Advance Access publication March 7, 2015

Political Analysis (2015) 23:212–224
doi:10.1093/pan/mpu038

What's in a Name? A Method for Extracting Information about Ethnicity from Names

J. Andrew Harris

Assistant Professor of Political Science, New York University - Abu Dhabi
e-mail: andy.harris@nyu.edu

Edited by R. Michael Alvarez

Questions about racial or ethnic group identity feature centrally in many social science theories, but detailed data on ethnic composition are often difficult to obtain, out of date, or otherwise unavailable. The proliferation of publicly available geocoded person names provides one potential source of such data—if researchers can effectively link names and group identity. This article examines that linkage and presents a methodology for estimating local ethnic or racial composition using the relationship between group membership and person names. Common approaches for linking names and identity groups perform poorly when estimating group proportions. I have developed a new method for estimating racial or ethnic composition from names which requires no classification of individual names. This method provides more accurate estimates than the standard approach and works in any context where person names contain information about group membership. Illustrations from two very different contexts are provided: the United States and the Republic of Kenya.

1 Introduction

Political scientists often consider theories about racial or ethnic identity at the local level, where detailed data on the ethnic or racial composition of the population are scarce (Hopkins 2010; Enos 2011; Kasara 2013).[1] At the same time, large numbers of locally geo-coded person names (e.g., voter registers or phone listings) are increasingly available. Given the relationship between person names and group identity, these data can be used to generate local estimates of group composition. In this article, I develop a method to estimate group proportions from a list of names that avoids the error-prone and tedious classification of individual names.

To provide a proof of concept, I begin in a context with copious information on names and racial demography: the United States. The U.S. data are rich enough to examine the performance of estimators in both Monte Carlo simulations as well as real data from Florida and North Carolina. I then apply the proposed method to names from the East African nation of Kenya, where existing direct measures of local ethnic composition (e.g., census or survey data) are, like many places in the developing world, unavailable or unsuitable for the research question. Based on King and Lu (2008) and Hopkins and King (2010), the method avoids individual classification of names in a list and instead focuses on modeling the proportions of each unique name in a list. This approach yields more efficient estimates of group proportions than approaches based on individual

Author's note: The author is grateful for comments from Andy Eggers, Arthur Spirling, Rachel Gould, Ben Ansell, Bernard Grofman, Gary King, Ken Benoit, Dominik Hangartner, Geoffrey Evans, and Lucy Barnes. Two anonymous reviewers provided excellent comments that resulted in a significantly improved manuscript. The author gratefully acknowledges his time at Nuffield College, Oxford University, as Postdoctoral Prize Research Fellow, during which much of this work was written. Computation for the research was carried out on the High Performance Computing resources at New York University–Abu Dhabi, with the enthusiastic support of Muataz Al-Barwani and Benoit Marchand. The replication archive for this article is available at the Political Analysis Dataverse as Harris (2014). Supplementary materials for this article are available on the *Political Analysis* Web site.

[1] UNSD (2003) reviews the collection and dissemination of data on race and ethnicity. The results paint a grim picture of the availability of ethnicity data, particularly in Africa, where over half of surveyed countries did not report data on ethnicity.

classification, and performs well even when the training and target populations differ. By reframing the estimation problem in this way, existing diagnostic measures and bias reduction techniques for regression to improve inferences that are unavailable with classification-based approaches can be used.

The next section outlines the estimation problem, describes issues with current approaches, and presents an alternative method for estimating group proportions. Section 3 presents results from Monte Carlo simulations, demonstrating that the proposed approach outperforms existing approaches. Section 4 discusses the collection of training data. Sections 5 and 6 present the applications in North Carolina and Kenya. Section 7 concludes. Code to implement these methods is available in the online appendix and on the author's website.

2 Estimating Ethnic Proportions from Names

The goal of the proposed method is to estimate the group (e.g., racial or ethnic) proportions of individuals in a list of names. Perhaps the most intuitive way to do this is to classify each name in the list into an ethnic group, and then aggregate those classifications into the desired proportions. Name classification begins by defining the relationship between an individual's name and a set of ethnic groups. One way to do this is to define a probability distribution across J ethnic groups for each name i. Then, we draw a classification for each individual person k with name i from this distribution:

$$G_k \sim P(\text{Group } j|\text{Name } i). \tag{1}$$

To estimate proportions, G_k for each of the K individuals in the list is drawn. Then, the proportion of group j as $\hat{\theta}_j = \frac{\sum_{k=1}^{K} I_{(G_k=j)}}{K}$, where $I(\cdot)$ is equal to one when the expression is true, zero otherwise, is estimated. In other words, the estimated proportion for group j is the number of individuals categorized in group j divided by the total number of individuals.

In practice, $P(\text{Group } j|\text{Name } i)$ is defined using a set of *training* data in which names are linked with groups. Dictionary-based approaches deterministically link names to ethnic groups; this may be done using a distillation of expert knowledge or simply assigning the most likely ethnic group to each name.[2] A second approach uses the full conditional distribution $P(\text{Group } j|\text{Name } i)$ to draw classifications for individual names, as in equation (1).[3] Either way, this mapping is then applied to *target* names (i.e., names without information on ethnicity) to generate classifications, which can then be aggregated into proportions.

If the quantity of interest is an individual person's racial or ethnic identity, these classification-based approaches may be appropriate. Aggregation of individual classifications is problematic when group proportions are the estimand. First, the estimated conditional in equation (1) may be different from the actual (unobserved) data-generating process for some or all of the names in the list. This causes the misclassification of some individuals' names into the wrong group. Since $\sum_{j=1}^{J} P(\text{Group } j|\text{Name } i) = 1$, misclassification errors in one group must be offset by errors in the other estimated parameters. Second, classification uses only the information in G_k—the individual classification for individual k's name—discarding information about a name's usage across groups. Third, and perhaps most importantly, each name classification is done independently of the others, discarding information about the other names in the target data.[4]

I introduce an alternative approach to estimate group proportions that builds on work in Hopkins and King (2010) and King and Lu (2008). This proposed method uses all information about the names in a target set simultaneously, avoids the need to compile a name dictionary or

[2] Mateos (2007) reviews a number of dictionary-based approaches.
[3] A growing literature in public health and machine learning explores ways to estimate the conditional in equation (1) to classify names into groups (Coldman, Braun, and Gallagher 1988; Ambekar et al. 2009; Treeratpituk and Giles 2012).
[4] Examples of recent work using a classify-and-aggregate approach include Rosenwaike (1994), Mateos (2011), Byrne and O'Malley (2012), Kasara (2013), Grofman and Garcia (2014), and Susewind (2015).

estimate $P(\text{Group } j|\text{Name } i)$, and allows for standard regression-based statistical fixes when such discrepancies threaten inference.

Instead of classifying the names of K individuals, the proportions of the N unique names in the list of people are modeled. The entire vector of names in the target data is represented as an N-length vector $P(\mathbf{Name})$, where each element of $P(\mathbf{Name})$ represents the proportion of name i in the list. For instance, a common name like Smith may make up 3% of the names in a list, whereas a name like Klein may only make up 0.5%. This N-length vector $P(\mathbf{Name})$ represents the target data. Then, $P(\text{Name} \mid \text{Group})$ is used as the training data. For example, suppose there are two groups—Black and White. Then, by the total law of probability:

$$\begin{aligned}
P(\text{Smith}) &= P(\text{Smith}|\text{White}) \times P(\text{White}) + P(\text{Smith}|\text{Black}) \times P(\text{Black}) \\
P(\text{Johns}) &= P(\text{Johns}|\text{White}) \times P(\text{White}) + P(\text{Johns}|\text{Black}) \times P(\text{Black}) \\
P(\text{Klein}) &= P(\text{Klein}|\text{White}) \times P(\text{White}) + P(\text{Klein}|\text{Black}) \times P(\text{Black}) \\
\vdots &= \vdots + \vdots
\end{aligned}$$

The target data on the left-hand side of the set of equations are calculated from names found on a voter register for a given precinct. The training data are the known or estimated conditional probabilities on the right-hand side, $P(\text{Smith}|\text{White})$, and so on. $P(\text{White})$ and $P(\text{Black})$ are the unknown estimand, and can be estimated like coefficients β_1 and β_2 in a linear regression, such that $\beta_1 + \beta_2 = 1$.

This approach can be generalized for N names and J groups, stacking the equations for each of the N names in the target data and rewriting in matrix notation:

$$\underset{N\times 1}{P(\mathbf{Name})} = \underset{N\times J}{P(\mathbf{Name}|\mathbf{Group})}\ \underset{J\times 1}{P(\mathbf{Group})}\,. \tag{2}$$

Equation (2) mirrors the standard matrix notation used for linear regression, where $P(\mathbf{Name})$ is the outcome variable; the matrix $P(\mathbf{Name}|\mathbf{Group})$ contains the independent variables; and $P(\mathbf{Group})$ are the J unknown group proportions. By representing the training data as the N-length vector $P(\mathbf{Name})$, all of the training data is used simultaneously to estimate the quantities of interest, rather than as the aggregation of a series of independent classifications.

Figure 1 presents a geometric example of the approach in three dimensions. The slopes governing the plane represent the group proportions—$\beta_1 = P(\text{White}) = 72\%$ and $\beta_2 = P(\text{Black}) = 28\%$ in this example data. The z-axis reports the proportion of each name in the target data, $P(\mathbf{Name})$. The x-axis reports the probabilities of each name in the White training population; the y-axis reports the probabilities of each name in the Black training population. The slope coefficients estimated in this way incorporate all covariate data contained in $P(\mathbf{Name}|\mathbf{Group})$.

While linear regression can be used to estimate the βs, this may give estimates greater than one or less than zero. Following King and Lu (2008) and Hopkins and King (2010), the proposed method uses constrained optimization (rather than linear regression) to estimate the coefficients, ensuring that the estimated group proportions sum to one.[5] Basic code to implement the method is available in the online appendix, with additional code available on the author's website and this journal's dataverse.

2.1 *Key Assumption*

The key assumption enabling estimation is that $P(\mathbf{Name}|\mathbf{Group})$ is the same in the training and the target data. Of course, this conditional for the target data cannot be observed, and the assumption may be violated in practice. For example, group j might use name i more frequently in the region where the target data are collected relative to the training data. Or, the researcher might

[5] Specifically, the proposed approach uses the quadratic programming methods described in Goldfarb and Idnani (1982) and Goldfarb and Idnani (1983) and implemented in the `R` package `quadprog` Turlach and Weingessel (2013).

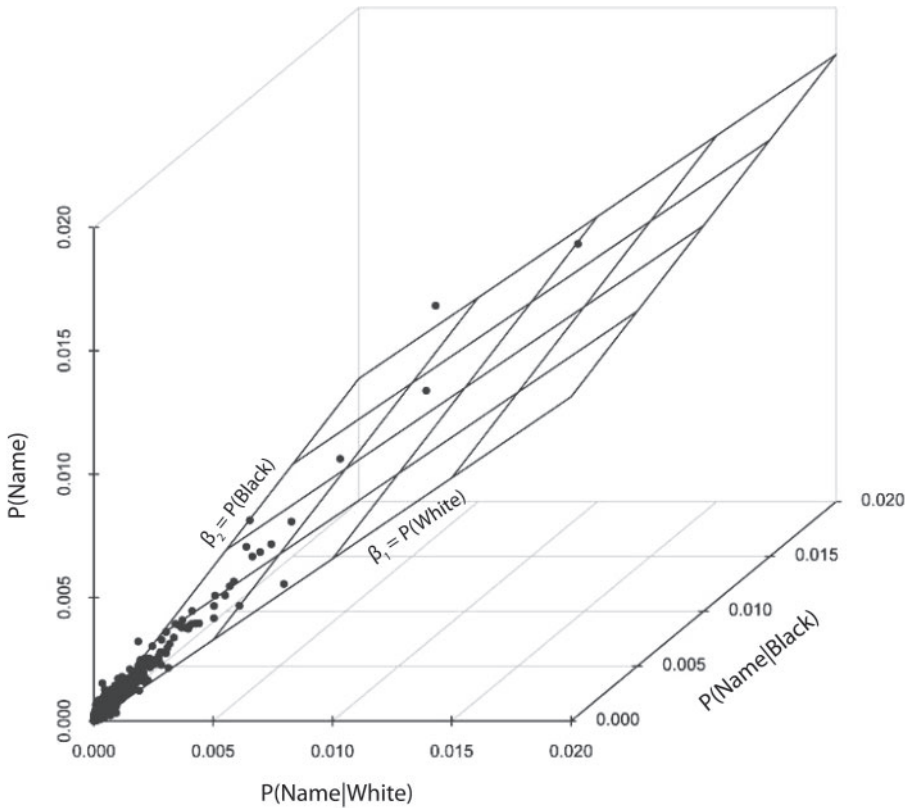

Fig. 1 Geometric Representation of the Approach: The target data—*P*(**Name**)—lie on the vertical *z*-axis, while the training data vectors *P*(**Name**|Black) and *P*(**Name** | White) lie on the horizontal *x*- and *y*-axes. The slopes governing the plane represent the estimated group proportions: $\beta_1 = P(\text{White}) = 72\%$ White and $\beta_2 = P(\text{Black}) = 28\%$ Black.

inadvertently include significant numbers of names from group *-j* when calculating *P*(**Name**|**Group** *j*), conflating names from two or more groups in the training data. Such violations are problematic for estimation via classify-and-aggregate as well. Indeed, these problems will likely be more acute for the classify-and-aggregate approach, as errors will be distributed across a relatively small *J*-length vector, affecting each name *i* classified.

Because the proposed approach is regression-based, existing methods to improve inferences using model-fit diagnostics (to identify cases where the key assumption may be violated) and outlier identification techniques (to reduce bias) can be used. These improvements are not available with classify-and-aggregate approaches. First, the standard R^2 measure can be used to assess the fit of the training data and estimated coefficients to the target data. A low R^2 implies a poor fit. Figure 4A in the online appendix shows a strong negative relationship between R^2 and mean absolute error (MAE) in county-level data from North Carolina and Florida: counties with a higher R^2 tend to have lower MAE. This implies that, at the very least, R^2 can be used to identify those sets of estimates that are well described by the training data, and estimates with lower R^2 can be selected for more scrutiny. As each context will have differing degrees of noise in names data, it is not possible to derive an analytic threshold below which estimates using the new approach should be questioned. Rather, researchers should use model fit as one way to detect problems.

Second, outlier-detection statistics like the one developed in Cook (1977) can be used to identify and remove or down-weight influential observations in the target vector *P*(**Name**). Since name proportions are directly related to individuals in the group, care must be taken not to remove "too much" density by removing influential observations, as is common practice. Instead, inverse exponential weighting is used to down-weight observations with a large Cook's distance. Table 1 below presents Monte Carlo evidence that this approach improves estimates across a wide range of

Table 1 Monte Carlo simulations show that the proposed method has better performance than classification-and-aggregation over a wide range of sample sizes and parameter values: Lower values represent better performance. Each cell is the ratio of mean absolute error (across 5000 simulations) for classify-and-aggregate to the proposed method (panels 1, 2, and 4) and the basic proposed method versus its weighted improvement (panel 3). Each cell tests against a different sample size-parameter set combination. Sample sizes vary across rows; parameter sets Θ_j vary across columns. The parameters used for testing are listed in the online appendix. Classify-and-aggregate only performs well (e.g., column 1 of panels 1 and 4) when the composition of the training data matches that of the target data, which does not hold in real-world applications

	Θ_1	Θ_2	Θ_3	Θ_4	Θ_5	Θ_6
Proposed Method vs. Classification: US-Unif.						
N = 2000	9.74	0.60	0.37	0.24	0.24	0.16
N = 4000	9.89	0.37	0.22	0.18	0.16	0.14
N = 8000	9.53	0.23	0.14	0.13	0.12	0.12
N = 16 000	8.89	0.15	0.10	0.10	0.08	0.09
N = 32 000	8.46	0.10	0.07	0.07	0.05	0.07
Proposed Method vs. Classification: US-Pop.						
N = 2000	0.48	0.57	0.85	0.19	1.25	0.19
N = 4000	0.35	0.35	0.50	0.14	0.86	0.16
N = 8000	0.23	0.22	0.32	0.11	0.61	0.13
N = 16,000	0.16	0.14	0.22	0.08	0.42	0.11
N = 32,000	0.10	0.09	0.15	0.06	0.28	0.08
Proposed Weighted vs. Unweighted: US						
N = 2000	0.97	0.87	0.79	0.92	0.85	0.88
N = 4000	0.95	0.84	0.76	0.89	0.84	0.86
N = 8000	0.90	0.80	0.79	0.87	0.84	0.87
N = 16,000	0.84	0.77	0.82	0.86	0.81	0.87
N = 32,000	0.80	0.76	0.85	0.79	0.78	0.85
Proposed Method vs. Classification: Kenya						
N = 2000	6.07	0.65	0.61	0.40	0.50	0.42
N = 4000	4.64	0.42	0.38	0.25	0.31	0.27
N = 8000	3.68	0.28	0.25	0.17	0.21	0.18
N = 16,000	3.41	0.18	0.16	0.12	0.15	0.13
N = 32,000	3.28	0.12	0.10	0.08	0.11	0.09

target parameter values and sample sizes. Similar results hold for Kenyan data, as shown in Table A8 in the online appendix.

3 Monte Carlo Simulations

A key problem with classify-and-aggregate approaches is their dependence on the composition of the training data. Suppose the training data contain a disproportionate number of people from group *j*. As a result, name *i* will have more chances to appear in group *j*, potentially inflating *P*(Group *j*|Name *i*) relative to the population. This is especially problematic in the kinds of contexts where the proposed approach can be most useful: regions where prior information about local group composition is unclear or unavailable.

The Monte Carlo simulations show that the proposed approach outperforms the classify-and-aggregate approach by avoiding this dependence on *P*(Group) in the training data, doing so across a range of parameter values and sample sizes.[6]

[6] Figure 1 in the online appendix provides a simulation showing how the performance of classify-and-aggregate approaches depends on the similarity between the target and training data sets: the more different *P*(Group) between the training and target data sets, the worse the performance of the classify-and-aggregate approach.

The simulations are based on names data from the 2000 United States Census reporting the counts of names by racial groups: White, Black, Asian Pacific Islander, American Indian, Mixed Race, Hispanic.[7] Estimations of racial proportions using the proposed method are compared with the aggregations of classifications using $P(\text{Group} \mid \text{Name})$ based on uniform population parameters.

The size of the target list of names ($N = 2000, 4000, 8000, 16000, 32000$) and the underlying population parameters used to generate the training data are varied to test the methods across a range of conditions. Tables A1 and A2 in the online appendix describes the thirty sample size/population parameter combinations used to test the relative performance of the estimators. The performance metric is MAE: the average absolute difference between the parameter estimates and the true parameters from which the training data are simulated. For economy of presentation, MAE ratios are presented in the body of the paper; raw results are included in the online appendix for both the U.S. and Kenyan data. A ratio below one indicates that the proposed estimator outperforms the classify-and-aggregate estimator in terms of MAE.

Table 1 presents the ratio of the average MAE for the proposed estimator to the average MAE of the alternative estimator. Three results stand out. First, the first column of the first and fourth panels of Table 1 suggests that the proposed estimator performs poorly relative to the standard approach. This is due to the fact that the parameters assumed in calculating $P(\text{Group}|\text{Name})$ are identical to the parameters generating the target data in column 1.[8] This illustrates one key drawback of the classify-and-aggregate approach: the parameters underlying the training data acutely influence the conditional used for classification.[9] Not surprisingly, when those parameters are—by chance—equal to the estimand, classify-and-aggregate performs well.

Second, the proposed estimator outperforms the alternative under all other combinations, with the exception of one combination (parameter set 5, $N = 2000$) in panel 2 of Table 1 where the proportion of White individuals in the target data mirrors the proportion White in the training data. Third, increasing the sample size improves the performance of the proposed estimator but not the classify-and-aggregate approach.[10]

Fourth, panel 3 of Table 1 suggests that the proposed bias reduction technique works well, providing improvements over the standard version of the proposed approach. This implies that other regression-based approaches to bias-reduction may provide further performance gains, though we leave that for future research. The bottom panel of Table 1 presents results similar to those of the U.S. data in panels 1 and 2 for simulations based on Kenyan data. Full Monte Carlo results for both the U.S. and Kenyan names are available in the online appendix.

A key reason that classify-and-aggregate does not work well with names lies in its fundamental reliance on $P(Group)$ represented in the training data, as illustrated by comparing the first columns of panels 1 and 4 in Table 1. The only difference between these two columns is the $P(Group)$ assumed in the training data—uniform for the top panel and U.S. population parameters in the second panel. The classify-and-aggregate approach appears to work relatively well in column 1 of Table 1 only because $P(Group)$ in the training data matches $P(Group)$ in the target data.

The Monte Carlo results demonstrate that the proposed method outperforms traditional classify-and-aggregate approaches across a broad range of parameter and sample-size values. Section 3 in the online appendix provides evidence that the proposed method works well on real-world data from counties in North Carolina and Florida.

4 Collection of Training Data

Training data collection differs from other kinds of texts. Hopkins and King (2010) deal with the large number of words in a corpus by classifying whole texts, and then decomposing texts

[7] These data are available at http://www2.census.gov/topics/genealogy/2000surnames/names.zip and discussed in Word et al. (nda, ndb).
[8] Identical evidence for Kenyan data is presented in Table A7 in the online appendix.
[9] This problem is well known in machine-learning and classification literatures (He and Garcia 2009; Sun, Wong, and Kamel 2009).
[10] Full tables of MC results in the appendix provide direct evidence of this result (Harris 2014).

into their constituent words to calculate P(Words|Category) to comprise the training data. As a result, the collection of training data proceeds like most categorization tasks: assign a given text to a group. This research has used two distinct approaches to compile training data, both of which have proved effective in practice. For the United States, training data were easily accessible from the Census Bureau. For the Kenyan data discussed below, names from ethnically homogeneous regions of the country were used to build the training data for each group. For example, suppose we know from prior contextual knowledge or available data that a vast majority of the population in a given region is from group j. Then, we use names from that region to form the training data for group j. Inadvertent inclusion of a small proportion of people that belong to group $-j$ does little to affect the training data estimates, given the relative size of group j in the training data.

5 Application: Racial Turnout in North Carolina

In this section, the proposed approach is applied to names from the North Carolina voter register to make inferences about racial voting behavior in North Carolina. The R × C ecological inference method described in Greiner and Quinn (2009) is used to focus on the empirical problem of estimating Black turnout in each of North Carolina's 100 counties for the 2012 presidential election. Ecological inference requires two kinds of data to estimate turnout-by-race: electoral outcomes and racial proportions. The latter is estimated using the standard and proposed approaches. The proposed method leads to better estimates of Black turnout than estimates based on the alternative approach to racial proportions.

Ecological inference has an important place in voting rights litigation, affecting states' electoral law and redistricting processes (Greiner 2007). As a result of the Voting Rights Act of 1965, legislation aimed at addressing persistent political discrimination of (mainly Black) minorities in the U.S. South, jurists sought ways to demonstrate the presence of racial disparities in turnout or registration. Because individual ballots are cast secretly, it cannot be directly observed whether one racial group votes differently than another. This application demonstrates one potential use of the proposed method for estimating racial proportions, and examines how it performs relative to the classify-and-aggregate approach.

Because North Carolina is partially covered by the Voting Rights Act, it collects comprehensive data on registration and turnout by race. These data allow *estimates* of Black turnout using name-based racial proportions to be compared with *actual* Black turnout percentages at the county level. In this way, whether the reduction in error provided by the proposed approach has an appreciable effect on estimates of black turnout can be tested.[11]

First, I gathered precinct-level data on *actual* Black registration and turnout in the 2012 presidential election from North Carolina 2012 voter history files, and calculated the *actual* percent turnout for Black voters for each county as the number of Black voters who voted divided by the total number of Black voters.[12] This measure of Black turnout was used as the "true" Black turnout in a given county, which was then estimated using ecological inference. Next, the precinct-level turnout and abstention counts required for ecological inference were gathered.[13] These precinct-level election results were complemented with precinct-level estimates of the number of Black, White, and Hispanic voters using voters' names with both the classify-and-aggregate approach

[11] In North Carolina and other jurisdictions covered by the Voting Rights Act, the collection of racial data on registration and turnout may end if such data become legally unnecessary, given the invalidation of the coverage formula of section 4 of the VRA. In the absence of another data source, the study of racial electoral behavior would be much more difficult. An anonymous reviewer helpfully pointed out that jurisdictions targeted by the "bail-in" mechanism in section 3 of the VRA may be another set of cases where names, but not racial data, are available.

[12] The November 2012 county-level voter registration files were acquired in electronic form from the North Carolina State Board of Elections, removing records with status "removed" or "denied" and maintaining those marked "active," "inactive," or "temporary."

[13] 2012 presidential election results and racial turnout and registration data by precinct are available at ftp://alt.ncsbe.gov/ENRS/. (NCSBOE 2012a, NCSBOE 2012b).

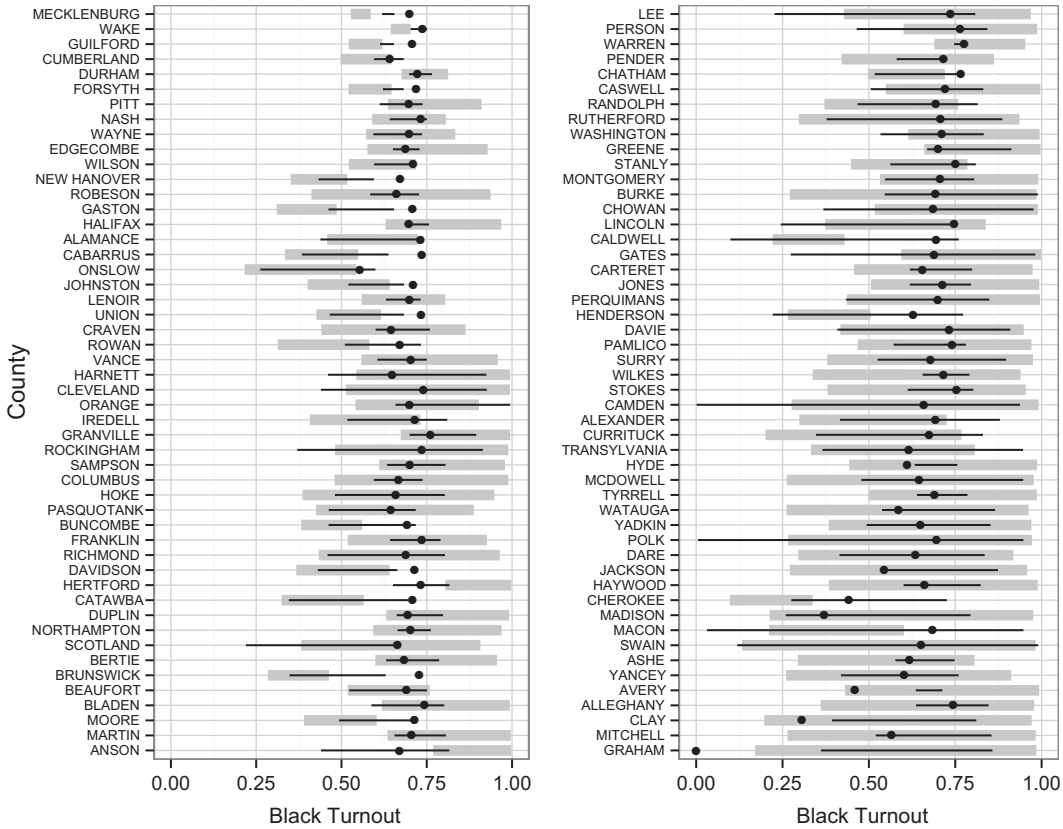

Fig. 2 Proposed approach enables more accurate and efficient inference of Black turnout using ecological inference: Actual Black turnout (black dots) for each of North Carolina's 100 counties displayed with 95% credible intervals for Black turnout estimated using data on racial composition from the improved approach (black lines) and the classify-and-aggregate approach (gray bars). Counties are ordered from top-left to bottom-right, decreasing in numbers of Black voters in each county. Error bars for both methods widen moving toward the bottom right, as those counties have both fewer Black voters (increasing measurement error in the group proportions for both approaches) and fewer precincts (leading to higher variance estimates of Black turnout). Both approaches—not to mention EI—perform poorly in counties like Graham (bottom right), where very few Blacks are registered to vote and Black turnout is low. In counties with more Blacks, the proposed method generates tighter credible intervals centered closer to the true value of Black turnout.

and the improved approach. Finally, I estimated county-level Black turnout using both sets of racial proportions via the ecological inference method in Greiner and Quinn (2009).[14]

Results suggest that the improved approach enables more accurate and more efficient results than the classify-and-aggregate approach. Figure 2 compares actual Black turnout (black dots) for each of North Carolina's 100 counties with the 95% credible intervals estimated using the improved method (black lines) and the classify-and-aggregate method (gray bars). How do the two different sets of racial estimates perform in recovering actual Black turnout? The credible interval from the improved approach's estimates contains actual Black turnout in 80 of 100 counties, compared with 75 of 100 using the classify-and-aggregate estimates. In the 20 counties where the credible intervals for the improved approach did not contain actual Black turnout, the median absolute difference between the point estimate and actual Black turnout is 15.7%; for the classify-and-aggregate approach, the difference is 23.5%. There is also evidence of attenuation in the point estimates of

[14] For each county, I ran one chain of 5 million samples, discarding the first 3 million as burn-in. Every 500th draw of the remaining two million draws was saved, resulting in 4000 draws for inference.

the classify-and-aggregate approach, which consistently underestimates the actual Black turnout. The mean difference between the point estimate and actual Black turnout for classify-and-aggregate is −17.3%, while it is only −3.9% for the proposed approach.

Credible intervals from the improved approach tend to be smaller than those from the classify-and-aggregate approach.[15] In 70 of 100 counties, credible intervals from both approaches covered actual Black turnout. On average, the intervals using the improved approach were 36% smaller than those using the classify-and-aggregate approach. The intervals from the improved approach were larger than those from the classify-and-aggregate approach in only 11 of those 70 counties.[16]

6 Application: Ethnic Displacement in Kenya

In this section, the proposed method is used to examine changes in voter registration by ethnic group between 2007 and 2010 in the 138 polling stations of Kuresoi constituency, Kenya. Following the 2007 general elections, Kenya's Rift Valley exploded in violence, and hundreds of thousands of individuals were displaced from their homes. Consistent with past violence, this outbreak—commonly referred to as the "Post-Election Violence" or PEV—is often construed as political violence designed to "gerrymander by moving people" (Anderson and Lochery 2008; Klopp and Kamungi 2008; Waki 2008). Do the data accord with these descriptions of the violence? Did the ethnic composition of the register actually change? To date, little systematic evidence has been brought to address these questions.

Ideal data to address these questions would enumerate the population prior to and after the violence, collecting information on where people live and their ethnic identity. In this case, census data are not a good temporal fit, since enumeration takes place every ten years. Changes over that ten-year period may be attributable to other factors during the period, such as differential population growth rates or labor migration.

More to the point, our interest lies in the political impact of violence on registered voters, not the whole population. Thus, working directly with voter registration records—and the names therein—makes sense. Even though voter registration records contain no direct information on the ethnic identity of voters, the voters' names can be used to make inferences about the ethnic composition of the register and how it changes across time.

Three ethnic groups in Kuresoi constituency in the central Rift Valley are focused upon: the Kalenjin, the group alleged to have perpetrated the violence; the Kikuyu, thought to be the primary targets; and the Kisii, another targeted group. Evidence taken by the Commission of Inquiry on Post Election Violence—an investigative panel established to investigate the circumstances surrounding the PEV—notes that the constituency has a history of "ethnic conflict pitting the Kalenjin on the one hand and the Kikuyu and Kisii on the other," tracing the roots of the 2007 violence to similar events in 1992 and 1997 (Waki 2008, 80). The organization of the violence in places like Kuresoi in 2007, however, was unprecedented in Kenyan history. For instance, (Waki 2008, 85) relates an instance where a "politician and a civic candidate were responsible for organizing and financing Kalenjin youth who were reportedly paid Kshs. 1,000 [about $11] for every house razed down belonging to the Kisii and Kikuyu in Kuresoi." It was this level of organization across much of Kenya's Rift Valley that led to charges against several high-level politicians at the International Criminal Court (Brown and Sriram 2012; Mueller 2014).

The hypotheses reflect the qualitative evidence cited above, which suggests that different ethnic groups were affected by the violence in different ways. Increases in the Kalenjin share of the voter register and a concomitant decrease in Kikuyu and Kisii registration are expected. To test these

[15] The width of the CIs is, in part, driven by the number of precincts in the county. The focus here is to hold data and methodology constant in order to evaluate the relative performance of estimates from the proposed estimator and the existing approach.

[16] Both approaches perform poorly in cases like Graham County, where Blacks comprised 0.3% of the population (relative to 22% for the state) according to 2012 estimates based on the 2010 census. Table A12 in the online appendix provides full information on the proportion and absolute number of Black voters in North Carolina, and shows very few Black registered voters in Graham county. Counties with few Blacks also tend to be rural and have few precincts, further increasing uncertainty.

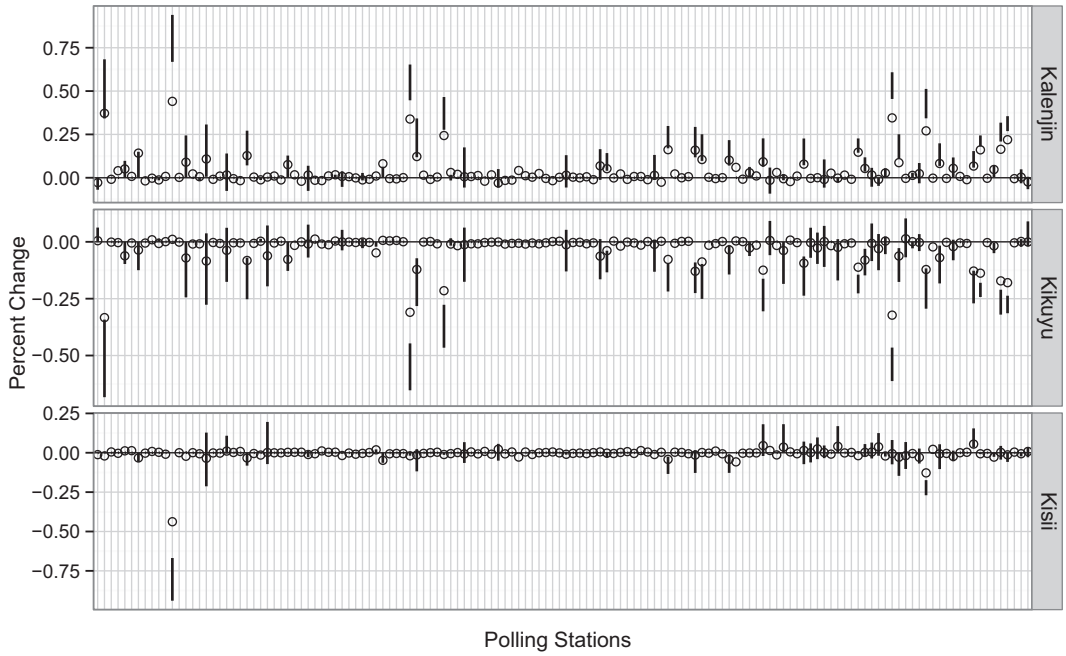

Fig. 3 Changes in the ethnic composition of the voter register, Kuresoi constituency, Kenya: The polling-station level findings show large significant decreases in Kisii (bottom panel) and Kikuyu (middle panel) registration and significant increases for the Kalenjin (top panel). Black lines represent the 99% bootstrapped confidence intervals from the new method and circles show the estimate using a dictionary-based approach. These results are consistent with narrative evidence that "outsider" ethnic groups like the Kisii and Kikuyu were forcefully ejected from Kuresoi constituency during the 2007 Post-Election Violence in Kenya. Results suggest that estimates from the dictionary-based approach are attenuated relative to those from the new approach.

hypotheses, the change in the ethnic composition of the voter registers at each of the 138 polling stations in Kuresoi constituency was estimated using both the proposed method with bias-reduction weighting and a name dictionary. The same basic training data were used for both the proposed method and the name dictionary (ECK 2007, IIEC 2010). For the proposed method, 1000 bootstrapped estimates of the difference in ethnic composition in 2007—before the violence—and 2010—after the violence—for each polling stations were generated. For the name dictionary, each name was linked to its most likely ethnic group.

Figure 3 presents the change in ethnic proportions on the voter register between 2007 and 2010 for the Kalenjin (top panel), Kikuyu (middle panel), and Kisii (bottom panel). Black lines represent the 99% bootstrapped confidence intervals from the new method, and circles show the estimate using a dictionary-based approach. The results largely confirm the qualitative evidence. In most polling stations where the composition changed, increases in Kalenjin correspond to decreases in *either* the Kikuyu *or* the Kisii, but rarely both. This is due to the segregated nature of ethnic groups, which often tend to cluster spatially, leading to relatively homogeneous polling station–level ethnic composition: Kalenjin, Kikuyu, and Kisii tend not to register to vote at the same polling station.[17] Another consequence of segregation is that it makes violent targeting easier.

Figure 3 allows estimates from the new method to be compared with those from the dictionary-based approach. Several interesting differences arise. First, dictionary-based approaches provide a point estimate, but no measure of uncertainty. Since the results are ordered by polling station size in 2007 from smallest (left) to largest (right), the behavior of the bootstrapped confidence intervals

[17] For some polling stations, both methods produce efficiently estimated changes of zero. This too is an artifact of the ethnic segregation.

can be evaluated. As expected, they tend to narrow moving from smaller polling stations—which have fewer unique names, smaller data matrices, and thus higher uncertainty—to larger polling stations. Second, virtually all of the dictionary-based estimates (open circles) are closer to zero than estimates from the proposed method. One explanation of this pattern is attenuation: as dictionaries deterministically link names to one ethnic group, they necessarily misclassify some names. Misclassification is a type of measurement error, which may cause the attenuation seen in Figure 3.

These results are the first micro-level quantitative estimates demonstrating how the PEV reshaped the ethnic composition of Kenya's voter register. Their consonance with existing narrative evidence provides some validation of the method in a context where, unlike the U.S. applications above, no direct validation data are available. In this way, the proposed method allows us to do something previously impossible: make valid inferences about ethnic change at the micro-level in rural Africa. We should note, however, that these results should not be interpreted as causal, or even to isolate decreases occurring as a consequence of violence occurring after the election. Other factors—ethnic differences in migration, mortality, and pre-election intimidation, to name but a few—surely all contribute to the results seen here. Future work will attempt to disentangle these and other factors.

7 Discussion

This article has presented a new approach for extracting information about ethnicity from names. Building on King and Lu (2008) and Hopkins and King (2010), the proposed method yields more efficient estimates than those based on individual classification and provides corrections to improve inferences when discrepancies between target and training data exist. Moreover, the approach avoids the labor-intensive compilation of a large dictionary of names and recognizes that names often do not fit into discrete, mutually exclusive categories.

The method is ripe for applications where names contain information about group identity or membership. While this article has focused on race and ethnicity, other potential cleavages (e.g., class or caste) could be estimated as long as naming conventions vary across these cleavages. More generally, the approach could be adapted to other problems with similar data types where category proportions are the quantity of interest.[18] Of course, the proposed method (along with classify-and-aggregate approaches) will not be applicable in cases where names contain little or no information about group identity. This might occur in a society where very few unique names exist (making inference difficult due to small sample size), or when most people, regardless of identity group, have similar names. For instance, in Kenya's Northeastern Province, many people are Muslim, and have adopted Islamic or Arabic names. Though there are many different ethnic groups in that region, names do a relatively poor job of differentiating those groups, because individuals' names come from a common religious source.

These advantages notwithstanding, using the method requires the researcher to make several informed decisions. First, choosing which identity groups to analyze is a compromise between the theoretical question of interest and the limits of data and method. For instance, Kenya has over forty distinct ethnic groups, some of which have virtually identical name profiles due to common religious heritage, and many of which are rarely found outside their "home" areas. Thus, the researcher must use contextual knowledge to focus estimation to relevant ethnic groups. Prior contextual knowledge of the politically relevant groups during the period in question enabled the decision of which parameters to estimate. This is akin to choices in the ecological inference literature to focus on larger groups to the exclusion of some smaller communities.

Another important choice centers on which names to use. Surnames were used in the North Carolina application, given that the census training data only includes surnames. In the Kenya

[18] Further research might examine ways to improve performance of the approach. For instance, one improvement might break estimation down into further subgroups (e.g., age groups or gender) to improve inference. This would require additional training data—conditional distributions for each subgroup (e.g., $P(\text{Name}|\text{Group} = \text{A and Gender} = \text{F})$). However, if the name-conditional distribution across, say, gender varies substantially across groups, then improvements could be significant. Thanks to an anonymous reviewer for suggesting this line of inquiry.

application, all available names were used. This improved inferences by enlarging the size of the target and training data and increasing overlap between training and target data. In other contexts, it may make sense to identify parts of names that signify identity and use only those stems in estimation. However, simplification of names does have downsides. Early experiments with estimating ethnic composition using name-substrings (e.g., *n*-grams), letter frequencies, or phonetic representations (e.g., the Soundex algorithm) provided little traction on the problem.

As with any statistical method, researchers must be cognizant of assumptions underlying inference. In this case, *P*(Name | Group) is assumed to be the same in the training and target data. As discussed above, this assumption may not always hold. Model fit statistics provide one way to assess whether the training data and estimated coefficients fit the target data well, but this approach must be complemented with contextual knowledge about the specific populations and data under examination. Very rarely do we know nothing about the places we study; thus, researchers should compare estimates with existing knowledge whenever possible to evaluate the veracity of estimates.

The two applications presented here suggest that the proposed method could work in many different contexts, as long as training data exist and names communicate information about group identity. Many other applications exist.[19] The proposed approach can also be seen as complementary to classification-based approaches, and may be useful when researchers want individual name classifications, but have little knowledge of local group composition. The method described here is a principled way to generate a prior distribution from which to improve individual name classifications. Both Enos (2012, 2014) and Grofman and Garcia (2014) provide avenues for improving classification-based approaches, though that work focuses on the United States, where high-quality demographic data from the census provide a strong prior upon which to base classifications. In much of the rest of the world, such detailed, high-resolution demographic data are unavailable. The proposed method provides a way to produce such estimates when names data exist.

Funding

The author acknowledges funding from an NSF Doctoral Dissertation Improvement Grant and Harvard University Fredrick Sheldon Traveling Fellowship during early stages of this research.

References

Ambekar, A., C. Ward, J. Mohammed, S. Male, and S. Skiena. 2009. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 49–58. KDD '09, New York, NY, USA: ACM.

Anderson, D., and E. Lochery. 2008. Violence and exodus in Kenya's Rift Valley, 2008: Predictable and preventable. *Journal of Eastern African Studies* 2(2):328–43.

Brown, S., and C. L. Sriram. 2012. The big fish won't fry themselves: Criminal accountability for post-election violence in Kenya. *African Affairs* 111(443):244–60.

Byrne, K., and E. O'Malley. 2012. What's in a name? Using surnames as data for party research. *Party Politics* 19(6):985–97.

Coldman, A. J., T. Braun, and R. P. Gallagher. 1988. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health* 42:390–95.

Cook, R. D. 1977. Detection of influential observation in linear regression. *Technometrics* 19(1):15–18.

Electoral Commission of Kenya. October 2007. Register of electors. Kuresoi Constituency.

Enos, R. D. 2011. *What tearing down public housing projects teaches us about the effect of racial threat on political participation*. Working Paper, Department of Government, Harvard University.

———. 2012. *Testing the elusive: A field experiment on racial threat*. Working Paper, Harvard University.

———. 2015. Forthcoming. What the demolition of public housing teaches us about the impact of racial threat on political behavior. *American Journal of Political Science*.

Goldfarb, D., and A. Idnani. 1982. Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*, ed. J. P. Hennart, 226–39. Berlin: Springer-Verlag.

———. 1983. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* 27:1–33.

[19] The Indian and Philipines governments print government gazettes with the names of appointees. See http://www.gov.ph/section/appointments-designations/ and http://www.egazette.nic.in/.

Greiner, D. J. 2007. Ecological inference in voting rights act disputes: Where are we now, and where do we want to be? *Jurimetrics* 47:115–67.
Greiner, D. J., and K. M. Quinn. 2009. R x C ecological inference: Bounds, correlations, flexibility, and transparency of assumptions. *Journal of the Royal Statistical Society, Series A* 172(1):67–81.
Grofman, B., and J. Garcia. 2014. Using Spanish surname to estimate Hispanic voting population in voting rights litigation: A model of context effects. *Election Law Journal* 13(3):375–93.
Harris, J. A. 2014. Replication data for: What's in a name? A method for extracting information about ethnicity from names. Dataverse Network, doi:10.7910/DVN/27691 (v1).
He, H., and E. Garcia. 2009. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions* 21(9):1263–84.
Hopkins, D. J. 2010. Politicized places: Explaining where and when immigrants provoke local opposition. *American Political Science Review* 104(1):40–60.
Hopkins, D., and G. King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1):229–47.
Interim Independent Electoral Commission. July 2010. Voter's register. Kuresoi Constituency.
Kasara, K. 2013. Separate and suspicious: Local social and political context and ethnic tolerance in Kenya. *Journal of Politics* 75(4):921–36.
King, G., and Y. Lu. 2008. Verbal autopsy methods with multiple causes of death. *Statistical Science* 23(1):78–91.
Klopp, J., and P. Kamungi. 2007. Violence and elections: Will Kenya collapse? *World Policy Journal* 24(4):11–18.
Mateos, P. 2007. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space, and Place* 13(4):243–63.
———. 2011. Uncertain segregation: The challenge of defining and measuring ethnicity in segregation studies. *Built Environment* 37(2):226–38.
Mueller, S. D. 2014. Kenya and the International Criminal Court: Politics, the election, and the law. *Journal of Eastern African Studies* 8(1):25–42.
NCSBOE. 2012a Voter statistics file.
———. 2012b Voting history file.
Rosenwaike, I. 1994. Surname analysis as a means of estimating minority elderly: An application using Asian surnames. *Research on Aging* 16(2):212–27.
Sun, Y., A. K. C. Wong, and M. S. Kamel. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23(4):687–719.
Susewind, R. 2015. What's in a name? Probabilistic inference of religious community from South Asian names. *Field Methods* 27(3):1–14.
Treeratpituk, P., and C. L. Giles. 2012. Name-ethnicity classification and ethnicity-sensitive name matching. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.
Turlach, B. A., and A. Weingessel. 2013. `quadprog`: Functions to solve quadratic programming problems.
UNSD. Ethnicity: A review of data collection and dissemination. Technical report, United Nations Statistics Division, Demographic and Social Statistics Branch, Social and Housing Statistics Section.
Waki, J. P. 2008. *Report of the Commission of Inquiry into Post Election Violence*. Nairobi, Kenya: Government Printer.
Word, D., C. Coleman, R. Nunziata, and R. Kominski n.d.a. Data accompanying "Demographic Aspects of Surnames from Census 2000." U.S. Census Bureau, Washington, DC.
———. n.d.b. Demographic aspects of surnames from Census 2000. Technical report, U.S. Census Bureau, Washington, DC.