Supplementary Appendix: Imai, Kosuke and Kabir Kahnna. (2016). “Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records.” *Political Analysis*

doi: 10.1093/pan/mpw001

A Appendix

A.1 Name Merging Procedure

In order to determine the prior probability $\Pr(R_i = r \mid S_i = s)$, we use the Census Surname List and Spanish Surname List. For any surname S_i that appears on the Spanish Surname List, we set the prior probability to 1 for Latinos and 0 for every other racial group. For the remaining surnames, we use the more comprehensive Census Surname List. However, not all of the surnames in the voter files appear in the Census Surname List. This sometimes occurs because surnames are “double barreled,” i.e. two names separated by a hyphen or space. We take the following steps in order to merge as many surnames as possible with the Census list. Each step only applies to names that were not matched in a previous step.

1. Capitalize all surnames and attempt to match with Census list.
2. Remove spaces from surnames and match again.
3. Split double-barreled names apart, and attempt to match first half of name.
4. Split double-barreled names apart, and attempt to match second half of name.
5. Impute priors for remaining names using overall U.S. race distribution.

A.2 Expectation-Maximization Algorithm

Define the following model of partisanship,

$$\psi^p_{R_i G_i X_i} = \Pr(P_i = p \mid G_i, R_i, X_i) \tag{11}$$

This model may be non-parametric, as done in this paper, or parametric (e.g., logistic regression). For the notational simplicity, define $\phi^r_{G_i X_i S_i} = \Pr(R_i = r \mid G_i, X_i, S_i)$, which is observed. Note that R_i is missing data. Then, the complete-data log-likelihood is,

$$\sum_{i=1}^{n} \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}} \mathbf{1}\{P_i = p, R_i = r\} \left(\log \psi^p_{r G_i X_i} + \log \phi^r_{G_i X_i S_i}\right) \tag{12}$$

Then, in the E-step, we take the expectation of the above complete-data log-likelihood function conditional on the observed data (i.e., the Q-function),

$$\sum_{i=1}^{n} \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}} \pi^{r}_{pG_iX_iS_i} \mathbf{1}\{P_i = p\} \left(\log \psi^{p}_{rG_iX_i} + \log \phi^{r}_{G_iX_iS_i} \right) \tag{13}$$

where

$$\begin{aligned} \pi^{r}_{pG_iX_iS_i} &= \Pr(R_i = r \mid P_i = p, G_i, X_i, S_i) \\ &= \frac{\psi^{p}_{rG_iX_i} P(X_i \mid R_i = r, G_i) P(G_i \mid R_i = r) \Pr(R_i = r \mid S_i)}{\sum_{r' \in \mathcal{R}} \psi^{p}_{r'G_iX_i} P(X_i \mid R_i = r', G_i) \Pr(G_i \mid R_i = r') \Pr(R_i = r' \mid S_i)} \end{aligned} \tag{14}$$

The M-step maximizes the Q-function with respect to the model ψ^{p}_{rgx}. In the non-parametric model as done in our empirical application, we update ψ^{p}_{rgx} as,

$$\hat{\psi}^{p}_{rgx} = \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i = g, X_i = x\} \pi^{r}_{pgxS_i} \mathbf{1}\{P_i = p\}}{\sum_{i=1}^{n} \mathbf{1}\{G_i = g, X_i = x\} \pi^{r}_{pgxS_i}}. \tag{15}$$

We repeat the E-step and M-step until convergence. Finally, equation (14) gives the predicted probability of individual race based on this methodology.

A.3 Probing the Conditional Independence Assumption

We probe the conditional independence assumption in equation (1) by comparing $P(S_i, G_i)$ against the product of the marginals $P(S_i) \times P(G_i)$. These two quantities should be equal to each other within a racial category under the conditional independence assumption. We compare the distribution of absolute residuals from this comparison with and without conditioning on race. Figure 2 presents the quantile-quantile plot. Conditioning on race substantially decreases absolute residuals for each racial group.

A.4 Comparing Precinct-Level Data from Census and Voter File

We examine whether the Census and voter file data yield comparable estimates of racial composition by precinct. One possible reason why the demographic information does not improve the performance of our methods is the potential discrepancy between the Census and voter file data. We plot Census and voter file estimates of race by precinct against each other in Figure 3, separately for males and females. With the exception of Asians, the two estimates are highly consistent with one another, suggesting that measurement error is not a problem at the precinct level.

We also reran our race predictions using voter file, rather than Census, estimates of age, sex, and precinct conditional on race. Doing so does not substantially reduce error rates, as shown in Table 3, suggesting that data issues do not explain the ineffectiveness of demographics in predicting race, over and above surname, geolocation, and party.

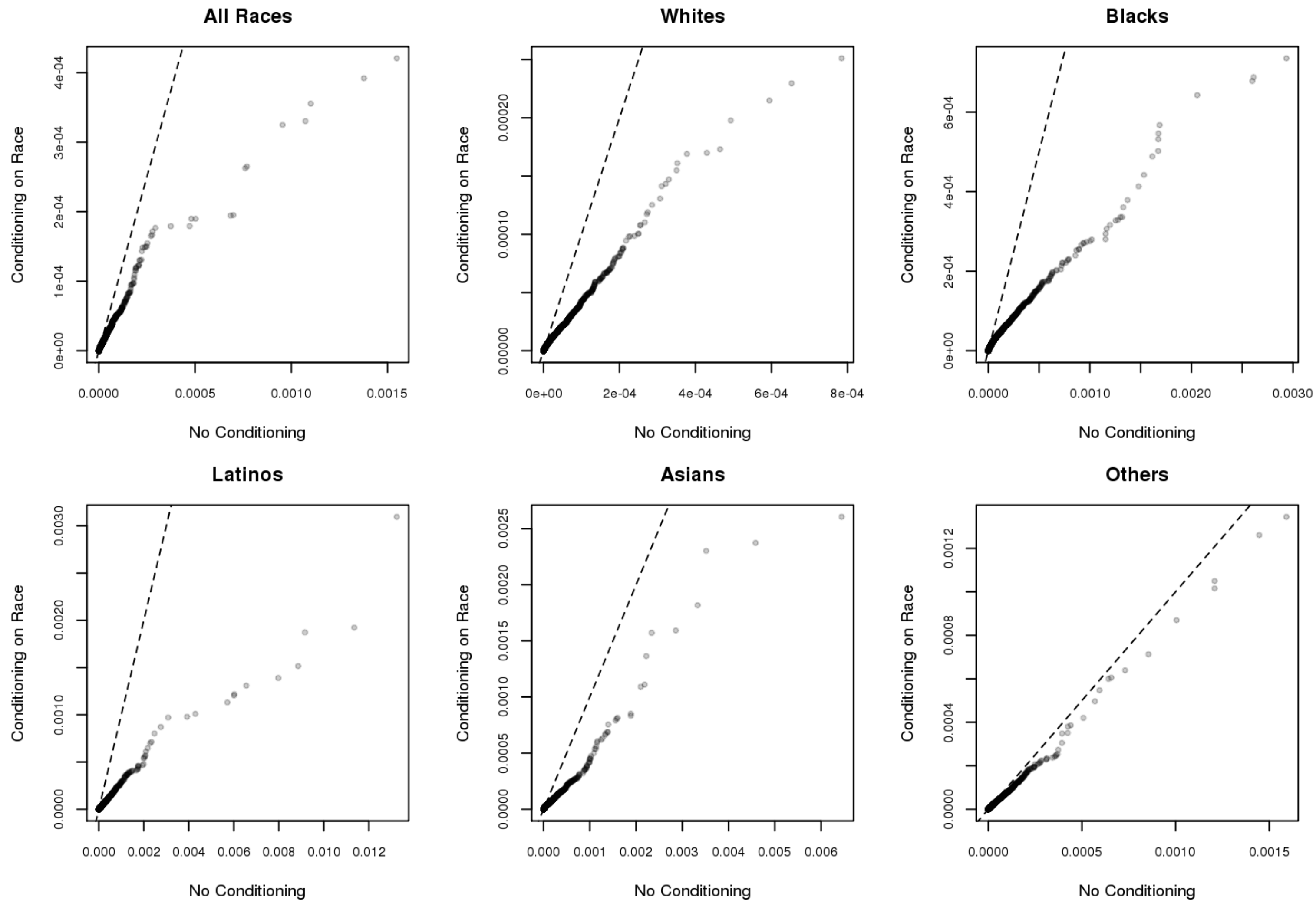

Figure 2: Quantile-quantile Plots of Distributions of Surname-by-County Absolute Residuals. These residuals represent the differences between $P(S_i, G_i)$ and $P(S_i)P(G_i)$, which are estimated from the data using the corresponding sample proportions. Plots compare the absolute residuals with (vertical axis) and without (horizontal axis) conditioning on race. Conditioning on race generally reduces the size of absolute residuals, suggesting that the conditional independence assumption may be appropriate.

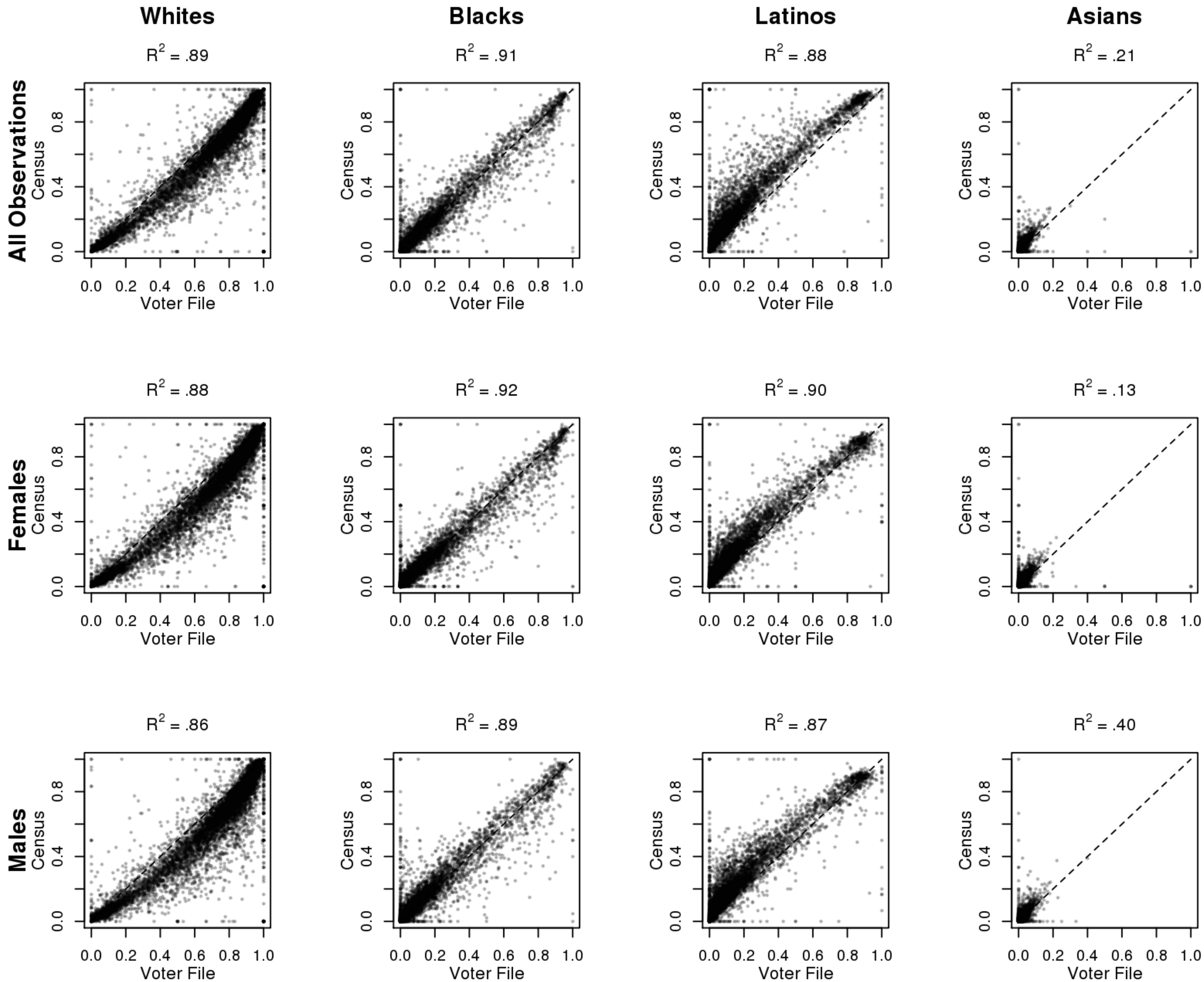

Figure 3: Comparison of Precinct-level Estimates of Race based on Census Data (vertical axis) and Voter File (horizontal axis) in Florida. The two estimates are relatively similar.

		Name, Precinct, Demographics		Name, Precinct, Party, Demographics	
		Census	Voter File	Census	Voter File
Overall	error rate	.159	.148	.151	.140
White (66%)	false negative	.056	.059	.059	.062
	false positive	.305	.267	.269	.231
Black (14%)	false negative	.394	.335	.305	.247
	false positive	.024	.028	.028	.032
Latino (14%)	false negative	.162	.139	.170	.147
	false positive	.037	.036	.036	.035
Asian (2%)	false negative	.571	.468	.571	.466
	false positive	.007	.006	.007	.006

Table 3: The Accuracy of Race Predictions Using the Aggregate Demographic Data in Each Precinct Based on Either the Census or Voter File Data. The results show that the use voter file does not substantially improve the predictions, thereby indicating that discrepancies between the Census and voter file data are unlikely to account for the ineffectiveness of aggregate demographic characteristics in improving the prediction of individual race.

A.5 Additional Empirical Results

	Overall	Whites		Blacks		Latinos		Asians	
Predictors	Error Rate	FN	FP	FN	FP	FN	FP	FN	FP
Name Only	.215	.047	.523	.839	.011	.193	.037	.540	.006
with survey									
Name, Precinct	.158	.060	.294	.381	.027	.150	.039	.519	.007
Name, Precinct, Demo	.159	.056	.305	.394	.024	.162	.037	.571	.007
Name, Precinct, PID	.151	.065	.257	.290	.033	.158	.038	.520	.007
Name, Precinct, Demo, PID	.151	.059	.269	.305	.028	.170	.036	.571	.007
without survey									
Name, Precinct, PID	.155	.059	.285	.362	.026	.148	.039	.516	.007
Name, Precinct, Demo, PID	.157	.058	.291	.370	.025	.160	.038	.563	.008
with survey									
Name, Block	.152	.059	.266	.320	.026	.155	.038	.533	.007
Name, Block, Demo	.186	.068	.247	.290	.029	.210	.039	.577	.009
Name, Block, PID	.145	.061	.237	.249	.029	.162	.037	.532	.007
Name, Block, Demo, PID	.180	.069	.229	.250	.030	.212	.039	.576	.010
without survey									
Name, Block, PID	.151	.061	.255	.301	.026	.153	.038	.524	.008
Name, Block, Demo, PID	.189	.074	.238	.277	.032	.213	.040	.570	.011

Table 4: Empirical Validation of Race Classification Using the Florida Registration Records. The table displays the overall classification error rate, and false negative (FN) and false positive (FP) rates for White, Black, Latino, and Asian voters using our proposed prediction method (with and without survey data about the conditional probability of party ID given race). We classify each voter to the racial category with the highest predicted probability. Each row corresponds to predictions based on different sets of information. We start with Census Bureau's surname list and then add information about the voter's precinct, demographics (age and gender), and party registration (PID). $N = 9,247,810$.

	Goodman's multivariate regression		Name-only classification		Bayesian classification	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Precincts						
Whites	.005	.071	−.003	.016	−.005	.016
Blacks	−.077	.147	−.006	.075	−.002	.075
Latinos	−.099	.236	.007	.034	.004	.038
Asians	.219	.683	−.008	.135	−.006	.133
Others	−.030	.479	−.012	.272	−.029	.253
Districts						
Whites	.011	.040	−.006	.011	−.003	.005
Blacks	−.110	.174	.002	.012	−.004	.011
Latinos	−.228	.413	.017	.021	.005	.012
Asians	.264	.763	−.001	.021	−.003	.020
Others	−.009	.499	−.011	.048	−.060	.078

Table 5: Additional Results for Bias and Root Mean Squared Error (RMSE) of Predicted Turnout by Race across 8,828 Precincts and 25 Congressional Districts in Florida. Goodman's multivariate regression, name-only classifications (based on the Census surname list), and our proposed Bayesian classifications. Precinct-level bias and RMSE are weighted by the number of voters in each precinct.

	Goodman's regression		King's EI		Name-only prediction		Bayesian prediction	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Whites	−.017	.065	.015	.022	−.002	.008	−.002	.007
Blacks	−.069	.130	−.071	.178	.005	.067	.003	.064
Latinos	−.259	.486	−.250	.364	.042	.092	.018	.074
Asians	−.192	.808	−.545	.612	.077	.167	.049	.151
Others	−.220	.580	−.266	.467	.056	.113	.028	.094

Table 6: Bias and Root Mean Squared Error (RMSE) of Predicted Turnout by Race across 2,567 Racially Homogenous Precincts in Florida. We evaluate Goodman's regression, King's EI, name-only prediction, and our proposed Bayesian prediction method. The Bayesian method outperforms the other methods, Goodman's regression and the EI in particular. While Goodman's regression and King's EI use only precinct-level turnout and racial composition, the proposed Bayesian methodology uses name, residence location, and party registration of voters. Bias and RMSE are weighted by number of voters in each precinct.

	Whites	Blacks	Latinos	Asians	Others
Name Only					
False Negative	.720	.717	.666	.645	.639
False Positive	.696	.723	.682	.650	.657
Difference	.024	$-.006$	$-.016$	$-.006$	$-.018$
Name, Precinct, and Party					
False Negative	.698	.714	.670	.646	.640
False Positive	.691	.671	.667	.648	.600
Difference	.007	.042	.003	$-.002$	.040

Table 7: Turnout among False Negatives and False Positives. The table displays the actual turnout rate among voters that we misclassify based on both the name-only and the Bayesian prediction based on name, precinct, and party registration. We calculate the turnout rate among both false negatives and false positives, as well as the difference between the two. We find that the differences are small on average, indicating that turnout is independent of classification error.