2018 Midterm Election:
Voter Turnout Prediction

Team Korea(U22)

Jinju Park
Donghwa Seo
Boram Shim
Dongseon You
Tae Uk You

CONTENTS

1. Data Extraction
2. Data + Model Selection
3. Model Explanation
4. Prediction
5. Limitations

Theoretical Basis (Literature Review)

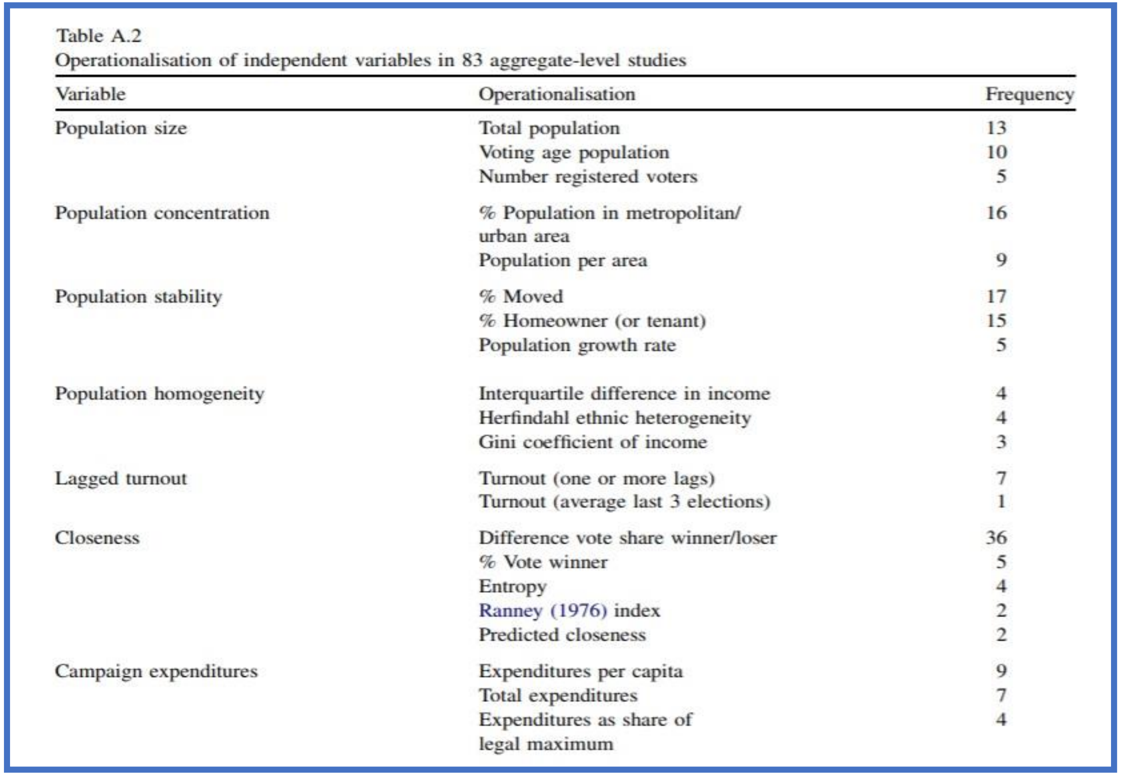


Table A.2
Operationalisation of independent variables in 83 aggregate-level studies

Variable	Operationalisation	Frequency
Population size	Total population	13
	Voting age population	10
	Number registered voters	5
Population concentration	% Population in metropolitan/ urban area	16
	Population per area	9
Population stability	% Moved	17
	% Homeowner (or tenant)	15
	Population growth rate	5
Population homogeneity	Interquartile difference in income	4
	Herfindahl ethnic heterogeneity	4
	Gini coefficient of income	3
Lagged turnout	Turnout (one or more lags)	7
	Turnout (average last 3 elections)	1
Closeness	Difference vote share winner/loser	36
	% Vote winner	5
	Entropy	4
	Ranney (1976) index	2
	Predicted closeness	2
Campaign expenditures	Expenditures per capita	9
	Total expenditures	7
	Expenditures as share of legal maximum	4

Geys, Benny. "Explaining voter turnout: A review of aggregate-level research." Electoral studies 25, no. 4 (2006): 637-663.

United States™ Census Bureau

National Weather Service

Demographic variables:

- Age
- Gender
- Race
- Income
- Education Level

Election Day Temperature

CNN.com

OFFICE OF THE MINNESOTA SECRETARY OF STATE STEVE SIMON

Competitiveness

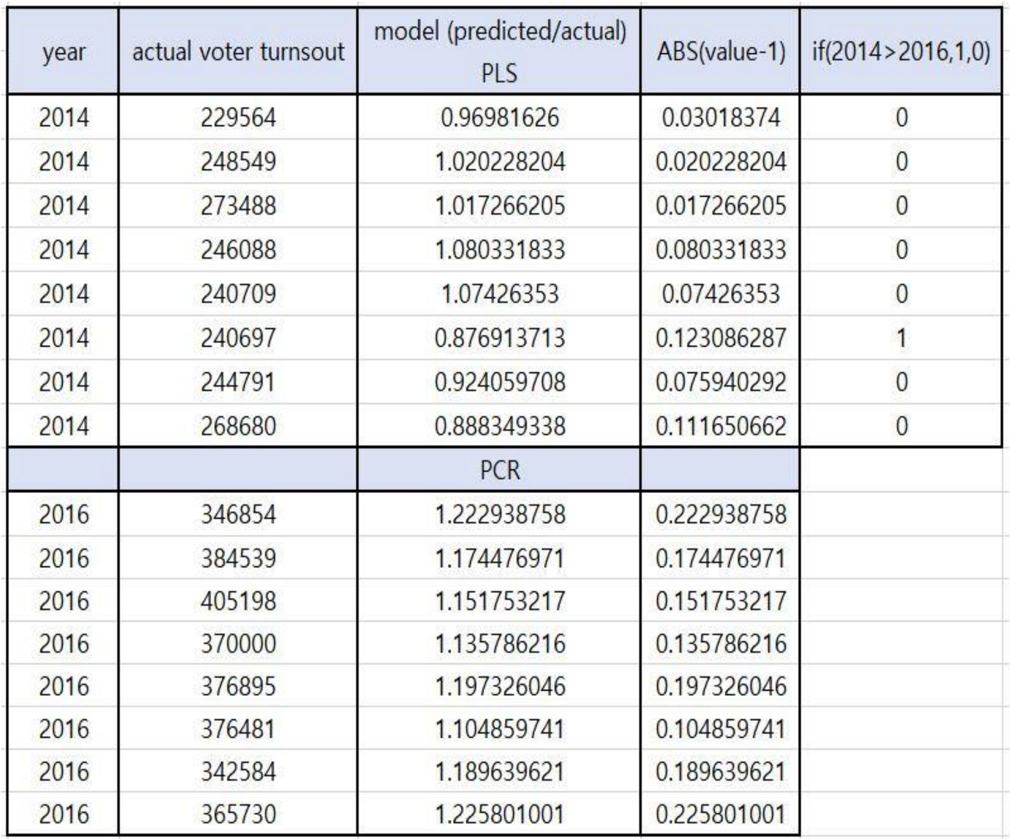
Actual Voter Turnouts

Variables Included

- Number of Actual Voter
- Age (18~29, 30~44, 45~54, 54 over)
- Average Temperature
- Competitiveness
- Congressional District
- Educational Level (Less than High School, High School, High School Graduate, Associate, Bachelor, and more)
- Gender (Male, Female)
- Income
- Race (White, Black, Asian, Hispanic, Other)
- Year (2006, 2008, 2010, 2012, 2014, 2016)

K-Fold Cross-Validation

- Specifically, the n observations (y1, x11,..,x1n), … , (yn, xn1,...xnn) are randomly divided in K groups or folds approximately equal size.
- 6 years as row, 8 districts as columns->48 observations
- 48 obs with 10 folds -> each fold: 4~5 observations are validation datasets, rest of them are train datasets
- Random sampling and leave one out CV

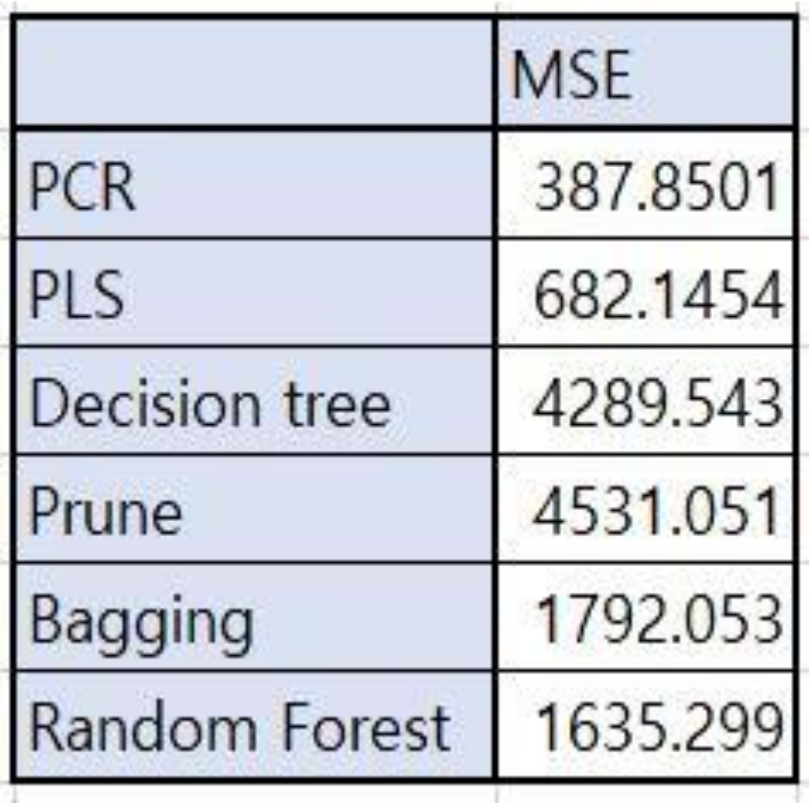


year	actual voter turnsout	model (predicted/actual) PLS	ABS(value-1)	if(2014>2016,1,0)
2014	229564	0.96981626	0.03018374	0
2014	248549	1.020228204	0.020228204	0
2014	273488	1.017266205	0.017266205	0
2014	246088	1.080331833	0.080331833	0
2014	240709	1.07426353	0.07426353	0
2014	240697	0.876913713	0.123086287	1
2014	244791	0.924059708	0.075940292	0
2014	268680	0.888349338	0.111650662	0
		PCR		
2016	346854	1.222938758	0.222938758	
2016	384539	1.174476971	0.174476971	
2016	405198	1.151753217	0.151753217	
2016	370000	1.135786216	0.135786216	
2016	376895	1.197326046	0.197326046	
2016	376481	1.104859741	0.104859741	
2016	342584	1.189639621	0.189639621	
2016	365730	1.225801001	0.225801001	

- Get ratio between predicted and actual voter turnouts.

- If (2014 > 2016)
  True : 1
  False : 0

- Prediction accuracy higher when only midterm election data used



	MSE
PCR	387.8501
PLS	682.1454
Decision tree	4289.543
Prune	4531.051
Bagging	1792.053
Random Forest	1635.299

PCR has <u>the lowest MSE</u>

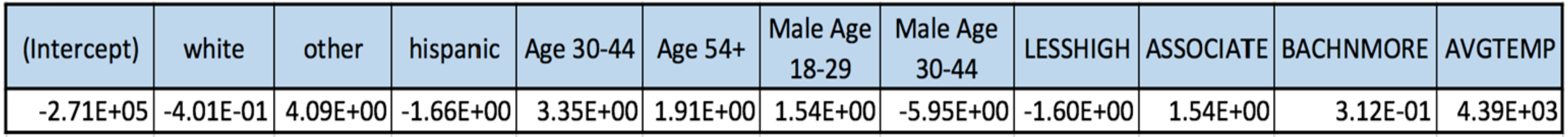
PCR (Principal Component Regression)

- 24 explanatory variables may correlate with each other, so we use it to reduce dimension
- Selected PC components serve as explanatory variables to find relationship with the response
- Choose the number of PC with lower cross validation MSE

2018 Midterm Election
Voter Turnout Prediction

District 1 :	280,245
District 2 :	288,797
District 3 :	330,244
District 4 :	296,466
District 5 :	315,381
District 6 :	291,445
District 7 :	272,583
District 8 :	293,737

Best Subset Selection



(Intercept)	white	other	hispanic	Age 30-44	Age 54+	Male Age 18-29	Male Age 30-44	LESSHIGH	ASSOCIATE	BACHNMORE	AVGTEMP
-2.71E+05	-4.01E-01	4.09E+00	-1.66E+00	3.35E+00	1.91E+00	1.54E+00	-5.95E+00	-1.60E+00	1.54E+00	3.12E-01	4.39E+03

- Selecting the best model with all possible predictors
- Selected 11 variables among 24 variables
- Temperature affects the most among variables

Limitations

- Data Availability e.g. campaign Spending
- Different Total Number of Each Variable
- Effects of Current President Favorability
- Redistriction of Minnesota Cong Districts in 2013
- Weather Data