Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata

Seth Flaxman (Department of Statistics, Oxford),*
Dougal Sutherland (Gatsby Computational Neuroscience Unit, University College London),
Yu-Xiang Wang (Machine Learning Department, Carnegie Mellon University),
and Yee Whye Teh (Department of Statistics, Oxford)

November 10, 2016

1 Introduction

The results of the 2016 US Presidential Election were, to put it mildly, a surprise. Pre-election polls and forecasts based on these polls pointed to a Clinton victory, a prediction shared by betting markets and pundits. In the aftermath of the vote, the main question asked is "why?" with answers ranging from the political to the economic to the social/cultural. These explanations are informed by incomplete information:

- Vote counts will not be finalized in many precincts until days or in rare cases weeks after the election. Very close popular vote totals yield winner-take-all results, a fact of the US's electoral system but one that can lead to winner-take-all explanations.
- The winner-take-all system means that the geography of voting patterns is often overly simplified, with states considered as homogeneous entities, leading to the ubiquitous red and blue maps popularized in 2000 (Gastner et al., 2005). The geography of voting patterns are also often based on overly broad demographic/geographic categories, such as "suburban whites" which do not necessarily reflect an up-to-date picture of who lives where.
- Exit polls (surveys of voters) conducted by Edison Research on behalf of a consortium of media outlets are available immediately, with sufficient sample coverage to yield representative samples in 28 US states and nationally. In their publicly available format they do not provide sub-state geographic

Note that in the months after the election, high quality survey results will become available (the 2016 American National Election Study will be a large-scale look at who voted and why, and the US Census Bureau collects data on who voted), along with voter rolls data.

In the meantime, our goal in this article is provide an answer to the question "what?" so as to inform discussions of "why?" All of the source code for our analysis is freely available online to

*For correspondence: flaxman@stats.ox.ac.uk

enable replication[1].

Our analysis is in two parts: building on Flaxman et al. (2015), we provide exit poll style contingency tables estimating the number of voters and non-voters by candidate and various subgroups. We are able to make much more finely-grained estimates, at the local level, than conventional exit polling. We make these estimates at the national, state, and local level. For example, Table **??** shows that 47% of voters were men while 53% were women. Among women, Clinton received 56% of the vote while Trump received 45% of the vote. Unlike exit polls, we also estimate the participation rate by demographic, calculating the fraction of "potential voters"—voting-age citizens—who voted for one of the two major party candidates. Equivalently, we calculate the fraction of Other / non-voting. Thus we see that 50% of eligible men participated, voting for either Clinton or Trump, while 53% of eligible women participated.

The second part of our analysis is an exploratory analysis of which group-level variables (i.e. distribution of age, joint distribution of income and race) were correlated with the final outcome. While previous approaches to this question have focused on correlations between group-level means[2], our novel approach makes it possible to investigate correlations between the full distributions of these group-level variables, observed at the local area level.

2 Methods

Given n local areas of interest (counties reporting electoral results merged with PUMA regions, following Flaxman et al. (2015)) we assume that we have:

- Samples of individual level demographic data from the American Community Survey:

$$\{\boldsymbol{x}_1^j\}_{j=1}^{N_1}, \dots, \{\boldsymbol{x}_n^j\}_{j=1}^{N_n} \tag{1}$$

- Election results $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$.

Election results for each local area i are summarized as counts vectors $\boldsymbol{y}_i$. We use the method of "distribution regression" (Szabo et al., 2015; Lopez-Paz et al., 2015; Flaxman et al., 2015) to regress from a set of samples $\{\boldsymbol{x}_i^j\}_{j=1}^{N_i}$ to a label $\boldsymbol{y}_i$. This entails encoding N_i samples as a high-dimensional feature vector, using explicit feature vectors and kernel mean embeddings (Smola et al., 2007; Gretton et al., 2012):

We use an additive featurization to compute mean embeddings $\boldsymbol{\mu}_i$ as follows for local area i:

$$\{\boldsymbol{x}_i^j\}_{j=1}^{N_i} = \{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \dots, \boldsymbol{x}_i^{N_i}\} \tag{2}$$

Each $\boldsymbol{x}_i^j$ is a vector of length d consisting of a mix of categorical and real valued variables:

$$\boldsymbol{x}_i^j = [x_{r1}^j, \dots, x_{rd}^j]^\top \tag{3}$$

[1] `www.github.com/dougalsutherland/pummeler` and `www.github.com/flaxter/us2016`

[2] `http://www.nytimes.com/2016/03/13/upshot/the-geography-of-trumpism.html`

We consider feature mappings $\phi_1, \ldots, \phi_d$ and the additive featurization in which feature vectors are concatenated:

$$\boldsymbol{\phi}(x_i^j) := [\phi_1(x_{r1}^j), \ldots, \phi_d(x_{rd}^j)]^\top \tag{4}$$

We also consider including interaction terms of the form: $\phi_{pq}(x_{rp}^j, x_{rq}^j)$ in $\boldsymbol{\phi}(x_i^j)$ in $\boldsymbol{\phi}$.

We use $\boldsymbol{\phi}$ to estimate mean embeddings, one for each region, where the weights w_j are the person weights, one per observation, reported in the census:

$$\boldsymbol{\mu}_i = \sum_{j=1}^{N_i} w_j \boldsymbol{\phi}(\boldsymbol{x}_i^j) \tag{5}$$

We model the outcome $\boldsymbol{y}_i$ as a function of covariate distribution vectors $\boldsymbol{\mu}_i$ using penalized multinomial regression with a softmax link function. Let

$$\boldsymbol{y}_i = [\text{Clinton votes}, \text{Trump votes}, \text{Non-votes and third party votes}]^\top$$

Then:

$$\boldsymbol{y} \sim \text{Multinomial}(\text{softmax}(\boldsymbol{\mu}\beta_1, \boldsymbol{\mu}\beta_2, \boldsymbol{\mu}\beta_3)) \tag{6}$$

where softmax generalizes the logistic link to multiple categories:

$$\text{softmax}_i(x_1, x_2, x_3) = \frac{\exp(x_i)}{\exp(x_1) + \exp(x_2) + \exp(x_3)} \tag{7}$$

We implemented the featurization using orthogonal random Fourier features (Felix et al., 2016) for real-valued variables and unary coding for categorical variables. We fit the penalized multinomial model using `glmnet` in R, crossvalidating to choose the α parameter (relative strength of L_1 vs. L_2 penalty) and sparsity parameter λ. We used `glmnet`'s built-in group lasso functionality meaning that $(\beta_1)_i, (\beta_2)_i, (\beta_3)_i$ would either all be simultaneously zeroed out by the L_1 penalty or not.

2.1 Inferring who voted

After obtaining maximum likelihood estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ from our fit of the model, we used subgroup populations to make predictions, calculating mean embeddings for each local area based only on the group under consideration. For example, to predict the vote among women in region i we calculated new predictor vectors restricting our summation to the N_i^{W} women in area i:

$$\boldsymbol{\mu}_i^{\mathrm{W}} = \sum_{j=1}^{N_i^{\mathrm{W}}} w_j \boldsymbol{\phi}(\boldsymbol{x}_i^j) \tag{8}$$

and then used $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and softmax to calculate the corresponding probabilities of supporting Clinton, Trump, and other among women in region i. To calculate expected vote totals, we multiply these probabilities by the estimated total number of women in each region i, calculated by summing the census weights:

$$\sum_{j}^{N_i^{\mathrm{W}}} w_j \tag{9}$$

2.2 Group-based exploratory data analysis

While exit poll style ecological inference can give us deep insights into various preselected demographic categories, exploratory data analysis can reveal unexpected patterns. We consider fitting the same models as above, but only using related subsets of the features, e.g. all of the categorical variables related to race or the interaction between age and income.

3 Data and implementation

Our analysis can be replicated using a python package we wrote called `pummeler`[3] with R replication scripts in our GitHub repository[4]. We obtained the most recent American Community Survey 5-year dataset, 2010-2014 and the 2015 1-year dataset and merged them. We excluded data from 2010 and 2011 because these years used the 2000 US Census geography, leaving us with four years of individual-level data. We adjusted the 2015 weights to match 2012-14 thus obtaining a 4% sample of the US population, consisting of 9,222,637 million individual observations. Electoral results by county were scraped from nbc.com on 9 November 2016 (the day after the election). Using the merging strategy described in Flaxman et al. (2015) we ended up with 979 geographic regions. Real-valued data was standardized to have mean zero and variance one and categorical variables were coded in unary, omitting a reference category to aid in model fit.

We obtained state-level exit polling data obtained from foxnews.com for the following demographics: age, sex, and education. We used these to increase the size and diversity of our sample. In addition to the 979 geographic regions labeled with election outcome data, we used a total of 249 state-level subgroups, e.g. women in Florida, calculating the subgroup feature vectors as described above by restricting to the individual's in the census matching the demographic reported in the exit poll. As shown in Figure 1, our model is not overfitting, either to true outcome data (black) or exit poll data (blue). The best α parameter we found was 0.05 (where 0 is pure ridge regression). The number of parameters in the best model was 415 out of a total of 11,112 features. Using this model, we made predictions for a variety of groups derived from the census, as detailed in the next section.

4 Results: inferring who voted with ecological inference

In this section we show exit poll-style results which we inferred using the ecological inference methods described above. The tables give national results, and we will soon add maps showing local trends. CSV files with all of our estimates are available at the local level on our github.

[3] `www.github.com/dougalsutherland/pummeler`
[4] `www.github.com/flaxter/us2016`

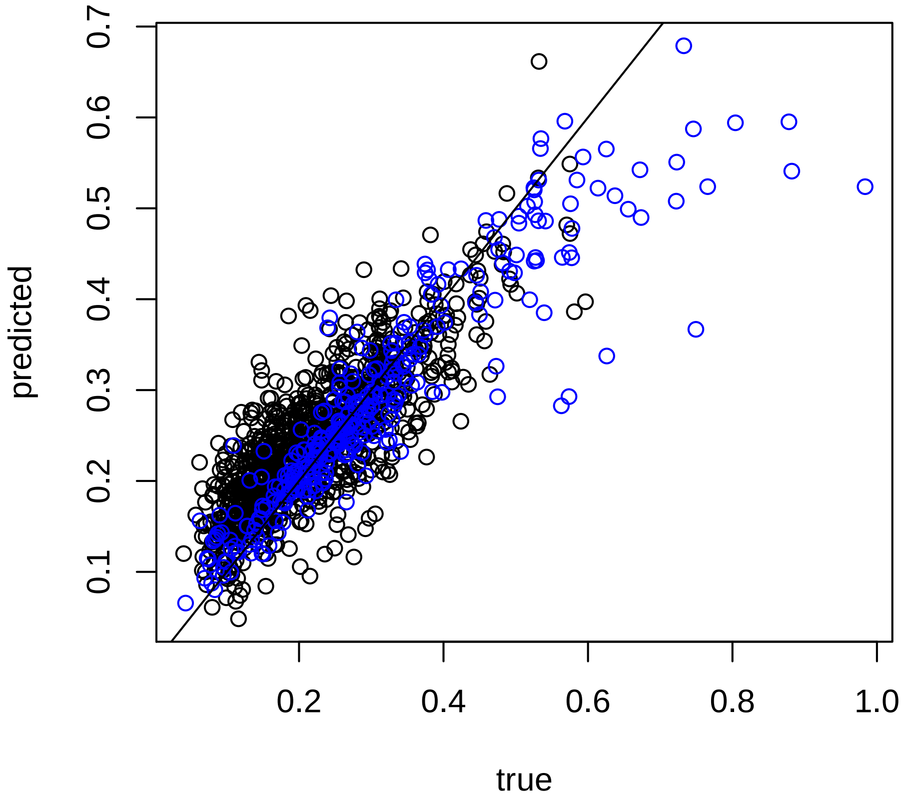

Figure 1: Fit of our model. Black dots are electoral outcomes by local area, while blue dots are exit polling (survey-based) state-level results by age, sex, and education.

	Clinton	Trump	Fraction of electorate	Participation Rate	Other / Non-voting
Men	0.45	0.55	0.47	0.50	0.50
Women	0.56	0.44	0.53	0.53	0.47

Table 1: Estimated demographic trends using our model. Columns show support for Clinton, Trump, fraction of the electorate made up by men (first row) vs women (second row), fraction of voting-age citizens who voted for Clinton or Trump, and fraction of voting-age citizens who voted for a third party or did not vote.

	Clinton	Trump	Fraction of electorate	Participation Rate	Other / Non-voting
18 <= age <= 29	0.62	0.38	0.17	0.42	0.58
30 <= age <= 44	0.54	0.46	0.25	0.54	0.46
45 <= age <= 64	0.46	0.54	0.39	0.58	0.42
age >= 65	0.45	0.55	0.18	0.47	0.53

	Clinton	Trump	Fraction of electorate	Participation Rate	Other / Non-voting
High-school or less	0.47	0.53	0.20	0.25	0.75
Some college	0.46	0.54	0.33	0.49	0.51
Bachelor's degree	0.51	0.49	0.30	0.84	0.16
Postgraduate education	0.58	0.42	0.17	0.85	0.15

	Clinton	Trump	Frac of elec	Part rate	Other / Non-voting
Born in the U.S.	0.42	0.58	0.93	0.62	0.38
Born in Puerto Rico, Guam, the U.S. Virgin Islands, or the Northern Marianas	0.54	0.46	0.01	0.42	0.58
Born abroad of American parent(s)	0.57	0.43	0.01	0.52	0.48
U.S. citizen by naturalization	0.74	0.26	0.06	0.42	0.58

	Clinton	Trump	Fraction	Participation	Other
hispanic	0.68	0.32	0.09	0.43	0.57
white	0.40	0.60	0.72	0.54	0.46
black	0.84	0.16	0.14	0.58	0.42
amerindian	0.55	0.45	0.00	0.26	0.74
asian	0.76	0.24	0.03	0.39	0.61
other/biracial	0.76	0.24	0.01	0.34	0.66

	Clinton	Trump	Frac	Part	Other / Non-voting
personal income <= 50000	0.53	0.47	0.64	0.44	0.56
50000 < personal income <= 100000	0.46	0.54	0.25	0.72	0.28
personal income > 100000	0.45	0.55	0.11	0.77	0.23

	Clinton	Trump	Part.	Other / Non-voting
Hearing difficulty	0.37	0.63	0.47	0.53
Vision difficulty	0.42	0.58	0.46	0.54
Independent living difficulty	0.50	0.50	0.28	0.72
Veteran service connected disability rating (percentage) $\geq 10\%$	0.07	0.93	0.82	0.18
Cognitive difficulty	0.52	0.48	0.23	0.77
Ability to speak English = well	0.73	0.27	0.31	0.69
Ability to speak English = not well or not at all	0.67	0.33	0.21	0.79
Gave birth to a child within the past 12 months	0.51	0.49	0.64	0.36
Grandparents living with grandchildren	0.46	0.54	0.44	0.56
Grandparents responsible for grandchildren	0.42	0.58	0.53	0.47
Insurance purchased directly from an insurance company	0.43	0.57	0.59	0.41
Medicare, for people 65 and older, or people with certain disabilities	0.44	0.56	0.46	0.54
Medicaid or similar	0.56	0.44	0.27	0.73
TRICARE or other military health care	0.26	0.74	0.72	0.28
VA used ever	0.21	0.79	0.70	0.30
Indian Health Service	0.69	0.31	0.22	0.78

	Clinton	Trump	Part	Other / Non-voting
Language other than English spoken at home	0.74	0.26	0.32	0.68
Mobility = lived here one year ago	0.45	0.55	0.55	0.45
Mobility = moved here from outside US and Puerto Rico	0.60	0.40	0.47	0.53
Mobility = moved here from inside US or Puerto Rico	0.56	0.44	0.48	0.52
Active duty military	0.45	0.55	0.56	0.44
Not enrolled in school	0.45	0.55	0.60	0.40
Enrolled in a public school or public college	0.61	0.39	0.39	0.61
Enrolled in private school, private college, or home school	0.66	0.34	0.53	0.47

	Clinton	Trump	Frac	Part	Other / Non-voting
18 <= age <= 29 & men	0.55	0.45	0.08	0.39	0.61
18 <= age <= 29 & women	0.71	0.29	0.08	0.37	0.63
30 <= age <= 44 & men	0.43	0.57	0.13	0.56	0.44
30 <= age <= 44 & women	0.55	0.45	0.14	0.58	0.42
45 <= age <= 64 & men	0.48	0.52	0.16	0.49	0.51
45 <= age <= 64 & women	0.58	0.42	0.21	0.61	0.39
age >= 65 & men	0.36	0.64	0.10	0.57	0.43
age >= 65 & women	0.49	0.51	0.10	0.48	0.52

	Clinton	Trump	Fraction	Participation	Other
hispanic men	0.66	0.34	0.04	0.39	0.61
white men	0.39	0.61	0.36	0.57	0.43
black men	0.77	0.23	0.05	0.44	0.56
amerindian men	0.50	0.50	0.00	0.28	0.72
asian men	0.74	0.26	0.01	0.37	0.63
hispanic women	0.71	0.29	0.04	0.39	0.61
white women	0.48	0.52	0.38	0.57	0.43
black women	0.87	0.13	0.07	0.61	0.39
amerindian women	0.50	0.50	0.00	0.26	0.74
asian women	0.79	0.21	0.02	0.40	0.60

	Clinton	Trump	Frac	Part	Other / Non-voting
personal income <= 50000 & men	0.56	0.44	0.25	0.37	0.63
personal income <= 50000 & women	0.63	0.37	0.36	0.40	0.60
50000 < personal income <= 100000 & men	0.40	0.60	0.15	0.67	0.33
50000 < personal income <= 100000 & women	0.53	0.47	0.13	0.84	0.16
personal income > 100000 & men	0.49	0.51	0.08	0.70	0.30
personal income > 100000 & women	0.62	0.38	0.03	0.80	0.20

5 Results: group-based exploratory data analysis

In order to explore which parameters were correlated with the outcome, we used penalized multinomial distribution regression using the same setup as above restricted to subsets of related categorical parameters, e.g. there are 6 different categories for marital status, as shown in Figure 4. In each case, we calculated the crossvalidated deviance for the multinomial logistic likelihood, which is a measure of fit allowing different models with different numbers of parameters to be compared. Smaller deviance is better. In Table 2 we show the top 25 performing features.

	feature	deviance	frac.deviance
1	RAC3P - race coding	0.04	0.86
2	ethnicity interacted with has degree	0.04	0.74
3	schooling attainment	0.04	0.72
4	ANC2P - detailed ancestry	0.04	0.83
5	OCCP - occupation	0.04	0.75
6	COW - class of worker	0.04	0.67
7	ANC1P - detailed ancestry	0.05	0.77
8	NAICSP - industry code	0.05	0.71
9	RAC2P - race code	0.05	0.70
10	age interacted with usual hours worked per week (WKHP)	0.05	0.69
11	sex interacted with ethnicity	0.05	0.65
12	MSP - marital status	0.05	0.61
13	FOD1P - field of degree	0.05	0.61
14	ethnicity	0.06	0.57
15	RAC1P - recoded race	0.06	0.54
16	sex interacted with age	0.06	0.57
17	has degree interacted with age	0.06	0.55
18	age interacted with personal income	0.06	0.76
19	sex interacted with hours worked per week	0.06	0.62
20	personal income interacted with hours worked per week	0.06	0.69
21	personal income	0.06	0.59
22	RACSOR - single or multiple race	0.07	0.42
23	has degree interacted with hours worked per week	0.07	0.59
24	hispanic	0.07	0.56
25	sex interacted with personal income	0.07	0.57

Table 2: Lower deviance values indicate a better fit. Larger fraction of deviance explained values are better. Models are fit to the outcome data as described in the text using only the categorical predictors or interactions of categorical predictors listed.

We visualized the most informative categorical features using ternary plots, considering not just their predictive ability for Clinton vs. Trump but also their predictive ability in terms of voting for other / not voting. The estimated coefficient for non-zero covariates is shown on the 3-dimensional probability simplex (visualized as a ternary plot) where the three corners represent the extreme case of 100% support for Clinton, Trump, or not voting. In Figure 2, for example, whites with degrees are close the Clinton corner, while whites without degrees are far from Clinton, and in between other / non-voting and Trump, slightly closer to Trump.

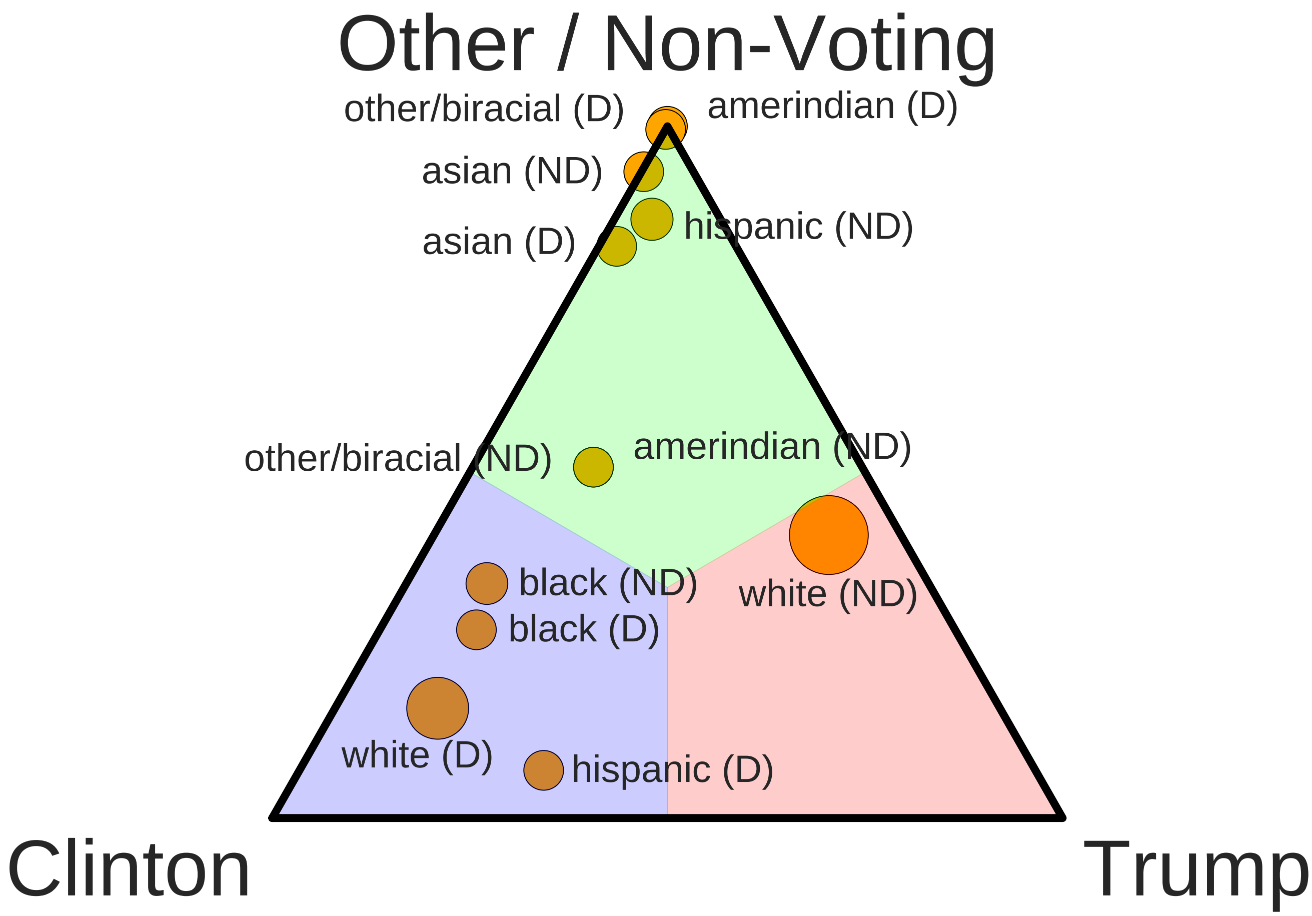

Figure 2: Ethnicity-has degree (D = has degree, ND = no degree)

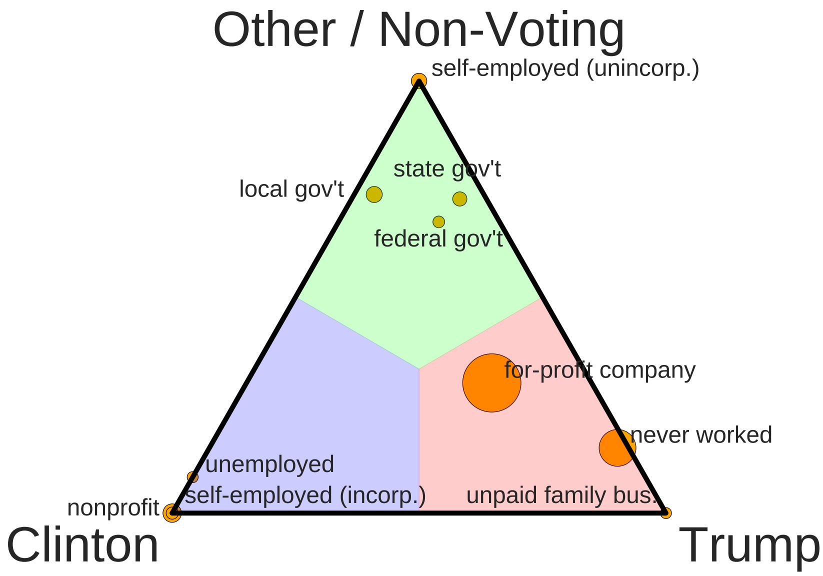

Figure 3: Class of worker `COW`

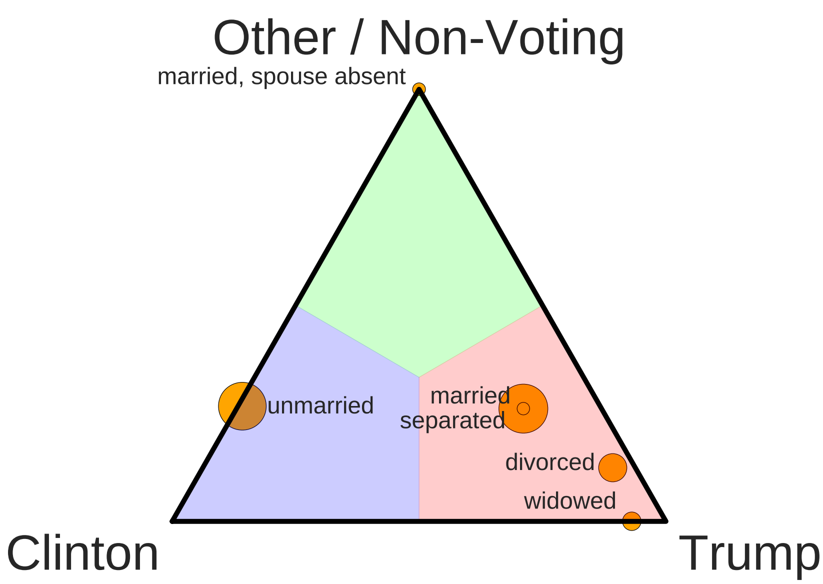

Figure 4: Marital status `MSP`

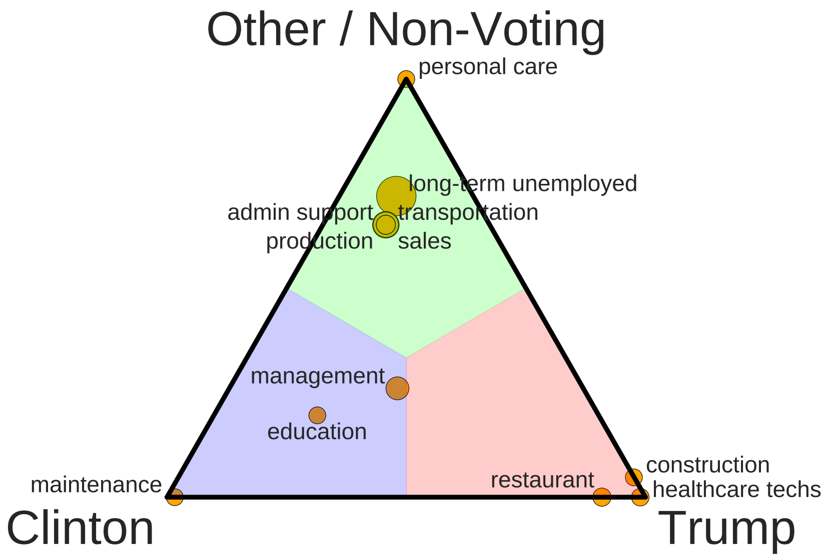

Figure 5: Occupation `OCCP`

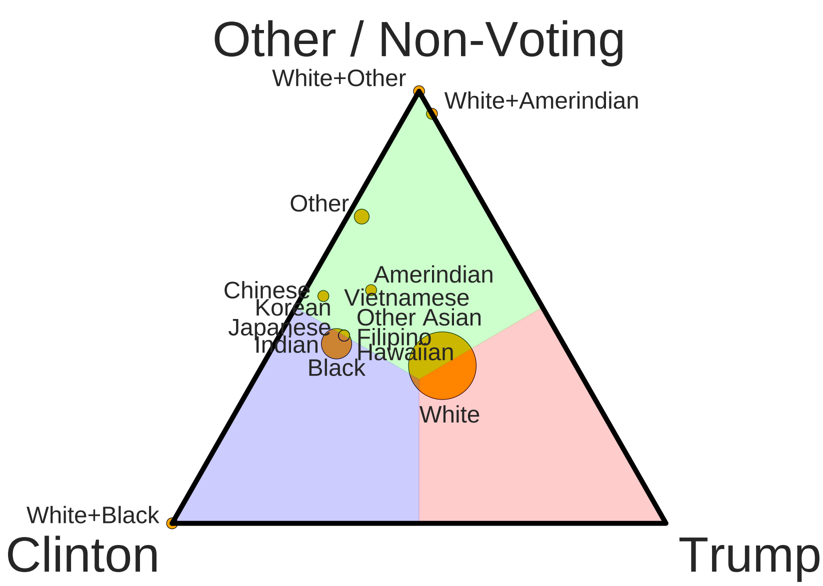

Figure 6: Race `RAC3P`

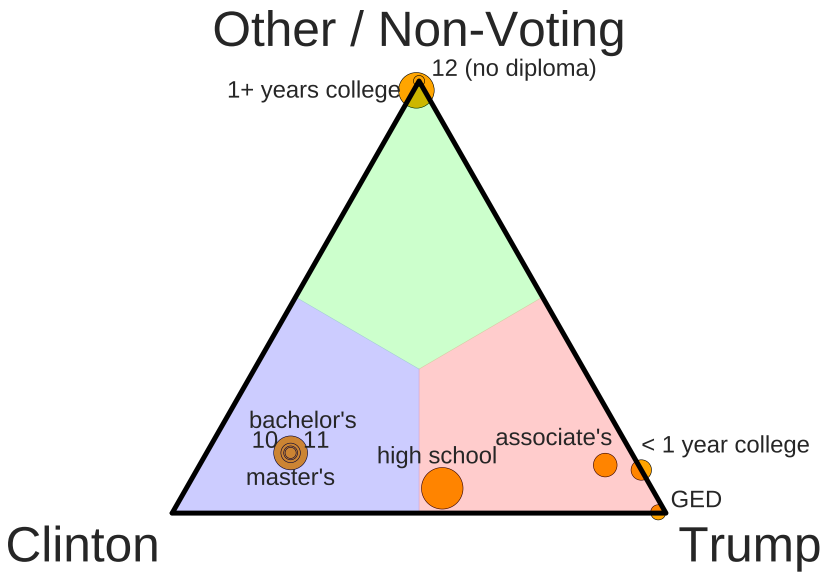

Figure 7: Education level `SCHL`

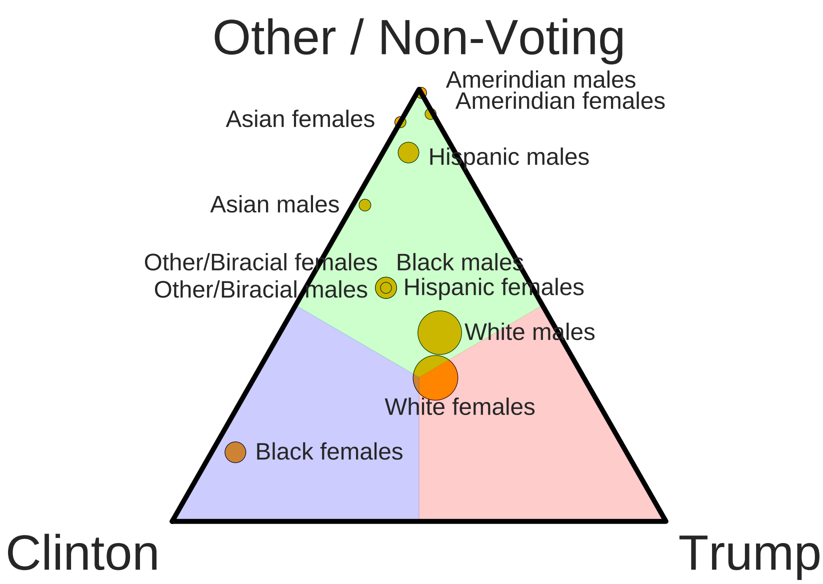

Figure 8: Sex-ethnicity

6 Conclusions

This is a work in progress—we are sharing it before the analysis is complete in order to get feedback and in the hopes that it can inform the thinking and analysis of others. In particular, we intend to make comparisons to 2012 data. We will also further explore the wealth of variables in the American Community Survey, add fine-grained spatial maps to complement the demographic modeling in Section 4 and visualize the distributions of real-valued covariates for the most important features discovered in Section 5. Finally, we will consider Bayesian alternatives to the penalized MLE approach in Section 5 so as to, e.g. infer posterior uncertainty regions over the location of our estimates in the ternary plots.

References

X Yu Felix, Ananda Theertha Suresh, Krzysztof Choromanski, Daniel Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. 2016.

Seth Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who Supported Obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298. ACM, 2015.

Michael T Gastner, Cosma Rohilla Shalizi, and Mark EJ Newman. Maps and cartograms of the 2004 us presidential election results. *Advances in Complex Systems*, 8(01):117–123, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning, JMLR: W&CP, Lille, France*, 2015.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

Z. Szabo, A. Gretton, B. Poczos, and B. Sriperumbudur. Two-stage Sampled Learning Theory on Distributions. *Artificial Intelligence and Statistics (AISTATS)*, February 2015.