Election modeling in 2024: a conformal inference approach

John Cherian [1] Lenny Bronner [2] Emmanuel Candès [1, 3]

[1]Department of Statistics, Stanford University [2]The Washington Post [3]Department of Mathematics, Stanford University

Summary

- Preliminary election results on the night of November 5th ⇝ who will win the presidency and Congress?
- **Ensembled prediction rule**: combines *fundamentals* model (based on obs. shifts in voter preferences) and *extrapolation* model (based on obs. vote counting process)
- **Assumption-lean inference**: *bootstrap* + *conformal inference* yield final prediction intervals

Set-up and notation

Data

Observe $\{X_i, R_{it}, D_{it}\}_{i=1}^N$ for N counties over times $t \in \{0, \ldots, 100\}$ (% reporting)

X_i : covariates for county i (racial composition, education, income...)

R_i : Republican votes in county i

D_i : Democratic votes in county i

Estimands: aggregate outcomes

$$\text{Margin}_{\text{PA}} = \frac{\sum_{i \in \text{PA}} D_{i,100} - R_{i,100}}{\sum_{i \in \text{PA}} D_{i,100} + R_{i,100}}$$

Goals

$$\widehat{\text{Margin}}_{\text{PA}} \approx \text{Margin}_{\text{PA}}$$

$$\mathbb{P}\left(\text{Margin}_{\text{PA}} \in \widehat{C}_{\text{PA}}\right) \approx 90\%$$

- Additional forecasts for Electoral College and Senate control

Background

Prior work

- Greben et al. (2006) cluster reporting units using previous elections and extrapolate from within-cluster observations
- Pavia et al. (2008) fit Gaussian Process regression with well-specified covariance kernel
- Cherian et al. (2021) aggregate county-level conformalized quantile regressions via equi-correlated Gaussian model

Problem

Prediction error distribution is *non-stationary* over elections

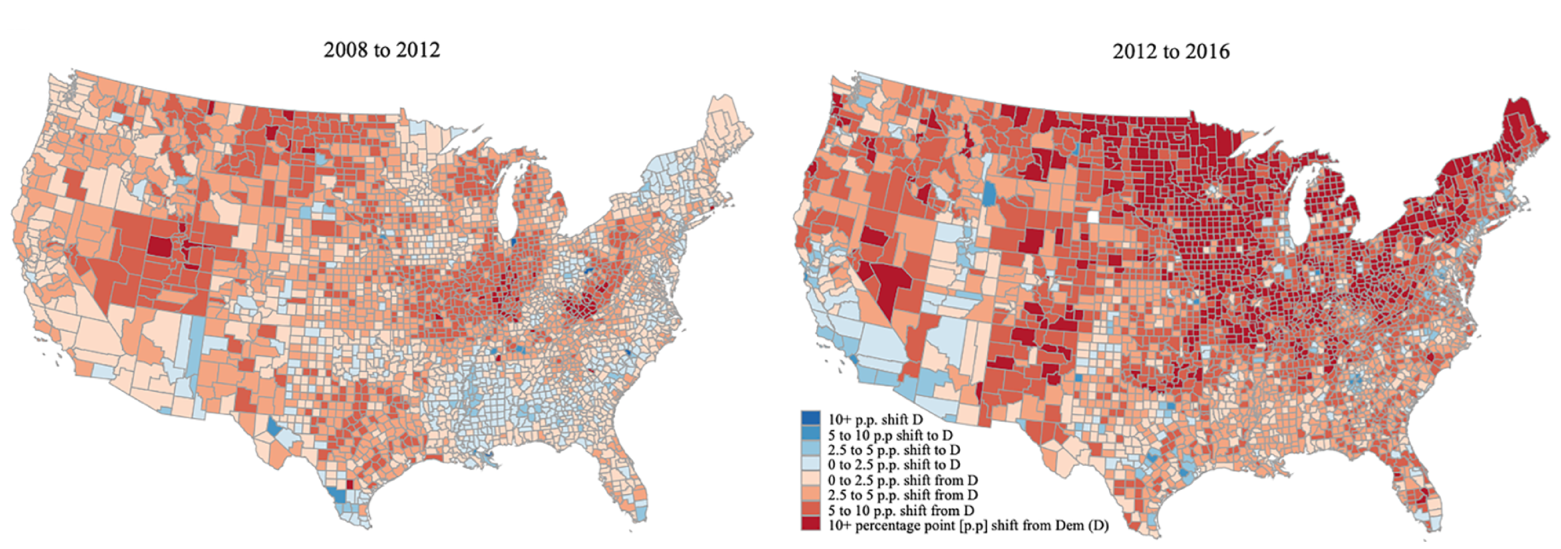
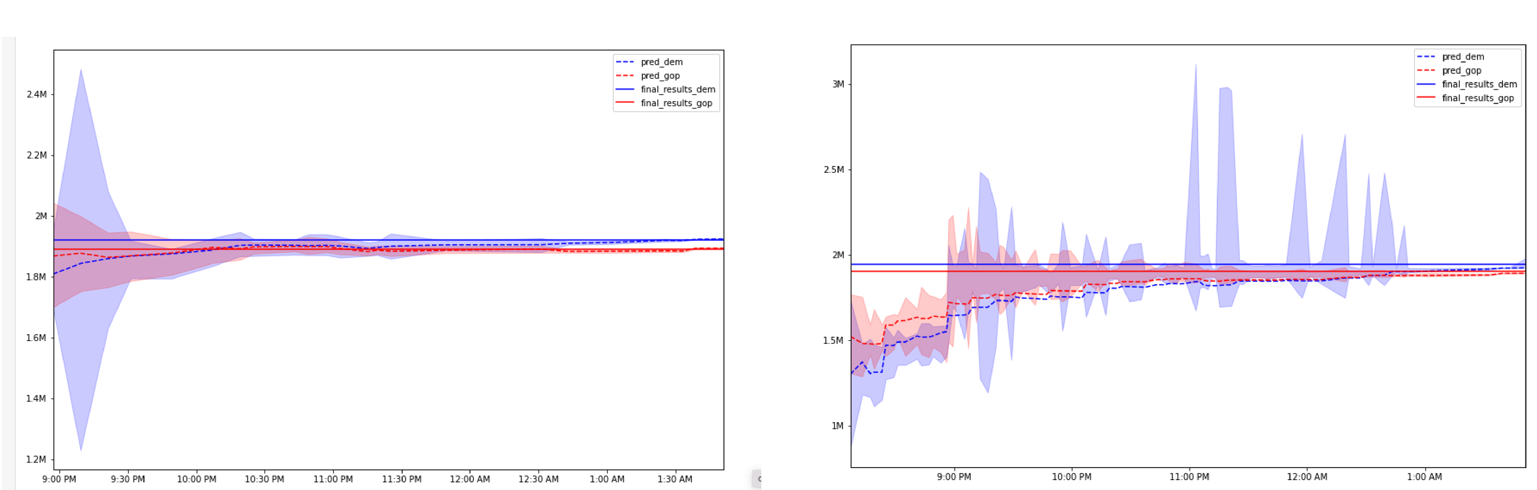


Figure: County vote swings in the 2008-2012 and 2012-2016 presidential election cycles.

Problem

Previous election model is **unstable**



Prediction rule

Estimand

$$Y_i := \underbrace{\frac{D_{i,100} - R_{i,100}}{D_{i,100} + R_{i,100}}}_{\text{unit margin}} \qquad Z_i := \underbrace{\frac{D_{i,100} + R_{i,100}}{D^b_{i,100} + R^b_{i,100}}}_{\text{turnout factor}}$$

$D^b_{i,100}, R^b_{i,100}$ are the previous election result in that county

Fundamentals prediction rule

- Using fully-reported counties, fit models $\widehat{f}_Y(\cdot)$ and $\widehat{f}_Z(\cdot)$ for Y_i and Z_i
- Yields estimator for aggregate margin:

$$\widehat{\text{Margin}}_{\text{PA}} = \frac{\sum_{i \in \text{PA}} w_i \cdot \widehat{f}_Y(X_i) \cdot \widehat{f}_Z(X_i)}{\sum_{i \in \text{PA}} w_i \cdot \widehat{f}_Z(X_i)} \qquad \text{where } w_i = D^b_{i,100} + R^b_{i,100}$$

- **Our approach**: $\widehat{f}(\cdot)$ uses cross-validated ridge regression where X_i includes previous unit margin, race, and education

Why ridge?

- AP data is imperfect (esp. early in election night) ⇝ **need good tools for outlier detection**
- Model is most important in early stages of election night ⇝ $\mathbf{n \approx 250}$

Extrapolation prediction rule

Motivation

- *Nearly* finished counties ⇝ does reported unit margin predict final unit margin?
- State-specific voting rules may lead to "*blue* or *red* shifts" (c.f. PA in 2020, CA in 2018)

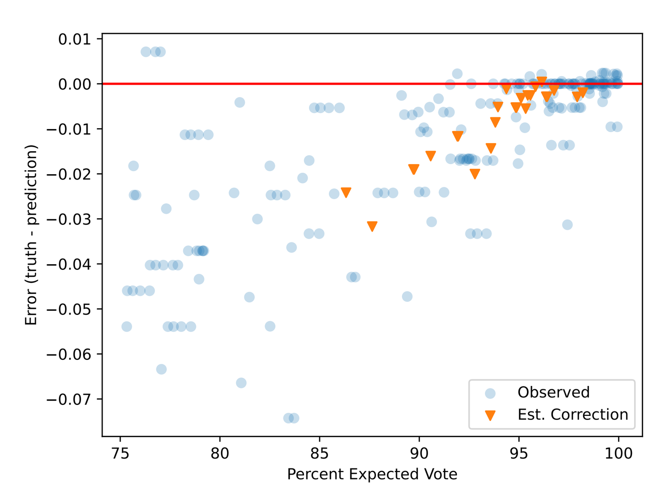


Figure: Extrapolation error in FL (2020 pres.)

$$\text{Error}_{\text{extrap.}} = \frac{D_{i,100} - R_{i,100}}{D_{i,100} + R_{i,100}} - \frac{D_{i,\text{current}} - R_{i,\text{current}}}{D_{i,\text{current}} + R_{i,\text{current}}}$$

- Within-state extrapolation error is predictable
- Est. correction obtained via local regression: $\widehat{h}_i$

Ensembled prediction rule

Variance-minimizing weights:

$$\widehat{Y}_i = \frac{\sigma_h^2 \cdot \widehat{f}_Y(X_i) + \sigma_f^2 \cdot \widehat{h}_i}{\sigma_h^2 + \sigma_f^2}$$

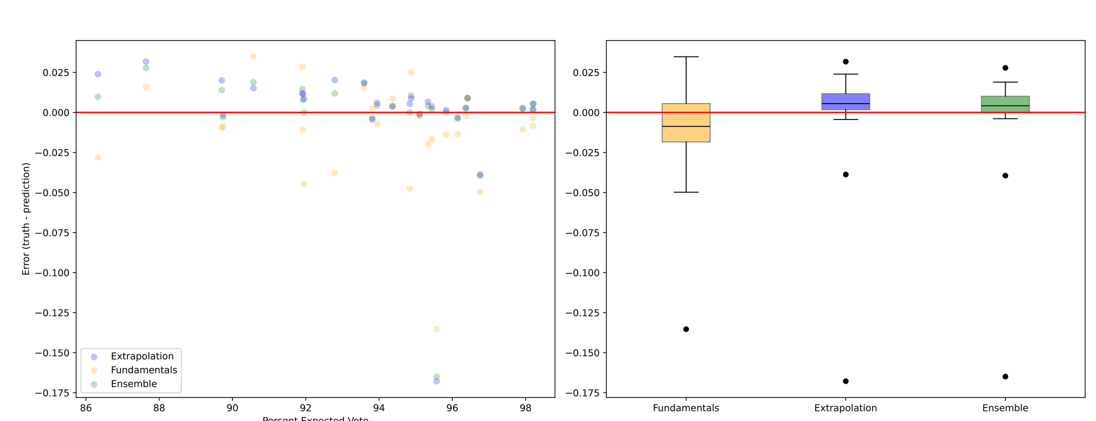


Figure: Prediction errors in FL (2020 pres.)

Predictive inference: background

Goal

Want a model P for the *joint* distribution of

$$\left\{Y_{n+i,100} - \widehat{Y}_{n+i}, Z_{n+i,100} - \widehat{Z}_{n+i}\right\}_{(n+i) \in \text{unobs.}} \sim P$$

- Assuming a statistical model, i.e., $P \in \{P_\theta\}_{\theta \in \Theta}$, is fraught
- Standard spatiotemporal methods (kriging, random effects model) have poor predictive coverage ⇝ Gaussian assumption is problematic

Predictive inference: our approach

Model-free methods (e.g., conformal inference) target *marginal coverage*

Theorem 2 (Gibbs, Cherian, & Candès, 2023)

Given any prediction rule $f(\cdot)$ and an exchangeable dataset $\{(X_i, Y_i)\}_{i=1}^{n+1}$ with Y_{n+1} unobs.,

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) \mid X_{n+1} \in G) = 1 - \alpha \qquad \text{for all } G \in \mathcal{G}$$

Reframing this work

- $\widehat{C}(\cdot)$ obtained via (modified) quantile regression (QR) on residuals (aka conformity scores)
- **Key insight**: conformal inference corrects over-fitting bias of high-dim. QR on prediction errors

Assumption

If I fit our prediction rule to *all of the data* on election night,

$$\left\{Y_{i,100} - \widetilde{Y}_i, Z_{i,100} - \widetilde{Z}_i\right\}_{i \in [N]}$$ are independent (but **not** identically distributed)

We can estimate $\widehat{Y}_i - \widetilde{Y}_i$ via **model-free bootstrap** (Politis (2015))

We can model heteroskedasticity in $Y_i - \widetilde{Y}_i$ via **conformal prediction**

Algorithm

1. Run conformal method (debiased QR) on leave-one-out residuals for $\alpha \in \{0.01, \ldots, 0.99\}$
   ⇝ CDF est. for $Y_{n+k} - \widehat{Y}_{n+k} \mid G_{n+k}$ and $Z_{n+k} - \widehat{Z}_{n+k} \mid G_{n+k}$
   - **Our approach**: run method over sub-groups that historically capture heteroskedasticity
2. Compute $\mathbf{U} = \left\{U_i^Y, U_i^Z\right\}$ by evaluating the estimated CDFs at the observed values of Y_i and Z_i
3. Create B datasets $\left\{X_i, Y_i^{(1)}, Z_i^{(1)}\right\}, \ldots, \left\{X_i, Y_i^{(B)}, Z_i^{(B)}\right\}$ by sampling (w/ replacement) from $\mathbf{U}$
4. Re-compute prediction rule $\left\{\widehat{Y}^{(b)}(\cdot), \widehat{Z}^{(b)}(\cdot)\right\}_{b=1}^B$ on bootstrap data sets
5. Sample B sets of *new* test errors $\left(\epsilon_{n+i}^{(b),Y}, \epsilon_{n+i}^{(b),Z}\right)$ from conformal model

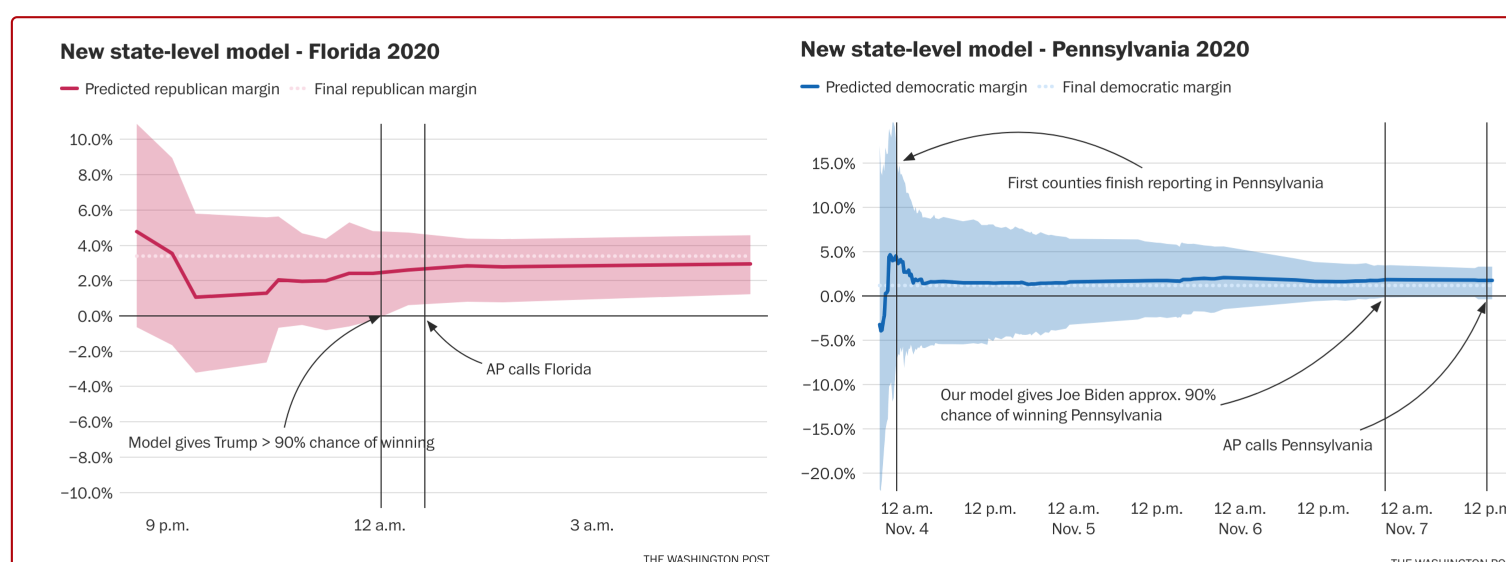
Bootstrap pivot

$$\widehat{\text{Margin}}_{\text{PA}}\left(\widehat{Y}_{n+i} + \epsilon_{n+i}^{(b),Y}, \widehat{Z}_{n+i} + \epsilon_{n+i}^{(b),Z}\right) - \widehat{\text{Margin}}_{\text{PA}}\left(\widehat{Y}_{n+i}^{(b)}, \widehat{Z}_{n+i}^{(b)}\right) \overset{d}{\approx}$$

$$\widehat{\text{Margin}}_{\text{PA}}(Y_{n+i}, Z_{n+i}) - \widehat{\text{Margin}}_{\text{PA}}\left(\widehat{Y}_{n+i}, \widehat{Z}_{n+i}\right)$$

6. Output

$$\widehat{C}_{\text{PA}} = \left[\widehat{\text{Margin}}_{\text{PA}} + Q_{\alpha/2}(\text{Pivot}), \widehat{\text{Margin}}_{\text{PA}} + Q_{1-\alpha/2}(\text{Pivot})\right]$$



Acknowledgments

Data Science: Dara Gold, Diane Napolitano

Engineering: Jen Haskell, Stewart Bishop, Dana Cassidy, Ben King, Alexis Barnes, Anthony Pesce, Claire Helms, Daniel Kao, Emily Liu

Graphics: Ashlyn Still