Understanding results of the Chilean presidential elections of 2017

Cristobal Montt-Volosky
Joaquin Perez-Lapillo
City, University of London

1. INTRODUCTION

1.1. DOMAIN OVERVIEW AND MOTIVATION

As in other countries – including the United States and United Kingdom, last presidential elections of 2017 in Chile surprised most analysts and political experts.

Results of the first round showed a surge of the left, thanks to the rising of a new coalition called Frente Amplio founded the same year. Along with them, centre left parties accounted for 55 percent of the first-round votes. On the other hand, right coalition candidates were led by Sebastian Pinera, a former president of Chile. The surprise came a month later when Pinera obtained almost 55 percent of votes on the second round against Alejandro Guillier –the centre left coalition candidate who came second on the first round— becoming president for the second time.

These interesting results motivated experts to come with different hypothesis. Some of them suggested that young voters did influenced the first-round results but not the second, and others proposed a change of preferences from left to right of middle-class adults, following the behaviour seen in the US and UK elections and referendum (El Mercurio, 2017).

To our knowledge, despite the interest cause by these results and some development of interactive tools to explore the outcome (Unholster, 2017), there has not been any research that has tried to understand this election at the level of geographical detail that we present here.

In this project, our aim will be to use a data-driven visual analytics approach to answer how population demographics could help explain the volatile results in the past presidential elections in Chile.

1.2. DATA

Several data sources were used to carry out this study. First, we obtained the presidential elections results from the National Electoral Service (SERVEL)[1]. This data has information on the number of votes for each candidate at each of the voting locations. Also, before each election, SERVEL releases the registered voters' dataset. Since the implementation of the automatic registration and voluntary vote in 2012, this data includes every adult in Chile eligible to vote. The data includes the residential address and the location at which they vote, which allowed us to merge both datasets base on the voting location. Previously, voter's addresses using Google Maps Geocoding API were geocoded using Google Maps Geocoding API. In order to enrich this data with demographic characteristics, we use the 2012 Census dataset available at the National Statistics Office (INE) webpage[2]. Census data was aggregated several times. First, we aggregated the data at the household level and then at the block

[1] https://www.servel.cl/resultados-definitivos-elecciones-presidencial-parlamentaria-cores-2017/
[2] http://www.ine.cl/estadisticas/censos/censos-de-poblacion-y-vivienda

level, obtaining the proportion of each demographic variable per block. Just for Santiago, data for more than 40,000 blocks was available. For simplification, we decided to create a hexagonal grid in which each hexagon´s area was of 0.9 km^2 resulting in 832 identical hexagons. Census block data was then aggregated for each hexagon and subsequently each voter was assigned to their corresponding hexagon based on their geographical coordinates[3]. The last step was to aggregate the results of the election for each hexagon.

We will restring our study to Santiago. Two are the main reasons. First, the city accounts for almost half of the registered voters, with a fair variation between voter's demographics and vote distributions. Second, the irregular shape of the whole country would make it very difficult to visualize the results in only one view. Furthermore, focusing in only one city, we can explore the voting behaviour between different parts in detail, gaining insights on which local demographics drove the election.

1.3. RESEARCH QUESTION

We would like to answer the following questions:

1. What are the most relevant demographic variables explaining the right votes of the first-round election?
2. Did participation play a role in explaining outcomes?
3. Which variables explain better the "swing to right" behaviour between rounds?
4. How does votes and demographic variables relate in terms of geographic areas?

2. TASKS AND APPROACH

2.1. ANALYTICAL TASKS

2.1.1. UNDERSTAND GEOGRAPHICAL PATTERNS OF VOTES OF THE FIRST ROUND

We will use visualization to identify where the population were most likely to vote right on the first-round elections. This will be conducted by creating a hexagon map of Santiago that would help us compare percentage of votes of equally sized regions, as opposed to a normal choropleth map. In addition, we will plot on the same hexagon map different demographics variables such as education levels and age groups as well as the participation rates as an exploratory way of identifying correlation between those variables and votes. This analytical task will help us answer research questions 1, 2 and 4.

2.1.2. IDENTIFY RELEVANCE OF DEMOGRAPHICS AND PARTICIPATION IN EXPLAINING RIGHT VOTES

The dataset was previously filtered (from more than 50 attributes) to obtain 7 demographic features from which we will identify the most relevant against right-vote intensity. This will be done with a combination of visual techniques, i.e. 2D scatterplots and linear regressions and the computation of correlation coefficients that will be also visualized in a correlation matrix. For more specification, we will include in the 2D scatterplots a colour differentiation by 5 macrozones of Santiago (north, south, west, east and centre). A correlation matrix between features will also be included to identify relationships between them. Finally, we will build a linear model to find

[3] We could have assigned each voter to their corresponding block, but this was not easy as the shapefile we used had spaces between each block polygon and, about 50% of the time, the geocoded voters fell outside any block. The faster solution we thought of, was to convert the spatial polygons of census blocks to spatial points, aggregate the data for each hexagon, and then assign each voter to their corresponding hexagon.

how well the group of features explain right votes, along with analysing significance of coefficients, multicollinearity, and how residuals distribute on the map.

This task is key to allow a quantitative measure of the relationship between demographic variables and votes, helping understand research question 1 and 2.

2.1.3. IDENTIFY THE RELATIONSHIP BETWEEN DEMOGRAPHICS AND PARTICIPATION TO SWING

Like in the previous analytical task, we will identify most relevant demographic features in terms of correlation with swing-to-right by combining visual and computational techniques such as 2D scatterplots and the calculation of correlation coefficients. A correlation matrix will be included, as well as a linear model to test the overall fit of the features explaining swing. This task is key to answer to research question 3.

2.2. APPROACH

2.2.1. LEFT AND RIGHT VOTE

The current political spectrum in Chile is clearly defined by four groups: left, centre-left, centre-right and right. For analytical and simplification purposes, we define "left" and "right" vote as a summation of left and centre-left votes and right and centre-right. This assumption reduces the complexity of the problem, allowing us to do meaningful visualizations and analysis.

For the first-round elections, the candidates were previously classified into the two groups matching their political party association to a left-right scale from previous work in the area ((Bargsted & Maldonado, 2018). The eight candidates were classified as follows:

$$Left\ vote = AG + BS + CG + MEO + EA + AN$$

$$Right\ vote = SP + JAK$$

Where:

AG = Alejandro Guillier; BS = Beatriz Sanchez; CG = Carolina Goic; MEO = Marco Enriquez-Ominami; EA = Eduardo Artes; AN = Alejandro Navarro; SP = Sebastian Pinera; JAK = Jose Antonio Kast.

The second-round election only considered the two most voted candidates of the first-round, Sebastian Pinera (right) and Alejandro Gullier (left). Consequently, on the second-round vote analysis we compare just votes of these two candidates.

2.2.1. SWING

As defined by political experts, swing is "*the statistical measure by which the switch of voters from one party to another on a national or constituency basis can be judged. It is calculated by adding the rise in one party's vote to the fall of the other and dividing by two.*" (Comfort, 1996). Using this definition, the formula of the swing vote between first and second round elections will be obtained by:

$$swing = \frac{(\%PartyA_2 - \%PartyA_1) - (\%PartyB_2 - \%PartyB_1)}{2}$$

Where:

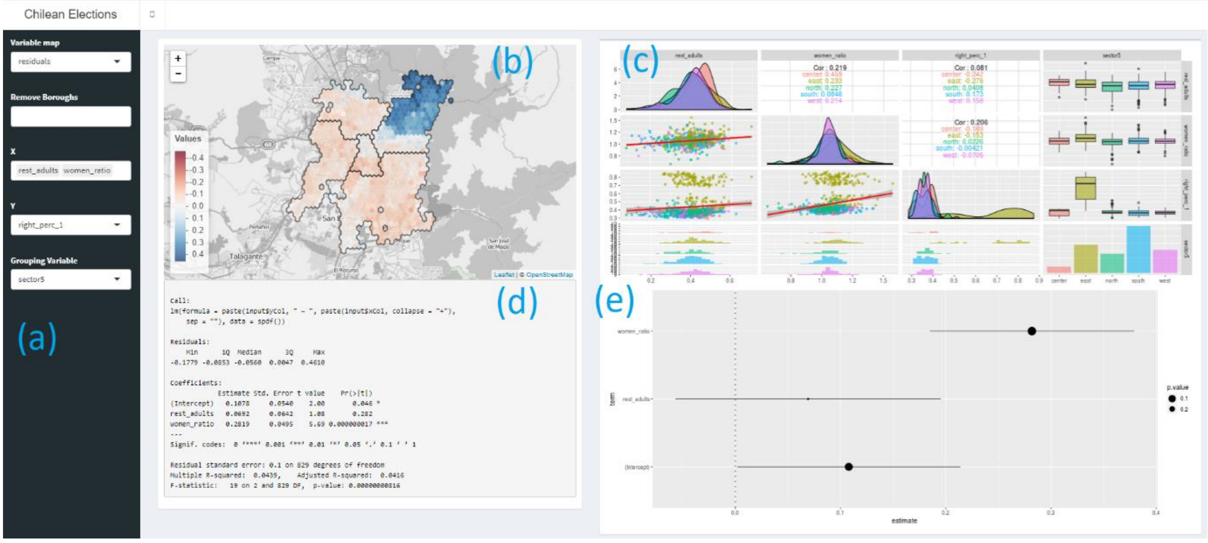
%PartyA (or B)$_{1\ (or\ 2)}$ = percentage of votes obtained by party A/B on the first/second round.

Given that we kept only two groups, the gains of the right will be the exact opposite to the loses of the left. Hence, the swing-to-right calculation can be simplified to:

$$swing = (\%Right_2 - \%Right_1)$$

The analysis of swing-to-right will allow us to identify which groups of the population of Santiago de Chile changed their political preferences between the two elections.

For aiding us to complete the previous tasks, we developed a tool for exploring our data in Shiny. The app can be found here. The elements of the interface are the following. (a) Sidebar to select inputs to be map and model. Variable map is by default set to 'residuals' which maps the residuals of a multiple regression into a hexagonal heatmap (b), other variables can be plotted too. (c) Shows a generalized pairs plot (Emerson, y otros, 2013) with correlation coefficients between the variables selected in input X and Y. (d) Is the output of the linear regression model Y ~ X. Finally (e), depicts the coefficients of the model as a plot, which allows faster inspection of the model results.



Variable map: residuals

Remove Boroughs

X: rest_adults women_ratio

Y: right_perc_1

Grouping Variable: sector5

```
Call:
lm(formula = paste(input$yCol, " ~ ", paste(input$xCol, collapse = "+"),
    sep = ""), data = spdf())

Residuals:
    Min      1Q  Median      3Q     Max
-0.1779 -0.0853 -0.0560  0.0047  0.4610

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)   0.1078     0.0540    2.00       0.046 *
rest_adults   0.0692     0.0642    1.08       0.282
women_ratio   0.2819     0.0495    5.69 0.000000017 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1 on 829 degrees of freedom
Multiple R-squared:  0.0439,    Adjusted R-squared:  0.0416
F-statistic:   19 on 2 and 829 DF,  p-value: 0.0000000816
```

Figure 1: Shiny App for Exploring Chilean Elections Results of 2017

3. ANALYTICAL STEPS

3.1. GEOGRAPHICAL ANALYSIS OF VOTES OF THE FIRST ROUND AND DEMOGRAPHICS

Figure 1 displays the percentage of right votes over the city of Santiago, showing that there are three boroughs that have a clear tendency to vote for right-side candidates: Las Condes, Vitacura and Lo Barnechea. The rest of the boroughs of Santiago tend to behave differently from these three.

Graphs of some demographics and participation are presented in Figure 2 as a way of characterizing boroughs of Santiago. A first observation is acknowledging that the three most right-side boroughs are also the ones that concentrate people with the highest education. A second plot also shows that most company owners and entrepreneurs have their homes in the three boroughs mentioned before.

Participation was calculated as the total number of votes (right and left) over the total number of individuals registered to vote in each voting location, this result was assigned to each voter and then mean participation was calculated for each hexagon. The graph of participation (Figure 3) on the first-round elections over Santiago shows that north-east boroughs are also the most participative ones.

In summary, we see that right-party votes are mostly determined by north-east boroughs (east macrozone) that concentrate people with higher education, owners of companies, and that are more participative than other areas of Santiago. We did not find a geographical relation between age groups, religion, and percentage of immigrants with right-party votes.

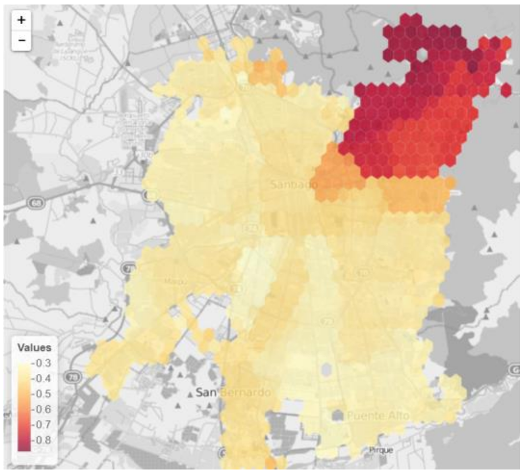


Figure 2: % of right votes on first-round election

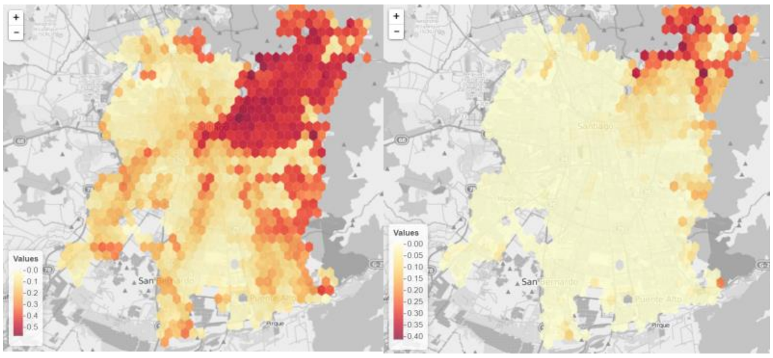


Figure 3: % of Higher Education (left) and % of Employers (right)

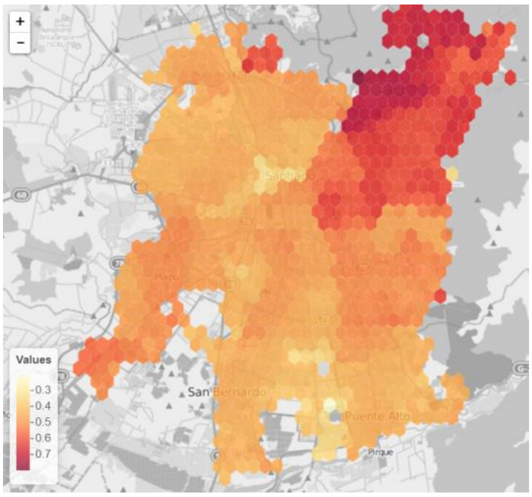


Figure 4: % of Participation on first-round election

3.2. MODELLING DEMOGRAPHICS AND PARTICIPATION AGAINST RIGHT VOTE ON FIRST-ROUND

An initial visual exploration on votes on the first-round showed the presence of two distinct distributions when the entire dataset was analysed. With a simple boxplot of the percentage of right votes over macrozones of Santiago we see that statistics for east boroughs are quite different from the rest (Figure 4). This was due to the presence of 3 east-side boroughs (Las Condes, Vitacura and Lo Barnechea) that are well known for having an extreme tendency to vote for right-party candidates, as discussed on the previous step.

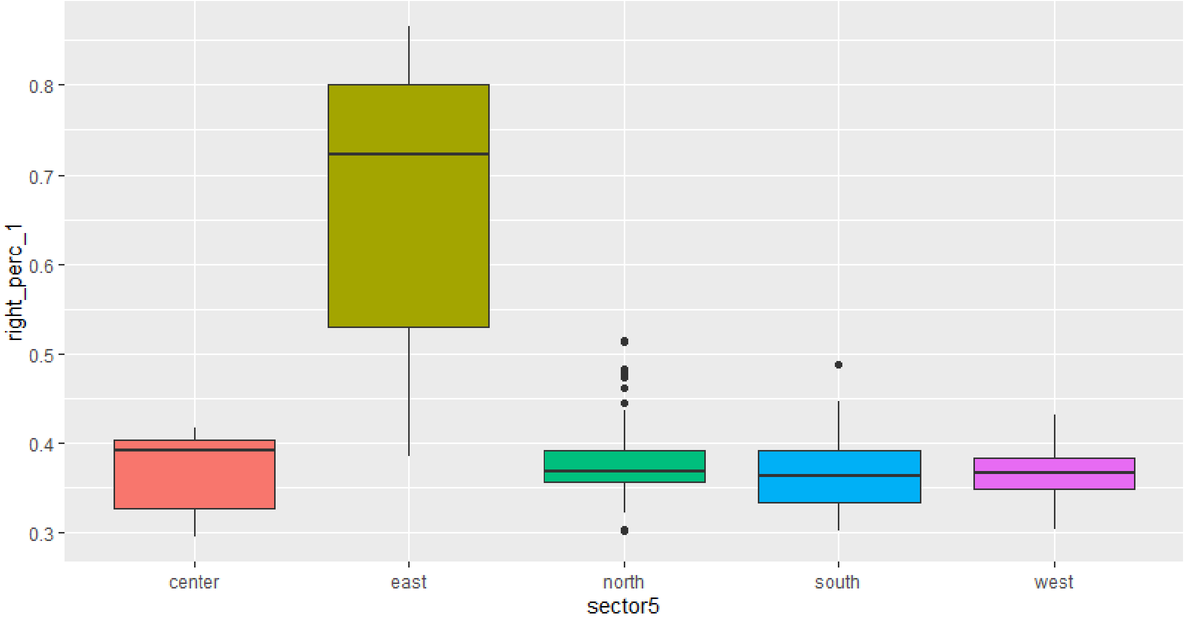


Figure 5: boxplot of % of right votes on the first-round election by macrozone of Santiago

Consequently, the following analysis of demographic features and participation against right-party vote will be performed excluding the 3 east-side boroughs mentioned before, which is a common decision taken by previous studies in the area (Ramirez, 2017).

Scatterplots of demographic features against right vote on the first-round election presented in Figure 4 show interesting insights that are reflected by correlation coefficients.

First, both the percentage of older adults[4] (rest_adults) and the percentage of women as household head over men (women_ratio) show a correlation coefficient of 0.20 meaning that there is only a weak positive relationship between them and right vote. Similarly, the percentage of immigrants also shows a weak positive relationship (0.22), while a better result is found for company owners (Employer) with a coefficient of 0.37.

An interesting result came when the percentage of protestants (religion_evangelicos) and the percentage of people belonging to ethnic groups (indigenous_pop) are plotted against right vote. We see a relatively strong correlation of -0.45 for the first one and -0.39 for the second, meaning that those areas that are highly protestants and where ethnic groups concentrate are less likely to vote right.

On the other hand, a strong positive correlation of 0.58 was found for percentage of individuals with higher education (EdNonTechnicalHighedu/Higher_Educated in model tables).

Percentage of participation also showed a strong positive relationship with a coefficient of 0.62. We now have more evidence to say that right vote is related to areas where people tend to participate more.

[4] Defined as older than 40 years old.

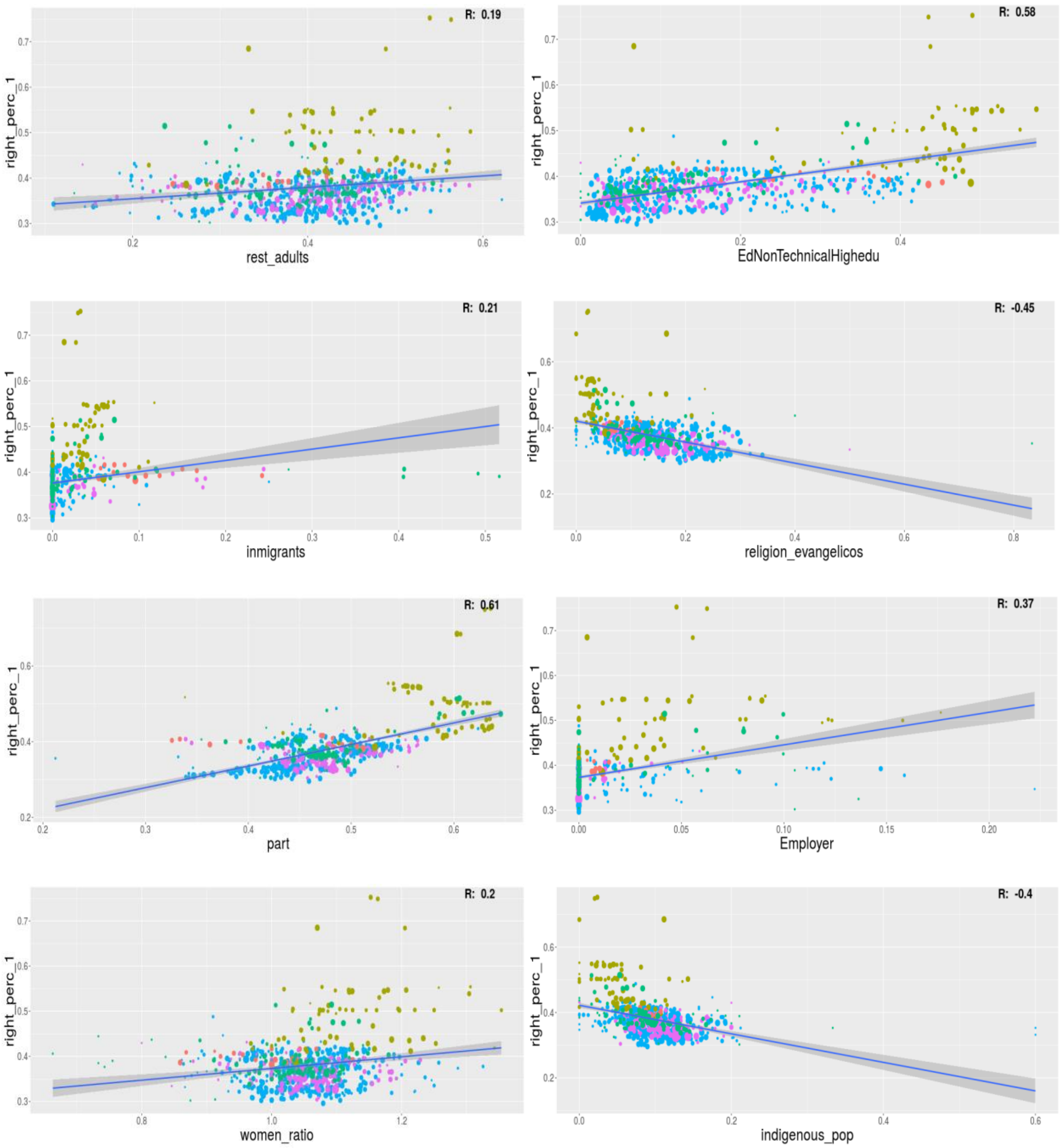


Figure 6: scatterplots of demographics and participation against right vote on first-round.

A correlation matrix between features explaining the right vote of the first-round election shows a strong relationship between some variables. From the matrix we can derive the relationships between participation and the rest of features, as for example its strong positive relation to high education (0.6). Literature on the subject confirms that education is commonly used as a predictor of participation (Blais, 2014; Soto Zazueta, 2014). This is especially interesting now that vote is not mandatory in Chile and participation rates have dropped significantly since the 2012 (PNUD, 2017; Mackenna, 2015). In addition, we see that presence of groups such as protestants and ethnics relate negatively to participation due to its negative relation to education.

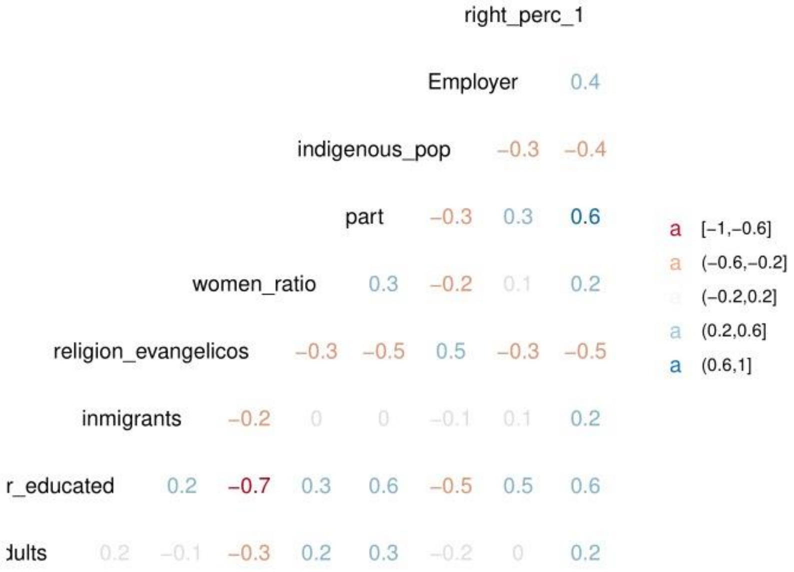


Figure 7: correlation matrix of first-round features

After having explored the relationship between each explanatory variable and right votes, we will build a linear model as an attempt to know how well these features explain the behaviour of right votes and to find which ones are statistically significant. Is it important to notice that we will continue excluding the 3 boroughs that are extreme right voters for consistency with the previous step.

In theory, a multiple linear regression tries to model the relationship between two or more explanatory features and a dependant variable by fitting a linear equation to the observed data[5]. For our problem, the model can be represented as:

$$right\_perc\_1 = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_n * X_n$$

Where:
right_perc_1= dependant variable; β_0= intercept; β_n=coefficients, and X_n= feature "n".

To measure how good the fit is when different features are included, we build 4 models:

1. A null model with just the intercept
2. Including demographics: older_adults, women_ratio, indigenous_pop, immigrants, and religion_evangelicos
3. All above plus income-related features (Higher education and Employer)
4. All above plus participation (part)

Table 1 show results for all 4 models. We see that the model including all variables achieve the highest R^2, explaining 50,8% of the variance of right votes on the first-round election.

In terms of significance of features, we found that 5 out of the 8 variables are significant ($p < 0.01$). The ones not found significant are rest_adults, women_ratio, and religion_evangelicos. It is interesting to note that, in model 2 religion_evangelicos is significant but in model 3, when we control for education it is not anymore. One possible explanation is that there is some confounding variable that explains both, probably socioeconomic

[5] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

status. Indeed, we created a model (not shown) in which we only included religion_evangelicos and its effect is significant, but when we added higher_educated, higher_educated was significant but religion was not anymore. We explored the hexagonal heatmap for each variable, and it was clear that were higher educated people reside, less evangelicos there are. Furthermore, the scatterplot showed that the correlation was quite strong -0.71, nonetheless, we did not see any problem when we searched for multicollinearity before performing the model. Something similar happened with older adults (rest_adults) when we added participation to our model, but in this case, we were not able to come with a plausible explanation from examining the heatmap and correlation matrix. The correlation between both is not high (0.29) and there does not seem to be any spatial relation between them. If we exclude from the model the non-significant features, R^2 slightly reduces to 50.6% and all features remain statistically significant, this would be the most parsimonious and final model (you can reproduce the model in the app).

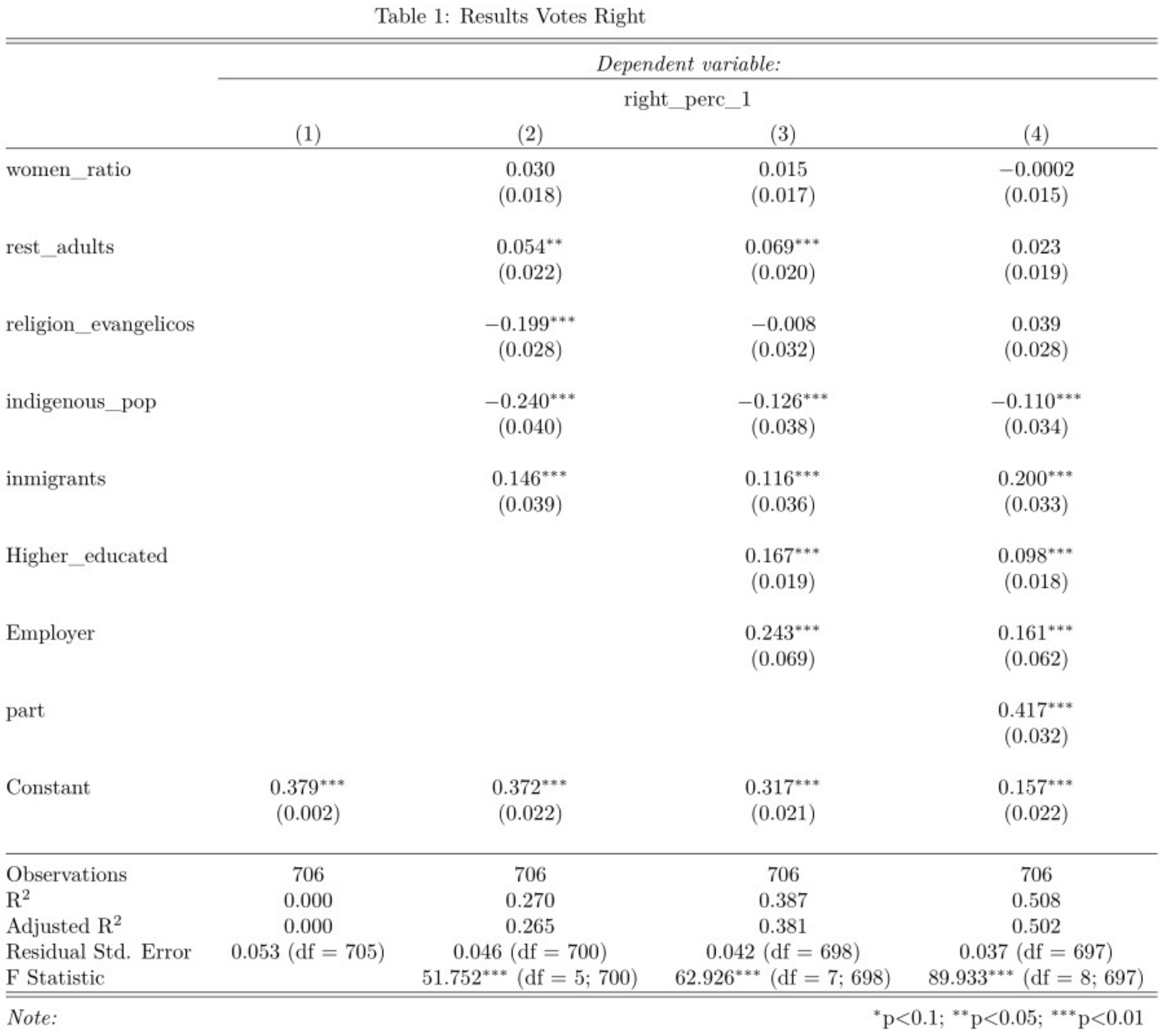


Table 1: Results Votes Right

	Dependent variable:			
	right_perc_1			
	(1)	(2)	(3)	(4)
women_ratio		0.030 (0.018)	0.015 (0.017)	−0.0002 (0.015)
rest_adults		0.054** (0.022)	0.069*** (0.020)	0.023 (0.019)
religion_evangelicos		−0.199*** (0.028)	−0.008 (0.032)	0.039 (0.028)
indigenous_pop		−0.240*** (0.040)	−0.126*** (0.038)	−0.110*** (0.034)
inmigrants		0.146*** (0.039)	0.116*** (0.036)	0.200*** (0.033)
Higher_educated			0.167*** (0.019)	0.098*** (0.018)
Employer			0.243*** (0.069)	0.161*** (0.062)
part				0.417*** (0.032)
Constant	0.379*** (0.002)	0.372*** (0.022)	0.317*** (0.021)	0.157*** (0.022)
Observations	706	706	706	706
R^2	0.000	0.270	0.387	0.508
Adjusted R^2	0.000	0.265	0.381	0.502
Residual Std. Error	0.053 (df = 705)	0.046 (df = 700)	0.042 (df = 698)	0.037 (df = 697)
F Statistic		51.752*** (df = 5; 700)	62.926*** (df = 7; 698)	89.933*** (df = 8; 697)
Note:				*p<0.1; **p<0.05; ***p<0.01

As seen in Figure 7, geographic distribution of residuals improves but not in a dramatic way when groups of features are included. Regarding the spatial distribution of residuals, this are not homogenous. The areas with slightly negative residuals are the ones where the model predicted a higher right vote. The borough of Pedro Aguirre Cerda is one of them (southwest from the city center), historically known for having a closer affiliation with left parties such as the Communist Party. On the other hand, the highest residuals are positive, and their location are in the surroundings of the boroughs we excluded due to their particular higher right vote percentage. In these areas, the model underestimates the percentage of voting right. Suggesting that there must be and spatially varying relationship between our explanatory variables and the percentage of right voting.

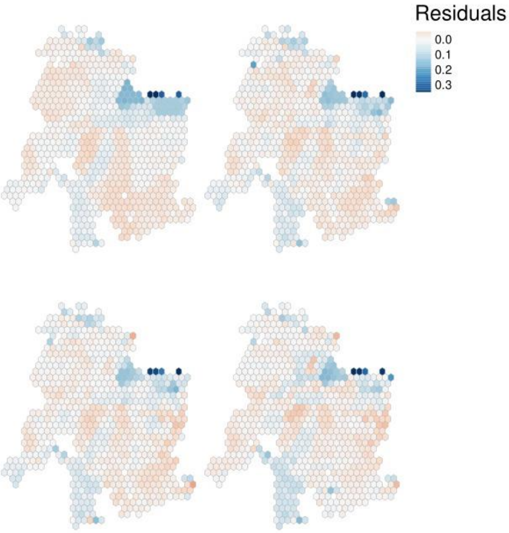


Figure 8: heatmap of residuals of 4 right vote models.

3.2. EXPLAINING SWING-TO-RIGHT BETWEEN ELECTIONS

To explain the swing behaviour that allowed the right to win presidency we will analyse its relationship with the same group of demographic variables and participation but considering all boroughs.[6]

As seen in Figure 8, areas where people changed their preferences in favour to the right between the first and second rounds concentrate on the periphery of Santiago. Of course, this exclude the north-east areas because they already had a high tendency to vote right on the first-round, having less space to grow in the same direction on the second-round.

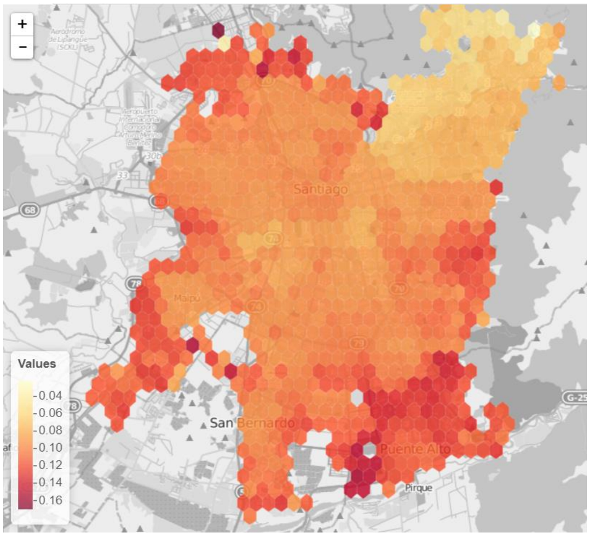


Figure 8: % of swing-to-right on second-round election

Scatterplots and computation of correlation coefficients show also some relevant observations (Figure 9). Negative correlation coefficients are found between women ratio (-0.19), older adults (-0.29), immigrants (-

[6] Unlike the case of right votes, swing did not show different distributions. That is why we explore swing vote considering all boroughs.

0.32), higher education (-0.47), and company owners (-0.56) against swing. This means that those areas with proportionally more male young adults, less educated people, and without many immigrants were more likely to change their vote to the right between first and second rounds. On the other hand, positive correlation coefficients of 0.41 and 0.51 were found for percentage of protestants and ethnic groups against swing.

With respect to participation, plotting the difference between rounds (diff_part)[7] against swing, we see that areas where participation decreased were the most likely to swing to right, as opposed to places were participation increased. This behaviour is reflected in a strong negative correlation coefficient of -0.61.

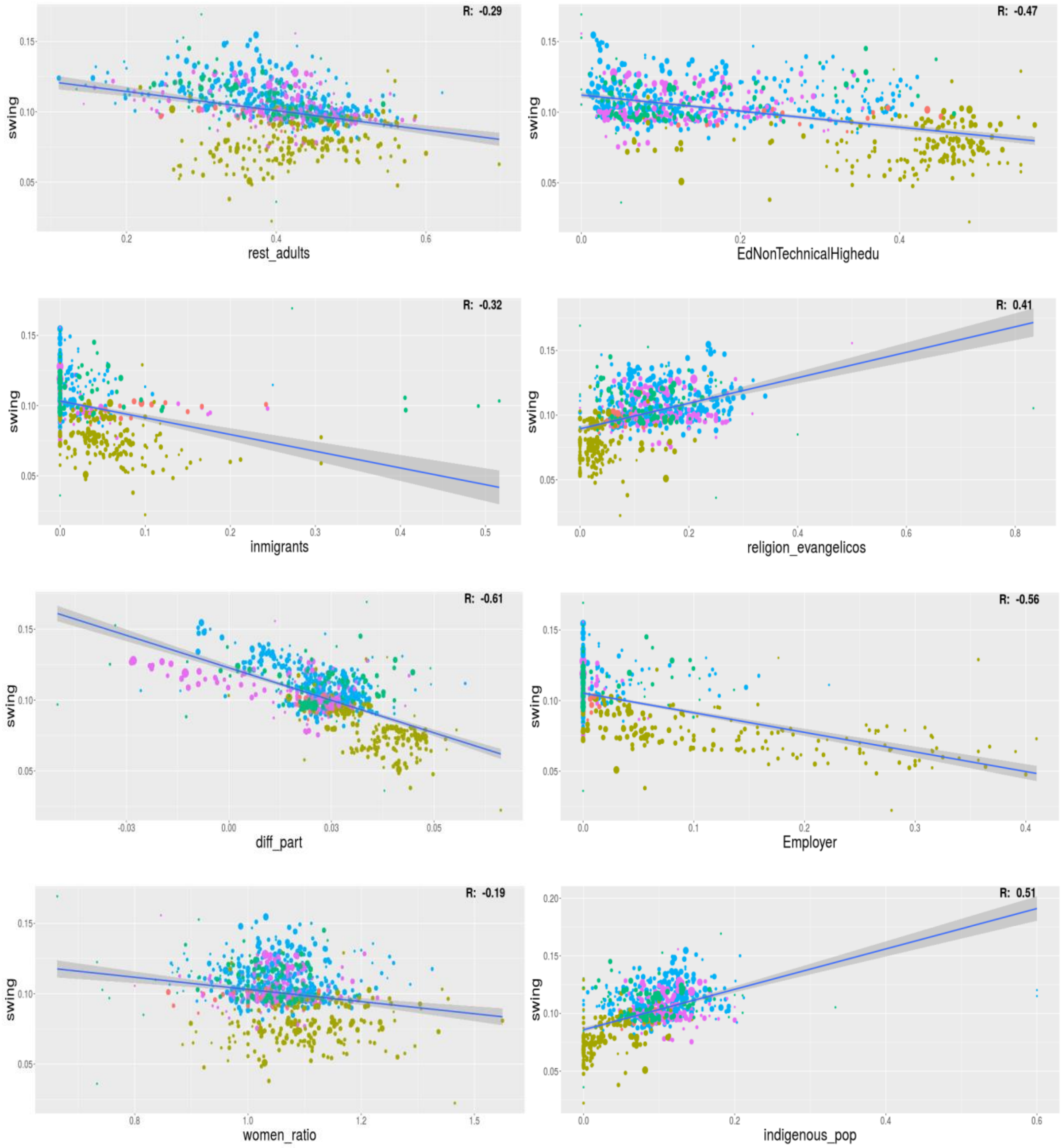


Figure 9: scatterplots of demographics and participation against swing-to-right vote

[7] diff_part= % participation second-round - % participation first-round.

In general, the correlation matrix between features in the swing context keeps the same structure in terms of positive/negative coefficients as seen in explaining the right votes of the first-round elections. Nevertheless, it shows some differences due to the consideration of all boroughs.

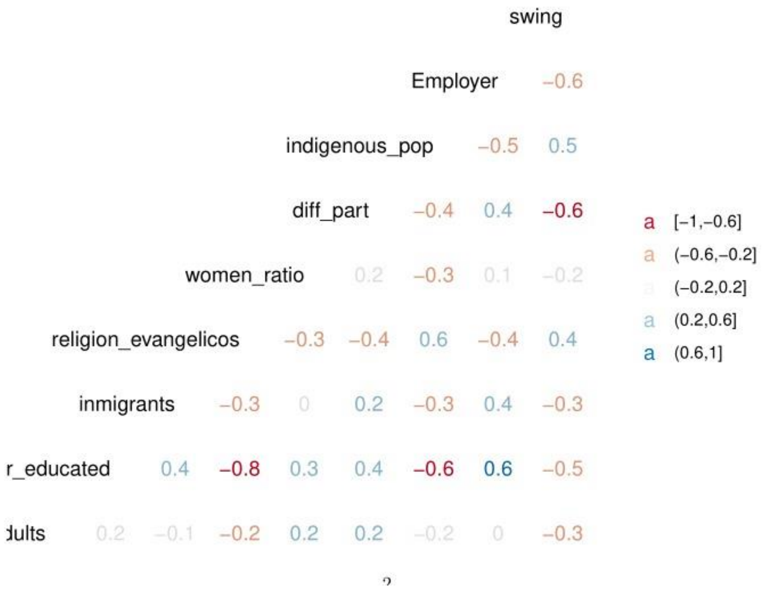


Figure 10: correlation matrix of swing features

Following the steps of the right vote exercise, we will now build a linear model to know how much of the behaviour of swing is being explained by the independent features. Mathematically, the model can be expressed as:

$$swing = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_n * X_n$$

Where:
swing= dependant variable; β_0= intercept; β_n=coefficients, and X_n= feature “n”.

To measure how good the fit is when different features are included, we build 4 models:

5. A null model with just the intercept
6. Including demographics: older_adults, women_ratio, indigenous_pop, immigrants, and religion_evangelicos
7. All above plus income-related features (Higher education and Employer)
8. All above plus difference in participation (diff_part)

Results on Table 2 show that the model including all variables achieve the highest R^2, explaining 54,3% of the variance of swing. In terms of significance of features, we found that all variables are significant ($p < 0.01$) except for higher education, women_ratio and religion_evangelicos.

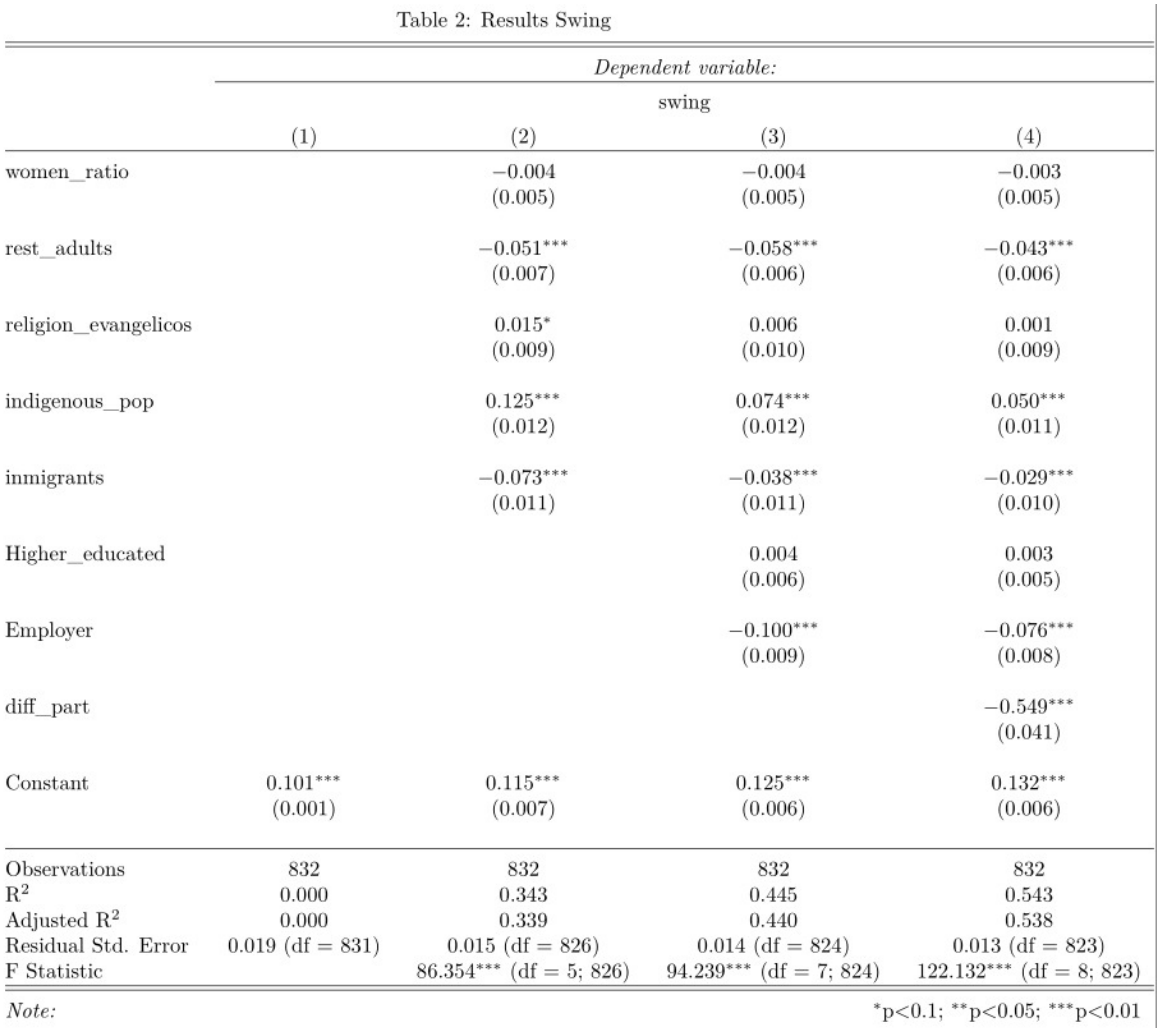


Table 2: Results Swing

	Dependent variable:			
	swing			
	(1)	(2)	(3)	(4)
women_ratio		−0.004	−0.004	−0.003
		(0.005)	(0.005)	(0.005)
rest_adults		−0.051***	−0.058***	−0.043***
		(0.007)	(0.006)	(0.006)
religion_evangelicos		0.015*	0.006	0.001
		(0.009)	(0.010)	(0.009)
indigenous_pop		0.125***	0.074***	0.050***
		(0.012)	(0.012)	(0.011)
inmigrants		−0.073***	−0.038***	−0.029***
		(0.011)	(0.011)	(0.010)
Higher_educated			0.004	0.003
			(0.006)	(0.005)
Employer			−0.100***	−0.076***
			(0.009)	(0.008)
diff_part				−0.549***
				(0.041)
Constant	0.101***	0.115***	0.125***	0.132***
	(0.001)	(0.007)	(0.006)	(0.006)
Observations	832	832	832	832
R^2	0.000	0.343	0.445	0.543
Adjusted R^2	0.000	0.339	0.440	0.538
Residual Std. Error	0.019 (df = 831)	0.015 (df = 826)	0.014 (df = 824)	0.013 (df = 823)
F Statistic		86.354*** (df = 5; 826)	94.239*** (df = 7; 824)	122.132*** (df = 8; 823)
Note:				*p<0.1; **p<0.05; ***p<0.01

As seen in Figure 11, geographic distribution of residuals improves significantly after including demographic features (change between upper-left and upper-right graphs). When income-related features and participation are included (lower-left and lower-right graphs), the distribution of residuals keeps improving but at a decreasing rate. If we compare the map of model residuals to the distribution of swing to right, we can see that, except for some individual hexagons for which the model performs particularly bad, the areas in which there was more swing vote, some places in the south east (Puente Alto) , and areas where there was more swing vote (the north east side of the map) is where the model seems to perform worst, underestimating swing vote in the former and overestimating swing vote in the latter.

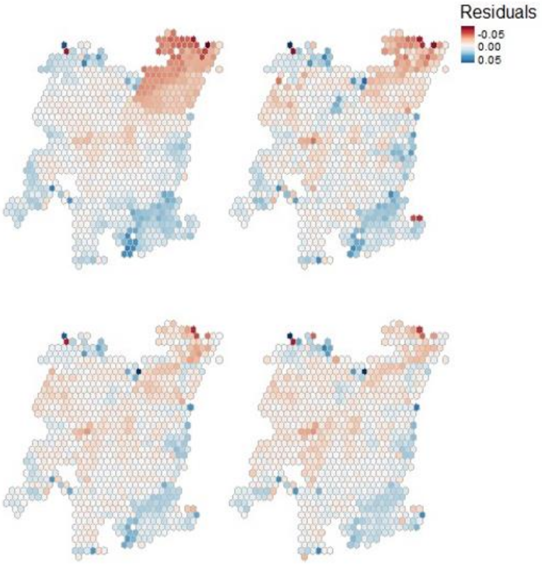


Figure 11: Heatmap of residuals of 4 swing models.

4. FINDINGS

4.1. CHARACTERIZATION OF RIGHT VOTES

Initially, we described how right votes are distributed among Santiago using a heatmap of hexagons. Results showed that east is the macrozone where right votes are more intensive, with a percentage of votes of over 70% in average, meanwhile the rest of the city votes 40%-50% right in average. Hence, we decided to conduct our posterior analysis excluding a group of 3 east-side boroughs.

A visual exploration of demographics showed that in macro zones where people are more likely to vote right coincide with areas where individuals with higher education live. Right votes are also concentrated on the same areas where company owners and entrepreneurs reside.

Percentage of participation is not equally distributed in the city, having a similar geographical representation than right votes. This approach also revealed that participation and education are related, which was later confirmed by correlation coefficients.

Computation of correlation coefficients confirmed the strong positive relationship between higher education and participation to right votes visually found before. A positive correlation was also found for presence of immigrants, older adults, and women to right votes. On the contrary, right votes correlate negatively to presence of ethnic groups and protestants.

Finally, by building a linear model we identify that the group of 8 variables is capable of explain 50,8% of the variance of right votes on the first-round elections, with 5 of the 8 variables being significant at $p < 0.01$. The most relevant variables in explaining right vote were participation (already discussed), indigenous population, immigrants, higher educated and employer. It is interesting to note that, not for every macrozone these variables have the same direction. In specific, for the city center, the direction of some of the variables is the opposite. The relationship between presence of employers and right vote is negative. The same happens with the effect of participation. While in the rest of the city higher participation is related to higher right vote, the contrary occurs in the city center, suggesting that there is something different about its population. Indeed, if we look at the correlation between older adults (rest_adults) and right vote, the coefficient is higher and negative compare to the other zones (-0.504).

4.2. EXPLAINING SWING-TO-RIGHT

Areas where people changed their preferences in favour to the right concentrate on the periphery of Santiago, excluding the north-east macrozone.

In terms of demographics, we found that swing happened in areas where people tend to be less educated, low percentages of older adults, employers and immigrants, and with high presence of ethnic groups and protestants.

Also changes in participation between first and second round elections played an important role in swing-to-right. Places where participation decreased showed higher swing as opposed to areas where it increased.

When building a linear model, we found that the list of 8 variables explained 54,4% of the variance of swing. If we exclude the non-significant variables from the model 54% percent of the variables is explained, therefore this model should be chosen over the more complex one.

Taking into consideration both models, explaining vote for right and swing, participation and difference in participation are the features with a larger effect on the outcome variables, with coefficients of 0.42 and -0.55 respectively. Therefore, they seem to be the most important variables in explaining both outcomes.

5. CRITICAL REFLECTION

5.1. IMPLICATIONS OF THE STUDY

So far, analysis of electoral outcomes and vote characterization in Chile has been done mainly at the borough or regional level, without considering geographical variations at a greater detail. To our knowledge, this is the first time that data from the voting locations, voter register and census data have been used in order to explain voting behaviour. This data allows us to explore voting behaviour variations at an arbitrary level of granularity. In our case, we created equally sized 832 hexagons in order to make computations faster, but any other higher number could have been chosen. Due to the high level of detail enable by the release of voter register data, issues of data privacy should be taken into consideration. Here, we did not include any data that could lead to identification of voters, nonetheless, this data is available in the dataset, that is, their name, residential address and personal ID number, therefore it should be use with caution.

Regarding the implications of the findings, various aspects seem relevant. In Chile, since the implementation of the automatic registration of voters and voluntary vote, there has been an intense debate on whether it was a good decision or not in terms of strengthening Chilean democracy. Our findings are in line with other research (Mackenna, 2015) which shows that participation is highly correlated with level of education. Areas where a higher percentage of higher educated people reside exert their right to vote more often than those who their credentials are lower, which undermines their representation not only in presidential elections but also in parliamentary and municipal elections.

Nevertheless, our approach has relevant limitations. As explained in the data section, we aggregated different data sources in several steps for a specific geographical area, hence we should be careful of drawing conclusions about individual voting behaviour based on our data. Moreover, by this data aggregation local variation is somewhat hidden since we took the average within each hexagon. This could lead to over stating the importance of social demographics over other possible explanations (Beecham, Slingsby, & Brundson, 2018). One possible solution could be to analyse our data at the individual level instead of aggregated to some area, having the voters register data would allow us to do these but visualization would be more challenging. Other limitation could be the use of participation and difference in participation as explanatory variables in our models. It could be argue that both variables as well as the outcome variables we explore (% right vote and % of swing to right) could be explained by the same variable or set of variables, and maybe we should have developed a model for them on their own or some sort of model that could have account for this interdependency (e.g. a Structural Equation Model).

Another limitation more specific to our application, is that, arguably, much of what could explained voting for the right and swing to right are not part of our model. Information on what happened between both rounds is not included, therefore contingencies that may have had a strong influence on presidential results are hidden from our results. For example, between both elections, the candidate from Frente Amplio Beatriz Sánchez, who had a surprisingly high voting percentage regardless the short life of her coalition, did not explicitly call her voters to vote for Alejandro Guillier in the second round (the highest voted left coalition candidate). Puente Alto, the borough where she did surprisingly well, was indeed one of the boroughs that swing right the most (south east corner of Santiago in our hexagonal map).

Despite the limitations mentioned above, we believe that the tool we developed could be use by experts in order, not to only explore how relevant measures -such as participation- could be understood by taking into consideration sociodemographic variables at the micro level, but also to assess the impact of public policies, e.g. reimplementing voluntary vote. Besides academics and public policy makers, the tool could be of interest to political advisors interested in understanding the voting profiles of certain geographical areas for different political party coalitions. To make our approach even more useful, we should extent its capabilities to be able to perform models for specific areas of the city, and in that way, allow to assess changes in the effects of variables between geographical regions.

5.2. REFLECTIONS ON THE VISUAL ANALYTICS APPROACH

Three were the main visual analytics tools we used for answering our questions, a) mapping sociodemographic variables on an hexagonal heatmap map, which allowed us to quickly distinguish their spatial distribution in Santiago, b) the pairwise comparison of explanatory and dependent variables, which quickly showed us the relationships between them and difference of these relations between macrozones and c) the output of different regression models which showed us the most relevant variables for explaining our research question. Plotting the residuals of this model on the hexagonal heatmap also showed us in which geographical areas this global model performed better. We think that these different tools used in conjunction permitted us to answer our research questions.

Compare to other works, in which the use of hexagonal heatmaps is an abstraction of administrative borders (e.g. the transformation of diverse shape local authorities into uniform hexagons (Beecham, Slingsby, & Brundson, 2018) in our work, due to the level of detail provided by our data, this level of abstraction was not necessary, each hexagonal area corresponds to that actual area and its demographic characteristics. We are not only able to see differences between boroughs but also within boroughs. As mentioned in the previous section, the level of detail of our data would have even allow us to visualize each individual voter, though, for practical reasons (faster analysis) we decided to aggregate the data at a higher level. What this type of visualization loses in terms of information, gains it in terms of easiness of interpretability and therefore we think it was the right approach.

Even though characterization of voters and explaining changes in their behaviour using census variables is a difficult task, we consider that the visual analytics approach, together with statistics modelling did achieve the goal of answering the research questions.

5.3. POSSIBLE APPLICABILITY TO OTHER DOMAINS

The approach followed in this study can be generalizable to any other domain in which spatial multivariate data is used. Other domains that share this possible application are many. It could be use on criminology, in order to understand how sociodemographic and characteristics of each region influence certain types of crime. The analysis of social data such as twitter could also be explored, etc. Hence, if there is data enriched with geographical data, the same approach we presented here could be applied. This approach could be summarize as follows: wherever there are measured attributes for which we know their spatial location, we can aggregate those attributes by some spatial subdivision and understand how those attributes vary along latitude and longitude coordinates. Furthermore, if time was included, as the evolution of certain trends over time, time varying spatio-temporal analysis could be performed. In the case of political elections, this could be achieved by linking the results for different elections over time and it would be a natural extension of this application.

REFERENCES

Altman, D. (2004). Redibujando el mapa electoral chileno: incidencia de factores socioeconómicos y género en las urnas. *Revista de ciencia política*.

Bargsted, M., & Maldonado, L. (2018). Party Identification in an Encapsulated Party: The Case of Postauthoritarian Chile. *Journal of Politics in Latin America*, 29-68.

Blais, A. G. (2014). Where does turnout decline come from? *European Jornal of Political Research*, 221-236.

Comfort, N. A. (1996). *Brewer's Politics: A Phrase and Fable Dictionary.* Weidenfeld Nicolson Illustrated.

El Mercurio. (18 de December de 2017). *Comentaristas opinan sobre las razones del triunfo de Piñera*. Obtenido de emol.com: https://www.emol.com/noticias/Nacional/2017/12/18/887836/Sebastian-Pinera-nuevo-Presidente-electo-Comentaristas-Emol-explican-las-razones-de-su-triunfo.html

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hoffman, H., & Wickham, H. (2013). The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics*, 79-91.

Mackenna, B. (2015). Composición del Electorado en Elecciones con Voto Obligatorio y Voluntario: Un Estudio Cuasi-Experimental de la Participación Electoral en Chile. *Revista Latinoamericana de Opinión Pública.*, 49-98.

PNUD. (2017). *Diagnostico Sobre La Participacion Electoral en Chile.* Santiago de Chile: Programa De Las Naciones Unidas Para el Desarrollo.

Ramirez, J. (2017). *Una mirada profunda a las elecciones primarias 2017.* Santiago de Chile: Serie Informe Sociedad y Politica - Instituto Libertad y Desarrollo.

Soto Zazueta, I. M. (2014). Determinantes de la participación electoral en. *Estudios Sociologicos, XXXII (95)*, 323-353.

Unholster. (2017). *DecideChile*. Obtenido de https://www.decidechile.cl