imputation

Jeremy Forbes

18/03/2019

Joining Census and election data

Differences between Census and electoral data

Between 2001 and 2016 there were six elections and four Censuses (see Figure @ref(fig:timeline)). Only in 2001 and 2016 did both events occur, and electoral boundaries are also redistributed regularly. Therefore, there are both temporal and spatial differences that need to be accounted for when joining the electoral two-party preferred vote from the four elections between 2004 and 2013 with Census data. For these elections a spatio-temporal imputation method is employed to derive a electoral socio-demographics which uses Census information from both before and after the election of interest.

Spatio-temporal imputation

To account for spatial differences, the piece-wise approximation method in @Goodchild1993 is appropriate for our problem. 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:

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

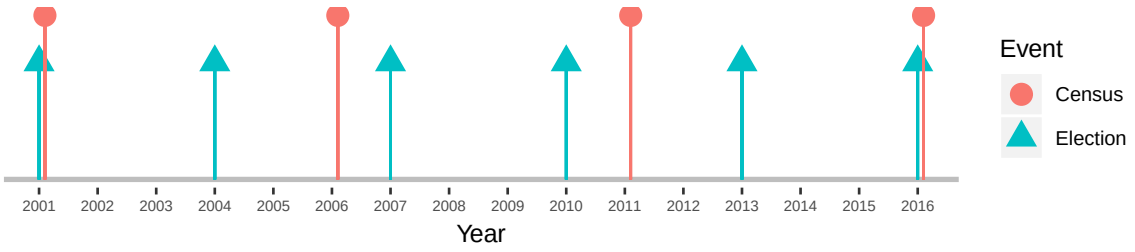

Figure 1: Timeline of Australian elections and Censuses. They do not always occur in the same year.

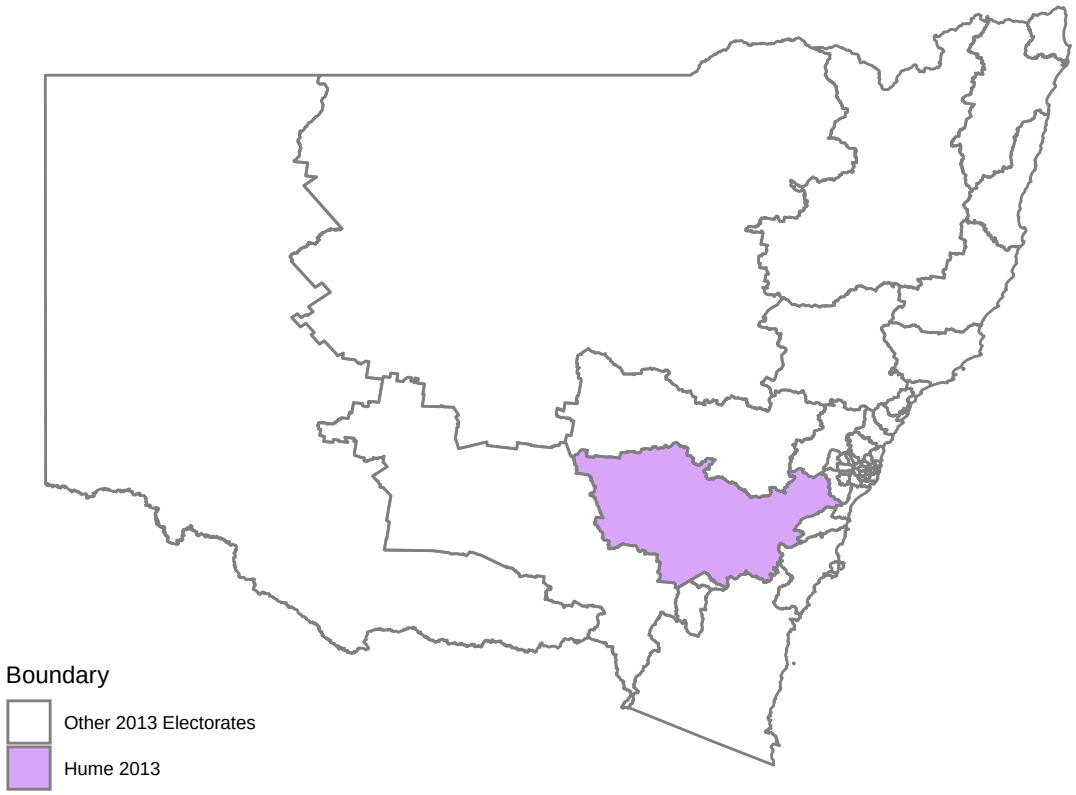

Figure 2: Some of the electoral boundaries in NSW for 2013, with the electoral boundary for Hume, shown in purple.

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

Socio-demographic information is now imputed for the target zones, but these are not appropriate for the 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.

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

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). Any of these resolutions could be used as source zones. For this study, electoral divisions are used and this imputation method is applied to each of the 2004, 2007, 2010 and 2013 elections. To demonstrate its functionality, consider the imputation of the socio-demographic variable *AusCitizen* for the electorate of Hume in New South Wales (NSW), at the time of the 2013 federal election. Figure @ref(fig:hume13) shows this region amongst other NSW electorates.

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.

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,

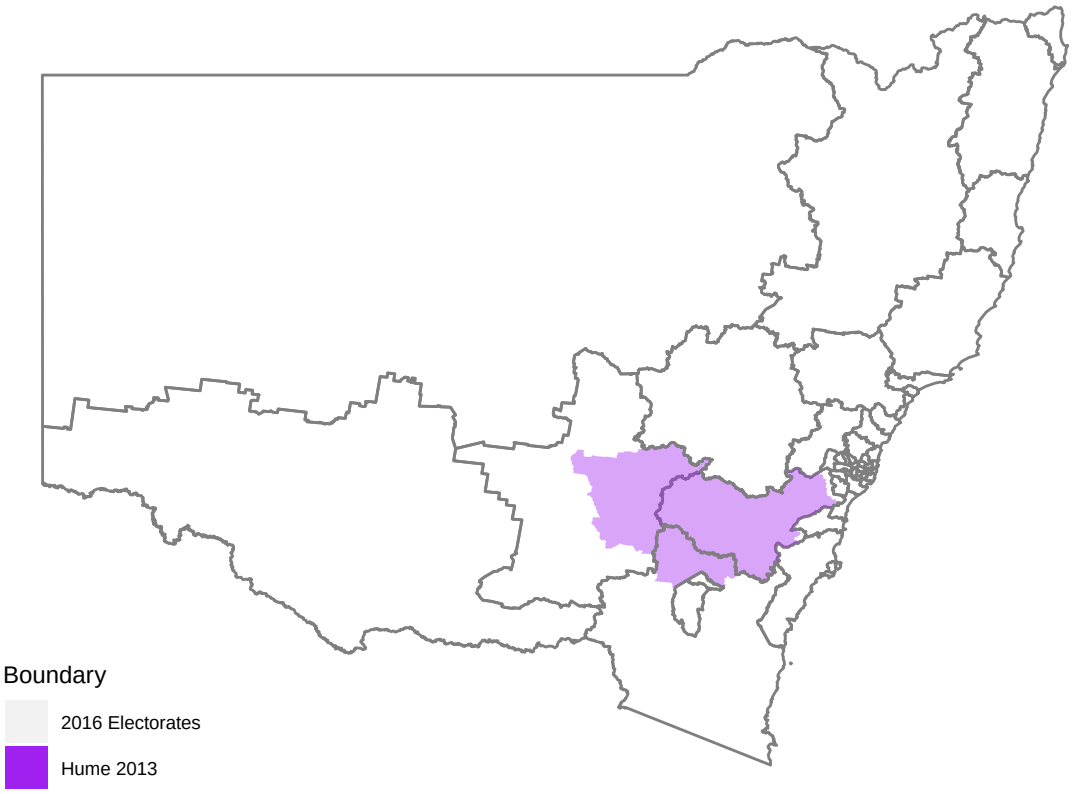

Figure 3: Census division boundaries in NSW for 2016, with the 2013 electoral boundary for Hume, shown in purple. The purple region is not contained within a single Census division.

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

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

The socio-demographic of interest is *AusCitizen*, which is 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.