SICSS-Oxford: Advances in Computational Experimental Social Sciences

Nuffield College Oxford

Ray Duch

June 22, 2019

Nuffield College Oxford Centre for Experimental Social Science - CESS

Computational methods in experimental design and analysis

- Part I: Brief introduction to CESS and Talk
- Part II: Micro-replications and experimental measurement error
- Part III: Designing virtual experiments with post-stratified Average Treatment Effects

Part I: Introduction

Nuffield Centre for Experimental Social Sciences (CESS)

- CESS Centres and Labs
  - Nuffield College Oxford
  - CESS Santiago
  - Nuffield CESS FLAME India
  - CESS China Nankai University
- CESS Online
  - Nuffield CESS Online facilities
  - Virtual non-deception CESS subject pool
  - UK, Ireland, India, China, U.S.,and Chile
- CESS in the field
- CESS Schools, Workshops, Visiting Post-/Pre-doc positions...

The Talk

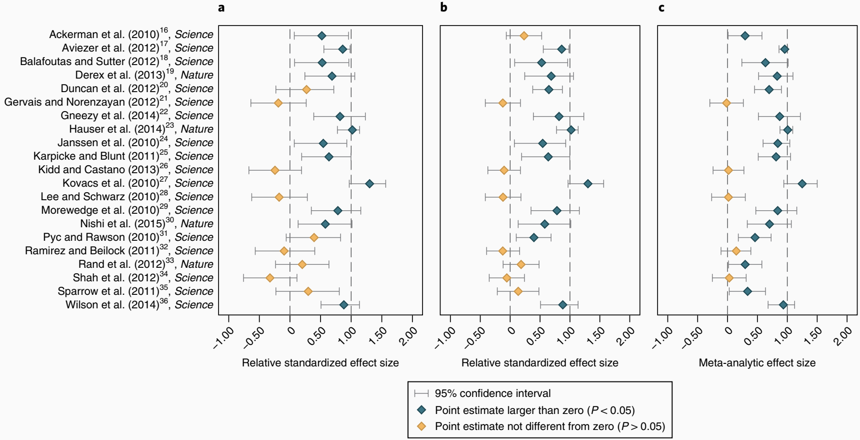
- Experimental perspective on themes you covered this week
  - Computational methods
  - Large data
  - Social media
  - Robustness/replication
- Part II
  - Detecting experimental measurement error
  - Vary the experimental context or mode
  - Machine learning for detection of heterogeneity
- Part III
  - Large-scale experimental interventions
  - Digital trace outcomes
  - Post-stratification with probability machines

Part II: Micro-replications

Data Generation

- Costs declining significantly
- Convenience samples are the norm
- Proliferation of data generation modes
- Democratic

There are Costs: Camerer et al 2018 Nature



Some Observations

- How do you know you have this experimental measurement error?
- You typically have no clue as to whether its an issue
- Note: this has nothing to do with external validity/representative sample/etc.

Micro-replications can help

- Maybe....
- But what micro-replication?
- In which micro-replication should you invest your research dollars?
- Multi- rather than single-mode replications are more informative of experimental measurement error

The Experimental Mode or Context

amazon mechanical turk

LUCID

respondi

KNOWLEDGE NETWORK

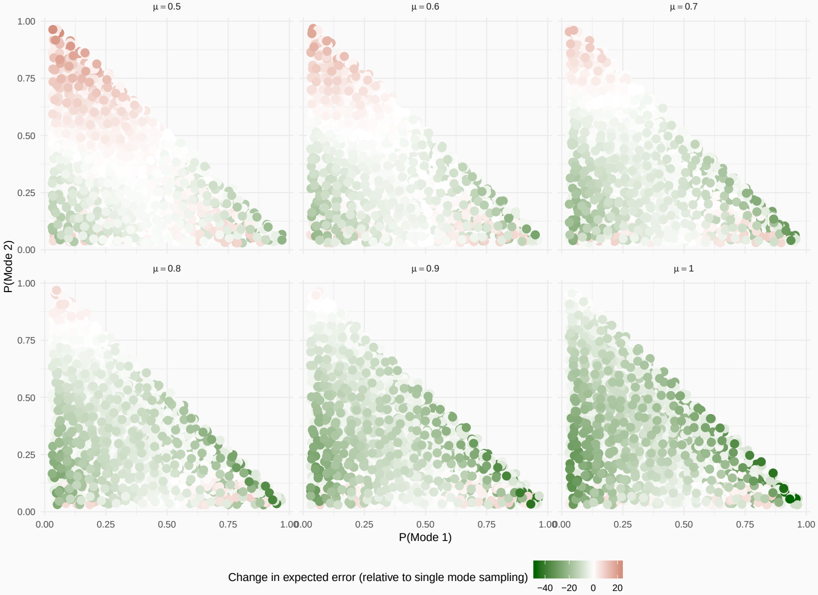
Centre for Experimental Social Sciences

ANES
American National Election Studies

Modes and Experimental Measurement Error

- do modes exaggerate measurement error, i.e., $ME_k > 0$
- resulting in $ATE_k^* = (ATE_T + ME_k)$
- multi-mode replication design may be informative when:
  - $ME_k \neq ME_{k^{\cdot}}$ and
  - there is a reasonably high probability the researcher can distinguish low from high error modes

Multiple-mode Replication Simulation



Illustrate: Lying Experiment (Duch Laroze Zakharov 2018)

- Outcome of interest: Lying about income from RET
- Treatment: Deduction rate that make it more expensive to lie
- Expectation: Lying declines if deduction rates rise

Lying Experiment Design (Duch Laroze Zakharov 2018

- 3 different tax rates (10%, 20% and 30%)
- Fixed at the group level
- Taxes are redistributed equally among group members
- Public good
- No excludability
- No social gains/losses
- No audits or fines
- 10 rounds
- Paid for one of them at random
- Fixed groups of 4 participants
- Random matching at the beginning

Design: each round

- RET: solve as many additions as possible in 60 sec
- two random two-digit numbers
- Information individual gross profit (before tax)
- Declare their income (to be taxed)
- Information individual net profit (after tax and redistribution)
- Differentiated by profit, tax and redistribution

Lying Experiments

CESS Oxford Lab	CESS Oxford Online	CESS Online UK	Mturk USA
CESS Oxford Subject Pool	CESS Oxford Subject Pool	CESS Online US Subject Pool	U.S. Crowd-sourced workers
1600 decisions/116 subjects	1367 Decisions/144 Subjects	696 Decisions/90 Subjects	2419 Decisions/390 Subjects
57% Rate of Lying	54% Rate of Lying	37% Rate of Lying	40% Rate of Lying

Conventional GLM Estimation

	Mode			
	Lab	Online Lab	Online UK	Mturk
Ability Rank	-0.500^{***}	-0.163^{***}	-0.163^{**}	-0.120^{***}
	(0.036)	(0.045)	(0.071)	(0.037)
20% Deduction	-0.123^{***}			
	(0.024)			
30% Deduction	-0.128^{***}	-0.184^{***}	0.042	0.018
	(0.025)	(0.025)	(0.038)	(0.021)
No Audit	-0.334^{***}	-0.127^{***}	-0.155^{***}	0.011
	(0.023)	(0.026)	(0.036)	(0.024)
Age	0.012^{***}	0.007^{**}	-0.0002	0.002^{**}
	(0.002)	(0.003)	(0.001)	(0.001)
Gender	0.002	0.100^{***}	-0.022	-0.004
	(0.022)	(0.025)	(0.035)	(0.020)
Constant	0.715^{***}	0.476^{***}	0.880^{***}	0.576^{***}
	(0.066)	(0.089)	(0.070)	(0.043)

BART Estimation

- Bayesian estimation strategy using tree-logic
- Highly flexible estimation strategy

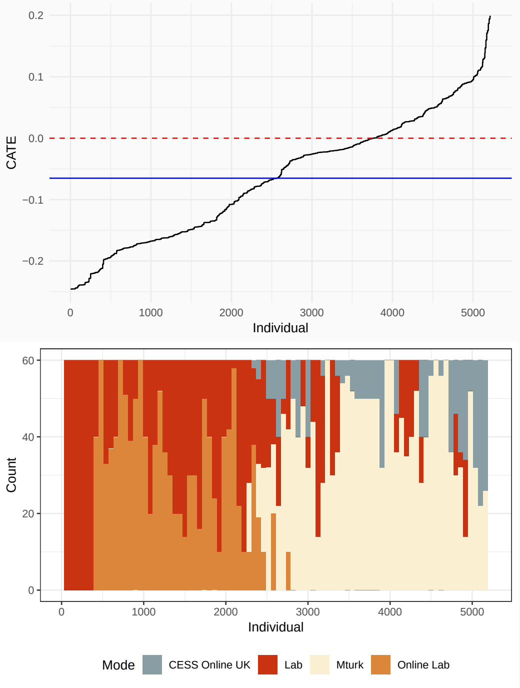
To recover individual estimates of treatment effect:

- Assume binary treatment
- Run BART on experimental data (the training set) to generate both model and predicted outcomes for observed data
- Invert treatment assignment of all observations, and pass through model (test set) to generate set of counterfactual predictions
- For each individual, i, $CATE = Y_{i,D=1} - Y_{i,D=0}$

BART: R Code

```r
# Separate outcome and training data
y <- df$report.rate
train <- df[,-1]

# Gen. test data where those treated become untreated, for use in calculating ITT
test <- train
test$treat.het <- ifelse(test$treat.het == 1,0,ifelse(test$treat.het == 0,1,NA))

# Run BART for predicted values of observed and synthetic observations
bart.out <- bart(x.train = train, y.train = y, x.test = test)

# Recover CATE estimates and format into dataframe
CATE <- c(bart.out$yhat.train.mean[train$treat.het == 1] - bart.out$yhat.test.mean[test$treat.het == 0],
          bart.out$yhat.test.mean[test$treat.het == 1] - bart.out$yhat.train.mean[train$treat.het == 0])

CATE_df <- data.frame(CATE = CATE)
covars <- rbind(train[train$treat.het == 1,c(2:5)], test[test$treat.het==1,c(2:5)])

CATE_df <- cbind(CATE_df,covars)
CATE_df <- CATE_df[order(CATE_df$CATE),]
CATE_df$id <- c(1:length(CATE))
```

All replication code available at `https://github.com/rayduch/Experimental-Modes-and-Heterogeneity`

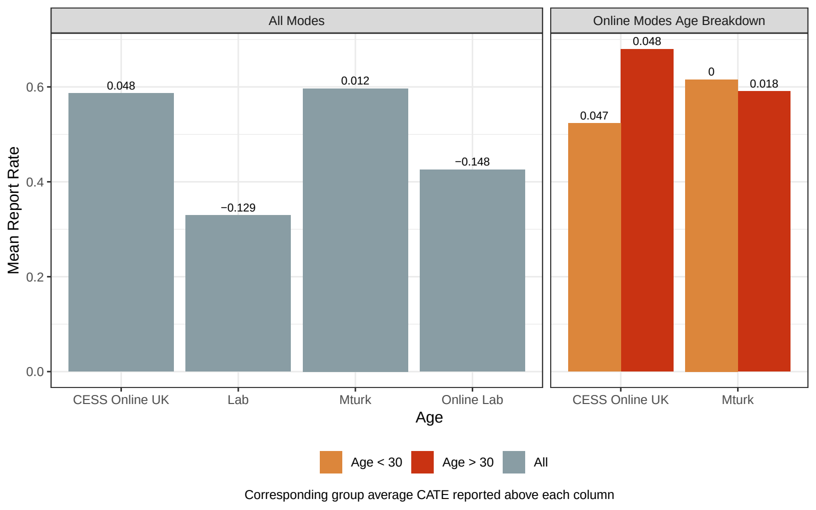


CATE
0.2
0.1
0.0
−0.1
−0.2
0
1000
2000
3000
4000
5000
Individual
Count
60
40
20
0
0
1000
2000
3000
4000
5000
Individual
Mode
CESS Online UK
Lab
Mturk
Online Lab

Real Effort Tasks Inter-Class Correlations Across Modes

Mode	(1)	(2)	(3)	(4)
Lab	0.768	0.768	0.636	0.85
	(0.018)	(0.018)	(0.039)	(0.047)
Lab Online	0.807	0.76	0.762	0.767
	(0.018)	(0.017)	(0.021)	(0.047)
Online UK	0.88	0.827	0.827	0.752
	(0.011)	(0.018)	(0.026)	(0.029)
MTurk	0.758	0.758	0.782	0.828
	(0.015)	(0.012)	(0.024)	(0.026)
Deduction Rate	10%	30%	10%	30%
Audited?	No	No	Yes	Yes

Comparing Percentages of Actual Earnings Reported



India Measurement Error Experiments

Coeff	S.E.	t-statistic	p	Mode	Error	Incentivised?
-0.74	0.47	-1.57	0.12	MTurk	Control	No
-0.83	0.47	-1.76	0.08	MTurk	High	No
-3.85	0.51	-7.52	0.00	CESS Online	Control	No
-3.23	0.49	-6.64	0.00	CESS Online	High	No
-1.16	0.49	-2.35	0.02	MTurk	Control	Yes
-1.00	0.33	-3.01	0.00	MTurk	High	Yes

Table 2: Induced measurement error model results

Part II Recommendations

- micro-replication with multiple-modes
- machine learning strategies for estimating CATE that might be mode-related
- incorporate measurement strategies in experimental design that might explain mode-related experimental measurement error

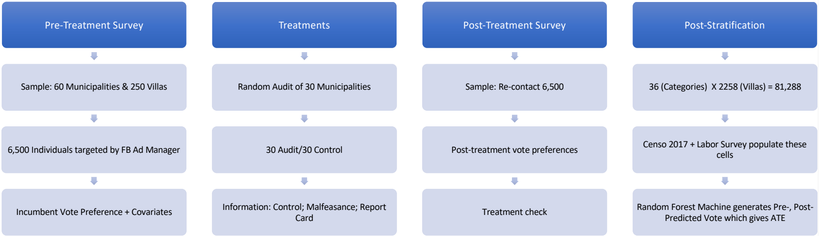
Part III: Designing virtual experiments with post-stratified Average Treatment Effects

The CESS Chile Audit Experiment Design

- Random Assignment of Audit "Shocks" to Municipal Voters
- Random Assignment of Information Treatments to targeted zip codes, aka Villas
- Efficient pre-and post-treatment measurement
- Post-stratification of ATE with random forest estimation
- Inspired by the Digital Trace and Vote India project that have focused on simple election outcomes rather than causal effects

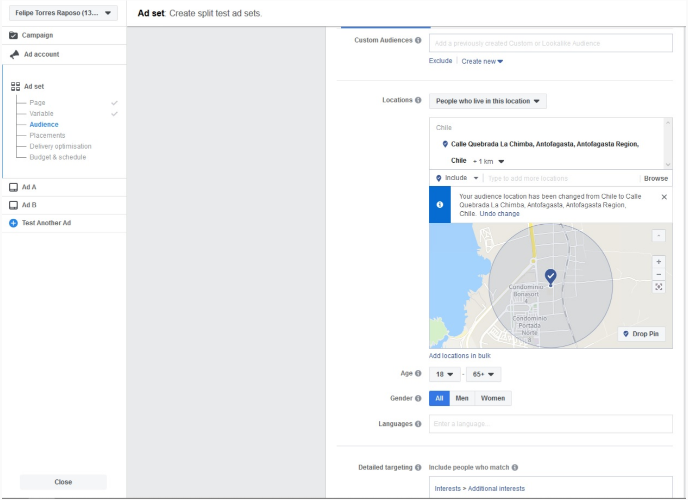
Chile 345 Municipalities

CESS Santiago Chile Audit Experiment Design



Facebook Sampling

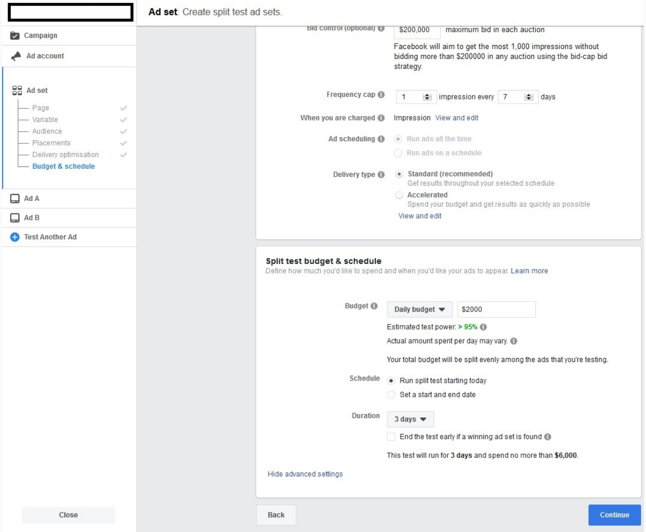
Facebook Ad Manager Dashboard



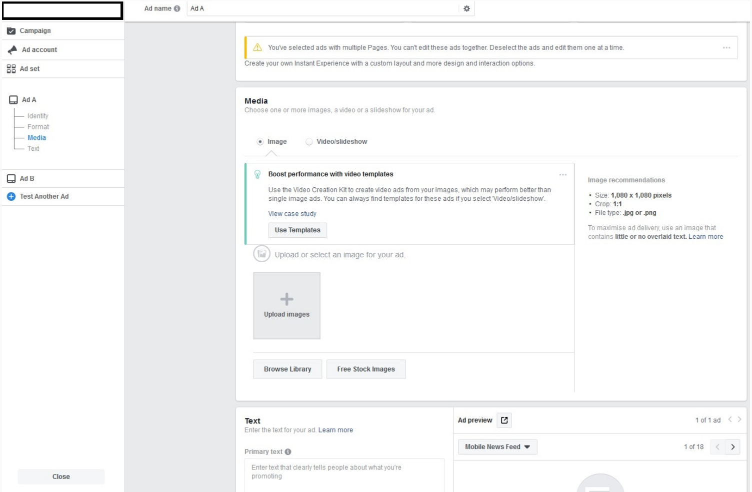
Pre-Treatment Facebook Sampling

- Facebook ad manager allows us to target subjects in specific Villas
- Banner recruitment ads sent to FB subscribers in specific Villas
- 6,500 pre-treatment online qualtrics interviews completed

Facebook Ad Manager Dashboard



Facebook Ad Manager Dashboard



Treatment Assignment and Sample

Treatments

- Treatment assignment and pre-treatment sample sizes are in the following Figure
- Contraloria will cooperate in randomly assigning audits to municipalities
- two X two factorial design will allow us to assess two different information treatments:
  - audit results
  - performance report
- control conditions
  - no audits
  - no information

Treatment Summary

Treatments

Block1: Ranking Level	High						Medium						Low					
Block 2: Municipal Level	Audit (15)				No Audit (15)		Audit (10)				No Audit (5)		Audit (5)				No Audit (5)	
Information treatments	No Info (40)	Audit + Perf. (40)	Audit (40)	Perf. (40)	No Info (10)	Perf. (20)	No Info (4)	Audit + Perf. (6)	Audit (6)	Perf. (4)	No Info (6)	Perf. (14)	No Info (2)	Audit + Perf. (4)	Perf. (2)	No Info (4)	No Info (4)	Perf. (6)
Pre-treatment Sample	1000	1000	1000	1000	250	500	100	150	150	100	150	350	50	100	50	100	100	150

Total sample: 6,500 individuals

Information Treatments

Malfeasance Information Treatment

Information Campaigns

- Treated Villas receive saturation information treatments
  - FB ads
  - Twitter ads
  - Instagram ads
  - Note: these are targeted at all social media in the treated Villas
- Post-Treatment Survey
  - Re-contact 6,500 initially surveyed subjects
  - Evaluations of municipal government and voting intentions

Average Treatment Effect Estimation

Average Treatment Effect Estimation

Three steps to obtaining area-estimates of ATE:

- identification of individual characteristics to use in ATE prediction model (socio-economic categories);
- estimation, at the individual-category level, of ATE;
- weighting predicted ATEs by cell counts and summing over the area of interest to recover estimates of ATE.
- Outcome of interest is change in support for municipal government incumbent as measured by the pre- and post-surveys.

Average Treatment Effect Estimation

- Individual-level characteristics X from pre-treatment survey;
- Missing values in X are imputed with a random-forest multiple-imputation strategy implemented via the packages `ranger` and `missForest`;
- define the individual-level category C_g, for categories $g = 1, ..., G$ as a unique realization of the set of variables which compose X, i.e. $C_g = \{X_1 = x_1, ..., X_m = x_m\}$.

Estimation Categories for Observed Data

	Categories (C)	Audit/Info	Audit/NoInfo
Gender	Two	600	575
Education	Two	600	575
Income	Three	400	382
Age	Three	400	383
Total	36	33	32

Table 3: Number of cell categories and number of subjects within cells for pre-treatment survey

Estimation of ATE: Change in Vote for Municipal Incumbent

- Pre-treatment Estimate
  - Estimate the probability of an individual-category voting for a municipal mayor prior to treatment;
  - quantity of interest is per group and is $\mathrm{P}_0 g\left(+|C\right)$
- Post-treatment Estimate
  - Estimate the probability of an individual-category voting for a municipal mayor prior to treatment;
  - quantity of interest is per group and is $\mathrm{P}_1 g\left(+|C\right)$

Average Treatment Effects: High Risk 30 Municipalities

	Audit Treatment Effect						
	Pre-Treatment			Post-Treatment			
	Audit/No Info	No Audit/No Info	Diff/No Info	Audit/No Info	No Audit/No Info	Diff/No Info	ATE
Male	0.6	0.5	-0.1	0.7	0.5	-0.2	-0.1
Female	0.6	0.5	-0.1	0.8	0.5	-0.3	-0.2

	Audit Information Treatment Effect						
	Pre-Treatment			Post-Treatment			
	No Info	Audit Info	Diff	No Info	Audit Info	Diff	ATE
Male	0.8	0.7	-0.1	0.8	0.5	-0.3	-0.2
Female	0.6	0.5	-0.1	0.6	0.3	-0.3	-0.2

Likelihood Estimation

Likelihood Estimation

- Train a probability machine to estimate proportion of individuals who express a positive opinion of their incumbent municipal politician before treatment (P_0) and following the treatment (P_1);

- Make a prediction for each group before and after treatment;

$$\hat{\mathrm{P}}_\tau\left(+|C\right) = \varphi\left(C\right); \qquad \forall\tau = \{0,1\} \tag{1}$$

- estimate the global error via MSPE1 and approximate distribution with Normal density:

$$\mathrm{P}_g\left(+|C\right) \sim \mathrm{N}\left(\varphi\left(C_g\right), \left(\hat{\sigma}_{\text{RMSE1}}\right)^2\right); \tag{2}$$

Area Estimation

Area Estimation: Overall National ATE

- Cell counts: $Q_g = \sum_h 1(x_h = C_g)$;
- Estimation of National treatment effect:

$$ATE = \frac{\sum_g (\hat{\mathrm{P}}_{1g}(+|C) - \hat{\mathrm{P}}_{0g}(+|C)) \times Q_g}{\sum_g Q_g}; \tag{3}$$

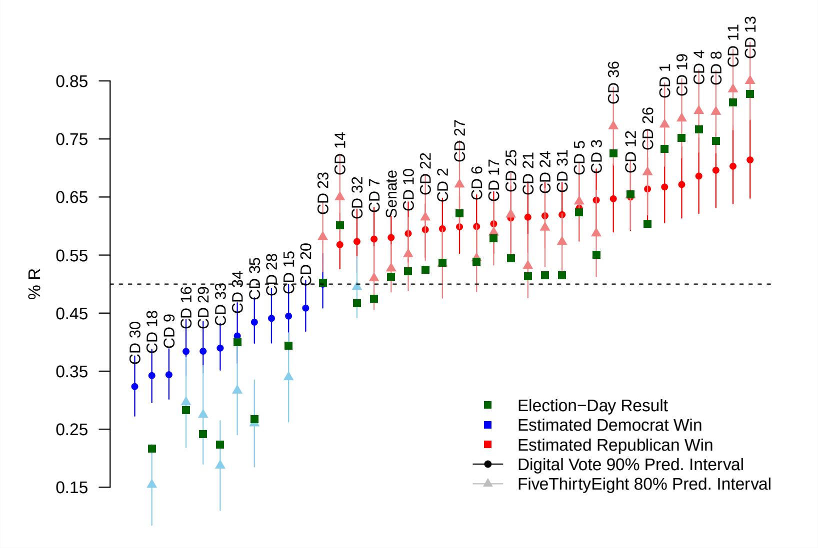
Area Estimation: Specific Municipal ATE

- Assume there are $s = 1....S$ municipal units;
- Municipal cell counts: $Q_{gs} = \sum_h 1(x_h^{muni} = C_{gs}, x^{muni} = s)$;
- Estimation of an s municipal treatment effect:

$$ATE_s = \frac{\sum_g (\hat{\mathrm{P}}_{1gs}(+|C) - \hat{\mathrm{P}}_{0gs}(+|C)) \times Q_{gs}}{\sum_{gs} Q_{gs}}; \quad (4)$$

Illustrated Results Cerina and Duch

Results



% R
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
CD 30
CD 18
CD 9
CD 16
CD 29
CD 33
CD 34
CD 35
CD 28
CD 15
CD 20
CD 23
CD 14
CD 32
CD 7
Senate
CD 10
CD 22
CD 2
CD 27
CD 6
CD 17
CD 25
CD 21
CD 24
CD 31
CD 5
CD 3
CD 36
CD 12
CD 26
CD 1
CD 19
CD 4
CD 8
CD 11
CD 13
Election−Day Result
Estimated Democrat Win
Estimated Republican Win
Digital Vote 90% Pred. Interval
FiveThirtyEight 80% Pred. Interval

Results

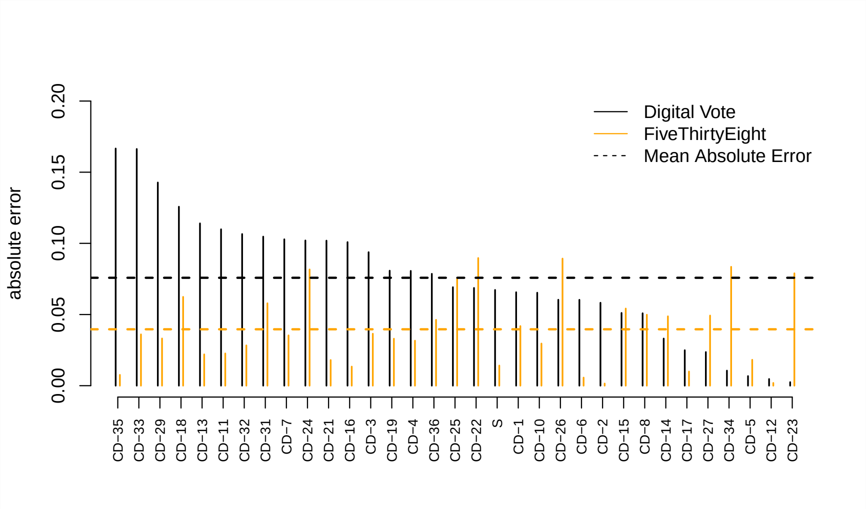
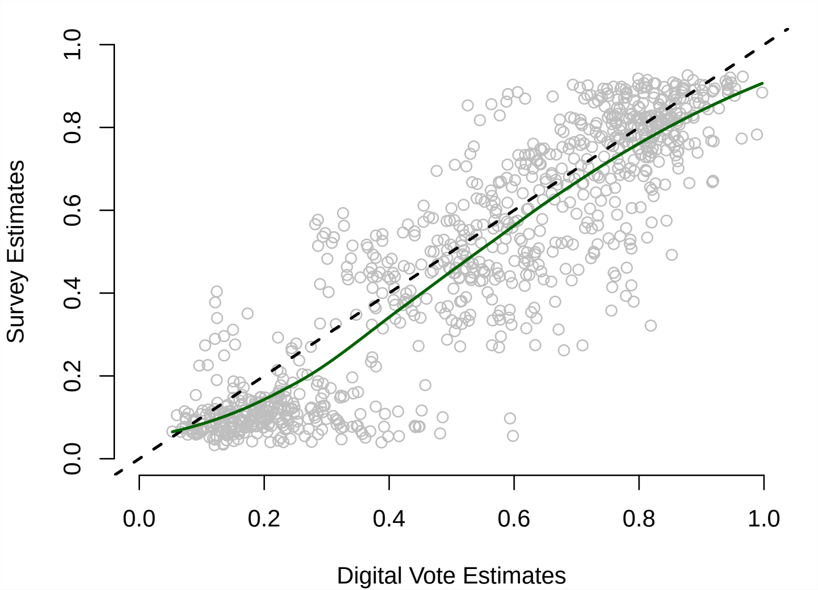


Figure 1: Mean Absolute Error: FiveThirtyEight and Digital Vote

Results



Part III: Recommendations

Part III: Recommendations

- Virtual sampling advantage: incorporates digital trace intelligence including geo-location
- FB Ad Manager: facilitates information treatments administered to very large samples of the population
- Area Post-stratification with machine learning: convenience sample + detailed census and survey data + random forest predictions = ATE for nation and sub-national units