Initial Report: Alberta Election Data 2015-2019

Sam Mikes
smikes@apple.com

Abstract

Election data for Alberta for 2015-2019 is collected from government open data sources and analyzed with unsupervised learning. Geographic correlation to 2016 Canada census.

1 Data Sources

Results are available for municipal, provincial, and federal elections. The initial analysis will use only the 2015 Provincial Election results to validate methods.

1.1 Provincial

2015 Election results and geography

`https://open.alberta.ca/opendata/2015-general-election-results`

1.2 2016 Census and Geography

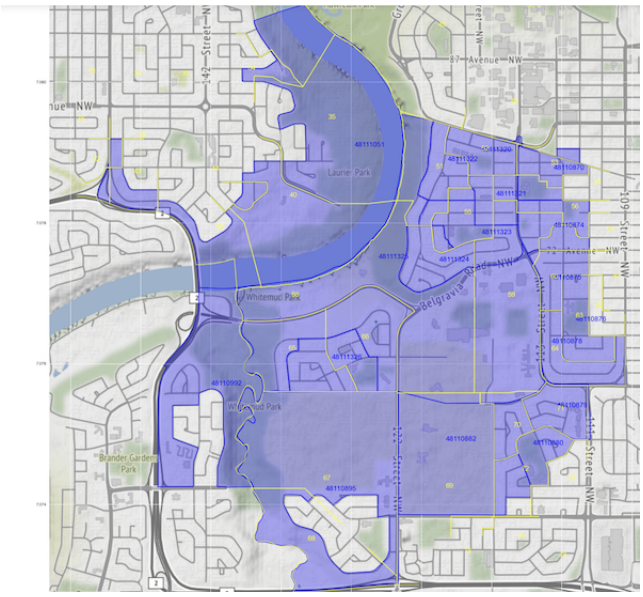
Canada Census data for Alberta is available by geography, organized by 'Census Divisions' (CDs: 20), 'Census Subdivisions' (CSDs: 425), 'Dissemination Areas' (DAs: 5803). Initially we hoped to use 'Dissemination Blocks', which are even smaller than DAs. Becuase Dissemination Blocks may be too small to maintain the privacy of census respondents, we chose to use only DA and larger-sized units.

2 Design

In the 2015 election, there were 87 electoral districts (EDs). For each district, results are reported by 'Poll', where a Poll may be a geographic subset of the district, or a virtual/logical entity such as 'Advance Votes', 'Mail-in Votes', 'Mobile Poll 1'.

Initial analysis attempted to connect results by poll to census data at dissemination areas. This problem (2247x3 features ; 5803 DAs ; 6269 Polls) was difficult to approach. Specifically, the multiple overlaps between census DAs and polls (see figure 1) were difficult to model.

Preprint. Work in progress.



To validate the approach, we followed the suggestion of Gu (personal communication) to first examine the 87 electoral districts in relation to the CSDs. This problem proved to be easily approached with PCA and linear regression, which suggests that dimension reduction should be applied to the larger problem (polls <-> DAs).

3 Data Conditioning

3.1 2015 Electoral Results

Alberta's 2015 provincial election results are reported in a human-friendly Excel format, with column headers varying by the presence of particular party candidates in any given district. A custom python script was written to produce a consistent, flattened summary (data/interim/2015-pge.csv).

For initial analysis, this was further reduced to a sum of votes cast by electoral districts for the three political parties of interest in 2015: New Democratic Party, Progressive Conservative Party, Wild Rose Party. Both the number of registered votes and the number of valid votes were retained for reference. (data/interim/votes-riding-2015.csv.gz)

3.2 2016 Census

The 2016 census reports direct variables (from immediate answers to questions); derived variables (such as median or average income); and coded variables. To simplify the initial analysis, the 2247 reported variables were reduced.

First, the distribution of variables across 'level' was evaluated. Level is indicated by the indentation in the 'census-value-structure.txt' file supplied by Statistics Canada. The distribution of features by level is documented in table 1.

level 0	108
level 1	440
level 2	936
level 3	1241
level 4	1457
level 5	1855
level 6	2133
level 7	2247

Table 1: Canada Census 2016 Variable Levels

We chose to cut at level 2, which meant that 936 variables were retained for further analysis. Because each variable may be reported as Total, Male, Female when appropriate, this is still >2000 features.

Variables with 'Average', 'Median' or 'Total' in the name were dropped, as these are presumptively derived variables. Variables which were always 0 in the Alberta CSD subset were also removed at this point. The resulting data was unstacked to form a single matrix (425 rows x 1232 columns) containing the census features at levels 0, 1, and 2 that were not obviously dependent on other features.

3.3 Geographic Correlation

Geographic correlation was determined by evaluating whether there was overlap between two geometries; if there was overlap, a binary feature (BIN-OVERLAP) was set for that combination of ED and CSD. Overlap was calculated in the commonly-used EPSG:3857 coordinate reference, because of the easy availability of free maps for that CRS. Overall, there were 756 overlaps, and the median overlap of 0.98 indicated that for half the census subdivisions, the subdivision lay almost completely within one ED. Similarly, the 25th percentile overlap (0.0097) indicated that 1/4 of overlaps were nearly insignificant.

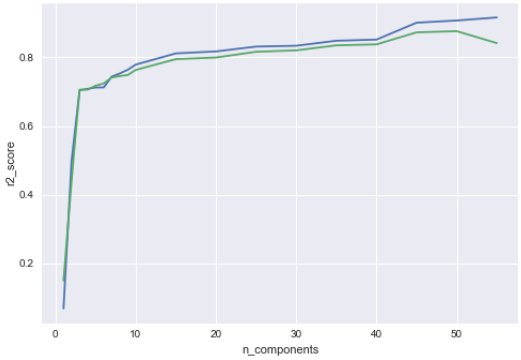
After correlation, the census variables were summed by ED, producing a raw data table of 87 EDs x 1232 census features.

3.4 Target Data

To investigate PCA, we chose two endpoints of interest: the proportion of votes cast in 2015 that went to the NDP (socialist) party on the one hand, and the proportion of votes that went to one of the two conservative parties (Wild Rose and PC). These very nearly sum to 1, so they cannot be considered as independent; more work should be done to check the validity of the PCA.

4 Analysis

Using Scikit-Learn, PCA followed by linear regression was applied to the data. The maximum possible rank is 87; accordingly, we evaluated the r^2 score for $n_{components}$ between 1 and 60. The initial result is that score rises rapidly at first, and flattens off at about 15 components. See figure 2.



4.1 Future Work

The next step is to standardize the dimension-reducing PCA matrix and use it to reduce the DA-level census data to a more tractable size. Then we can investigate whether any meaningful difference is made by weighting the DAs equally (as we have done so far), or in proportion to the overlap area between the DA and the Poll to the total area of the DA.

Separately, an investigation into which census variables are most relevant to the computed principal components should be done, to produce a model with more meaningful independent variables.

Other endpoints such as the fraction of votes for each Conservative party, or voter turnout should also be evaluated in order to confirm that the resolved principal components robustly predict more than one result of interest.

5 Conclusion

When using census data to predict results, one must first focus on an appropriate dimension reduction.