Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016

Jeremy Forbes*, Dianne Cook, Rob J Hyndman

Department of Econometrics & Business Statistics, Monash University, Australia

Summary

We examine the relationships between electoral socio-demographic characteristics and two-party preferences in the six Australian federal elections held between 2001 and 2016. Socio-demographic information is derived from the Australian Census which occurs every five years. Since a census is not directly available for each election, an imputation method is employed to estimate census data for the electorates at the time of each election. This accounts for both spatial and temporal changes in electoral characteristics between censuses. To capture any spatial heterogeneity, a spatial error model is estimated for each election, which incorporates a spatially structured random effect vector. Over time, the impact of most socio-demographic characteristics that affect electoral two-party preference do not vary, with age distribution, industry of work, incomes, household mobility and relationships having strong effects in each of the six elections. Education and unemployment are amongst those that have varying effects. All data featured in this study has been contributed to the eechidna R package (available on CRAN).

Keywords: federal election, census, Australia, spatial modelling, imputation, data science, socio-demographics, electorates, R, eechidna

*Corresponding author. Email: jeremyforbes1995@gmail.com

1. Introduction

Australia has changed in many ways over the last two decades. Rising house prices, country-wide improvements in education, an ageing population, and a decline in religious affiliation, are just a few facets of the country's evolving socio-demographic characteristics. At the same time, political power has moved back and forth between the two major parties. In the 2007 and 2010 federal elections, the Australian Labor Party (hereafter Labor) was victorious, whereas the 2001, 2004, 2013 and 2016 elections were won by the Liberal National coalition (hereafter Liberal). The two-party preferred vote, a measure of support between these two parties, fluctuated between 47.3% and 53.5% (in favour of the Liberal party) over this period. This study explores how electoral (aggregate) socio-demographic characteristics relate to two-party preferences, and whether their effects have changed over time.

The Australian Electoral Commission (AEC) divides Australia into 150 regions, called electorates, with each corresponding to a single seat in the House of Representatives. If a party wins a majority of seats, they become the governing party. Data on the socio-demographics of these electorates are derived from the Australian Census, and vote counts are obtained from Australian federal elections. Joining these two data sources is problematic as there is an inherent asynchronicity in the two types of events. A census is conducted by the Australian Bureau of Statistics (ABS) every five years, whereas federal elections, conducted by the AEC, usually occur every three years or so. The first problem addressed is that of constructing appropriate census data for the 2004, 2007, 2010 and 2013 elections—election years in which a census did not occur. The predominant approach in previous studies was to join voting outcomes to the nearest census, without accounting for any temporal differences (see Davis & Stimson 1998; Stimson, McCrea & Shyy 2006; Liao, Shyy & Stimson 2009; Stimson & Shyy 2009). Furthermore, electoral boundaries change regularly, so spatial discrepancies also arise when matching with electoral data. To obtain appropriate 'census-like' data for these four elections, electoral socio-demographics are constructed using a spatio-temporal imputation that combines areal interpolation (Goodchild, Anselin & Deichmann 1993) and linear time-interpolation. Collecting and wrangling the raw data, along with the imputation process, are detailed in Section 2. All data and associated documentation relating to this procedure are available in the eechidna R package (Forbes et al. 2019), providing a resource for any future analysis.

Previous work on modelling Australian federal elections has found that aggregate socio-demographics are relatively good predictors of voting outcomes. Forrest et al. (2001) used multiple regression to model the Liberal and Labor primary vote for polling booths in the Farrer electorate in 1998 as a function of census variables from 1996. Stimson, McCrea &

Shyy (2006), Stimson & Shyy (2009) and Stimson & Shyy (2012) used principal component analysis of polling booths in the 2001, 2004 and 2007 elections respectively, also finding that socio-demographic characteristics of polling booths are linked to their two-party preferred vote. In contrast, Stimson & Shyy (2009) models the polling booth swing vote (change in the two-party preferred vote) in the 2007 election, finding that little of the swing vote can be explained by census data.

Instead of analyzing a single election in isolation, this paper employs a consistent model framework across six elections so that temporal changes in the effects of socio-demographics can be observed. Each federal election is modelled with a cross-sectional dataset, where each observation is one of the 150 electorates. This dataset consists of the two-party preferred vote (as the response variable) and a set of common socio-demographic variables (as the explanatory variables). To prepare these datasets, socio-demographic variables are first standardised, and then a principal component analysis is used to group many of the variables into 'factors'. To account for the inherent spatial structure of the data, a spatial error model is then estimated for each election. In interpreting these models, it is important to be mindful of the ecological fallacy. Insights are being drawn at the electorate level and cannot be inferred for another disaggregate level (in particular, drivers of individual voter behaviour may vary from what is observed at the electorate level).

The paper is organised as follows. Section 2 describes the data collection, joining and cleaning, while model details are discussed in Section 3. Section 4 describes the inference conducted to determine significance of effects and how these change over time, as well as including details on model robustness. Section 5 summarises the work. Two supplementary sections document the contributions of others to this work and the software.

2. Data collection, wrangling and imputation

2.1. Collecting the data

The voting outcome of interest is the electoral two-party preferred vote, which is provided by the Australian Electoral Commission (AEC) for the 2001, 2004, 2007, 2010, 2013 and 2016 elections via the AEC Tally Room. The AEC divides Australia into 150 regions, called electorates, with each corresponding to a single seat in the House of Representatives. Voting is compulsory in Australia, and each voter assigns a numbered preference to each available candidate in their electorate. The two-party preferred vote is determined by a tally of these preferences where, by convention, only the ranks of the Labor and Liberal candidates are considered. This is recorded as a percentage preference in favour of the Liberal party.

Socio-demographic variables are derived from the Australian Census of Population and Housing (census), which is a survey of every household in Australia, recording information such as age, gender, ethnicity, education level and income. There have been four censuses so far in the 21st century, conducted in 2001, 2006, 2011 and 2016. The Australian Bureau of Statistics (ABS) conducts the census and publishes aggregated information. The ABS uses electoral boundaries as defined by the AEC at the time of each census, which may not match those in place at the subsequent and previous elections. From the available census information aggregated at the electorate level, 50 socio-demographic variables are defined for each of the electorates to be used in the analysis. These variables include information relating to electoral age distributions, income, education qualifications, employment industries and job types, religion, birthplace, household characteristics and relationships.

Raw data is sourced online from the AEC and ABS websites in .csv and .xlsx files. The formats of these files differ over the years, making extracting the appropriate information a big task. The functions available in the dplyr (Wickham et al. 2019b) and readxl (Wickham et al. 2019a) R packages are particularly useful, as they provide fast consistent tools for data manipulation and functions to import .xlsx files. The 2001 and 2006 census data are published in a format where the information for each electorate is held in a separate document making it difficult to use the dplyr tools. Instead, cells have to be selected from each individual file to construct the desired variables. All scripts required for the data wrangling process can be found in the github repository for the eechidna R package (Forbes et al. 2019), along with the raw data. The eechidna package makes this study entirely reproducible and provides a resource to help wrangle data for future censuses and elections, when they become available.

2.2. Joining census and election data

Differences between census and election data

Between 2001 and 2016 there were six elections and four censuses (see Figure 1). Electoral boundaries are redistributed regularly by the AEC, meaning that only in the years where both a census and an election occur are all boundaries likely to match—the case for the 2001 and 2016 elections. Therefore, for the four elections between 2004 and 2013, both temporal and spatial differences in electorates need to be accounted for when joining the electoral two-party preferred vote with census data. For these elections a spatio-temporal imputation method is employed to obtain electoral socio-demographics. This method uses census information from both before and after the election of interest.

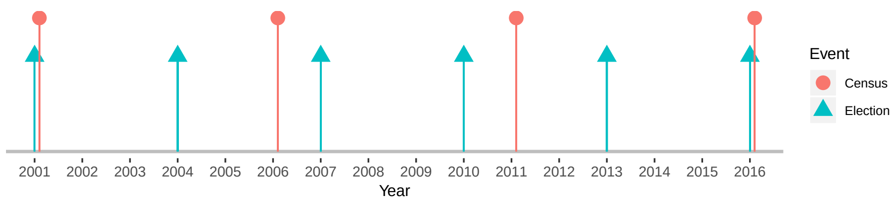

Figure 1. Timeline of Australian elections and censuses. They do not always occur in the same year.

Spatio-temporal imputation

For each election, neighbouring census information has to be combined in some way so that it represents the boundaries in place at the time of the election. This is done by taking the electoral boundaries and imputing the corresponding socio-demographic characteristics for each of the neighbouring censuses, thereby addressing the spatial aspect. Next, to deal with the temporal component, characteristics at the time of the election are constructed using linear interpolation between the spatially imputed neighbouring census variables.

The finest level of disaggregation available for census data is the region classification called Statistical Area 1 (SA1). In 2016, Australia was divided into over 55,000 SA1s. Consider each of these SA1 regions as a source zone, $s = 1, \ldots, S$, for which socio-demographic information is available. For simplicity, let each source zone be wholly summarised by its centroid. A set of target zones, $t = 1, \ldots, T$, are defined as regions for which information is to be imputed—these are the electoral boundaries for a particular election.

Take the example of the Melbourne Ports electorate from the 2013 federal election, illustrated in Figure 2. The purple region in this figure represents the target zone and the source zones are the centroid locations from the 2016 census SA1 areas.

Furthermore, let $I_{s,t}$ be an indicator variable, for which $I_{s,t} = 1$ if the centroid of source zone s falls within target zone t, and 0 otherwise. Additionally, let the population of the source zone s be P_s.

In order to calculate socio-demographic information for each of the target zones, a weighted average of source zones is taken using their populations as weights. Denote a given census variable for the target zone by C_t, and the same census variable for the source zone as D_s.

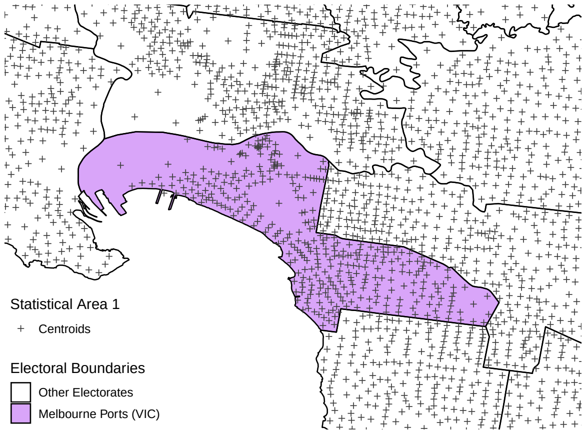

Figure 2. The electoral boundaries for Melbourne Ports (shown in purple) and surrounding electorates, with centroids for Statistical Area 1 regions from the 2016 census overlaid. The centroids falling within the purple region are attributed to Melbourne Ports.

Then, estimate C_t using

$$\hat{C}_t = \frac{\sum_{s=1}^{S} I_{s,t} \times D_s \times P_s}{\sum_{s=1}^{S} I_{s,t} \times P_s}, \quad \text{for each } t = 1, \ldots, T.$$

This concludes the spatial imputation of the socio-demographic characteristics for one target zone (a single electoral boundary), at the time of only one of the neighbouring censuses. This process is repeated for all of the target zones, and then for the other neighbouring census.

To account for temporal changes, linear interpolation is used between census years to get the final estimate of a census variable for the target zone in the election year. Let y_1 be the year of the census preceding an election, let y_2 be the year of the election, and y_3 be the year of the census that follows. Add this year subscript to the census variable estimate $\hat{C}_t$, resulting in $\hat{C}_{t,y}$. Linear interpolating between these census years results an imputed value for the election year, given by

$$\hat{C}_{t,y_2} = \frac{y_3 - y_2}{y_3 - y_1}\hat{C}_{t,y_1} + \frac{y_2 - y_1}{y_3 - y_1}\hat{C}_{t,y_3}.$$

Implicitly this assumes that population characteristics change in a linear manner over time.

Continuing with the example of Melbourne Ports in the 2013 election, the estimate for a given census variable in 2016, $\hat{C}_{\text{MelbPorts},2016}$ would be obtained by computing the weighted average of this variable amongst the SA1s within the purple region shown in Figure 2. This would be repeated with the 2011 census SA1s to obtain $\hat{C}_{\text{MelbPorts},2011}$, from which the final estimate is given by

$$\hat{C}_{\text{MelbPorts},2013} = \frac{3}{5}\hat{C}_{\text{MelbPorts},2011} + \frac{2}{5}\hat{C}_{\text{MelbPorts},2016}.$$

This is done for each of the socio-demographic variables, and is repeated for each of the 149 remaining target zones corresponding with 2013 electorates.

3. Modelling

From this imputation process, electoral socio-demographic variables are available for each of the six elections and can be joined with their corresponding two-party preferred votes. Before choosing an appropriate model, two issues with the socio-demographic variables need to be addressed. First, variable scales change over the years, making it important to standardise variables. Second, many variables represent similar information and where appropriate, will be grouped together. To determine which variables should be grouped, principal component analysis (PCA) is used to guide the construction of specific factors. The intuition here is that PCA will identify which variables covary, from which intuitive groupings of variables can be chosen to combine into individual variables. Details are given in Section 3.2. After these steps, a model specification is chosen.

3.1. Standardizing variables

Many of the socio-demographic variables have changing scales over the years. For example, inflation-adjusted median rental prices increased across almost all electorates, with median rent of 225 dollars per week placing an electorate in the 90th percentile in 2001, but only the 45th percentile in 2016. In order for socio-demographic effects to be comparable across years, all explanatory variables are standardised to have mean zero and variance one within each election year. By standardizing, each variable is reported as a relative measure compared to all other electorates in the same year. (Note that the log values were standardised for the variables `Judaism`, `Indigenous`, `Islam` and `Buddhism`.)

3.2. Creating factors

There are only $N = 150$ observations (electorates) in each election and $p = 50$ socio-demographic variables in each cross-section, with many variables representing similar information about an electorate. Any model that uses all variables would face problems with multi-collinearity and over-fitting, which would likely lead to erroneous conclusions regarding variable significance. To address this, a subset of variables that represent similar information are combined into a single variable, which will be referred to as a 'factor'.

A factor is created from a group of variables if there is an intuitive reason as to why these variables should represent similar information and if there is evidence to suggest that they covary. For example, a potential group would be variables relating to electoral incomes—median family, household and personal incomes. To determine which variables covary, principal component analysis is used on a combined dataset of socio-demographic variables from all six elections. The only variables exempted from the principal component analysis are the four variables representing age brackets (the proportion of the population aged 0–19 years old, 20–34 years old, 45–54 years old and 55 years plus), which are included in the model as separate variables.

Only the first four principal components from the combined dataset are considered, as the scree plot levels off after the fourth component. Variables that have a large loading in a particular component are deemed to covary, with a loading with magnitude greater than 0.15 being considered large. Each principal component is considered separately. If a subset of variables have large loadings (positive or negative) in a given component, and there is an intuitive reason as to why they should be grouped together, then this subset of variables will be combined to become a factor. Note that more than one factor can be deduced from a principal component (i.e. multiple non-overlapping subsets of variables), and that any variables not included in a factor are not discarded.

Six factors are created using this approach. These are: `Incomes` (median personal income, median household income, median family income); `Unemployment` (unemployment rate, labour force participation rate); `PropertyOwned` (proportion of dwellings that are owned, proportion of dwellings that are mortgages, proportion of dwellings that are rented, proportion of dwellings that classified as government housing); `RentLoanPrice` (median rental payment amount, median loan repayment amount); `FamHouseSize` (average household size, ratio of people to families, incidence of single person households, incidence of households containing a couple with kids, incidence of households containing a couple without kids); and `Education` (high school completions, undergraduate and postgraduate degrees, proportion of employed people working as professionals, proportion of jobs in finance, proportion of workers who are labourers, proportion of workers who work as a tradesperson, diploma and certificate qualifications).

For each of these groupings, a factor is created by taking a weighted sum of the variables. The weightings are allocated on the basis of whether the variable had a positive or negative loading in the principal component from which the grouping was identified. Variables with a positive loading are allocated a weight of $+1$ and those with negative loadings are allocated a weight of -1. After computing these weighted sums, the factor is standardised to have mean zero and variance one, within each election.

The final predictor set contains $p = 32$ variables which are listed in Table 1. (Note that the factor creation procedure reduces the variable set to $p = 33$, however the `Pop_55_plus` age bracket is not included as a variable to avoid multicollinearity, because the other three age brackets are included.)

3.3. Regression incorporating spatially dependent errors

An identical model specification is used for each of the six elections, with each election modelled separately. Separate models are preferred to a single model because of how frequently electoral boundaries change, noting that electorates with the same name across elections are not guaranteed to represent the same geographic region. Therefore any fixed or random effects models would be difficult to estimate without implementing consistent boundaries, which would require further imputation (of voting information). The separate models also allow the socio-demographic effects to be estimated separately for each election year, facilitating analysis of temporal changes in variable effects. This can be considered a special case of a longitudinal model where all coefficients are time-varying and heteroskedasticity is time-varying.

For each cross-section, let the response $\boldsymbol{y}$ be the vector two-party preferred vote in favour of the Liberal party; for example, $y_i = 70$ represents a 70% preference for Liberal, 30% for Labor, in electorate i. Although y_i lies in the interval $(0, 100)$, observed values are never close to 0 or 100 (minimum 24.05% and maximum 74.90%), so there is no need to formally impose the constraint of $y_i \in [0, 100]$. Furthermore, the responses are found to be spatially correlated in each election (Moran's I test, $p \leq 7 \cdot 10^{-15}$). This is not surprising as electorates are aggregate spatial units, and hence the spatial structure of the data must be modelled appropriately.

The spatial error model (Anselin 1988) is chosen because it captures spatial heterogeneity by incorporating a spatially structured random effect vector (LeSage, Kelley Pace & Pace 2009). In this context, the random effect can be thought of as capturing the effect of any characteristics that neighbourhoods share that have not been addressed by the independent variables included in the model.

Spatial weights are calculated in accordance with the assumption that an electorate is equally correlated with any electorate that shares a part of its boundary. Let ρ be the spatial autoregressive coefficient, $\boldsymbol{v}$ be a spherical error term, $\boldsymbol{W}$ be a matrix of spatial weights (containing information about the neighbouring regions), $\boldsymbol{X}$ be a matrix of socio-demographic

Table 1. Estimated spatial error model parameters (standard errors) for each of the six election years.

	2001	2004	2007	2010	2013	2016
ρ	0.53^{***}	0.33^{**}	0.21	0.17	0.27	0.39^{**}
	(0.15)	(0.16)	(0.18)	(0.17)	(0.17)	(0.17)
AusCitizen	-3.94^{*}	-1.39	-2.18	-1.28	-3.89	-2.66
	(2.27)	(2.44)	(2.21)	(2.69)	(2.51)	(2.61)
Pop_00_19	0.49	2.66	9.39^{***}	5.25	3.31	0.88
	(2.54)	(3.91)	(3.63)	(3.64)	(2.91)	(2.62)
Pop_20_34	-8.04^{***}	-7.72^{***}	-8.34^{***}	-11.68^{***}	-9.29^{***}	-9.21^{***}
	(1.80)	(2.21)	(2.18)	(2.90)	(2.62)	(2.37)
Pop_35_54	-2.64^{***}	-2.78^{***}	-3.62^{***}	-3.13^{***}	-2.76^{**}	-2.13^{**}
	(0.84)	(0.89)	(0.83)	(1.10)	(1.11)	(1.06)
BornAsia	3.58^{*}	-1.09	0.66	-1.78	-1.08	-0.14
	(2.09)	(2.52)	(1.99)	(2.74)	(2.54)	(2.17)
BornMidEast	-1.02	-1.75	-0.98	-1.00	-1.66	-1.31
	(1.00)	(1.17)	(1.09)	(1.33)	(1.23)	(1.11)
BornSEEuro	-1.63	-3.17^{*}	-1.07	-2.04	-2.89^{***}	-2.53^{***}
	(1.37)	(1.68)	(1.06)	(1.29)	(1.11)	(0.97)
BornUK	0.29	0.31	0.32	0.28	-0.15	-0.61
	(1.02)	(1.04)	(0.87)	(1.06)	(0.99)	(0.99)
BornElsewhere	-4.13	-1.51	-1.03	2.45	-4.21	-2.17
	(3.14)	(3.62)	(3.18)	(4.13)	(3.90)	(3.76)
Buddhism	-0.07	0.80	0.58	-0.14	-0.43	-1.16
	(1.31)	(1.54)	(1.39)	(1.66)	(1.60)	(1.58)
Christianity	-1.70	-1.01	-0.45	0.13	2.03	3.76^{**}
	(1.62)	(1.75)	(1.60)	(1.85)	(1.68)	(1.83)
CurrentlyStudying	-2.20^{*}	-0.01	-0.14	1.35	0.32	0.22
	(1.22)	(1.50)	(1.39)	(1.41)	(1.35)	(1.56)
DeFacto	-3.24	-2.25	-4.67^{**}	-7.75^{**}	-7.82^{**}	-10.39^{***}
	(2.07)	(2.62)	(2.27)	(3.09)	(3.08)	(3.15)
DiffAddress	3.06^{***}	2.75^{**}	0.73	2.55	2.27	5.20^{***}
	(0.94)	(1.20)	(1.24)	(1.79)	(1.67)	(1.51)
Distributive	1.60	1.89^{*}	0.50	0.62	1.59	1.31
	(1.06)	(1.14)	(0.99)	(1.27)	(1.20)	(1.18)
Education	-0.37	-0.26	-6.72^{**}	-7.31^{*}	-7.31^{**}	-8.55^{**}
	(2.35)	(3.34)	(3.00)	(3.90)	(3.63)	(3.37)
Extractive	3.74^{***}	4.96^{***}	4.64^{***}	6.46^{***}	5.97^{***}	6.38^{***}
	(1.43)	(1.47)	(1.20)	(1.45)	(1.35)	(1.38)
FamHouseSize	1.94	-2.55	-6.47^{**}	-3.84	-3.12	-2.00
	(2.61)	(3.66)	(3.28)	(3.87)	(3.52)	(3.06)
Incomes	4.36^{***}	2.42	5.52^{**}	5.63^{*}	8.02^{***}	12.70^{***}
	(1.69)	(3.00)	(2.42)	(3.15)	(2.78)	(2.64)
Indigenous	1.26	1.96	2.41	2.38	0.46	-0.22
	(1.61)	(1.89)	(1.59)	(2.00)	(1.88)	(1.90)
Islam	-0.75	-0.91	-0.60	-2.01	-0.88	-1.09
	(1.14)	(1.28)	(1.14)	(1.41)	(1.26)	(1.30)
Judaism	1.32	0.93	1.47	0.28	1.35	1.15
	(1.01)	(1.08)	(0.92)	(1.10)	(1.02)	(0.97)
ManagerAdmin	2.62^{***}	4.67^{***}	7.47^{***}	7.05^{***}	5.93^{***}	5.64^{***}
	(0.67)	(1.06)	(0.95)	(1.16)	(1.06)	(0.97)
Married	-3.93	-2.72	-9.35^{***}	-10.12^{***}	-7.91^{**}	-9.47^{**}
	(2.51)	(3.56)	(3.12)	(3.55)	(3.57)	(3.85)
NoReligion	-0.73	0.04	1.32	0.37	1.41	2.94
	(1.50)	(1.65)	(1.51)	(1.75)	(1.74)	(2.03)
OneParentHouse	-4.77^{***}	-3.23	-6.55^{***}	-7.03^{***}	-5.32^{***}	-4.94^{**}
	(1.49)	(1.99)	(1.81)	(2.04)	(1.97)	(2.03)
OtherLanguage	-1.02	6.88	6.21	7.80	10.13^{**}	9.98^{**}
	(3.00)	(4.93)	(3.97)	(5.25)	(5.09)	(4.26)
PropertyOwned	-2.01	-0.30	0.74	-1.92	-1.05	0.73
	(1.35)	(1.49)	(1.36)	(1.74)	(1.67)	(1.48)
RentLoanPrice	-2.17	0.37	1.23	3.08	1.36	-2.04
	(1.46)	(1.93)	(1.76)	(2.23)	(2.20)	(2.07)
SocialServ	3.31^{***}	2.85^{**}	3.46^{***}	3.72^{**}	2.98^{**}	4.04^{***}
	(1.27)	(1.40)	(1.17)	(1.46)	(1.28)	(1.15)
Transformative	2.30	4.71^{***}	4.58^{***}	4.55^{**}	3.63^{**}	4.05^{***}
	(1.48)	(1.77)	(1.51)	(1.87)	(1.67)	(1.47)
Unemployment	-3.39^{**}	-3.47^{**}	-0.40	-0.68	0.81	1.93
	(1.37)	(1.69)	(1.45)	(1.80)	(1.47)	(1.32)
Constant	50.80^{***}	52.63^{***}	47.31^{***}	49.92^{***}	53.52^{***}	50.46^{***}
	(0.76)	(0.59)	(0.44)	(0.52)	(0.54)	(0.64)
Residual Standard Error (GLS)	4.34	4.82	4.32	5.30	4.82	4.76
Observations	150	150	150	150	150	150

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

covariates, $\boldsymbol{\beta}$ be a vector of regression coefficients and $\boldsymbol{a}$ be a spatially structured random effect vector.

Let

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{a},$$

and

$$\boldsymbol{a} = \rho\boldsymbol{W}\boldsymbol{a} + \boldsymbol{v},$$

where $\boldsymbol{v} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I_n})$, and hence

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I}_n - \rho\boldsymbol{W})^{-1}\boldsymbol{v}.$$

Estimation of the above spatial error model is undertaken using feasible generalised least squares.

Table 1 details the estimated model coefficients and their estimated standard errors, for each of the six elections. An interpretation of these estimated values is provided in the next section.

4. Results

4.1. Spatial autoregressive parameter

The spatial autoregressive coefficient ρ was positive and significant in the 2001, 2004 and 2016 elections (Figure 3). In these three elections, there is evidence to suggest that neighbours shared some influential characteristics outside the explanatory variables, which affected the two-party preferred vote. Conversely, in the other three elections, the spatial effect was weaker and insignificant (although still positive).

4.2. Country-wide trend

Since all socio-demographics were standardised to have a mean of zero and a variance of one, the intercept in each model can be interpreted as the estimated two-party preferred vote for an electorate with mean characteristics (aside from `Judaism`, `Indigenous`, `Islam` and `Buddhism`, where it assumes the mean of the log value). Figure 4 shows that the baseline of party preference varied over the elections, with the biggest swing occurring in the 2007 election where the mean electorate shifted more than five percentage points in favour of the Labor party.

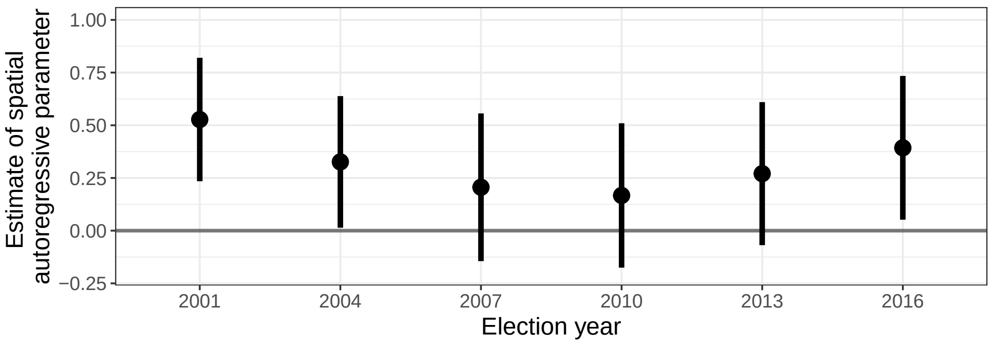

Figure 3. Estimates of the spatial autoregressive parameter for each of the six elections, reported with their individual 95% confidence intervals. In 2001, 2004 and 2016 there was a significant spatial component.

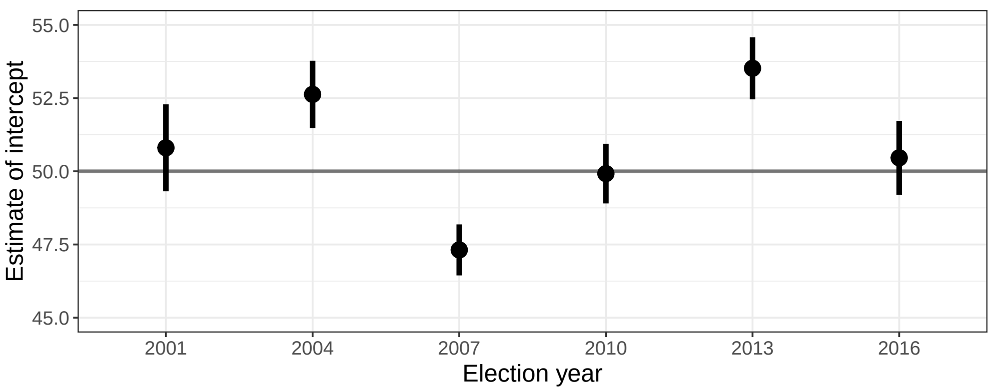

Figure 4. Estimated intercept for each election, which represents the two-party preferred vote for an electorate with mean characteristics.

4.3. Influential socio-demographics

To investigate the socio-demographics that had a strong effect on the two-party preferred vote, partial residual plots were used and shown in Figures 5 and 6. Partial residuals, for a given variable, are the residuals from the fitted model with the estimated effect of that variable added to it. These plots show the direction, size and significance of an estimated effect, as well as any deviations from linearity. In each plot, the slope of the prediction line matches the estimated coefficient and the shaded region represents a 95% confidence band. Plots were computed using the method in Breheny & Burchett (2017). If a horizontal line can be drawn through the confidence band, then the effect was insignificant. The estimated intercept was also added to the partial residuals for interpretability. Plots for each election are faceted in Figures 5 and 6 to compare the effects over time. Only socio-demographics that had a significant effect in at least two elections are displayed.

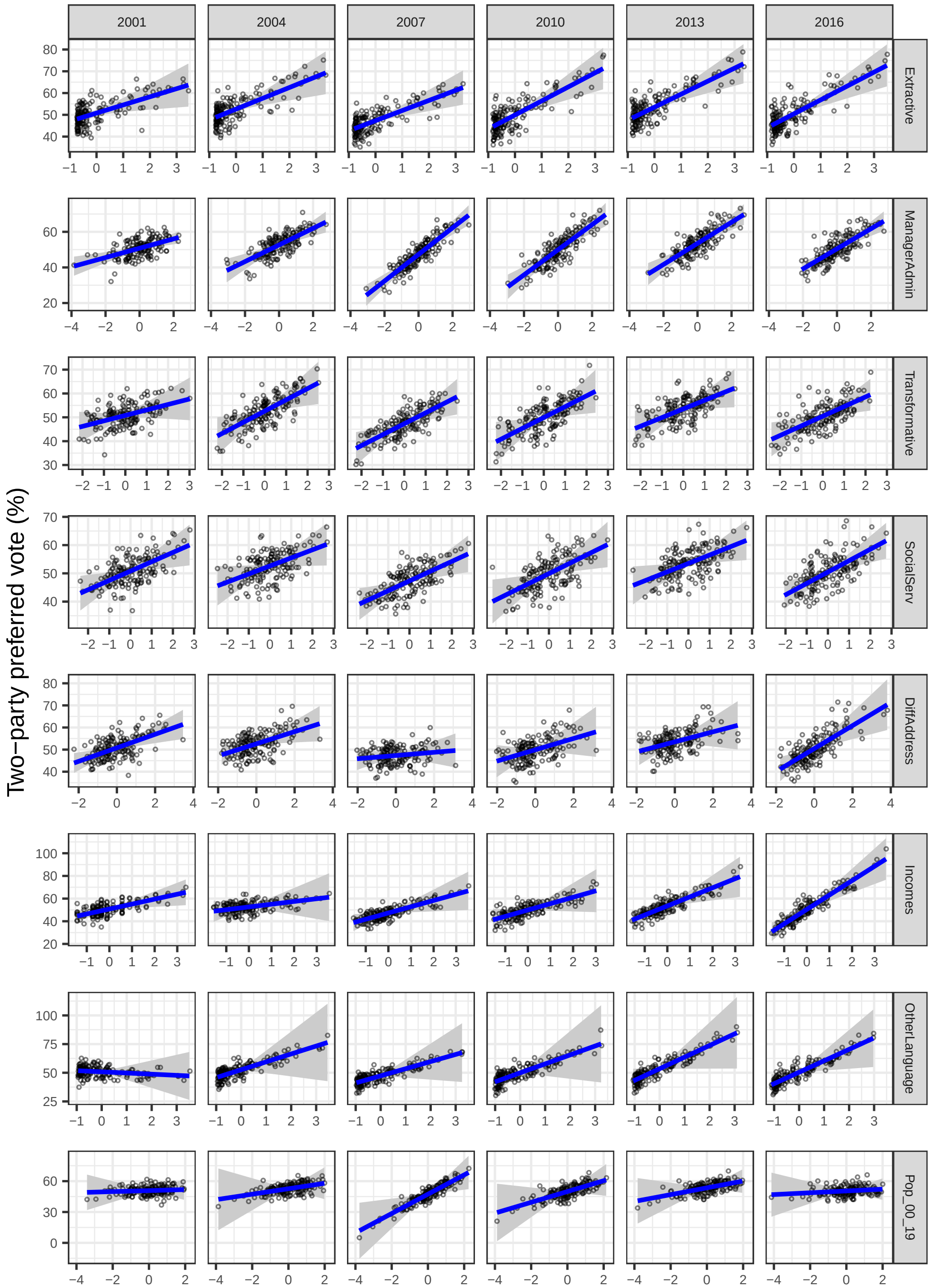

Figure 5. Partial residual plots by election year for a selection of predictors. Linear model with 95% confidence bands overlaid. Most predictors had a positive relationship: the larger the value the more likely the electorate preferenced Liberal. The relationships were relatively robust over time, with the exception of `DiffAddress`, `Incomes`, `OtherLanguageHome` and `Pop_00_19`.

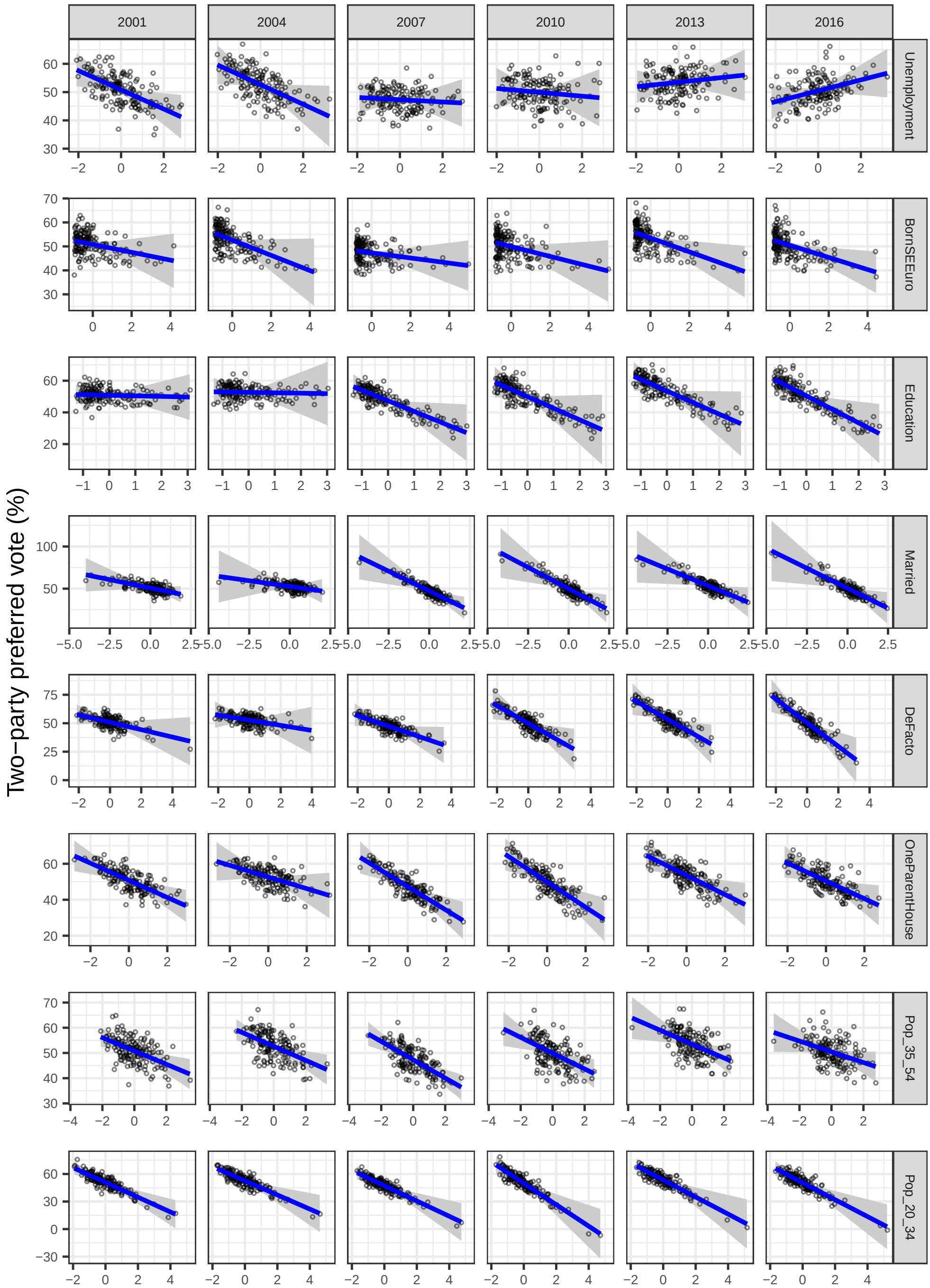

Figure 6. Partial residual plots by election year for a selection of predictors. Linear model with 95% confidence bands overlaid. Several predictors had a negative relationship: the larger the value the more likely the electorate preferenced Labor. Most relationships were relatively stable over elections, except `Unemployment` and `Education`.

Industry and type of work

Electorates with higher proportions of workers in mining, gas, water, agriculture, waste and electricity (grouped as `Extractive` industries) were consistently linked with higher support for the Liberal party, with the magnitude of this effect slightly increasing over the years (see row 1 in Figure 5). This is unsurprising, as the Liberal party maintained close ties with these traditional energy industries, and typically presented policies to reduce taxation on energy production. Furthermore, electorates with more workers in construction or manufacturing industries (`Transformative`) were also more likely to support the Liberal party (see row 3 in Figure 5), from 2004 onwards.

Similarly, the proportion of workers in managerial, administrative, clerical and sales roles (`ManagerAdmin`), was also a significant predictor of two-party preference vote across all six elections, where higher proportions of people working these jobs increased Liberal support.

Of these job related variables, the most surprising effect is that associated with the proportion of workers in education, healthcare, social work, community and arts (`SocialServ`). Typically the Labor party has had more generous funding schemes affecting these areas of work, so one might expect `SocialServ` to have had a negative effect on two-party preference. However, in every election this effect was found to be positive and significant.

Income and unemployment

Typically the Labor party has campaigned on more progressive policies, often including tax reform that adversely affects higher income earners, and more generous social assistance programs. Perhaps it is due to these policies that higher income electorates were more likely to support the Liberal party, as the `Incomes` factor had a positive effect on Liberal preference (see row 6 in Figure 5). This effect was significant in every election aside from 2004 and 2010. `Unemployment` however, was not as influential. In 2001 and 2004, electorates with higher unemployment aligned with Labor, but over time this shifted towards support for the Liberal party, culminating in a positive (insignificant) effect in 2016.

Age

The older Australian population has often been considered to be more conservative, and the left leaning political parties (including Labor) have typically had a stronger appeal to younger people. This effect was indeed observed across all six elections, as electorates with higher proportions of people aged between 20 and 34 (`Pop_20_34`) aligned strongly with Labor preference (bottom row in Figure 6). Larger populations of 35 to 54 year olds (`Pop_35_54`) were also associated with Labor, but the magnitude of this effect was far smaller. Populations under 20 years of age was only significant in 2007, where `Pop_00_19` increased Liberal support.

Education

From 2007, electorates with higher education levels were associated with support for the Labor party, with this effect being significant in 2007, 2013 and 2016 and only marginally insignificant in 2010. In the elections before 2007, education had a negligible effect (see row 3 in Figure 6). Additionally, student populations (`CurrentlyStudying`) did not affect electoral party preference in any election (not shown).

Diversity

Larger migrant populations from Asia, the Middle East, South-Eastern Europe, the United Kingdom and elsewhere, were either associated with Labor support, or had no effect. Of these areas, only South-Eastern European populations significantly affected party preference, with larger populations associating with Labor in 2013 and 2016 (row 2, Figure 6). Speaking other languages (aside from English) however, appears to have had a far stronger effect, as observed through the `OtherLanguage` variable. Electorates with more diverse speech were linked with higher support for the Liberal party from 2004 onwards, with this effect being significant in 2013 and 2016 (see row 7, Figure 5). Furthermore, none of the variables relating to religious beliefs aside from Christianity had a material effect in any election (this includes the Buddhist, Muslim, Jewish, non-religious and Indigenous Australian populations). The relationship between Christian populations (`Christianity`) and the Liberal party strengthened over the years, becoming positive and significant in 2016.

Households

In 2001, 2004 and 2016, higher proportions of people that recently (in the past five years) moved house (`DiffAddress`) increased electoral support for the Liberal party (see row 5 in Figure 5). This was somewhat surprising as one might expect house ownership and rental prices to be linked to two-party preference, rather than household mobility (`PropertyOwned` and `RentLoan` were not significant in any election).

Higher proportions of single parent households were associated with Labor support in all elections (albeit insignificant in 2004, see row 6 in Figure 6), whereas family and household sizes (via the `FamHouseSize` variable) did not appear to be associated with either party.

Relationships

From 2007 onwards, the percentage of people in both marriages (`Married`) and de facto relationships (`DeFacto`) were found to be strong predictors of the two-party preferred vote in favour of the Labor party. In 2001 and 2004 neither of these variables were significant (see rows 4 and 5 in Figure 6).

4.4. A closer look at the residuals

Residuals by state

It is often hypothesised that states have had a systematic bias towards one of the two major parties. Boxplots of residuals grouped by state (Figure 7) showed that the data reflects this to only a limited extent. Tasmania and the Australian Capital Territory appeared to have a bias towards Labor, whereas the South Australia and the Northern Territory tended towards voting Liberal. However, there were relatively few electorates in each of these states (five, two, eleven and two respectively), so this apparent result may be due to incumbent effects rather than an actual state-specific bias.

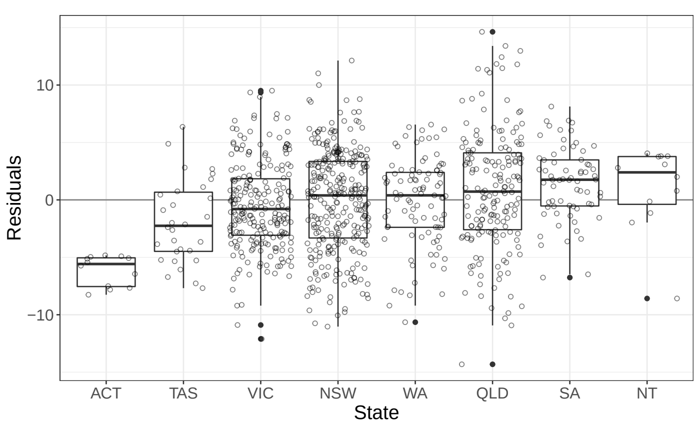

Figure 7. Boxplot of residuals by state with jittered points. States ordered by median residual. A state-specific bias present only in the smaller states appeared to have not been captured by the model.

Residuals by party incumbency

The incumbent party appeared to have a distinct advantage at the next election. The boxplots in Figure 8 show that if either of the Labor or Liberal parties won the seat at the previous election, the electorate was likely to vote in their favour at the subsequent election, over and above any socio-demographic effects—this effect has not been captured by the model.

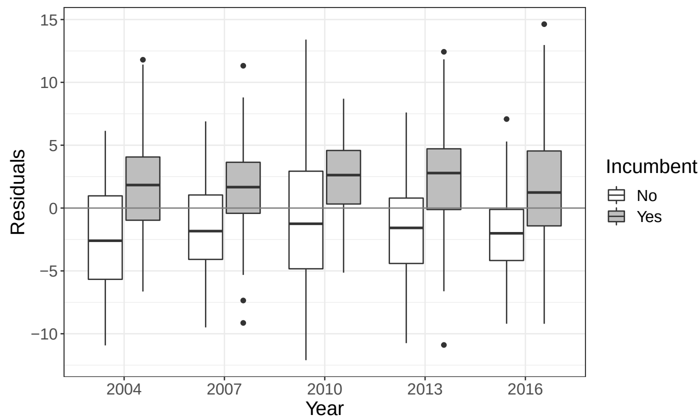

Figure 8. Boxplot of residuals for incumbent and non-incumbent parties each year. An incumbent advantage was evident and had not been captured by the model.

4.5. Robustness

Multicollinearity

Three robustness checks were conducted to confirm model stability. First, a model for each election was re-estimated using only the variables that were found to be significant in at least one of the six elections. The estimated coefficients of the variables in the re-estimated models all fell within their respective 95% confidence intervals from the full models. The second check involved the ten largest pairwise correlations. For each pair, a model for each election was re-estimated omitting one of the two variables. It was found that for each of these pairs, the estimated effect of the remaining variable in the reduced model was within the 95% confidence interval from the full model. The final check was a visual exploration of different variable projections using a tour (Wickham et al. 2011) for each election. No definitive signs of multicollinearity were observed, and as expected (given the nature of spatial data), there was some clumping of electorates for certain projections.

Influential and outlier electorates

Based on the distribution of the Cook's distance values and the distribution of hat values, a Cook's distance greater than 0.1 was considered to be influential, and a hat value greater than 0.5 was considered to have high leverage. Electorates fitting these criteria were flagged and investigated to examine the characteristics driving these values.

The electorate of Sydney (NSW) had a large Cook's distance and high leverage from 2001 to 2007, due to its diverse population (languages, birthplace and religion), high density of young adults (20 to 34 years old), high number of defacto relationships, high income, high household mobility and small amount of workers in extractive and transformative jobs. It remained a strong supporter of the Labor party and the extent of this support was underpredicted by the model, making it an outlier. Nearby in metropolitan NSW, the electorate of Wentworth was found to be an outlier in the 2013 and 2016 elections. Although historically Liberal, its two-party vote jumped by over 10 percentage points in 2010 without experiencing any notable changes in its socio-demographic makeup—implying that this may be the direct effect of its Liberal member, Malcolm Turnbull, becoming the leader of the Liberal party. In the elections that followed, the model underpredicted Wentworth's Liberal support.

Lingiari, an electorate making up almost all of the Northern Territory, had consistently high leverage (all years) and was an outlier in all but the 2013 election due to its large Indigenous population, low rates of property ownership and few workers in management or administrative jobs. Fowler (NSW) had a diverse population with a high proportion of migrants, many Buddhists and Muslims, as well as a high proportion of single parent households. These characteristics explain its high leverage in 2001, 2004, 2010 and 2013, and its strong Labor support made it influential in 2001, 2004 and 2010. Other electorates with large Cook's distance were Canberra (ACT) and Durack (WA) in 2013, and Solomon (NT) in 2016.

All of the electorates examined were not unduly influential in the model and therefore no action was required.

5. Conclusion

This paper explored the effects of electoral socio-demographic characteristics on the two-party preferred vote in the 2001–2016 elections, using information from the corresponding Australian federal elections and censuses. As a census did not always occur in the same year as an election, census data for each of the 2004–2013 elections were generated by employing a method of spatio-temporal imputation. This method imputes electoral socio-demographics for the electoral boundaries in place at the time of the corresponding election—an approach that is distinctly different from previous work on modelling election outcomes, where census and election data has typically been joined without addressing their temporal differences. Before estimating a model, these socio-demographic variables were standardised (to adjust for changing variable scales) and subsets of variables (representing similar information) were combined into factors, resulting in a reduced predictor set. A spatial error model was then estimated for each election, accounting for the inherent spatial structure of the data.

Across the past six elections, most of the socio-demographics driving the electoral two-party preferred vote were found to remain steady, whilst a few (typically weaker) effects varied over time. Industry and type of work were particularly influential. Energy-related and manufacturing/construction jobs, as well as administrative roles and jobs in education and social services were strongly linked with the Liberal party in all elections. Incomes had a similarly consistent effect, with higher income areas supporting Liberal. Higher levels of unemployment shifted from a weak association with Labor to a significant Liberal effect over the years, and higher education levels were associated with Labor from 2007 (although marginally insignificant in 2010). Electorates with large populations 20 to 34 years were strongly associated with Labor, whilst the 35 to 54 year old bracket also increased Labor support, but to a lesser extent. It was also found that birthplace diversity slightly favoured Labor, relationships (both marriages and de facto relationships) aligned with Labor preference from 2010 onwards, and the influence of Christian populations trended towards Liberal support whilst other religions had negligible effects. Family and household sizes had minimal influence, although electorates with more single parent households were linked with Labor support. Furthermore, the spatial effects were found to be positive in all elections and significant in 2001, 2004 and 2016, meaning that other characteristics that neighbours had in common (outside of the variables in the model) appeared to be influential in those years.

The findings in this paper complement the existing literature by modelling temporal trends, which as far as the authors are aware, has not been done previously for Australian elections using a regression framework. It is also the first study to model any Australian election since 2010 using census information.

Additionally, a key contribution of this research is the wrangling of raw data and imputation of data sets for the 2004, 2007, 2010 and 2013 elections, which have been contributed to the eechidna R package—providing a rich, accessible data resource for any future Australian electoral analysis.

6. Acknowledgements

This paper was produced using R Markdown (Allaire et al. 2019) and knitr (Xie 2015). All corresponding code for this paper can be found in the github repository github.com/jforbes14/eechidna-paper, and the data used is available in the eechidna package (Forbes et al. 2019). All raw data was obtained from the Australian Electoral Commission, the Australian Bureau of Statistics and the Australian Government.

The authors would like to sincerely thank the editor and associate editor of the Australian & New Zealand Journal of Statistics and the two anonymous reviewers for providing helpful comments and suggestions on earlier drafts of this manuscript. Additionally, the authors would like to thank Anthony Ebert, Heike Hofmann, Thomas Lumley, Ben Marwick, Carson Sievert, Mingzhu Sun, Dilini Talagala, Nicholas Tierney, Nathaniel Tomasetti, Earo Wang and Fang Zhou, all of whom have contributed to the eechidna package.

7. Software

All election and census datasets, along with electoral maps and more, are available in the eechidna (Exploring Election and Census Highly Informative Data Nationally for Australia) R package, which can be downloaded from CRAN. The eechidna package makes it easy to look at the data from the Australian Federal elections and censuses that occurred between 2001 and 2019. This study contributed a large revision to the eechidna package, which included the addition of election and census data for 2001–2010, voting outcomes for polling booths and imputed census data for election years. For more details on using eechidna, please see the articles (vignettes) on the github page: ropenscilabs.github.io/eechidna.

References

ALLAIRE, J., XIE, Y., MCPHERSON, J., LURASCHI, J., USHEY, K., ATKINS, A., WICKHAM, H., CHENG, J., CHANG, W. & IANNONE, R. (2019). *rmarkdown: Dynamic documents for R*. URL https://rmarkdown.rstudio.com. R package version 1.12.

ANSELIN, L. (1988). *Spatial econometrics: methods and models*, vol. 4. Dordrecht, Netherlands: Springer Science & Business Media.

BREHENY, P. & BURCHETT, W. (2017). Visualization of regression models using visreg. *The R Journal* **9**, 56–71. URL https://journal.r-project.org/archive/2017/RJ-2017-046/index.html.

DAVIS, R. & STIMSON, R. (1998). Disillusionment and disenchantment at the fringe: explaining the geography of the one nation party vote at the queensland election. *People and Place* **6**, 69–82.

FORBES, J., COOK, D., EBERT, A., HOFMANN, H., HYNDMAN, R.J., LUMLEY, T., MARWICK, B., SIEVERT, C., SUN, M., TALAGALA, D., TIERNEY, N., TOMASETTI, N., WANG, E. & ZHOU, F.

(2019). *eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia.* URL https://CRAN.R-project.org/package=eechidna. R package version 1.4.0.

FORREST, J., ALSTON, M., MEDLIN, C. & AMRI, S. (2001). Voter behaviour in rural areas: a study of the Farrer electoral division in southern New South Wales at the 1998 federal election. *Australian Geographical Studies* **39**, 167–182.

GOODCHILD, M.F., ANSELIN, L. & DEICHMANN, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* **25**, 383–397.

LESAGE, J., KELLEY PACE, R. & PACE, R.K. (2009). *Introduction to spatial econometrics.* Boca Raton, Florida: Chapman and Hall/CRC.

LIAO, E., SHYY, T. & STIMSON, R. (2009). Developing a web-based e-research facility for socio-spatial analysis to investigate relationships between voting patterns and local population characteristics. *Journal of Spatial Science* **54**, 63–88.

STIMSON, R., MCCREA, R. & SHYY, T. (2006). Spatially disaggregated modelling of voting outcomes and socio-economic characteristics at the 2001 australian federal election. *Geographical Research* **44**, 242–254.

STIMSON, R. & SHYY, T. (2012). And now for something different: modelling socio-political landscapes. *Annals of Regional Science* **50**, 623–643.

STIMSON, R. & SHYY, T.K. (2009). A socio-spatial analysis of voting for political parties at the 2007 federal election. *People and Place* **17**, 39–54.

WICKHAM, H., BRYAN, J., KALICINSKI, M., VALERY, K., LEITENNE, C., COLBERT, B., HOERL, D. & MILLER, E. (2019a). *readxl: Read excel files.* URL https://CRAN.R-project.org/package=readxl. R package version 1.3.1.

WICKHAM, H., COOK, D., HOFMANN, H. & BUJA, A. (2011). tourr: An R package for exploring multivariate data with projections. *Journal of Statistical Software, Articles* **40**, 1–18. doi:10.18637/jss.v040.i02. URL https://www.jstatsoft.org/v040/i02.

WICKHAM, H., FRANÇOIS, R., HENRY, L. & MÜLLER, K. (2019b). *dplyr: A grammar of data manipulation.* URL https://CRAN.R-project.org/package=dplyr. R package version 0.8.0.1.

XIE, Y. (2015). *Dynamic documents with R and knitr.* Boca Raton, Florida: Chapman and Hall/CRC, 2nd edn. URL http://yihui.name/knitr/.