17 November 2019

Dear Dr Holland,

Thank you for your helpful comments on our paper titled 'Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016'.

We have revised our paper to address the issues raised. Below are detailed summaries of the changes made in response to the points raised by yourself, the Associate Editor and by the anonymous referees.

Kind regards,

Jeremy Forbes, Di Cook and Rob Hyndman

Editor's Comments

1.   *Address reviewer 1's question about why spatial imputation is required given that the census is reported at a low level of aggregation (lines 150-151) that could presumably be clumped to give a good approximation to electorate boundaries.*

We now use Census data at the lowest level of aggregation available — the geographic region type of 'Statistical Area 1' — to approximate the electoral boundaries (see Section 2.2). Some spatial imputation is still required as the SA1 boundaries do not exactly sit within the electorate boundaries. For this we adopt a method similar to referee 1 suggests. Even with this change in imputation method, our results are effectively the same as the original submission.

2.   *I note that you do mention the ecological fallacy (lines 285-286) but reviewer 1's reaction suggests that you should expand your remarks here.*

We have expanded our remarks and moved them to the introduction (Section 1) to give them more prominence. We are explicit in stating that "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)".

3.   *It appears that your approach to variable selection has confused everybody. My understanding of what you have done is that: You begin with 65 variables (are they all the available census data or a subset?). Then you standardise the variables. Then you remove some variables to do with Age Bands and Christianity (how many variables out of the 65 does this leave? Were these variables totally removed or lumped in some way? Christianity still appears in table 2 so they obviously weren't all removed)*

We have now explained our approach to variable selection in more detail and made some additional changes to provide greater clarity. Previously we started with 65 variables, but many of these variables were removed in the variable selection process because they provided additional granularity that we did not want to pursue. We have now refined the initial subset of variables considered for analysis in this paper to 50 Census variables. Census data is provided as a head count for categories for many different questions, and there are thousands of variables that could be constructed from the raw data. These 50 variables are chosen because they appear in previous studies of Australian elections (those mentioned in the introduction) and we believe that they capture many key characteristics of an electorate - age, income, employment, religion, birthplace and relationships.

4. *...Then you use the PCA to identify variables that covary. If variables appear together in a PC with loading higher than 0.15 they are combined together into a factor, e.g. a new variable Income replaces two previously existing variables. (In the description of this it is unclear how many variables are being combined in each factor, the 'ands' and commas make things tricky, .e.g is 'renting and government housing' one variable or two? Do the factors use the loading weights or just the signs?)*

We have made an effort to clarify this process in Section 3.2, which now details each of the variables being combined into each factor. For consistency, only commas are used to separate the variable descriptions.

The factors are calculated using the signs of the loading weights from PCA, but not the magnitudes. A total of 23 variables are combined into 6 different factors.

5. *The steps above get you from 65 variables down to 30 (I can count 10 variables that get accounted for via the creation of factors, but that doesn't get from 65 to 30 so presumably some other variables have been removed? Or is this difference all in the age bands and variants of Christianity? More clarity is required)*

This is reconciled in Section 3.2. We begin with 50 variables. Six factors are created using a total of 23 variables. One of the age brackets (population aged 55 and over) is removed, and the other three age brackets are kept (to avoid multicollinearity). This results in a final predictor set of 32 variables. Any variables that are not contained in a factor are not discarded, and are kept in the model as explanatory variables.

6. *As you note in the paper, it seems likely that there are still issues with multi-collinearity. Figs 6 and 7 show the variables with significant effects, do the results change much if you fit a model with just these 13 variables instead of the 30?*

This was a very helpful suggestion to explore, and as such an additional section on robustness has been included in the revised paper (Section 4.4). When we fit a

model using just the variables with significant effects, we find that the results are essentially unchanged. Each of the estimated coefficients from the reduced model (containing only the subset of significant variables) lie within the 95% confidence interval from its corresponding full model (containing all 32 variables).

We have also included two further multicollinearity checks (see Section 4.4). No definitive signs of multicollinearity are observed in any of these three checks.

7. *Other minor points*
- *line 120, each of the*
- *line 129, is it really the boundaries that are the target /source zones or the areas defined by these boundaries?*
- *line 241, must be modelled*
- *line 261-266, is there anything in the fact that the spatial coefficient is significant only in the years where no imputation was done?*

The first three minor issues have been corrected. For the last issue, the spatial coefficient is now significant in 2001, 2004 and 2016, which no longer corresponds only to years without imputation.

Dear Associate Editor,

Thank you for your helpful comments on our paper titled 'Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016'. We have revised our paper to address the issues you raised and have detailed a summary of the changes made in response below.

Kind regards,

Jeremy Forbes, Di Cook and Rob Hyndman

Associate Editor's Comments

1. *The main thing I found confusing about this paper was the variable selection procedure. It seems to be heading in the direction of four PCs, then it veers into six factors before ending with 30 variables. This all requires greater clarity and justification.*

We have reworked Section 3.2 to better explain this process.

We begin with 50 explanatory variables. Many of these variables represent similar information and hence we argue, could be grouped together. PCA is used to see which variables covary, and so we use it to determine the subsets of variables that should be combined into individual factors.

Each of the first four PCs are examined to determine which variables have large loadings. If there is an intuitive reason as to why a subset of variables with large loadings (in a given PC) should be grouped together, then they are combined into a factor using a weighted sum. The weights are simply +1 if the loading in that PC is positive, and -1 if it is negative.

Six factors are created using a total of 23 variables. One of the age brackets (populated aged 55 and over) is removed, and the other three age brackets are kept (to avoid multicollinearity). This results in a final predictor set of 32 variables. Any variables that are not contained in a factor are not discarded and are kept in the model as explanatory variables.

2. *The authors opt for "visualization" of the fixed effects rather than the usual table of estimated parameters. Perhaps this is a good choice given that there are so many. But then spatial parameters are exhibited in a huge table. This seems inconsistent.*

Partial residual plots are presented to give the reader a visual feel for the direction, size and significance of an estimated effect, as well as the distribution of the data points (for that covariate). Whilst there is duplication of information here, we think that the partial residual plots play an important role in illustrating the effect of a given variable across elections, and that the estimated model parameters should be included for completeness.

3. *I found the writing unclear in a few places. What are the "time-varying restrictions" on a variable (line228)?*

We have revised our explanation relating to the model specification (Section 3.3). Each election is modelled as a separate cross section, which can be considered a special case of a longitudinal model where all coefficients are time-varying and heteroskedasticity is time-varying. A time-varying coefficient is a coefficient that a different estimate for each time period $t = 1, \ldots, T$. We have removed the note about "time-varying restrictions".

4. *And the description of a partial residual (line 276) doesn't make sense to me.*

Partial residuals, for a variable X1, are the residuals from the fitted model with the estimated effect of X1 added to each. Partial residual plots show the direction, size and significance of an estimated effect, as well as any deviations from linearity (see Section 4.3).

5. *Some of the interpretations of results seem a little naïve. For example, does Figure 4 really justify concluding that "electorates effectively voted independently" except for 2001 and 2016? (Is it a coincidence that the spatial autoregression is greatest in 2001 and 2016 where there is no need for imputation?)*

In retrospect, we agree that this is perhaps a naïve conclusion. It is possible that unobserved variables are being captured in the spatial effect, and as such we have removed this interpretation from the revised paper. Additionally, on revision of the imputation method, we find that the spatial effect is significant in 2004 as well as 2001 and 2016.

Dear Anonymous Referee (1),

Thank you for your comments on our paper titled 'Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016'. The feedback you provided was very beneficial, as it helped us refine our analysis and make some adjustments to the paper. We have responded to the points you made below, and in many cases, have made changes to address your concerns.

Kind regards,

Jeremy Forbes, Di Cook and Rob Hyndman

Referee 1's Comments

1. *Its goal is to account for – though the word 'explain' is used! – patterns of voting at Australian federal elections, using ecological data, but with no reference to the ecological fallacy.*

We have expanded our remarks on the ecological fallacy (previously in section 4.3) and moved them to the introduction (Section 1) to give them prominence. We are explicit in stating that "drivers of individual voter behaviour may vary from what is observed at the electorate level", which may not have been clear enough in the original submission.

2. *The paper is written in virtual total ignorance of the large amount of work done on Australian voting patterns; apart from reference to a few papers by geographers there is not a single mention of work by political scientists – of which there is a great deal. As a consequence there is no theoretical background whatsoever, let alone hypotheses to be tested. It is banal empiricism: throw a lot of data (what are called 'electoral socio-demographic variables') at something and see what comes out.*

We formed our approach to exploring this question by reviewing the Australian literature in this area, which has been referenced accordingly. As this is an exploratory piece, we wanted to let the data reveal the relationships themselves, rather than imposing and testing explicit hypotheses.

3. *The banality of this approach is illustrated by the finding that 'more de facto relationships' are associated with greater support for Labor. No rationale for this is offered – why should people in such relationships be more likely to vote Labor? And while there may be a statistically significant relationship (i.e. slope coefficient) it may be substantively trivial – a fact hidden because the data are transformed. If there is a small SD then the 'real' difference between a place with*

many de facto relationships and another with few will be small and will not account for much of the absolute variation in Labour support at all. Further, since all of the relationships with individual variables are partial, it may well be that when 'all other variables are held constant' the difference in Labour support is trivial.

Since this study is of electorates, rather than individual voters, changing an electorate's demographics can have an effect on two-party preference beyond just the head count of population whose demographics are changed. In particular, the effect of de facto relationships on the electorate's vote is not necessarily just as a result of the votes from people who are in a de facto relationship.

4. *Much is done – unnecessarily in my view – in preparing the data through imputation. This has two components. First there is the creation of census data for the 150 constituencies because boundaries of the large census reporting areas may not coincide with the constituencies' at the time of an election. Imputation is deployed by assuming that if 50% of area A is in district X then 50% of A's population is allocated to X without any recognition that area A may be internally heterogeneous. But that is in any case unnecessary. Australian census data are reported – as the authors note – for very small areas with populations of only a few hundred each, and there are shape files associated with these areas. It is thus a straightforward task to amalgamate the small areas into the electoral districts and get the 'real' data for each without any (or at best a very, very small amount of) imputation.*

We have revised our imputation method to reflect this suggestion. We now use Census data at the lowest level of aggregation available — the geographic region type of 'Statistical Area 1' — to obtain the approximation of electoral boundaries (see Section 2.2). Some spatial imputation is still required as the SA1 boundaries do not exactly match the electorate boundaries. Despite these changes, our results are effectively the same as the initial submission.

5. *The authors recognise that there is likely to be collinearity among 65 Census variables – why those; we are never told?*

The (now) 50 Census variables are chosen because they appear in previous studies of Australian elections (those mentioned in the Introduction), and we believe that they capture key characteristics of an electorate - age, income, employment, religion, birthplace and relationships.

Note that we have refined the subset of variables considered for analysis in this paper to 50 Census variables. Previously we started with 65 variables, but many of these variables were removed in the variable selection process because they provided additional granularity that we did not want to pursue.

6. *And so, as others do, they did a pca of the data matrix, reducing it to four components – we are not told what rotation was employed to achieve simple structure. We are not given the results of this analysis, merely a naming of the six factors identified – how do you get six 'factors' from four components?*

We have reworked Section 3.2 to better explain this process.

Our approach is to combine the six election datasets together into a single dataset of 900 observations, containing 50 socio-demographic variables. Many of these variables represent similar information and hence we argue, could be grouped together. PCA is used to determine which variables covary, and so we use it to guide the selection of variables to be combined into individual factors. Each of the first four PCs are examined to see which variables have large loadings. If there is an intuitive reason as to why a subset of these variables with large loadings (in a given PC) should be grouped together, then they are combined into a single factor using a weighted sum. The weights are simply +1 if the variable's loading in that PC is positive, and -1 if it is negative. Hence, the factors are not the principal components themselves, rather they are each a weighted sum of a subset of variables.

The subsets that comprise each of the six factors are explicitly detailed in paragraph four of Section 3.2.

7. *My expectation was that they would then fit a model using the four components as the independent variables, thus avoiding any collinearity and confounding. But they didn't. Instead they selected 30 of the 65 variables – which must still be collinear; the resulting equation in Table 2 clearly implies this given the large number of (a) insignificant relations (normal with collinearity – see a recent paper in Quality and Quantity), (b) unreasonably large coefficients (-8.07 for de facto, for example), and (c) very large standard errors for many of the variables.*

In response to this suggestion, we have now added an assessment of multicollinearity within the robustness section of the revised paper (Section 4.5). Three robustness checks are conducted as follows:

- We fit a model using just the variables that have a significant effect in at least one election. In this model, each of the estimated coefficients lie within the 95% confidence interval from its corresponding full model (containing all 32 variables).

- We also checked the effect of omitting a variable that is contained in the ten largest pairwise correlations. For each pair, a model for each election is refit omitting one of the two variables. It is found that for each of these pairs, the estimated effect of the remaining variable in the reduced model lies within the 95% confidence interval from the full model.

- A third and final check is the visual exploration of different variable projections using a tour (for each election). No definitive signs of multicollinearity are observed.

8. *The authors fit a spatial autoregressive model, on the assumption that neighbouring places are more likely to be similar than distant places (the 'first law of geography'). This is only a significant finding at two of the elections and is wrongly interpreted: it does not necessarily show that then 'the political climate of an electorate appears to be affected by the attitudes of its neighbours' but, much more likely, that neighbouring places share characteristics in common that affect voting behaviour but are not captured by the independent variables included in the model. Spatial autocorrelation may reflect spatial diffusion processes – but in which direction? – but is much more likely to reflect an under-specified model.*

In retrospect, we agree that this is perhaps a naïve conclusion. It is possible that unobserved variables are being captured in the spatial effect, which is the interpretation that we have now included. We have removed the old interpretation from the revised paper. Additionally, on revision of the imputation method, we find that the spatial effect is significant in 2004 as well as 2001 and 2016.

Dear Anonymous Referee (2),

Thank you for your comments on our paper titled 'Spatial modelling of the two-party preferred vote in Australian federal elections: 2001–2016'.

We have revised our paper to address the issues raised. Below are detailed summaries of the changes made in response to each of your points

The feedback you provided was very beneficial, as it helped us refine our analysis and make some adjustments to the paper. We have responded to the points you made below, and in many cases, have made changes to address your concerns.

Kind regards,

Jeremy Forbes, Di Cook and Rob Hyndman

Referee 2's Comments

1. *It seems very strange to assume that population is uniform within electorates (line 133) so that weighting is proportional to area, and not to the number of people in an area of intersection. An electorate on the urban fringe may be strongly affected by loss of a relatively small but highly populated area but little affected by incorporating a large rural area. A simple indirect way to get at this might be to address the question: Would the results change much if A_{s,t} was log area, for each s,t ?*

We have revised our imputation method to avoid this assumption. We now use Census data at the lowest level of aggregation available - the geographic region type of 'Statistical Area 1' (SA1) - to obtain the approximation of electoral boundaries (see Section 2.2). There are on average approximately 400 people in each SA1. This approach will successfully capture any densely populated areas on the fringe of an electorate. Our original method was chosen because we believed that the imputation method would not materially affect results. Despite these changes, our results are effectively the same as the previous submission.

2. *It seems to me that the use of Principal Components analysis is unclear. (a) PC appears to have been used as a model selection criterion, with the variables related to the first four PCs included into the model in Table 2. However typically when principal components are computed they only involve the explanatory variables, not the response, and so there is no guarantee that all the first four PCs (and only them) will be related to the response. For example using the eechidna data for 2016 I found the fifth PC was extremely significant. But this is a minor point.*

We have used principal components to inform which variables should be combined into factors, rather than use them directly as explanatory variables. This is done as a step to address any (near) multicollinearity in the explanatory variables. Any variables that are not used in creating a factor still remain in the model. Of course, there is no guarantee about any of these factors being related to the response. We have now explained our variable selection procedure in more detail in the revised paper (Section 3.2).

3.   *The variables which load highly on a PC are likely to be highly correlated, and so if these high-loading variables are used together in the regression model we may again have multicollinearity. I found this was the case when applying the author's regression model to the 2016 data. This may be why several of the predictors in Table 2 are never related significantly related to the response in any of the years.*

By taking a subset of variables with high loadings on a principal component and combining them into a single variable, we are trying to address potential multicollinearity. Two factors that have been created from the same PC will be correlated, which could mean that (near) multicollinearity persists. However, we have conducted three robustness tests for multicollinearity (see Section 4.5), and no conclusive signs of multicollinearity are observed.

The first check involved fitting a model using just the variables with significant effects. We find that the estimated coefficients in the reduced model (containing only the subset significant variables) all lie within the 95% confidence interval from its corresponding full model (containing all 32 variables).

We also checked the effect of omitting a variable that is contained in the ten largest pairwise correlations. For each pair, a model for each election is refit omitting one of the two variables. It is found that for each of these pairs, the estimated effect of the remaining variable in the reduced model lies within the 95% confidence interval from the full model.

A third and final check is the visual exploration of different variable projections using a tour (for each election). No definitive signs of multicollinearity are observed.

4.   *Conversely there could be a variable which is independent of the other variables (so representing 1 / 65th of the explanatory information) but by dint of that fact not be included in the first four PCs. In short, the narrative leaves me unconvinced that all the variables included are necessary or that all the relevant variables are included in the regression.*

Any variables that are not used in creating a factor still remain in the model. The four PCs are used only to inform which variables are correlated and should be combined into a factor. For example, median household, personal and family incomes all have large positive loadings in the first principal component, so we

combine them into a factor called `Incomes`. Any variable that does not have a large loading in any of the four PCs are not discarded - they remain in the model. It is of course possible that there exist variables beyond what we have considered that could affect two party preference. However, we believe that the variables we have used cover the core demographic components of an electorate - age, income, employment, religion, birthplace and relationships.

5. *I presume the "factors" referred to lines 210-220 were created as the first principal component of the constituent raw variables?*

In our revised submission, we have made an effort to clarify what we have defined a "factor" to be (Section 3.2). A "factor" is created when there is evidence to suggest that a subset of variables are explaining similar information, and if they were all included in the model as separate variables, there would be issues with multicollinearity. To inform which variables should be grouped, we use the first four PCs from PCA. In a single PC, if there are is a subset of variables with large (positive or negative) loadings that intuitively could represent similar information, we combine them into a factor using a weighted sum, where the weights are +1 or -1 depending on the sign of the loading in the that PC. This means that many factors can be created from a single PC, and that any variables that do not appear in a factor are still included in the model as stand-alone variables.

6. *The description of the variables included in "Education" is not clear.*

We have provided more details on the variables that are combined into each of the factors in Section 3.2, including `Education`. `Education` is a weighted sum of seven variables that represent: high school completions, undergraduate and postgraduate degrees, proportion of employed 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.

7. *I could not get the variable 'OtherLanguageHome' in 2016 to be significant, but it also had a high VIF with other explanatory variables. Please check this.*

Because of the revised imputation method, the data is now slightly different (however the conclusions do not change). We have checked that `OtherLanguageHome` is significant in 2016.

8. *I am not convinced it is reasonable to summarise the effect of age by median age. Perhaps age effects are more at the edges of the age distribution than in the middle? I would have thought the percentage of the population in different age bands e.g. 20-34 years and Over 65years, might be significant. For 2016 the % aged 20-34 years was highly significant and removed median age.*

This was a very useful suggestion. Instead of using median age, we summarise age of an electorate using its distribution amongst age brackets 0-19, 20-34, 35-54 and 55+ years. We find that higher populations in the 20-34 and 35-54 age brackets are associated with electoral support for Labor in all of the six elections, whilst populations 0-19 years have weakly positive associations with the Liberal party (aside from 2007, when this relationship was significantly positive).

9. *Is there any tendency for the vote in an election to depend on how long the incumbent has been in power? In your cross-sectional models this would be confounded with any other effects in the intercept, but a comment would be nice to know. It would help in terms of interpreting the variations in intercept, and setting a reader's mind to rest that the issue has been considered. Perhaps a something like a simple regression of average vote (or the intercept) on length of incumbency could address the issue.*

This is an interesting question. We have included some boxplots of the residuals from each election, split by whether the party was the incumbent party. We find that the incumbent party has a clear advantage of around 3-5 percentage points. This can be found in Section 4.4.