

# 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 preference in the six Australian federal elections held between 2001 to 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, spatio-temporal imputation 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 that can be thought of as the unobserved political climate in each electorate. Over time, the impact of most socio-demographic characteristics that affect electoral two-party preference do not vary, with industry of work, incomes, household mobility and de facto relationships having strong effects in each of the six elections. Education and unemployment are amongst those that have varying effects. It is also found that between 2004 and 2013, the spatial effect was insignificant, meaning that electorates voted effectively independently. 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*

## 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,

---

\*Corresponding author. Email: [jeremyforbes1995@gmail.com](mailto:jeremyforbes1995@gmail.com)

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 (Labor) was victorious, whereas the 2001, 2004, 2013 and 2016 elections were won by the Liberal National coalition (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 preference, 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 does 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, McCrear & 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 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, McCrear & 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 standardized, and then a principal component analysis is used to group 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 (this means that 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. These pre-processing steps and model details are discussed in Section 3. Section 4 describes the inference conducted to determine significance of effects and how these change over time. 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 were 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 were defined for each of the electorates to be used in the analysis.

Raw data was 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 (respectively). 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 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 was employed to obtain electoral socio-demographics. This method uses Census information from both before and after the election of interest.

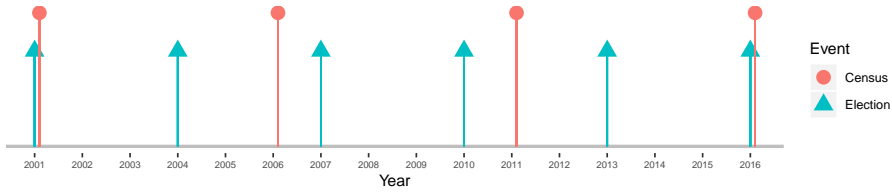


Figure 1. Timeline of Australian elections and Censuses. They do not always occur in the same year.

## 122 *Spatio-temporal imputation*

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

129 The finest level of disaggregation available for Census data is the region classification called  
 130 Statistical Area 1 (SA1). In 2016, Australia was divided into over 55,000 SA1s. Consider each  
 131 of these SA1 regions as a source zone,  $s = 1, \dots, S$ , for which socio-demographic information  
 132 is available. For simplicity, let each source zone be wholly summarised by its centroid. A set  
 133 of target zones,  $t = 1, \dots, T$ , are defined as regions for which information is to be imputed  
 134 - these are the electoral boundaries for a particular election. This is described in the context  
 135 of a single election, and a single neighbouring Census. Take the example of the Melbourne  
 136 Ports electorate from the 2013 federal election, illustrated in Figure 2. The purple region in  
 137 this figure represents the target zone and the source zones are the centroid locations from the  
 138 2016 Census SA1 areas.

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

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

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

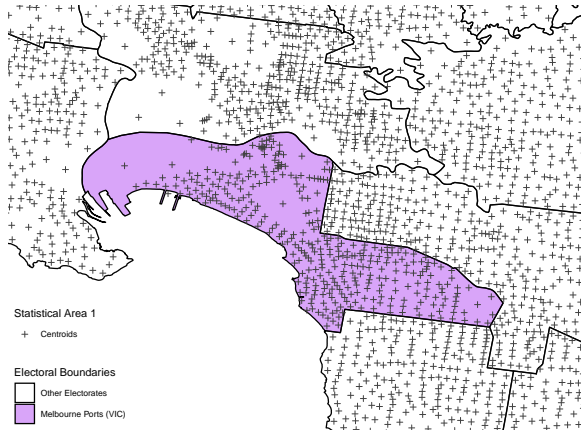


Figure 2. The electoral boundaries for Melbourne Ports (shown in purple) and surrounding electorates, with centroids for Statistical Area 1 regions from the 2016 Census overlaid. The centroids falling within the purple region are attributed to Melbourne Ports.

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,  $\hat{C}_{MelbPorts,2016}$  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  $\hat{C}_{MelbPorts,2011}$ , from which the final estimate is given by

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

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

### 3. Modelling

Following this 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 standardize variables. Second, many variables represent similar information and where appropriate, should be grouped together. To determine which variables should be grouped, principal component analysis (PCA) is used. The intuition here is that PCA will identify which variables covary, from which intuitive groupings of variables can be chosen to combine into a single variable. 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 standardized 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.

#### 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 represent similar information about an electorate. Any model that uses all variables would face problems with multicollinearity and over-fitting, which would likely lead to erroneous conclusions regarding variable significance. To address this, 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 to be included in the model as separate variables by construction.

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 a large loading in a given component, and there is an intuitive reason as to why they should be grouped together, then this subset will be combined to become a factor. Six factors are created using this criteria. These are: `Incomes` (median personal income, median household income, median family income); `Unemployment` (unemployment rate, labor 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, university qualifications, proportion of people working as professionals, proportion of jobs in finance, proportion of workers who are laborers, 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 standardized to have mean zero and variance one, within each election.

There are  $p = 33$  variables in the resultant set obtained from this procedure. One of the age brackets (`Pop_55_plus`) is removed from this set to avoid multicollinearity, because the other three age brackets are included. Hence, the final predictor set contains  $p = 32$  variables.

---

\*It is appropriate to compute principal components on a combined dataset of all six elections because when computed separately for each election, scree plots level off after four components and the loadings of the first four components are similar across the elections.



### 3.3. Regression incorporating spatially dependent errors

An identical model specification is used for each of the six elections, with each election modelled separately. This allows the socio-demographic effects to be estimated separately for each election year, facilitating analysis of temporal changes in variable effects. This approach is preferable to using a single longitudinal model because it avoids imposing any time-varying restrictions on the effect of a variable. To gain any advantage over modelling each election as a separate cross-section, a pooled cross-sectional model would have to force the effect of at least one variable to be invariant over time (e.g. forcing the estimated coefficient of the variable `Unemployed` to be the same in 2013 and 2016 elections). This would not allow for conclusions to be drawn about how the effects of these variables change over time. Furthermore, the panel approach is avoided because of how frequently electoral boundaries change, noting that electorates with the same name across elections are not guaranteed to represent the same geographical 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).

For each cross-section, let the response  $\mathbf{y}$  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 unobserved political climate in each electorate, where this climate is correlated with the climate in neighbouring electorates, under the assumption that the climate is independent of electoral socio-demographics.

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  $\rho$  be the spatial autoregressive coefficient,  $\mathbf{v}$  be a spherical error term,  $\mathbf{W}$  be a matrix of spatial weights (containing information about the neighbouring regions),  $\mathbf{X}$  be a matrix of socio-demographic covariates,  $\boldsymbol{\beta}$  be a vector of regression coefficients and  $\mathbf{a}$  be a spatially structured random effect vector.

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

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

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

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

256 and

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

257 where  $\mathbf{v} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ , and hence

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

258 Estimation of the above spatial error model is undertaken using feasible generalized least  
259 squares.

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

## 262 4. Results

### 263 4.1. Spatial autoregressive parameter

264 The spatial autoregressive coefficient  $\rho$  is positive and significant in the 2001, 2004 and 2016  
265 elections (Figure 3). In these three elections, there is evidence to suggest that the political  
266 climate of an electorate appears to be affected by the attitudes of its neighbours, affecting the  
267 electorate's two-party preference. Conversely, in the other three elections, the spatial effect  
268 weakens to become insignificant (although still positive).

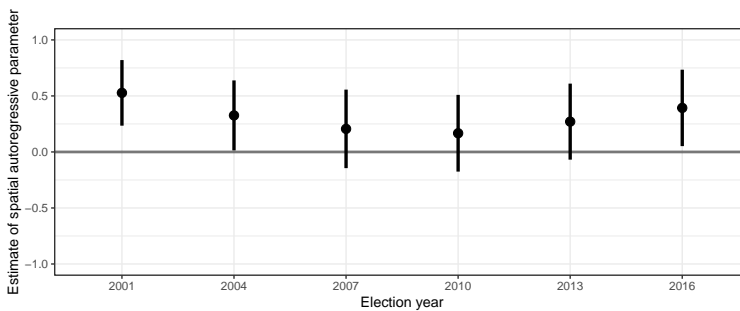


Figure 3. Estimates of the spatial autoregressive parameter for each of the six elections, reported with their individual 95% confidence intervals. Only in 2001 and 2016 is there a significant spatial component.

### 269 4.2. Country-wide trend

270 Since all socio-demographics have been standardized to have a mean of zero and a variance  
271 of one, the intercept in each model can be interpreted as the estimated two-party preferred

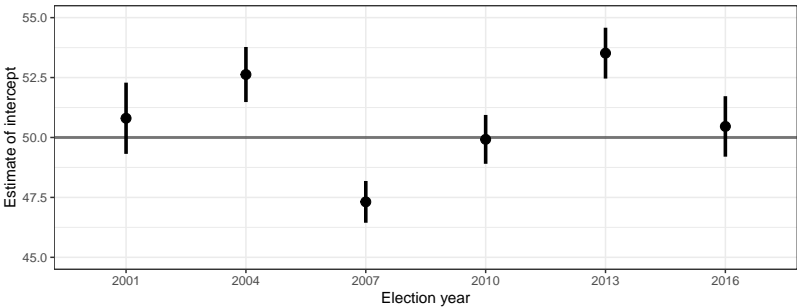


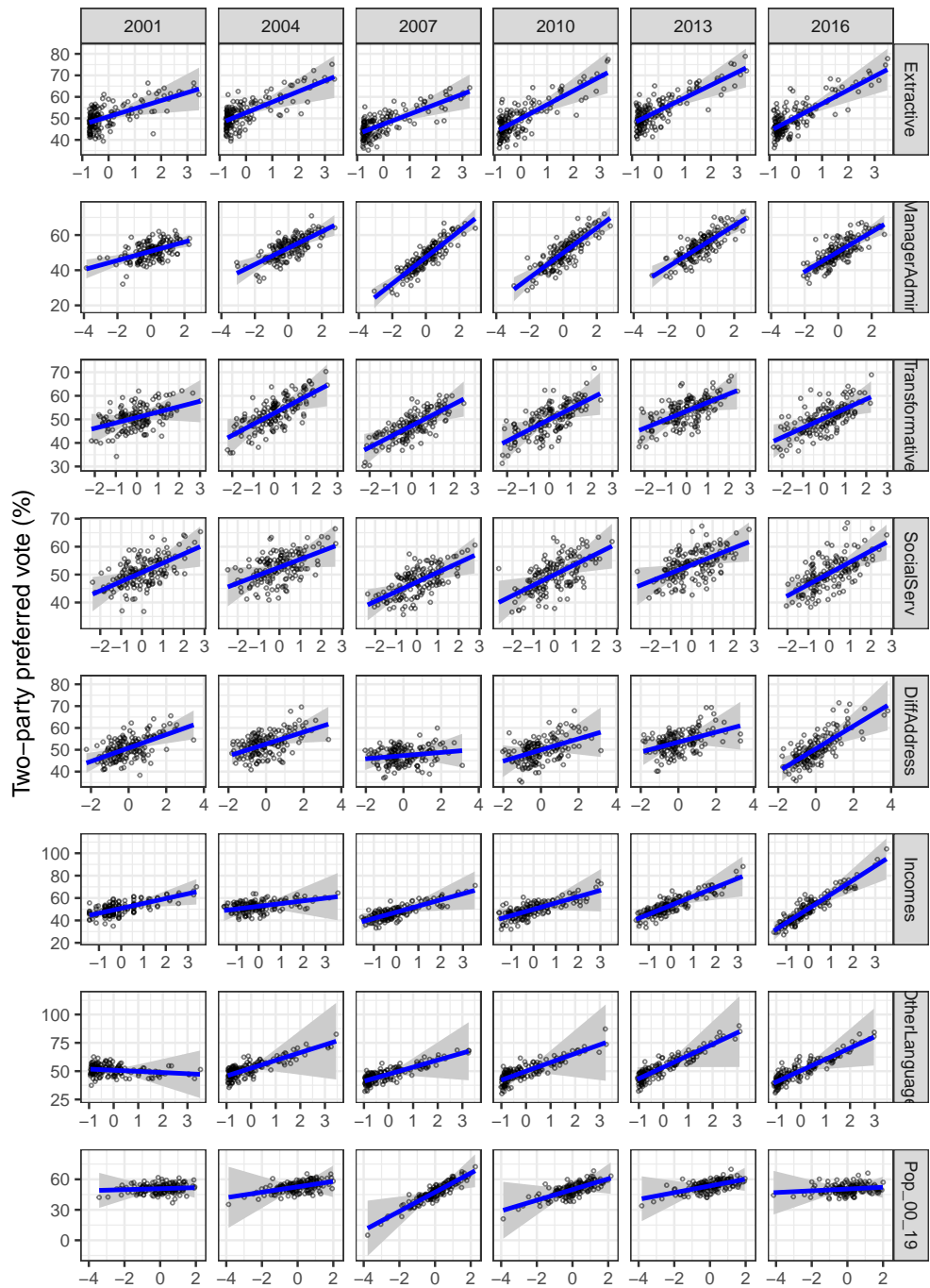
Figure 4. Estimated intercept for each election, which represents the two-party preferred vote for an electorate with mean characteristics.

vote for an electorate with mean characteristics<sup>†</sup>. Figure 4 shows that the baseline of party preference has 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.

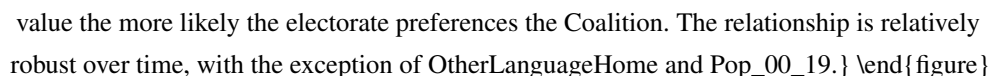
4.3. Influential socio-demographics

`\begin{figure}[h]`

<sup>†</sup>Mean of all variables aside from Judaism, Indigenous, Islam and Buddhism, where it assumes the mean of the log value.



Partial residual plots by election year for a selection of predictors. Linear model with 95% confidence bands overlaid. Most predictors have a positive relationship: the larger the

value the more likely the electorate preferences the Coalition. The relationship is relatively robust over time, with the exception of `OtherLanguageHome` and `Pop_00_19`. 

To investigate the socio-demographics that have a strong effect on the two-party preferred vote, partial residual plots are used and shown in Figures 4.3 and 5. 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 are computed using the method in Breheny & Burchett (2017). If a horizontal line can be drawn through the confidence band, then the effect is insignificant. The estimated intercept is also added to the partial residuals for interpretability. Plots for each election are faceted to compare the effects over time in Figures 4.3 and 5. Only socio-demographics that have a significant effect in at least two elections are displayed in Figures 4.3 and 5.

### ***Industry and type of work***

Electorates with higher proportions of workers in mining, gas, water, agriculture, waste and electricity (grouped as `Extractive` industries) are consistently linked with higher support for the Liberal party, with the magnitude of this effect slightly increasing over the years (see row 3 in Figure 4.3). This is unsurprising, as the Liberal party has close ties with these traditional energy industries, and typically present policies to reduce taxation on energy production. Furthermore, electorates with more workers in construction or manufacturing industries (`Transformative`) are also more likely to support the Liberal party (see row 4 in Figure 4.3), from 2004 onwards.

Similarly, the proportion of workers in managerial, administrative, clerical and sales roles (`ManagerAdmin`), is also a significant predictor of two-party preference vote across all six elections, with a higher proportion of people working these jobs increasing 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 more generous funding schemes affecting these areas of work, so one might expect `SocialServ` to have a negative effect on two-party preference.

However, in every election this effect is found to be positive and significant.

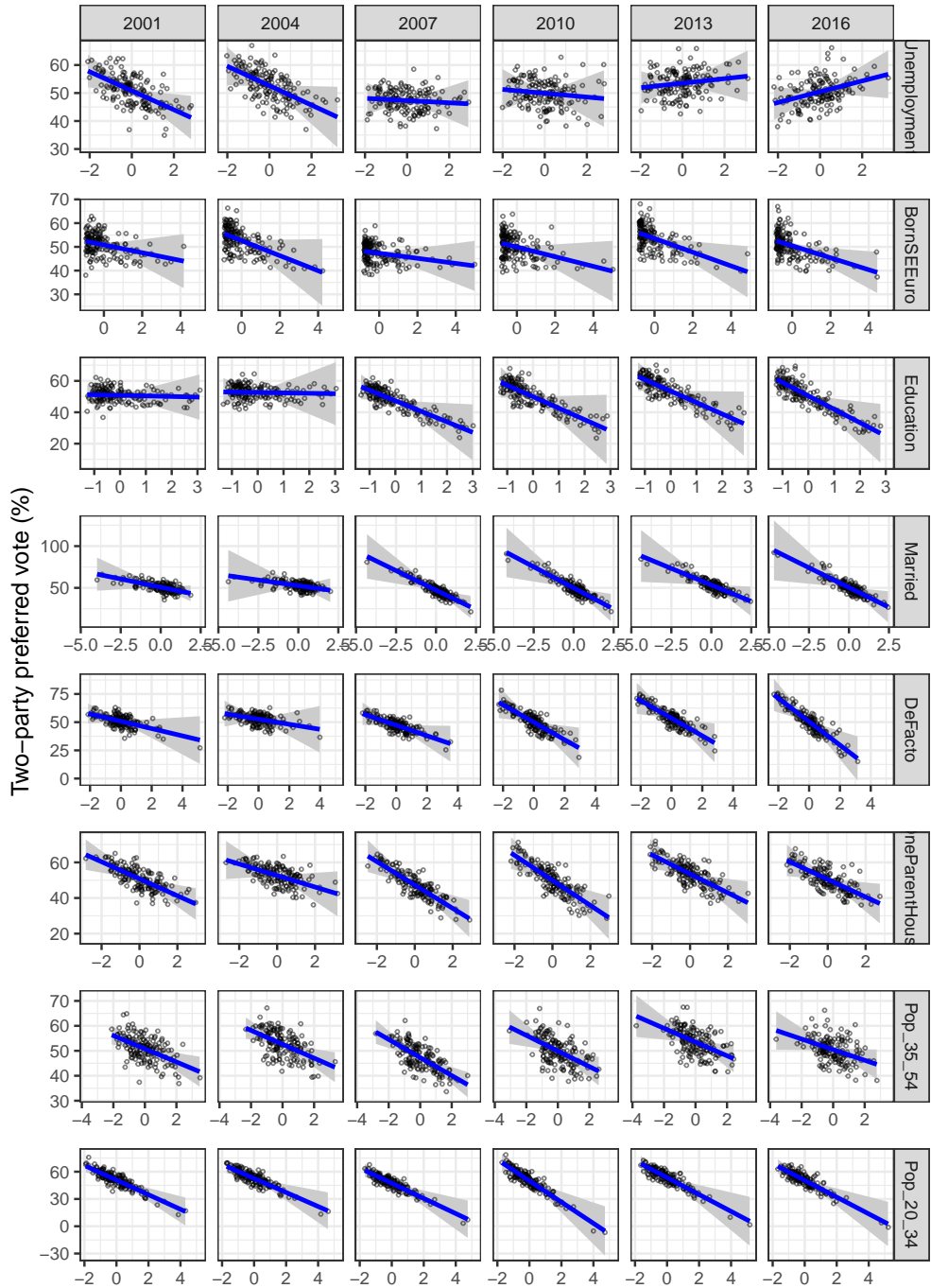


Figure 5. Partial residual plots by election year for a selection of predictors. Linear model with 95% confidence bands overlaid. Several predictors have a negative relationship: with larger values indicating the electorate more likely preferences Labor. Most relationships are relatively stable over elections, except Unemployment and Education.

### ***Income and unemployment***

Typically the Labor party campaigns on more progressive policies, which often include 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 appear more likely to support the Liberal party, as the `Incomes` factor has a positive effect on Liberal preference (see row 1 in Figure 4.3). This effect is significant in every election aside from 2004 and 2010.

Unemployment however, is not as influential. In 2001 and 2004, electorates with higher unemployment align with Labor, but over time this shifts towards support for the Liberal party, culminating in a positive (insignificant) effect in 2016.

### ***Age***

The older Australian population is often believed to be more conservative, and the left leaning political parties (including Labor) typically have a stronger appeal to younger people. This effect is indeed observed across the six elections, with populations between 20 and 34 years of age (`Pop_20_34`) being very strongly aligned with Labor preference (bottom row in Figure 5). Larger populations of 35 to 54 year olds (`Pop_35_54`) are also associated with Labor, but the magnitude of this effect is far smaller. Populations under 20 years of age is only significant in 2007, where `Pop_00_19` increased Liberal support.

### ***Education***

Since 2007, electorates with higher education levels are associated with supporting 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 has a negligible effect (see row 3 in Figure 4.3). Additionally, student populations (`CurrentlyStudying`) do 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, are either associated with Labor support, or have no effect. Of these areas, only South-Eastern European populations significant affect party preference, with larger populations associating with Labor in 2013 and 2016 (row 2, Figure 5). Speaking other languages (aside from English) however, appears to have a far stronger effect, as observed



through the `OtherLanguage` variable. Electorates with more diverse speech are associated with higher support for the Liberal party from 2004 onwards, with this effect being significant in 2013 and 2016 (row 7, Figure 4.3). Furthermore, none of the variables relating to religious beliefs aside from Christianity have a material effect in any election (this includes the Buddhist, Muslim, Jewish, areligious and Indigenous Australian populations). The association between Christian populations (`Christianity`) and the Liberal party steady increases over the years, becoming positive and significant in 2016.

### ***Households***

In 2001, 2004 and 2016, higher proportions of people that have recently (in the past five years) moved house (`DiffAddress`) increased electoral support for the Liberal party (see row 5 in Figure 4.3). Having controlled for characteristics of house ownership and rental prices (`PropertyOwned` and `RentLoan` - both of which have no effect), this effect is somewhat surprising.

Higher proportions of single parent households are associated with Labor support in all elections (albeit insignificant in 2004, see row 6 in Figure 5), whereas the electoral family and household sizes (via the `FamHouseSize` variable) do not appear to be associated with either party in any election.

### ***Relationships***

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

#### ***4.3.1. Robustness***

Two robustness checks are conducted to confirm model stability. First, a model for each election is fit using only the variables that are found to be significant in at least one of the six elections. Let this model be the reduced model, and the original model be the full model. The estimated coefficients of the variables in the reduced models all fall within their respective 95% confidence intervals from the full models, confirming that the significant effects are robust to variable selection.

The second check involves computing pairwise correlations amongst the explanatory variables, and seeing if the removal of one variable (from a pair) affects the other's estimated effect. For example, the variables `Born_Asia` and `BornElsewhere` have a pairwise correlation of 0.93 - the largest of any pair. A model is fit omitting `BornElsewhere` (the reduced model). The estimated coefficient of `Born_Asia` is then compared with the confidence interval for `Born_Asia` in the full model. It is found that for each of the top ten pairwise correlations, the estimated effect in the reduced model lies within the confidence interval from the full model. This second check indicates that variables with the highest pairwise correlation are not masking the effects of each other.

#### 4.4. A closer look at the residuals

##### *Residuals by state*

It is often hypothesized that states have systematic differences that cause their electorates to vote differently. Boxplots of residuals grouped by state (Figure 6) reveal that the data reflects this to a limited extent. Tasmania and the Australian Capital Territory appear to have a bias towards Labor, whereas the Northern Territory tends towards voting Liberal. However, there are relatively few electorates in each of these states (five, two and two respectively), so this apparent result may be due to incumbent effects rather than an actual state-specific bias.

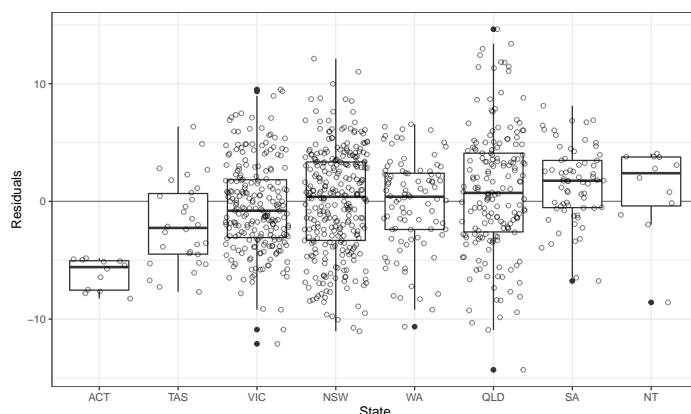


Figure 6. Boxplot of residuals by state with jittered points. States ordered by median residual. A state-specific bias present only in the smaller states appears to have not been captured by the model.

### *Residuals by party incumbency*

The incumbent party appears to have a distinct advantage at the next election. The boxplots in Figure 7 show that if either of the Labor or Liberal parties won the seat at the previous election, the electorate is likely to vote in their favour, over and above any socio-demographic effects — this effect has not been captured by the model.

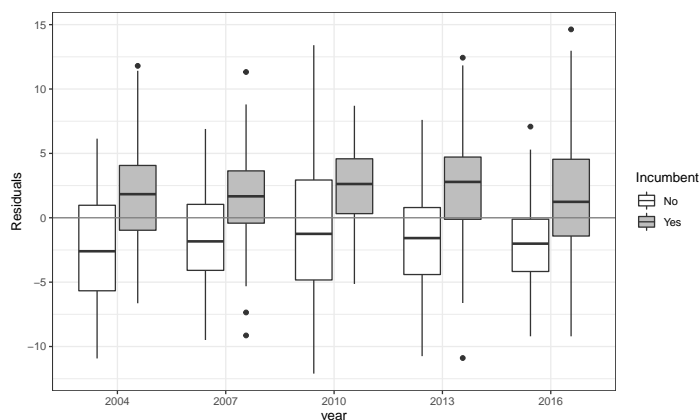


Figure 7. Boxplot of residuals for incumbent and non-incumbent parties each year. An incumbent advantage is evident and has not been captured by the model.

### *Outlier electorates*

Based on the distribution of the Cook's distance values, a Cook's distance greater than 0.1 is considered to be influential and a potential outlier. The electorate of Sydney (NSW) has a large Cook's distance from 2001 to 2010, due to its diverse population (language, birthplace and religion), high number of defacto relationships, high income, high household mobility and small amount of workers in extractive and transformative jobs. It has remained a strong supporter of the Labor party and the Liberal vote is severely overpredicted by the model, making it an outlier. Nearby in metropolitan NSW, the electorate of Wentworth is 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 since, the model underpredicts Wentworth's Liberal support.

Lingiari, an electorate taking up almost all of the Northern Territory, is an outlier in the all but the 2013 election due to its large Indigenous population, young age profile and low rates of

property ownership. Fowler (NSW) has a diverse population with a high proportion of migrants, many Buddhists and Muslims, and has strong Labor support, making it influential in 2001, 2004 and 2010. Other electorates with large Cook's distance are Canberra (ACT) and Durack (WA) in 2013, and Solomon (NT) in 2016.

## 5. Conclusion

This paper explores 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 does not always occur in the same year as an election, Census data for the 2004–2013 elections are generated by employing a method of spatio-temporal imputation. This imputes electoral socio-demographics for the electoral boundaries in place at the time of the election — an approach that is distinctly different from previous work on modelling election outcomes, where Census and election data are typically joined without addressing their temporal differences. Before estimating a model, these socio-demographic variables are standardized (to adjust for changing variable scales) and many variables (representing similar information) are combined into factors, resulting in a reduced predictor set. A spatial error model is then estimated for each election, accounting for the inherent spatial structure of the data.

Across the past six elections, most of the socio-demographics that drive the electoral two-party preferred vote are found to remain steady, whilst a few (typically weaker) effects vary over time. Industry and type of work are particularly influential. Energy-related and manufacturing/construction jobs, as well as administrative roles and jobs in education and social services are strongly linked with the Liberal party in all elections. Incomes have a similarly consistent effect, with higher income areas supporting Liberal. Higher levels of unemployment shift from weak association with Labor to a significant Liberal effect over the years, and higher education levels are associated with Labor from 2007 (although marginally insignificant in 2010). Electorates with large populations 20 to 34 years are strongly associated with Labor, whilst the 35 to 54 year old bracket also increases Labor support, but to a lesser extent. It is also found that birthplace diversity slightly favours Labor, relationships (both marriages and de facto relationships) align with Labor preference from 2010 onwards, and the influence of Christian populations has trended towards Liberal support whilst other religions have negligible effects. Family and household sizes have minimal influence, although electorates with more single parent households are linked with Labor support. Furthermore, the neighbourhood (spatial) effects are found to be positive in all elections and

significant in 2001, 2004 and 2016, meaning that neighbouring party preference appears to be an influential factor 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 the raw data and imputed 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 future Australian electoral analysis.

## 6. Acknowledgements

This paper was produced using *RMarkdown* (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](https://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.

## 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 2016. 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/](https://ropenscilabs.github.io/eechidna/).

The authors would like to sincerely 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.

## 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. 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 1.3.0.
- 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*. 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., 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/>.