Ecological Inference

Simone Zhang

March 2017

With thanks to Gary King for slides on EI.

What is ecological inference?

Definition: Ecological inference is the process of using aggregate (i.e., "ecological") data to infer *discrete* individual-level relationships of interest when individual-level data are not available.

Why is this a problem?

- Ecological fallacy: believing that relationships that hold at the group level necessarily also hold at the individual level
  - In *Suicide* (1897), Émile Durkheim tries to study the relationship between religion and suicide rates.
  - Durkheim finds that suicide rates are higher in more Protestant Prussian provinces than in more Catholic ones
  - Durkheim uses this to conclude that Protestants are more likely to commit suicide
  - Note here Durkheim is analytically interested in one grouping (by religion), but actually has data available on a different grouping (by geography)
- Another example:
  - Simpson's paradox: relationships at individual level can reverse at the aggregate level (Berkeley admissions example)

If you can avoid making ecological inferences, do so!

1. **Public policy:** Applying the Voting Rights Act.
2. **History:** Who voted for the Nazi's?
3. **Marketing:** What types of people buy your products?
4. **Banking:** Are banks complying with red-lining laws? Are there areas with certain types of people who might take out loans but have not?
5. **Candidates for office:** How do good representatives decide what policies they should favor? How can candidates tailor campaign appeals and target voter groups?
6. **Sociology:** Do the unemployed commit more crimes or is it just that there are more crimes in unemployed areas?
7. **Economics:** With some exceptions, most theories are based on assumptions about individuals, but most data are on groups.

If you can avoid making ecological inferences, do so!
Some of those who aren't so lucky:

1. **Public policy:** Applying the Voting Rights Act.
2. **History:** Who voted for the Nazi's?
3. **Marketing:** What types of people buy your products?
4. **Banking:** Are banks complying with red-lining laws? Are there areas with certain types of people who might take out loans but have not?
5. **Candidates for office:** How do good representatives decide what policies they should favor? How can candidates tailor campaign appeals and target voter groups?
6. **Sociology:** Do the unemployed commit more crimes or is it just that there are more crimes in unemployed areas?
7. **Economics:** With some exceptions, most theories are based on assumptions about individuals, but most data are on groups.

If you can avoid making ecological inferences, do so! Some of those who aren't so lucky:

8. **Education:** Do students who attend private schools through a voucher system do as well as students who can afford to attend on their own?
9. **Atmospheric physics:** How can we tell which types of the vehicles actually on the roads emit more carbon dioxide and carbon monoxide?
10. **Oceanography:** How many marine organisms of a certain type were collected at a given depth, from fishing nets dropped from the surface down through a variety of depths.
11. **Epidemiology:** Does radon cause lung cancer?
12. **Changes in public opinion:** How to use repeated independent cross-sectional surveys to measure individual change?

Notation for the problem

We can think of the problem as the following:

	Vote	**No vote**	
Black	β_i^b	1 - β_i^b	X_i
White	β_i^w	1 - β_i^w	$1 - X_i$
	T_i	$1 - T_i$	

- We observe:
  - T_i : turnout (fraction who people who voted in precinct i)
  - X_i : fraction of people who are Black in precinct i
  - N_i: total number of voting age people
- We do not observe (but are interested in):
  - β_i^b: fraction of Black residents who voted
  - β_i^w: fraction of white residents who voted

Two older methods

1. Method of bounds (Duncan and Davis 1953): a deterministic solution
2. Goodman's regression (Goodman 1953): a statistical solution

Method of bounds

General idea: Use information about the proportion who voted and the proportion who are Black to bound the quantities of interest, β_i^b and β_i^w.

We can assemble the following 100% confidence intervals:

$$\beta_i^b \in \left[max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right), min\left(\frac{T_i}{X_i}, 1\right)\right]$$

$$\beta_i^w \in \left[max\left(0, \frac{T_i - X_i}{1 - X_i}\right), min\left(\frac{T_i}{1 - X_i}, 1\right)\right]$$

Method of bounds

Walking through the first 100% confidence interval:

$$\beta_i^b \in \left[max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right), min\left(\frac{T_i}{X_i}, 1\right)\right]$$

Note that:

- $\frac{T_i-(1-X_i)}{X_i}$ bounds β_i^b from below with the case that everyone who was white voted
- $\frac{T_i}{X_i}$ bounds β_i^b from above with the case that everyone who was black voted
- You get a lower bound of 0 if $T_i < 1 - X_i$ (fraction white $>$ fraction who voted)
- You get an upper bound of 1 if $T_i > X_i$ (fraction black $<$ fraction who voted)

Goodman's regression

Start with the following identity (*not an assumption*)

$$T_i = X_i \beta_i^b + (1 - X_i)\beta_i^w$$

- Recover B^b and B^w, district-wide turn-out by race, by regressing T_i on X_i and $(1 - X_i)$
- Constancy assumption: $Cov(\beta_i^b, X_i) = Cov(\beta_i^w, X_i) = 0$. In substantive terms, it means voters of a given race will vote the same way regardless of the racial composition of their local precincts.
- Violations lead to biased estimates including ones that fall out of the deterministic bounds.

Note the equation above can be rearranged:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right) \beta_i^b$$

Thus β_i^w can be expressed as a linear function of β_i^b

King (1997)'s approach

- Combine the two approaches: use statistical approach to extract information within bounds
- Using the fact that proportions must be in [0,1] and $\beta_i^w = \left(\frac{T_i}{1-X_i}\right) - \left(\frac{X_i}{1-X_i}\right)\beta_i^b$ we get Figure 0.1(b):

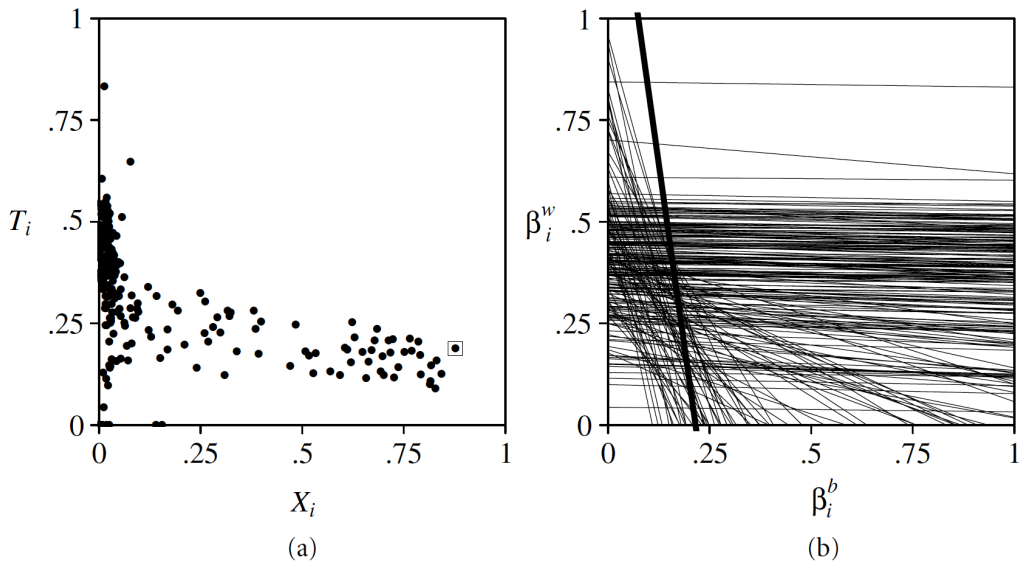
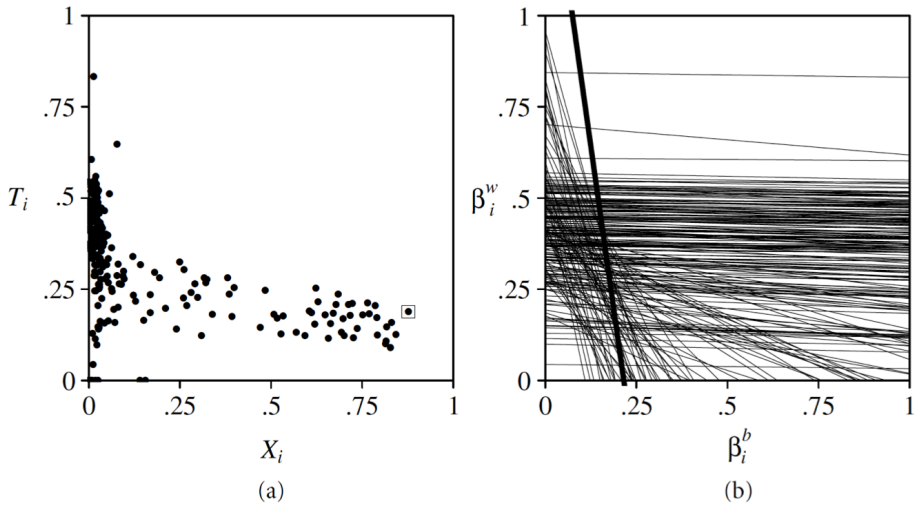


Figure 0.1. Two views of the same data: (a) a scatterplot of the observables, X_i by T_i; (b) this same information as a tomography plot of the quantities of interest, β_i^b by β_i^w. Each precinct i that appears as a point in (a) appears instead as a line (because of information lost due to aggregation) in (b). For example, precinct 52 appears as the dot with a little square around it in (a), and as the dark line in (b). The data are from King (1997: Figures 5.1, 5.5).

What does this buy us?



- Narrows search area from full square to line within the square
- We can then use information from other observations to learn more about β_i^b and β_i^w in a given precinct, given three assumptions:
  - Single cluster of β_i^b and β_i^w points
  - No spatial autocorrelation: $T_i|X_i$ are independent over observations
  - No (a priori) aggregation bias: X_i is mean independent of β_i^b and β_i^w

The model

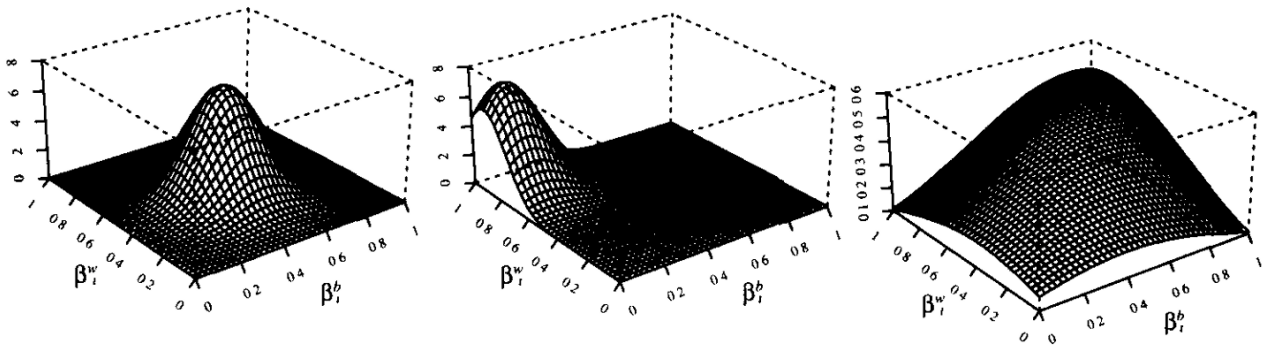


Figure 6.2 Truncated Bivariate Normal Distributions, $\mathrm{TN}(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma})$. Graph (a) is a distribution relatively unaffected by the truncation bounds. The graphs (b) and (c) are strongly truncated. Parameter values $\breve{\mathfrak{B}}$, $\breve{\mathfrak{B}}$, $\breve{\sigma}_b$, $\breve{\sigma}_w$, and $\breve{\rho}$ are indicated beneath each graph.

- Model β_i^w and β_i^b as generated by a truncated bivariate normal distribution conditional on X_i
  - *Note this is unimodal, hence the single cluster assumption*
  - *Will need to estimate the parameters of this density by forming the likelihood*

The model

- For a given precinct, we can get posterior distribution of the quantities of interest from conditioning on T_i
  - *We get a univariate distribution from the slice of the surface above the line that defines the relationship between β_i^w and β_i^b*

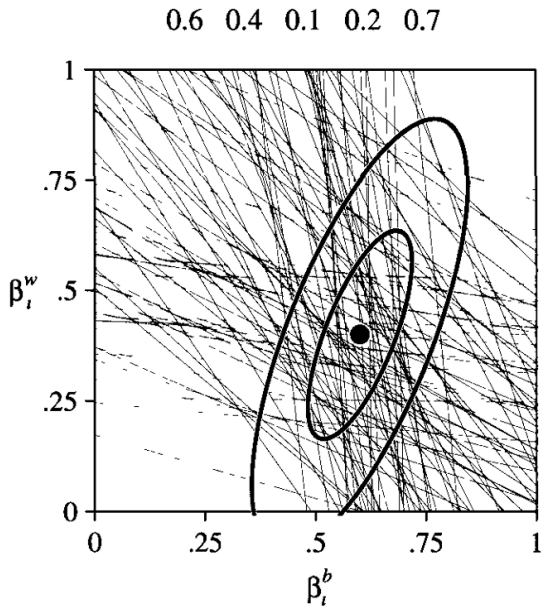


Figure 6.3 A Tomography Plot. Each coordinate represents a unique value of β_i^b and β_i^w. Each line traces out, for a particular combination of T_i and X_i, the algebraic expression in Equation 6.27. The contour lines on the figure represent the truncated bivariate normal distribution of β_i^b and β_i^w, with parameters $\breve{\mathfrak{B}}^b$, $\breve{\mathfrak{B}}^w$, $\breve{\sigma}_b$, $\breve{\sigma}_w$, and $\breve{\rho}$ indicated at the top of the graph. (See Figure 6.4 for a surface plot representation of these contours, which may be more familiar.)

The model

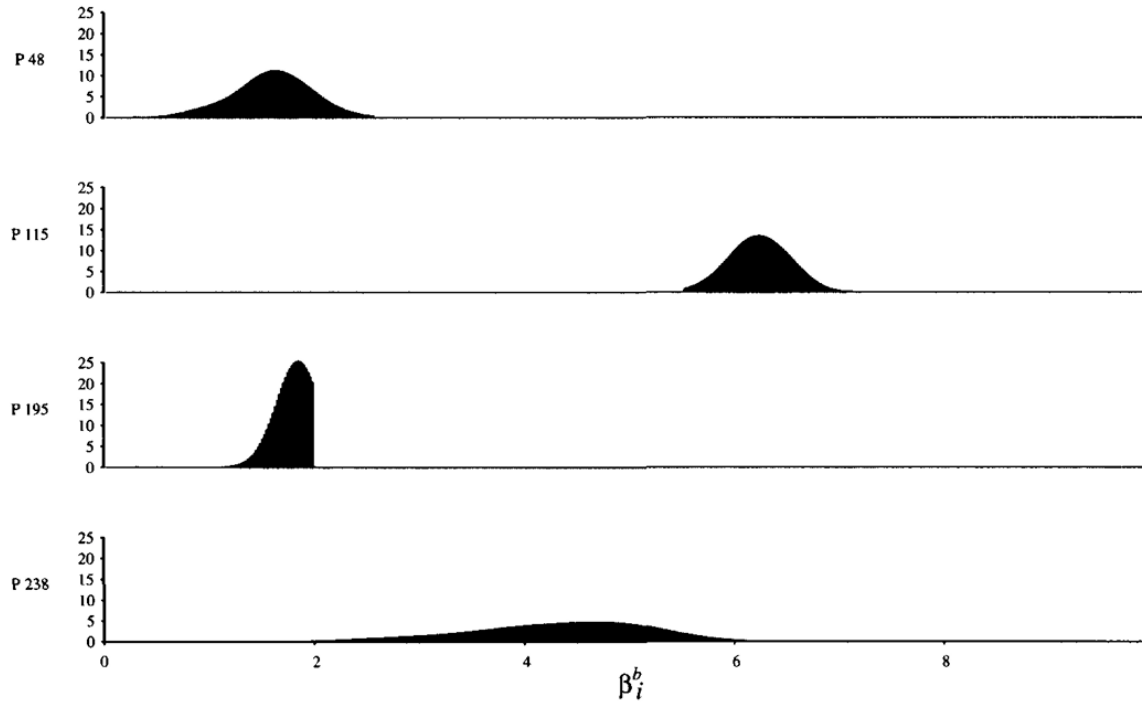


Figure 8.1 Posterior Distributions of Precinct Parameters β_i^b. Proportion of blacks voting in selected precincts from Pennsylvania's 8th senatorial district in 1990. The first two posteriors are symmetric; the third is strongly affected by its upper bound; and the last has a wide variance and is thus less informative. These figures are "density estimates" (smooth versions of histograms) drawn using the simulations $\tilde{\beta}_i^b$.

Parameterizing the truncated bivariate normal

To arrive at an algebraic expression for the probability distribution, start with the *untruncated* bivariate normal distribution

$$N(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) = (2\pi)^{-1} |\breve{\Sigma}|^{-1/2} exp\left[-\frac{1}{2} (\beta_i - \breve{\mathfrak{B}})' \breve{\Sigma}^{-1} (\beta_i - \breve{\mathfrak{B}}) \right]$$

We need to modify this to get a truncated density:

$$TN(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) = N(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}), \breve{\Sigma}) \frac{\mathbf{1}(\beta_i^b, \beta_i^w)}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}$$

- $\mathbf{1}(\beta_i^b, \beta_i^w)$ truncates: indicator that equals 1 if β_i^w and $\beta_i^b \in [0, 1]$
- $R(\breve{\mathfrak{B}}, \breve{\Sigma})$ keeps volume under truncated distribution to 1: dividing by volume under the untruncated distribution over the unit square

$$R(\breve{\mathfrak{B}}, \breve{\Sigma}) = \int_0^1 \int_0^1 N(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) d\beta^b d\beta^w$$

Deriving the Likelihood Function

1. The story of the model is that we learn things in order
   (a) (As in regression), everything is conditional on X_i, which means we learn it first.
   (b) Then the world draws β_i^b and β_i^w from a truncated normal, but we don't get to see them.
   (c) Finally, we learn T_i, which is computed via the accounting identity deterministically: $T_i = \beta_i^b X_i + \beta_i^w (1 - X_i)$.
2. The random variable is then T (given X), which is truncated bivariate normal
3. The five parameters of the truncated bivariate normal need to be estimated:

$$\breve{\psi} = \{\breve{\mathfrak{B}}^b, \breve{\mathfrak{B}}^w, \breve{\sigma}_b, \breve{\sigma}_w, \breve{\rho}\} = \{\breve{\mathfrak{B}}, \breve{\Sigma}\}$$

These are on the untruncated scale (and not quantities of interest) since:

$$\mathsf{TN}(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) = N(\beta_i^b, \beta_i^w | \breve{\mathfrak{B}}, \breve{\Sigma}) \frac{\mathbf{1}(\beta_i^b, \beta_i^w)}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}$$

Deriving the Likelihood Function

4. (From simulations of these parameters, we will compute quantities of interest: β_i^b, β_i^w.)
5. The likelihood:

$$
\begin{aligned}
L(\breve{\psi}|T) &\propto \prod_{X_i \in (0,1)} \mathrm{P}(T_i|\breve{\psi}) \\
&= \prod_{X_i \in (0,1)} \left(\frac{\text{What we observe}}{\text{What we could have observed}} \right) \\
&= \prod_{X_i \in (0,1)} \left(\frac{\text{Area above line segment}}{\text{Volume above square}} \right) \\
&= \prod_{X_i \in (0,1)} \left(\frac{\text{Area above line}}{\text{Volume above plane}} \right) \frac{\left(\frac{\text{Area above line segment}}{\text{Area above line}} \right)}{\left(\frac{\text{Volume above square}}{\text{Volume above plane}} \right)} \\
&= \prod_{X_i \in (0,1)} N(T_i|\mu_i, \sigma_i^2) \frac{S(\breve{\mathfrak{B}}, \breve{\Sigma})}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}
\end{aligned}
$$

Deriving the Likelihood Function

where

$$E(T_i|X_i) \equiv \mu_i = \breve{\mathfrak{B}}^b X_i + \breve{\mathfrak{B}}^w (1 - X_i)$$
$$V(T_i|X_i) \equiv \sigma_i^2 = (\breve{\sigma}_w^2) + (2\breve{\sigma}_{bw} - 2\breve{\sigma}_w^2)X_i + (\breve{\sigma}_b^2 + \breve{\sigma}_w^2 - 2\breve{\sigma}_{bw})X_i^2$$

$$S(\breve{\mathfrak{B}}, \breve{\Sigma}) = \int_{\max\left(0, \frac{T-(1-X_i)}{X_i}\right)}^{\min\left(1, \frac{T_i}{X_i}\right)} N\left(\beta^b \middle| \breve{\mathfrak{B}}^b + \frac{\omega_i}{\sigma_i}\epsilon_i,\ \breve{\sigma}_b^2 - \frac{\omega_i^2}{\sigma_i^2}\right) d\beta^b$$

How to Calculate Quantities of Interest

Use the knowledge that simulations for observation i must come from its tomography line:

(a) By the story of the model, if we know T_i, we learn the entire tomography line (since X_i is known ex ante).

(b) So we will condition on T_i to make a prediction from the tomography line.

(c) We could use rejection sampling (discard simulations of β_i^b, β_i^w that are not on the tomography line), but this would take forever.

(d) Alternative algorithm for drawing simulations of β_i^b and β_i^w.

i. Find the expression for $P(\beta_i^b | T_i, \breve{\psi})$ analytically, which is a particular truncated univariate normal (see King, 1997: Appendix C).
ii. Draw $\breve{\psi}$ from its posterior or sampling density (the same multivariate normal as always).
iii. Insert the simulation into $P(\beta_i^b | T_i, \breve{\psi})$ and draw out one simulated β_i^b.
iv. Compute β_i^w, if desired, deterministically from reformulated accounting identity:

$$\tilde{\beta}_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right) \tilde{\beta}_i^b$$

Model Extensions

Allowing parameters to vary as functions of measured covariate Z_i:

$$\breve{\mathfrak{B}}_i^b = \left[\phi_1(\breve{\sigma}_b^2 + 0.25) + 0.5\right] + (Z_i^b - \bar{Z}^b)\alpha^b$$

$$\breve{\mathfrak{B}}_i^w = \left[\phi_2(\breve{\sigma}_w^2 + 0.25) + 0.5\right] + (Z_i^w - \bar{Z}^w)\alpha^w$$

Relaxes mean independence assumption:

$$E(\beta_i^b | X_i, Z_i) = E(\beta_i^b | Z_i)$$

$$E(\beta_i^w | X_i, Z_i) = E(\beta_i^w | Z_i)$$

Imai and Khanna (2016)

- Alternate approach: combining individual-level data with aggregate data to find quantities of interest
- Useful when you want to understand the relationship between a covariate and a behaviour/outcome but you do not have both at the individual level (here, voting and ethnicity)
- Useful when you have:
  1. Individual-level data that includes information on the behaviour/outcome of interest along with some individual-level covariates (here, geocoded voter registration records)
  2. Aggregate demographic data that relates the individual-level covariates you have to the covariate you lack (here, Census Surname List, Census block-level racial composition data)

Basic Bayesian approach in Elliott et al. (2009)

Key assumption:

$$G_i \perp S_i | R_i$$

(once we know race, surname does not provide additional information on where person i lives)

By Bayes' rule:

$$P(R_i = r | S_i = s, G_i = g) = \frac{P(G_i = g | R_i = r, S_i = s) P(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} P(G_i = g | R_i = r') P(R_i = r' | S_i = s)}$$

- $P(R_i = r | S_i = s, G_i = g)$: Probability of race r given surname and geolocation
- $P(G_i = g | R_i = r)$: Probability of geolocation g given race (racial composition of geolocation)
  - $P(G_i = g | R_i = r) = \frac{P(R_i = r | G_i = g) P(G_i = g)}{\sum_{g' \in \mathcal{G}} P(R_i = r | G_i = g') P(G_i = g')}$
- $P(R_i = r | S_i = s)$: Probability of race given surname (racial composition of surnames)
- Denominator: $P(G_i = g | S_i = s)$

Imai and Khanna (2016) Extensions

Adding in additional covariates (X_i) - age and gender - requires the following conditional independence assumption:

$$\{G_i, X_i\} \perp S_i | R_i$$

(once we know a voter's race, surname does not provide additional information about geolocation or demographics)

$$P(R_i = r | S_i = s, G_i = g, X_i = x) =$$

$$\frac{P(G_i = g, X_i = x | R_i = r) P(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} P(G_i = g, X_i = x | R_i = r') P(R_i = r' | S_i = s)}$$

Imai and Khanna (2016) Extensions

Adding in party registration requires the following two assumptions:

$$\{G_i, P_i, X_i\} \perp S_i | R_i$$

$$\{G_i, X_i\} \perp P_i | R_i$$

$$P(R_i = r | S_i = s, G_i = g, P_i = p, X_i = x) =$$

$$\frac{P(G_i = g, X_i = x | R_i = r) P(P_i = p | R_i = r) P(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} P(G_i = g, X_i = x | R_i = r') P(P_i = p | R_i = r) P(R_i = r' | S_i = s)}$$

Alternative strategy

Assumption:

$$\{X_i, P_i\} \perp S_i | G_i, R_i$$

$$P(R_i = r | S_i = s, G_i = g, P_i = p, X_i = x) =$$

$$\frac{P(P_i = p, X_i = x | G_i = g, R_i = r) P(G_i = g | R_i = r) P(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} P(P_i = p, X_i = x | G_i = g, R_i = r') P(G_i = g | R_i = r) P(R_i = r' | S_i = s)}$$

where:

$$P(P_i = p, X_i = x | G_i = g, R_i = r) =$$
$$P(P_i = p | X_i = x, G_i = g, R_i = r) P(X_i = x | G_i = g, R_i = r)$$

ROC Curves

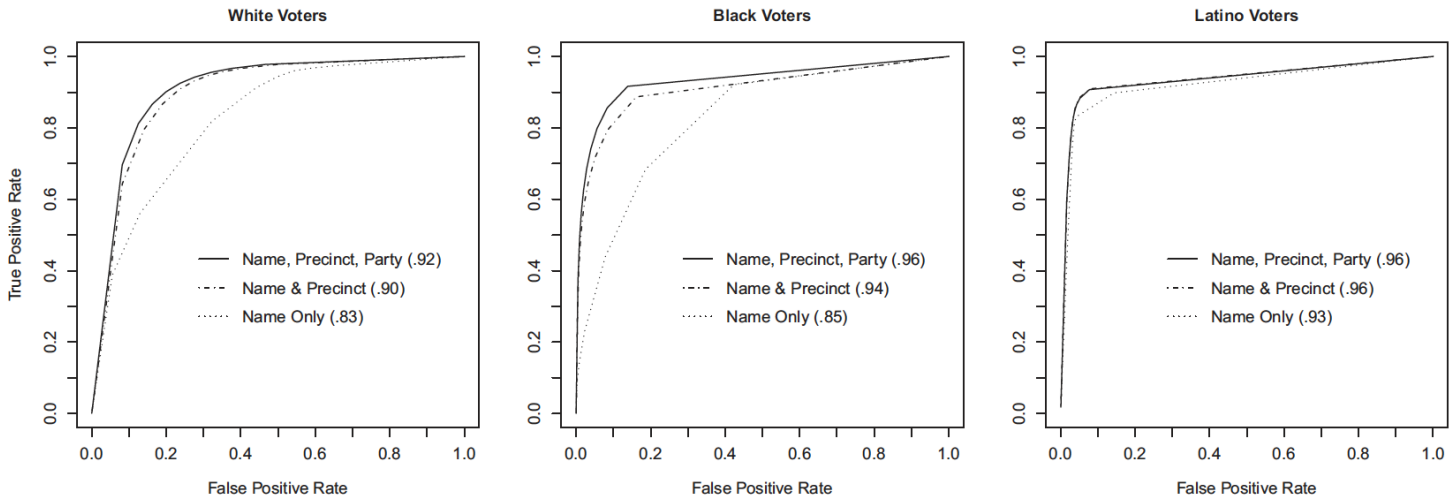


Fig. 1 ROC curves for the proposed race prediction methods. ROC curves plot true positive rate (vertical axis) against false positive rate (horizontal axis) for all possible thresholds used for classification. The area under the ROC curves, given in the legend, summarizes the overall classification success. Among White and Black voters, using voter precinct (denoted as "Precinct") in addition to surname ("Name") substantially improves classification accuracy. Adding voter party registration ("Party") results in further improvements. Among Latino voters, surname alone yields a high success rate and adding other information produces minor improvements.

Comparing approaches

Table 2 Bias and RMSE of predicted turnout by race across 8,828 precincts and 25 congressional districts in Florida

	Goodman's regression		*King's EI*		*Name-only prediction*		*Bayesian prediction*	
	Bias	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
Precincts								
Whites	0.003	0.069	0.041	0.062	−0.003	0.015	−0.003	0.012
Blacks	−0.102	0.162	−0.133	0.217	−0.009	0.043	−0.007	0.039
Latinos	−0.114	0.251	−0.163	0.250	0.016	0.042	0.011	0.035
Asians	0.017	0.713	−0.470	0.550	0.041	0.116	0.040	0.111
Others	−0.214	0.499	−0.338	0.450	0.068	0.109	0.048	0.094
Districts								
Whites	0.008	0.037	0.047	0.058	−0.007	0.012	−0.001	0.004
Blacks	−0.147	0.197	−0.215	0.267	0.009	0.020	−0.006	0.010
Latinos	−0.272	0.463	−0.300	0.354	0.045	0.052	0.017	0.021
Asians	0.072	0.808	−0.459	0.530	0.055	0.058	0.043	0.046
Others	−0.229	0.527	−0.342	0.448	0.073	0.078	0.042	0.053

Notes: Goodman's regression, King's EI, name-only prediction (based on the Census Surname List), and our proposed Bayesian prediction method. Although Goodman's regression and King's EI use precinct-level turnout and racial composition data only, the proposed Bayesian methodology uses the name, residence location, and party registration of voters. Precinct-level bias and RMSE are weighted by the number of voters for each precinct. Generally, the proposed Bayesian method performs best, though the name-only prediction also yields a reasonable performance.