Spatial modelling of the electoral two-party preferred vote in Australia. A study of federal elections between 2001 and 2016 via the `eechidna` `R` package.

Jeremy Forbes, Di Cook & Rob Hyndman

2019-03-24

Contents

Chapter 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 (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 characteristics relate to two-party preference, and whether their effects have changed over time. Electoral socio-demographics are derived from the Census, and vote counts are obtained from federal elections.

Joining these two data sources is problematic as there in an inherent asynchronicity in the two events. A Census is conducted by the Australian Bureau of Statistics (ABS) every five years, whereas a federal election (conducted by the Australian Electoral Commission (AEC)) usually occurs every three years. The first problem addressed is that of obtaining 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 is to join voting outcomes to the nearest Census, without accounting for any temporal differences (see Davis and Stimson (1998), Stimson et al. (2006), Liao et al. (2009) and Stimson and Shyy (2009)). Furthermore, electoral boundaries change regularly, so spatial discrepancies also arise when matching electoral data. To obtain appropriate Census data for these four elections, electoral socio-demographics are imputed using a spatio-temporal imputation that combines areal interpolation (Goodchild et al., 1993) and linear time-interpolation. The process of collecting and wrangling the raw data is outlined in section 2, and the imputation process is detailed in section 3. 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 have found that aggregate socio-demographics are relatively good predictors of voting outcomes. Forrest et al. (2001) does this using multiple regression of the Liberal and Labor primary vote for polling booths in the Farrer electorate in 1998. Stimson et al. (2006), Stimson and Shyy (2009) and Stimson and Shyy (2012) use 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. On the contrary, Stimson and Shyy (2009) models the polling booth swing vote (change in the two-party preferred vote) in the 2007 election, finding that little of 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, where each federal elections is modelled with a cross-sectional data set. The use of a regression framework to examine these socio-political relationships over time is seemingly absent from previous Australian studies. It also appears that no study has attempted any type of statistical analysis of socio-demographics in conjunction with voter behaviour in Australia since 2007, making this paper distinctly different from those previous.

The cross-sectional data set for each election consists of the two-party preferred vote (response variable), and socio-demographic variables (explanatory variables) that characterise each electorate. To obtain these cross-sections, socio-demographic variables are first standardized, and then principal components are used to group variables into "factors". To account for the inherent spatial structure of the data, a spatial error model is fit for each election. These steps are discussed in section 4. In section 5 inference is conducted on the models to see which effects are significant, how effects change over time and which electorates have abnormal voting behaviour.

Chapter 2

Data collection

2.1 Voting 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 divide 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.

2.2 Census data

Socio-demographic variables are derived from the 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 in the 21st century, being that in 2001, 2006, 2011 and 2016. The Australian Bureau of Statistics (ABS) conducts the Census and publishes aggregated information. The ABS approximation of electorates at the time of the Census is chosen. From this aggregate information, 67 socio-demographic variables are computed for each of the electorates.

2.3 Downloading and wrangling

Raw data is sourced online from the AEC and ABS websites in .csv and .xlsx files. The format of these files change 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 very useful, as they provide fast consistent tools for data manipulation and functions to import .xlsx files (respectively). The 2001 and 2006 Census data are however published in a format where each electorate has a separate document, making it difficult to use the `dplyr` tools and 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.

Chapter 3

Joining Census and election data

3.1 Differences between Census and election data

Between 2001 and 2016 there were six elections and four Censuses (see Figure 3.1). Electoral boundaries are redistributed regularly by the AEC, meaning that only in the years where both a Census and election occur will the boundaries match - the case for the 2001 and 2016 election. 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.

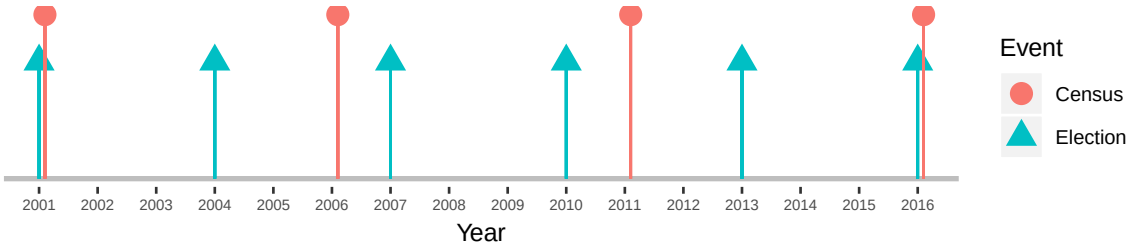

Figure 3.1: Timeline of Australian elections and Censuses. They do not always occur in the same year.

3.2 Spatio-temporal imputation

To account for spatial differences, the piece-wise approximation method in Goodchild et al. (1993) is adopted. Consider a map of source zones $s = 1, ..., S$, for which socio-demographic information is available, and a set of target zones $t = 1, ..., T$ for which information is to be imputed. In this context the map of electoral boundaries at the time of a Census would be the source zones, and the boundaries at the time of the election would be the target zones. Denote the area of intersection between source zone s and target zone t as $A_{s,t}$, the population of the source zone s as U_s, and the population of intersection between source zone s and target zone t as $P_{s,t}$.

Compute each $A_{s,t}$ and estimate population of the intersection:

$$\hat{P}_{s,t} = \frac{U_s * A_{s,t}}{\sum_{t=1}^{T} A_{s,t}}$$

This assumes that populations are uniformly distributed within each source zone.

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

$$\hat{C}_t = \frac{\sum_{s=1}^{S} D_s * \hat{P}_{s,t}}{\sum_{s=1}^{S} \hat{P}_{s,t}}$$

This assumes that each individual in a source zone assumes the aggregate characteristics of the zone.

Applying this to each of the target zones addresses the spatial component, as it imputes the required socio-demographic for the desired electoral boundaries. However these are applicable at the time of the Census (source year) and are not yet appropriate for the election (target year).

Denote year y, with a Census falling on y_1 and y_3, and an election on year y_2, and add this subscript to the Census variable estimate, $\hat{C}_{t,y}$. 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 y_2. This assumes that population evolves in a linear manner over time.

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

3.3 Applied

Publically available Census data is aggregated and there are different resolutions accessible, ranging from SA1 (over 50,000 zones) to electoral divisions (150 zones). For this study electoral divisions are used as source zones, and this imputation method is applied for each of the 2004, 2007, 2010 and 2013 elections. To demonstrate its functionality, consider the imputation of socio-demographic variables for the electorate of Hume in New South Wales (NSW), at the time of the 2013 federal election. Figure 3.2 shows this region amongst other NSW electorates.

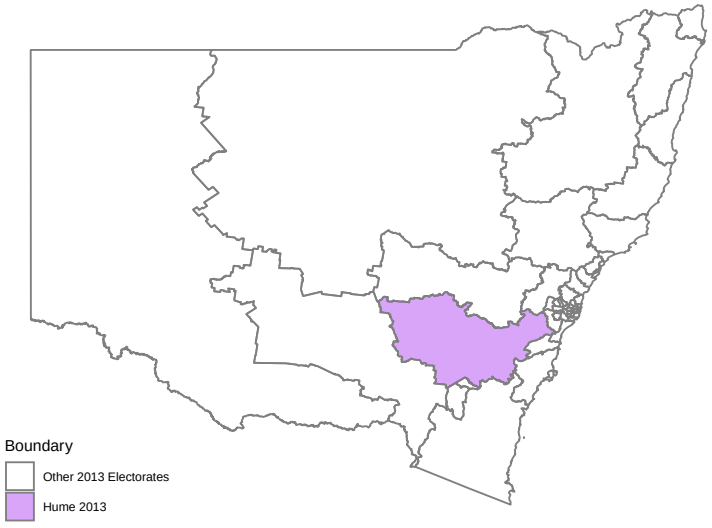

Figure 3.2: Some of the electoral boundaries in NSW for 2013, with the electoral boundary for Hume, shown in purple.

The Censuses neighbouring the 2013 election are those in 2011 and 2016, and the Hume boundary is changed, as seen by plotting the Hume boundary (purple) in the 2013 election over the divisions in 2016 (see Figure

3.3).

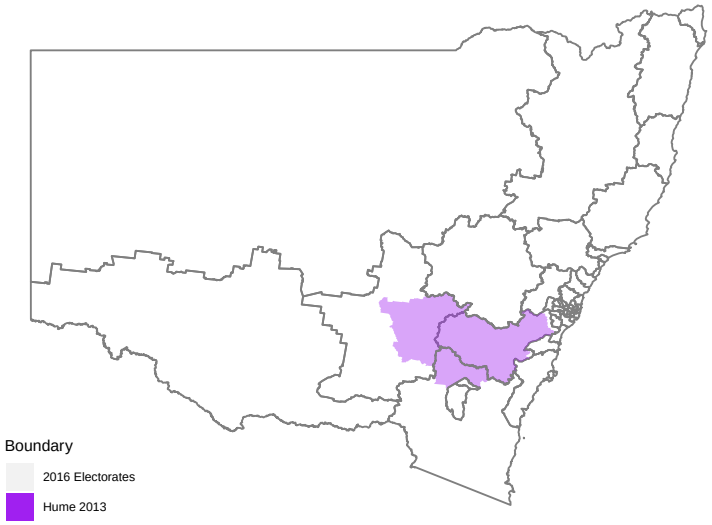

Figure 3.3: Census division boundaries in NSW for 2016, with the 2013 electoral boundary for Hume, shown in purple. The purple region is not contained within a single Census division.

There are many electorates in 2016 that intersect with the purple region (Hume boundary for 2013), these include the divisions of Riverina, Eden-Monaro and Hume, along with smaller intersecting areas with Fenner, Calare, Gilmore and Whitlam. To impute Census information for this purple region, calculate the percentage of each 2016 electorate that intersects with the purple region, which is then used to estimate intersection populations $\hat{P}_{s,t}$.

Electorate (2016)	Percentage	Population in Electorate	Estimated Population Allocated to Purple Region: $\hat{P}_{s,t}$
HUME	96.54%	150643	145427
RIVERINA	25.11%	155793	39117
EDEN-MONARO	11.09%	147532	16358
CANBERRA	0.28%	196037	548
FENNER	0.23%	202955	474
WHITLAM	0.06%	152280	92
GILMORE	0.06%	150436	86
CALARE	0.01%	161298	21

Now consider the socio-demographic *AusCitizen* - the proportion of people in the region who are Australian citizens.

DivisionNm	AusCitizen (%): D_s	Estimated Population Allocated to Purple Region: $\hat{P}_{s,t}$
HUME	90.02	145427
RIVERINA	89.11	39117
EDEN-MONARO	88.00	16358
CANBERRA	85.48	548
FENNER	83.64	474
WHITLAM	89.52	92
GILMORE	89.03	86
CALARE	87.56	21

Then taking a weighted average of *AusCitizen* using the estimated population as weights yields $\hat{C}_{Hume,2016} = 89.65\%$. Repeating this process using the 2011 Census and electoral boundaries yields $\hat{C}_{Hume,2011} = 91.00\%$

Finally, linearly interpolate between 2011 and 2016 to arrive at the 2013 estimate:

$$\begin{aligned} \hat{C}_{Hume,2013} &= \frac{3}{5} \cdot \hat{C}_{Hume,2011} + \frac{2}{5} \cdot \hat{C}_{Hume,2016} \\ &= \frac{3}{5} \cdot 91.00\% + \frac{2}{5} \cdot 89.65\% \\ &= 90.46\% \end{aligned}$$

This is done for each of the socio-demographic variables, and repeated each of the 2013 electorates.

Chapter 4

Modelling

4.1 Data pre-processing

With socio-demographic information now available for each electorate, each election is joined to the data corresponding with its two-party preferred vote. Socio-demographic variables within each election year are standardized to have mean zero and variance one, to adjust for changing variable scales. For example, inflation-adjusted median rental prices increased across almost all electorates, with median rent of 200 dollars per week placing an electorate in the 90th percentile in 2001, but only the 30th percentile in 2016.

4.1.1 Dimension reduction

With only $N = 150$ observations (electorates) in each election and $p = 65$ socio-demographic variables in each cross-section, any model using all variables would face serious problems with multi-collinearity and over-fitting, likely leading to erroneous conclusions regarding variable significance. Therefore a form of dimension reduction is adopted before models are fit.

Socio-demographic variables[1] that represent similar information are combined into "factors" using principal component analysis (PCA). The scree plots of the principal components for each election all level off after four components, and the loadings of these four components are similar across the elections. Principal components are then computed on the combined set of socio-demographics across all six elections. A factor is created by combining several variables all have large loadings in a particular component and when there is an intuitive reason as to why these variables could represent common information. A loading with magnitude greater than 0.15 is considered large. After computing these sums, each factor is again standardized to have mean zero and variance one, within each election.

Consider the `Incomes` factor as an illustration. Independent of principal components, we may suspect that median personal income, median household income and median family income are providing similar information about the financial wellbeing of an electorate. Their loadings in the first principal component are large (0.20, 0.21 and 0.22 respectively), which provides the evidence needed to combine these variables into a single factor, which is called `Incomes`.

This process reduces the predictor set to $p = 30$.

[1] A preliminary step involved removing all age bands, because age is represented by median age, and to remove variables relating to particular denominations of Christianity.

4.2 Model framework

An identical model specification is used across the six elections, with each election modelled separately. This allows for the socio-demographic effects to be estimated separately for each year, allowing for interpretation of temporal changes in these effects. This is preferable over a single longitudinal model because it avoids any concerns of undue bias stemming from an incorrectly imposed time-varying restriction on any variable. Without such restrictions, a pooled cross-sectional model does not yield any distinct advantage over separate cross-sections. The panel approach is avoided because of how frequently electoral boundaries change, meaning that electorates that have 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 requiring further imputation.

For each cross-section, let the response variable be the two-party preferred vote in favour of the Liberal party, denoted Y, with $Y = 70$ representing a 70% preference for Liberal, 30% for Labor. Although Y lies in the interval $(0, 100)$, observed values are never very close to 0 or 100 (minimum 24.05% and maximum 74.90%), so there is no need to impose the constraint of $Y \in [0, 100]$. Furthermore, the response is found to be spatially correlated in each election (Moran's I test, $p \leq 7 \cdot 10^{-15}$). This is expected, as electorates are aggregate spatial units, and hence the spatial structure of the data must modelled appropriately.

The spatial error model (Anselin, 1988) is chosen because captures spatial heterogeneity by incorporating a spatially structured random effect vector (LeSage et al., 2009). In this context, the random effect can be thought of as capturing the unobserved political climate in each electorate, where the climate is correlated with the climate in neighbouring electorates. This functions under the assumption that the climate is independent of electoral socio-demographics, and that an electorate is equally correlated with any electorate that shares a part of its boundary. Spatial weights are calculated in accordance with these assumptions. The spatial error model is specified as follows:

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

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

and

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

where

$$\boldsymbol{v} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

.

so it can be written

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

Estimation is done using generalized least squares (using the `spdep` `R` package (Bivand et al., 2013)).

Table 4.1 details the resultant estimated model coefficients and their estimated standard errors for each of the six elections. These are interpreted in the next section.

insert table here

Chapter 5

Results

Bibliography

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*.

Davis, R. and 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(3):69–82.

Forbes, J., Cook, D., Ebert, A., Hofmann, H., Hyndman, R., Lumley, T., Marwick, B., Sievert, C., Sun, M., Talagala, D., Tierney, N., Tomasetti, N., Wang, E., and Zhou, F. (2019). *eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia*. R package version 1.3.0.

Forrest, J., Alston, M., Medlin, C., and 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(2):167–182.

Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and planning A*, 25(3):383–397.

LeSage, J., Kelley Pace, R., and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Liao, E., Shyy, T., and 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(2):63–88.

Stimson, R., McCrea, R., and Shyy, T. (2006). Spatially disaggregated modelling of voting outcomes and socio-economic characteristics at the 2001 australian federal election. *Geographical research*, 44(3):242–254.

Stimson, R. and Shyy, T. (2012). And now for something different: modelling socio-political landscapes. *Ann Reg Sci*, 50:623–643.

Stimson, R. and Shyy, T.-K. (2009). A socio-spatial analysis of voting for political parties at the 2007 federal election. *People and place*, 17(1):39–54.

Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitenne, C., Colbert, B., Hoerl, D., and Miller, E. (2019a). *readxl: Read Excel Files*. R package version 1.3.1.

Wickham, H., François, R., Henry, L., and Müller, K. (2019b). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.0.1.