MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

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

Jeremy Forbes, Dianne Cook, Rob J Hyndman

May 2019

Working Paper 08/19

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

Jeremy Forbes
Monash University
Melbourne, Australia
Email: jeremyforbes1995@gmail.com
Corresponding author

Dianne Cook
Monash University
Melbourne, Australia

Rob J Hyndman
Monash University
Melbourne, Australia

8 May 2019

JEL classification: C31, C33, D72

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

Abstract

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, 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 socio-demographic characteristics relate to two-party preference, and whether their effects have changed over time.

Data on electoral socio-demographics 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 Australian Electoral Commission (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 (**DavisStimson98Stimson06Liao09Stimson09**). 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 (**Goodchild1993**) and linear time-interpolation. Collecting and wrangling the raw data, along with the imputation process, are detailed in Section **??**. All data and associated documentation relating to this procedure are available in the `eechidna` `R` package (**eechidna**), 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. **Forrest01** 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. **Stimson06**, **Stimson09** and **Stimson12** 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, **Stimson09** 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. The cross-sectional dataset for each election used here consists of the two-party preferred vote (as the response variable), and a set of common socio-demographic variables (as the explanatory variables) that characterize each electorate. 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.

The paper is organised as follows. Section **??** describes the data collection, joining and cleaning. These pre-processing steps and model details are discussed in Section **??**. Section **??** describes the inference conducted to determine signficance of effects and how these change over time. Section **??** 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, 65 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