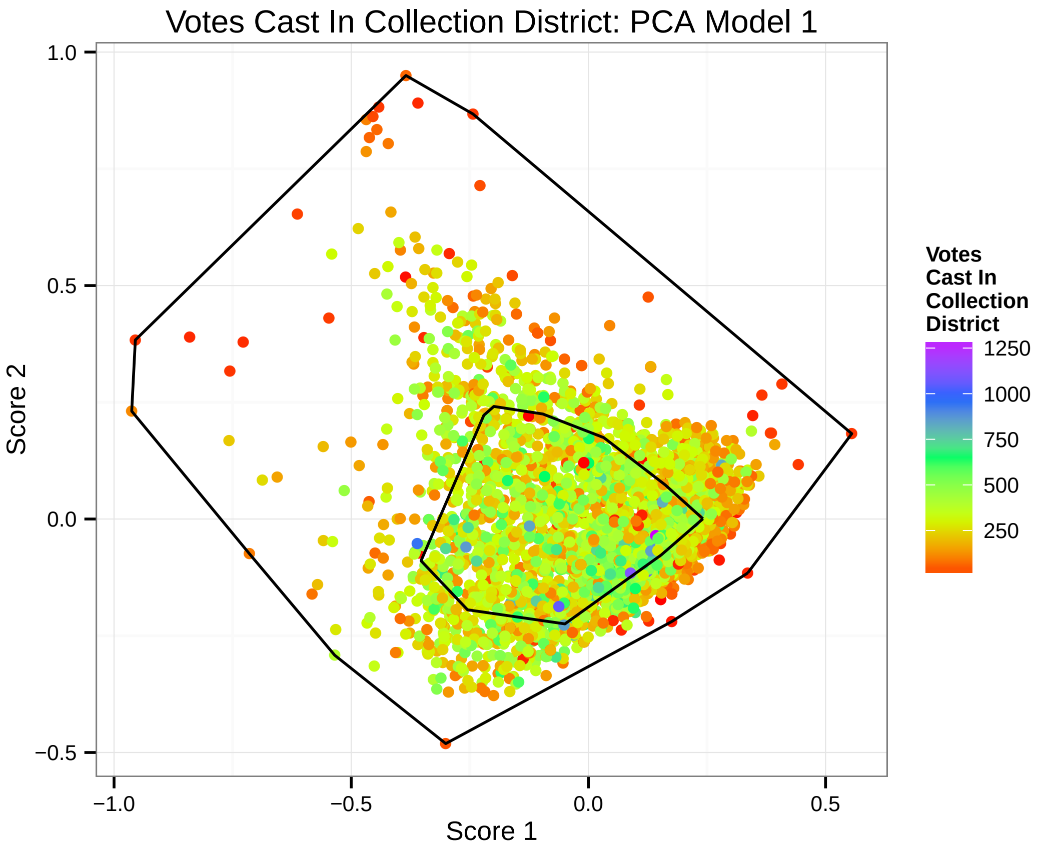


Figure 3.8.7: Plot of the first two components of each collection district under PCA Model 1. Each collection district is coloured according to the number of votes that were cast in it.

same across both school and non-school education.

This suggests that the two predictors categories for school and non-school education might be giving us similar information, and we may not need to use both categories in our analysis. This is supported by the heat map of covariances in Figure 3.7.1.

Figure 3.8.6 again shows a strong pattern for the language spoken at home. In this two-dimensional projection of the principal components, the broad direction of change appears to be approximately perpendicular to the direction of change in the education predictor categories.

Finally, we colour each collection district on the same plot by the number of votes that were cast in it in the 2010 state election[4] (Figure 3.8.7).

This figure shows that the most of the extreme collection districts are also very small, and thus in the context of electoral redistributions are relatively unimportant. These are also the ones most susceptible to be affected by the ABS introduced random error.

However there are still many very large collection districts that fall outside the convex hull. Hence on a model fit performed on electoral districts, we will always be extrapolating our results to sizeable collection districts.

3.9 Conclusion

Analysis of the 2006 Census data has shown that there are strong correlations between our predictors. When principal component analysis is performed, the first few principal components appear to adequately capture the structure and variance of the data.

These issues will be explored in the next chapter when we apply a multinomial regression model to infer voting behaviours from electoral and census data when looking at the electoral district level.

[4] The source of this data is the voter location information already mentioned, and this will be discussed in detail in Section 5.1.

Chapter 4

Modelling at Electoral District Level

If both the Census and election results data existed for collection districts, a relatively simple model could be used to predict the support for each party. One approach would be to find a linear solution using a form of logistic regression adapted for our multinomial variables. Techniques already exist for this, although their development is by no means complete [27].

Rather than having data all aggregated to the same level, we have electoral results data at polling place and electoral district levels, and Census data at the collection district and electoral district levels.

In order to apply the techniques of multinomial logistic regression, we must choose to work with the data at a level at which they all exist: the electoral district level.

Once multinomial logistic regression is performed on the aggregated data, we can use this model to predict the support for each party at collection district level. These predictions are likely to be relatively inaccurate, as to do this we extrapolate from electoral districts to the much smaller collection districts. However, we expect that they will have some predictive power, including at least the direction of influence of some of our predictors.

To perform this analysis we only use a small amount of the information contained in the dataset, as we perform regression on only 47 data points, each corresponding to an electoral district.

Performing this simpler analysis first with a smaller set of data is useful because it provides for the development of a toolbox to interrogate the results of our larger analysis in later chapters.

The analysis gives us some early results that, although expected to be naive, will hopefully give an indication as to the kinds of factors that can influence voting habits in collection districts.

Notation

Let CD be the set of collection districts, PP be the set of polling places, and ED be the set of electoral districts in the area of interest. For South Australia, $|CD| = 3247$[1] (but after processing and removing empty collection districts we work with 3212 unique collection districts), $|PP| = 746$[2], and $|ED| = 47$.

In each election, each elector is able to order the candidates in order of preference. Given the motivation for this thesis, and to keep the model manageable, we only need to know which of the two major parties each elector preferences. Thus, for our purposes, we group each elector according to whether they preference the ALP above the Liberal Party, the Liberal Party above the ALP, or lodge an informal ballot.

[1] At the time of the 2006 Census.

[2] At the 2010 state election.

Let P be the set of parties that an elector can preference. In this case, $P = \{ALP, LIB, INF\}$.

We let π_c^p be the proportion of people in collection district $c \in CD$ that vote for party $p \in P$, and it is this that we are trying to estimate.

4.1 Statistical Background

We first review the basic statistical methodologies of regression. The review in Sections 4.1.1 and 4.1.2 is generic and so the notation used is not consistent with the rest of this thesis.

4.1.1 Multiple Linear Regression

Multiple linear regression is a technique to study the relationship between a response variable Y and a set of explanatory variables $X_1, \ldots, X_n$. Notationally, we write x_{ij} to refer to the value of the j^{th} explanatory variable for the i^{th} subject.

The statistical model for multiple linear regression is then

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is a normally distributed term representing the individual deviations from the predicted mean.

For the i^{th} subject, the predicted response is then $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_{ij}$, and we refer to the difference between the observed and predicted response for the i^{th} subject as the residual, written $e_i = y_i - \hat{y}_i$.

We then choose the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ that minimise the sum of the squares of all the residuals, and take these as our parameter estimates.

An explanation of techniques for linear regression and calculating the parameter estimates can be found in Chapter 11 of Moore *et al.* [28].

Multiple linear regression is not suitable for this work because the response variable is count data, with three possible outcomes, rather than a continuous variable.

4.1.2 Multiple Logistic Regression

We now review an extension of multiple linear regression that allows us to consider cases where the response variable has only two possible outcomes. For example, these two outcomes could be that an elector votes for either the ALP or the Liberal Party. We construct a model for the estimation of probabilities (in this context, the probabilities that a randomly chosen person in each collection district will vote for each party).

Multiple linear regression as outlined in the previous section is not appropriate for this, as there is nothing to constrain our predicted responses to the interval $[0, 1]$. The extension of multiple linear regression known as logistic regression will restrict our predictions to this interval.

Generally speaking, we refer to one of the outcomes as 'success' and the other as 'failure' and represent them by 1 and 0 respectively. If we have n independent observations, then we have an underlying binomial distribution and the object is then to estimate the probability of success π using regression techniques.

In the same way that we modelled the mean of the response variable as a linear function of the explanatory variables in multiple linear regression, we now model the probability π in terms of the explanatory variables. However, we are not able to simply use the model

$$\pi = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij},$$

since extreme values of x_{ij} will give values of π that do not lie between 0 and 1.

Rather than working with π directly, we use odds. That is, the ratio of proportions for the two outcomes. If we write $\hat{\pi}$ for the proportion of observed successes, then the odds are

$$\text{odds } = \frac{\hat{\pi}}{1-\hat{\pi}}.$$

We then transform the odds and use the log odds (also known as logit), call this η, and model it as a linear function of the explanatory variables. Hence the model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij},$$

where π_i is a binomial proportion.

We use the deviance residuals to check the suitability of the model. These residuals should take approximately a standard normal distribution [14, p139].

The right hand side of the model remains a linear function and so the same techniques of multiple linear regression that were discussed in the previous section to choose estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ of the parameters apply. However, in order for sensible interpretations to be made from the fitted model, the log odds need to be transformed back and expressed in terms of π.

To implement this model we use the `glm` function in the `stats` package of R, part of R's base installation [37].

For a more detailed look at logistic regression, refer to Chapter 14 of Moore *et al.* [28] or Kleinbaum and Klein [24].

4.1.3 Multinomial Logistic Regression

In the previous section we considered response variables that can have one of only two outcomes. In this investigation, however, we have a response variable that can take any one of three outcomes (but always takes one of these outcomes) — a voter can vote for any of the three parties in the set $P = \{ALP, LIB, INF\}$. We now consider a further extension of the log-linear models that will allow us to model our situation.

The multinomial logistic regression model effectively works as a set of parallel logistic regression models. One category is chosen to be the baseline category and we fit the other categories by comparing it to that baseline. In our model we choose ALP as the baseline category, as it is first alphabetically.

So the model becomes

$$\eta_i^p = \log\left(\frac{\pi_i^p}{\pi_i^{ALP}}\right) = \beta_0^p + \sum_{j=1}^{n} \beta_j^p x_{ij}, \quad i \in CD, PP, ED \text{ and } p \in \{LIB, INF\}.$$

After fitting the model for η_i^{LIB} and η_i^{INF}, we have three equations in three unknowns for each i, specifically

$$\eta_i^{LIB} = \log\left(\frac{\pi_i^{LIB}}{\pi_i^{ALP}}\right), \eta_i^{INF} = \log\left(\frac{\pi_i^{INF}}{\pi_i^{ALP}}\right), \text{ and } \sum_{p \in P} \pi_i^p = 1,$$

which we solve simultaneously to find π_i^{ALP}, π_i^{LIB}, and π_i^{INF}.

More information about multiple logistic regression can be found in Fox [17].

The `glm` function cannot fit multinomial logistic models and so for implementation we use the `multinom` function in the `nnet` package in R [39]. This function uses neural networks, a form of supervised learning, to perform the model fit.

4.2 Method of Model Validation

It is difficult to interpret the results of the multinomial regression fit, because of the transformations that are applied to the data. Before attempting to interpret the coefficients we look to validate the model by investigating the predictions produced by the model. We want to ensure that the predictions are consistent with our preexisting knowledge of the relationship between voting behaviour and demographics, from both the literature and election results.

For each model we consider a number of statistics to validate the model. We first consider the distributions of the predicted values of π_c^{ALP}, π_c^{LIB} and π_c^{INF} for the collection districts c.

We expect the empirical distributions of the sets $\{\hat{\pi}_c^p : c \in CD\}$ for $p \in \{ALP, LIB\}$ to be unimodal with means approximately equal to the mean vote for each of these parties across the entire state.

We also expect the distribution of $\hat{\pi}_c^{INF}$ to be unimodal, with a mean approximately equal to the statewide informal vote. Given the low proportion of informal voters, we also expect the spread of this distribution to be small.

Additionally, we consider the predictions for collection districts in three different electoral districts: Cheltenham, Chaffey, and Norwood. These districts are chosen for case studies as they give a mix of demographics and a combination of metropolitan and rural seats, marginal and safe seats, and ALP and Liberal held seats. A discussion of the characteristics of these electoral districts will be provided shortly.

We calculate a single summary statistic for each collection district to reduce dimensionality and allow for simpler comparisons. We define

$$\pi_i^{DIFF} = \pi_i^{LIB} - \pi_i^{ALP}, \qquad i \in CD, PP, ED,$$

to be the difference between the probability that a person in region i votes for the Liberal Party and the probability they vote for the ALP.

Electoral districts are visualised using shapefiles that contain coordinate information for both electoral districts and collection districts. The electoral district shapefile was supplied by ECSA [31] and the collection district shapefile is available from the ABS [32].

Using spatial objects in R and the package `rgdal` [9] we combine this data with our model fit and produce maps of collection districts coloured by $\hat{\pi}_c^{DIFF}$. In fact, we are able to colour the maps based on any variable we have.

A positive value of $\hat{\pi}_c^{DIFF}$ means that collection district c has been predicted to have more Liberal Party voters than ALP voters, and a negative value corresponds to a prediction of more ALP voters than Liberal Party voters. We thus colour the collection districts on the map on a gradient from blue (for Liberal voting districts) to red (for ALP voting districts).

Cheltenham

Cheltenham is an electoral district in the western suburbs of Adelaide. The seat was won by the ALP in 2010 and is regarded as a safe seat for the Labor Party, with 66.1% of formal votes cast preferencing the ALP over the Liberal Party.

The member of parliament for the seat is Jay Weatherill, who in 2010 held the ministries of Environment and Conservation, Early Childhood Development, Aboriginal Affairs and Reconciliation, and assisted the Premier in Cabinet Business and Public Sector management [12]. Weatherill later became Premier of South Australia following the resignation of Mike Rann from the position.

One of the biggest local issues in Cheltenham at the time was the controversial St Clair oval land swap, in which the Government and local council allowed a housing development on the St Clair oval site. Opposition to this from the community was strong. Independent candidate Henrietta Child ran in Cheltenham on a platform that included saving the St Clair oval [30]. Child received 7.7% of the first preference votes in the 2010 election, more than the Greens.

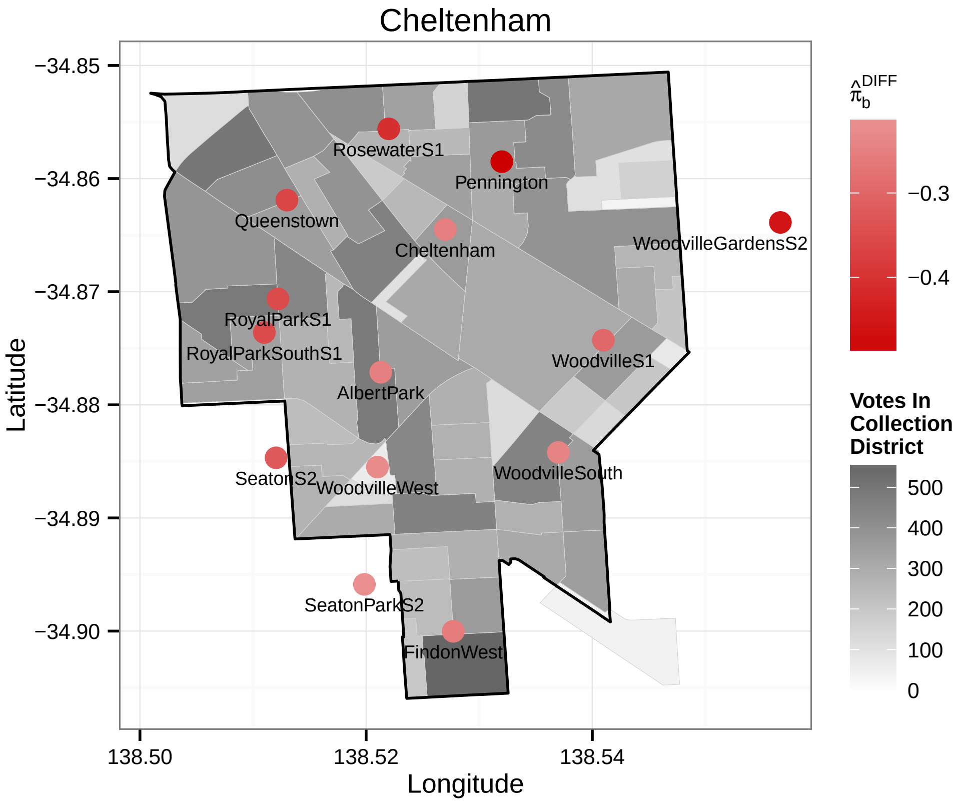


Figure 4.2.1: Cheltenham, with collection districts coloured according to the number of votes cast in it and polling places coloured by π_b^{DIFF} for each $b \in PP$.

Polling Place b	π_b^{DIFF}	Polling Place b	π_b^{DIFF}
AlbertPark	-0.25	RoyalParkSouthS1	-0.35
Cheltenham	-0.25	SeatonParkS2	-0.22
FindonWest	-0.26	SeatonS2	-0.32
Pennington	-0.50	WoodvilleGardensS2	-0.46
Queenstown	-0.36	WoodvilleS1	-0.30
RosewaterS1	-0.41	WoodvilleSouth	-0.24
RoyalParkS1	-0.35	WoodvilleWest	-0.23

Table 4.2.1: Values of π_b^{DIFF} for each polling place b in Cheltenham.

Using the aggregate statistics previously calculated, and described in detail in Table 3.4.1, Cheltenham is not a wealthy electoral district, ranking 37^{th} out of the 47 electoral districts on the household income aggregate statistic. The district sits nearer the middle on education, ranking 27^{th} for school education and 31^{st} for non-school education. There are many non-English speakers in Cheltenham, and it ranks 4^{th} on this aggregate statistic.

Figure 4.2.1 shows a map of Cheltenham, with each of the polling places marked and coloured by the value of π_b^{DIFF} for that polling place b. Collection districts on the map are shaded according to the number of votes that were cast by residents of that district.[3] Table 4.2.1 gives the value of π_b^{DIFF} for each of the polling places.

While there is no requirement that a person vote in any particular polling place, the data suggests that it is most likely that they will vote in the one nearest to them. In validation we can use the polling places as a proxy to give us an impression of the geographic support for each party in Cheltenham.

The three polling places shown outside of the boundary of the district are shared polling places with neighbouring electoral districts. Additionally, the polling places of Royal Park, Royal Park South, Woodville, and Rosewater are also all shared polling places that happen to lie within the boundaries of Cheltenham.

The figure shows strong ALP support across the whole of Cheltenham, with this being strongest in the northeast of the electoral district.

Norwood

Norwood is in the inner eastern suburbs of Adelaide, bordering the Adelaide parklands and CBD on its western edge. The River Torrens passes through Norwood

[3] The source of this data is the voter location information already mentioned, and this will be discussed in detail in Section 5.1.

and forms part of its northern border. The electoral district is regarded as marginal, and changed hands from the ALP to the Liberal Party at the 2010 election with the Liberal Party receiving 54.9% of the two-party preferred vote (excluding informal votes). The winner of the seat in 2010 was Steven Marshall, who would go on to become the South Australian Leader of the Opposition in 2013.

The seat of Norwood was regarded as a key seat in the 2010 election, and it was contested by eight candidates [13]. Norwood was renamed Dunstan for the 2014 election.

Norwood is a highly educated electoral district, ranking 4^{th} for school education and 5^{th} for non-school education. It is also relatively high in both income and numbers of people that speak languages other than English, ranking 11^{th} and 10^{th} on these aggregate statistics respectively.

Figure 4.2.2 shows a map of Norwood, with each polling place marked and coloured by π_b^{DIFF} for that polling place and each collection district coloured by the number of votes cast by people living in it. Table 4.2.2 shows the precise values of π_b^{DIFF}.

Support for the Liberal Party is strongest in the western side of the district. As you move across the district to the east, support for the Liberal Party weakens and there are some polling places in which the ALP received more votes than the Liberal Party.

Chaffey

The electoral district of Chaffey is a regional seat in the Riverland district of South Australia, and includes the towns of Renmark, Berri, Barmera, Loxton, and Waikerie.

The seat has amongst the lowest education levels, ranking 45^{th} out of 47 on the

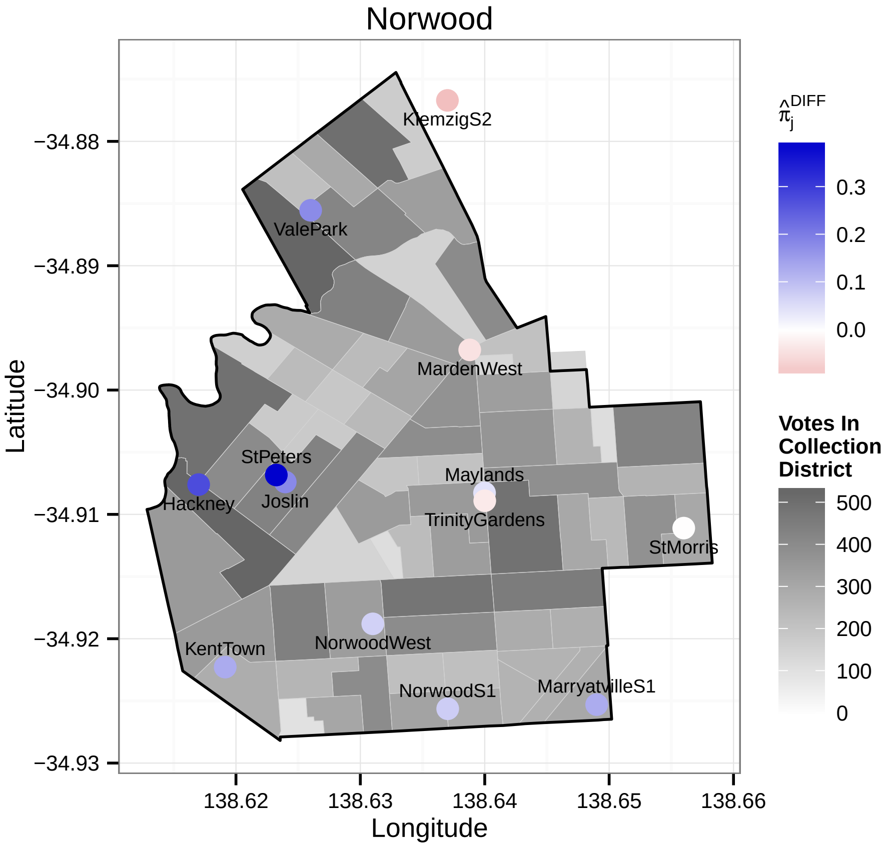


Figure 4.2.2: Norwood, with collection districts coloured according to the number of votes cast in it and polling places coloured by π_b^{DIFF} for each $b \in PP$.

Polling Place b	π_b^{DIFF}	Polling Place b	π_b^{DIFF}
Hackney	0.27	NorwoodS1	0.08
Joslin	0.18	NorwoodWest	0.07
KentTown	0.13	StMorris	-0.00
KlemzigS2	-0.10	StPeters	0.39
MardenWest	-0.04	TrinityGardens	-0.03
MarryatvilleS1	0.13	ValePark	0.18
Maylands	0.04		

Table 4.2.2: Values of π_b^{DIFF} for each polling place b in Norwood.

school education aggregate statistic, and 43^{rd} for non-school education. Chaffey has relatively low income, ranking 34^{th} on household income, and sits in the middle for language, ranking 26^{th} on the language spoken at home aggregate statistic.

Until the 2010 election the seat was held by Karlene Maywald of the Nationals. Maywald served as a minister in the Labor government from 2004 until her exit from the parliament, with portfolios at various points including the River Murray, Regional Development, Small Business, Consumer Affairs, Science and the Information Economy. At the 2010 election the Liberal Party won the seat from the Nationals [11].

Chaffey recorded a two-party preferred vote of 77.8% to the Liberal Party against 22.2% to the ALP in 2010. The ALP received only 7.2% of the primary vote.

This seat is interesting in part because of the strength of support for parties other than the ALP or Liberal Party. In the 2010 election, the ALP came fourth in order of first preference votes, and was beaten by both the Nationals and Family First. The final two *candidate* preferred count after preferences were distributed was between the Liberal Party and the Nationals.

On how-to-vote cards distributed to voters, the ALP and the Greens both directed preferences to the Nationals, and Family First and the Nationals directed preferences to the Liberals.

The strength of support for the Nationals in Chaffey in combination with this formulation of preference tickets means that the true two party preferences between the ALP and the Liberal Party may have been masked in the results.

It is reasonable to suspect that some voters whose most preferred party was the ALP, followed by the Nationals, followed by the Liberal Party (perhaps with some other candidates mixed in) would choose to vote tactically and actually preference the Nationals first, knowing that the ALP stood no chance of winning the election.

If they were to do this and follow a Nationals preference ticket, also not an unreasonable assumption, they would end up ranking the ALP below the Liberal Party, despite the fact that the ALP was their favourite party.

This potential distortion in the two-party preferred figures from the true preferences of the electors may mean that Chaffey is difficult to predict using our models.

Figure 4.2.3 shows a map of Chaffey with polling places marked and coloured by π_b^{DIFF} for that polling place and collection districts coloured by the number of votes that were cast by voters that live in it, and Table 4.2.3 shows the precise values of the π_b^{DIFF}.

Because Chaffey is a regional seat and it is more difficult for many voters to get to a polling place on election day, there are a number of prepoll polling places set up around the district for people to vote in at specific times before polling day. Most of the votes cast in this way are treated as declaration votes, except for the 'Riverland Mobile' and 'Gerard Mobile' polling places, which contain ordinary votes. These two polling places are not included on the map as no fixed location is given for them in the ECSA dataset. In total 123 votes were cast in them.

Note also that there is one very large collection district across the north of Chaffey. The large area of this collection district can give a false impression about its importance. As can be seen from the shading in the map, that collection district contains very few votes.

The figure shows very strong Liberal party support across all of Chaffey. The strength of support for the Liberal party is comparatively weakest in the more populated areas of the electoral district.

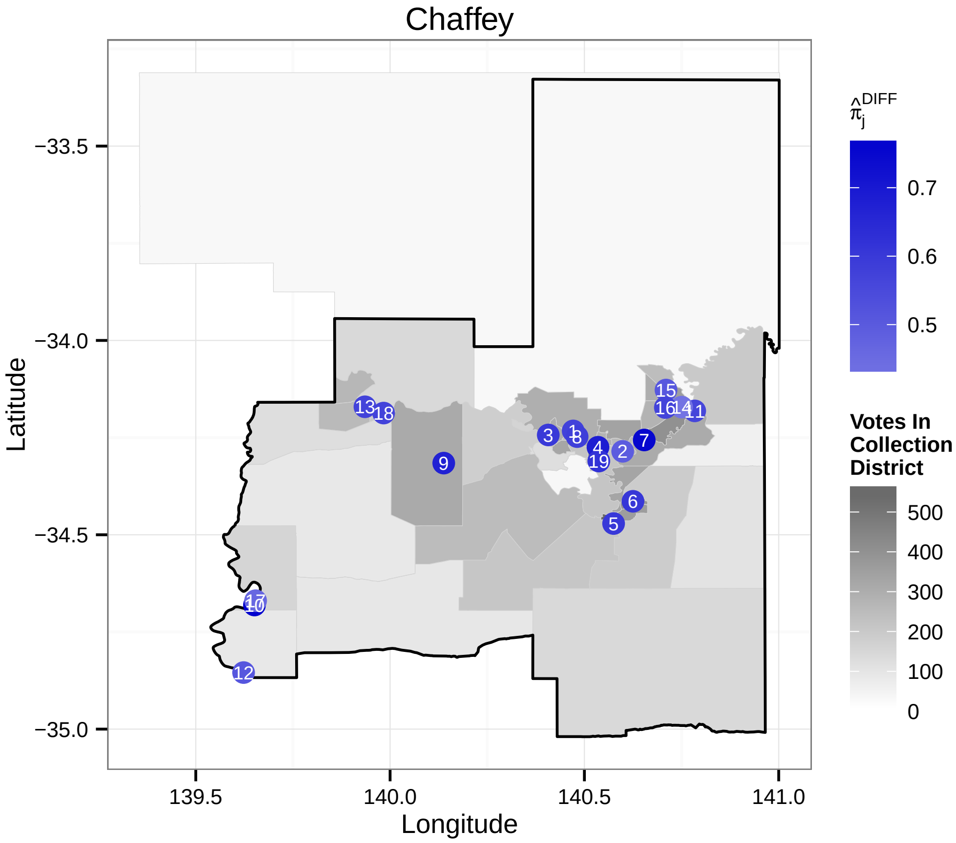


Figure 4.2.3: Chaffey, with collection districts coloured according to the number of votes cast in it and polling places coloured by π_b^{DIFF} for each $b \in PP$.

	Polling Place b	π_b^{DIFF}		Polling Place b	π_b^{DIFF}		Polling Place b	π_b^{DIFF}
1	Barmera	0.58	8	Monash	0.59	15	RenmarkNorth	0.51
2	Berri	0.49	9	Moorook	0.67	16	RenmarkWest	0.57
3	Cobdogla	0.60	10	Nildottie	0.78	17	SwanReach	0.46
4	Glossop	0.69	11	Paringa	0.56	18	Waikerie	0.57
5	Loxton	0.60	12	Purnong	0.51	19	Winkie	0.62
6	LoxtonNorth	0.61	13	Ramco	0.57			
7	Lyrup	0.76	14	Renmark	0.42			

Table 4.2.3: Values of π_b^{DIFF} for each polling place b in Chaffey.

4.3 Model 4.1 - with original 17 predictors

We first fit the multinomial logistic model using the full set of 17 predictors, listed in Table 3.3.1 in Section 3.3, and call this Model 4.1 (as it is the first model of Chapter 4). The model has the form

$$\log\left(\frac{\boldsymbol{\pi}^p_{ED}}{\boldsymbol{\pi}^{ALP}_{ED}}\right) = \boldsymbol{\eta}^p_{ED} = \beta_0 + X^{ED}\boldsymbol{\beta}, \qquad p \in \{LIB, INF\},$$

where X^{ED} is the design matrix containing the values of the 17 predictors for each of the electoral districts, $\boldsymbol{\pi}^p_{ED} = (\pi^p_e, e \in ED)$, and $\boldsymbol{\eta}^p_{ED} = (\eta^p_e, e \in ED)$.

We will perform inference for the collection districts by applying the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ under the model to the collection district data and calculating

$$\hat{\boldsymbol{\eta}}^p_{CD} = \beta_0 + X^{CD}\hat{\boldsymbol{\beta}}, \qquad p \in \{LIB, INF\},$$

where X^{CD} is the design matrix of predictors for each of the collection districts.

The coefficients and standard errors of the Model 4.1 fit can be found in Section A.2.1 in Appendix A.

We first validate the model fit by looking at the resultant predicted $\hat{\pi}^p_c$ for $c \in CD$ in Figure 4.3.1, and checking them against our existing understanding of the political situation.

We can see from these figures that the distribution of predicted values for the ALP and Liberal Party roughly approaches a uniform distribution over the interval $[0, 1]$, with some tapering off at the extremes.

According to this fit, roughly 10% of collection districts have at most 10% of voters supporting the ALP. Given that most elections are closely fought, it seems unbelievable that such a large proportion of collection districts would have such a lopsided result.

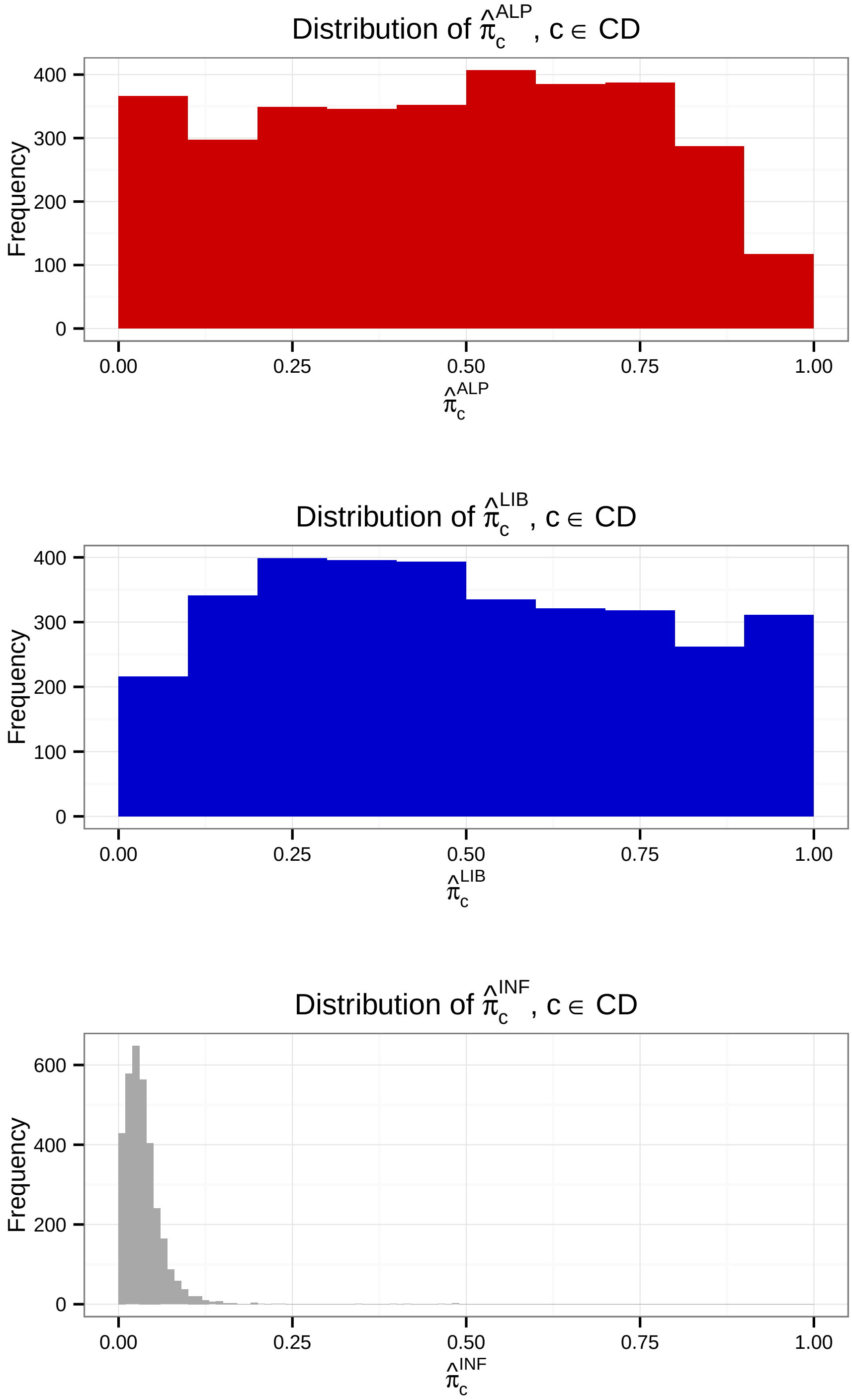


Figure 4.3.1: Distribution of predicted values for $\hat{\pi}_c^p$, for $p \in P$ and $c \in CD$, under Model 4.1

The bulk of the predicted values for $\hat{\pi}_c^{INF}$ are below 0.05, which is as we would expect given the low amount of informal voting in South Australian elections. However there are some predictions for $\hat{\pi}_c^{INF}$ as high as 0.5, which is unbelievable.

Figure 4.3.2 shows the electoral district of Cheltenham, with collection districts coloured by $\hat{\pi}_c^{DIFF}$. The map is mostly coloured in shades of red, as we expect. However the collection districts are not coloured consistently with the polling places in Figure 4.2.1. While we don't expect the patterns of polling places to translate directly to collection districts when we use the model for prediction, we do expect there to be fewer strongly blue collection districts.

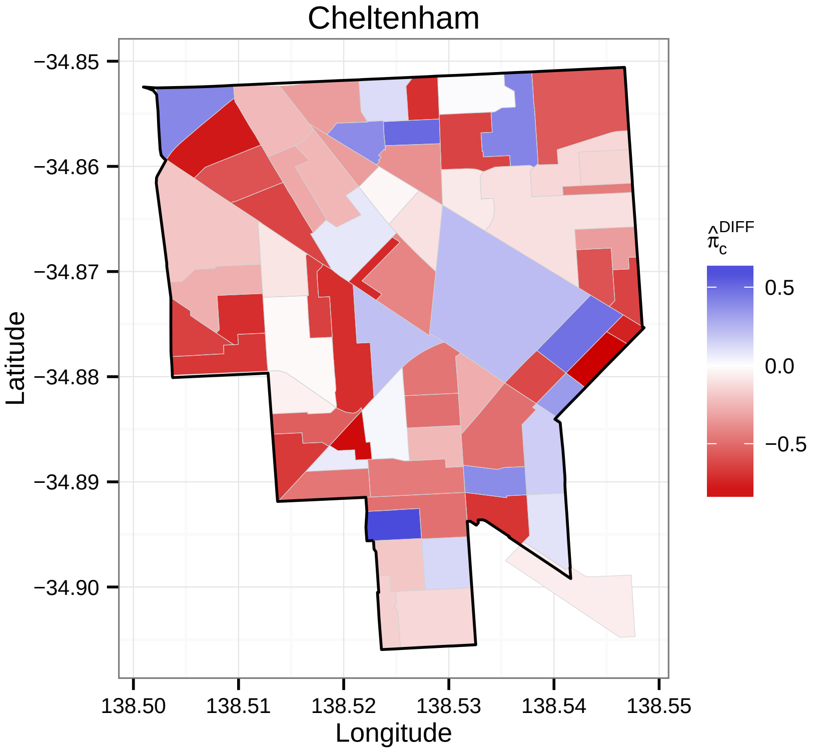


Figure 4.3.2: The electoral district of Cheltenham, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.1.

Figure 4.3.3 shows the regional electoral district of Chaffey. As expected, Chaffey is coloured very strongly blue in the plot. The pattern is consistent with the polling

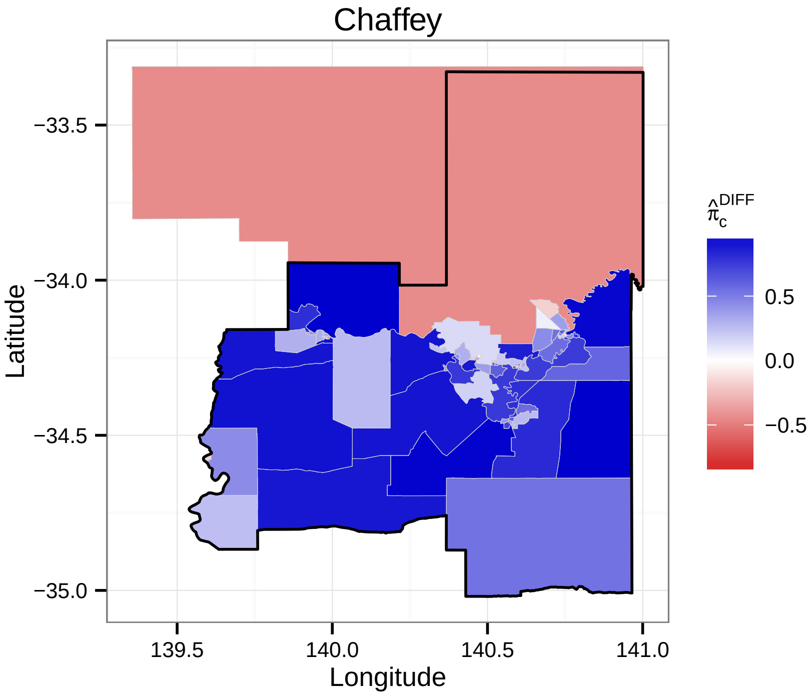


Figure 4.3.3: The electoral district of Chaffey, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.1.

places in Figure 4.2.3, with strong blue collection districts in more rural areas and weaker (but still strong) Liberal support in the towns.

We now turn to Norwood, which had a tighter margin of victory in 2010. Figure 4.3.4 shows Norwood, again with collection districts coloured by $\hat{\pi}_c^{DIFF}$.

We see from the plot that this electoral district contains both red and blue collection districts, mixed seemingly arbitrarily across the map. This is suspicious as we would expect neighbouring collection districts to have broadly similar demographics, and therefore not have wildly different fitted values.

In particular, consider collection district 4120602, in College Park on the western side of Norwood. It can be seen labelled on Figure 4.3.4 with an asterisk.

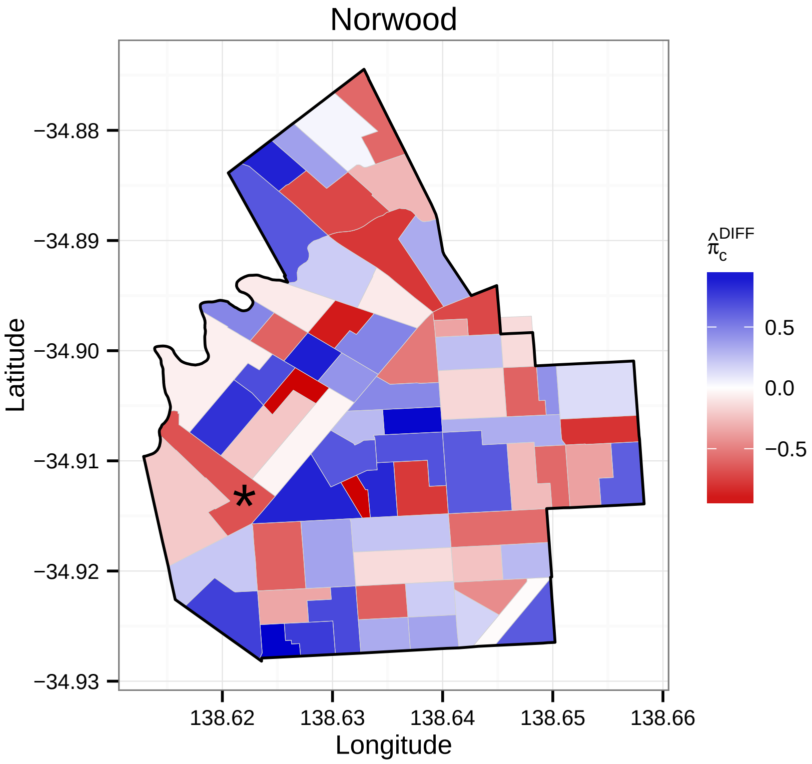


Figure 4.3.4: The electoral district of Norwood, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.1.

The prediction for the ALP vote in this district according to this model is 0.88. This is a very wealthy and highly educated collection district. Counter to the prediction, based on the literature review in Chapter 2 and the polling place results shown in Figure 4.2.2, we expect this collection district to vote strongly for the Liberal Party.

This area is also likely to be inconsistent with Figure 4.2.2, as the polling places that showed the strongest Liberal support in the election are near many strong ALP leaning collection districts in the prediction.

This model fit has given us results that are clearly not sensible in closely fought electoral districts, although they may be more accurate for safer electoral districts. Rather than proceed with assumption checking and interpretation of our coefficients, we now return to the issues with the data identified in Chapter 3.

In particular, there are problems with the extrapolation of highly aggregated data to the much smaller collection districts, and of performing the model fit using only 47 pieces of data. The data is also highly correlated which could cause problems in fitting.

We are unable to tackle the extrapolation issue at the moment as it is inherent to the model we are using. More detailed information and techniques will be required to resolve that issue.

For now though, we can reduce the correlations in the predictor data by using the principal component analysis results, as discussed in Section 3.8, as the predictors. In the next section we apply the same process to the modified data and validate the results again.

4.4 Model 4.2 - with all principal components

We first fit the multinomial regression using all 17 principal components. The coefficients, standard errors, and significant levels for this model fit are shown in Section A.2.2 in Appendix A.

The model is

$$\boldsymbol{\eta}^p_{ED} = \beta_0 + P^{ED}_{17}\boldsymbol{\beta}, \qquad p \in \{LIB, INF\},$$

and then apply the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ to calculate predictions for collection districts

$$\hat{\boldsymbol{\eta}}^p_{CD} = \beta_0 + P^{CD}_{17}\hat{\boldsymbol{\beta}}, \qquad p \in \{LIB, INF\},$$

where P^{ED}_{17} and P^{CD}_{17} are the design matrices containing the scores for each of the 17 principal components for each of the electoral districts and collection districts respectively.

We then perform model selection to determine the number of significant components that should be included in the regression. We use stepwise model selection by AIC. This is implemented automatically using the `stepAIC` function, part of the `MASS` package in R [39].

This model selection suggests that *all* of the 17 principal components are significant. Therefore, we do not remove *any* of the components. This is a somewhat surprising result given the correlations in the dataset we have previously investigated.

We produce the same figures for this model as were produced for Model 4.1 for this second model. In all cases the figures look near identical to their analogues in the first model. For completeness they are included in Section A.3 of Appendix A.

It is interesting that reorienting the predictors to reduce the correlation from the data has virtually no impact on the model fit. Later we tackle the theory of extrapolation as a cause for poor model fit, but in the next section we consider fitting the model to a reduced set of predictors.

4.5 Model 4.3 - with two principal components

In Table 3.8.1 we showed that over 92% of the variance in the data can be explained using only the first two principal components, so we now fit the multiple logistic model using only these two components, despite the fact that Model 4.2 showed that all the principal components are statistically significant. We call this Model 4.3.

With just these first 2 principal components, we fit

$$\boldsymbol{\eta}^{p}_{ED} = \beta_0 + P_2^{ED}\boldsymbol{\beta}, \qquad p \in \{LIB, INF\},$$

and then apply the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ to calculate predictions for collection districts

$$\hat{\boldsymbol{\eta}}^p_{CD} = \beta_0 + P_2^{CD}\hat{\boldsymbol{\beta}}, \qquad p \in \{LIB, INF\},$$

where P_2^{ED} and P_2^{CD} are the design matrices containing the scores for the first two principal components for the electoral districts and collection districts respectively.

Figure 4.5.1 shows the distribution of the predicted values $\hat{\pi}^p_c$ for each collection district c and party p, as in the first model fit. This time, the predicted values approach something much closer to a normal distribution, as we expect. The distribution of predicted informal votes also looks more sensible as there are no collection districts with a huge predicted $\hat{\pi}^{INF}_c$.

Figures 4.5.2 to 4.5.4 show the maps of the electoral districts of Cheltenham, Norwood, and Chaffey, with each collection district c coloured by its prediction for $\hat{\pi}^{DIFF}_c$ under Model 4.3.

The map of Cheltenham in Figure 4.5.2 shows that our predictions in this district have been pushed further toward the ALP, with almost all collection districts coloured red now. Before, they were a mix of values, and most collection districts were more extreme than they are now (strong for either the ALP or the Liberal Party).

The map of Norwood in Figure 4.5.3 shows a vast improvement on the predictions given by the first two multinomial regression models, with predictions that broadly fit our expectations.

As you move from the west of the electoral district to the east, you move from areas that strongly support the Liberal Party to collection districts that back the Labor party strongly. The collection district encompassing College Park is now coloured deep blue, aligning with our expectations.

Chaffey (Figure 4.5.4), on the other hand, shows predictions looking considerably

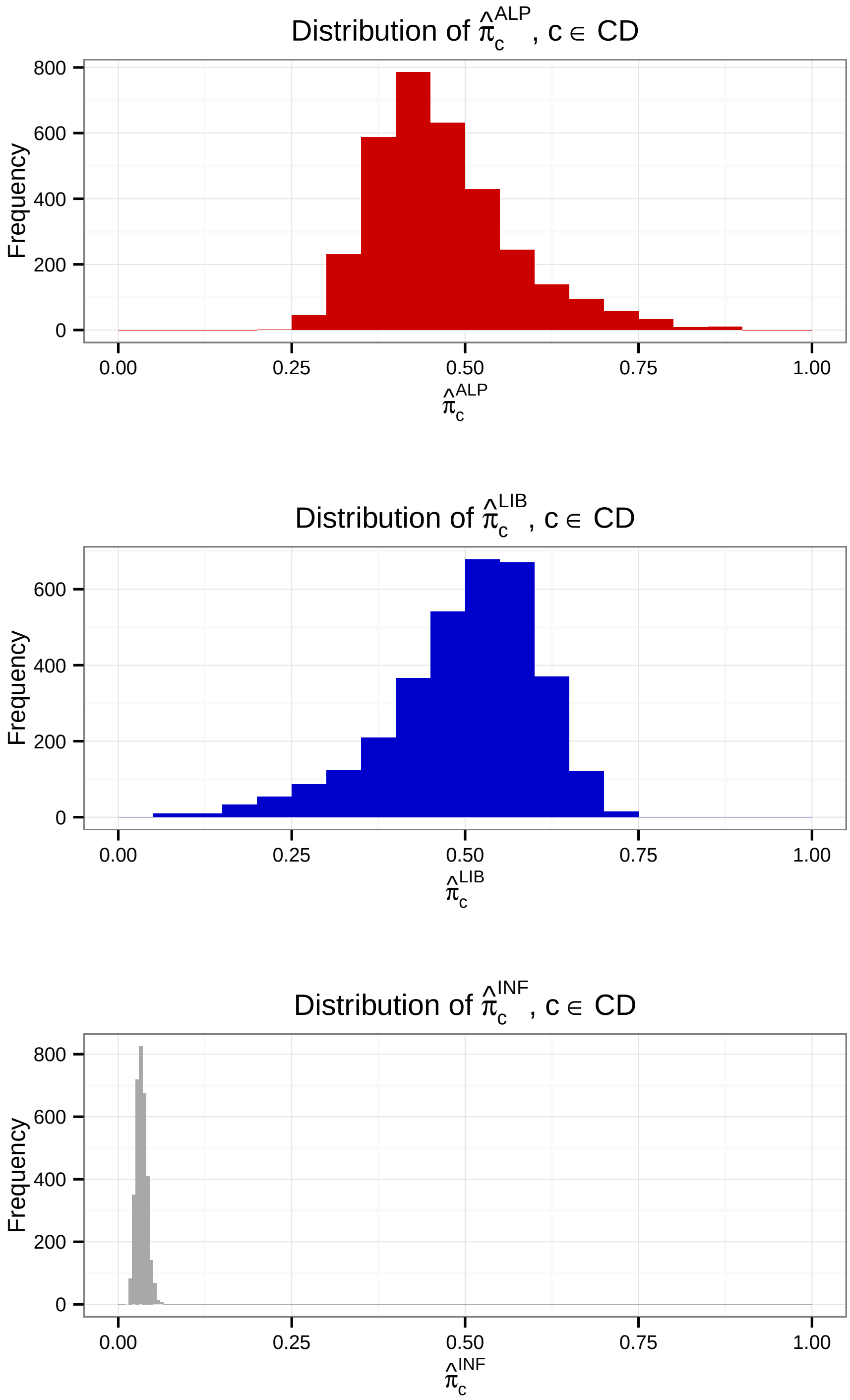


Figure 4.5.1: Distribution of predicted values for $\hat{\pi}_c^p$, for $p \in P$ and $c \in CD$, under Model 4.3.

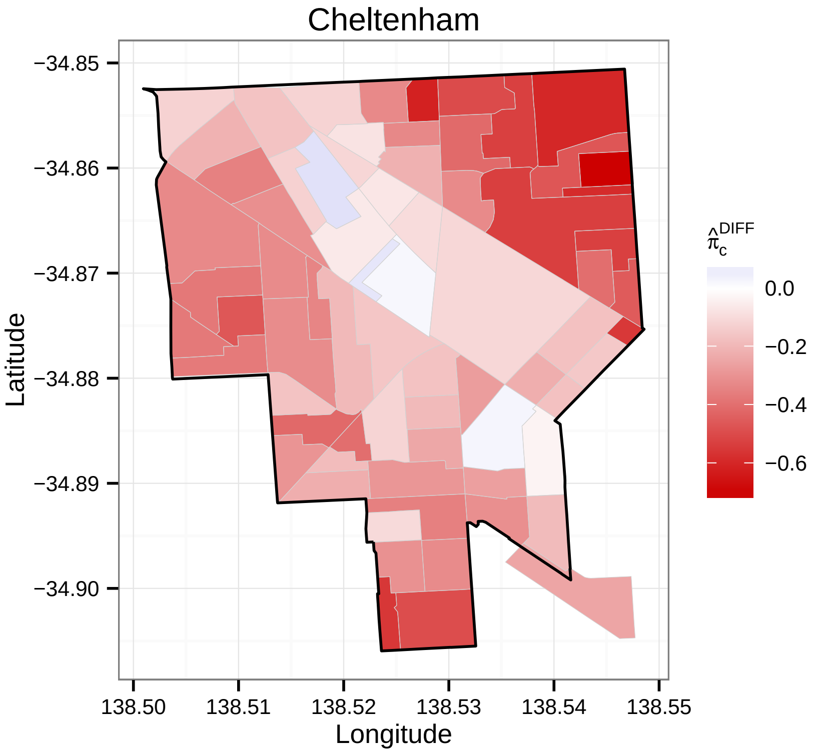


Figure 4.5.2: The electoral district of Cheltenham, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.3.

worse than those given by the previous models. Where before the vast majority of collection districts showed strong support for the Liberal Party, we now have a set of predictions where the seat appears much more marginal, and there are many collection districts in the more urbanised areas of Chaffey showing unexpectedly strong support for the ALP. The predictions for this electoral district do not appear sensible.

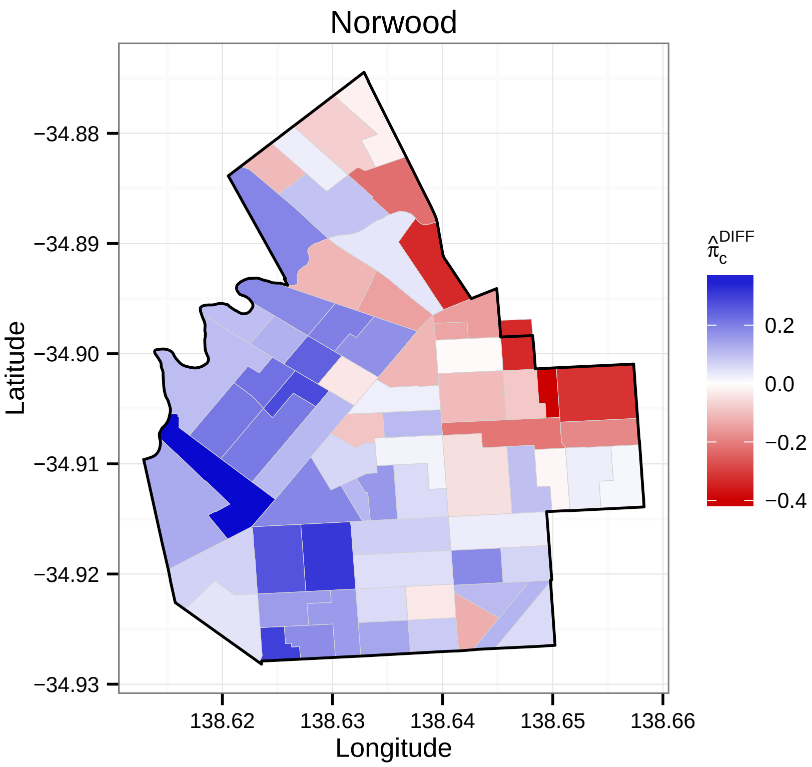


Figure 4.5.3: The electoral district of Norwood, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.3.

4.6 Discussion

In this chapter we have fit a family of multinomial regression models using a restricted set of the data available. We have also developed a technique for model validation that can be used for future refinements.

While the three multinomial regression models examined thus far have clear deficiencies, they are also giving predictions that have some of the correct qualitative features and so there is merit in further refining this work.

Due to the fact that Model 4.2 gave predictions that were no better than Model 4.1, and is considerably more difficult to interpret given the predictors are all linear

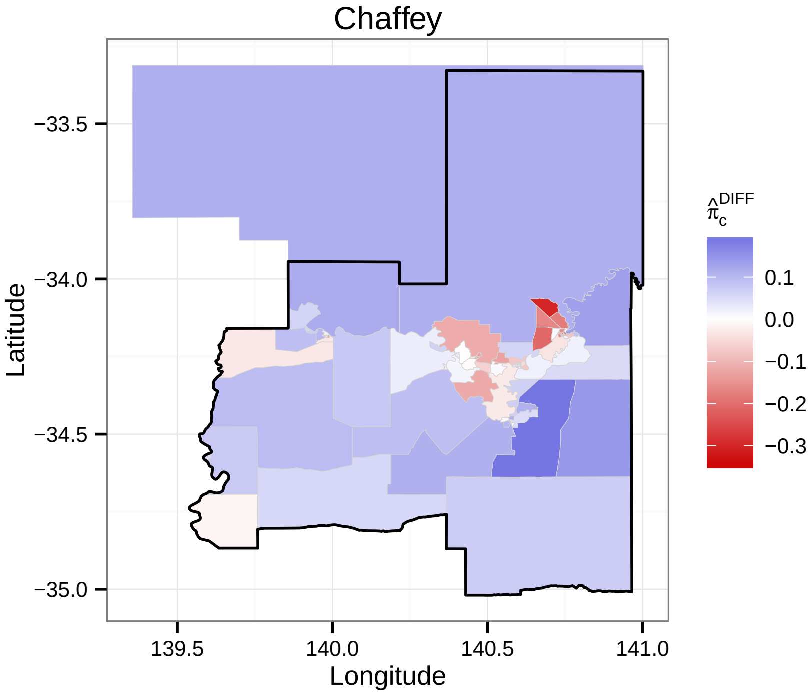


Figure 4.5.4: The electoral district of Chaffey, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 4.3.

combinations of our original set of predictors, we do not continue to consider models in the same form as Model 4.2.

In the next chapter we attempt to reduce the 'extrapolation' effect in the data that has already been discussed and make use of the more detailed data available. We develop a model that works with the polling place data directly, and fit analogues of Models 4.1 and 4.3.

Chapter 5

Modelling at Polling Place Level

Up to this point we have only considered two spatial layers in the system - the collection districts and the electoral districts. In this chapter we incorporate polling places, the third level at which we have information.

Recall that for collection districts we know the demographic data from the ABS but not voting data, for polling places we have voting data but not demographic data, and for electoral districts we have both.

As previously discussed, the EDBC have a pre-existing method of projecting the votes at polling place level to collection districts using the voter location data (introduced in Section 2.4.1). This is the method that we aim to improve by incorporating demographic data.

We can employ this same projection technique, in reverse, to model values for the predictors for the group of people that vote at each polling place, by aggregating the collection district predictors to polling place level.

With these predictors for polling places we can fit the multinomial regression model as in Chapter 4, to the 746 polling places rather than the 47 electoral districts. This should give better predictions for the π_c^p's, $c \in CD$, and should reduce the potential for over-fitting.

In this chapter we consider, check and clean the voter location data. We then develop the tools for estimating the predictors for polling places and then perform multinomial regression at that level. Finally we compare all of the regression models from both this chapter and Chapter 4 to see if there is a clearly best model.

5.1 Voter Location Data

When an elector votes in an election, they have a choice to cast an ordinary vote in any of the polling places in the electoral district that they reside in, or a declaration vote through a number of different means.

The voter location data is supplied by ECSA and describes the distribution of voters in each collection district across each polling place, including declaration voters. There are two different datasets that give us this information - one for the ordinary votes and one for the declaration votes.

5.1.1 Method of Voter Location Data Generation

On the day of an election, one or more electoral rolls are kept for each polling place, and each voter is crossed off an electoral roll at the polling place at which they present to vote. Attached to each person's entry on the electoral roll is the collection district in which they reside. Declaration votes are treated as all having been cast in one additional polling place per electoral district.

Therefore ECSA knows in which polling place each voter voted (although of course they do not know how they voted).

Table 5.1.1: Sample Voter Location data showing ordinary votes.

Collection District	Electoral District	Polling Place	Number of Votes
4010101	617	2180	6
4010102	617	630	1
4010102	617	2180	41
4010103	617	226	1
4010103	617	2180	13
⋮	⋮	⋮	⋮

Following each election, ECSA collects all of the electoral rolls and has them electronically scanned by a third party to determine which electors voted in each polling place. For each collection district, they then calculate the number of people in each collection district that voted in each polling place. These numbers form the voter location data.

This process is also used to identify people who voted more than once, amongst other things.

For the ordinary votes (that is, votes cast in polling places), a sample of the records is contained in Table 5.1.1.

The first row of this dataset tells us that 6 voters who were enrolled in collection district 4010101 and electoral district 617 presented at polling place 2180.

The dataset for declaration votes is very similar and tells us the number of people enrolled in each collection district who cast a declaration vote. We code a separate polling place to contain the declaration votes for each electoral district. For example, declaration votes cast by electors enrolled in electoral district 601 are assigned to polling place 601000, those in electoral district 602 are assigned to polling place

Table 5.1.2: Sample Voter Location data showing declaration votes.

Collection District	Electoral District	Number of Votes	Polling Place
4010101	617	1	617000
4010102	617	9	617000
4010103	617	13	617000
4010104	617	1	617000
4010105	617	16	617000
4010106	617	7	617000
⋮	⋮	⋮	⋮

602000, and so on.

A sample of the dataset for declaration votes is presented in Table 5.1.2. The final column, for polling place, has been generated by the author.

5.1.2 Data Verification and Cleaning

We first seek to verify the accuracy of the voter location data by comparing it to the published election results. We sum the voter location data over each polling place. Every polling place is allocated within exactly one electoral district as shared polling places are identified with multiple IDs to keep votes in distinct electoral districts separate.

Once the voter location data is aggregated in this way we can compare it to the actual number of votes that were cast in each polling place. A sample of comparisons is shown in Table 5.1.3.

As can be seen in the table, the number of votes that were actually cast in each

Table 5.1.3: Sample of comparisons between votes recorded in voter location data and votes actually cast in the election.

Polling Place	Elec. Dist.	Votes Cast	Votes in Location Data	Mismatch
5	601	955	958	-3
6	601	415	413	2
7	601	1557	1560	-3
8	601	612	612	0
10	601	902	901	1
11	601	2418	2412	6
⋮	⋮	⋮	⋮	⋮

polling place does not precisely match the number of votes recorded in the voter location data for each polling place. The 'mismatch' in each polling place is defined to be the number of votes cast in that polling place minus the number of votes recorded in the voter location data.

Errors in the voter location data can be caused in multiple ways, including for example:

- Human error in crossing off incorrect names on electoral rolls, or crossing off too many or too few names.
- Voters entering a polling place but not casting a ballot for the House of Assembly at all (that is, walking out of the polling place with their House of Assembly ballot paper). This is a legitimate and legal thing to do[1] but introduces error into the voter location data. Anecdotal advice from ECSA indicates that this happens more commonly in safe electoral districts than marginal ones.

[1] It is a requirement that every person on the electoral roll present at a polling place, or cast a declaration vote, but not that they actually fill out a valid vote, or submit a vote at all.

- Voters illegitimately claiming a declaration vote in an electoral district that they are not enrolled in. In these cases the House of Assembly ballot is rejected, but their ballot for the Legislative Council is still counted as the whole of the state is counted as a single electoral district for the Legislative Council.

Figure 5.1.1 shows the size of the mismatch for each polling place, plotted against the size of the polling place. Most of the mismatches lie close to zero, meaning that in most polling places the number of votes recorded and the number of votes actually cast are broadly similar. However, there are a small number of polling places in which the mismatch is very large.

The polling place with the greatest variance from the true number of votes cast is the Gilles Plains East polling place in Florey (polling place 2149 in electoral district 615). It recorded a mismatch of -408. That is, 408 *more* people were recorded as voting in this polling place in Florey than actually did vote.

This polling place happens to be a shared polling place with the electoral district Torrens (coded as polling place 647 in electoral district 643), and this polling place recorded a mismatch of 404. That is, 404 *fewer* people were recorded as voting in this polling place in Torrens than actually did vote.

Consider the map of Torrens as shown in Figure 5.1.2. This map shows all of the collection districts in which votes were recorded in the voter location data as being cast in Torrens, all shaded black. The red line shows the boundary of Torrens. The map indicates votes being cast in over 20 collection districts that are not located in Torrens. This is not permitted under the rules of the elections. Additionally, all of the votes indicated in these collection districts were recorded as being cast in the shared polling place 647.

This situation, where the voter location data indicates people voting in an electoral district that they do not live in, is repeated in four other electoral districts. All

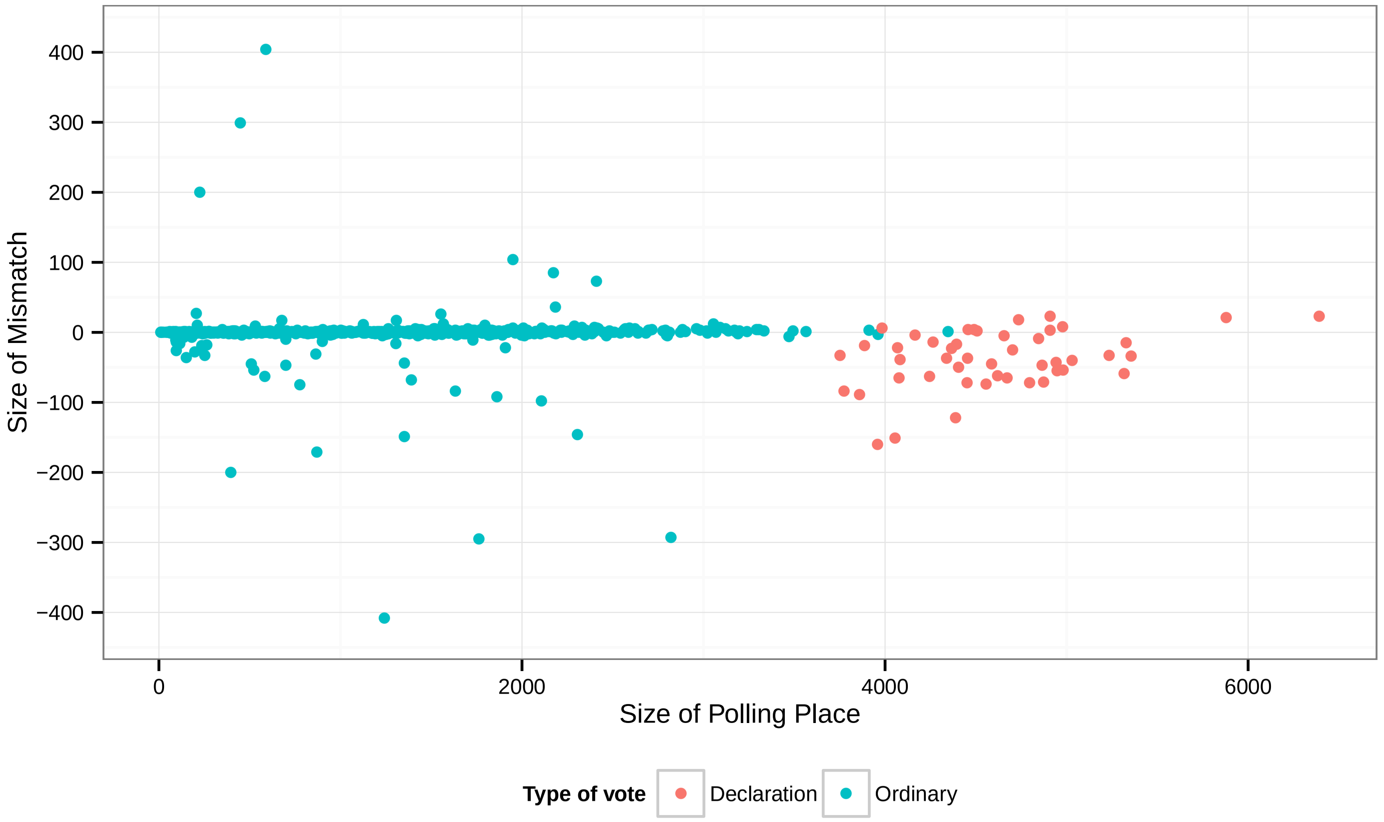


Figure 5.1.1: Plot of polling places, with the size of the polling place plotted against the size of the mismatch. That is, the difference between the number of votes recorded in that polling place in the voter location data, and the number of votes actually cast in the polling place. A positive mismatch means that more people were recorded in the voter location data than actually voted in the polling place. Polling places are coloured according to whether they contain ordinary or declaration votes.

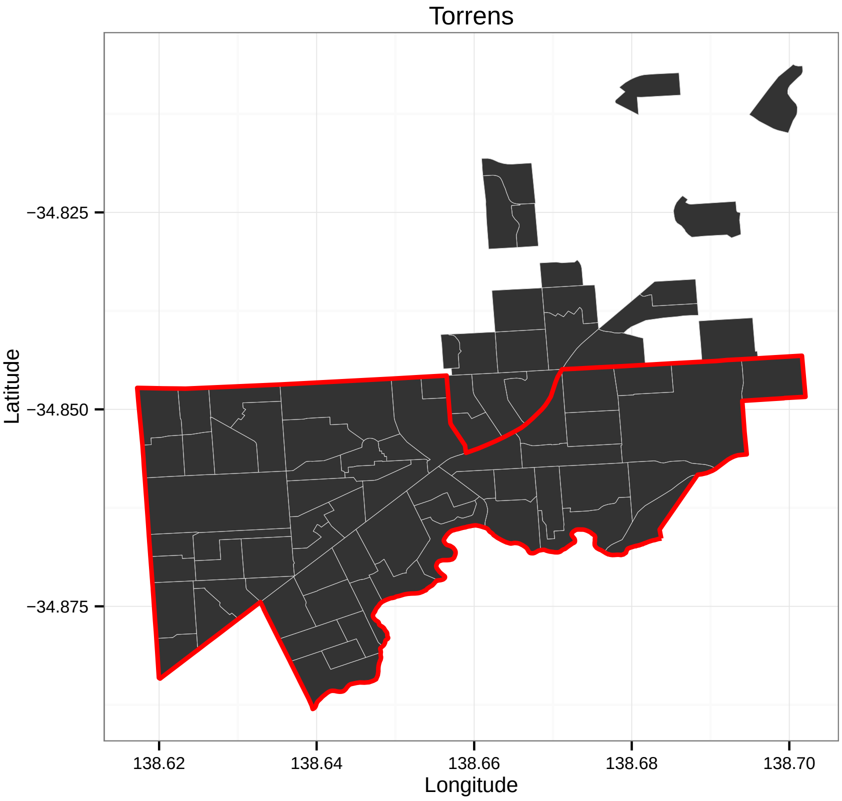


Figure 5.1.2: The electoral district Torrens, outlined in red. Collection districts that are recorded in the voter location data as having had at least one person vote in Torrens are displayed. Note that this does not mean a voter in each of the marked collection districts *actually* voted in Torrens, just that the voter location data records them doing so.

of these electoral districts also have shared polling booths with large mismatches. Figure 5.1.3 shows these five pairs of polling places in which votes are indicated as coming from outside an electoral district, and the complementary shared polling place.

In all five of these shared polling places, the positive and negative mismatches align very closely.

This set of facts naturally lead to the conclusion that the electoral rolls were mislabeled before being scanned, and voters were unintentionally allocated to the wrong

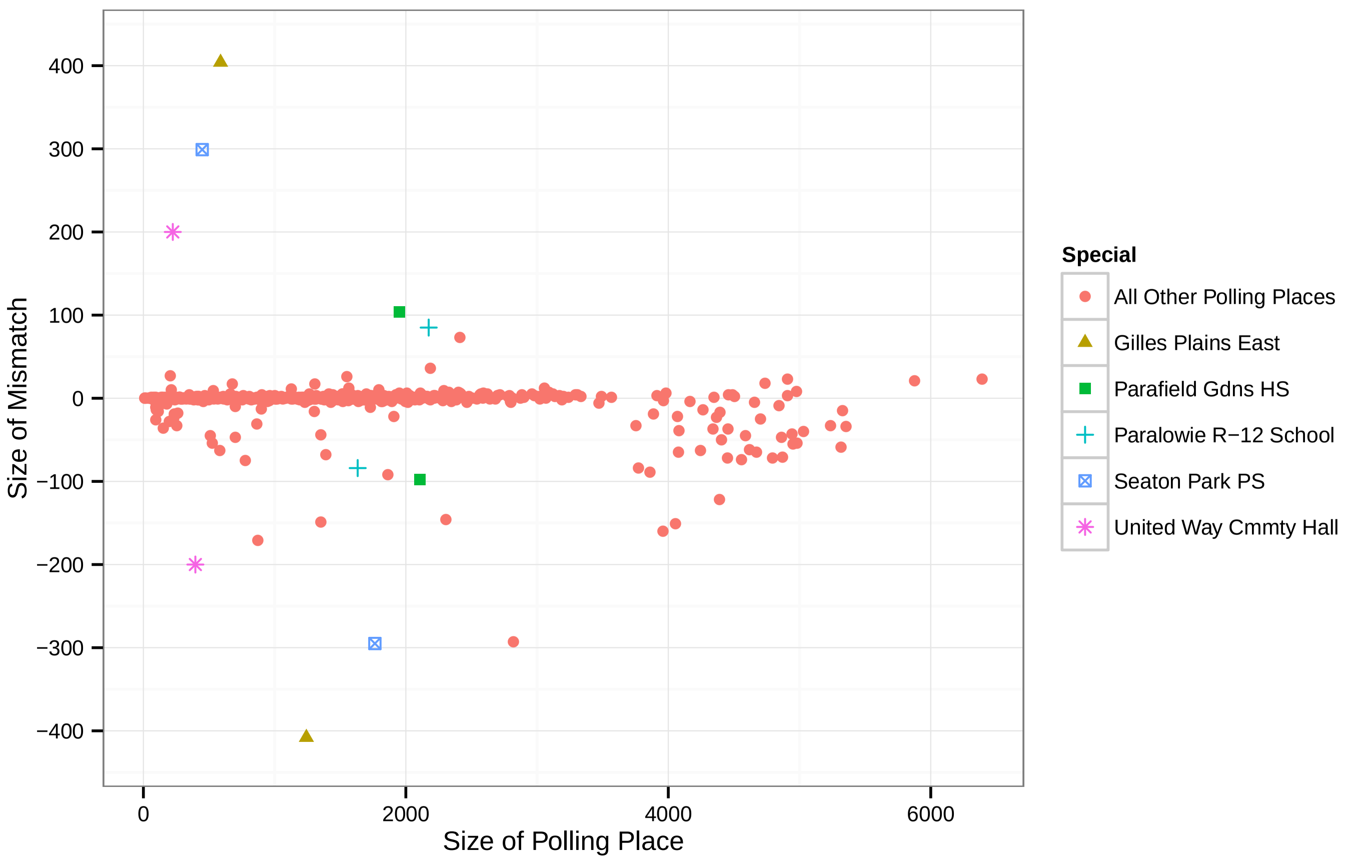


Figure 5.1.3: Plot of polling places, with the size of the polling place plotted against the size of the mismatch. That is, the difference between the number of votes recorded as voting in that polling place in the voter location data, and the number of votes actually cast in the polling place. A positive mismatch means that more people were recorded in the voter location data than actually voted in the polling place. Polling places are differentiated if they are among the shared polling places that indicate people voting in the incorrect electoral district.

polling place (and hence the wrong electoral district).

This conclusion is confirmed by ECSA and so we switch the polling place allocations in each of these five shared polling places. Doing this results in mismatches as shown in Figure 5.1.4.

It is important to note that this error in data collection in no way affected the result of the election, and no voter actually voted in the wrong electoral district because of this. It merely affects the voter location data.

As can be seen in Figure 5.1.4, there are still a number of large mismatches in the data, and these mismatches are mostly negative, meaning there are fewer votes recorded in the voter location data than were actually cast. Table 1.1.1 in Appendix A shows the 30 polling places with the greatest mismatches (the mismatches furthest from zero).

There is a clear pattern in these polling places, with the six greatest mismatches being in Electoral District 614, which is Flinders. In fact, 23 of the 30 largest mismatches are in Flinders.

After querying this with ECSA the following explanation was offered: the electoral rolls used were labelled with the polling place they were to be used in, but in 2010 in Flinders the electoral rolls were not allocated to polling places in accordance with this standard. This means that when the rolls were collated after the election and scanned, ordinary votes in Flinders were allocated to the incorrect polling places. ECSA are aware of this issue, and have made steps to correct this for subsequent elections.

Additionally there was no significant impact to the work of the electoral commission or the boundaries commission as a result of this error in roll allocations, since the electoral division is an incredibly safe one for the Liberal Party (in 2010 the Liberal Party received 76.2% of the two-party preferred vote) and the boundary of Flinders

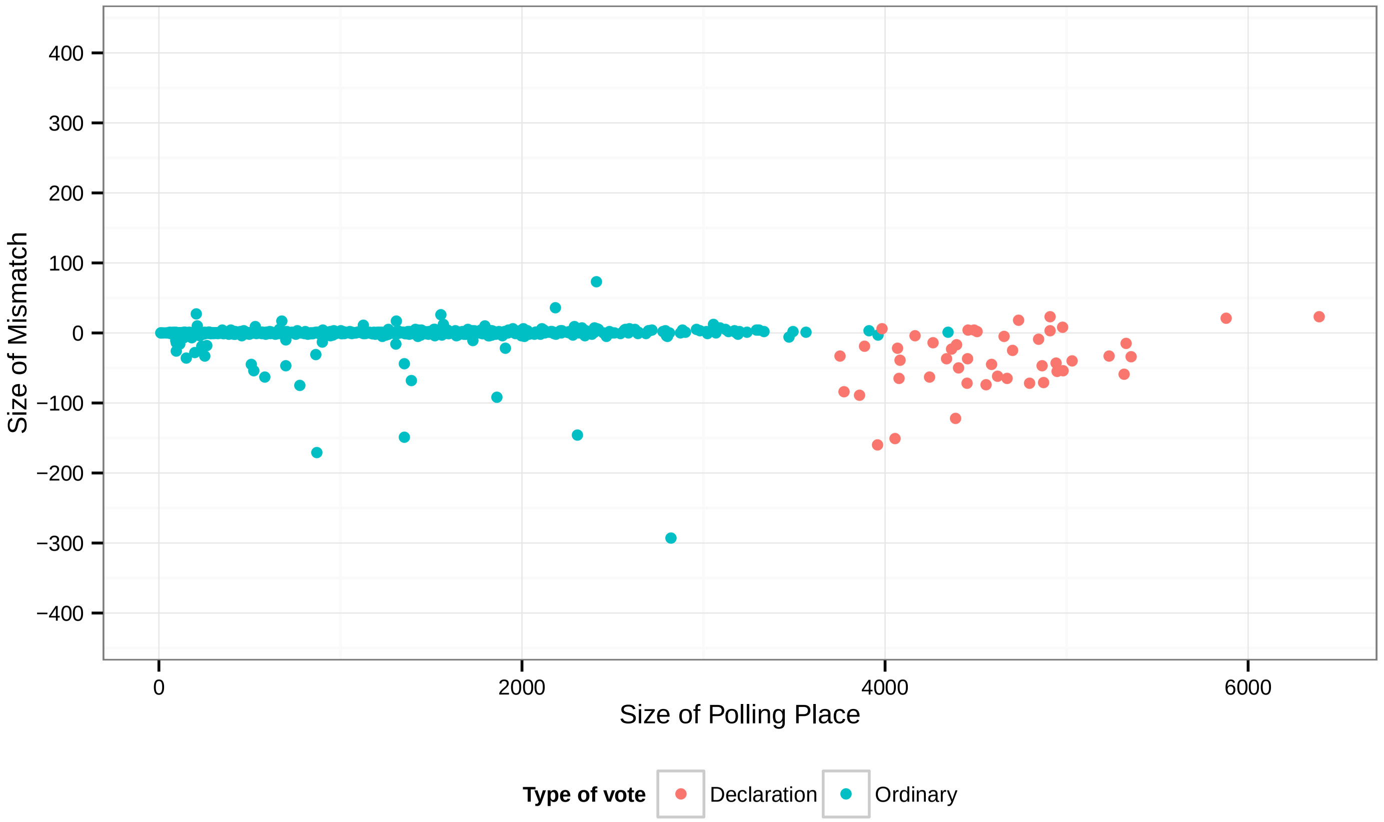


Figure 5.1.4: Plot of polling places, with the size of the polling place plotted against the size of the corrected mismatch, after allocations in the pairs of shared polling places identified in Figure 5.1.3 are switched. The mismatch is the difference between the number of votes recorded as voting in that polling place in the voter location data, and the number of votes actually cast in the polling place. A positive mismatch means that more people were recorded in the voter location data than actually voted in the polling place. Polling places are coloured by whether they contain declaration or ordinary votes.

was not changed in the redistribution immediately following the 2010 election.

It is unclear why the voter location data more often records fewer votes than were cast, because if the only mistake in the conduct of the election was in the distribution of rolls we would expect the mismatches to be centred around zero. Given the fact that we have no way of reversing the error already present in the data, we are forced to conclude that the data is irreparable and have to proceed with it as-is.

Finally, we consider the declaration votes, and the fact that the spread of the mismatches is much greater for these than for ordinary votes, as can be seen in Figure 5.1.4. The mismatches are also predominantly below zero for declaration votes.

These larger negative mismatches are not surprising to ECSA, due in part to the process of completing a declaration vote, and the way in which they are counted.

The set of declaration votes contains a number of different types of vote, two of which are postal votes and absentee votes. When an elector applies for a postal vote they receive their ballot papers in the mail. The voter completes their ballots, then places them inside an envelope. The envelope has a detachable flap and the voter must sign a declaration on this flap. They then put the envelope into another envelope that is pre-paid and pre-addressed to ECSA and post it back to ECSA.

Absentee votes are cast by people who vote on election day, but at a polling place that is not located in their electoral district. The process for actually completing an absentee ballot is very similar to that for postal votes, with the exception that the envelopes do not need to be posted to ECSA as they are completed in a polling place.

After the declaration vote arrives with the electoral commission and all polls have closed, the declaration vote envelopes are all considered by the returning officer. Each voter's envelope is considered and, if it is accepted, the detachable flap identifying the elector is removed from the envelope (without actually opening the envelope) and the elector is marked off the roll as casting a declaration vote.

Reasons a declaration vote could be refused include the elector having already cast an ordinary vote in the election, inegibility to vote in the election, an unsigned declaration, and the signature on the declaration envelope not matching the elector's name.

Once this process is completed, the now anonymous declaration vote envelopes are opened and the ballots in them are counted. This means that if an accepted envelope is empty, then a vote would be recorded in the voter location data but no corresponding vote would actually be counted.

It is possible for an elector to walk into a polling place on the day of an election, get their name marked on the electoral roll, and then walk out of the polling place with their ballots, but anecdotal evidence from ECSA suggests that this is much more likely for declaration votes.

When voting in a polling place there is significant pressure to complete and lodge formal votes, with ECSA staff directing voters to private booths in the polling place, and then ballot boxes at the exit staffed by a polling attendant who directs voters to place their ballots in the correct box. None of this infrastructure exists for people completing postal votes. ECSA suggests this leads to a higher rate of non-voting, and that this is more common in safe seats than marginal ones.

We conclude that there is no reason to believe that there are errors in the voter location data for declaration votes. We also note that given trends in Australian elections of increasingly higher proportions of declaration votes, we would expect the size of the mismatch to increase in future. This could cause problems were we to try and replicate this work for elections held subsequent to this research.

5.2 Demographics of Polling Places

With the voter location data we can establish estimates for the demographics of the group of people who voted in each polling place. We use the ECSA records of how many people in each collection district actually voted at each polling place and calculate estimates for the predictors for each polling place b. To do this we employ essentially the same method used by the EDBC to estimate the π_c^p's in collection districts, but in reverse.

Recall that X^{CD} is the design matrix of predictors at collection district level, as previously used and generated from the ABS dataset. We now develop X^{PP}, the design matrix of predictors at *polling place level.*

We build the polling place data by weighting the collection district data according to the proportion of voters attending that polling place that came from each collection district.

We let l_b be the number of votes cast in polling place b, and l_{cb} be the number of votes cast in polling place b, by people that live in collection district c.

To calculate X^{PP} we apportion the estimated votes in collection districts according to the voter location data. The value of predictor m in polling place b can be estimated using the relations

$$\left[X^{PP}\right]_{bm} = \sum_{\{c|l_c>0\}} \frac{l_{cb}\left[X^{CD}\right]_{cm}}{l_b} \qquad \text{for all } b \in PP, p \in P.$$

We construct the $|CD| \times |PP|$ matrix L such that $[L]_{cb} = l_{cb}$ for $c \in CD$ and $b \in PP$, and then calculate the matrix $L^{CD\to PP}$ where

$$\left[L^{CD\to PP}\right]_{cb} = \frac{l_{cb}}{l_b}.$$

We can then simply compute the estimate

$$X^{PP} = \left(L^{CD\to PP}\right)^T X^{CD}.$$

Hence $L^{CD \to PP}$ can be thought of as the matrix which transforms collection districts to polling places.

With these modelled predictors at polling place level, we return to the multinomial regression models of the previous chapter, and fit them to the 746 polling places rather than only 47 electoral districts. In the next sections we perform the polling place level version of Models 4.1 and 4.3.

5.3 Model 5.1 - with original 17 predictors

We now fit the multinomial regression model fitting the 17 predictors to the election results in the 746 polling places, and call this Model 5.1. This is a direct analogue of Model 4.1 from the previous chapter, but at the polling place level. The model is

$$\boldsymbol{\eta}^p_{PP} = \beta_0 + X^{PP}\boldsymbol{\beta}, \qquad p \in \{LIB, INF\},$$

and we then apply the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ to calculate predictions for collection districts

$$\hat{\boldsymbol{\eta}}^p_{CD} = \hat{\beta}_0 + X^{CD}\hat{\boldsymbol{\beta}}, \qquad p \in \{LIB, INF\},$$

where X^{PP} and X^{CD} are the design matrices containing the values of the predictors for each of the polling places and collection districts respectively.

We again include the model output in Section A.2.4 of Appendix A, and consider the same diagnostic plots as were used in the previous chapter to validate the results.

The distribution of the predicted values for each of the parties for each of the collection districts are shown in Figure 5.3.1.

In contrast to Model 4.1, we can see that the distributions of the π^p_c are unimodal for all $p \in P$. We see that the distribution of $\hat{\pi}^{ALP}_c$ is skewed to the left, and the distribution of $\hat{\pi}^{LIB}_c$ is skewed to the right. The means of both distributions appear

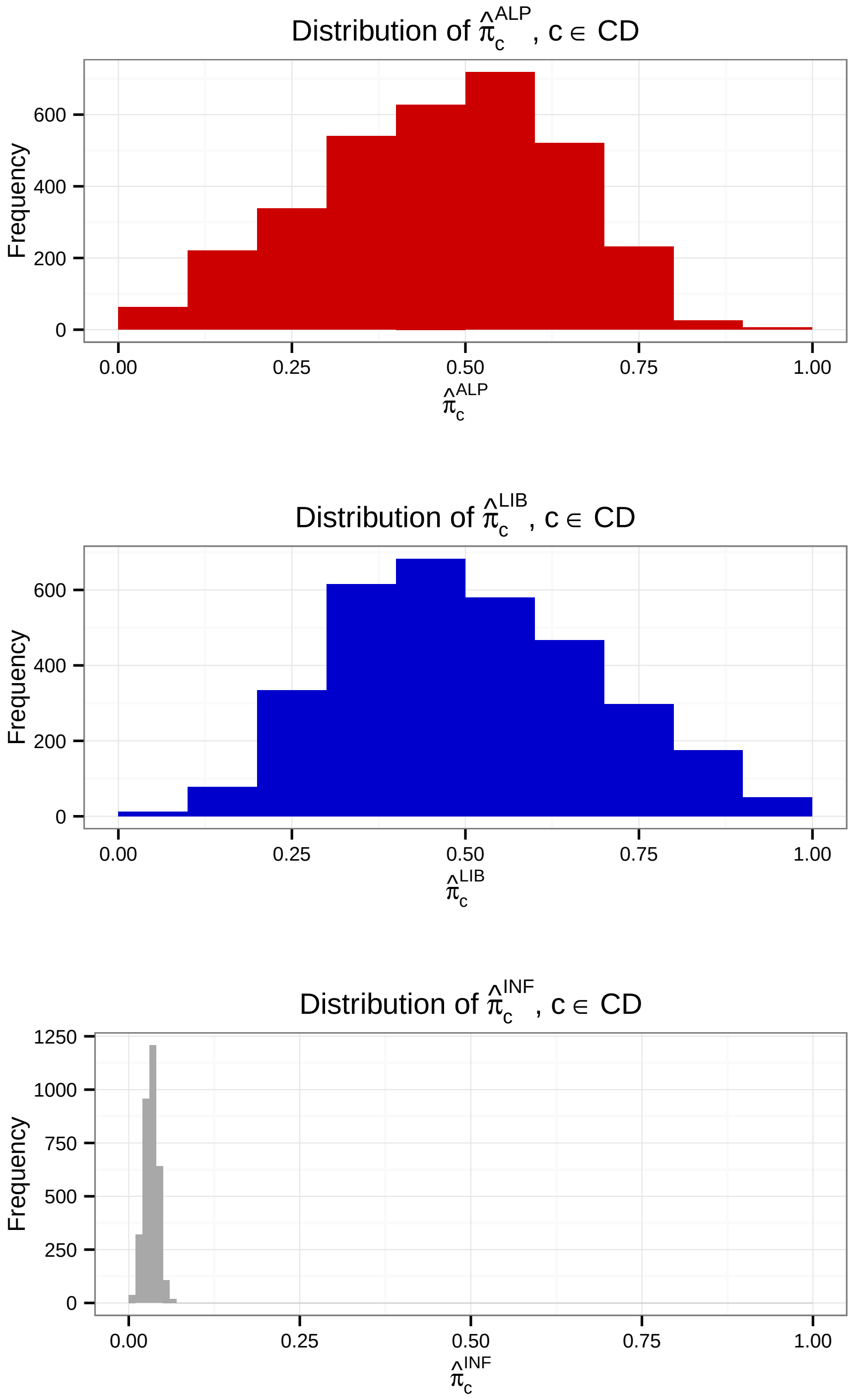


Figure 5.3.1: Distribution of predicted values for $\hat{\pi}_c^p$, for $p \in P$ and $c \in CD$, under Model 5.1.

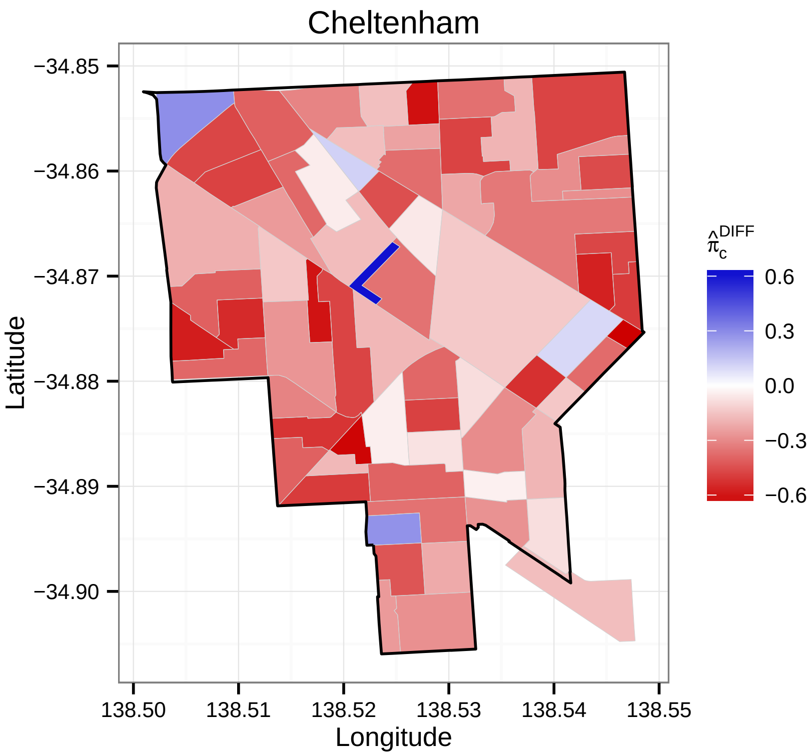


Figure 5.3.2: The electoral district of Cheltenham, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 5.1.

to be in the vicinity of 0.5, as we expect. There is again a wide spread of predictions, with them lying across the entire interval $[0, 1]$. The spread of $\hat{\pi}_c^{INF}$ is very narrow and centred around the true mean for informal votes.

As before we also consider predictions for the three chosen electoral districts: Cheltenham (Figure 5.3.2), Chaffey (Figure 5.3.3), and Norwood (Figure 5.3.4).

Virtually all of the collection districts in Cheltenham are coloured red and there are pockets of extremely strong Labor support. The figure is consistent with the results data and our understanding of voting patterns in the electoral district.

In particular there is one very deep blue prediction under Model 5.1. The collection district in question is 4101004, and the prediction for $\pi_{4101004}^{ALP}$ is 0.186.

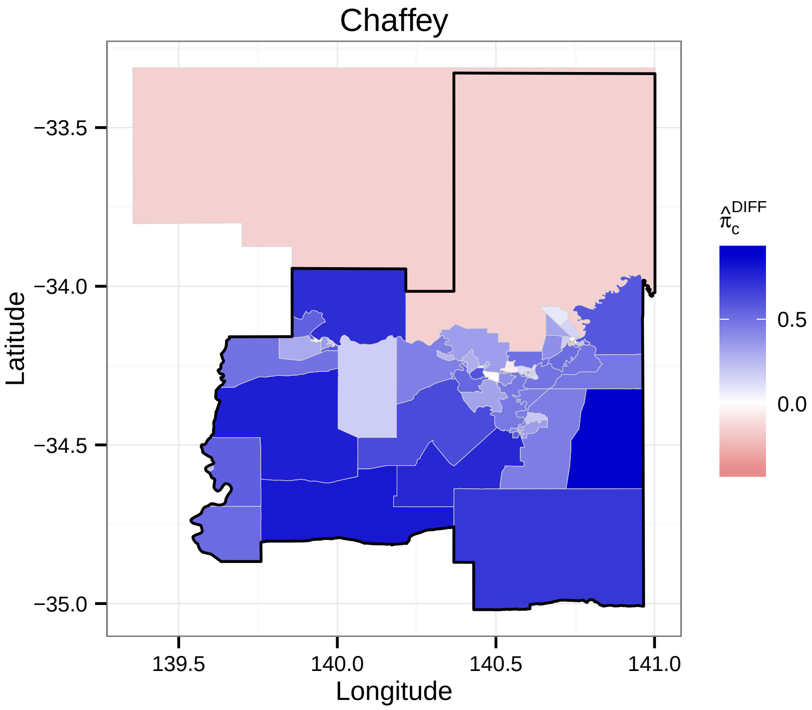


Figure 5.3.3: The electoral district of Chaffey, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 5.1.

This collection district has significantly fewer non-English speakers than the mean in Cheltenham, with approximately 11.2% of people speaking a language other than English at home, compared to the average of collection districts in Cheltenham of around 28.5%. This is the third lowest proportion of non-English speakers in the seat.

It is also the collection district with the third highest household income, according to our aggregate statistics, and is the second highest on our measure of non-school education.

The fact that this collection district is an extremity within Cheltenham is likely the reason for the very different prediction from its neighbours. However it seems

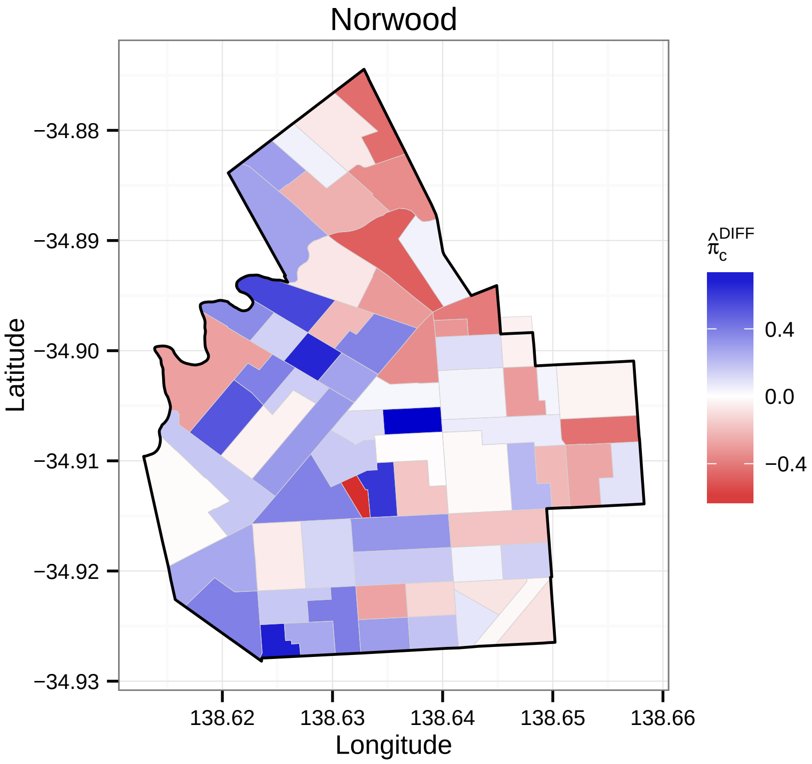


Figure 5.3.4: The electoral district of Norwood, with each collection district c coloured according to $\hat{\pi}_c^{DIFF}$, $c \in CD$, under Model 5.1.

unbelievable that a collection district would record a vote for the ALP of less than 20%, while all of its neighbours recorded a vote for the ALP over 50%.

Chaffey also appears as expected, with most collection districts coloured blue and indicating strong support for the Liberal party. This support is strongest in the more rural areas of the district.

Norwood is a more marginal seat, and so there is more local variation in the predictions for the seat, but the broad trends of the seat are again what we expect: stronger Liberal support in the west of the district closer to the Adelaide CBD, and stronger Labor support on the eastern side of the seat.

Generally speaking, predictions under this model appear sensible. We consider more statistics to measure model fit shortly, but in the next section we fit one more multinomial regression model, the analogue of Model 4.3.

5.4 Model 5.2 - with two principal components

Recall that in the previous chapter we chose to discard Model 4.2, fit on a set of principal components chosen by model selection, as it did not perform significantly better than the predictors themselves, and is considerably more difficult to interpret.

We still fit the polling place level analogue of Model 4.3, in which we fit the model to the polling places against only two predictors, the first two principal components. To be consistent with the method of generating polling place predictors, we use the loadings from the principal component analysis performed on *collection districts* to calculate the first two scores for each polling place, rather than the principal component analysis performed on electoral districts in Models 4.2 and 4.3.

We fit the model

$$\boldsymbol{\eta}^p_{PP} = \beta_0 + P_2^{PP}\boldsymbol{\beta}, \qquad p \in \{LIB, INF\},$$

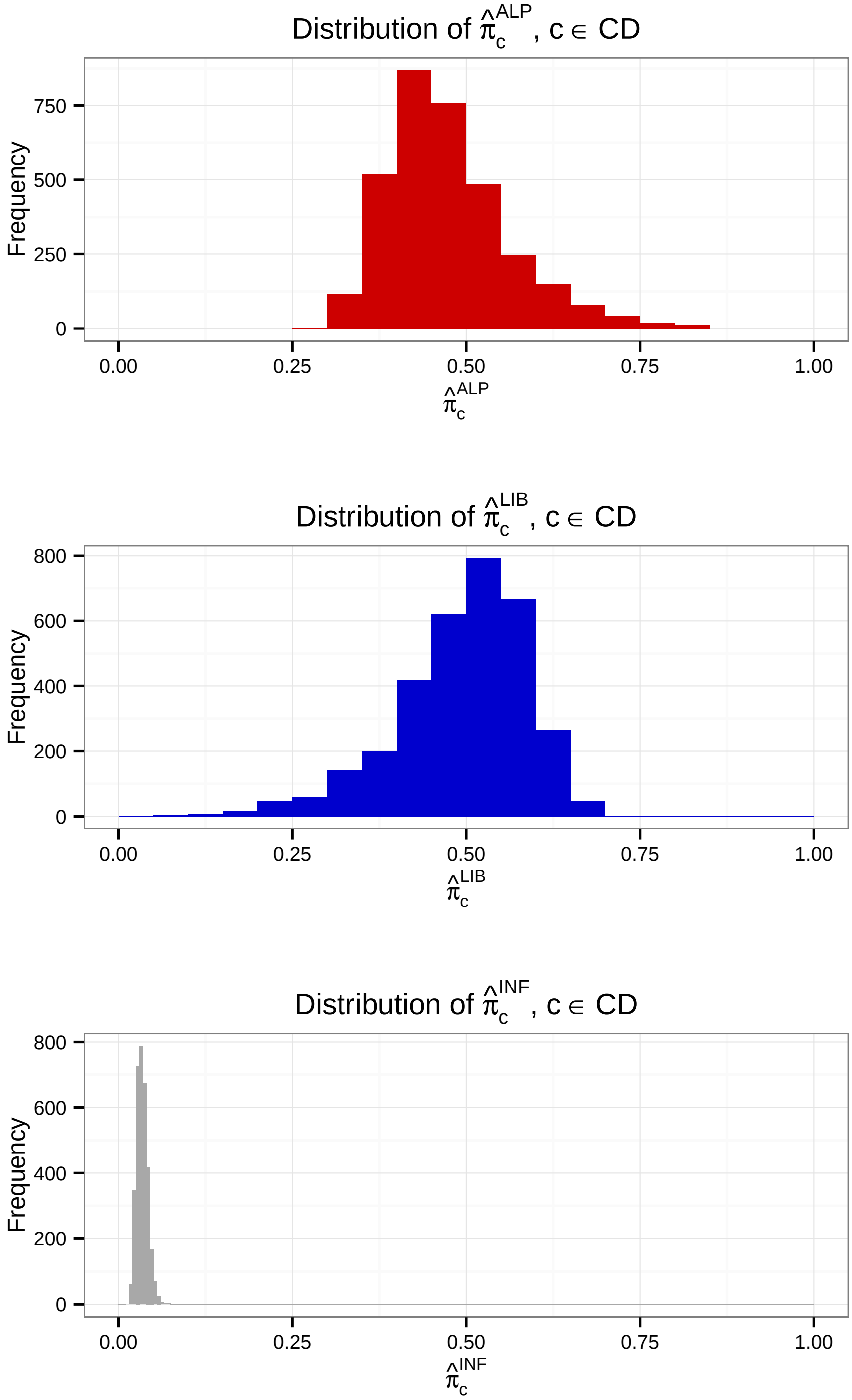
and then apply the estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ to calculate predictions for collection districts

$$\hat{\boldsymbol{\eta}}^p_{CD} = \hat{\beta}_0 + P_2^{CD}\hat{\boldsymbol{\beta}}, \qquad p \in \{LIB, INF\},$$

where P_2^{PP} and P_2^{CD} are the design matrices containing the scores for each of the 2 principal components for each of the polling places and collection districts respectively.

Once again we validate the model with the same tools. The distributions of each of the π^p_c's are shown in Figures 5.4.1.

Figure 5.4.1 shows a right skewed unimodal distribution of $\hat{\pi}^{ALP}_c$, $c \in CD$. The spread of predictions is narrower than the spread for Model 5.1. Again the mean



Distribution of $\hat{\pi}_c^{ALP}$, $c \in$ CD

Distribution of $\hat{\pi}_c^{LIB}$, $c \in$ CD

Distribution of $\hat{\pi}_c^{INF}$, $c \in$ CD

Figure 5.4.1: Distribution of predicted values for $\hat{\pi}_c^p$, for $p \in P$ and $c \in CD$, under Model 5.2

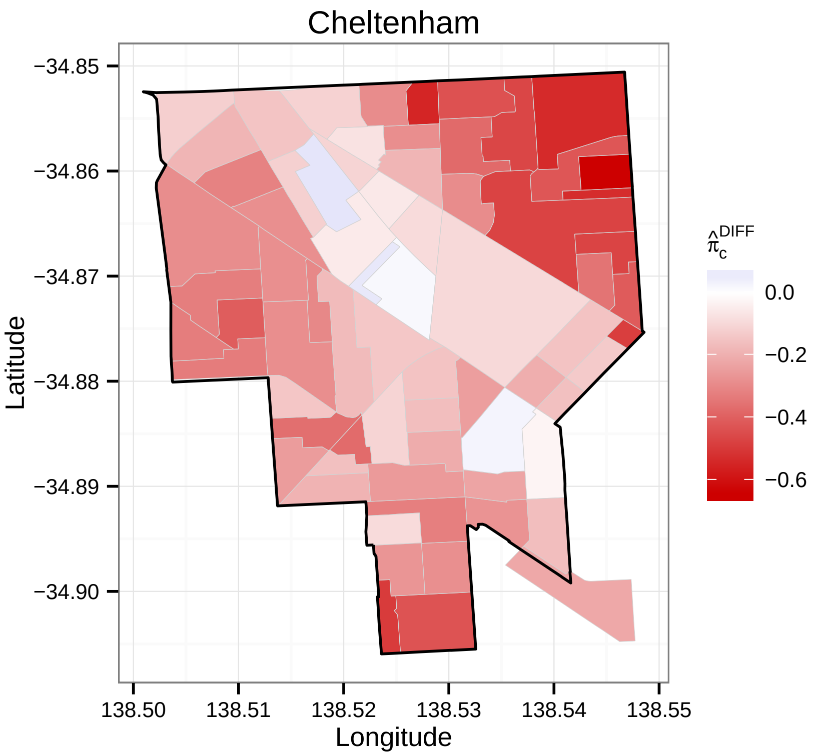


Figure 5.4.2: The electoral district of Cheltenham, with collection district i coloured according to $\hat{\pi}_i^{DIFF}$ under Model 5.2.

appears roughly in the place we would expect, although the majority of collection districts lie below the mean due to the long tail.

As we would expect, the predictions for $\hat{\pi}_c^{LIB}$ shows essentially the same pattern and appear roughly like $1 - \hat{\pi}_c^{ALP}$. This is unsurprising and a good sign for the model fit.

The distribution of $\hat{\pi}_c^{INF}$ again shows a unimodal distribution, and the long tail we saw in Model 4.1 is now largely gone. The mean and spread are again in the vicinity of what we would expect.

The predictions for Cheltenham shown in Figure 5.4.2 are again in line with our expectations and show strong Labor support across the entire electoral district.

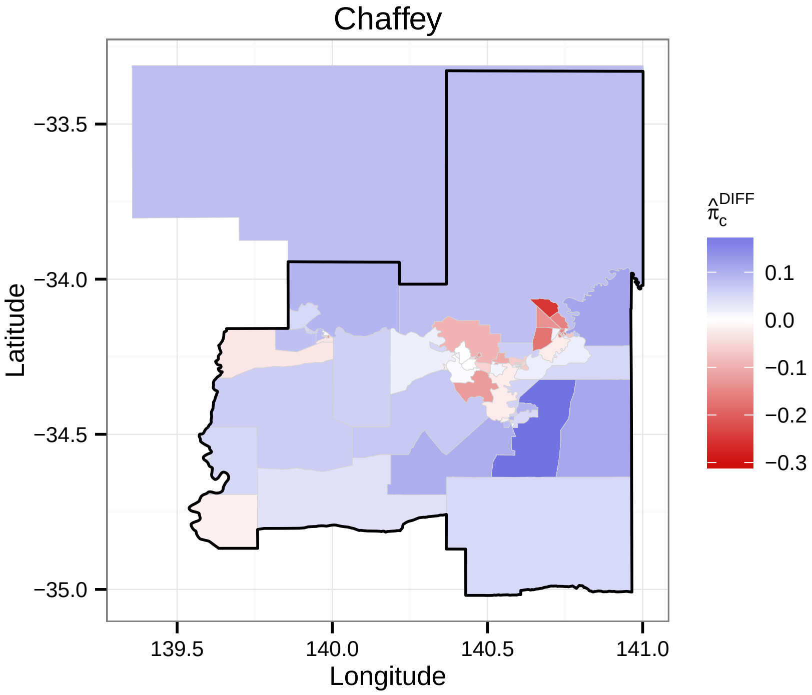


Figure 5.4.3: The electoral district of Chaffey, with collection district i coloured according to $\hat{\pi}_i^{DIFF}$ under Model 5.2.

There is considerably less variation in the predictions here, with the plot showing more uniform trends across the district than the previous model.

In Chaffey (Figure 5.4.3) we see collection districts that are largely coloured as we expect, but the predictions are much more marginal than we would expect given the two-party preferred margin in the seat. We would expect more collection districts to be coloured a much darker shade of blue in this seat, and so we do not hold much faith in these predictions.

Finally, Norwood (Figure 5.4.4) behaves in a similar way to Cheltenham, with predictions in line with expectations and showing a gradual transition from Liberal support to ALP support moving from west to east. Again there is less variation in

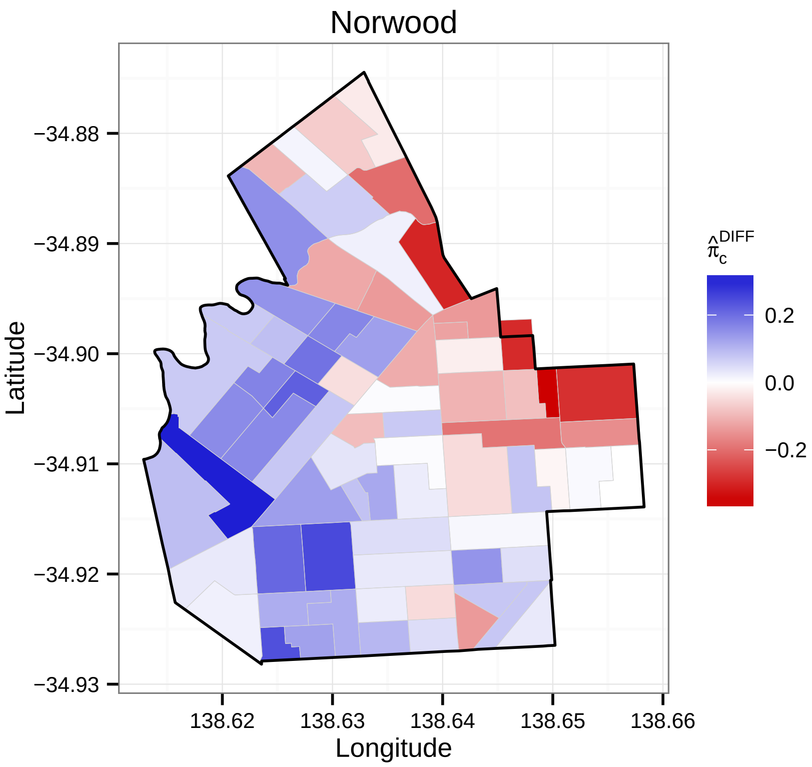


Figure 5.4.4: The electoral district of Norwood, with collection district i coloured according to $\hat{\pi}_i^{DIFF}$ under Model 5.2.

the predictions under Model 5.2 than under Model 5.1.

Overall, the predictions in this model do appear sensible but we see a shrinking in the differentiation between collection districts due to the severe reduction in the number of predictors.

5.5 Comparing Multinomial Regression Models

We now compare all five multinomial regression models with the aid of some new summary statistics, as well as the visualisations of predictions.

5.5.1 Summary Statistics

We consider a number of summary statistics on our models to see numerically how close the predictions are to the real election results. This is intended to give an idea of the goodness of the models, where they perform well and where they perform poorly.

The predictions for the π_e^p in each of the sample electoral districts we have used as case studies are calculated by aggregating the π_c^p for the collection districts in that electoral district, and for the state of South Australia as a whole to give a simple and intuitive comparison. The predictions for the three districts and also for the statewide vote are shown in Table 5.5.1, along with the actual 2010 election result.

Table 5.5.1 shows that on average, all models perform relatively well on predicting the two-party preferred vote across the state, with the models fit to polling places performing better than the models fit to electoral districts.

Different models have different strengths and weaknesses when it comes to the three sample electoral districts considered. Consider Model 4.3, which does a relatively good job predicting the vote in Cheltenham, but an incredibly poor job in Chaffey where the model predicts the ALP winning despite the fact that it is incredibly safe for the Liberal Party on two-party preferred terms. In fact, none of the models do a particularly good job fitting to the seat of Chaffey (recall this may be in part due to the abnormal results in that electoral district).

The seat of Cheltenham is fit relatively well across all models, with the standout performer being Model 5.1, which fit the raw predictors to the polling places.

Seat-wide, Norwood is very well predicted across the board, with Models 4.1, 4.2 and 5.1 being the standouts.

Table 5.5.1: Predictions $\hat{\pi}_e^p$ for the subset of the electoral districts e containing Cheltenham, Chaffey, Norwood, and also the whole of South Australia, under each of the five multinomial regression models, compared to the true π_e^p from the 2010 election results. These summary statistics are calculated by weighting the predictions $\hat{\pi}_c^p$ according to the number of votes in each collection district c.

District	Party	2010 Result	Multinomial Regression Model				
			4.1	4.2	4.3	5.1	5.2
South Australia	ALP	0.469	0.475	0.475	0.462	0.471	0.468
	Lib	0.498	0.490	0.490	0.505	0.496	0.499
	Inf	0.033	0.035	0.035	0.033	0.033	0.033
Cheltenham	ALP	0.632	0.596	0.596	0.615	0.624	0.600
	Lib	0.324	0.352	0.353	0.341	0.331	0.356
	Inf	0.043	0.051	0.051	0.043	0.045	0.044
Chaffey	ALP	0.215	0.274	0.274	0.491	0.337	0.492
	Lib	0.754	0.689	0.688	0.471	0.631	0.470
	Inf	0.031	0.037	0.037	0.038	0.032	0.038
Norwood	ALP	0.436	0.457	0.457	0.463	0.458	0.474
	Lib	0.530	0.512	0.512	0.507	0.513	0.497
	Inf	0.034	0.030	0.030	0.029	0.029	0.030

5.5.2 Comparing ‘goodness of fit’

We now seek to find a way of comparing the overall appropriateness of our various models. The obvious way to do this is by comparing the predicted fit of the entire models to the observed elections results with a single summary statistic.

We have predicted votes for each party at collection district level, but have no observed votes at this level. We need to aggregate our predictions to polling place level, at which we do have observed votes.

Table 5.5.2: Overall error E of each of the five multinomial regression models.

Model	Description	E
4.1	Regression on 17 predictors (ED level)	66,109,303
4.2	Regression on 17 principal components (ED level)	66,117,400
4.3	Regression on first 2 principal components (ED level)	107,625,736
5.1	Regression on 17 predictors (PP level)	39,329,162
5.2	Regression on first 2 principal components (PP level)	108,369,447

Then we can measure the closeness between our predicted model and the observed election as

$$E = \sum_{b \in PP} \left(\sum_{p \in P} \left| Y_b^p - \hat{Y}_b^p \right| \right)^2 . \tag{5.5.1}$$

We call E the overall error for the model, and we prefer models for which E is smaller. There is no way that E will ever be zero, as $\sum_{p \in P} \left| Y_b^p - \hat{Y}_b^p \right|$ must be greater than or equal to the mismatch in that particular polling place for each polling place b. The sum of the squares of the mismatches is 357,337, and this is therefore a lower bound on E.

For each of the five multinomial regression fits performed at the start we now compute E. The results are summarised in Table 5.5.2.

The first thing to notice is the magnitude of E for all models. An error of 60,000,000 corresponds to the model being around 284 votes out in each polling place, across each of the three parties. This is an error rate of around 20% of the size of an average polling place. However, note that we are essentially double counting each error, as an overestimate of the vote for one party must be offset by an underestimate for one of the other parties, so this figure can be misleading. Regardless of this, the magnitude of the error E is still larger than we would like.

By a significant margin, the ‘closest’ model to the real election is Model 5.1. Models 4.1 and 4.2 perform similarly, as do Models 4.3 and 5.2. The large errors in Models 4.3 and 5.2 are likely to be due to the significantly limited differentiation between predictions, caused by the fact that we only use 2 predictors. The fact that the predictions of Models 5.1 are significantly ‘closer’ than those for Models 4.1 and 4.2 is probably due to the huge increase in information that we use in Model 5.1 by fitting to the polling places rather than just the electoral districts.

We return to these numerical measures of fit again at the end of this chapter, after comparing the visual indications of fit side by side.

As future research, we could examine the robustness of these models and the predictions they produce by performing cross validation, but this is outside the scope of this study.

5.5.3 Graphical Analysis of Fit

We now compare the visualisations of predictions from each of the five models to gain an understanding of the relative strengths and weaknesses of each of the models. Side by side plots of each of the visualisations are included in Appendix A.4 to aid in comparisons.

First we consider the set of distributions of predictions for the major parties, shown in Figures A.4.1 and A.4.2. The side-by-side plots are arranged with the first three models performed at electoral district level on the left, and the two models performed at polling place level on the right. Models 4.1 and 5.1 are analogues of each other, as are Models 4.3 and 5.2, so these are shown on the same row to ease comparison.

It is unrealistic that the predictions for collection districts would be distributed approximately uniformly between zero and one. As such, the only plausible distributions are those given by Models 4.3, 5.1 and 5.2.

Similarly, looking at the predictions for the informal vote in each collection district, compared side by side in Figure A.4.3, we immediately conclude that the predictions given by Models 4.1 and 4.2 are implausible. There should be absolutely no collection districts that actually record informal vote proportions anywhere near as high as 0.5 (the highest informal voting rate in any polling place is approximately 0.07). Far too many predictions are implausible. The other three models are all plausible, but note that Model 5.1 has a wider spread, with predictions getting close to zero, and a small number above 0.1.

We move now to our predictions for our three case study electoral districts. Starting with Cheltenham (Figure A.4.4), a seat safely held by the ALP, we see a diversity of predictions. Models 4.1 and 4.2 give very similar predictions, as do Models 4.3 and 5.2, so we will consider these two pairs together.

While the predictions for Models 4.1 and 4.2 do not look outlandish, there is a lot of questionable variation in Cheltenham, with a surprising number of Liberal leaning collection districts. Models 4.3 and 5.2 present a much more homogenous view of the electoral district, with a particularly strong area for the ALP in the North-East corner. Model 5.1 presents a similar view of Cheltenham, but the support for each party swings more dramatically in each collection district.

Ultimately, Models 4.3, 5.1 and 5.2 align most closely with our expectations of the voting pattern in Cheltenham, and none of them seem significantly more plausible than the others.

Looking at Norwood (Figure A.4.5), we make the same conclusion as for Cheltenham: Models 4.3, 5.1 and 5.2 appear most sensible. Models 4.1 and 4.2 are implausible, and one reason for this is that the collection district including the suburb of College Park is coloured deep red in the plot.

In Chaffey (Figure A.4.6), we see a different story. This is an incredibly safe rural electoral district for the Liberal Party, but the predictions given by Models 4.3 and 5.2 indicate that the seat is marginal. This is not a plausible prediction. On the other hand, all of Models 4.1, 4.2 and 5.1 give predictions that could be plausible, with strong Liberal support across the whole of Chaffey, easing up a little in the more urbanised collection districts.

5.5.4 Preferred Model

We now draw some conclusions about the five models we have considered.

It is clear from all the measures of model appropriateness that we have chosen that Models 4.1 and 4.2 are poor models. In particular, we expect very few collection districts to uniformly support one party, and so having the π_c^p's for each of the major parties distributed nearly uniformly is highly implausible.

Turning to Models 4.3 and 5.2, while many of the predictions look relatively sensible, there is a glaring problem in Chaffey, where the models predict the ALP actually winning the seat (although this is an abnormal seat as it is not a contest between the ALP and Liberal Party). This is an absurd prediction for a seat in which the ALP performs so poorly. In addition, our measure for the closeness of the models' prediction to the actual results at polling place level shows that Models 4.3 and 5.2 are the furthest from the real election, and by a long way.

Model 5.1, while far from perfect, is the most robust in giving reasonable looking predictions. Its biggest shortcoming is in the prediction for Chaffey, a very safe Liberal seat on two-party preferred terms, even though it still greatly outperforms Models 4.3 and 5.2 on this measure. The supremacy of Model 5.1 is in many ways unsurprising, as it retains all the information contained in the predictors, and the regression is performed on the greatest number of pieces of information we are able to work with, minimising extrapolation and over-fitting.

It is interesting to note that despite the strength of the first two principal components in explaining the variance of the dataset, and the high level of correlation between our predictors, the most useful model we have is a model in which we use the original predictors and *not* the principal components. Hence the correlation does not appear to be a major cause of the problems seen in Chapter 4.

5.6 Discussion

In this chapter we have verified and corrected some of the most glaring errors in the voter location data provided by ECSA. We have developed a model for approximating predictors and predictions at polling place level using the voter location data, and used this to refine the multinomial regression models and use more information.

We have found that one of these refined models, Model 5.1, is significantly better than those in the previous chapter, and we adopt this model as our preferred model and use it as a benchmark for further work in this thesis.

In the next chapter we approach the formulation of the model in a different way, and introduce spatial information in an attempt to further improve the quality of the model.

Chapter 6

Spatially-aware Model

So far we have been considering the ‘absolute’ model

$$\boldsymbol{\eta}_{PP} = \beta_0 + X^{PP}\boldsymbol{\beta},$$

where we then predict the $\boldsymbol{\eta}_{CD}$ by taking the fitted coefficients on this model and calculate

$$\boldsymbol{\eta}_{CD} = \hat{\beta}_0 + X^{CD}\hat{\boldsymbol{\beta}}.$$

This approach ignores the spatial nature of the system entirely, but we know that there is a spatial element to the way voters decide to cast their ballots. Local issues and the perceived quality of individual candidates can affect how people vote, partly explaining why two people with similar demographics may vote in different ways.

For example, as discussed in Section 4.2, there were local issues in Cheltenham that could have influenced voters in that electoral district, and in Chaffey the election was complicated by the strength of the Nationals and weakness of the ALP vote.

In this chapter we consider a different way of working with the predictors, incorporating some of this spatial information into the model. The basic approach is to measure the demographics of each collection district and polling place by comparing it to its neighbours.

At its simplest, this means we compare the income in each electoral district to the state average income and use these differences to predict π_e^p for each $e \in ED$ and $p \in P$. We can then continue to build up a nested system by looking at the polling place and collection district levels. This allows us to retain some of the spatial structure in the data.

This is a more complicated model, and so to simplify implementation we consider this as an ordinary logistic regression model by ignoring the party containing all informal votes. Extending this model to a multinomial regression environment is left to future work.

As a side effect of fitting a logistic regression model, we gain the full toolbox and a more developed understanding of model diagnostics and assumption checking from the literature. We also can more simply interpret the coefficients in the model in a way that is accessible to a wider, non-mathematical audience.

Additionally, both ECSA and the EDBC work with results and boundaries in two-party preferred terms where they do not consider informal votes, and working with only two parties brings our predictions into line with how they report them. For the rest of this thesis, we set $P = \{ALP, LIB\}$.

To enable a like-for-like comparison between modelling techniques, in this chapter we also fit the logistic regression version of our benchmark multinomial model, Model 5.1, by eliminating the informal votes from consideration. This gives us a way to compare between these two model formulations.

6.1 Model Description

We construct three vectors that will become offset terms in the model, as follows:

- $\boldsymbol{\delta}_{ED} = \operatorname{logit}\left(\pi_{SA}^{ALP}\right) \times \mathbf{1}_{47}$, where π_{SA}^{ALP} is the state-wide proportion of voters that preference the ALP over the Liberal party and $\mathbf{1}_{47}$ is the 47 element vector containing only ones;
- $\boldsymbol{\delta}_{PP}$ — a vector of length $|PP|$, where each element b corresponds to $\operatorname{logit}\left(\pi_{ED(b)}^{ALP}\right)$, and $ED(b)$ is the function that maps polling place b to its electoral district. That is, this vector contains an element for each polling place, and that element is the logit of the proportion of voters that preferenced the ALP above the Liberal party in the corresponding electoral district; and
- $\boldsymbol{\delta}_{CD}$ — a vector of length $|CD|$, where each element $c \in CD$ corresponds to

  $$\operatorname{logit}\left(\sum_{b \in PP} \frac{l_{cb}\pi_b^{ALP}}{l_c}\right), \qquad c \in CD,$$

  where l_c is the number of votes cast by people living in collection district $c \in CD$.

  That is, this vector contains an element for each collection district, and that element is a convex combination of the proportion of people that preferenced the ALP in all the polling places in which voters from that collection district voted. This is precisely the calculation that the EDBC use to estimate the π_c^p.

  If we construct the new matrix

  $$\left[L^{PP\to CD}\right]_{cb} = \frac{l_{cb}}{l_c},$$

  then $\boldsymbol{\delta}_{CD} = \operatorname{logit}\left(L^{PP\to CD}\boldsymbol{\pi}_{PP}^{ALP}\right)$.

We now specify the model at each of the three layers.

Electoral District Level

We fit the π_e^{ALP}'s for electoral districts $e \in ED$ using the differences in predictors between electoral districts and the state average.

$$\boldsymbol{\eta}_{ED}^{ALP} - \boldsymbol{\delta}_{ED} = X^{ED-SA}\boldsymbol{\beta}^{ED-SA}$$

where X^{ED-SA} is the matrix containing the difference between the predictors for each electoral district and the mean of the predictors for all electoral districts in the state.

In this model we just need to fit to the $\boldsymbol{\eta}_{ED}^{ALP}$ as for each electoral district e our prediction for π_e^{LIB} is precisely $1 - \pi_e^{ALP}$.

Moving the constant state average term to the right hand side, the model becomes

$$\boldsymbol{\eta}_{ED}^{ALP} = \boldsymbol{\delta}_{ED} + X^{ED-SA}\boldsymbol{\beta}^{ED-SA}, \tag{6.1.1}$$

and so the $\boldsymbol{\delta}_{ED}^{p}$ term acts as an offset term in the model.

We can think of this model as *perturbing* the state average within each electoral district based on its demographics to fit the $\boldsymbol{\eta}_{ED}$.

Polling Place Level

The next layer of the nested model involves predicting at polling place level. The next level in the hierarchy, is

$$\boldsymbol{\eta}_{PP}^{ALP} = \boldsymbol{\delta}_{PP} + X^{PP-ED}\boldsymbol{\beta}^{PP-ED}, \tag{6.1.2}$$

where the matrix X^{PP-ED} contains the differences between the predictors for each polling place and the predictors for that polling place's electoral district.

In this level, the offset term $\boldsymbol{\delta}_{PP}$ acts as a sort of seat specific term. We can think of the model as perturbing the electoral district's ALP vote using the predictors of the polling place relative to other polling places in the seat.

Collection District Level

The third level of the model, predicting at collection district level, would be

$$\boldsymbol{\eta}_{CD}^{ALP} = \boldsymbol{\delta}_{CD} + X^{CD-PP}\boldsymbol{\beta}^{CD-PP}, \tag{6.1.3}$$

where the matrix X^{CD-PP} contains the differences between the predictors for each collection district, and the estimate for the polling place found by projecting the collection district predictors up to the polling places using the voter location data.

Mathematically, the matrix of predictors can be written

$$X^{CD-PP} = X^{CD} - L^{PP\to CD}\left(L^{CD\to PP}\right)^{T} X^{CD}.$$

However, as in the case of the multinomial regression models, we do not know the $\boldsymbol{\eta}_{CD}^{ALP}$ and so cannot implement a model fit.

Instead we consider inference in the same way as before, by assuming that $\beta_{PP-ED} = \beta_{CD-PP}$ and use the estimate found with the polling place model to predict for collection districts. That is, we compute

$$\boldsymbol{\eta}_{CD}^{ALP} = \boldsymbol{\delta}_{CD} + X^{CD-PP}\hat{\boldsymbol{\beta}}^{PP-ED},$$

and use the result to find estimates of the π_{ALP}^{p} for each collection district c.

Note that in the previous models, all of the elements of the matrices X^{CD}, X^{PP}, and X^{ED} are in the interval $[0, 1]$. Under this proposed model, all elements of the matrices X_{ED-SA}, X_{PP-ED}, and X_{CD-PP} are in the interval $[-1, 1]$.

6.1.1 Implementation

Multiple logistic regression is a general linear model and we can perform the model fit using the `glm` function in the `stats` package in R [37]. This function is a generalised version of the linear regression function `lm` which provides for the fitting of many different families of function.

6.2 Model 6.1 - Electoral district level

We now perform the model fit as described in Equation (6.1.1) on the 47 electoral districts as a proof of concept, and call this Model 6.1. No interaction terms are included in the model as there are only 47 observations.

We first examine the modelling assumptions and model fit, starting by considering the fitted values. We save examination of the coefficients to the polling place level model, which has more information.

Figure 6.2.1 shows predicted values for the 47 electoral districts, plotted against their observed values. For illustrative purposes and ease of interpretation, these are expressed as the number of votes for the ALP, rather than in terms of $\pi_e^{ALP}, e \in ED$. The predicted proportion of the ALP vote is converted to a predicted number of votes by multiplying by the total number of votes in that electoral district.

The dotted lines on the plot are positioned 1000 votes either side of the identity line (where the prediction exactly matches the observed votes). 1000 votes is approximately 5% of the size of an average electoral district, and so the lines give a very crude indication of electoral districts which have been poorly fitted. We can see that the fitted values in general increase linearly as the observed ALP vote increases, but with some large residuals and a far from perfect fit.

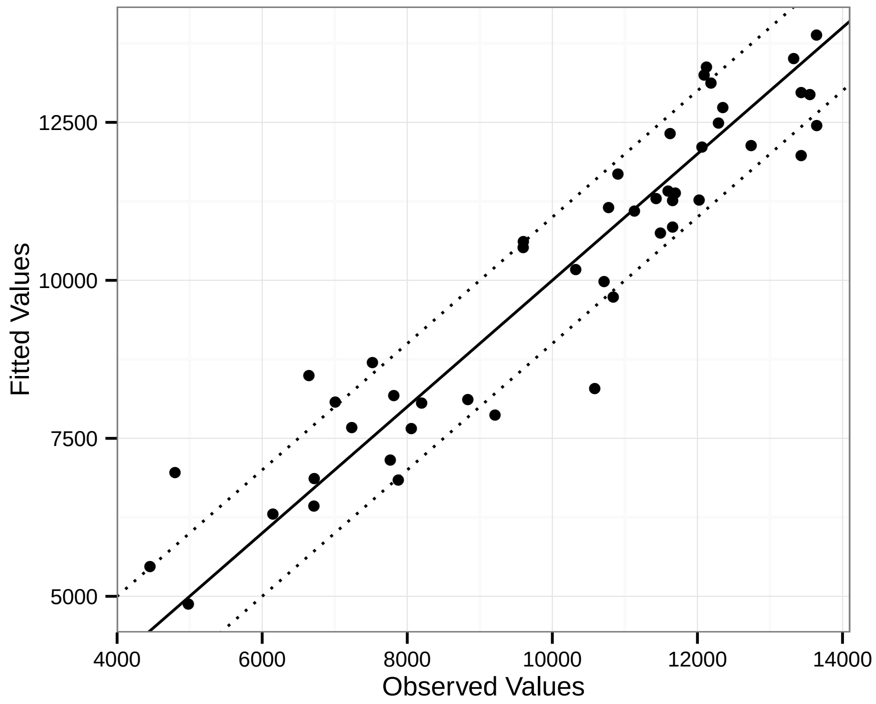


Figure 6.2.1: Predicted values plotted against observed values for the 47 electoral districts fitted under Model 6.1, where both predicted and observed values are converted from proportions to real numbers of votes. The solid line on the plot is the identity line. The two dotted lines show a deviance of 1000 votes from the solid line (this is about 5% of the size of an average electoral district).

We now consider the modelling assumptions, through the analysis of residuals. Consider the plot of deviance residuals against predicted values, shown in Figure 6.2.2, to assess these. The deviance residuals are produced by the `glm` package in R [37].

We examine the plot for residuals that are randomly scattered about zero and homoscedastic to check the assumption that the residuals have mean zero and constant variance.

Given that there are only 47 observations, a small number of outliers can make a large difference to the appearance of the plot, but we can see from Figure 6.2.2 that the residuals are roughly uniformly scattered about the zero line, and, aside from

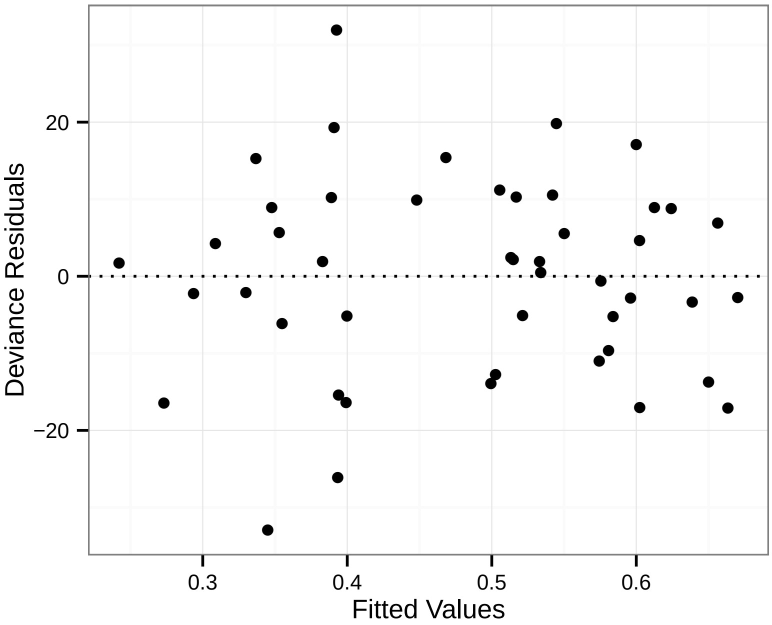


Figure 6.2.2: Plot of the deviance residuals for the 47 electoral districts fitted under Model 6.1.

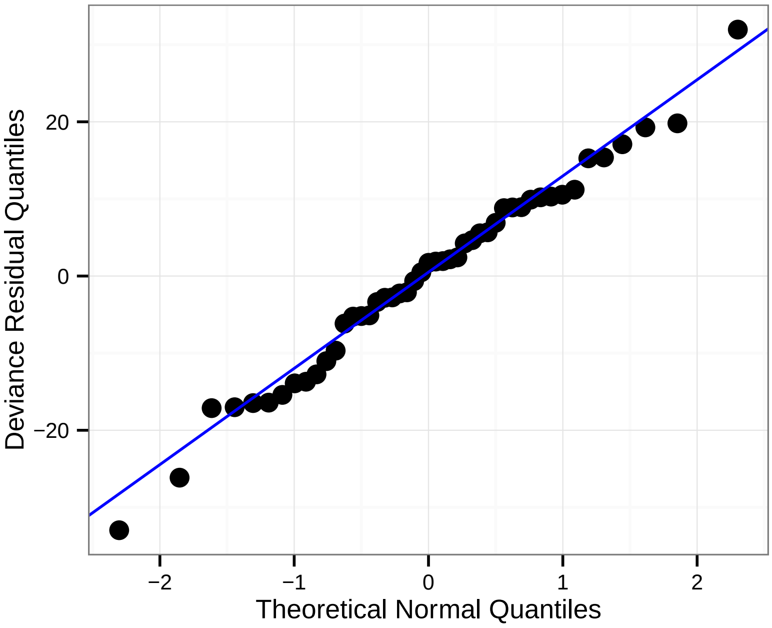


Figure 6.2.3: Normal Q-Q Plot of deviance residuals for the 47 electoral districts fitted under Model 6.1.

a few outliers, are even bounded by ±20. This indicates that the assumptions that the errors in the model have mean zero and constant variance are reasonable.

To check the normality assumption [14, p139] we examine the normal Q-Q Plot of deviance residuals, as shown in Figure 6.2.3. This plot compares the distribution of the residuals to a theoretical normal distribution.

If the residuals were normally distributed we would expect to see the points on the plot lie along the blue identity line in the figure. The points on the figure *do* lie roughly on this line, and so the assumption that the residuals are normally distributed is reasonable.

Other modelling assumptions are more inherent to the system - that we know the values of the predictors without error and that the errors are independent and are not correlated with the predictors.

The values of the predictors are drawn directly from the ABS. For predictors with very small values the ABS introduces random noise to ensure that individuals are not identifiable from the data. At electoral district level there are enough people in each category that this should not be a problem.

The only other error in the data is caused by individuals making errors when they complete their census form or knowingly falsify data. Given that this data is the most comprehensive set of demographic data in Australia, and is routinely used in planning decisions for all sorts of purposes, we can only assume that there exists no dataset with this information that is *more* accurate. We believe it to be a reasonable assumption that the rate of individuals introducing errors to this data is very small.

We therefore assume that we do know the values of the predictors with no error.

The errors themselves are calculated from the fitted values of the 47 electoral districts. There is no reason to believe that these errors would be correlated to each other or the predictors so we regard these as reasonable assumptions to make.

We now consider the goodness of fit of Model 6.1, by considering the deviance of the model as shown in the following R output.

```
Null deviance: 62370  on 47  degrees of freedom
Residual deviance:  7546  on 30  degrees of freedom
```

The deviance under this model is much lower than the deviance under the 'Null Model' (that in which we fit using only one intercept coefficient). However the residual deviance is still very large and we have a p-value of $P(X > 7546) \approx 0$, where $X \sim \chi^2_{30}$.

This implies that the model does not explain all of the variance in the system, but it does have some predictive power of voting intention.

We conclude that an electoral district's demographics for household income, education level and language spoken at home do not give a perfect indication of that electoral district's voting inclinations for the 2010 SA state election, but they are a useful and significant guide.

Having established this new spatially-aware model at an electoral district level, we now shift focus to consider the relative model at the polling place level, rather than exploring and interpreting the regression coefficients.

6.3 Model 6.2 - Polling place level

We now perform the polling place level model fit as described in Equation (6.1.2), and hereafter refer to this model as Model 6.2.

In this model we fit the votes in the 746 polling places using the relative differences in the proportions of predictors between the polling places and the electoral district they are in. In other words, our predictors now tell us how different each polling place is from its electoral district. The offset term is the log-odds of the ALP vote in the relevant electoral district.

Introducing interaction terms in the model is worthwhile future work, but is outside the scope of this thesis.

As we did in Chapter 4, we also attempt to simplify the model using stepwise model selection by AIC. This is again implemented using the `stepAIC` function, part of the `MASS` package in R [39]. The coefficients for the full model with all 17 predictors, along with corresponding standard errors and p-values, are shown in Section B.1 of Appendix B.

After performing model selection we are left with the predictors and coefficients shown in Table 6.3.1.

We return to the interpretation of coefficients shortly, but we again start by considering the modelling assumptions. As in the case of the previous model, we look at the plot of fitted values against the observed votes in the polling places, shown in Figure 6.3.1.

Dotted lines are positioned on the plot at a distance from the identity line approximately 5% of the size of a polling place (68 votes) to get an idea of the closeness of the fit. Again we see that the fitted values in general increase linearly as the observed ALP vote increases, but there appear to be a larger proportion of residuals greater than this 5% error.

Assumptions about the mean and variance of the residuals are tested using the plot of deviance residuals, shown in Figure 6.3.2. In this figure each of the polling places is coloured according to whether it contains ordinary or declaration votes. Again, we hope for homoscedasticity and roughly random scatter about zero.

Table 6.3.1: Model output for Model 6.2, after stepwise model selection by AIC is performed to reduce the number of predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
NegativeOrNilIncome	-9.4995	0.5290	-17.96	0.0000 ***
X1.499	2.9326	0.1060	27.68	0.0000 ***
X500.999	1.2931	0.1387	9.33	0.0000 ***
X1000.1399	-2.3605	0.1806	-13.07	0.0000 ***
X1400.1999	3.3302	0.1821	18.28	0.0000 ***
X2500.2999	-1.0105	0.3418	-2.96	0.0031 **
X3000.3499	-2.3768	0.5751	-4.13	0.0000 ***
MoreThan4000	-8.0823	0.5383	-15.01	0.0000 ***
DoesNotSpeakEnglishAtHome	1.6920	0.0590	28.67	0.0000 ***
LessThanYear12	-4.4600	0.1432	-31.15	0.0000 ***
Year12OrEquivalent	-2.0922	0.2367	-8.84	0.0000 ***
BDAD	-4.9003	0.3025	-16.20	0.0000 ***
PDGD	11.1814	0.5600	19.97	0.0000 ***
CL	5.2615	0.2546	20.67	0.0000 ***
Significance codes:	· p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

We can see that there is a difference in variance between the polling places that contain ordinary votes, and those that contain declaration votes. This could be a false impression of lower variance because there are many fewer polling places with declaration votes. It could also be an indication that people casting declaration votes vote differently, which would be a very interesting result. This theory is supported anecdotally by evidence from ECSA. If this were true it could have significant impacts on future elections, as we know that the proportion of people that cast declaration votes is growing in jurisdictions across Australia.

For future work it would be worth considering a new factor in the model, detailing the type of votes the polling place contains to explore this theory.

We test the assumption that the residuals are normally distributed by considering

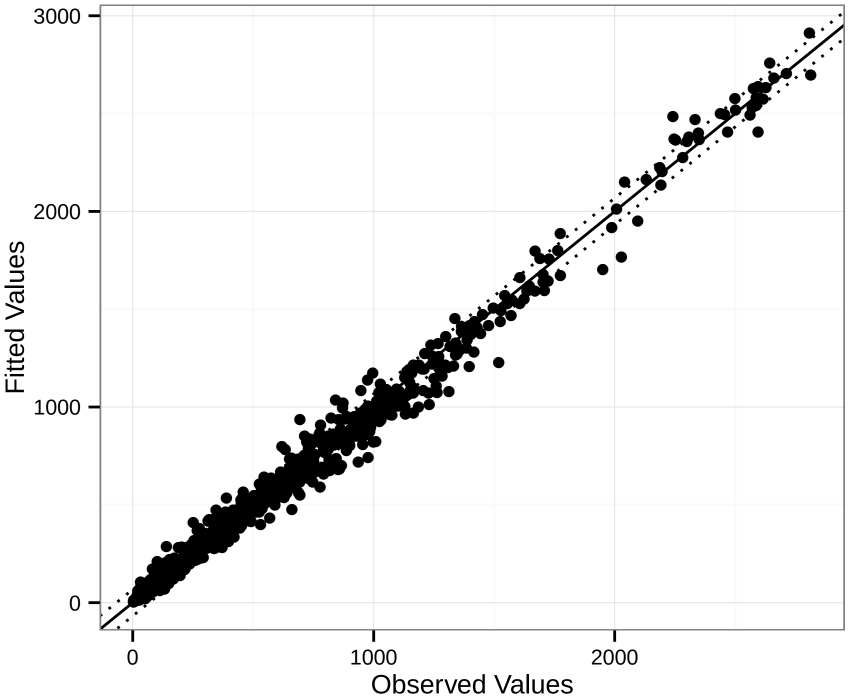


Figure 6.3.1: Predicted values plotted against observed values for the 746 polling places fitted under Model 6.2, where both predicted and observed values are converted from proportions to real numbers of votes. The solid line on the plot is the identity line. The two dotted lines show a deviance of 68 votes from the solid line (this is about 5% of the size of an average polling place).

the normal Q-Q plot, shown in Figure 6.3.3.

We see that most of the plot shows linearity, but it veers away from normality at both tails. We conclude that the residuals are roughly normally distributed, but we may have some problems trying to predict at the extremes.

Again we consider the model fit by analysing the residual deviance as shown in the R output below.

```
Null deviance: 17809.5  on 746  degrees of freedom
Residual deviance:  9620.4  on 732  degrees of freedom
```

Again we see a lower deviance under Model 6.2 than the ‘Null Model’, but the

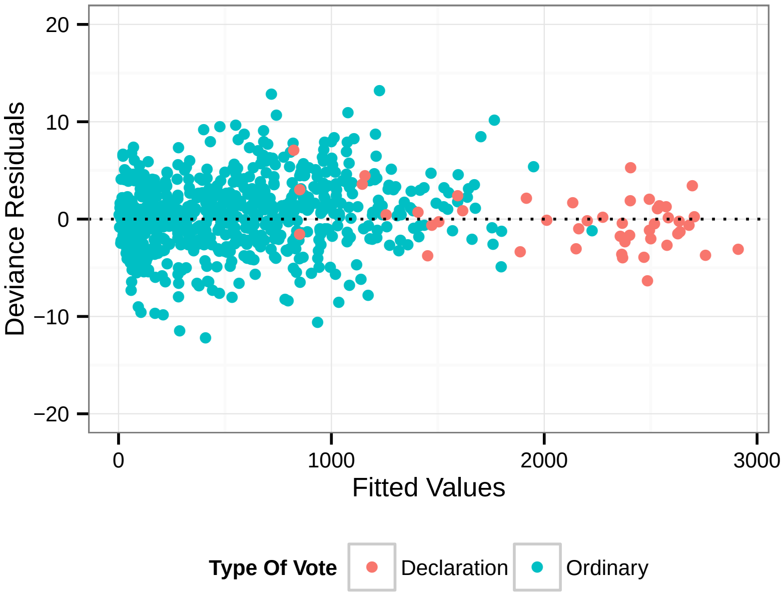


Figure 6.3.2: Plot of the deviance residuals for the 746 polling places fitted under Model 6.2.

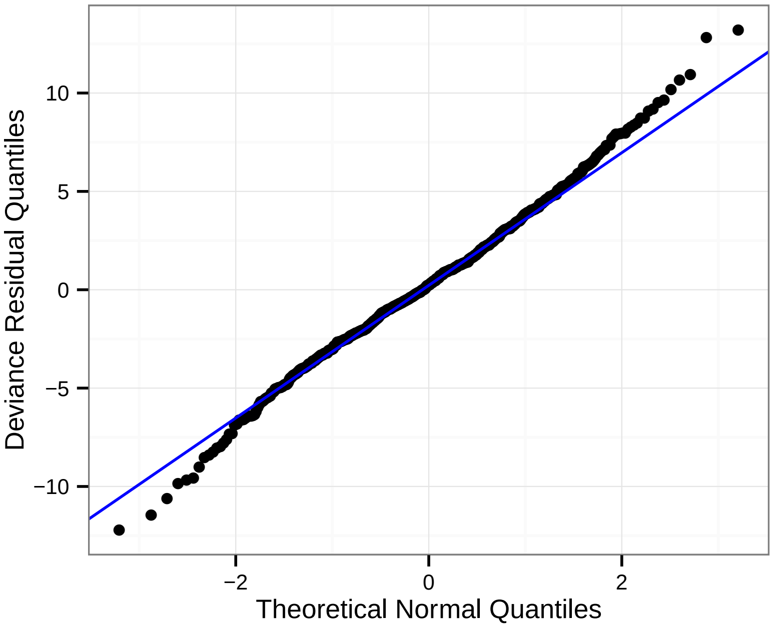


Figure 6.3.3: Normal Q-Q Plot of deviance residuals for the polling places, fitted under Model 6.2.

residual deviance is large and we have a still p-value of $P(X > 9620.4) \approx 0$, where $X \sim \chi^2_{732}$.

As for Model 6.1 we conclude that the predictors have some predictive power but they do not tell the whole story of voting intention.

6.4 Model 6.3 - logistic analogue of Model 5.1

To give us a basis for comparison between Model 6.2 and the multinomial regression models we fit

$$\boldsymbol{\eta}^p_{PP} = X^{PP}\boldsymbol{\beta}^{PP}, \tag{6.4.1}$$

with the parties $P = \{ALP, LIB\}$. This is the logistic regression version of our multinomial regression models. We fit the analogue of Model 5.1, using all 17 predictors at polling place level and call this Model 6.3. There is no offset term in this model as Model 5.1 was fit using only absolute predictors.

We will compare Model 6.3 to Model 6.2 to test the strengths of this new approach to prediction.

After performing stepwise model selection by AIC we have the predictors and coefficients contained in the output in Table 6.4.1. The coefficients for the full set of predictors, prior to model selection being performed, is contained in Section B.2 of Appendix B.

Note that the coefficients in this model are not directly comparable to the coefficients in Model 6.2 as the structures of the predictors are completely different.

Just as before, we first consider the modelling assumptions. Figure 6.4.1 shows the fitted values for the model plotted against the observed votes in the polling places. The dotted lines remain at a distance 68 votes from the identity line (approximately

Table 6.4.1: Model output for Model 6.2, after stepwise model selection by AIC is performed to reduce the number of predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.2470	0.1430	15.71	0.0000 ***
NegativeOrNilIncome	-14.6515	0.4903	-29.88	0.0000 ***
X1.499	4.0064	0.0840	47.69	0.0000 ***
X500.999	4.4170	0.1192	37.05	0.0000 ***
X1000.1399	-5.8250	0.1543	-37.74	0.0000 ***
X1400.1999	6.0090	0.1819	33.04	0.0000 ***
X2000.2499	3.9979	0.2835	14.10	0.0000 ***
X3000.3499	1.8958	0.5138	3.69	0.0002 ***
MoreThan4000	-7.8051	0.4971	-15.70	0.0000 ***
DoesNotSpeakEnglishAtHome	-1.4280	0.1922	-7.43	0.0000 ***
SpeaksEnglishAtHome	-3.3004	0.1855	-17.79	0.0000 ***
LessThanYear12	-4.0199	0.1347	-29.83	0.0000 ***
Year12OrEquivalent	3.7121	0.1778	20.87	0.0000 ***
BDAD	-14.4366	0.2351	-61.41	0.0000 ***
PDGD	14.3836	0.4841	29.71	0.0000 ***
CL	8.3223	0.1864	44.65	0.0000 ***
Significance codes:	· p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

5% error) and this allows us to compare the fitted values to those of Model 6.2.

The predicted values do increase roughly linearly with the size of the polling place. However, comparing Figures 6.3.1 and 6.4.1 shows that the fit for Model 6.3 is worse, and the predictions in this model lie much less tightly around the identity line than in Model 6.2.

We check the assumptions about the distribution of the residuals with the plot of deviance residuals against predicted values (Figure 6.4.2).

Again we hope for the plot to show homoscedastic residuals that are randomly scattered about zero to indicate that the residuals have mean zero and constant

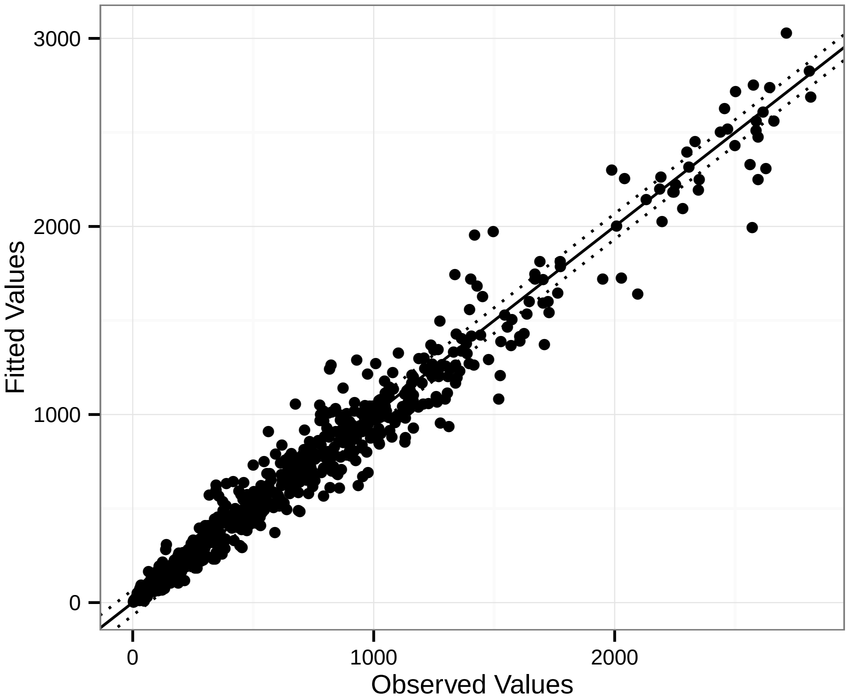


Figure 6.4.1: Predicted values plotted against observed values for the 746 polling places fitted under Model 6.3, where both predicted and observed values are converted from proportions to real numbers of votes. The solid line on the plot is the identity line. The two dotted lines show a deviance of 68 votes from the solid line (this is about 5% of the size of an average polling place).

variance. There are a small number of outlier predictions but for the most part these assumptions are met. Most of the predictions are contained within ±10, but by comparing this plot to Figure 6.3.2, the equivalent plot for Model 6.2, we again see that the residuals have a wider spread in this absolute model.

The normality assumption is assessed with the aid of the normal Q-Q Plot of deviance residuals (Figure 6.4.3) which compares the distribution of the residuals to a theoretical normal distribution.

Here we see another deterioration in the quality of the modelling assumptions. There are some clear deviations from normality on either end of the plot, and these are significantly worse than the smaller deviations observed for the relative model fit

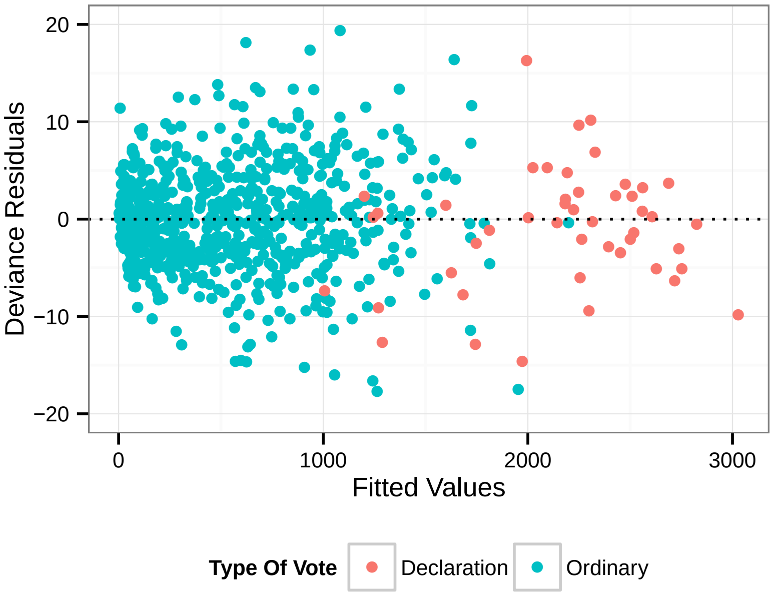


Figure 6.4.2: Plot of the deviance residuals for the 746 polling places fitted under Model 6.3.

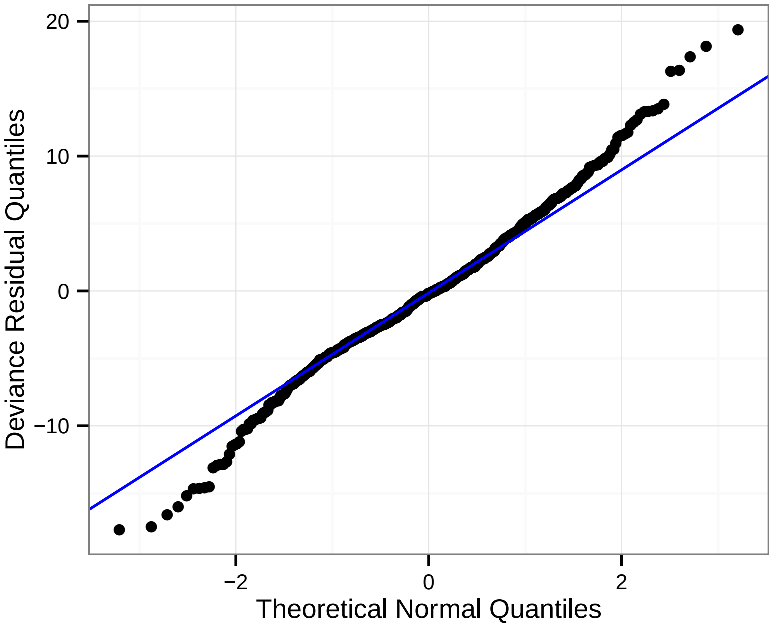


Figure 6.4.3: Normal Q-Q Plot of deviance residuals for the polling places, fitted under Model 6.3.

(Figure 6.3.3).

This modelling assumption is clearly being stretched under Model 6.3. This may mean that the assumptions were violated under the multinomial regression models.

For the other modelling assumptions (that we know the values of the predictors without error and that the errors are independent and are not correlated with the predictors), the same comments from the previous model apply equally to this model and we do not believe they have been violated.

The deviance of the model is shown in the R output below.

```
Null deviance: 78055  on 745  degrees of freedom
Residual deviance: 20909  on 730  degrees of freedom
```

The residual deviance under Model 6.3 is more than twice that for Model 6.2. Once again the p-value is very low: $P(X > 20909) \approx 0$, where $X \sim \chi^2_{730}$.

Although this model is not entirely useless, we conclude that Model 6.2 provides a better fit, and the modelling assumptions are being stretched under Model 6.3.

6.5 Model Predictions

We again consider how close the predictions under these two models are to the actual results of the 2010 election. Table 6.5.1 shows the predicted proportion of people that vote for the ALP under both models for each of our test electoral districts.

Note that the equivalent table considered for the multinomial regression models (Table 5.5.1) has different proportions for the ALP under the column showing actual results. This is because in this chapter we are modelling only two parties and have removed the informal votes. The actual results are calculated by dividing the number of votes the ALP received in each jurisdiction by the number of *formal* votes cast (as opposed to before where all votes cast were included).

To calculate the predictions under the models we first calculate an estimate from the election results for the number of informal votes in each collection district, by distributing the informal votes in each polling place to collection districts using the voter location data. This estimate is subtracted from the number of votes cast in each collection district, and thus the proportion of votes that the ALP received in each of our jurisdictions of interest can be estimated.

Table 6.5.1: Predictions $\hat{\pi}_e^p$ for the electoral districts Cheltenham, Chaffey, Norwood, and also the whole of South Australia, under Models 6.2 and 6.3, compared to the true π_e^p from the 2010 election results (calculated after excluding all informal votes). These summary statistics are calculated by weighting the predictions $\hat{\pi}_c^p$ according to the number of votes in each collection district c.

District	2010 ALP Result	Model	
		Model 6.2	Model 6.3
South Australia	0.484	0.486	0.488
Cheltenham	0.661	0.658	0.654
Chaffey	0.222	0.227	0.349
Norwood	0.452	0.453	0.472

We can see that Model 6.2 performs better than Model 6.3 in all cases. For the most part it gives a small improvement, but in the case of Chaffey, Model 6.2 gives an estimate that is much closer to the actual results than Model 6.3.

Recall that Chaffey was a seat that all of the multinomial regression models struggled to predict. One possible reason for this is that there were significant local issues at play affecting voters decisions, including the fact that the seat was held until 2010 by The Nationals, and the ALP traditionally have a very poor result in the seat.

We now return to the tools for model validation developed previously for the multi-

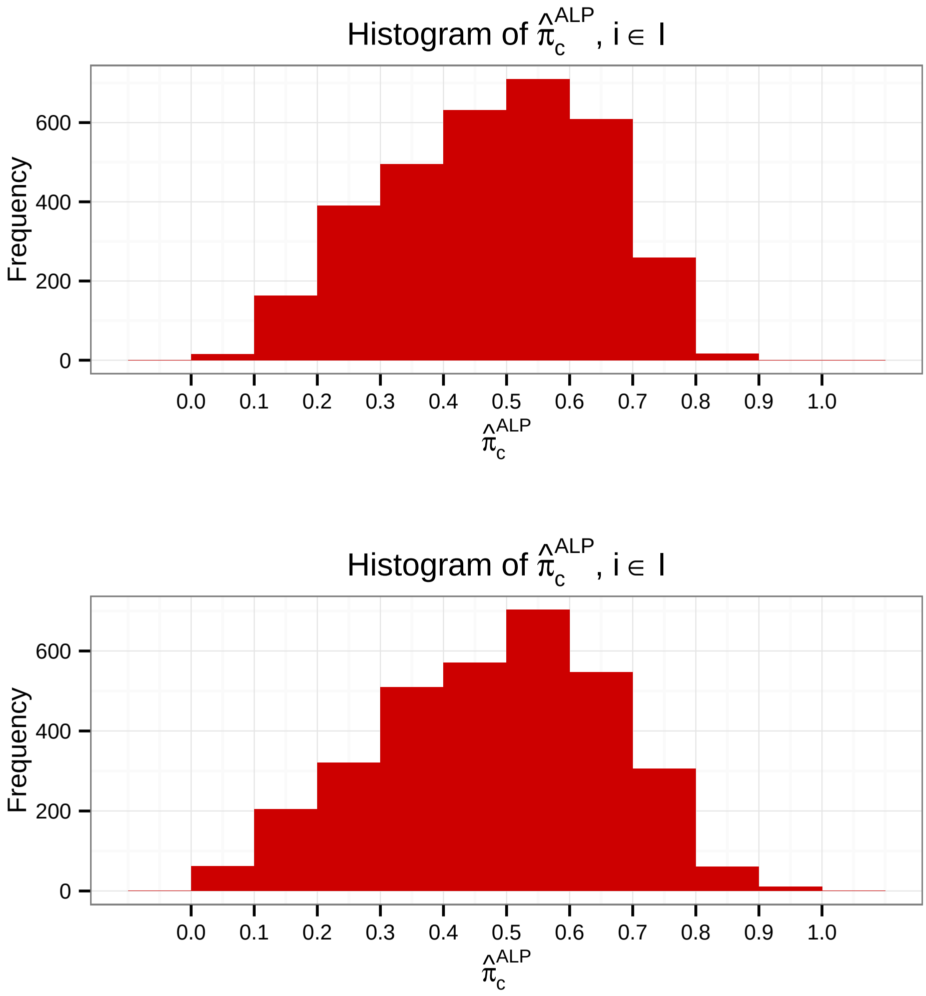


Figure 6.5.1: Distribution of predicted values for $\hat{\pi}_c^{ALP}$, for $c \in CD$, under Model 6.2 (top) and Model 6.3 (bottom).

nomial regression models. Figure 6.5.1 shows the distributions of the predicted values for $\pi_c^{ALP}, c \in CD$, for Models 6.2 and 6.3 respectively.

The figures show that the two distributions have very similar shapes and locations. The biggest difference is a smaller spread of predictions under Model 6.2. We still expect a very small number of predictions at either extreme. Because of this Model 6.2 gives us a more believable set of predictions.

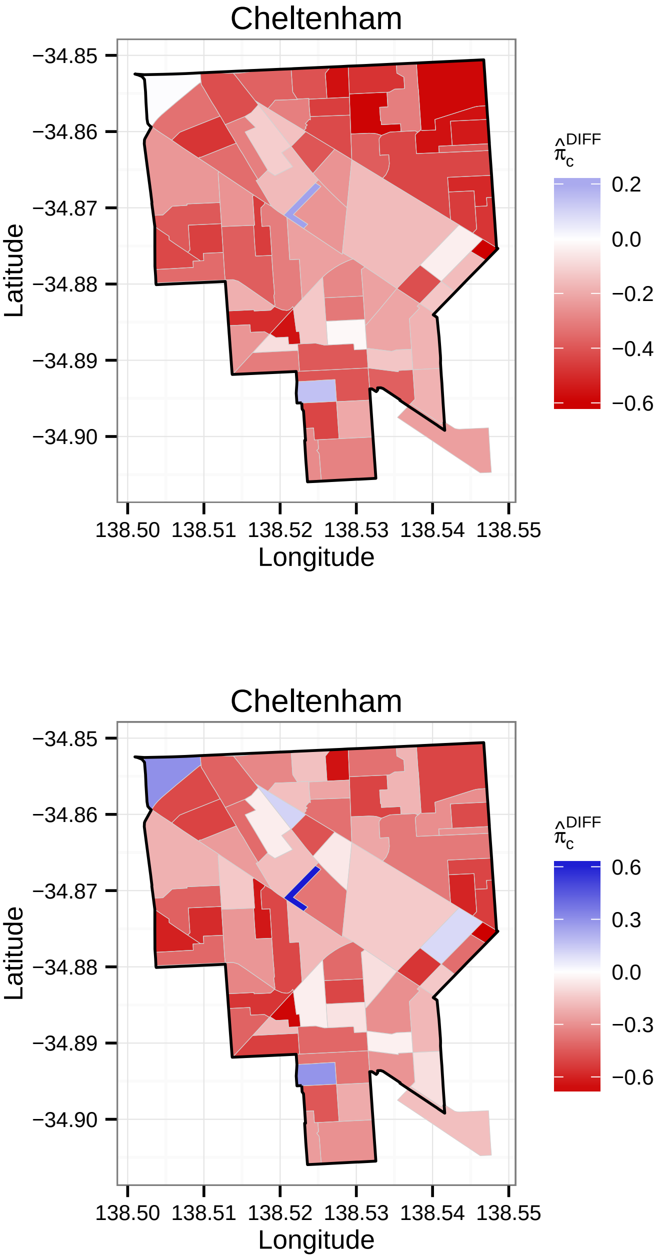


Figure 6.5.2: The electoral district of Cheltenham, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top) and Model 6.3 (bottom).

Figure 6.5.2 shows the predictions for Cheltenham under Models 6.2 and 6.3. We see similar patterns, with less extremities under Model 6.2. Again we see the very deep blue prediction under Model 6.3 that we saw under Model 5.1. The predictions for this collection district are $\pi^{ALP}_{4101004}$ are 0.191 under Model 6.3 and 0.416 under Model 6.2. We regard the prediction given by Model 6.2 as more plausible for this collection district.

Across the whole of the electoral district we regard Model 6.2 as having given more sensible predictions, as the strength of ALP support more closely follows the indicative distribution shown by polling places in Figure 4.2.1 under this model.

Figure 6.5.3 shows the predictions for π_c^{DIFF} for collection districts in Norwood, under Models 6.2 and 6.3 respectively. There is very little obvious difference between the predictions for Norwood under each model, but Model 6.2 is again less extreme. Both sets of predictions appear plausible.

Finally, Figure 6.5.4 contains the predictions for collection districts in Chaffey. Here we see a dramatic difference between the two models. Where Model 6.3 gives us wildly implausible predictions similar to those of Model 5.1, Model 6.3 predicts that almost all collection districts record very strong Liberal Party support, as we expect, and as reflected in the polling place results in Figure 4.2.3.

6.5.1 Goodness of fit

We again calculate an overall measure E of the closeness between our predicted elections and the actual 2010 results. This time we only need to consider how close the prediction for the ALP in each polling place is, as the prediction for the Liberal Party will differ from the total number of votes by precisely $\hat{Y}_b^{ALP} + M_b$, where M_b is the mismatch in the voter location data in polling place b.

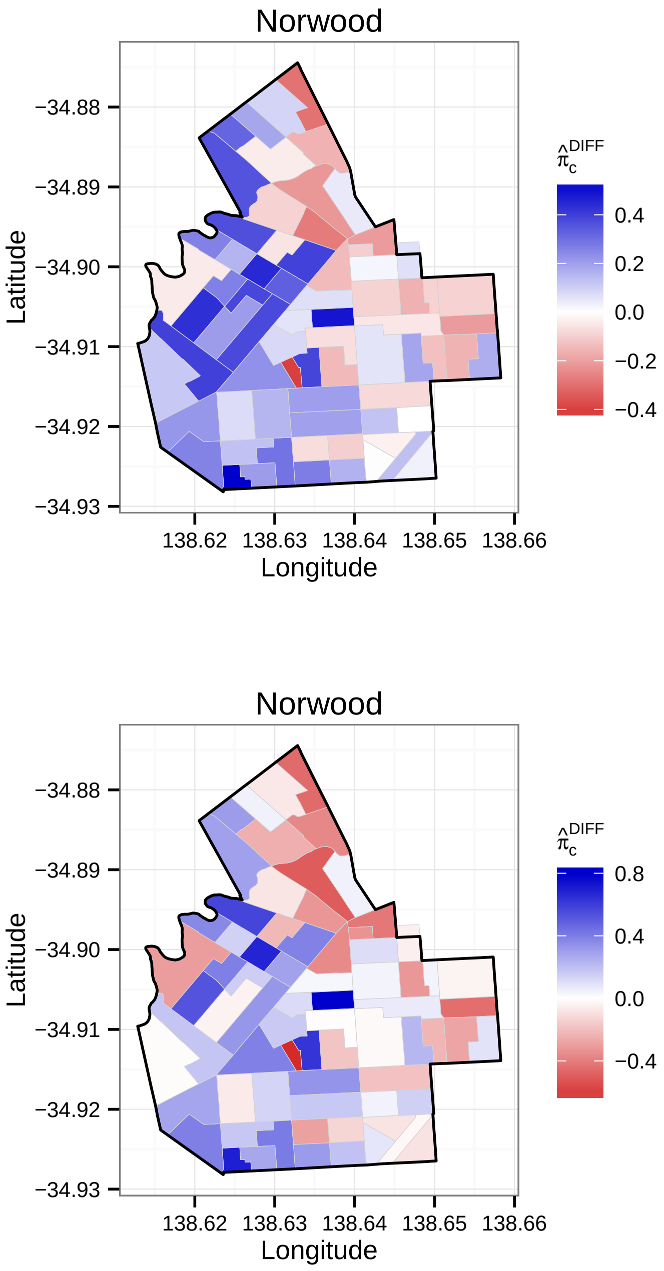
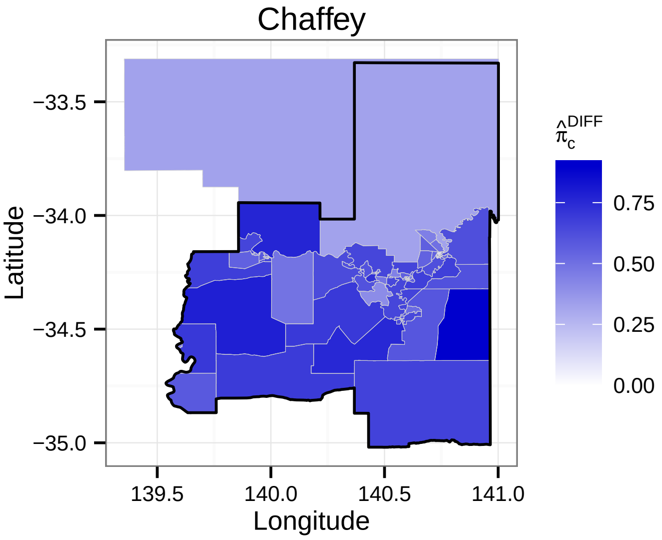


Figure 6.5.3: The electoral district of Norwood, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top) and Model 6.3 (bottom).



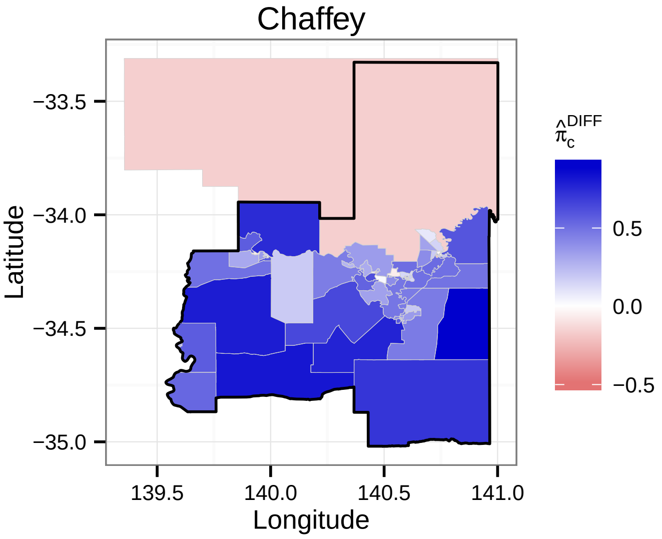


Figure 6.5.4: The electoral district of Chaffey, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top) and Model 6.3 (bottom).

So we can measure the closeness between our predicted model and the observed election as

$$E = \sum_{b \in PP} \left(Y_b^{ALP} - \hat{Y}_b^{ALP} \right)^2 . \tag{6.5.1}$$

Again we prefer models for which E is smaller. It should be noted that this error term cannot be compared to the errors for the multinomial regression models. The values for E for each of the models is shown in Table 6.5.2.

The mean of $\left| Y_b^{ALP} - \hat{Y}_b^{ALP} \right|$ for $b \in PP$ is 30.61 under Model 6.2, and 73.03 under Model 6.3 .

Table 6.5.2: The overall error E for each of Models 6.2 and 6.3, with E calculated as per Equation (6.5.1).

Model	E
Model 6.2	1,509,046
Model 6.3	9,210,688

It can immediately be seen that Model 6.2 gives us a set of predictions that are overall much closer to the 2010 results than Model 6.3.

6.6 Preferred Model

With the development of spatially aware models in this chapter we have seen an incremental improvement in the quality of predictions produced by our model, and the goodness of the model fit overall.

Because Model 6.3 is a direct analogue of our preferred multinomial regression model in the single response logistic environment, we conclude that Model 6.2 gives another incremental improvement over the multinomial regression models, and this is now our preferred model.

The improvements in Model 6.2 appear to be more pronounced for areas that previous models struggled to fit, and make less of an impact on the areas that were well predicted, which are largely the marginal areas.

6.7 Coefficients

The coefficients of the logistic regression models are difficult to interpret directly. The transformation applied to the data means we cannot think about changes in the predictors altering the predictions linearly. The structure of the predictors, and the fact that each predictor category must sum to less than one (making the predictors partially dependent on each other) also make it difficult to interpret the coefficients.

The usefulness in the model is largely for prediction for collection districts, but nonetheless we now consider the coefficients of our preferred model, Model 6.2, to see what they tell us about voting intention in broad terms. The full model output has previously been included in Table 6.3.1.

There are large negative coefficients at either extreme of the Household Income predictor category, indicating that collection districts with extremely high household incomes and those with negative or nil income are relatively more likely to prefer the Liberal Party over the ALP. It should be noted, however, that the negative or nil income category is very small. In more than 2000 collection districts this predictor is empty, so we urge caution in the interpretation of this coefficient.

The effect of the other predictors is less clear as the direction of influence changes repeatedly within the category, and one of the predictors in the centre of the category ($2000-2499) was excluded following model selection. This suggests that the behaviour of groups of people is more dependent on the numbers of people at each extreme than it is on variations nearer the average.

One of the two predictors in the Language Spoken At Home category was excluded after model selection, and there remains a positive coefficient for the proportion of people that do not speak English at home. This indicates that collection districts with higher proportions of non-English speaking people have a higher proportion of ALP voters.

Both of the coefficients in the School Education category are negative, but the coefficient for the proportion of people that have not completed Year 12 is less than the coefficient for the proportion of people that have. This indicates that if a collection district increases the proportion of people that have not completed Year 12 (at the expense of the proportion that have), then more people are likely to vote for the Liberal Party.

Finally, the Non-School Education category shows a similar pattern, though interpretation is trickier because there are many people that have not completed any tertiary education and a person could be in the 'Other' predictor for many different reasons. Nonetheless we see that of the two predictors for university qualifications, voters holding lower qualifications (such as Bachelor Degrees) are more likely to vote for the Liberal Party than those holding higher qualifications (such as Postgraduate Degrees). Holding a Certificate Level qualification and no qualification that fits into the other predictors makes you more likely to preference the ALP.

6.8 Discussion

In this chapter we attempted to approach the model in a different way, and incorporate some of the spatial information that we have available. Model 6.2 has shown another incremental improvement on previous models and we are able to predict the vote for each of the ALP and the Liberal Party in each polling place to within an average of around 30 votes.

In the next chapter we compare the predictions for collection districts under Model 6.2 to the predictions that were used by the EDBC in the redistribution following the 2010 state election, and further investigate the credibility and usefulness of Model 6.2.

Chapter 7

Comparing to EDBC predictions

In the previous three chapters we developed a series of models to predict π_c^p for the collection districts c and parties p. Of the models that were investigated, our preferred is Model 6.2.

In this chapter we compare the predictions under this preferred model to the estimates that were used by the EDBC in their work following the 2010 election.

Recall that the EDBC estimates are $L^{PP\to CD}\boldsymbol{\pi}_{PP}^{ALP}$, as discussed in Section 6.1. In calculating these estimates, the EDBC assume that polling places are homogeneous. That is, every voter who visits an individual polling place has the same probability of voting for each party as anyone else that visits that polling place.

Since there is no statistical model underlying these predictions, and the method relies on the results at all 746 polling places to calculate estimates, this model is saturated. By saturated, we mean that the EDBC takes the 746 polling place results as individual polling place parameters. These are then deterministically transformed using some linear algebra to calculate estimates for the collection districts. There is also no noise structure in the predictions, and this makes it difficult to reliably use the estimates for forecasting.

Model 6.2 uses the EDBC estimate as a base, but then layers on top the demographic predictors. The predictors were chosen after reviewing the literature and expert opinion in the field [26], and we make no assumptions about the predictors prior to performing the model fit.

Statistically speaking, Model 6.2 is a better model than the EDBC model. It has many less parameters and is hence more robust and allows prediction where the EDBC method does not.

We now investigate briefly whether Model 6.2 could be more useful, and whether the differences between the two models are important.

We define the vectors $\pi^p_{6.2}$ and π^p_{EDBC} to be of length $|CD|$ and contain π^p_c for $c \in CD, p \in P$ under Model 6.2 and the existing EDBC methodology respectively. Then $\pi^{DIFF}_{6.2} = \pi^{LIB}_{6.2} - \pi^{ALP}_{6.2}$ and $\pi^{DIFF}_{EDBC} = \pi^{LIB}_{EDBC} - \pi^{ALP}_{EDBC}$.

Given the EDBC methodology involves shifting collection districts between electoral districts to meet the requirements of the fairness clause, we are most interested in knowing how different the two predictions are in votes (rather than proportions), and what effect moving a collection district would have on the electoral districts margin. For this reason we calculate

$$\lambda_c = N_c \left[\pi^{DIFF}_{6.2} - \pi^{DIFF}_{EDBC}\right]_c, \tag{7.0.1}$$

where N_c is the number of votes in collection district c, according to the voter location data. It can easily be shown that $\lambda_c = 2N_c \left[\pi^{LIB}_{6.2} - \pi^{LIB}_{EDBC}\right]_c$.

So λ_c compares the margins between the Liberal Party and the ALP under each of the two models. It is the number of votes by which this margin would change (in favour of the Liberal Party) in the electoral district containing c if the prediction from Model 6.2 had been used instead.

In other words, in collection district c under Model 6.2 we predict that the Liberal

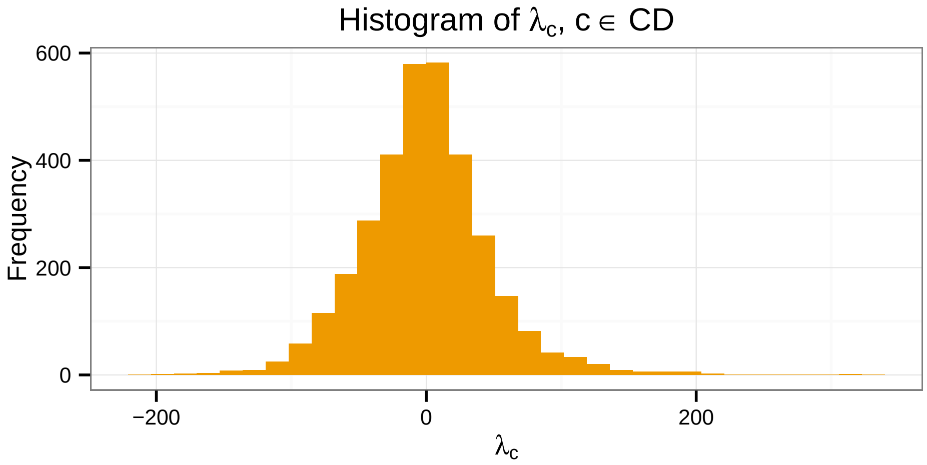


Figure 7.0.1: Histogram of λ_c for $c \in CD$, where λ_c is as defined in Equation (7.0.1).

Party will receive $\frac{1}{2}\lambda_c$ more votes than was predicted using the EDBC methodology, and the ALP will receive $\frac{1}{2}\lambda_c$ less votes than was predicted using the EDBC methodology.

We first consider the distribution of λ_c for all $c \in CD$. A histogram of the distribution is shown in Figure 7.0.1.

The figure shows a unimodal distribution centred on zero. For the most part the predictions between the models are similar ($|\lambda_c|$ is less than 50 in 78.2% of collection districts, and less than 20 in 41.4% of them), but there is a very large spread. There are a small number of collection districts in which $|\lambda_c|$ is more than 200, meaning that the predictions for Y_c^p differ by over 100 votes.

Figures 7.0.2 - 7.0.4 show the predictions for π_c^{DIFF} under both Model 6.2 and the EDBC method, and the values for λ_c for each collection district c for each of Cheltenham, Norwood, and Chaffey respectively.

In all cases, the figures show much less variation in the EDBC predictions than those produced by Model 6.2. This is no surprise as Model 6.2 works by perturbing the

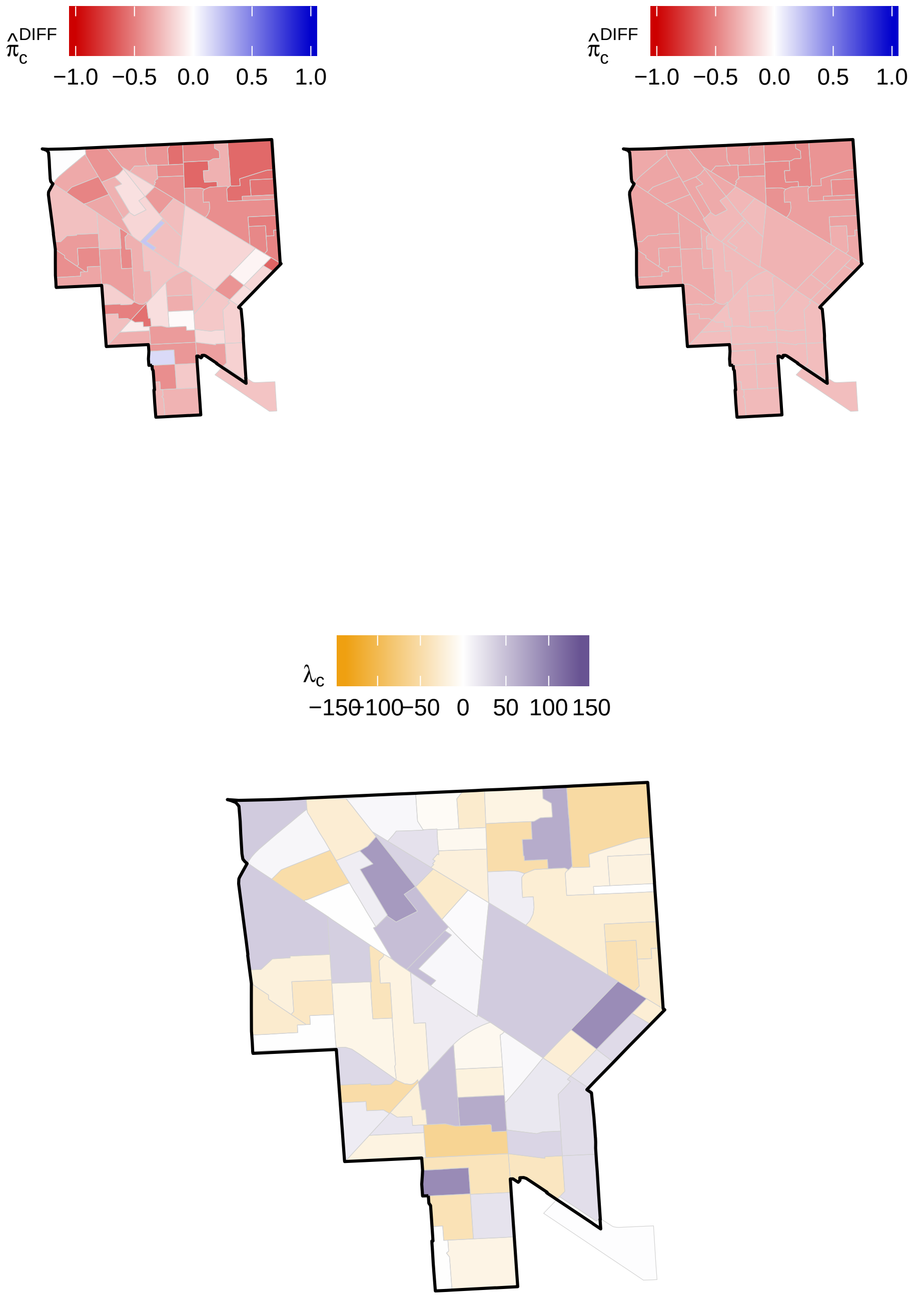


Figure 7.0.2: Cheltenham, with collection districts c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top left), $\hat{\pi}_c^{DIFF}$ from EDBC method (top right), and λ_c (bottom).

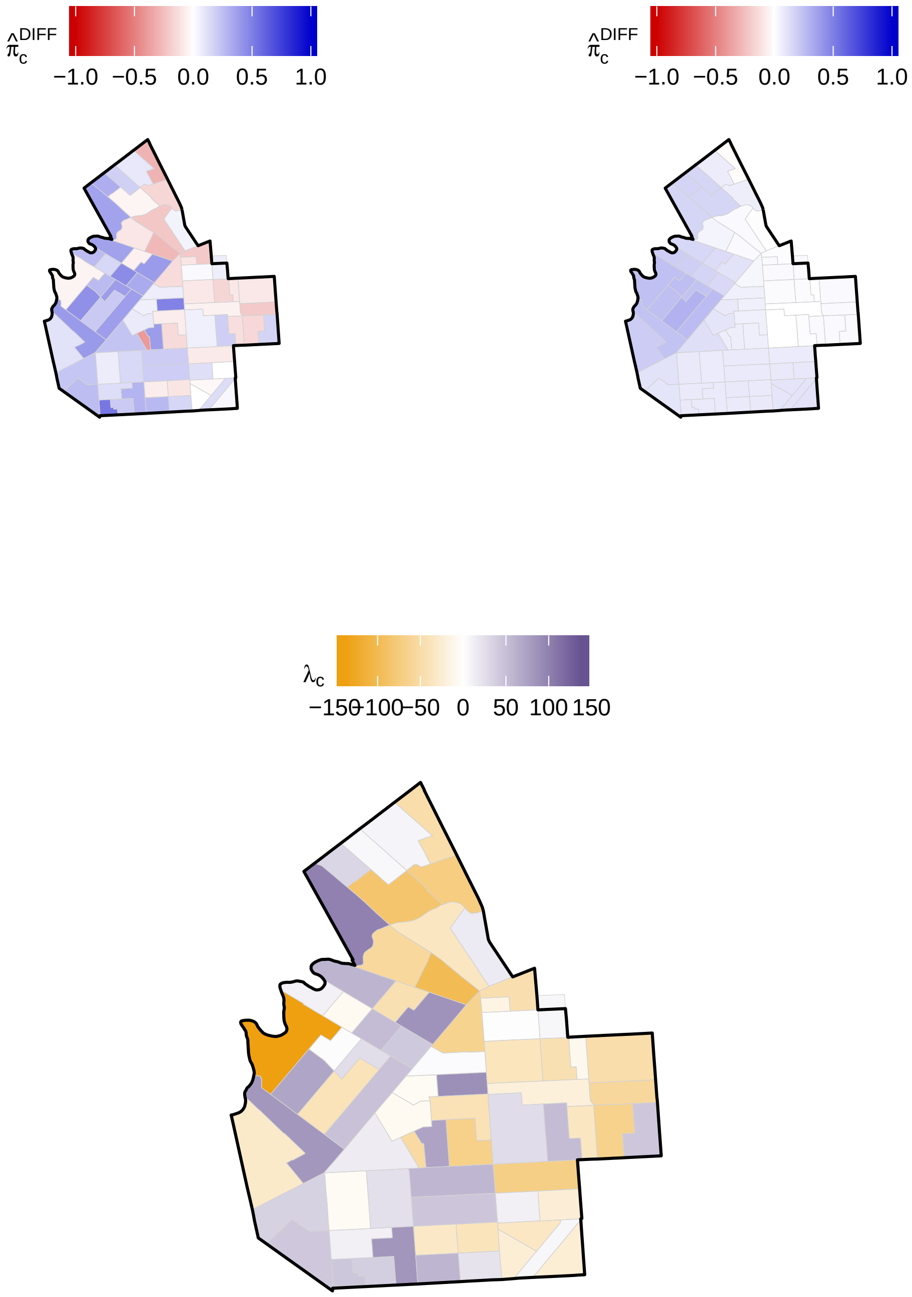


Figure 7.0.3: Norwood, with collection districts c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top left), $\hat{\pi}_c^{DIFF}$ from EDBC method (top right), and λ_c (bottom).

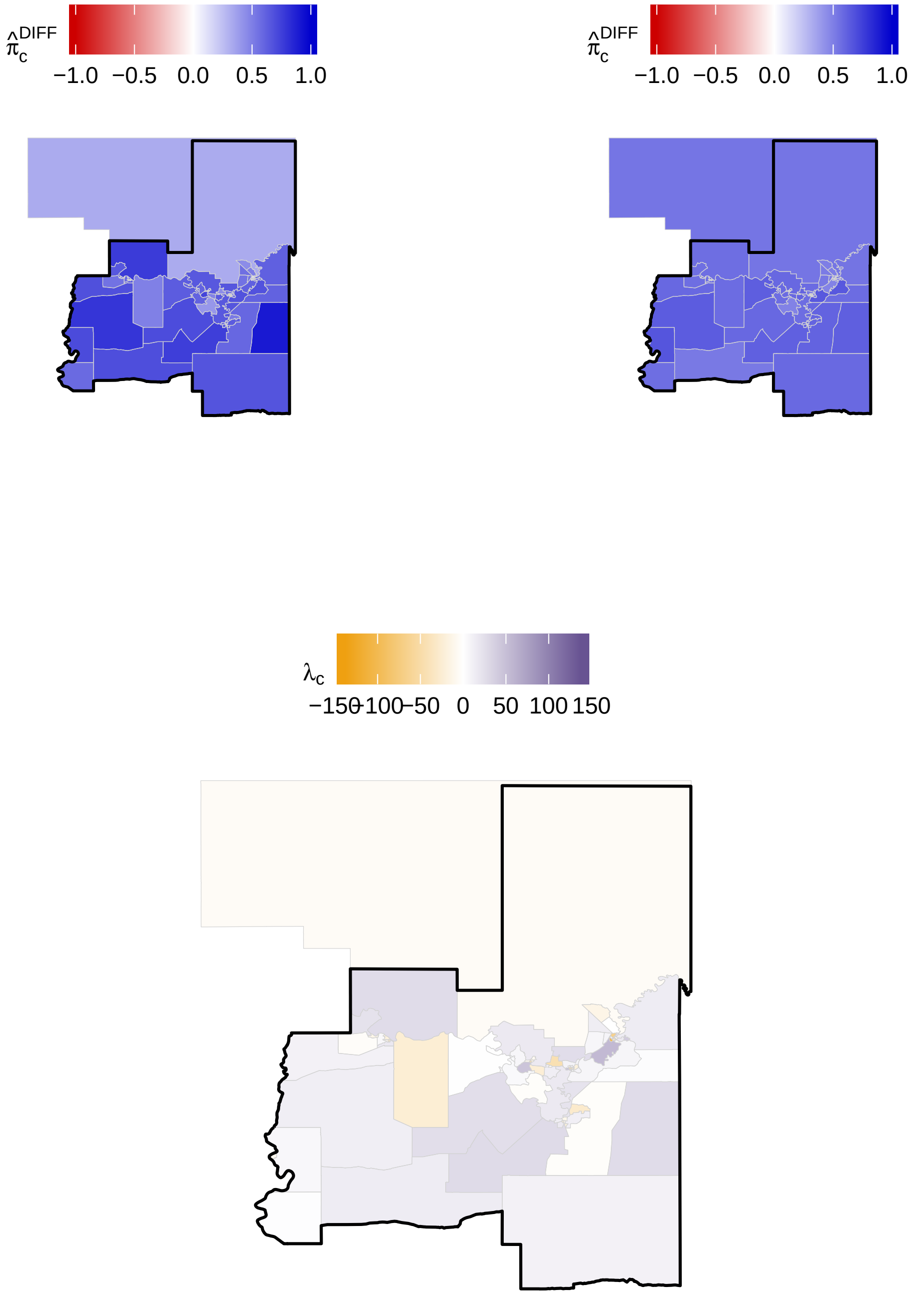


Figure 7.0.4: Chaffey, with collection districts c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 6.2 (top left), $\hat{\pi}_c^{DIFF}$ from EDBC method (top right), and λ_c (bottom).

EDBC predictions.

In Cheltenham, the values of λ_c all lie in the interval $[-62.71, 94.2]$ and in Chaffey the range is $[-101.6, 56.1]$. In Norwood the difference between the predictions from Model 6.2 and the EDBC predictions is much larger, with the range of λ_c being $[-140.7, 102.1]$. For all electoral districts, the absolute value of the means of λ_c are less than 3.

It is difficult to assess the credibility of each set of predictions using only these three electoral districts. Instead, we now consider predictions across the entire of the metropolitan area of Adelaide[1].

Maps displaying these predictions are contained in Appendix C as fold out pages at the back of this thesis. One reason that we consider this region is that this region is the most likely place that changes can be made by the EDBC to meet the fairness requirements. Most electoral districts outside this region are safe (and held by the Liberal Party) and it would be very difficult to make them more marginal without dramatic boundary changes (and those changes would necessarily involve some of these electoral districts).

Figure C.1 shows the predictions for π_c^{DIFF} for each collection district c in the region, under Model 6.2. Figure C.2 shows the equivalent EDBC predictions.

The figures clearly show that the EDBC predictions are much less variable than those from Model 6.2, across the whole area. The EBDC predictions are all very similar to the electoral district results, to the extent that they are unrealistically homogeneous.

Along some electoral boundaries there are some stark differences in the EDBC pre-

[1] Specifically, the electoral districts of Adelaide, Ashford, Bragg, Cheltenham, Colton, Croydon, Elder, Enfield, Florey, Hartley, Lee, Morphett, Norwood, Playford, Port Adelaide, Ramsay, Torrens, Unley, Waite, West Torrens, and Wright.

dictions between collection districts that are neighbours but lie in different electoral districts. For example, in Bragg (in the South-East of the map) the EDBC predictions are a deep blue, but this changes dramatically to a neutral colour over the border to Hartley and Norwood (in the East of the map). It does not seem reasonable that the predictions could change so dramatically at an electoral boundary.

The predictions under Model 6.2 at electoral boundaries do not show such stark changes, which seems more reasonable.

The Model 6.2 predictions contain some collection districts that lean very strongly one way or the other, and many of these collection districts are also very small (in terms of the number of votes). An example of this is the parklands surrounding the Adelaide CBD in the electoral district of Adelaide, which is collection district 4120803 and is coloured bright red on the map. In this collection district under Model 6.2 we have $\hat{\pi}^{ALP}_{4120803} \approx 0.95$, but there are only 14 votes in it.

We need to be careful to not put too much importance on these very small collection districts, particularly because in the context of electoral redistributions we care about the number of votes that are being moved. As we saw in Chapter 3, very small collection districts tend to have very extreme looking demographics and so result in extreme predictions.

To avoid treating small collection districts as more important than they are, we adjust both the Model 6.2 and EDBC predictions by multiplying them by the number of votes that were cast in the collection district. This gives the predicted margin between the Liberal Party and ALP in votes.

Figures C.3 and C.4 contain these predicted margins, from Model 6.2 and the EDBC method respectively. This reduces many of the most extreme predictions under Model 6.2, and appears to also reduce the differential in predictions across electoral boundaries.

Finally, Figure C.5 shows the difference between the two sets of predictions, using the values of λ_c for each collection district c. For the most part, the differences are relatively small, but there are some large values of λ_c in the metropolitan area. These large differences in particular warrant further examination.

We cannot definitively say which set of predictions is 'better' without further research, but based on the evidence explored in this chapter it seems likely that there is considerable variation in the predictions for collection districts that is not present in the EDBC predictions. We believe that Model 6.2 and further refined models based on this are worth further exploration.

Chapter 8

Conclusion

In this thesis we have attempted to develop a method for calculating more accurate predictions for the strength of support for each of the two major parties in South Australia, at the collection district level. We have used the 2010 state election results as our case study, and used results from the 2006 Census of Population and Housing as the basis of the predictors in our models. We have compared the predictions for collection districts under our preferred model to the actual predictions that were used in the electoral redistribution conducted following the 2010 election.

We have seen incremental improvement in the quality of our predictions through the course of eight different logistic regression model fits. The final model predicts the ALP vote in each polling place to within a mean of 30.61 votes. The mean number of votes for the ALP across all the polling places is 638.1, so this represents an error of around 4.8%.

Because of the method the EDBC use to calculate predictions, their estimates are identical to the polling place results when aggregated. This is because the EDBC method overfits the predictions at collection districts by using a weighted average of all polling place results, rather than using only a small number of predictors as in all of the models in this thesis. The EDBC method is therefore not as robust for forecasting as our method, and is also less robust to future demographic changes.

In the course of this thesis we have verified that there is a statistically significant link between education, income, and language spoken at home. From this dataset of aggregated voting and demographic information we will never be able to predict individual voting behaviour, and the coefficients of the model need to be interpreted within this context. However we have made some progress towards answering this interesting supplementary question in this thesis, and justified our choices of predictors, in a contemporary South Australian context.

After comparing the predictions from our preferred model to the EDBC's predictions, we believe ours to be credible and to warrant further exploration. We cannot make definitive statements about the accuracy of the two sets of predictions as we do not know the actual election results on a collection district by collection district basis.

8.1 Ideas for future research

There are many directions that this research could now be taken, and many questions that could be explored. Some examples include:

- **Model with more predictors**. This is a natural extension to this work, and the most obvious predictor categories to include next include age, gender, country of birth, religious affiliation, labour force status, and occupation.
- **Include a factor in the model identifying a polling place as containing either ordinary or declaration votes**. Recall that 47 of the 746 polling places (one for each electoral district) collect the declaration votes cast in the election. We believe that the voting behaviour for declaration voters is

different to that for ordinary votes (based partly on anecdotal evidence from ECSA). Introducing this factor would be an attempt to capture this difference in voting behaviour.

- **Introduce interaction terms**. Thus, far none of our modelling has included two or more way interaction terms, largely because this vastly increases the number of predictors that need to be fitted. This could be explored with the aim of an improved model fit.

- **Extend the spatially-aware models to the multinomial regression model**. To simplify the model in Chapter 6 we removed all informal votes from the system. We could look at trying to model them by returning to the multinomial regression model, and see if there are any interesting results regarding the informal votes.

- **Predict the effect of actual boundary changes using Model 6.2**. We could predict the effect to the margins in electoral districts from the boundary changes made after the 2010 election, under our set of predictions, to see how similar this is to the EDBC's predicted effect. In particular we could test whether any seats have the party holding them notionally changed under these predictions.

- **Research to determine the actual π_c^p in collection districts**. If we knew the actual values for π_c^p in collection districts, we would be better able to determine the accuracy of Model 6.2's predictions. It is likely that the best way to determine this is through polling in a number of test collection districts. It makes sense to do this soon after an election, or possibly as an exit poll on election day.

- **Fit Model 6.2 to the 2014 election, using results from the 2011 Census**. This gives us another dataset to test the methods on. Being a more contemporary example, it would also be more interesting to the EDBC when they begin the next electoral redistribution process in the second half of 2015.

- **Extend the research to a longitudinal study of more elections**. If we had predictions for the coefficients over a greater time period, we would be able to look at the relative importance and direction of the predictors over time, and better explain historical changes in voting patterns through demographics.

Appendix A

Supplementary Material for Chapters 4 & 5

This appendix contains supplementary material for the work in Chapters 4 and 5 regarding the data cleaning of the voter location data and the five multinomial regression models.

A.1 Mismatches in Voter Location Data

Table 1.1.1: The 30 polling places that have the greatest difference between the actual number of votes cast in the polling place (shown in the ‘Actual Votes’ column) and the number of votes recorded in the voter location data (shown in the ‘Voter Location column). The difference between these two is shown in the ‘Mismatch’ column.

	Polling Place	Electoral District	Voter Location	Actual Votes	Mismatch
1	170	614	2528	2821	-293
2	158	614	700	871	-171
3	175	614	1202	1351	-149
4	169	614	2160	2306	-146
5	163	614	1771	1863	-92
6	155	614	702	777	-75
7	134	612	2484	2411	73
8	154	614	1324	1392	-68
9	180	614	519	582	-63
10	166	614	470	524	-54
11	157	614	653	700	-47
12	156	614	465	510	-45
13	171	614	1308	1352	-44
14	176	614	116	152	-36
15	2172	645	2221	2185	36
16	153	614	220	253	-33
17	174	614	834	865	-31
18	172	614	169	197	-28
19	2179	614	233	206	27
20	161	614	70	96	-26
21	493	634	1578	1552	26
22	164	614	209	233	-24
23	471	631	1886	1908	-22
24	177	614	216	235	-19
25	160	614	245	263	-18
26	575	623	694	677	17
27	2271	623	1324	1307	17
28	179	614	99	115	-16
29	627	641	1288	1304	-16
30	167	614	83	97	-14

A.2 Multinomial Regression Models

All tables in this section were created with texreg [25].

A.2.1 Model 4.1

Table 1.2.1: Model output for Model 4.1. The numbers in parentheses are the standard errors of the parameter above.

	LIB	INF
(Intercept)	-8.95^{***}	-8.44^{***}
	(0.29)	(0.79)
NegativeOrNilIncome	36.78^{***}	19.49^{***}
	(1.51)	(4.25)
X1.499	-5.54^{***}	1.36^{**}
	(0.18)	(0.48)
X500.999	-3.36^{***}	-6.42^{***}
	(0.30)	(0.83)
X1000.1399	9.06^{***}	11.98^{***}
	(0.40)	(1.10)
X1400.1999	-12.54^{***}	-9.52^{***}
	(0.53)	(1.41)
X2000.2499	2.08^{*}	7.51^{***}
	(0.82)	(2.28)
X2500.2999	14.73^{***}	7.53^{**}
	(0.96)	(2.63)
X3000.3499	-62.35^{***}	-18.92^{*}
	(2.66)	(7.42)
X3500.3999	-4.46	-27.83^{***}
	(2.71)	(7.84)
MoreThan4000	48.68^{***}	23.23^{***}
	(1.83)	(5.41)
DoesNotSpeakEnglishAtHome	13.60^{***}	4.72^{***}
	(0.39)	(1.07)

	LIB	INF
SpeaksEnglishAtHome	14.23^{***}	4.54^{***}
	(0.36)	(0.99)
LessThanYear12	-1.79^{***}	3.43^{***}
	(0.25)	(0.69)
Year12OrEquivalent	-13.88^{***}	0.92
	(0.36)	(1.02)
BDAD	30.49^{***}	-2.60
	(0.53)	(1.49)
PDGD	-35.21^{***}	13.67^{***}
	(1.10)	(3.30)
CL	-7.96^{***}	-2.28^{*}
	(0.35)	(0.97)
AIC	1597640.43	1597640.43
BIC	1598066.31	1598066.31
Log Likelihood	-798784.22	-798784.22
Deviance	1597568.43	1597568.43
Num. obs.	141	141

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.2.2 Model 4.2

Table 1.2.2: Model output for Model 4.2. The numbers in parentheses are the standard errors of the parameter above.

	LIB	INF
(Intercept)	0.06^{***}	-2.65^{***}
	(0.00)	(0.01)
Comp.1	0.52^{***}	0.40^{***}
	(0.01)	(0.04)
Comp.2	-2.77^{***}	0.18^{***}
	(0.02)	(0.05)
AIC	1630721.17	1630721.17
BIC	1630792.15	1630792.15
Log Likelihood	-815354.58	-815354.58
Deviance	1630709.17	1630709.17
Num. obs.	141	141

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.2.3 Model 4.3

Table 1.2.3: Model output for Model 4.3. The numbers in parentheses are the standard errors of the parameter above.

	LIB	INF
(Intercept)	0.07^{***}	-2.63^{***}
	(0.00)	(0.01)
Comp.1	0.55^{***}	0.51^{***}
	(0.01)	(0.04)
Comp.2	-2.80^{***}	0.20^{***}
	(0.02)	(0.05)
Comp.3	5.03^{***}	1.75^{***}
	(0.05)	(0.15)
Comp.4	-5.24^{***}	-1.69^{***}
	(0.08)	(0.22)
Comp.5	0.78^{***}	0.40
	(0.12)	(0.33)
Comp.6	8.40^{***}	3.43^{***}
	(0.15)	(0.41)
Comp.7	-17.92^{***}	-6.82^{***}
	(0.19)	(0.50)
Comp.8	20.57^{***}	13.69^{***}
	(0.34)	(0.92)
Comp.9	-0.23	5.75^{***}
	(0.42)	(1.18)
Comp.10	-6.59^{***}	-3.26^{**}
	(0.45)	(1.23)
Comp.11	6.09^{***}	-6.44^{***}
	(0.48)	(1.33)
Comp.12	-12.55^{***}	-9.84^{***}
	(0.87)	(2.33)
Comp.13	0.57	-19.82^{***}
	(1.08)	(3.07)
Comp.14	21.59^{***}	-16.28^{***}
	(1.10)	(3.37)

	LIB	INF
Comp.15	-67.42^{***}	-18.41^{***}
	(1.53)	(4.32)
Comp.16	-24.90^{***}	-33.37^{***}
	(2.92)	(8.51)
Comp.17	-64.79^{***}	-12.78
	(3.08)	(8.62)
AIC	1597640.43	1597640.43
BIC	1598066.30	1598066.30
Log Likelihood	-798784.21	-798784.21
Deviance	1597568.43	1597568.43
Num. obs.	141	141

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.2.4 Model 5.1

Table 1.2.4: Model output for Model 5.1. The numbers in parentheses are the standard errors of the parameter above.

	LIB	INF
(Intercept)	-2.23^{***}	-4.08^{***}
	(0.14)	(0.40)
NegativeOrNilIncome	14.71^{***}	6.36^{***}
	(0.49)	(1.35)
X1.499	-4.00^{***}	-0.27
	(0.08)	(0.23)
X500.999	-4.38^{***}	-1.95^{***}
	(0.12)	(0.32)
X1000.1399	5.82^{***}	3.77^{***}
	(0.16)	(0.43)
X1400.1999	-5.99^{***}	-3.32^{***}
	(0.18)	(0.51)
X2000.2499	-4.00^{***}	-3.32^{***}
	(0.29)	(0.82)
X2500.2999	0.03	2.88^{**}
	(0.33)	(0.92)
X3000.3499	-1.85^{***}	-0.73
	(0.56)	(1.61)
X3500.3999	-0.05	0.50
	(0.62)	(1.82)
MoreThan4000	7.84^{***}	3.49^{*}
	(0.53)	(1.57)
DoesNotSpeakEnglishAtHome	1.42^{***}	1.04
	(0.19)	(0.54)
SpeaksEnglishAtHome	3.29^{***}	0.94
	(0.19)	(0.52)
LessThanYear12	4.00^{***}	2.11^{***}
	(0.14)	(0.37)
Year12OrEquivalent	-3.75^{***}	-0.05
	(0.18)	(0.49)

	LIB	INF
BDAD	14.46^{***}	1.57^{*}
	(0.23)	(0.65)
PDGD	-14.36^{***}	-2.94^{*}
	(0.49)	(1.39)
CL	-8.32^{***}	-1.40^{**}
	(0.19)	(0.51)
AIC	1598271.75	1598271.75
BIC	1598697.68	1598697.68
Log Likelihood	-799099.87	-799099.87
Deviance	1598199.75	1598199.75
Num. obs.	2238	2238

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.2.5 Model 5.2

Table 1.2.5: Model output for Model 5.2. The numbers in parentheses are the standard errors of the parameter above.

	LIB	INF
(Intercept)	0.03^{***}	-2.65^{***}
	(0.00)	(0.01)
Comp.1	-0.88^{***}	-0.35^{***}
	(0.01)	(0.03)
Comp.2	-2.25^{***}	0.43^{***}
	(0.02)	(0.04)
AIC	1632728.47	1632728.47
BIC	1632799.45	1632799.45
Log Likelihood	-816358.23	-816358.23
Deviance	1632716.47	1632716.47
Num. obs.	2238	2238

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.3 Model 4.2 Validation

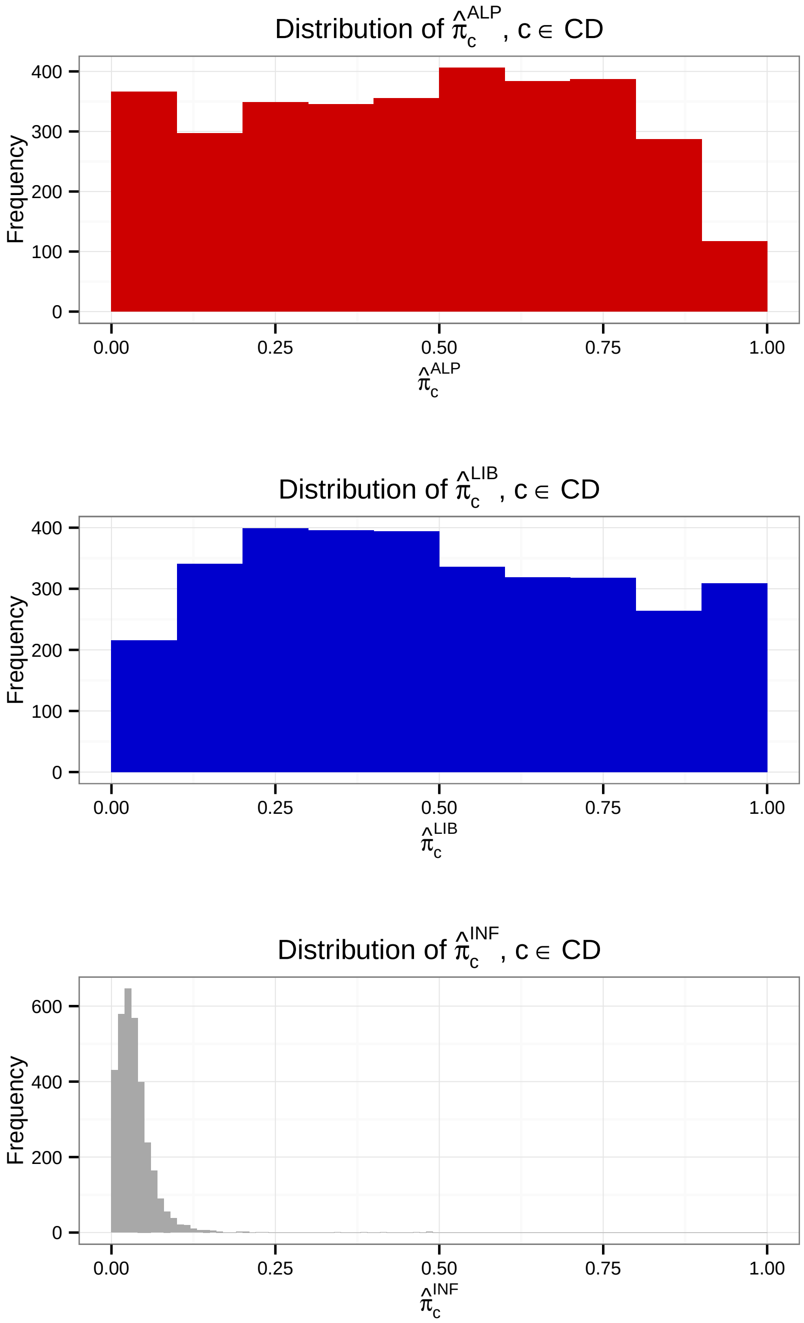


Figure A.3.1: Distribution of predicted values for $\hat{\pi}_c^p$, for $p \in P$ and $c \in CD$, under Model 4.2.

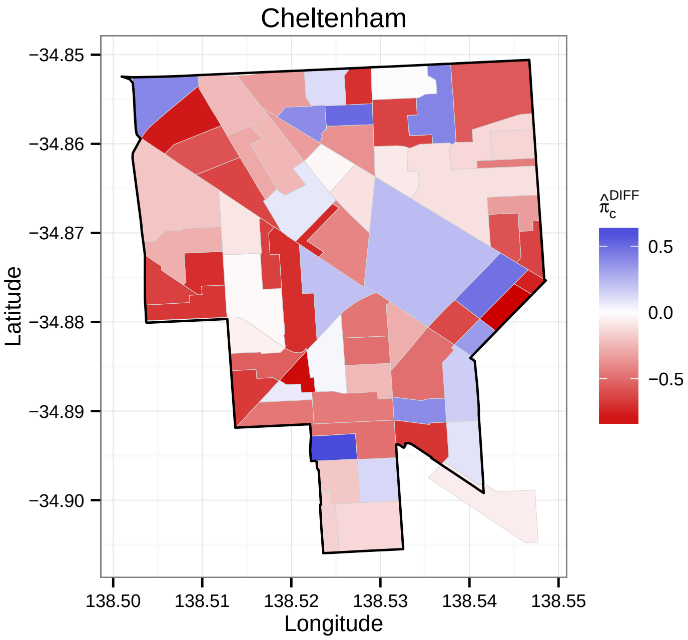


Figure A.3.2: The electoral district of Cheltenham, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 4.2.

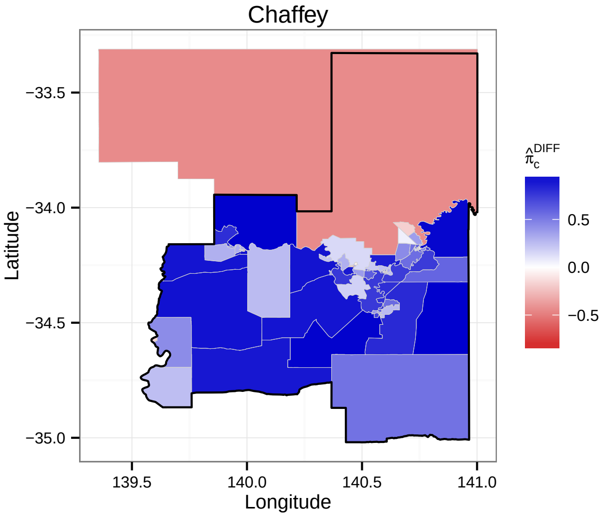


Figure A.3.3: The electoral district of Chaffey, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 4.2.

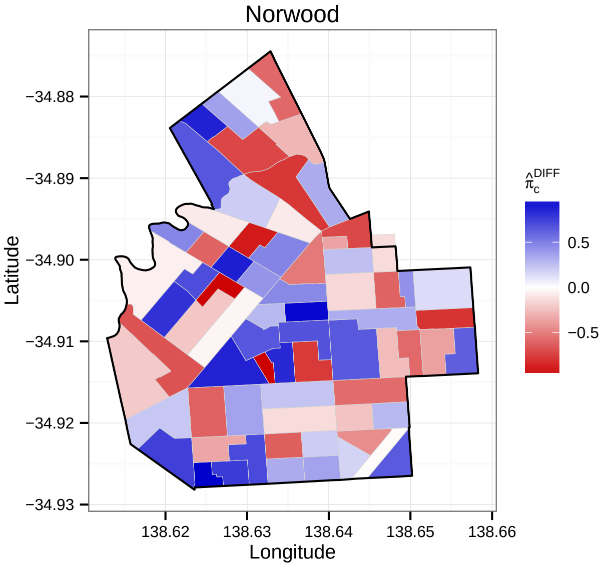


Figure A.3.4: The electoral district of Norwood, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$ under Model 4.2.

A.4 Side by side views of multinomial graphics

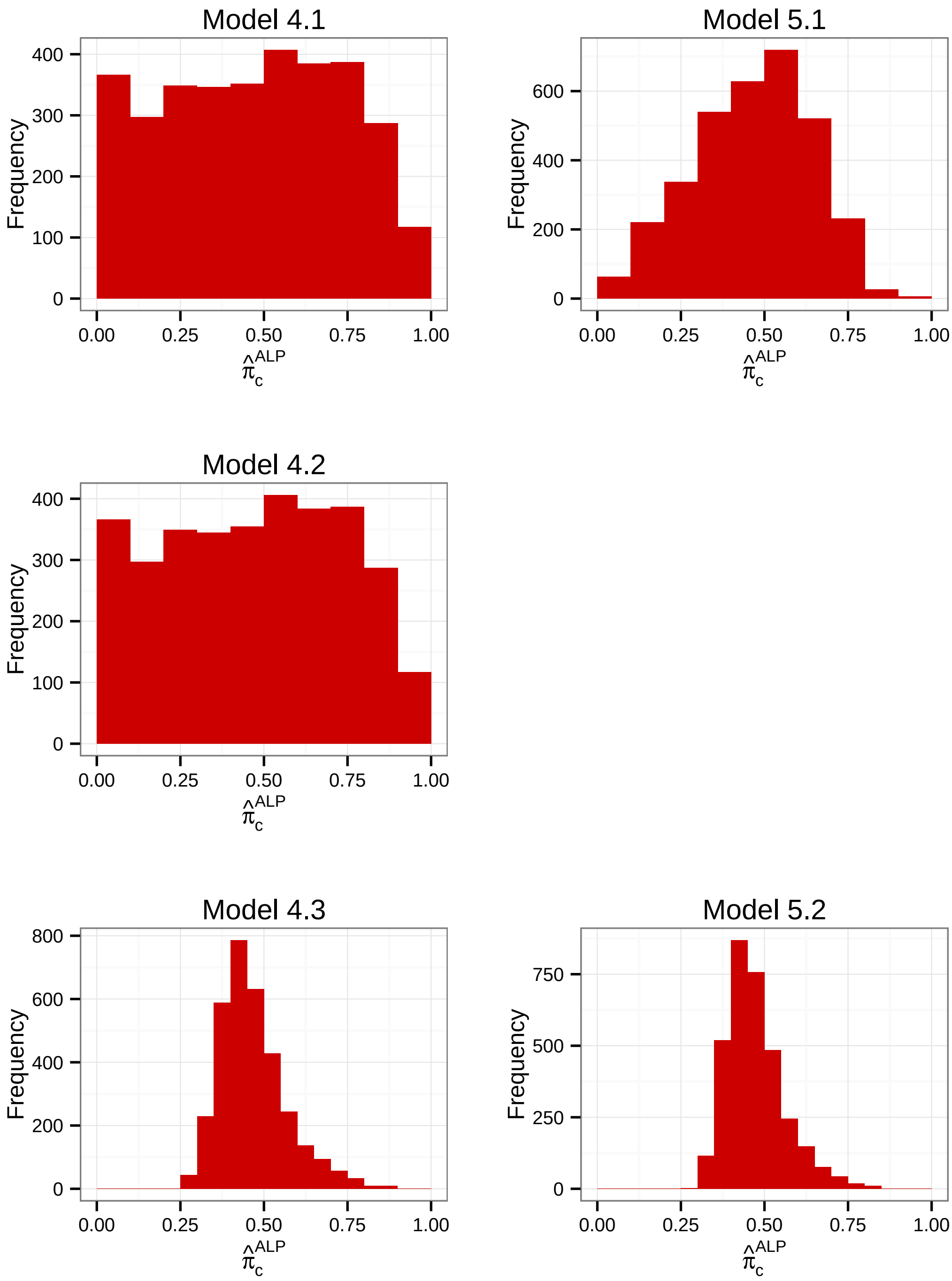


Figure A.4.1: Side by side views of the distribution of predicted values for $\hat{\pi}_c^{ALP}$, for $c \in CD$, under the five multinomial regression models.

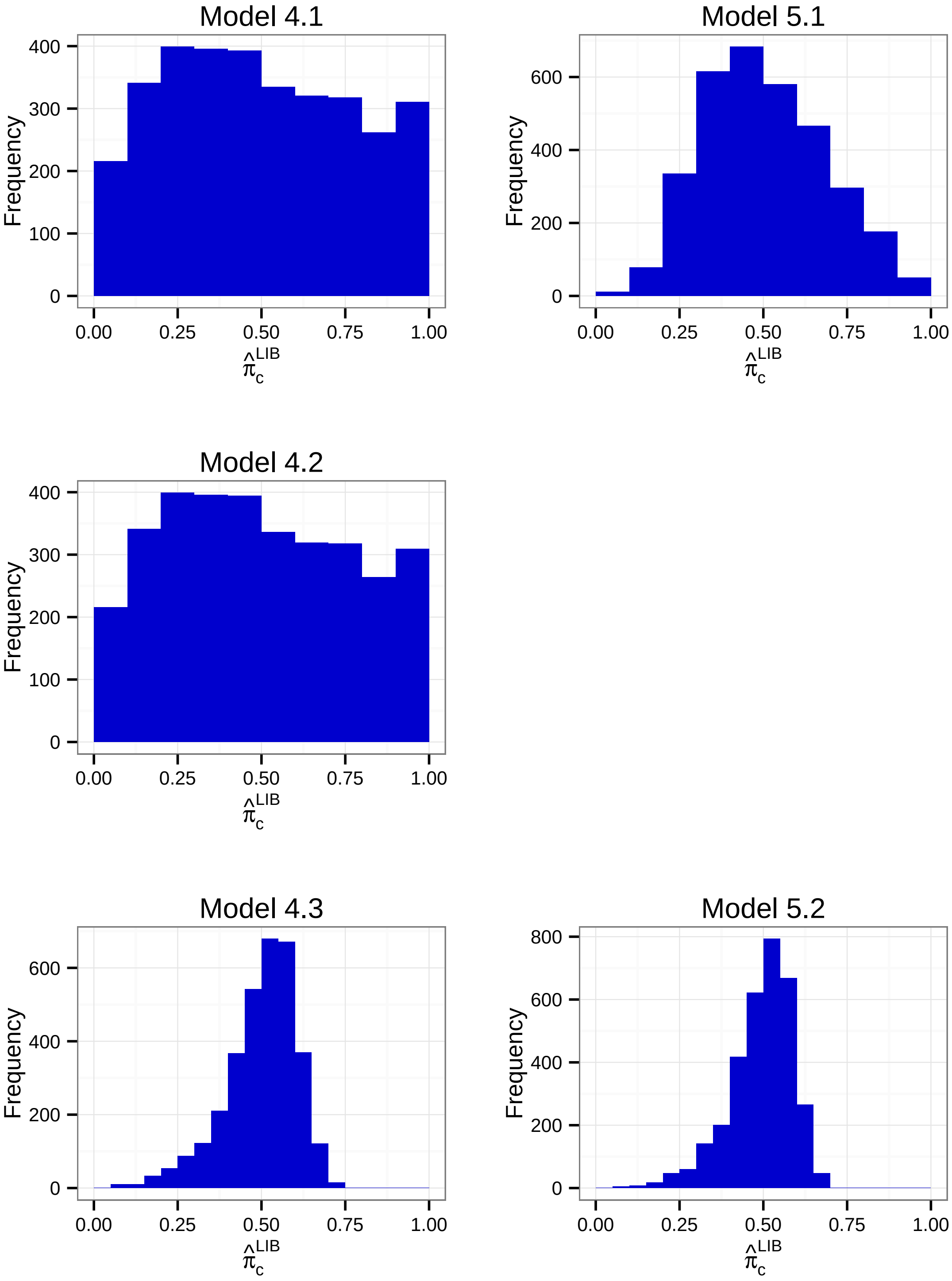


Figure A.4.2: Side by side views of the distribution of predicted values for $\hat{\pi}_c^{LIB}$, for $c \in CD$, under the five multinomial regression models.

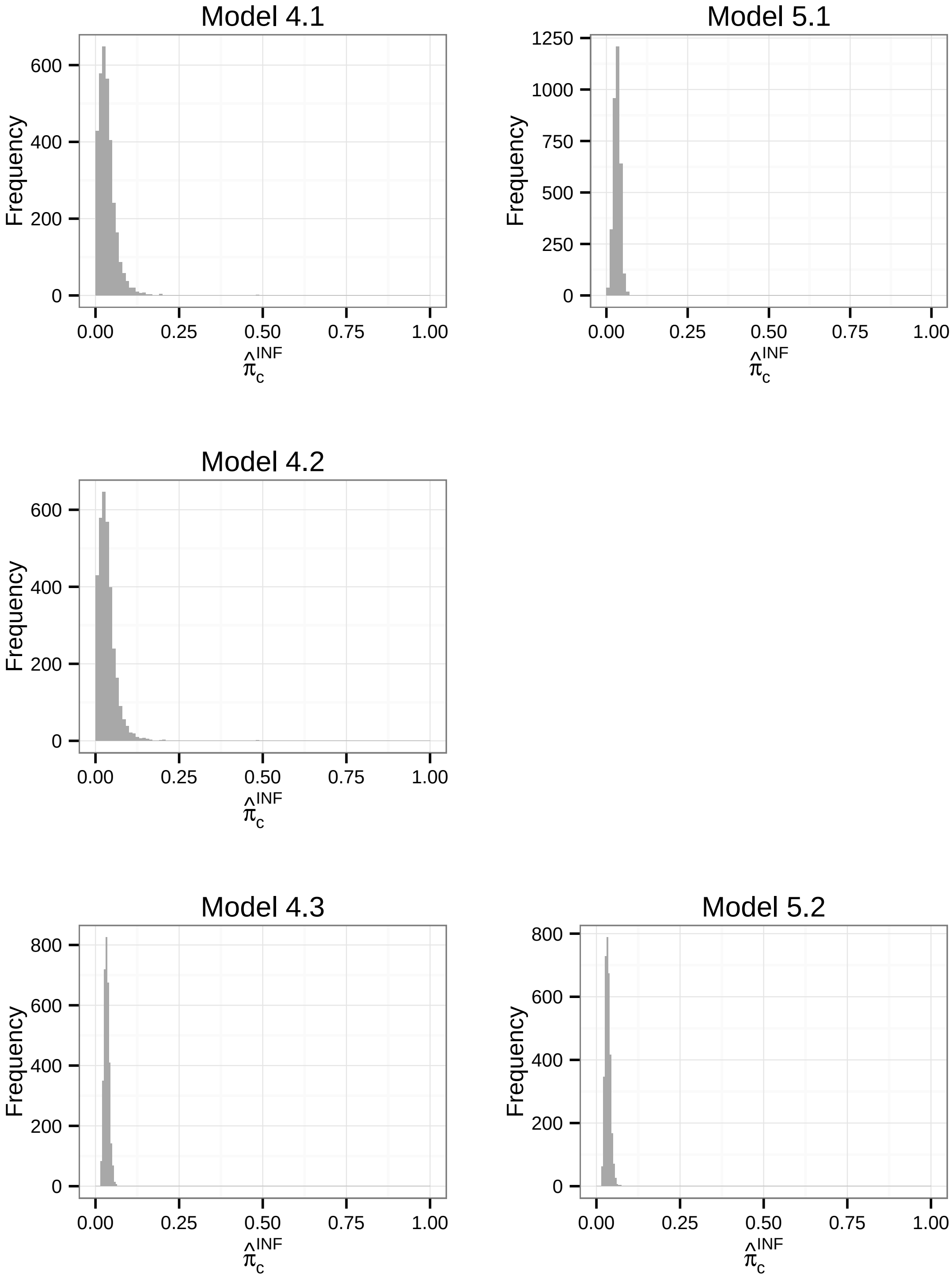


Figure A.4.3: Side by side views of the distribution of predicted values for $\hat{\pi}_c^{INF}$, for $c \in CD$, under the five multinomial regression models.

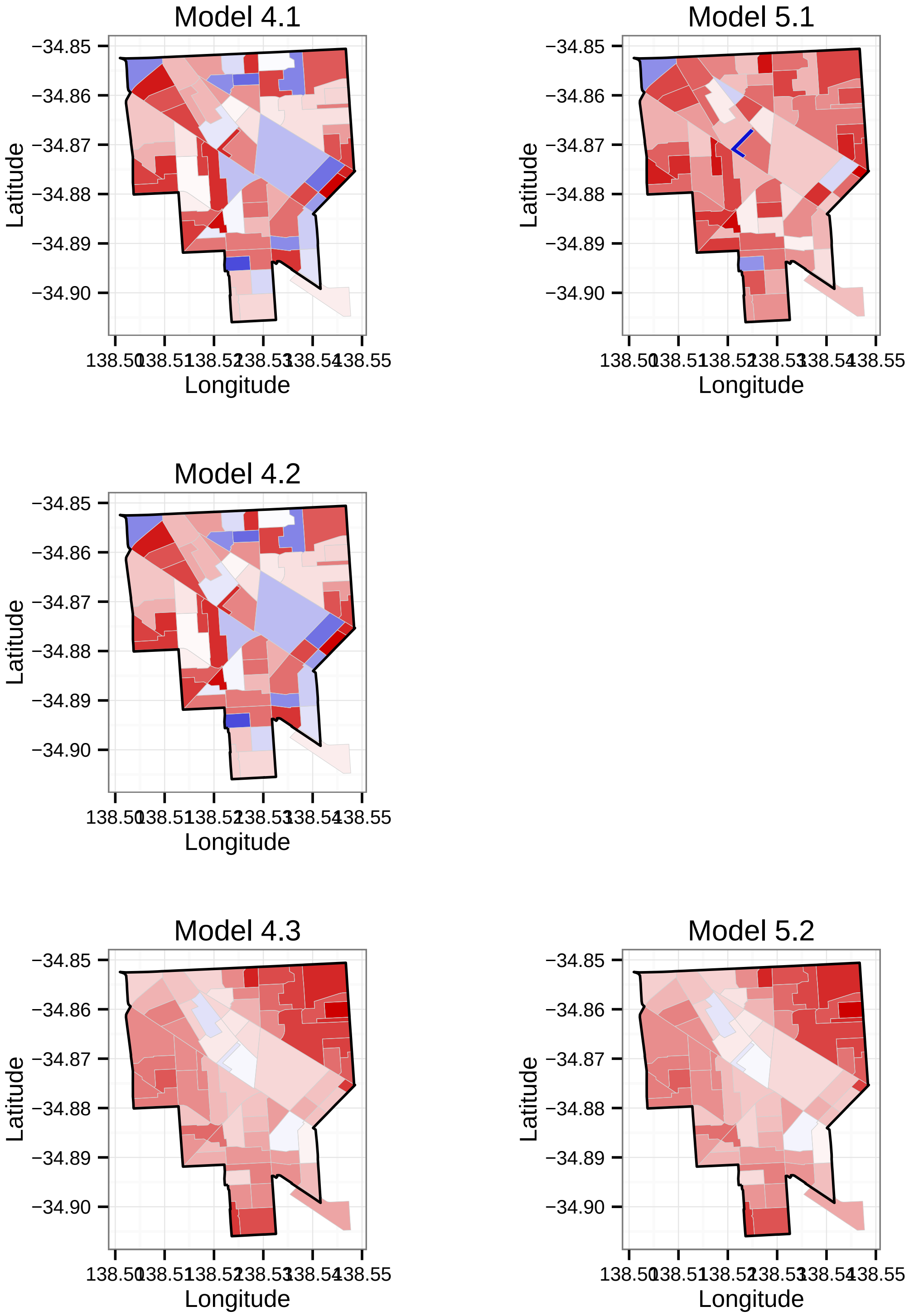


Figure A.4.4: Side by side views of the electoral district of Cheltenham, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$, under the five multinomial regression models.

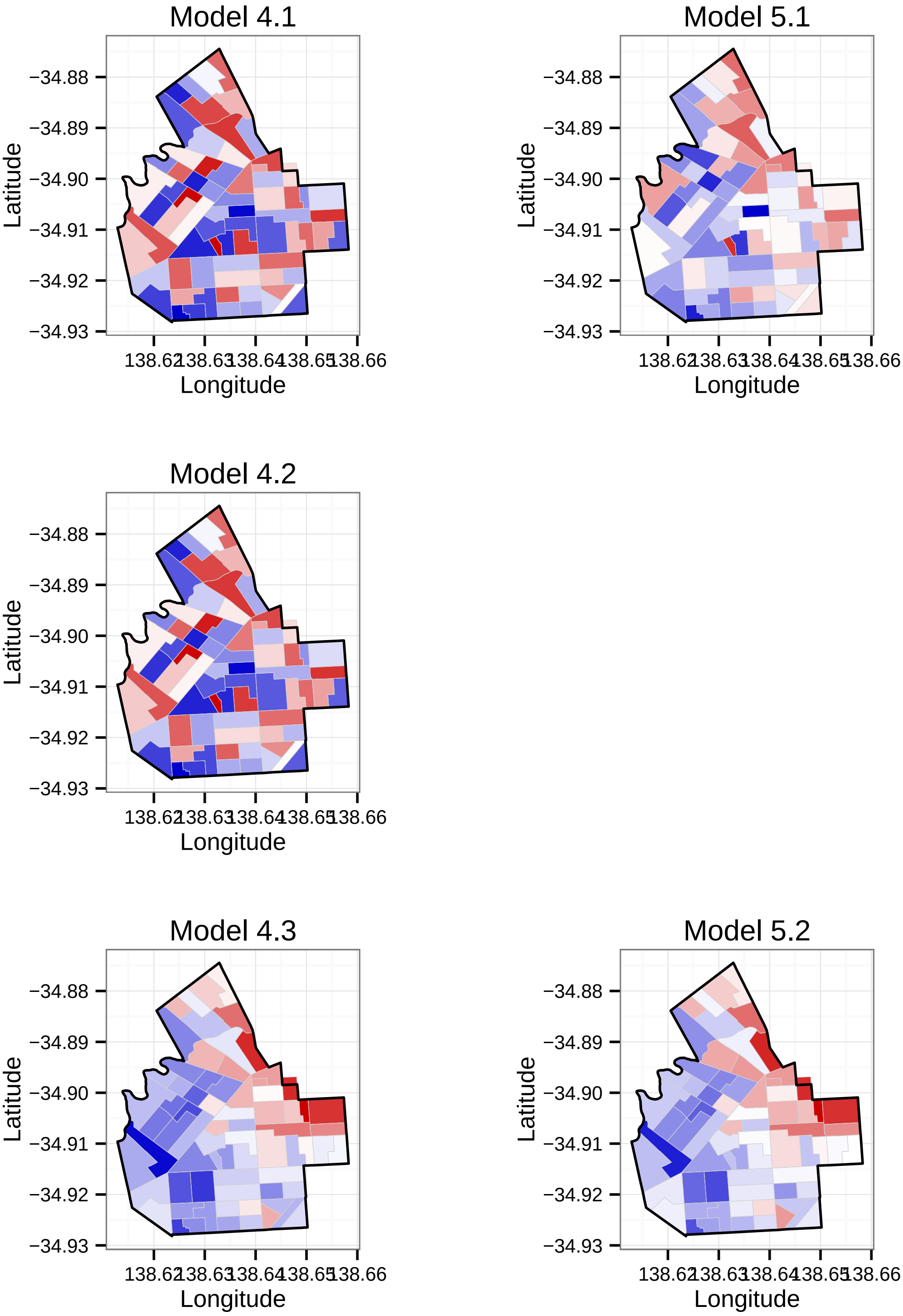


Figure A.4.5: Side by side views of the electoral district of Norwood, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$, under the five multinomial regression models.

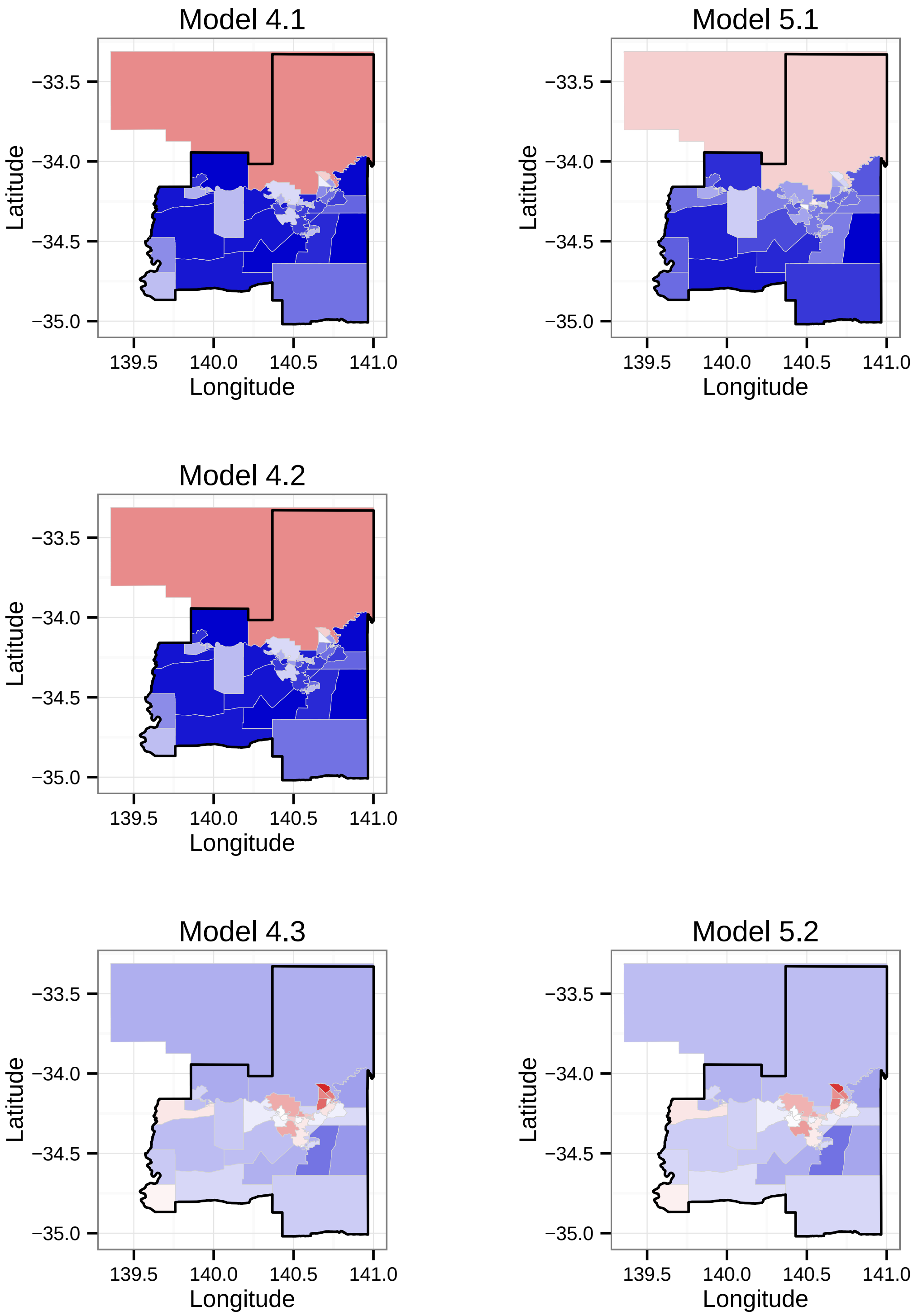


Figure A.4.6: Side by side views of the electoral district of Chaffey, with collection district c coloured according to $\hat{\pi}_c^{DIFF}$, under the five multinomial regression models.

Appendix B

Supplementary Material for Chapter 6

B.1 Full Model 6.2 Output

Table 2.1.1: Model output for Model 6.2 using the full set of predictors before model selection is performed.

	Estimate	Std. Error	z value	Pr(>|z|)
NegativeOrNilIncome	-9.6145	0.5347	-17.98	0.0000 ***
X1.499	2.8948	0.1077	26.88	0.0000 ***
X500.999	1.2350	0.1427	8.65	0.0000 ***
X1000.1399	-2.3593	0.1837	-12.84	0.0000 ***
X1400.1999	3.3827	0.2072	16.32	0.0000 ***
X2000.2499	-0.5069	0.3423	-1.48	0.1386
X2500.2999	-0.8409	0.3617	-2.33	0.0201 *
X3000.3499	-2.1939	0.5845	-3.75	0.0002 ***
X3500.3999	-0.9207	0.6423	-1.43	0.1517
MoreThan4000	-7.9508	0.5808	-13.69	0.0000 ***
DoesNotSpeakEnglishAtHome	1.8897	0.2193	8.62	0.0000 ***
SpeaksEnglishAtHome	0.1966	0.2088	0.94	0.3463
LessThanYear12	-4.5209	0.1654	-27.33	0.0000 ***
Year12OrEquivalent	-2.1688	0.2472	-8.77	0.0000 ***
BDAD	-4.8273	0.3054	-15.81	0.0000 ***
PDGD	11.2307	0.5651	19.88	0.0000 ***
CL	5.2439	0.2590	20.25	0.0000 ***
Significance codes:	· p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

B.2 Full Model 6.3 Output

Table 2.2.1: Model output for Model 6.3 using the full set of predictors before model selection is performed.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.2496	0.1441	15.61	0.0000 ***
NegativeOrNilIncome	-14.6501	0.4907	-29.85	0.0000 ***
X1.499	4.0073	0.0850	47.16	0.0000 ***
X500.999	4.4148	0.1211	36.46	0.0000 ***
X1000.1399	-5.8261	0.1583	-36.82	0.0000 ***
X1400.1999	6.0118	0.1839	32.68	0.0000 ***
X2000.2499	4.0051	0.2924	13.70	0.0000 ***
X2500.2999	-0.0352	0.3276	-0.11	0.9145
X3000.3499	1.8996	0.5563	3.42	0.0006 ***
X3500.3999	0.0783	0.6241	0.13	0.9002
MoreThan4000	-7.8191	0.5274	-14.83	0.0000 ***
DoesNotSpeakEnglishAtHome	-1.4304	0.1941	-7.37	0.0000 ***
SpeaksEnglishAtHome	-3.3026	0.1876	-17.61	0.0000 ***
LessThanYear12	-4.0208	0.1360	-29.56	0.0000 ***
Year12OrEquivalent	3.7118	0.1780	20.86	0.0000 ***
BDAD	-14.4354	0.2352	-61.37	0.0000 ***
PDGD	14.3781	0.4878	29.48	0.0000 ***
CL	8.3255	0.1885	44.16	0.0000 ***

Significance codes: · p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Bibliography

[1] Constitution Act, 1934, SA, s83(1-3).

[2] Electoral Act, 1929-1976, SA, s125(14).

[3] *1976 Report of the Electoral Districts Boundaries Commission*. Government Printer, 1976.

[4] *2007 Report of the Electoral Districts Boundaries Commission*. March 2007.

[5] *South Australian Parliamentary Debates*, 8 October 1975.

[6] The Australian National University. Australian Election Study. `http://aes.anu.edu.au/aes`.

[7] Clive Bean and Ian McAllister. The Australian election survey: the tale of the rabbit-less hat. Voting behaviour in 2007. *Australian Cultural History*, 27(2):205, October 2009.

[8] Kenneth Benoit. Electoral laws as political consequences: Explaining the origins and change of electoral institutions. *Annual Review Of Political Science*, 10:363–390, 2007.

[9] Roger Bivand, Tim Keitt, and Barry Rowlingson. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.8-16. `http://CRAN.R-project.org/package=rgdal`, 2014.

[10] William Bowe. Why the ‘wrong’ party could win SA’s state election. `http://www.crikey.com.au/2014/01/24/why-the-wrong-party-could-win-sas-state-election/`, January 2014.

[11] ABC News (Australian Broadcasting Corporation). Chaffey - 2010 South Australian Election. `http://www.abc.net.au/elections/sa/2010/guide/chaf.htm`, March 2010.

[12] ABC News (Australian Broadcasting Corporation). Cheltenham - 2010 South Australian Election. `http://www.abc.net.au/elections/sa/2010/guide/chel.htm`, March 2010.

[13] ABC News (Australian Broadcasting Corporation). Norwood - 2010 South Australian Election. `http://www.abc.net.au/elections/sa/2010/guide/norw.htm`, March 2010.

[14] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. CRC Press, third edition, 2008.

[15] Maurice Duverger. *Political parties: their organization and activity in the modern state*. J. Wiley, 1954.

[16] Deon Filmer. The structure of social disparities in education: Gender and wealth. *Policy Research Report on Gender and Development, Working Paper Series No. 5*, November 1999.

[17] John Fox and Sanford Weisberg. *An R companion to applied regression.* SAGE Publications Inc., 2nd edition, 2011.

[18] Murray Goot and Ian Watson. Explaining Howard's success: Social structure, issue agendas and party support, 1993–2004. *Australian Journal of Political Science*, 42(2):253–276, June 2007.

[19] Antony Green. 2014 South Australian post-election pendulum. `http://blogs.abc.net.au/antonygreen/2014/03/2014-south-australian-post-election-pendulum.html`, March 2014.

[20] D. H. Jaensch. Under-representation and the 'gerrymander' in the Playford era. *The Australian Journal of Politics and History*, 17(1):82–95, 1971.

[21] Dean Jaensch. Community access to the electoral processes in South Australia since 1850. *South Australian State Electoral Office*, 2002.

[22] Dean Jaensch. Swinging voters stymie fairness clause. `http://www.adelaidenow.com.au/news/opinion/dean-jaensch-swinging-voters-stymie-fairness-clause/story-fni6unxq-1226858544336`, March 2014.

[23] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York Inc., 1986.

[24] David G Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. Springer, 3rd edition, 2010.

[25] Philip Leifeld. texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8):1–24, 2013.

[26] Professor Clement Macintyre. Personal communication. 2012-2015.

[27] P. McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

[28] David S. Moore, George P. McCabe, and Bruce Craig. *Introduction to the practice of statistics*. W. H. Freeman and Company, 7th edition, 2012.

[29] Jenni Newton-Farrelly. From gerry-built to purpose-built: Drawing electoral boundaries for unbiased election outcomes. *Representation*, 45(4):471–484, 2009.

[30] City of Charles Sturt Ratepayers and Residents. Save St Clair Update - Good luck to our candidates! http://www.charlessturtratepayers.org/2010/02/09/save-st-clair-update-good-luck-to-our-candidates/, February 2010.

[31] Electoral Commission of South Australia. 2010 electoral district boundaries shapefile. Supplied by Deputy Electoral Commissioner David Gully by email.

[32] Australian Bureau of Statistics. Australian Standard Geographical Classification (ASGC), Digital Boundaries, 2006. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1259.0.30.002Main+Features12006?OpenDocument.

[33] Australian Bureau of Statistics. Census dictionary 2006. http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/bf9bec7e072fde1eca257230001c24d8/$FILE/29010_2006%20(reissue).pdf.

[34] Australian Bureau of Statistics. Census household form. http://www.abs.gov.au/AUSSTATS/abs@.nsf/bb8db737e2af84b8ca2571780015701e/05d78f82343a3ae5ca25720900078b62/$FILE/2006%20Census%20Form.pdf.

[35] Australian Bureau of Statistics. Tablebuilder. http://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder.

[36] Australian Bureau of Statistics. 2011 Census of Population and Housing: Basic Community Profile, South Australia. http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/communityprofile/4?opendocument&navpos=220, 2012.

[37] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[38] Oxford University Press. “oneiromancy, n.”. OED Online. December 2014. `http://www.oed.com/view/Entry/258270?redirectedFrom=oneiromancy`.

[39] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002.

[40] Karl R. White. The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, 91(3):461–481, 1982.

[41] Dennis Woodward, Andrew Parkin, and John Summers, editors. *Government, politics, power and policy in Australia.* Pearson Australia, 9th edition, 2010.

Appendix C

Supplementary Material For Chapter 7

This appendix contains five foldout maps that accompany Chapter 7. Each map contains the collection districts of the same 21 electoral districts in the metropolitan area: Adelaide, Ashford, Bragg, Cheltenham, Colton, Croydon, Elder, Enfield, Florey, Hartley, Lee, Morphett, Norwood, Playford, Port Adelaide, Ramsay, Torrens, Unley, Waite, West Torrens, and Wright.

The collection districts in the maps are coloured by:

- Figure C.1: The predictions for π_c^{DIFF} under Model 6.2.
- Figure C.2: The EDBC predictions for π_c^{DIFF}.
- Figure C.3: The size of the margin, in votes, between the Liberal Party and the ALP under Model 6.2.
- Figure C.4: The size of the margin, in votes, between the Liberal Party and the ALP according to the EDBC predictions.

- Figure C.5: The value of λ_c, or in other words, the difference between the margin under Model 6.2, and the margin according to the EDBC predictions.

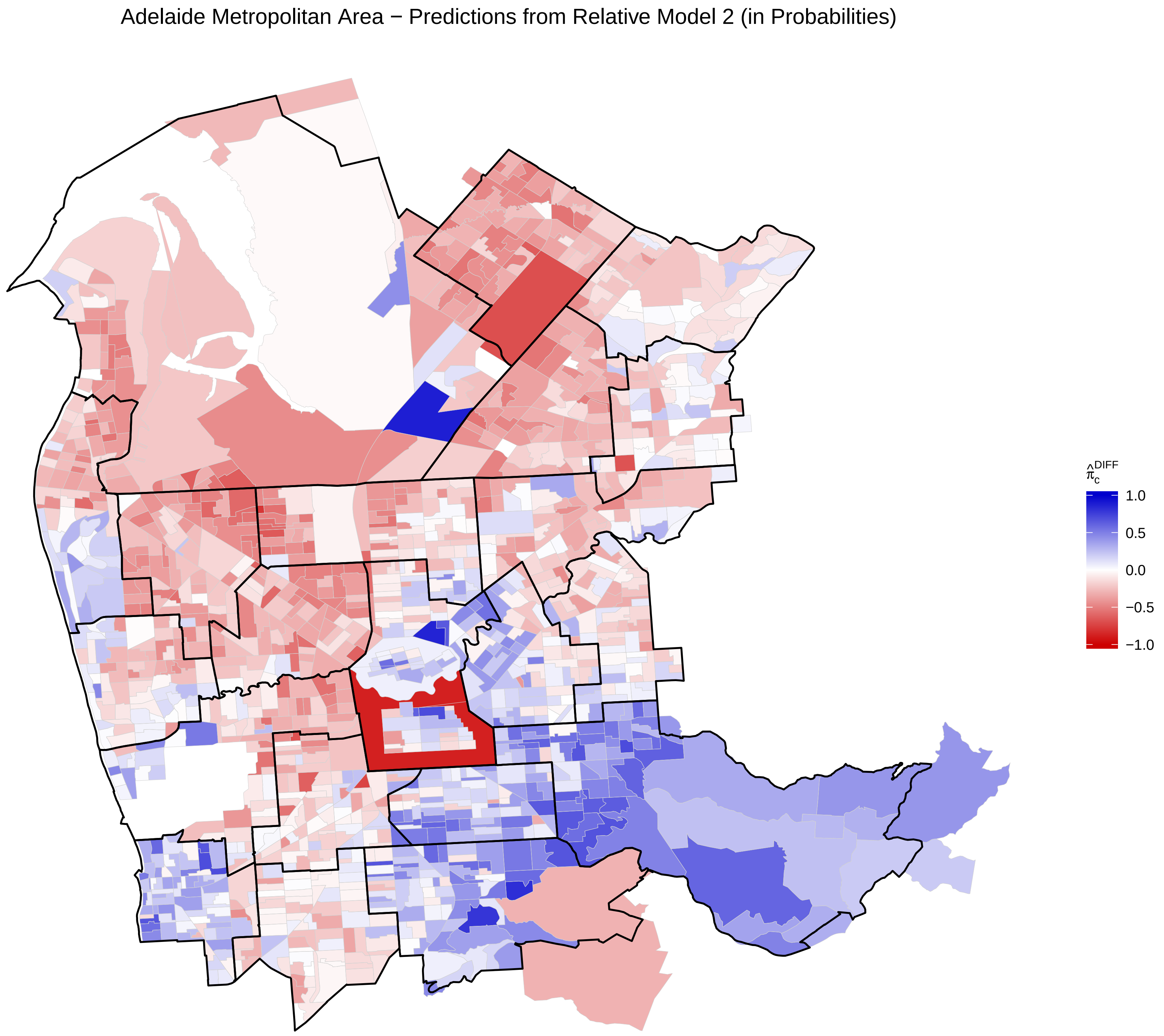

Adelaide Metropolitan Area − Predictions from Relative Model 2 (in Probabilities)
$\hat{\pi}_c^{DIFF}$
1.0
0.5
0.0
−0.5
−1.0

Figure C.1: Map showing the predictions for π_c^{DIFF} under Model 6.2 for collection districts $c \in CD$. 21 electoral districts in metropolitan Adelaide are included.

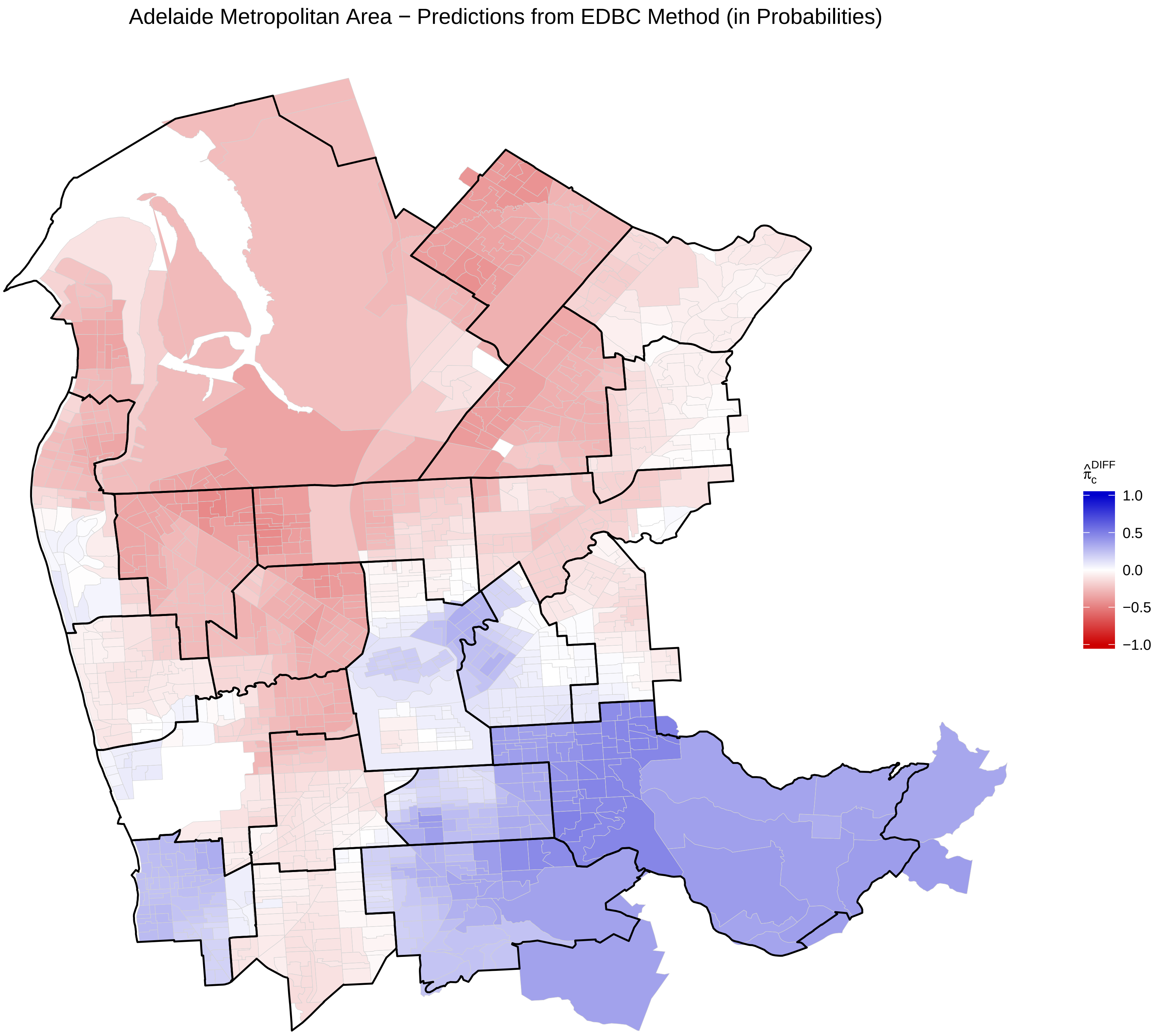

Adelaide Metropolitan Area − Predictions from EDBC Method (in Probabilities)
$\hat{\pi}_c^{DIFF}$
1.0
0.5
0.0
−0.5
−1.0

Figure C.2: Map showing the EDBC predictions for π_c^{DIFF} for collection districts $c \in CD$. 21 electoral districts in metropolitan Adelaide are included.

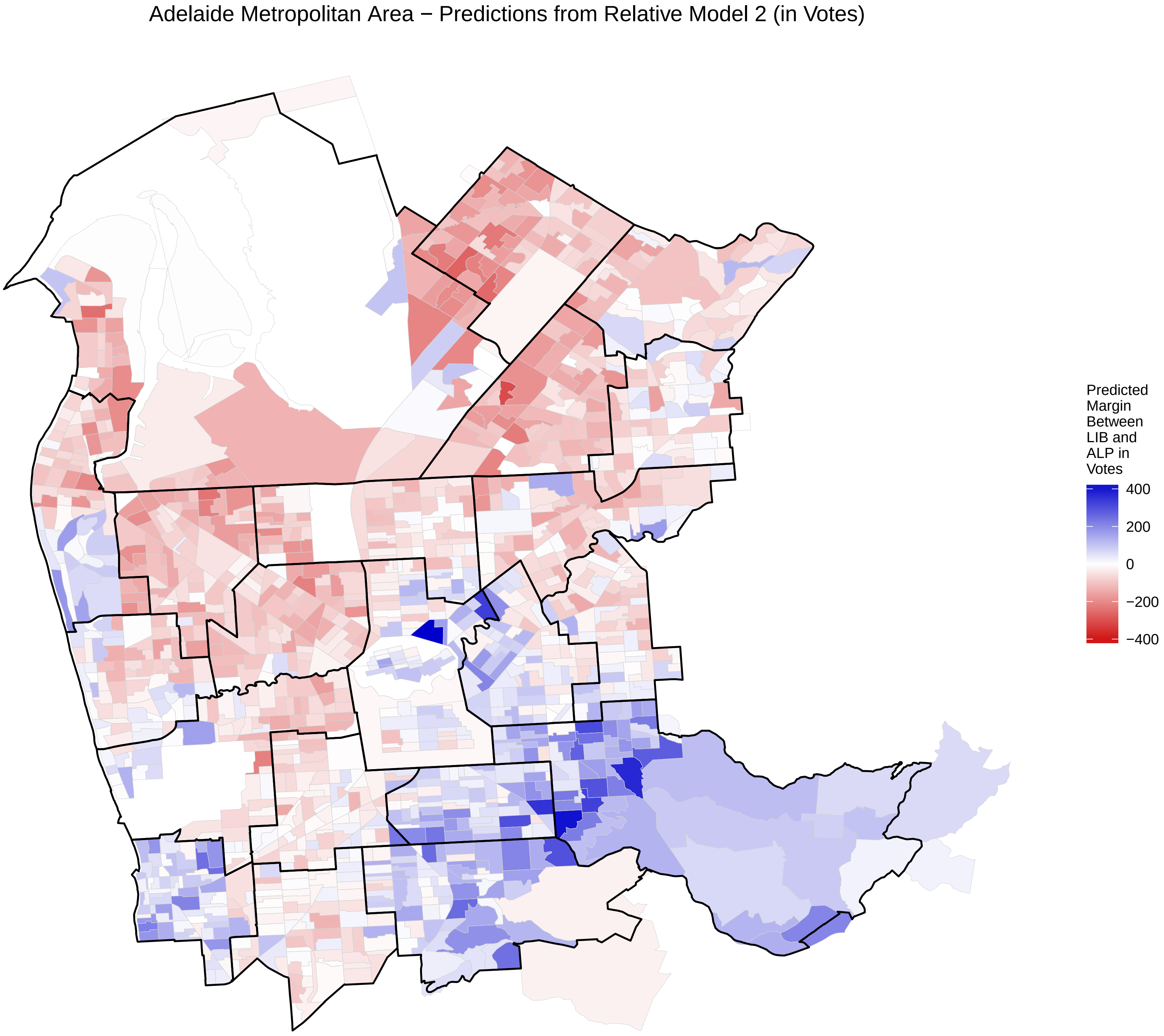

Adelaide Metropolitan Area − Predictions from Relative Model 2 (in Votes)
Predicted Margin Between LIB and ALP in Votes
400
200
0
−200
−400

Figure C.3: Map showing the size of the margin, in votes, between the Liberal Party and the ALP under Model 6.2. 21 electoral districts in metropolitan Adelaide are included.

Adelaide Metropolitan Area − Predictions from EDBC Method (in Votes)

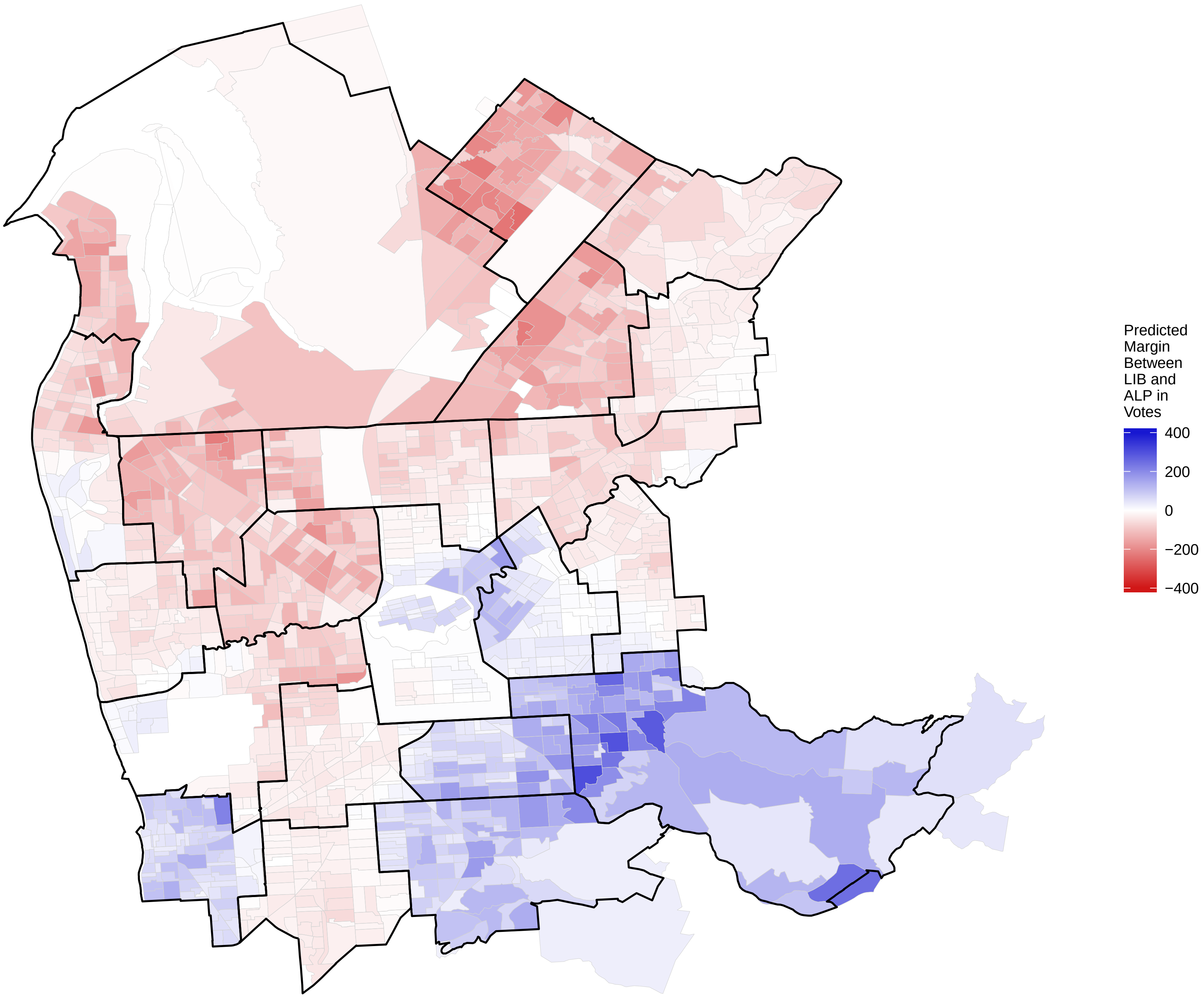


Figure C.4: Map showing the size of the margin, in votes, between the Liberal Party and the ALP according to the EDBC predictions. 21 electoral districts in metropolitan Adelaide are included.

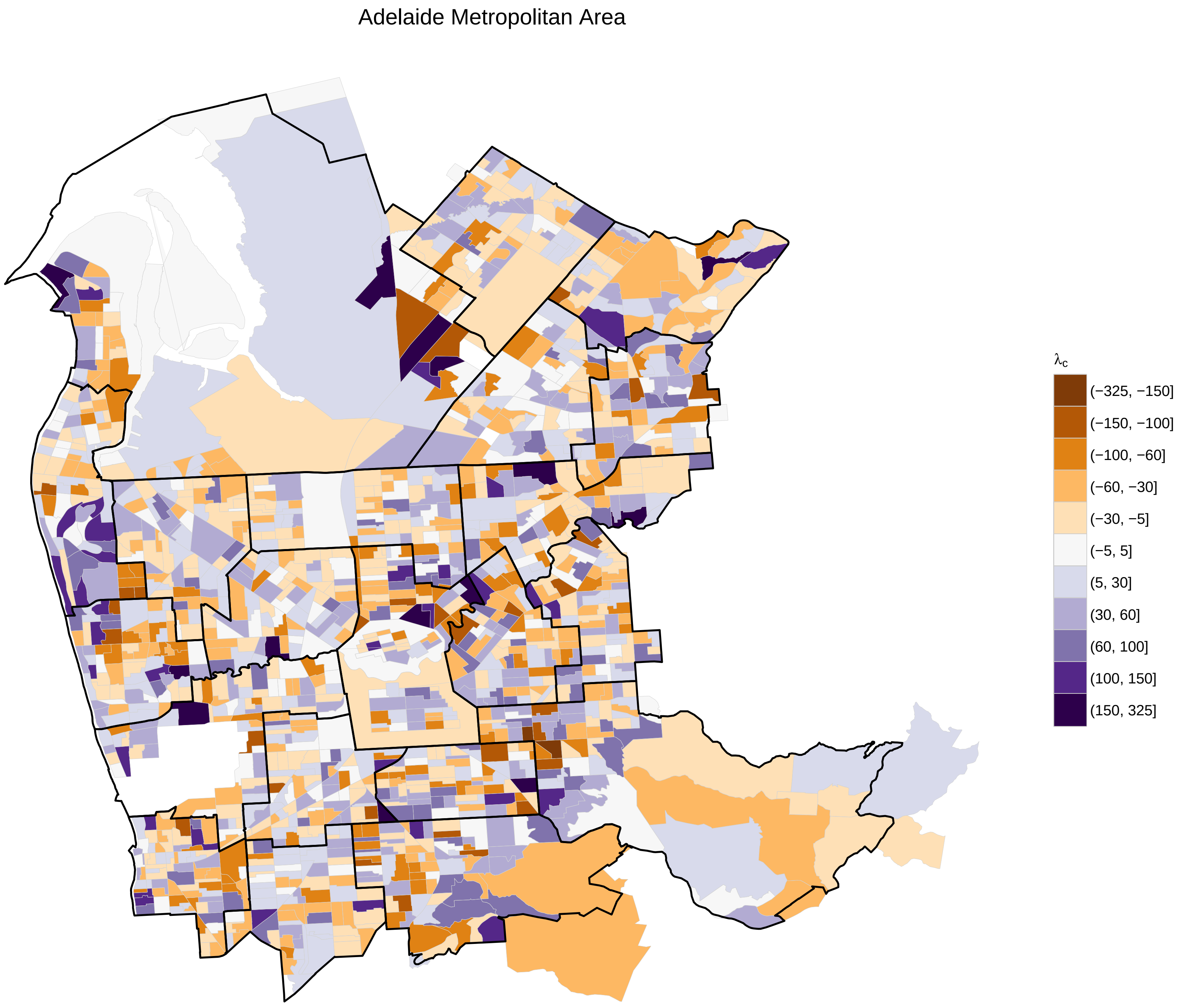


Figure C.5: Map showing the value of λ_c for $c \in CD$ (as defined in Equation 7.0.1), or in other words, the difference between the margin under Model 6.2 and the margin according to the EDBC. 21 electoral districts in metropolitan Adelaide are included.