British Journal of Political Science (2025), **55**, e110, 1–21
doi:10.1017/S0007123424000772

British Journal of
Political Science

ARTICLE

Measuring Descriptive Representation at Scale: Methods for Predicting the Race and Ethnicity of Public Officials

Diana Da In Lee[1] and Yamil Ricardo Velez[2]

[1]Center for the Study of Democratic Politics, Princeton University, Princeton, NJ, USA and [2]Department of Political Science, Columbia University, New York, NY, USA
Corresponding author: Diana Da In Lee; Email: dl9388@princeton.edu

(Received 29 August 2023; revised 25 June 2024; accepted 22 October 2024)

Abstract

Ethnicity and race are vital for understanding representation, yet individual-level data are often unavailable. Recent methodological advances have allowed researchers to impute racial and ethnic classifications based on publicly available information, but predictions vary in their accuracy and can introduce statistical biases in downstream analyses. We provide an overview of common estimation methods, including Bayesian approaches and machine learning techniques that use names or images as inputs. We propose and test a hybrid approach that combines surname-based Bayesian estimation with the use of publicly available images in a convolutional neural network. We find that the proposed approach not only reduces statistical bias in downstream analyses but also improves accuracy in a sample of over 16,000 local elected officials. We conclude with a discussion of caveats and describe settings where the hybrid approach is especially suitable.

Keywords: ethnoracial prediction; Bayesian imputation; machine learning; neural networks; descriptive representation

Political science has devoted an increasing amount of attention to the study of minority groups, with topics ranging from the behavior of voters (Tate 1991; Gay 2001; Barreto, Segura and Woods 2004; Barreto 2007; Hajnal 2010) to candidates (Shah 2014; Juenke and Shah 2015; Fraga 2018; Atsusaka 2021). Previously, a paucity of data on the racial and ethnic identities of voters or elected officials necessitated the use of aggregate data to infer individual-level behavior. But with recent statistical innovations, researchers are able to study individual-level, rather than group-level, political outcomes by imputing group categories. Namely, Bayesian Improved Surname Geocoding (BISG), which uses Bayes' Rule to combine geographic and surname information, has become a popular approach due to its computational efficiency. The development of various machine learning models for group classification has further facilitated the growth of research on topics such as racial disparities in turnout (Fraga, 2016b, 2018), public opinion (Menshikova and Tubergen 2022), campaign contributions (Grumbach and Sahn 2020), and descriptive representation (Enos 2016; Schwemmer and Jungkunz 2019).

However, an analysis of various prediction methods reveals unique limitations. First, under certain conditions, Bayesian predictions that rely on geographic information can lead to biased estimates in downstream analyses (Argyle and Barber 2024). Studies using BISG to generate measures of outcome variables, for example, can produce systematic measurement errors if prediction errors are correlated with covariates in a downstream regression model. When independent variables are geographic in nature – as they often are in studies of descriptive representation and minority turnout – biased and overconfident inferences are a possible consequence. Second, an alternative Bayesian approach that solely relies on surnames tends to generate imbalanced accuracy, with particularly high error rates among Blacks. Third, machine learning techniques, particularly image-based models, improve accuracy for racial categories but have difficulty distinguishing Hispanics from non-Hispanic whites. We propose an alternative approach that improves upon existing classification methods while sidestepping issues that arise due to the use of geography-dependent prediction methods. Specifically, we combine surname-based Bayesian estimation with image-based predictions generated by a convolutional neural network model. Using predicted probabilities from both methods as features, the proposed classification approach not only mitigates geography-driven biases but also yields a higher out-of-sample accuracy rate than alternatives.

An empirical validation study using data on over 16,000 US local elected officials shows that the proposed method results in lower error rates than other techniques. We also find that the proposed method has an overall classification success rate of 0.93 or greater across the four main ethnoracial categories in the Census – the strongest performance among all prediction methods evaluated. Furthermore, we find that our method reduces bias in settings involving geographic predictors.

Our approach contributes to the growing literature on ethnoracial prediction that has revealed impressive accuracy for machine learning approaches. Comparing the performance of BISG against a number of machine learning models, Decter-Frain (2022) finds that "even a simple switch from BISG to multinomial logistic regression can substantially improve individual predictions" (p. 4, see also Wong et al. 2020; Cheng et al. 2023). Similarly, Argyle and Barber (2024) incorporate BISG predictions into a random forest model to reduce misclassification bias. Our proposed method contributes to this increasing use of machine learning tools by incorporating rich information embedded in images. We conclude with a discussion of future directions, ethical and practical considerations, and recommendations for applied researchers. We also provide step-by-step instructions in online materials to facilitate its use.

An Overview of Existing Methods

Possessing individual-level race and ethnicity information is crucial for modeling turnout and vote choice, among other outcomes. However, publicly available data sets often lack this information, and scholars have to rely on a variety of prediction tools to construct proxies. Table 1 highlights the diversity in existing methods commonly adopted in the literature in terms of both statistical and practical approaches.

We first provide an overview of existing methods, which can largely be divided into two statistical approaches: Bayesian imputation and machine-learning techniques. We then introduce our proposed method, highlighted in the bottom row of Table 1. We demonstrate its use via empirical application and detail its benefits and limitations in comparison to other established methods.

Prior to proceeding, we note the fundamental differences between the two statistical approaches. Bayesian imputation relies on unsupervised learning, and thus, does not require labels or classifications. By contrast, machine-learning approaches are supervised methods that require a sufficient amount of labelled data.[1]

The Bayesian Approach

The Bayesian approach has been the most popular method due to its parsimony and usability, especially with the introduction of the R package *wru* (Imai and Khanna 2016). To quantify its

[1] This fundamental distinction implies that the assumptions each method must satisfy to generate accurate predictions differ substantially, with the former relying on conditional independence assumptions following Bayes' Rule and the latter relying on the size and representativeness of labelled data.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

Table 1. Assessment of existing prediction methods and the proposed hybrid approach. The 'hybrid method' is used as shorthand to describe the proposed prediction method introduced in this paper.

	Data & Method		Quantity of Interest			Performance/Efficiency					Academic Usage[a]
	Data	Statistical Approach	Concept of Race	Target	Racial Categories	Accuracy[b]	Downstream Bias[c]	Time	Cost	Applicability	
BSO	Surname	Bayes' Rule	Self-Reported	Any	5	4	Modest	Fast	Free	US-centric	24%
BISG	Surname + Geolocation	Bayes' Rule	Self-Reported	Any	5	3	Geography-Driven	Fast	Free	US-centric	36%
fBISG	Full name + Geolocation	Bayes' Rule	Self-Reported	Any	5	2	Geography-Driven	Fast	Free	US-centric	–
Text-based ML	Full name	Supervised Learning	Self-Reported	Any	5+	6	Modest	Fast	Free	Global	27%
Image-based ML	Public Image	Supervised Learning	Perceived	Public Figures	5+	4	Image-Driven	Moderate	Free	Global	5%
Hybrid	**Surname + Public Image**	**Supervised Learning**	**Perceived**	**Public Figures**	**5+**	**1**	**Modest**	**Moderate**	**Free**[d]	**Global**	**–**

[a] Based on a literature review of published articles in social science. See Appendix B for details.
[b] 1 = High; 6 = Low. Based on the overall classification rate identified in Section 'Applications'.
[c] Refers to biases when predictions are used as model outcomes. See Section 'The Bayesian Approach' for details.
[d] Costs may be incurred from data labelling. See Section 'Practical Considerations' for further discussions.

popularity, we reviewed over ninety recently published articles in the social sciences that utilized racial prediction methods and found that 60 percent of the articles implemented either Bayesian Surname-Only (BSO) or Bayesian Improved Surname Geocoding (BISG) methods.[2]

BSO infers race using the racial distribution of surnames from American Decennial Census Surname Files. While this method performs well for Hispanics/Latinos because surnames are a strong predictor of ethnic self-identification, it is less capable of distinguishing Blacks from non-Hispanic whites. For example, using Florida voter registration records, Imai and Khanna (2016) find false negative rates to be above 0.83 among Blacks and false positive rates to be above 0.52 among non-Hispanic whites for the surname-only Bayesian prediction.

BISG improves upon BSO by combining information from the Census surname list with data on local area demographics (Elliott et al. 2009; Imai and Khanna 2016). In practice, the Bayesian approach provides posterior probabilities across ethnoracial categories recorded by the Census – e.g., Asian, Black, Hispanic/Latino, white, and Other – for each individual.[3] BISG produces fairly accurate predictions of individuals' self-reported ethnoracial identification, although accuracy varies by group. For example, Elliott et al. (2009) estimate a weighted average correlation coefficient of 0.76 between predictions and self-reported data, which is 41 percent more accurate than BSO. Using the Florida voter file, Imai and Khanna (2016) report a similar accuracy rate, as well as improved estimates of aggregate-level turnout rates for different racial and ethnic groups, when compared to standard statistical methods.

In recent years, several approaches have aimed to improve the standard BISG methodology. Namely, a fully Bayesian Improved Surname Geocoding (fBISG) improves upon BISG by both accounting for Census measurement error and incorporating additional data on first and middle names taken from voter files of six Southern states that offer data on racial self-identification (Imai, Olivella and Rosenman 2022). Similarly, zipWRU extends BISG by drawing on ZIP code Tabulation Area (ZCTA) (DeLuca and Curiel 2022), as opposed to the larger units of aggregation that are commonly used in this research. Although the Bayesian approach is technically applicable to contexts beyond the USA, its adoption has surged within the USA, propelled by the availability of Census data as well as the convenience of R packages such as *wru*.

Despite its utility, the risks of systematic measurement error, due to the use of geographically-informed predictions, might become more prevalent as research on race and politics expands to answer a wider set of questions. To provide a running example, consider a typical question from the descriptive representation literature such as how many minority legislators are elected to a given office. In the USA, minority candidates face challenges in running for office, or getting elected, even when their districts have large co-ethnic populations (Swain 1995; Canon and Posner 1999). For example, prior to 2014, Ferguson, MO had never elected a Black mayor, and only two Black members had ever served on the council despite the fact that 67 percent of the city is African American (Eligon 2015). But if a researcher were to use BISG to estimate the racial diversity of the local government, it would disproportionately assign a high predicted probability to the Black category for *all* elected officials on the council due to the city's large Black population. In fact, BISG classifies 36 percent of candidates who ran for local elections in Ferguson as Black, when the true proportion is only 5 percent.[4] This

[2] See Appendix B for a detailed discussion of literature review.

[3] The Other category typically groups three categories used in the Census data: American Indian/Alaska Native (AIAN), Some Other Race, and Two or More Races. While it is theoretically possible to separate these groups, current practices often fail to do so. The *wru* package – the most commonly used BISG estimation package – collapses these groups as such because they jointly constitute a small share of the population. Failure to generate distinct predictions for these groups has several problematic implications. First, ignoring the populations that are exceptionally difficult to predict with BISG continues to stall research on inequality and political participation among these groups. Second, the inability to predict multi-racial individuals poses additional concerns, given that an increasing number of legislators and individuals in the United States identify as multi-racial (Lemi 2021). However, a newer R package developed by McCartan et al. (2023) separates AIAN from Other.

[4] We describe the Ferguson case further in Appendix C.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

example demonstrates how geography-informed predictions can understate the extent of racial inequality in local politics.

This issue arises in research that analyzes ethnoracial predictions as a function of geographic factors. Of course, using a predicted measure as a dependent variable would not introduce bias if the measurement error only adds stochastic variation (Berry and Feldman 1985; Hausman 2001). However, when measurement error is correlated with independent variables, biased estimates are a consequence. First, non-stochastic measurement error occurs when the observed outcome variable systematically measures some other variable(s) in addition to the true outcome, causing the observed outcome to vary on the basis of the other variable(s). If BISG is used to predict outcomes, this is a possible issue, as this outcome measure will also capture systematic variation due to the ethnoracial composition of each location. Second, if a downstream regression model includes a predictor that covaries with the measurement error, this can lead to biased estimates even after adjusting for name and location (Argyle and Barber 2024; McCartan et al. 2023).[5]

Applications involving geographic variables are common in political science research that examines the relationship between candidate ethnicity and district racial composition, or between district racial composition and minority turnout. For example, Hankinson and Magazinnik (2023) employ BISG to estimate the race of California city councillors, Grumbach and Sahn (2020) and Alvarez, Katz and Kim (2020) use BISG for political donors, Abott and Magazinnik (2020) and Kogan, Lavertu and Peskowitz (2021) for school board members, and Conroy and Green (2020) and Shah and Davis (2017) for prospective candidates – all of which include geography-specific independent variables in downstream analyses.

In Online Appendix D, we provide a simulation result that documents the potential magnitude and direction of bias when BISG is inappropriately used (that is, in cases where geographic variables are included as independent variables and there is systematic measurement error). Specifically, we simulate a downstream model where BISG predictions of legislator ethnorace are regressed on a geographic predictor. We find that the bias, while varying in magnitude, is generally upward when BISG is used and downward when surname-only predictions are used.

Machine Learning Approaches

Political scientists and social scientists more broadly have increasingly adopted machine-learning algorithms. Based on our literature review, text-based machine learning methods constitute 27 percent of published articles examined, which is comparable to the rate of BSO's use. Namely, many empirical papers employ Long Short-Term Memory Models (LSTM) that exploit the sequence of letters in a name. Trained on a diverse set of training data such as voter files and Wikipedia, LSTM has shown higher accuracy than conventional BISG, particularly among Blacks and Asians (Sood and Laohaprapanon 2018). LSTM has been implemented to examine racial disparities in academic institutions (Marschke et al. 2018), public opinion among Twitter users (Menshikova and Tubergen 2022), and vulnerability to natural disasters (Messager et al. 2021).

Automatic estimation of demographic attributes (for example, age, gender, and race) from images has become another topic of growing interest and controversy, with many potential applications. Our literature review illustrates this growing demand, as we find that 5 percent of articles, all published since 2019, have leveraged an image-based machine learning algorithm. Current research relies primarily on convolutional neural network (CNN) models, which have demonstrated strong performance on challenging prediction tasks. In essence, these deep learning algorithms take an input image, assign importance to various latent features, and classify demographic attributes based on this information. CNNs capture spatial dependencies in an image through the application of filters of varying levels of complexity that are learned by the

[5] Furthermore, McCartan et al. (2023) formally show that using BISG probabilities as weights or thresholding the BISG probabilities still leads to biased estimates.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

algorithm. As examples of how CNNs have been used in recent work, Schwemmer and Jungkunz (2019) investigate descriptive representation in digital platforms by annotating the gender and ethnoracial identities of TED Talk speakers using Kairos diversity recognition software, and Lyu et al. (2021) employ DeepFace API to predict the ethnic and racial identities of Twitter users when examining their opinion toward racially motivated hate crimes in the USA.

However, one of the shortcomings of the image-based approach is that it is less able to distinguish between Hispanics/Latinos and non-Hispanic whites, presumably because it relies on skin tone as a feature to predict ethnoracial identity, and Latinos are a multi-racial community. This sensitivity is, in part, due to the quality of the data set that the model is trained on. In the realm of deep learning, having a substantial amount of data in terms of both quantity as well as ethnoracial diversity is crucial for effectively training CNNs. However, the availability of such data sets has been lacking due to the challenges involved in annotating such data (Abdulwahid 2023). Because ethnorace cannot be quantitatively measured, human annotators must establish ground truth values (Sulaiman and Kocher 2022).

A growing number of training databases have been introduced to close this gap. For example, the FairFace data set is comprised of over 108,000 images balanced across seven different ethnoracial groups (Karkkainen and Joo 2021). Other similar data sets include VGG-Face2, which contains 2.6 million images with annotations for six different ethnoracial groups, and the VGG-Face2 Mivia Ethnicity Recognition (VMER) data set, which improves upon VGG-Face2 by obtaining ethnoracial classifications from three independent annotators belonging to different groups (Greco et al. 2020).

Supervised models trained on these data sets with a more granular set of ethnoracial categories could improve estimation methods. Aggregating ethnoracial groups with distinct cultures can limit research on categories not covered by existing classification schemes that rely on the Census (see footnote 3). For example, Middle Eastern and North African (MENA) Americans are categorized as white in the Census, but racial self-identification among MENA individuals typically does not reflect this (Maghbouleh, Schachter and Flores 2022). Similar research on Asian American identity also finds distinct political behaviors and self-identification patterns between South and East Asian Americans (Yamashita 2022). While conventional Bayesian approaches use five ethnoracial categories, certain pre-trained models (like FairFace and VGG-Face) distinguish between Middle Eastern and white individuals and Southeast Asian and East Asian individuals. This finer-grained classification could be helpful for researchers studying political behavior within these subgroups.

Furthermore, the fact that the image-based approach relies on perceptual classifications as ground truth may serve as a complement to the Bayesian approach, which is typically evaluated using self-reported data. Studies of ethnic politics have long been interested in the interaction between self-identification and perceived identification of others (Habyarimana et al. 2007; Harris and Findley 2014). This difference may be useful for research that examines two "ingredients" of identity and how they might affect political outcomes such as representation.

Proposed Method: Hybrid Approach

We propose a general supervised learning approach that improves prediction accuracy while mitigating the downstream bias observed with the previous methods. That is, we use probabilities obtained from both BSO and image-based CNN as features in a simple multinomial logistic model and generate predictions of individual ethnorace. The use of BSO predictions not only improves prediction accuracy among Hispanics – a group where image-based methods perform poorly – but also mitigates the potential bias in downstream analyses embedded in other Bayesian methods that rely on geographic information. The use of image-based predictions also improves accuracy for Blacks, a group where the Bayesian approach performs poorly. By allowing the strengths of one

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

to compensate for the weaknesses of the other, we show that this hybrid approach is able to not only reduce bias in downstream analyses but also increase the accuracy of the individual race/ethnicity prediction.

The proposed approach proceeds as follows. First, given data on names, we obtain predicted probabilities for each individual using BSO, which can be done using publicly available tools such as *wru*. Second, we collect images online using full names as search engine keywords. We then use a pre-trained CNN model to generate predicted probabilities of racial and ethnic categories based on the collected images. Third, we train a multinomial logistic model using these two sets of predicted probabilities as features to generate predictions.

In the following sections, we demonstrate the proposed methodology using large-scale data on local elected officials in the USA.[6] We show that, compared to other prediction techniques used in the literature, the proposed classification approach not only mitigates geography-driven biases but also yields a higher out-of-sample accuracy rate. We then discuss in detail the potential risks and comparable benefits associated with the hybrid approach.

Applications

To demonstrate the proposed methodology, we use a large-scale database of local elected officials. At the time of this study, the data set contained over 16,000 unique elected city councillors and mayors in the USA between 1950 and 2021, representing over 800 unique cities across the country. Importantly, it includes full names as well as ethnoracial identities of the elected officials classified as white, Black, Asian, Hispanic/Latino, or other. The data set also includes district identifiers, election year, and party affiliation (if available).[7]

Using this data set, we validate our method in two ways. First, we use the individual-level ethnoracial information in the data set as ground truth and compare the performance of the proposed approach, as well as other commonly used prediction techniques. Second, we test for bias in downstream analysis by conducting two descriptive analyses examining minority representation in government offices. That is, we assess how much estimates computed using our proposed approach differ from those that rely on the ground truth classification.

Procedures

We now document the implementation of our method. First, we use *wru* to predict the ethnoracial identities of all legislators in our data set using the surname-only Bayesian method (BSO) and calculate the posterior probabilities across five ethnoracial classifications: non-Hispanic white, non-Hispanic Black, non-Hispanic Asian, Hispanic/Latino, and Other.[8]

Second, we run a pre-trained CNN model to predict individual ethnoracial categories using publicly available images. To collect facial images associated with each name, we use the first and last names of elected officials as keywords and search for the associated images on a public search engine.[9]

[6] In Appendix I, we show additional empirical applications that replicate a recent study on gentrification and descriptive representation by Lee and Velez (2023b) as well as ethnoracial predictions of the Members of Parliament in the UK.

[7] The ethnoracial identities of the elected officials are collected from multiple external sources, including official lists from non-profit organizations, official city council websites, news articles, and replication materials of published journal articles. In cases where information is still missing, we rely on human-labelled data from Amazon's Mechanical Turk. We validate the performance of the MTurk-coded ethnoracial classification on a random sample of 100 names, which shows an error rate of 0.04. See Appendix E for details.

[8] As noted in footnote 3, the Other category in *wru* combines several distinct ethnoracial groups and the error rates associated with this category may be understated. For example, a biracial candidate may be lumped into a single race and a Native American candidate may be lumped into the white category based on her surname.

[9] We use a privacy-enhancing search engine that does not personalize search results or show results from content farms. This search engine returns a consistent set of images for a given keyword regardless of a web server setting. For each name, we collect 5 most recent images and take the average of the probabilities weighted by relevancy. See Appendix F for details.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

To predict the ethnoracial classifications of the collected images, we use a CNN model pre-trained on the FairFace data set. As discussed, the FairFace data set was constructed with the goal of mitigating racial bias in existing public face data sets and includes samples that are equally distributed among seven different ethnicity groups, annotated by at least three workers for each image (Karkkainen and Joo 2021). Studies find that a neural network trained on FairFace data sets generalizes better than the same model trained with other pre-existing training data sets such as UTKFace and LFWA+ (Greco et al. 2020).[10]

We emphasize that our method does not use facial images tied to particular individuals, but a collection of individual faces associated with a given name. Collecting photos of each elected official and generating predictions based only on those images is possible but we chose an alternative approach for two reasons. First, we attempt to mitigate the ethical concerns around the academic use of personally identifiable photos by carrying out a procedure where images are collected, machine classification is performed, and images are subsequently deleted.[11] Second, independent of ethical considerations, collecting photos for every individual is difficult or often impossible. Through an automatic collection of publicly available images on a large scale, we demonstrate that our method is able to increase efficiency while maintaining high performance, even when the images may not be of a particular individual in the data set.

Using the pre-trained weights from FairFace, we generate predictions for the following classifications: white, Black, East Asian, Indian, Southeast Asian, Middle Eastern, and Hispanic/Latino. To be consistent with the ethnic and racial classifications of the ground truth data – and the categorizations used in Bayesian approaches – we combine the Middle Eastern category with white and all Asian categories into a single Asian category. While we choose to aggregate the ethnoracial categories as such in order to evaluate the prediction performance across different methods, we encourage researchers to utilize more granular categorization in their research.

Lastly, we train a multinomial logistic model using the ground truth ethnoracial classifications as a dependent variable and the following predictions as features: 1) the predicted probabilities of five ethnoracial categories obtained from BSO and 2) the predicted probabilities of four ethnoracial categories obtained from the image-based predictions. We use 70 percent of the data as a training set and test its performance on the remaining 30 percent of the data set.

Prediction accuracy

We first validate the prediction accuracy of the proposed method. To do so, we compare the ethnoracial predictions from the following methods against the ground truth: 1) a surname-only Bayesian approach (BSO); 2) a Bayesian Improved Surname and Geocoding (BISG); 3) a fully Bayesian Improved Surname and Geocoding (fBISG); 4) a full name-based LSTM;[12] 5) image-only predictions based on FairFace; and, finally, 6) the proposed hybrid method that combines predictions 1) and 5).

We calculate the overall error rate, which represents the proportion of officials whose ethnoracial classification is incorrect, as well as the false positive (Type I error) and false negative (Type II error) rates. The results are shown in Table 2. We find that our proposed method reduces the overall error rate to 11.2 percent compared to 18.1 and 18.8 percent from BSO and the image-based prediction, respectively. The results also show that the proposed method is more accurate than fBISG (14.6 percent).

As expected, BSO yields high false negative rates among Blacks as well as high false positive rates among whites but performs well among Hispanics/Latinos.[13] Conversely, the image-based

[10] The pre-trained weights can be accessed via https://github.com/dchen236/FairFace. See Appendix F for a description of the FairFace model. We provide open-source Python scripts that allow users to implement the prediction.

[11] We further discuss the ethical considerations in Section 'Ethical Guidelines'.

[12] We use the *ethnicolr* python package developed by Sood and Laohaprapanon (2018).

[13] Out of 736 officials who are Black, 88 percent of them were predicted as white using BSO.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

Table 2. Overall classification error, Type I error and Type II error rates for each ethnoracial category across six prediction methods. The lowest error rate in each row is in bold. Results are based on out-of-sample data ($N = 4{,}976$)

		BSO	BISG	fBISG	LSTM	Image	Hybrid
	Overall Error Rate	0.181	0.155	0.146	0.229	0.188	**0.112**
Asian (3%)	False Negative	0.331	0.269	**0.238**	0.485	0.285	0.277
	False Positive	0.005	0.006	0.010	0.009	0.016	**0.001**
Black (15%)	False Negative	0.904	0.504	0.398	0.981	0.428	**0.370**
	False Positive	0.008	0.043	0.046	**0.011**	0.027	0.025
Hispanic (10%)	False Negative	0.136	**0.113**	0.118	0.532	0.909	0.142
	False Positive	0.020	0.029	0.032	0.008	**0.006**	0.018
White (73%)	False Negative	0.035	0.085	0.095	**0.026**	0.038	0.049
	False Positive	0.555	0.317	**0.252**	0.747	0.529	0.267
Other (0%)	False Negative	1.000	1.000	1.000	1.000	1.000	1.000
	False Positive	0.0002	0.0004	0.000	0.000	0.000	0.000

prediction yields high false negative rates among Hispanics/Latinos but shows a significant improvement in predicting the identity of Black legislators when compared to the surname-only Bayesian approach.[14] Importantly, the proposed method shows a significant improvement over the two prediction methods alone; specifically, the false negative rates among Blacks are reduced to their lowest values compared to the two methods alone. Among Hispanics/Latinos, the proposed method decreases the false negative rate by a factor of 6 compared to the image-based method. The proposed method also improves predictions among Asian Americans, yielding the lowest false positive rates when compared to all other methods.

Table 2 evaluates the performance of each method based on the plurality-based ethnorace assignment, which obscures the uncertainty embedded in the distribution of probabilities. To address this concern, we provide two additional diagnostics: tetrahedron diagrams and receiver-operated curve (ROC) analyses.[15] First, the tetrahedron diagrams in Figure 1 display the distribution of predicted probabilities across four ethnoracial categories from each prediction method. A tetrahedron diagram is a three-dimensional plot that takes a pyramid form with four vertex corners. Given that there are four primary ethnoracial categories generated by each method, these diagrams are useful for inspecting the composition of the probabilities and capturing the uncertainty associated with them.[16]

Each data point in the diagram represents an individual i with colours indicating the magnitude of the predicted probabilities allocated to each category, from low (blue) to high (red). For example, the red data points near the corner *H(ispanic)* represent individuals whose predicted probabilities of being Hispanic/Latino are close to 1. Conversely, the blue data points in the center represent individuals whose predicted probabilities are rather evenly distributed across the four categories, indicating high prediction uncertainty. The diagrams on the top row display the composition of predicted probabilities only for those individuals whose highest predicted probability correctly predicts their actual ethnoracial category (that is, true positives and true negatives). The diagrams on the bottom are limited to those individuals whose highest predicted probabilities incorrectly predict their ethnoracial category (that is, false positives and false negatives).[17]

Both BISG and fBISG show a relatively higher level of uncertainty as compared to BSO, which is unsurprising as the high probability assigned to the white category gets distributed to other

[14] Out of 485 officials who are Hispanics/Latinos, 83 percent of them were predicted as white using the image recognition model.

[15] In Appendix G, we also show calibration plots that assess the agreement between prediction and true observations in different deciles of the predicted values.

[16] The prediction for the Other category is excluded from the diagram and the probabilities across the four remaining ethnic categories are normalized to sum up to 1.

[17] In Appendix G, we include diagrams that further separate out the false positives from the false negatives.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

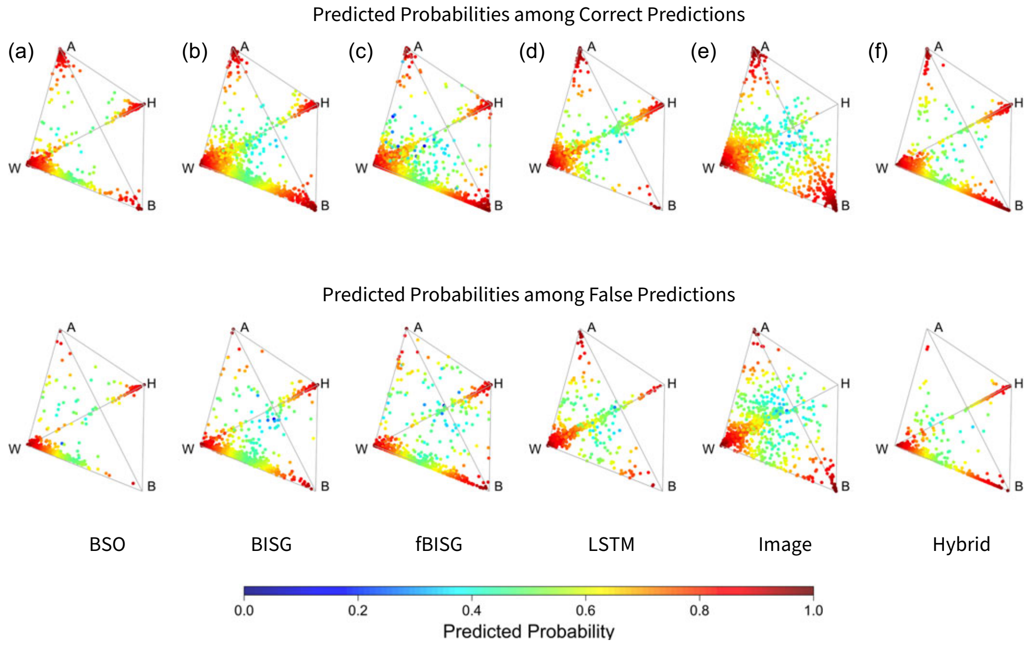

Figure 1. Tetrahedron diagrams of predicted probabilities across four ethnoracial categories (*A* = Asian, *B* = Black, *H* = Hispanic/Latino, and *W* = white).

ethnoracial categories once the city's racial composition is accounted for. All models, however, assign high probabilities to incorrect ethnoracial categories quite frequently, as illustrated by the concentration of red points in the bottom diagrams. One exception is the image-based method: the reduction of red points along the corners of *A(sian)* and *B(lack)* illustrates that when the officials are incorrectly predicted to be of these categories, the predicted probabilities associated with these ethnoracial categories tend to be low. This is also reflected in the hybrid method that incorporates image-based predictions. Nonetheless, we find that the magnitude of predicted probabilities does not necessarily correlate with whether or not the prediction can be made correctly based on the plurality-based ethnorace assignment.

Next, we examine the ROC curve for each prediction method, as shown in Figure 2. Rather than classifying individuals on the basis of their greatest predicted probability, ROC curves display the true positive rate against the false positive rate for a number of different classification thresholds between 0 and 1. The performance of a prediction model is evaluated using the area under the ROC curve, the AUC, which summarizes the overall classification success.

The results show that our method yields the highest AUC score – 0.93 or greater – across all racial categories: BSO alone does not perform well among Blacks and whites, while the image-based method alone does not perform well among Hispanics/Latinos. But combined, the proposed hybrid method is able to reach a higher level of accuracy. With the proposed method, it is possible to reduce the false positive rate among Hispanics/Latinos to 0.05 while maintaining the true positive rate above 0.91. This means that the combined method correctly classifies over 91 percent of Latinos, while only misclassifying 5 percent of non-Latinos as Hispanics/Latinos. This is a significant improvement over the image-based method alone, which generates a false positive rate of 0.42 if it were to maintain the same true positive rate as the combined method. Similarly, the proposed method allows for a reduction of the false positive rate among Blacks to 0.10 while maintaining the true positive rate above 0.80. BSO alone would generate a false positive rate of 0.24 if it were to maintain the same true positive rate.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

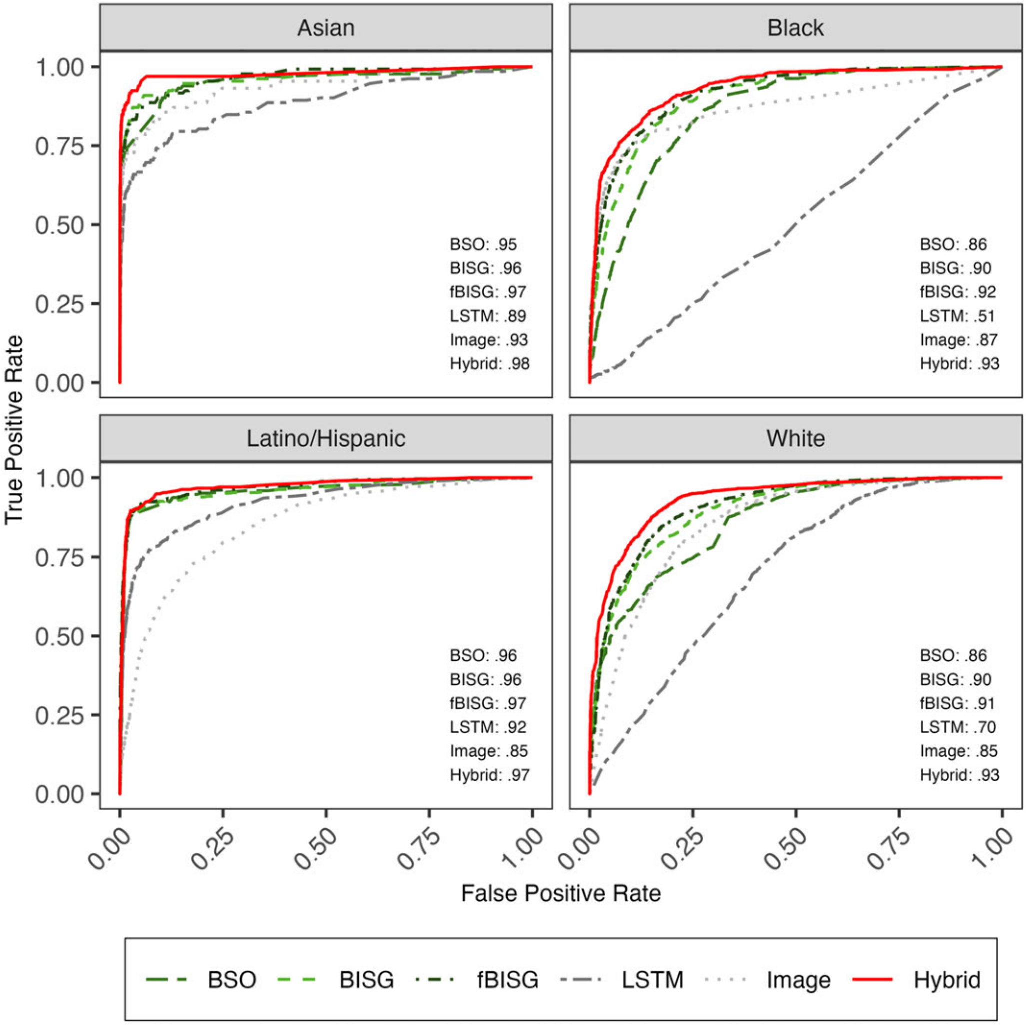

Figure 2. ROC curves across different ethnoracial prediction methods. The AUC scores for each method are shown within each plot.

Evaluating downstream bias

In this section, we assess how well the proposed method describes actual patterns of minority representation in local politics. We do so by examining a popular research question in the descriptive representation literature: the relationship between local ethnoracial composition and the success of co-ethnic candidates. Studies have long theorized that larger co-ethnic populations increase the likelihood of descriptive representation (Shingles 1981; Lublin 1997; Ocampo 2018; Atsusaka 2021). In addition to being the central inquiry in representation studies, this research question views the ethnoracial identity of elected officials as an outcome. Consequently, many empirical studies employ regression models that predict the ethnoracial makeup of politicians while considering factors such as the size of the co-ethnic population and other district-specific covariates (Fraga 2016a; Grumbach and Sahn 2020; Komisarchik and White 2022). This research question therefore offers an ideal empirical setting for evaluating downstream bias when predicted ethnorace is used as an outcome.

Before testing bias in downstream analyses, we first depict how descriptive representation has evolved over time within city councils. Figure 3 shows *loess* curves fitted through the proportion of

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

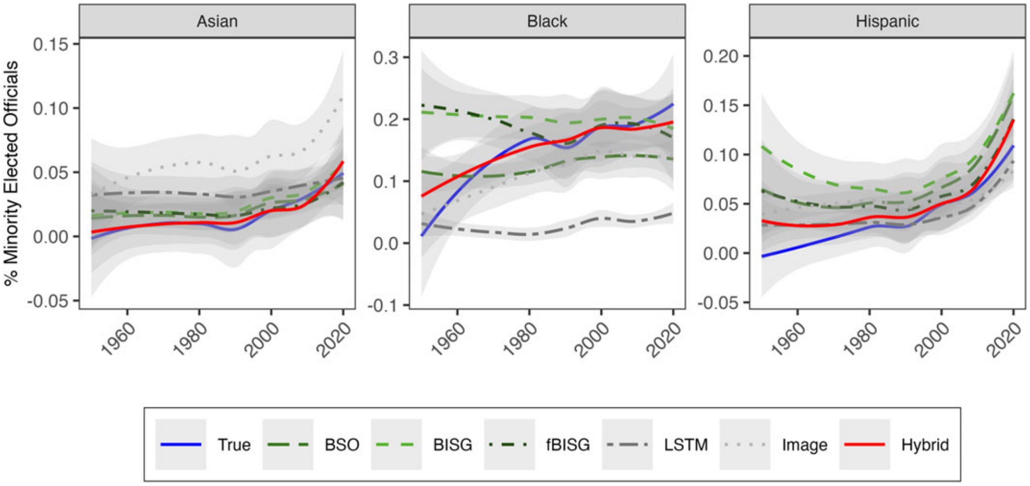

Figure 3. The proportion of non-white officials over time predicted with different ethnoracial category prediction methods.

Asian, Black, and Hispanic elected officials over time. The solid blue curve indicates an estimate based on the ground truth, which we use as a benchmark to evaluate the remaining six curves that are based on predictions.[18]

Examining the ground truth trend (blue line), we find that local government has experienced an increase in the share of Black and Hispanic legislators since the 1950s, but Asian representation shows stagnation until the 2000s and has experienced a slight uptick in recent years. However, the degree to which a researcher can observe such a pattern depends on which prediction method is used to measure the outcome. We observe that the over-time trend estimated with the proposed method most closely tracks the benchmark across all ethnoracial categories. In particular, all other methods diminish the overall increasing trend in Black representation, suggesting that the ability of Black voters to achieve descriptive representation in local government has remained stagnant over the past several decades. Furthermore, BISG and fBISG generally overestimate minority representation, especially during the earlier years when actual minority representation tends to be low. This may be due to the fact that the Census Surname File is only available for 2000 and 2010; thus, the ethnoracial identities of legislators prior to 2000 are predicted with lower accuracy. In comparison, the ability of the hybrid method to produce a fairly accurate result is noteworthy, given that it largely relies on recently collected images.

We now test for bias in the downstream analysis by estimating the relationship between minority population size and ethnoracial diversity in local government. The goal here is to investigate whether the proposed prediction, when used as a dependent variable, can reduce bias in a model that includes a geography-specific factor as an independent variable. We estimate a multinomial logistic model that regresses the ethnoracial category of an elected official i in state s and decade t ($Y_{i,t}$) on k population share in her city. Formally, for $k \in K = \{$Asian, Black, Hispanic$\}$,

$$ln\left(\frac{Pr(Y_i = k)}{Pr(Y_i = White)}\right) = \alpha + \beta_A Asian_{i,t} + \beta_B Black_{i,t} + \beta_H Hispanic_{i,t} + \theta \mathbf{X}_{i,t} + \delta_s + \gamma_t \qquad (1)$$

where $Asian_{i,t}, Black_{i,t}$ and $Hispanic_{i,t}$ represent Asian, Black, and Hispanic/Latino city population, respectively, $\mathbf{X}_{i,t}$ represents a collection of covariates (for example, incumbency and

[18] For each prediction method, we use the probability summed method (DeLuca and Curiel 2022) to estimate the proportion of minority elected officials.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

office title), and δ_s and γ_t denote state and decade fixed effects, respectively. We use the white category as the baseline category. The coefficients $\hat{\beta}_A$, $\hat{\beta}_B$, and $\hat{\beta}_H$ capture the adjusted correlation between the racial/ethnic composition of the population and the composition of the city council.

We run this model six times, each time with a different $Y_{i,t}$ but with the same set of predictors. We first use the ground truth ethnoracial classification for $Y_{i,t}$ and use the $\hat{\beta}_k$ obtained in this model as a benchmark ('benchmark model'). We then create $Y_{i,t}$ using each of the six prediction methods to run the same model ('comparison model') and compare the $\hat{\beta}_k$ against the benchmark estimate.

We choose this particular model in order to reflect the most commonly employed modelling approach in similar research. For example, Fraga (2016a) uses a generalized estimating equation with election year indicators and within-district clusters. Both Komisarchik and White (2022) and Grumbach and Sahn (2020) use a difference-in-difference design including year and geography fixed effects.[19] In Appendix H, we also run alternative specifications and show that the relative RMSE is most stable when the model's outcome is predicted with the proposed hybrid method.

Furthermore, because the multinomial method only allows for a nominal outcome variable, studies typically assign each individual the racial category with the highest predicted probability. This plurality-based method has become a common approach in studies. Indeed, our literature review shows that almost one-third of political science studies involving Bayesian racial prediction use the plurality-based method in their downstream analyses.[20] While we follow this general approach in our application, we recognize that this approach does not account for the uncertainties embedded in the entire distribution of predicted probabilities (Clark, Curiel and Steelman 2021; McCartan et al. 2023). In Appendix H, we present three additional downstream analyses: 1) a Dirichlet regression that takes into account the entire predicted probability distribution as an outcome; 2) a weighted model using the predicted probability as analytic weights; and 3) a thresholding approach that limits the data to individuals whose predicted probability exceeds a predetermined threshold.[21] We find that the results are substantively similar across approaches.[22]

Figure 4 displays $\hat{\beta}_k$ and 95 percent confidence intervals (CIs) from the six comparison models separately for each outcome category k. The black horizontal lines and the grey shaded areas represent $\hat{\beta}_k$ and the 95 percent CIs from the benchmark model. Similar to the simulation results in Appendix D, the estimates using geography-dependent predictions BISG and fBISG are consistently higher, while the estimates obtained from BSO and the image-based method are consistently lower than the benchmark estimate. By contrast, estimates from the hybrid method are closer to the 'benchmark' estimate in virtually all cases.

Substantively, the benchmark model predicts that a Black population shift from 1 percent (10th percentile of the data) to 42 percent (90th percentile) increases the probability of electing a Black official by 32 percentage points (pp). This 'marginal effect' decreases to 22pp under the hybrid method but increases to 48pp and 38pp under BISG and fBISG, respectively, and decreases to 2pp under BSO. The results are even more striking for Hispanic representation, where the estimated coefficient of the hybrid method is lower than that of the benchmark only by 0.0063: the benchmark model predicts that a Hispanic population shift from 2 percent (10th percentile of the data) to 42 percent (90th percentile) increases the probability of electing a Hispanic elected official by 13pp. This predicted probability is nearly identical under the hybrid method (0.57pp higher)

[19] The inclusion of state fixed effects also comports with Decter-Frain (2022) who finds the performance of BISG varies substantially across states.

[20] This number increases to almost half when we exclude studies that do not explicitly discuss the handling of the prediction probabilities in their paper.

[21] We also present results based on Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), which estimates conditional distributions of outcome by predicted race (McCartan et al. 2023). See Appendix H.

[22] While weighting and thresholding estimators are commonly used by researchers in attempts to capture the uncertainty inherent in the prediction, the estimates may still be biased (Chen et al. 2019; McCartan et al. 2023).

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

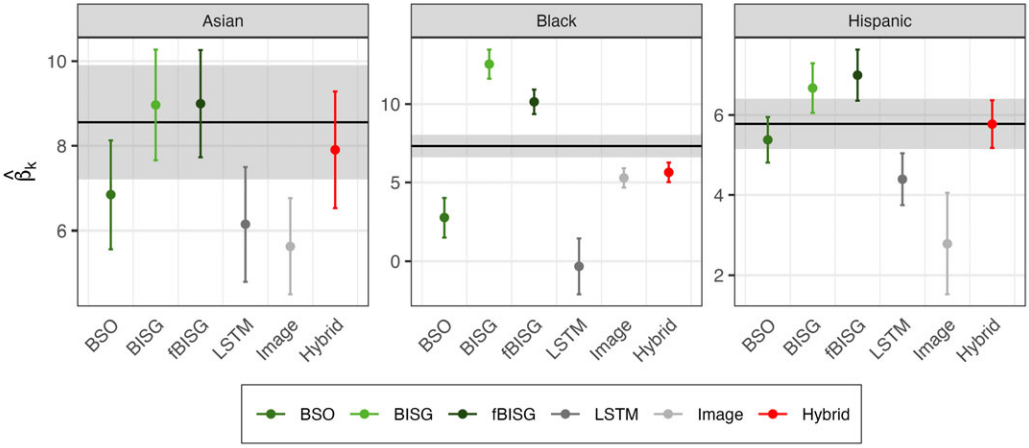

Figure 4. Multinomial logistic regression results. The solid black line and shaded grey area represent estimates and 95 percent CIs from the benchmark model. See Appendix H for full regression results.

while it increases to 17pp and 18pp under BISG and fBISG, respectively. Although the results are noisier, similar patterns are observed for Asian representation.

Assumptions

Comparing across six estimation methods in a sample of over 16,000 local elected officials, we find that the hybrid approach yields concrete advantages over alternatives with respect to predictive accuracy and bias. However, using a supervised machine learning approach with images as additional information requires additional theoretical and statistical assumptions. We address each of these assumptions below.

Image data generating process

Our proposed method relies on a collection of publicly available images associated with individual names, which assumes that these images are representative. We assess potential violations of this assumption by exploring three aspects of the data generating process for images: availability, relevance, and diversity.[23]

First, we assess whether image availability for a given name may influence accuracy. If the total number of available images (for example, popularity of names) varies significantly by racial groups and if availability is correlated with misclassification errors, the image-based predictions would yield lower and imbalanced accuracy. To evaluate this concern, we collect up to 50 recent images per individual, remove those that do not include faces, and exclude exact duplicates. For a given name, we are able to gather 19, 21, 22, and 24 images on average for Asian, Black, Hispanic, and white individuals, respectively, finding comparable levels of availability across racial groups. Further, we find that fewer available images generate *higher accuracy* for minority groups, while the reverse is true among non-Hispanic whites, implying that having more images does not necessarily improve accuracy. Therefore, we rely on the five most recent images rather than using all 50 images collected, which reduces the overall classification error rate from 0.199 to 0.188, as reported in Table 2.

Second, we assess whether variation in image relevance biases accuracy. If the collected images are irrelevant and if such irrelevance is correlated with prediction errors, the accuracy rate may be

[23] We provide further details in Appendix F.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

reduced and unbalanced across groups. We define relevance as the order in which the images appear in the search engine, with images that appear earlier in order as more relevant. We find a mild positive correlation between accuracy and image relevance: the earlier the images appear, the more likely the predictions from those images are correct. To minimize the classification error due to relevance (or a lack thereof), we take the average of the probabilities *weighted* by the relevance when we aggregate the predictions from five images for a given name.[24]

Third, we assess the diversity of images. If a search engine returns facial images unrelated to the name itself, the collected images may add further noise to the prediction. While we cannot assess this directly by identifying the 'true' racial and ethnic identification of each image, we evaluate this issue indirectly by comparing the dispersion of predicted probabilities generated from images against those generated from BSO. If the dispersion is higher for the image-based predictions, we treat the images as having more diverse ethnoracial identities relative to the true distribution of race for a given name in the USA, which may impact the prediction accuracy. We find that the dispersion pattern of image-based predictions is similar to BSO: when the image-based predictions assign lower predicted probabilities across all racial categories, so does BSO. Importantly, among names where BSO is highly uncertain (for example, when BSO assigns 50 percent to one and 50 percent to another racial category), the image-based predictions are more aligned with the ground truth.

Despite the superior performance on metrics such as accuracy and downstream statistical bias, scholars should always be mindful of systematic measurement error. For example, if racially mixed districts tend to elect people who look racially ambiguous, the image-based predictions may also be correlated with the district's racial composition. In Appendix F, we put this to the test and find no evidence of this non-linear expectation.

Neural network architecture

While users may choose a CNN model of their choice, we choose the FairFace architecture, as it is trained on one of the most "complete data sets currently available for ethnicity recognition" (Greco et al. 2020). Across a wide range of image data sets, the model trained on the FairFace data set increases accuracy for various racial groups, whereas preserving the architecture and training with other data sets leads to asymmetric accuracy (Karkkainen and Joo 2021). Another advantage of using FairFace is that its pre-trained model is publicly available, making it easy for users to implement the model without having to build their own. However, as a robustness check, we also generate predictions using the VGG-Face architecture (Parkhi, Vedaldi and Zisserman 2015) and find substantively similar, but slightly lower, predictive performance (see Appendix F). New training data sets and more refined deep network architectures are currently being developed, however, and future research may explore these other models to further improve on the prediction accuracy presented in this paper.[25]

Supervised learning model

Although more sophisticated machine learning algorithms for multi-class classification exist, we choose a multinomial logistic model because of its ability to generate fairly accurate predictions despite its simplicity. We find that estimating a super-learner ensemble using LASSO, Random Forest, and eXtreme Gradient Boosting, as base learners generate error rates that are similar to those from the multinomial logistic regression model proposed in the main text.[26] Similarly,

[24] This approach marginally reduces the overall classification error rate from 0.197 to 0.188, as reported in Table 2.

[25] For example, studies find that fine-tuning a pre-trained VGG-Face architecture with VMER data set results in a higher generalization capability on a different test set (Greco et al. 2020).

[26] We also find that training our models with a different set of features consistently yields lower error rates relative to the other prediction methods (see Appendix G).

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

Decter-Frain (2022) finds that a multinomial logistic regression improves upon BISG just as well as tree-based prediction models (for example, Bayesian Additive Regression Trees).

Concept of interest

To date, the use of racial predictions in social science has predominantly relied on administrative data such as Census and voter files, focusing primarily on self-reported identity as the ground truth. However, racial identity is comprised of a complex dynamic involving self-identification and ascription (Huddy 2001; Lee 2008), a dimension which prior studies have largely overlooked.

As elaborated in Appendix E, the validation of our method using a local elections data set heavily relies on the ethnoracial identity of elites perceived by annotators. Therefore, by operationalizing perceived identity as the ground truth, our method also focuses on an additional aspect of identities – ascription – that is worthy of further exploration. This perspective is especially relevant in studies of elites, where candidates and elected officials often strategically use different identities to signal their proximity to various constituents (Sriram and Grindlife 2017; Kreiss, Lawrence and McGregor 2020; Janusz 2021; Burnett and Kogan 2022). For example, former governor of South Carolina, Nikki Haley, listed her race as white on her 2001 voter registration form, even though she publicly identifies herself as Asian or Indian American (Fraga 2016a). Similarly, Stephanie Murphy, the first Vietnamese-American woman to be elected to Congress – who consistently presented herself as Asian – would receive a 'white' classification under BSO due to her English-sounding name. This discrepancy is most pronounced among Black politicians whose names often obscure their public portrayal. For example, both Representative Hakeem Jeffries (D-NY) and Senator Tim Scott (R-SC) are predicted as Black in the FairFace model, while BSO predicts them as white. Thus, using perceived identity as the ground truth may offer a more accurate reflection of both how politicians present themselves in public as well as how voters perceive politicians.

Still, integrating two distinct definitions of racial identity into a unified classification outcome implies that this method may not be optimal for studies aiming to isolate one particular facet of racial identity. For example, research on substantive representation among minority legislators critically hinges on the assumption that these legislators self-identify as members of racial or ethnic minority communities (Broockman 2013; White, Nathan and Faller 2015). Therefore, while our method introduces new dimensions of identity to the imputation literature, the selection of a prediction model should carefully account for the particular concept of identity that the researcher intends to explore.

Practical Considerations

Our proposed method stands as one among numerous existing predictive tools, each accompanied by unique advantages and disadvantages. In this section, we explore both the benefits and limitations of the proposed method vis-à-vis other commonly used techniques in terms of applicability and practicality.

Applicability

The proposed method broadens the set of prediction tools that can be used in the USA. With the recent implementation of differential privacy in the U.S. Census, racial composition data at a granular geographic level (for example, block) may no longer be reliable for future research (U.S. Census Bureau 2023). Newer approaches, such as those explored here, may be promising tools to address these challenges.

The proposed hybrid method may also be useful in contexts outside the USA, where micro-level data are difficult to collect. Across the globe, detailed data on the ethnic or racial composition of the

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

population are scarce, while large collections of names (for example, phone books and voter records) are ubiquitous (Harris 2015). A United Nations review of data collection on race and ethnicity demonstrates a dearth of ethnicity data, particularly in Africa, where over half of the surveyed countries did not report data on ethnic backgrounds (UNDC 2003), and in Europe, where some countries prohibit collecting data on race and ethnicity (Escafré-Dublet and Simon 2011; Simon 2013). In such instances, names and/or images are often the only publicly available resources to generate predictions and answer pressing questions related to group disparities in government. In fact, recent studies have adopted various name-based algorithms to predict race and ethnicity in countries including UK (Webber 2007; Kandt and Longley 2018), the Netherlands (Bodewes, Agyemang, and Kunst 2019), Scotland (Lakha, Gorman, and Mateos 2011), France (Mazières and Roth 2018), Kenya (Harris 2015), and Brazil (Monasterio 2017), to name a few.[27] In Appendix I, we demonstrate the applicability of our proposed method by applying it to members of the UK Parliament.

Cost and Time

In terms of cost, the Bayesian approach relies on Census data available to the public, as well as names included in voter files. The proposed method uses publicly available images, as well as an open-source CNN model, to increase precision without incurring significant data collection costs. However, training the multinomial model requires annotating true ethnorace for a portion of the desired data set. While researchers may do so themselves at zero cost, crowdsourcing the annotation may generate the training set much more quickly at a relatively low cost.[28]

In terms of implementation time, the Bayesian method takes less than one minute to predict 1,000 names using the fBISG method via wru whereas FairFace predicts 1,000 images in just under three minutes. Additionally, the hybrid method requires extra time to collect images: our data collection record shows that it takes about thirty seconds to collect the twenty most recent images while simultaneously checking for whether faces are present, as well as removing duplicates. However, as discussed under the Section ‘Assumptions’, using as few as five images per name continues to maintain superior accuracy rates over other methods. Hence, it is possible to increase speed with only a marginal loss in accuracy. To facilitate the use of the hybrid method, we also provide step-by-step instructions and open-source Python scripts.

Ethical Guidelines

Even though the proposed method is geared towards public figures, a group that has historically been subject to weaker human subjects protections, scholars should always be mindful of ethical concerns regarding the use of administrative and public data for research purposes. On the one hand, using potentially sensitive information, such as names and addresses in BISG/fBISG (and images in convolutional neural networks), can lead to accuracy gains. However, such sensitive information can be exposed in a data breach, putting human subjects at risk. On the other hand, maximizing human subjects’ protection at all costs may lead to sub-optimal accuracy in domains where accuracy is crucial. For example, one could redact any personally identifiable information from administrative data sets. However, this approach would render predictions impossible and could have detrimental consequences when using such statistics in domains like public policy, redistricting, or other settings that demand accurate estimates of group behaviors.

The balance between accuracy and human subjects protection is a decision that ought to involve various stakeholders, including institutional review boards (IRBs), researchers, and communities under study. In ongoing research, we find that the general public views research

[27] In a review of 120 published articles in health science that imputes an individual ethnorace based on names, Chin et al. (2023) find that 45 percent of them are studies conducted outside of the USA.

[28] We paid crowdsource workers (MTurk) approximately $0.30 per councillor to provide data on race and ethnicity.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

involving ethnoracial classifications of elites as more ethically permissible than research classifying individual voters, and that accurate research designs are seen as more ethically permissible than those presenting a less accurate view of racial inequality (Lee and Velez 2023a). While low-stakes research without direct policy implications may entail a stronger justification for the use of administrative information or public images, research that directly informs public policy and litigation might require a different approach, given the critical importance of accurate descriptive estimates.

Whether the data are collected by administrative units (for example, names and addresses) or publicly available (for example, images), stringent data security plans are necessary. We strictly followed data collection procedures approved by the IRB. In Appendix A, we describe our data collection procedure and how confidentiality was maximized by using anonymized identifiers and deleting images immediately after predictions were generated. Researchers interested in applying the proposed method to a more general sample must make explicit the ethical concerns and weigh possible benefits against potential social or political costs in advance. They should also abide by applicable laws in their country governing the use of biometric information by researchers.

Discussion

The stakes of accurately estimating populations are high. Policies may be justified on the basis of descriptive statistics about protected classes. Litigation involving voting rights depends on the best available data about joint distributions between race and voting behavior. Redistricting depends on precise counts of populations. Naturally, the importance of obtaining estimates that faithfully represent reality privileges the need for accuracy when addressing questions of inequality.

By combining the predictive power from the surname-only Bayesian method with an image-based prediction model, our proposed method increases predictive accuracy compared to other techniques widely used in the literature and may minimize bias due to geography-induced systematic measurement error. Our method yields an overall classification success of 0.93 or greater across the four main ethnoracial categories – the highest among all other prediction methods evaluated. It also recovers point estimates that are closest to benchmarks based on hand-coded data. While we choose particular models to generate features in our application (that is, FairFace for image-based and BSO for surname-based predictions), we leave the choice of particular CNN models or alternative name-based methods other than BSO to the discretion of researchers.[29]

The selection of prediction methods involves balancing theoretical, statistical, and practical considerations. The hybrid method excels in overall accuracy, independence from geolocation data, and applicability to contexts where Census data may be unavailable but may demand more computational resources. Furthermore, the proposed method proves effective in mitigating bias in downstream analyses, a concern that arises in contexts where predictions are used as outcomes that correlate with geography-specific covariates. Nonetheless, geography-dependent prediction methods maintain high accuracy rates and remain valuable outside of these scenarios.[30]

Researchers may find this method useful for addressing a variety of questions related to ethnic politics, including minority candidate emergence, patterns of descriptive representation, campaign finance, and coalition-building for ethnic minority candidates. Moreover, as the proposed method utilizes both perceptual and self-prescribed indicators of identity, it opens

[29] For example, in Appendix I, we use the LSTM model instead of BSO to generate surname-only predictions for members of the UK Parliament.

[30] In particular, a reduction in false negative rates among Blacks in fBISG is comparable to that of the hybrid approach in Table 2. Furthermore, while the estimation of racial disparities in other topics such as turnout and policy impact may still be biased if the outcome is correlated with the residuals of the racial predictions, a new identification strategy introduced by McCartan et al. (2023) can be used to correct for bias under the assumption that the outcome is conditionally independent of surname given (unobserved) ethnorace, geography, and other observed characteristics.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

avenues to examine the interplay between the two ingredients of identity and how they might affect political outcomes at the level of individuals and institutions.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0007123424000772.

Data availability statement. Replication data for this paper can be found at https://doi.org/10.7910/DVN/KRTVHD. Python scripts to implement the hybrid method can be found at https://github.com/DianaDaInLee/PredEthnorace.

Acknowledgements. This study was approved by IRB at Columbia University (IRB-AAAT8619). The data collection was supported by the Russell Sage Foundation. We would like to thank Yuki Atsusaka, James Bisbee, Naoki Egami, Donald P. Green, Macartan N. Humphreys, Santiago Olivella, Michael Peress, Chris Warshaw, Justin de Benedictis-Kessner, Dan Thompson, and participants at PolMeth XXXIX for helpful comments.

Financial support. The data collection was supported by the Russell Sage Foundation.

Competing interests. None.

References

Abdulwahid AI (2023) Classification of ethnicity using efficient CNN models on MORPH and FERET datasets based on face biometrics. *Applied Sciences* **13**(12), 7288.

Abott C and Magazinnik A (2020) At-large elections and minority representation in local government. *American Journal of Political Science* **64**(3), 717–733.

Alvarez RM, Katz JN and Seo-young Silvia Kim S-YS (2020) Hidden donors: The censoring problem in US federal campaign finance data. *Election Law Journal: Rules, Politics, and Policy* **19**(1), 1–18.

Argyle LP and Barber M (2024) Misclassification and bias in predictions of individual ethnicity from administrative records. *American Political Science Review*, **118**(2), 1058–1066.

Atsusaka Y (2021) A logical model for predicting minority representation: application to redistricting and voting rights cases. *American Political Science Review* **115**(4), 1210–1225.

Barreto M (2007) İsí se puede! Latino candidates and the mobilization of Latino voters. *American Political Science Review* **101**(3), 425–441.

Barreto MA, Segura GM and Woods ND (2004) The mobilizing effect of majority-minority districts on Latino turnout. *American Political Science Review* **98**(1), 65–75.

Berry WD and Stanley Feldman S (1985) *Multiple regression in practice*. SAGE Publications.

Bodewes AJ, Agyemang C and Kunst AE (2019) All-cause mortality among three generations of Moluccans in the Netherlands. *European Journal of Public Health* **29**(3), 463–467.

Broockman DE (2013) Black politicians are more intrinsically motivated to advance blacks' interests. *American Journal of Political Science* **57**(3), 521–36.

Burnett CM and Kogan V (2022) Do nonpartisan ballots racialize candidate evaluations? Evidence from 'who said what?' experiments. *Party Politics* **28**(3), 541–553.

Canon DT and Posner RS (1999) *Race, redistricting, and representation: The unintended consequences of black majority districts*. Chicago: University of Chicago Press.

Chen J, Kallus N, Mao X, Svacha G and Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency,* 339–348.

Cheng L, Gallegos IO, Ouyang D, Goldin J and Ho D (2023) How redundant are redundant encodings? Blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 667–686.

Chin MK, N Đoàn LN, Russo RG, Roberts T, Persaud S, Huang E, Fu L, Kui KY, Kwon SC and Yi SS (2023) Methods for retrospectively improving race/ethnicity data quality: A scoping review. *Epidemiologic Reviews* **45**(1), 127–139.

Clark JT, John A, Curiel JA and Tyler S, Steelman TS (2021) Minmaxing of Bayesian improved surname geocoding and geography level ups in predicting race. *Political Analysis* **30**(3), 1–7.

Conroy M and Green J (2020) It takes a motive: communal and agentic articulated interest and candidate emergence. *Political Research Quarterly* **73**(4), 942–956.

Decter-Frain A (2022) How should we proxy for race/ethnicity? Comparing Bayesian improved surname geocoding to machine learning methods. *arXiv* 2206.

DeLuca K and Curiel JA (2022) Validating the applicability of Bayesian inference with surname and geocoding to congressional redistricting. *Political Analysis,* 1–7. https://doi.org/10.1017/pan.2022.14.

Eligon J (2015) *Election in scarred Ferguson carries hope of 'new tomorrow'*. https://www.nytimes.com/2015/04/05/us/election-in-scarred-ferguson-carries-hope-of-new-tomorrow.html.

Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P and Lurie N (2009) Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* **9**(2), 69–83.

Enos RD (2016) What the demolition of public housing teaches us about the impact of racial threat on political behavior. *American Journal of Political Science* **60**(1), 123–142.

Escafré-Dublet A and Simon P (2011) Ethnic statistics in Europe: The paradox of colorblindness. *European multiculturalisms: Cultural, religious, and ethnic challenges*, 213–37.

Fraga BL (2016a) Candidates or districts? Reevaluating the role of race in voter turnout. *American Journal of Political Science* **60**(1), 97–122.

Fraga BL (2016b) Redistricting and the causal impact of race on voter turnout. *The Journal of Politics* **78**(1), 19–34.

Fraga BL (2018) *The turnout gap: Race, ethnicity, and political inequality in a diversifying America*. Cambridge University Press.

Gay C (2001) The effect of black congressional representation on political participation. *American Political Science Review* **95**(3), 589–602.

Gilens M (2012) *Affluence and influence: Economic inequality and political power in America*. Princeton University Press.

Greco A, Percannella G, Vento M and Vigilante V (2020) Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications* **31**, 1–13.

Grumbach JM and Sahn A (2020) Race and representation in campaign finance. *American Political Science Review* **114**(1), 206–221.

Habyarimana J, Humphreys M, Posner DN and Weinstein JM (2007) Placing and passing: Evidence from Uganda on ethnic identification and ethnic deception. *American Political Science Association, Chicago* **30**, 1–25.

Hajnal Z (2010) *America's uneven democracy: Race, turnout, and representation in city politics*. Cambridge University Press.

Hankinson M and Magazinnik A (2023) The supply-equity trade-off: The effect of spatial representation on the local housing supply. *The Journal of Politics* **85**(3), 1033–1047.

Harris AS and Findley MG (2014) Is ethnicity identifiable? Lessons from an experiment in South Africa. *Journal of Conflict Resolution* **58**(1), 4–33.

Harris JA (2015) What's in a name? A method for extracting information about ethnicity from names. *Political Analysis* **23**(2), 212–224.

Hausman J (2001) Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives* **15**(4), 57–67.

Huddy L (2001) From social to political identity: A critical examination of social identity theory. *Political psychology* **22**(1), 127–156.

Imai K and Khanna K (2016) Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* **24**(2), 263–272.

Imai K, Olivella S and Rosenman ETR (2022) Addressing census data problems in race imputation via fully Bayesian improved surname geocoding and name supplements. *Science Advances* **8**(49), eadc9824.

Janusz A (2021) Electoral incentives and elite racial identification: Why Brazilian politicians change their race. *Electoral Studies* **72**, 102340.

Juenke EG and Shah P (2015) Not the usual story: The effect of candidate supply on models of Latino descriptive representation. *Politics, Groups, and Identities* **3**(3), 438–453.

Kandt J and Longley PA (2018) Ethnicity estimation using family naming practices. *PLoS One* **13**(8), e0201774.

Karkkainen K and Joo J (2021) Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of The IEEE/CVF Winter Conference on Applications of Computer Vision, 1548–1558.

Kogan V, Lavertu S and Peskowitz Z (2021) How does minority political representation affect school district administration and student outcomes? *American Journal of Political Science* **65**(3), 699–716.

Komisarchik M and White A (2022) Throwing away the umbrella: Minority voting after the Supreme Court's Shelby decision. Working Paper. Available from https://arwhite.mit.edu/sites/default/files/images/vra_post_shelby_current.pdf.

Kreiss S, Lawrence RG and McGregor SC (2020) Political identity ownership: Symbolic contests to represent members of the public. *Social Media+ Society* **6**(2), 1–5.

Lakha F, Gorman DR and Mateos P (2011) Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland. *Public Health* **125**(10), 688–696.

Lee DDI and Velez YR (2023a) *Ethical use of administrative data in inequality research*. Preprint. In Review. https://osf.io/snm7p.

Lee DDI and Velez YR (2023b) Rising tides or political ripcurrents? Gentrification and minority representation in 166 cities. *Urban Affairs Review* **60**(3), 956–982.

Lee Diana Da In (2024) "Replication Data for: Measuring Descriptive Representation at Scale: Methods for Predicting the Race and Ethnicity of Public Officials", https://doi.org/10.7910/DVN/KRTVHD, Harvard Dataverse, V1.

Lee T (2008) Race, immigration, and the identity-to-politics link. *Annu. Rev. Polit. Sci.* **11**,457–478.

Lemi DC (2021) Do voters prefer just any descriptive representative? The case of multiracial candidates. *Perspectives on Politics* **19**(4), 1061–1081.

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press

Lublin D (1997) *The paradox of representation*. Princeton: Princeton University Press.

Lyu H, Fan Y, Xiong Z, Komisarchik M and Luo J (2021) Understanding public opinion toward the #stopasianhate movement and the relation with racially motivated hate crimes in the US. *IEEE Transactions on Computational Social Systems* **10**(1), 1–12. https://doi.org/10.1109/TCSS.2021.3136858.

Maghbouleh N, Schachter A and Flores RD (2022) Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be white. *Proceedings of the National Academy of Sciences* **119**(7), e2117940119.

Marschke G, Nunez A, Weinberg BA and Yu H (2018) Last place? the intersection of ethnicity, gender, and race in biomedical authorship. *Aea papers and proceedings* **108**, 222–227.

Mazières A and Roth C (2018) Large-scale diversity estimation through surname origin inference. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* **139**(1), 59–73.

McCartan C, Goldin J, Ho DE and Imai K (2023) Estimating racial disparities when race is not observed. *arXiv preprint arXiv:2303.02580*.

Menshikova A and van Tubergen F (2022) What drives anti-immigrant sentiments online? a novel approach using Twitter. *European Sociological Review* **38**(5), 694–706.

Messager ML, Ettinger AK, Murphy-Williams M and Levin PS (2021) Fine-scale assessment of inequities in inland flood vulnerability. *Applied Geography* **133**, 1–11.

Monasterio, Leonardo. 2017. Surnames and ancestry in Brazil. *PloS one* **12** (5): e0176890.

Ocampo AX (2018) The wielding influence of political networks: Representation in majority-Latino districts. *Political Research Quarterly* **71**(1), 184–198.

Parkhi OM, Vedaldi A and Zisserman A (2015) Deep face recognition. In Xie X, Jones MW and Tam GKL (eds), *Proceedings of the British Machine Vision Conference*. BMVA Press, 41.1–41.12.

Schwemmer C and Jungkunz S (2019) Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks. *Political Research Exchange* 1(,): 1–23.

Shah PR (2014) It takes a black candidate: A supply-side theory of minority representation. *Political Research Quarterly* **67**(2), 266–279.

Shah PR and Davis NR (2017) Comparing three methods of measuring race/ethnicity. *Journal of Race, Ethnicity, and Politics* **2**(1), 124–139.

Shingles RD (1981) Black consciousness and political participation: The missing link. *American Political Science Review* **75**(1), 76–91.

Simon P (2013) Collecting ethnic statistics in Europe: A review. In Simon P and Piche V (eds), *Accounting for Ethnic and Racial Diversity*, London: Routledge, pp. 10–35.

Sood G and Laohaprapanon S (2018) Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109*.

Sriram SK and Grindlife S (2017) The politics of deracialisation: South Asian American candidates, nicknames, and campaign strategies. *South Asian Diaspora* **9**(1), 17–31.

Sulaiman MA and Kocher IS (2022) A systematic review on evaluation of driver fatigue monitoring systems based on existing face / eyes detection algorithms. *Academic Journal of Nawroz University* **11**(1), 57–72.

Swain CM (1995) *Black faces, black interests: The representation of African Americans in Congress*. Cambridge: Harvard University Press.

Tate K (1991) Black political participation in the 1984 and 1988 presidential elections. *American Political Science Review* **85**(4), 1159–1176.

U.S. Census Bureau (2023) *Why the Census Bureau chose differential privacy*. U.S. Census Bureau, March.

UNDC (2003) *Ethnicity: A review of data collection and dissemination*. Technical report. United Nations Statistics Division, Demographic Social Statistics Branch, Social and Housing Statistics Section.

Webber R (2007) Using names to segment customers by cultural, ethnic or religious origin. *Journal of Direct, Data and Digital Marketing Practice* **8**, 226–242.

White AR, Nathan NL and Faller JK (2015) What do I need to vote? Bureaucratic discretion and discrimination by local election officials. *American Political Science Review* **109**(1), 129–142.

Wong KO, Zaiane OR, Davis FG and Yasui Y (2020) A machine learning approach to predict ethnicity using personal name and census location in Canada. *PloS one* **15**(11), e0241239.

Yamashita L (2022) ‘I just couldn’t relate to that Asian American narrative’: How Southeast Asian Americans reconsider panethnicity. *Sociology of Race and Ethnicity* **8**(2), 250–266.

Cite this article: Lee DDI and Velez YR (2025) Measuring Descriptive Representation at Scale: Methods for Predicting the Race and Ethnicity of Public Officials. *British Journal of Political Science* 1–21. https://doi.org/10.1017/S0007123424000772

https://doi.org/10.1017/S0007123424000772 Published online by Cambridge University Press