Understanding US counties: Clustering and Classification over voting, socio-economic and public health data

CHRISTINE E. DOIG-CARDET, Universitat Politècnica de Catalunya
DIEGO GARCIA-OLANO, Universitat Politècnica de Catalunya

This paper presents an analysis on the 2012 US presidential election results by county based on their socio-economic, demographics, education and public health data. The goal of the analysis is to explore similar counties based on the gathered dataset and modeling two classifiers: a first one that predicts whether a county voted democrat or republican and a second one that predicts blue islands (democrat counties surrounded by republican ones).

The problem presents highly unbalanced classes. We have implemented our modeling algorithms with and without balancing the training sets to assess how well they perform in both scenarios.

The paper is organized as following: Section 1 introduces the problem and the dataset; Section 2 explains the pre-processing performed; Section 3 presents the feature selection undertaken; Section 4 describes the exploratory analysis: Section 5 summarizes the results for both classifiers; and Section 6 presents the conclusions of the work.

1. INTRODUCTION

This analysis continues prior work done for a web visualization project [1]. The data set is comprised of 2012 US presidential election data results by county [2] combined with socio-economic data available from Measure Of America, a project of the Social Science Research Council [3], and with additional public health and socio-economic data from County Health Rankings provided by the University of Wisconsin Population Health Institute [4].

This project comprises three different analyses: a clustering task and two different classification problems. Prior to the analysis, we perform a basic inspection of the dataset and pre-process the data to treat missing values, outliers and detected errors. To explore the different counties, we perform a PCA and cluster them based on their attributes values to observe significant similarities and differences. The first classification problem predicts whether a county voted republican or democrat, while the second one, classifies a county as a blue-island or not. A blue-island is a county that voted democrat, which is surrounded by republican counties.

1.1 Initial dataset description

The election data set contained columns such as Party1, Candidate1, Party2, Candidate2, etc whose values could change per county, so we pulled out and stored them in a standardized way how many votes Democrats and Republicans got per county. This analysis doesn't take into consideration third party candidates which on a national level makes little difference, but can influence particularly smaller county results. Additionally, Alaska for voting purposes doesn't have the same concept of counties as the rest of the US, and thus the election data for it was summarized only at the state level, and hence not used for this analysis of counties.

All three data sets are linked by their Federal Information Processing Standards (FIPS) code.

The election results for the states in the northeast of the US (CT, MA, ME, NH, RI, and VT) are reported slightly different than other states because the FIPS codes within the state don't map one to one to counties as they do in other states. Thus in our analysis, we need to combine and sum up the voting results for counties with the same FIPS code. The necessity for this step actually went unnoticed during the development of the initial web visualization project [4] associated with an early

version of this data, and in effect over counted the number of democrat counties in the final tally.

The socioeconomic data from the Measure of America report was straightforward to merge using the FIPS code, as was the County Health data, but whereas the Measure data was for the most part "complete", having few missing values, the County Health data included some columns which were unfortunately too sparse and needed to be removed completely (particularly, AIDS and Homicide rates).

We then added a column, which specified if a county was a "blue island", a democrat-voting county surrounded on all sides by republican voting counties. This process was done manually using the results provided by the web visualization which used GIS tools to detect neighboring counties.

At this point, we had our dataset [5] consisting of 3113 counties (rows) and 95 variables that we analyzed further using the R programming language. There are 2427 republican counties, 686 democrat and 84 blue islands.

2. PRE-PROCESSING

The first thing we noticed is that many of the socio-economic variables taken from County Health Ratings included a percentage value along with an intrastate quartile value. We removed these intrastate quartile values due to their redundant nature. Additionally for analysis purposes, we removed some information that was simply identification information (state name, county name, fipscode) and gave rows identifiers based on their FIPS code.

We additionally noticed that we had 3 variables (“physicianratio", "mphratio", and "dentistratio") which were ratios of the form "1909:1", "23123:0", and referred to how many physicians/mental health providers/dentists were available per person in a given county. For these we changed these to be a continuous value (of the form 1/1909, and 0/23123 for the two instances above). We also noticed that the populations of some of the counties were excessively high, thus spotting an incorrect step in the prior combining of FIPS codes for northeastern states, and fixed those instances using a prior dataset (see code for specifics). We store these changes in [6], which consists now in 3113 counties and 63 variables, a full description of the variables can be found in Appendix A.

2.1 Treatment of missing values

For a small subset of values which had been given negative values erroneously during the prior work (smokers, motordeath, violentcrime, fastfoodresp and sickdays) and which correspond with NA's in the original raw source files, we set these to be missing, using the 'mice' package in R to imputate them.

2.2 Basic inspection of the dataset

After this, we separate variables by categories for visualization purposes. We define four categories: education, socio-economic, health and demographic. For each of the categories, we perform a pairs plot and boxplot by party. In this subsection we will just summarize the main outcomes from our exploratory analysis by winning party to make it easier to the reader to follow. The complete resulting plots can be found in Appendix B.

In the education category we found that "hs" and "lesshs" are redundant variables which have to sum to one (Fig.1). We decide to eliminate "lesshs". A visual inspection of plots by election winner shows that "graduate" and "leastbach" variables means are higher and there is more variance in the "education_index" for democrat counties (Fig.2).

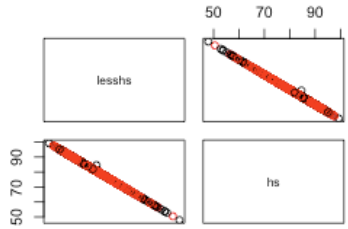

Fig. 1. “lesshs” and “hs” pair plot.

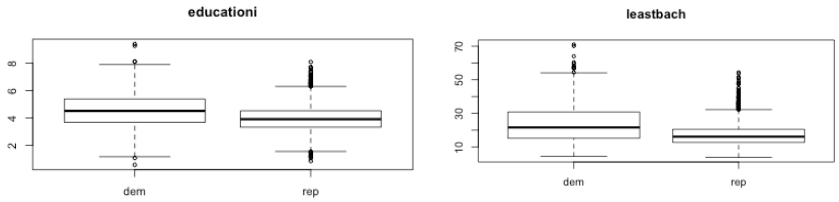

Fig. 2. “educationi” and “leastbach” boxplot by winning party.

In the social-economic category we found that "earnings" and "income_index" are redundant variables. We decide to eliminate "income_index", since earnings is easier to interpret. A visual inspection of plots by election winner shows that counties where democrats won have an upper tail for "violentcrimes", and have more variance in “earnings”, "poverty" and "freelunch", while republican counties have higher values in "farm_workers".

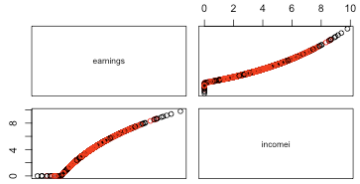

Fig. 3. “earnings” and “incomei” pair plot.

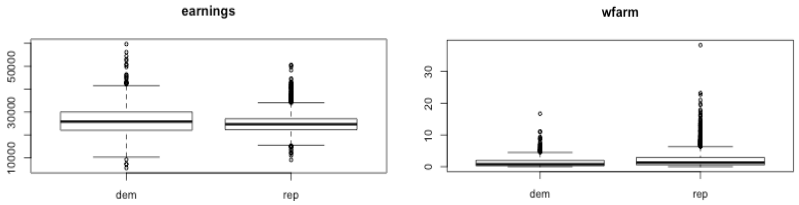

Fig. 4. “earnings” and “wfarm” boxplot by winning party.

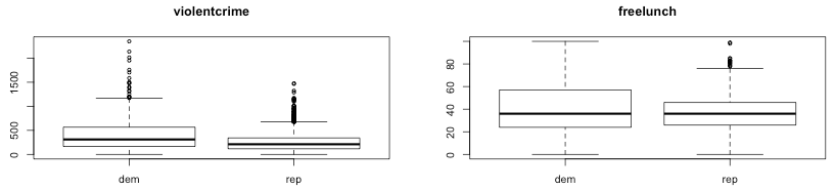

Fig. 5. “violentcrime” and “freelunch” boxplot by winning party.

Further exploring the “violentcrime” variable, we noticed that it was highly skewed and didn’t present a normal distribution, so we performed a Box-cox transformation on it as shown in Fig. 6.

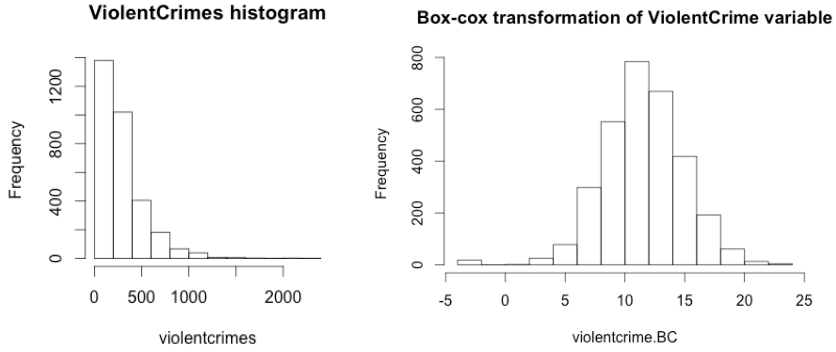

Fig. 6. Box-Cox transformation of “violentcrime” variable

In the health category we found that the variable "mental_health_providers" has six counties that are over five standard deviations away from the mean, the six counties are FIPS: 36031 (Essex County, NY), 36041 (Hamilton County, NY), 36073 (Orleans County, NY), 36095 (Schoharie County, NY), 36115 (Washington County, NY) and 36119 (Westchester, NY). Interestingly, all of these counties are in the state of New York. We confirmed the values from the published data in Measure Of America to make sure there were not induced errors from our side. In the case of 36073 (Orleans, NY) there is a mental health provider for every 38 citizens, a huge number, 42 times standard deviation distance from the mean. After confirming that these are real values, we decided not to remove the outliers.

Performing a histogram for the variable "stdsper100", we noticed an outlier, having a value of 2394 STDsper100. This instance corresponds to FIPS 46041, Dewey, South Dakota, an indian reservation, which also happens to have a surprising number in teen's birth rate of 99.

We plotted the "teenbirthrate" variable and observed values over 100, which seemed odd (Fig. 7). We confirmed that the value is possible since the variable is computed as the number of teen births divided by the number of females ages 15-19 and multiplied by 1,000. So the variable is actually over 1,000, making values over 100 possible.

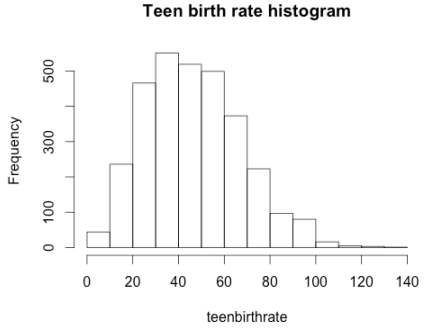

Fig. 7. Teen birth rate histogram

From the health variables boxplots by party, we could visually observe that democrat counties tend to have higher values in "healthy_foods", "dentist_ratio", "STDsper100" and higher variance in "adult_obese", while republican counties have higher values in "health_cost", "fairpoorheath", "mentaldays", "ambulatorycare" and "diabetic" (Fig.8).

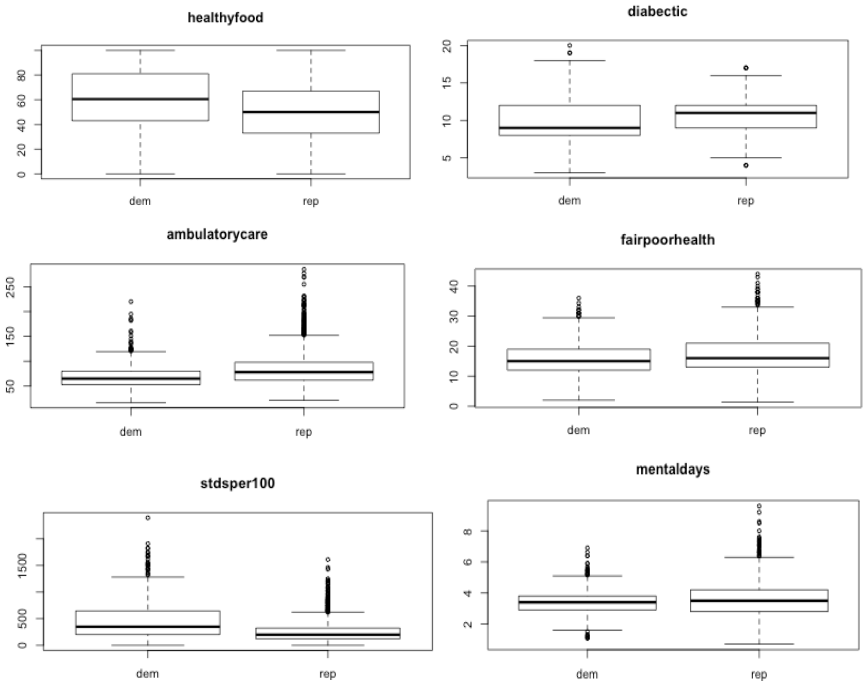

Fig. 8. Health variables Boxplots by winning party

Finally, in the demographics category, we can observe that democrats have higher values in "nonenglish", "african-american" and "single_parent" and more variance in "under_18", while republicans are generally, higher in "over65", "rural" and "white" (Fig.9).

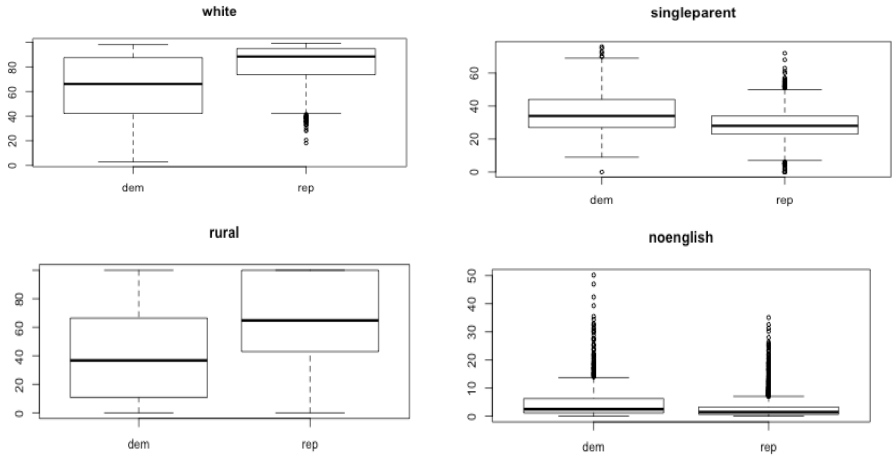

Fig. 9. Demographic variables by winning party.

3. FEATURE SELECTION

Even taking out some of the redundant variables, like “incomei” and “less_hs”, the dataset still has 58 descriptive variables. We will run our models with the entire dataset, but we also want to compare those results with a simplified model. Therefore, we decide to try out different feature selection procedures.

3.1 F-Fisher

We order the variables by F-Fisher correlation and take the top 10 most correlated for each of the two classification problems.

For the Republican/Democrat classifier, these are the top10 highly variables: white, wcontruct, rural, stdsper100, singleparent, graduate, preschl, leastbach, over65, ambulatorycare.

For the blue islands classifiers, these are the top10 highly correlated variables: gini, stdsper100, wtransport, white, excsdrinking, graduate, physicianratio, singleparent, leastbach, wconstruct.

3.2 CFS Filter

We also try getting a subset using the correlation filter CFS from the FSelector R package.

For the Republican/Democrat classifier we get the following model with 23 variables:

winner ~ leastbach + graduate + preschl + illiteracy + belowpov + wconstruct + unemployed + violentcrime + adultobese + excsdrinking + motordeath + stdsper100 + physicianratio + noemotionalsupport + mphratio + white + afric + other + popul + lowbirthrate + singleparent + female + rural

For the blueislands classifier we get the following model with just 6 variables:

blueisland ~ gini + excsdrinking + stdsper100 + nativ + asian + other.

3.3 Boruta

We also tried to implement a Filter + Wrapper feature selection method with the Boruta R package, but the obtained results were that all the 58 attributes were confirmed, so no subset could be obtained.

4. EXPLORATORY ANALYSIS : CLUSTERING U.S. COUNTIES

4.1 PCA

For exploratory purposes, we perform a Principal Component Analysis of the entire dataset, setting the county winning party as a qualitative supplementary variable.

Following the last elbow rule, there are 8 significant dimensions, which account for 61% of the variance.

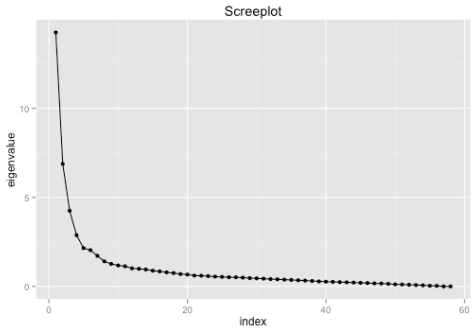

Fig. 9. Screeplot of the eigenvalues

We can plot the first two dimensions that account for 36% of the variance and observe that, in these first two dimensions, the counties are not separable and overlap a lot. However, from the centroids of the modalities (dem/rep), we can see that they lie in different plane quadrants. The centroid for democrats lies in the second quadrant, while the centroid for the republican counties, lies in the fourth quadrant (Fig.10). It would be interesting to interpret what these two axes represent, for that we need the variables factor map (Fig.11).

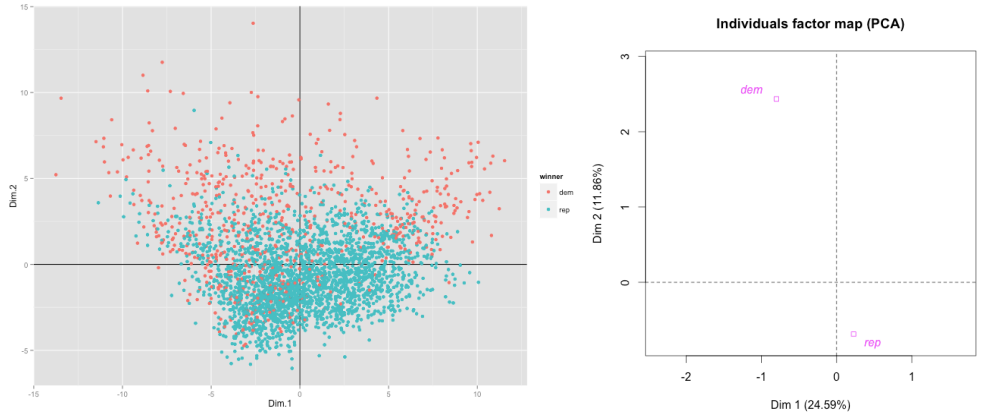

Fig. 10. Individuals factor map by winning party and centroids

From the variables factor map (Fig.11), we observe that the most correlated variables and highly contributing to the first dimension are those ones related with health and education like “fairpoorhealth”, “sickdays”, “diabetic” or “physicallyinactive” on the positive direction and “hs”, “educationi”, “mammography” in the negative direction. The second dimension is more represented by variables dealing with demographics, like “over65”, “white”, “rural”. From the position of the centroids, we can say that republican counties are less educated, have poor health, are rural, with a high number of people over 65 and white. On the other hand, counties where democrats won, have a more diverse and educated population. This can just be interpreted in very general terms, since we have seen in the individuals factor map that the individuals are not separable with just these two dimensions.

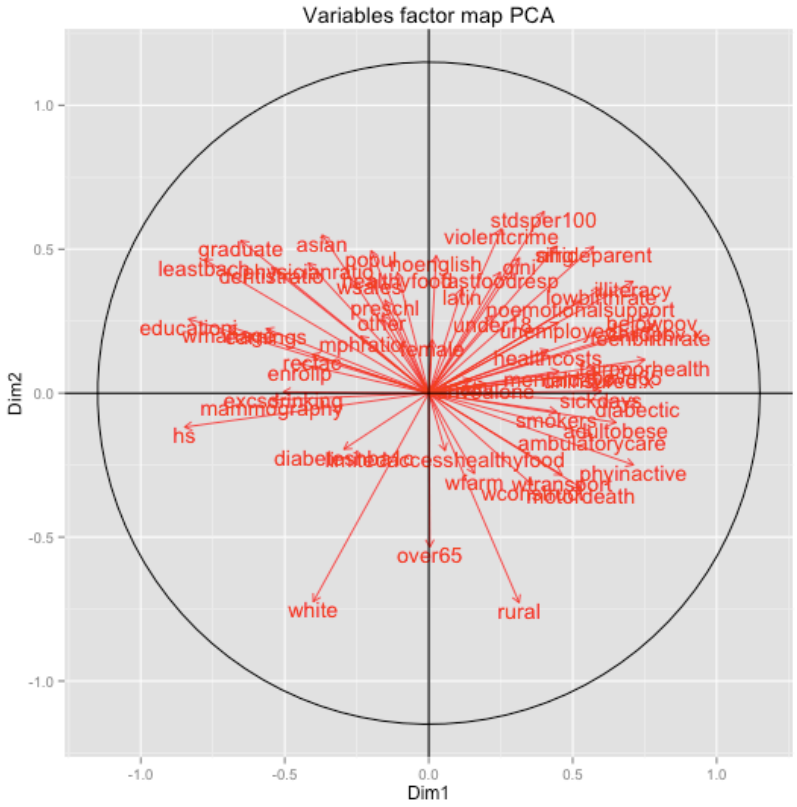

Fig. 11. Variables factor map

4.2 Clustering

Next, we would like to cluster counties by their variables, to assess whether democrat and republican counties, naturally form clusters. This is an unsupervised method; although we do have the label for each county, we are just going to used it to compare the result of the cluster with the real labels that we already got.

We perform two hierarchical clusters: one with two clusters and the second one with three, decided from Fig. 12.

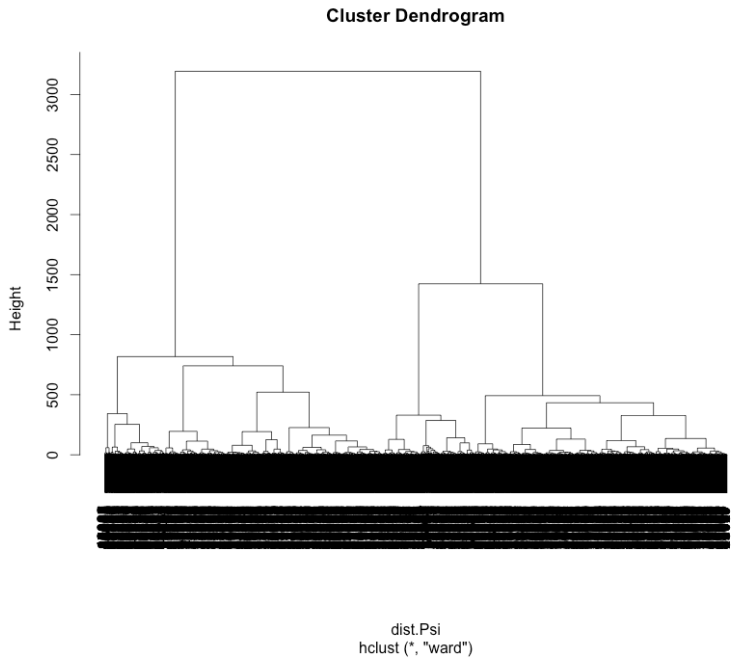

Fig. 12. Hierarchical clustering dendrogram

We just plot the first principal components for visualization purposes, although the clustering is performed in the entire dataset. From the results, we can observe that the clustering doesn't group counties very well by winning party (Fig.13). We also tried three clusters to see if it improved the grouping by winning party. It didn't (Fig.14).

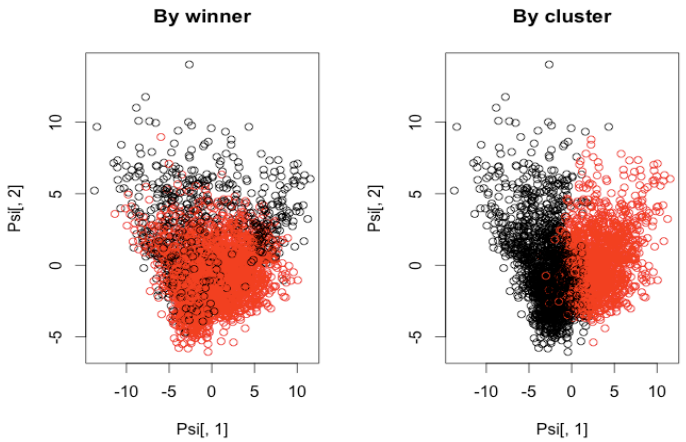

Fig. 13. Clustering result (two clusters) comparison with actual winning party

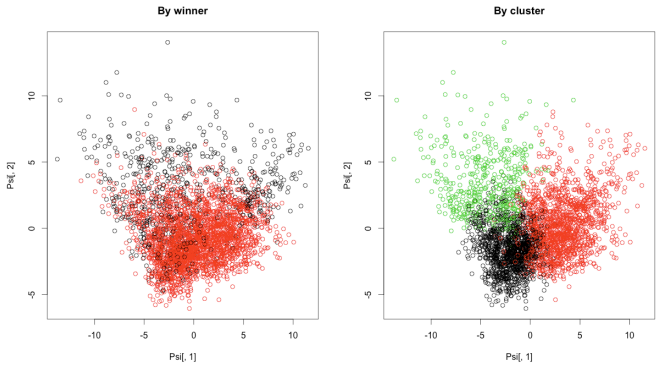

Fig. 14. Clustering result (three clusters) comparison with actual winning party

5. PREDICTIVE ANALYSIS : CLASSIFIERS

We want to build two classifier that are able to predict if a county voted democrat or republican and if a county is a "blue island", respectively, based on their socio-economic, educational, health and demographic variables. In order to do so, we consider the following family of classifiers: Naïve Bayes, Logistic Regression, Kernel SVM and Random forest. Some of these classifiers have parameters that can be tuned or have themselves different models (e.g. linear, quadratic or rbf kernel SVM). We can also perform. The full implementation of all these possible combinations of models can be found in the code file. In this section we'll present the validation protocol used and the results obtained.

As we have mentioned before, for both of our classifiers, we have unbalanced classes. There are 2427 republican counties, 686 democrat and 84 blue islands. So, in the first classifier democrats represent a 22% of the counties and in the second one, blue islands 2,7%. There are several methods to improve a classifiers performance in the case of imbalanced classes, like subsample the majority class, oversample the minority class or creating synthetic examples. We have chosen to use the SMOTE package to test whether balancing classes has an effect on the model accuracy. Therefore, we will train each of our different models with the initial imbalanced training dataset and with the balanced one using the SMOTE package.

As described in Section 3, given the amount of variables, we can do feature selection to simplify the model. For each of the different models and balancing method, we will compare the results from running the models on the entire dataset or by getting the subset from the CFS filter.

5.1 Classifying U.S. Counties in Democrat or Republican

Table 1 summarizes the results obtained for classifying counties into democrat or republican. When looking at these results we should remember that our dataset contains 22% of democrat counties, therefore a classifier that always predicted republican would have a 22% of error.

Model	Balanced training dataset	Subset variables	Validation error (%)	Democrats VA accuracy
Naïve Bayes	Imbalanced	All features	16,70%	62,99%
	SMOTE-Balanced	All features	17,07%	60,93%
	Imbalanced	CFS Subset	16,78%	63,26%
	SMOTE-Balanced	CFS Subset	17,35%	60,92%
Logistic Regression	Imbalanced	All features	12,33%	75,85%
	SMOTE-Balanced	All features	15,78%	61,11%
	Imbalanced	CFS Subset	14,17%	75,38%
	SMOTE-Balanced	CFS Subset	19,96%	54,05%
Random Forest	Imbalanced	All features	**11,40%**	**83,38%**
	SMOTE-Balanced	All features	11,68%	82,86%
	Imbalanced	CFS Subset	12,49%	79,93%
	SMOTE-Balanced	CFS Subset	13,90%	66,61%
Linear Kernel SVM	Imbalanced	All features	11,89%	66,67%
	SMOTE-Balanced	All features	**14,63%**	**84,62%**
	Imbalanced	CFS Subset	14,46%	48,76%
	SMOTE-Balanced	CFS Subset	19,77%	80,83%
Quadratic Kernel SVM	Imbalanced	All features	13,99%	67,15%
	SMOTE-Balanced	All features	18,33%	62,96%
	Imbalanced	CFS Subset	15,11%	61,02%
	SMOTE-Balanced	CFS Subset	18,17%	67,82%
RBF Kernel SVM	Imbalanced	All features	**10,61%**	**58,87%**
	SMOTE-Balanced	All features	**12,38%**	**84,32%**
	Imbalanced	CFS Subset	**11,41%**	**61,87%**
	SMOTE-Balanced	CFS Subset	15,59%	79,73%
RBF Kernel SVM with cost 5	SMOTE-Balanced	All features	13,18%	80,34%
	SMOTE-Balanced	CFS Subset	13,34%	75,97%
Quadratic Kernel SVM with cost 5	SMOTE-Balanced	All features	16,56%	70,25%
Linear SVM with cost 5	SMOTE-Balanced	All features	16,39%	82,24%

Table. 1. Classification results for democrat/republican

To select a model, we will not only focus on the best validation error result, but also on the accuracy on predicting the minority class, and when possible, selecting the simplest model (subset, better than all features).

The best models in terms of validation error are:

- RBF kernel SVM with imbalanced classes and all features - 10,61%.
- Random Forest with imbalanced classes and all features – 11,40%
- RBF Kernel SVM with imbalanced classes and subset – 11,41%

From these top validation error, the only one having an acceptable accuracy on finding democrat counties would be the Random Forest with 83,38%.

Taking a look at the best models in terms of democrats validation accuracy are:

- Linear Kernel SVM with SMOTE balancing and all features – 84,62%
- RBF Kernel with SMOTE balancing and all features – 84,32%
- Random Forest with imbalanced classes and all features – 83,38%

If we value simplicity the most, we could get a Random Forest model using a subset features generated from a CFS filter on imbalanced data 12,49% validation error and 79,93% on democrats accuracy or a SVM using an rbf kernel and cost 1 with a subset of features on balanced data 15,59% and 79,73%.

As mentioned above, our accuracy and overall error rate could be improved by including all variables, with one of the top models, but at the cost of complexity.

To improve the performance/reduce complexity further we could try:
1) couple the random forest with boosting or the svm with bagging,
2) different values of the cost parameter for the svm to tune it and
3) tuning parameters to random forest
4) see if we can find better subsets of variables further using backwards search or another method

For the reasons mentioned above, we select the Random Forest model with a subset of features on imbalanced data. The test error obtained is 11,4% with an accuracy of 61,81% of democrats and 96,66% of republicans.

5.2 Classifying U.S. Counties in Blue islands.

Table 2 summarizes the results obtained for classifying counties blue islands. When looking at these results we should take into account that our dataset contains 2.7% of blue islands, therefore a classifier that always predicted not a blue island will have a 97.3% of accuracy.

Model	**Balanced training dataset**	**Priors**	**Validation error (%)**	**Blue islands VA accuracy**
Random Forest	Imbalanced	No	2,57%	7,38%
	Imbalanced	Yes	2,49%	10,23%
	SMOTE-Balanced	Yes	4,70%	40,71%
	SMOTE-Balanced	No	3,61%	23,80%
SVM RBF with cost 5	Imbalanced	-	2,89%	10,52%
	SMOTE-Balanced	-	**9,00%**	**57,15%**

Table. 2. Classification results for blue islands

The SVM RBF does a better job at detecting the blue islands than Random Forest at the cost of the validation error, even higher than the a classifier that will always predict “no-blue-island”. To try to improve the results, we decide to try tuning the cost parameter for the SVM RBF model. The results are shown in Table 3.

Cost	1	2	3	4	5	6	7	8	9	10
Blue island accuracy	66,2%	66,4%	59,27%	52,1%	52,76%	52,7%	48,8%	47,1%	42,1%	50,3%
VA Errors	12,9%	12,3%	10,7%	10,5%	10,2%	9,7%	10,1%	9,5%	10,4%	9,6%

Table. 3. Results tuning the cost parameter for an SVM Kernel RBF classifier

The shows we should be using cost=8 for the best overall error validation (9,5% error) but that we should use cost=1 for best blue island accuracy 66,2%. We choose cost=5 since we want to balance the blue island accuracy with the overall error.

We then, refit the SVM kernel rbf model and we get a 3,69% error and 29,4% on blue island accuracy, which are terrible results, a higher overall error than the percentage of the minority class and only a finding 29,4% of the blue islands.

6. CONCLUSIONS

In this project, we analyzed U.S. counties based on their 2012 presidential election results, socio-economic, demographics, education and public health data. The goal of the analysis was to explore similar counties based on the gathered dataset and modeling two classifiers: a first one that predicted whether a county voted democrat or republican and a second one that predicts blue islands (democrat counties surrounded by republican ones).

For the democrat/republican classifier our final selected model had an error of 11,4% with an accuracy of 61,81% of democrats and 96,66% of republicans, while the blue islands classifier had an overall error of 3,69% and 29,4% accuracy on the blue islands.

We observed that the SMOTE method for dealing with unbalanced classes had little effect in improving the results, as also noticed in other papers like [7]: “SMOTE does not attenuate the bias towards the classification in the majority class for most classifiers when data are high-dimensional”.

REFERENCES

[1] http://www.diegoolano.com/electionmap/
[2] http://www.theguardian.com/news/datablog/2012/nov/07/us-2012-election-county-results-download#data
[3] http://www.measureofamerica.org/
[4] http://www.countyhealthrankings.org/rankings/data
[5] finaldata-withblueislands.csv
[6] finaldata-classificationset.csv
[7] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3648438/