**Swings are far from uniform**

Last week I came across a [Bayesian state space model of two-party-preferred voting intention](http://freerangestats.info/blog/2019/03/02/aust-election-1) for Australian federal elections. It treats the surveys from various polling firms as imperfect (potentially systematically imperfect) measurements of an unobserved latent variable of “true” voting intention, which manifests itself only every few years in the form of election results. The model estimates that latent value for every day that measurements exist, and going forward with increasing uncertainty past when there are measurements.

The most likely day for the next federal election seems to be 18 May 2019. In theory the House of Representatives election could be as late as November, but as half the Senate needs to go to the polls by end of May and most people prefer to have fewer elections not more the government is highly likely to opt for a simultaneous election of the House and Senate in May.

To turn my model into an actual election forecast I need a way of converting the distribution of probable two-party-preferred results on a given day into seats in the House of Representatives and hence probabilities for who will form a new government. In doing this we come up against the issue of the moderately high degree of spatial variation in voting behaviour in Australia (as elsewhere), in combination with the spatially-based conversion of vote into seats (ie regional divisions).

A key tool used in election analysis is to think in terms of the “swing” in vote in each division of the electorate, comparing vote in the last election to that in the next. This is just a practical application of a standard method in longitudinal statistical analysis. Analysis based on differential “swings” is basically the same idea as “difference in differences” analysis in econometrics and other social sciences.

It’s common as we come towards an election for pundits to project how a given swing would impact on seats if the swing were “uniform” ie the same across all divisions. But of course everyone knows swings are far from uniform. Here we can look at all the swings since the 1993 election (Australian Labor Party (ALP) victory –) to 2016 (Liberal-National Coalition under Malcolm Turnbull hanging on to government by the skin of their teeth with help from the cross bench):

It’s interesting to see in this way how John Howard lost nearly all of his substantial 1996 gains in the 1998 election, the biggest average anti-government swing in this data to not result in a change of government. Then regained a lot in the 2001 election (those of us who lived through it will need [little reminder](https://en.wikipedia.org/wiki/2001_Australian_federal_election) that this was the election of the September 11 attacks in the USA, and the ‘Children Overboard’ and Tampa incidents). Whether one thinks it good or bad, we can agree that if Beazley had snuck in in 1998 and Howard had been a one-term Prime Minister, Australian history would have gone down quite a different path.

This highlights one of the obvious issues with swings – big swings in one direction are predictive of a swing back to the centre. But the main aim of the chart above is to show how the individual divisions’ swings vary more than the average (of course).

After controlling for the average swing, the standard deviation of swings is still 3.2 percentage points. This adds a big chunk of uncertainty to how an average swing will translate into seats. A sitting member could reasonably expect their own swing (for or against) to be within plus or minus six percentage points of the overall swing, but in any given election we’d expect several division-level swings to be even more divergent than that.

Here’s the R code that produces that chart and that conclusion of 3.2 percentage points. It uses my new ozfedelect R package which is the workhorse I intend to use for re-usable functionality and data for Australian federal election analysis. So far it includes tidied polling data from 2007 to present (updated 11 March 2019), division-level two-party-preferred election results back to 1993, and “simple features” versions of electoral boundaries back to 2010 (simplified to 10% of their original size). Thanks to Wikipedia (and the polling firms of course), the Australian Electoral Commission, and the Australian Bureau of Statistics respectively for making the original data available.

*Post continues after code extract*

library(ozfedelect) # with data

library(tidyverse)

library(scales)

library(grid)

library(Cairo)

library(sf)

# Some checks.

# -7.5 for Aston in 2013 meant a swing against the Labor govt

# - 10.1 for Bass in 2016 meant a swing against the Lib/Nat govt

results\_2pp\_div %>%

select(division\_nm, swing\_to\_govt, election\_year) %>%

spread(election\_year, swing\_to\_govt)

#---------------------------explore distribution of swings---------

# Let's try to understand those historical swings

d <- results\_2pp\_div %>%

group\_by(election\_year, incumbent) %>%

mutate(avg\_swing = sum(swing\_to\_govt \* total\_votes) / sum(total\_votes)) %>%

ungroup() %>%

filter(abs(swing\_to\_govt) < 20) %>%

mutate(year = fct\_reorder(as.ordered(election\_year), avg\_swing))

# We are interested for future forecasting models in the variance of division-level swings

# that are not explained by the nation-wide swing

model <- lm(swing\_to\_govt ~ avg\_swing, data = d)

confint(model)

coef(model) # of course the slope is 1 and the intercept is 0, basically - by design

# the interesting thing is actually the residual standard error:

summary(model)

# what's the residual standard deviation of division-level swings?

residual\_sd <- summary(model)$sigma

avgs <- distinct(d, avg\_swing, year, incumbent)

annotation\_col <- "grey50"

update\_geom\_defaults("label", list(family = main\_font, fill = "white", colour = annotation\_col, alpha = 0.8))

# Main chart:

d %>%

ggplot(aes(x = avg\_swing / 100, y = swing\_to\_govt / 100)) +

geom\_smooth(se = FALSE, method = "lm") +

geom\_vline(xintercept = 0, colour = annotation\_col) +

geom\_hline(yintercept = 0, colour = annotation\_col) +

geom\_point(aes(colour = incumbent)) +

geom\_text(data = avgs, y = -0.18, aes(label = year, colour = incumbent)) +

labs(caption = "Source: Australian Electoral Commission data, analysed by freerangestats.info.") +

scale\_x\_continuous("Overall swing towards incumbent government", label = percent\_format(accuracy = 1)) +

scale\_y\_continuous("Division level swings to government", label = percent\_format(accuracy = 1)) +

annotate("label", x = -0.053, y = 0.06, label = "Strongly change\ngovernment") +

annotate("label", x = -0.045, y = -0.065, label = "Narrow escape for\nHoward in 1998") +

annotate("label", x = 0.0055, y = 0.04, label = "Strongly retain\ngovernment") +

scale\_colour\_manual("Incumbent government:",

values = c("ALP" = "#e53440", "Lib/Nat" = "#1c4f9c")) +

ggtitle("Individual seats face more voting uncertainty than Australia as a whole",

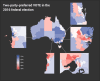
"Each point represents the swing in a single seat; residual standard deviation of about 3.2 percentage points around the nation-wide swing.")

**Swings vary spatially as well as over time**

Election analysts love our choropleth maps, but they pose a challenge because of the uneven distribution of people over land. This is a particularly acute problem for Australia.

In trying to draw a meaningful electorate choropleth map for Australia I first tried a cartogram (to distort borders so each electorate is represented as visually equal while retaining some of its basic shape) but this was too big a job for the algorithm in the cartogram R package. Fair enough. Some of those divisions are pretty large, including Durack in northern Western Australia, at 1.6 million square kilometres for its 100,000 electors.

So instead of a cartogram, I’ve opted instead for a bunch of call-outs of the five largest cities. Here we can see the two-party-preferred *votes* in the 2016 election:

[](http://freerangestats.info/img/0146-2pp-votes-2016.svg)

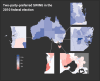
*Clicking on the image will open a scalable vector graphics version in a new tab.*

This call-out method seems effective in counteracting the big sea of blue that one gets from looking at the simple map of Australia, with Liberal-National Coalition support dominant in rural areas outside the Northern Territory.

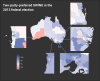
*Note for those more used to American politics. In Australia, the centre-left party (ALP) has red as its colour, and the right use blue, unlike in the USA. Unlike in many other countries the Democrats in the USA did not grow out of a nineteenth and twentieth century socialist movement and hence never self-identified with the colour red, and when TV stations got to allocate colours they seemingly did it arbitrarily. Also, while we’re talking about such things, the Liberal Party in Australia is a party of the right or centre-right, and historically referred to economic, individual/family liberalism in contrast to the collective action advocated by the labor movement and their party. End of digression.*

So, here’s how this mapping approach looks with today’s issue of interest, division-level variation in swings.

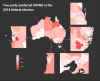
In 2010, we see the ALP under Julia Gillard taking a thumping and retaining government only as a minority in the House, with support from three independents and an Australian Green. Counter-swings to the ALP in the Melbourne and Adelaide areas (the two call-outs at the bottom of the map) saved the ALP from an outright loss:

[](http://freerangestats.info/img/0146-2pp-swing-2010.svg)

Then we see ALP under Kevin Rudd comprehensively losing their remaining wafer-thin hold of government nearly across the board in 2013. The average swing here was less than that against Howard in 1998, but the government had no buffer at all so the result was a clear win for the Liberal-National Coalition under Tony Abbott:

[](http://freerangestats.info/img/0146-2pp-swing-2013.svg)

Then in 2016 we saw most areas swing back, with Malcolm Turnbull’s narrow retain on government already mentioned:

[](http://freerangestats.info/img/0146-2pp-swing-2016.svg)

The ozfedelect package includes a function that does all the grunt work of these maps, so the end-user code that produces them is extremely simple:

*Post continues after code extract*

ozpol\_infographic(2016, variable = "swing\_to\_govt")

ozpol\_infographic(2013, variable = "swing\_to\_govt")

ozpol\_infographic(2010, variable = "swing\_to\_govt")

ozpol\_infographic(2016)

ozpol\_infographic(2013)

ozpol\_infographic(2010)

**Swing and vote each have a relationship with division-level census variables**

A common thing to do with either swing or vote variables is to compare them to division-level indicators of social, economic and demographic variables, typically available from the Census. Hugh Parsonage’s invaluable Census2016.DataPack contains tidied versions of the ABS’ Census Data Packs, and I’ve further summarised a small subset of the variables available into a one-row-per-division collation in ozfedelect. Here’s what we see when we compare those variables to the most recent swing in 2016:

The clearest thing to see here is that divisions with a high proportion of young persons (aged 19 and under) tended to swing away from the Liberal-National Coalition. If we put these variables, and the state territory (as a factor) into a simple model, we can see that there is a distinct state-based flavour to the swings, that divisions with a high number of people born in Australia tended to swing towards the government, and that those with many young people (which might be a proxy for either or both of students or young families) swung against it.

In addition to addressing the individual/regional problem or ecological fallacy, ideally, I would be comparing “swing” (which is effectively the first-difference in the “vote” time series) not against absolute levels of such variables but changes in them – so we were looking at the first-difference for both variables, and could see if a change in number of young people was leading to a change in vote. But this would need more data tidying than I’ve done yet.

For the record, here is how these same seven variables compare to the more natural comparator, the simple two-party-preferred vote:

As might be expected from a casual observation of Australian politics where the class origins of the main parties remain stubbornly persistent, areas with these characteristics are more likely to vote for Liberal-National Coalition:

* Higher proportion of adults have year 10 or more education
* Higher proportion of people were born in Australia
* Higher proportion of people speak only English at home
* Higher average personal income

Again, don’t mistake these area-based findings for individual-level conclusions. That may come later if I have time (this is just a hobby project after all; but I note the AES data is freely downloadable and will be hard to keep my hands off!).

Code for all these comparisons to Census data. Again, the hard yards have been done by Parsonage’s Census2016.DataPack and my ozfedelect.

*Post continues after code extract*

#----------------------- explore relationship to census variables------------

# a tiny bit of munging data together, into wide form for modelling:

d1 <- results\_2pp\_div %>%

filter(election\_year == 2016) %>%

left\_join(div\_census\_2016, by = "division\_nm") %>%

select(division\_nm, state\_ab, lib\_nat\_percentage, swing\_to\_govt, young\_persons:only\_english\_spoken\_home) %>%

mutate(state\_ab = fct\_relevel(state\_ab, "NSW"))

# and narrow form for plotting:

d2 <- d1 %>%

gather(variable, value, -division\_nm, -lib\_nat\_percentage, -swing\_to\_govt, -state\_ab)

#------------charts--------------

# chart of vote:

d2 %>%

ggplot(aes(x = value, y = lib\_nat\_percentage / 100)) +

facet\_wrap(~variable, scale = "free\_x") +

scale\_y\_continuous("Two-party-preferred vote for Liberal/National Coalition") +

geom\_smooth(method = "gam") +

geom\_point(aes(colour = state\_ab)) +

ggtitle("Vote compared to census variables by electoral division",

"2016 federal election") +

labs(colour = "",x = "")

# chart of swing:

d2 %>%

ggplot(aes(x = value, y = swing\_to\_govt / 100)) +

facet\_wrap(~variable, scale = "free\_x") +

scale\_y\_continuous("Two-party-preferred swing towards Liberal/National Coalition\n(incumbent government)") +

geom\_smooth(method = "gam") +

geom\_point(aes(colour = state\_ab)) +

ggtitle("Swing compared to census variables by electoral division",

"2016 federal election") +

labs(colour = "",x = "")

#----------Modelling -----------

d3 <- d1 %>%

select( -division\_nm, - lib\_nat\_percentage) %>%

mutate\_at(vars(young\_persons:only\_english\_spoken\_home), scale)

mod1 <- lm(I(swing\_to\_govt / 100) ~ ., data = d3)

# chart of results:

confint(mod1)[-1, ] %>%

as.data.frame() %>%

mutate(var = rownames(.),

var = gsub("state\_ab", "", var),

var = gsub("\_", " ", var)) %>%

rename(lower = `2.5 %`,

upper = `97.5 %`) %>%

mutate(mid = (lower + upper) / 2,

var = fct\_reorder(var, mid)) %>%

ggplot(aes(x = lower, xend = upper, y = var, yend = var)) +

geom\_vline(xintercept = 0, colour = "steelblue") +

geom\_segment(size = 3, colour = "grey") +

scale\_x\_continuous("Impact of change in one standard deviation in census variable on swing",

label = percent) +

labs(y = "",

caption = "Source: ABS Census data, AES election results, analysis by http://freerangestats.info") +

ggtitle("Division-level variables related to a swing to the Liberal-National Coalition",

"Comparing the 2016 results to 2013 by electoral division (or 'seat').

Conclusions about individual characteristics relating to vote should be drawn only with great caution.")

**And the answer is…**

Finally, enough exploring, and on to the answer. What’s the probability of an ALP government after the election, if it is in late May? The latest version of the model, with one more poll than my last blog post, and now extending only to late May not to November, looks like this:

The mean predicted two-party preferred vote for ALP at end of May is 52.48 percent and the standard deviation is 2.08 percentage points. We can combine the expected distribution of votes at the time, with the historical distribution of individual divisions’ swings on top of that, to get an estimate of the chance of a change of government.

So here’s my current best effort at converting modelled two-party-preferred vote into seats:

This gives the ALP a healthy 83% chance of winning; better than my stab in the dark a week ago on social media of 71%, when I had been thinking in terms of a November election which I now realise is unlikely.

Beware that this is a *very* simple model of seat allocation, and one I’m a bit untrusting of. For example, it doesn’t consider redistributions in electoral boundaries since 2016 at all. And I haven’t taken into account at all the *six* current independents or members of minor parties, which surely has an impact; never mind other seats where high profile independents might have a chance of knocking out a major party member that would not happen under straight two-party-preferred dominance.

So treat this carefully for now, this is just a first step towards a proper model. Over the next few weeks I hope to robustify it and productionise into something that can easily update itself as we move towards election day.

Final snippet of R code that did this simulation, based on a swing from the 2016 division results.

*Post continues after code extract*

#==============crude simulation===============

# so plausible to

# a) model the overall swing to the govt

# b) simulate for individual seats a randomness of N(0, 3.2) on top of that overall swing

# c) add those swings to the 2016 results

# Important - this misses out on two things that are needed:

# a) 2 new electorates this time around

# b) seats held by other parties

last\_result <- results\_2pp\_div %>%

filter(election\_year == 2016) %>%

summarise(alp\_2pp\_2016 = sum(alp\_votes) / sum(total\_votes)) %>%

pull(alp\_2pp\_2016)

# last election the ALP got 49.6% of the 2pp. Currently on track to about 52.48 with sd of about 2.08.

# So an average swing against the govt of N(2.88, 2.08), and division-level randomness on top of that.

baseline <- results\_2pp\_div %>%

filter(election\_year == 2016) %>%

select(division\_nm, alp\_percentage) %>%

mutate(link = 1)

nsims <- 1e5

set.seed(321)

sims <- tibble(sim = 1:nsims, link = 1) %>%

mutate(avg\_swing = rnorm(n(), -2.88, 2.08)) %>%

full\_join(baseline, by = "link") %>%

select(-link) %>%

mutate(extra\_swing = rnorm(n(), 0, residual\_sd)) %>%

group\_by(sim) %>%

# we scale the extra\_swing to be mean zero so we aren't accidentally changing the average swing:

mutate(extra\_swing = extra\_swing - mean(extra\_swing),

total\_swing = avg\_swing + extra\_swing,

alp\_percentage\_2019 = alp\_percentage - total\_swing) %>%

ungroup()

sims\_by\_div <- sims %>%

group\_by(sim) %>%

summarise(avg\_swing = unique(avg\_swing),

number\_seats\_alp = sum(alp\_percentage\_2019 > 50),

prop\_seats\_alp = mean(alp\_percentage\_2019 > 50)) %>%

ungroup()

m <- sims\_by\_div %>%

summarise(m = round(mean(prop\_seats\_alp > 0.5) \* 100, 1)) %>%

pull(m)

# draw histogram of results

sims\_by\_div %>%

ggplot(aes(x = number\_seats\_alp)) +

geom\_histogram(alpha = 0.5, binwidth = 1, colour = "grey") +

geom\_vline(xintercept = 75.5, colour = "steelblue") +

scale\_x\_continuous("Number of House of Representative seats won by ALP") +

scale\_y\_continuous("Number of simulations\n(out of 100,000)", label = comma ) +

ggtitle(paste0(m, "% probability of ALP win in the 2019 Federal Election"))

That’s all for now. Comments are very welcome, including spotting any mistakes; I’ve been working through this at some pace and there’s no doubt the checking and testing is inadequate so I would be surprised if something isn’t wrong somewhere, so I’ll be pleased, not upset, if someone tells me what!

As usual, here are the contributors to R that made all this possible:

thankr::shoulders() %>%

mutate(maintainer = str\_squish(gsub("<.+>", "", maintainer))) %>%

knitr::kable() %>%

clipr::write\_clip()

| **maintainer** | **no\_packages** | **packages** |
| --- | --- | --- |
| Hadley Wickham | 18 | assertthat, dplyr, ellipsis, forcats, ggplot2, gtable, haven, httr, lazyeval, modelr, plyr, rvest, scales, stringr, testthat, tidyr, tidyverse, usethis |
| R Core Team | 11 | base, compiler, datasets, graphics, grDevices, grid, methods, splines, stats, tools, utils |
| Gábor Csárdi | 9 | callr, cli, crayon, desc, pkgconfig, processx, ps, remotes, sessioninfo |
| Kirill Müller | 5 | DBI, hms, pillar, rprojroot, tibble |
| Jim Hester | 4 | devtools, pkgbuild, pkgload, readr |
| Winston Chang | 4 | extrafont, extrafontdb, R6, Rttf2pt1 |
| Jim Hester | 3 | fs, glue, withr |
| Lionel Henry | 3 | purrr, rlang, tidyselect |
| Yixuan Qiu | 3 | showtext, showtextdb, sysfonts |
| Yihui Xie | 3 | highr, knitr, xfun |
| Edzer Pebesma | 2 | sf, units |
| Jeroen Ooms | 2 | curl, jsonlite |
| Dirk Eddelbuettel | 2 | digest, Rcpp |
| R-core | 1 | nlme |
| Vitalie Spinu | 1 | lubridate |
| Michel Lang | 1 | backports |
| Simon Wood | 1 | mgcv |
| Achim Zeileis | 1 | colorspace |
| Gabor Csardi | 1 | prettyunits |
| Simon Urbanek | 1 | Cairo |
| James Hester | 1 | xml2 |
| Justin Talbot | 1 | labeling |
| Roger Bivand | 1 | classInt |
| Jennifer Bryan | 1 | readxl |
| Kevin Ushey | 1 | rstudioapi |
| Max Kuhn | 1 | generics |
| Stefan Milton Bache | 1 | magrittr |
| Martin Maechler | 1 | Matrix |
| Charlotte Wickham | 1 | munsell |
| Matthew Lincoln | 1 | clipr |
| Marek Gagolewski | 1 | stringi |
| Jeremy Stephens | 1 | yaml |
| Deepayan Sarkar | 1 | lattice |
| Claus O. Wilke | 1 | cowplot |
| Rasmus Bååth | 1 | beepr |
| Jennifer Bryan | 1 | cellranger |
| Hugh Parsonage | 1 | Census2016.DataPack |
| Alex Hayes | 1 | broom |
| Peter Ellis | 1 | ozfedelect |
| David Meyer | 1 | e1071 |
| Brian Ripley | 1 | class |
| Simon Urbanek | 1 | audio |
| Jim Hester | 1 | memoise |