

## EXIT POLLING AND RACIAL BLOC VOTING: COMBINING INDIVIDUAL-LEVEL AND $R \times C$ ECOLOGICAL DATA

BY D. JAMES GREINER<sup>†</sup> AND KEVIN M. QUINN<sup>‡</sup>

*Harvard Law School<sup>†</sup> and UC Berkeley School of Law<sup>‡</sup>*

Despite its shortcomings, cross-level or ecological inference remains a necessary part of some areas of quantitative inference, including in United States voting rights litigation. Ecological inference suffers from a lack of identification that, most agree, is best addressed by incorporating individual-level data into the model. In this paper, we test the limits of such an incorporation by attempting it in the context of drawing inferences about racial voting patterns using a combination of an exit poll and precinct-level ecological data; accurate information about racial voting patterns is needed to assess triggers in voting rights laws that can determine the composition of United States legislative bodies. Specifically, we extend and study a hybrid model that addresses two-way tables of arbitrary dimension. We apply the hybrid model to an exit poll we administered in the City of Boston in 2008. Using the resulting data as well as simulation, we compare the performance of a pure ecological estimator, pure survey estimators using various sampling schemes, and our hybrid. We conclude that the hybrid estimator offers substantial benefits by enabling substantive inferences about voting patterns not practicably available without its use.

---

\*The authors extend great thanks to Jayanta Sircar at the Harvard School of Engineering for allowing us to use the Harvard Crimson Grid gratis, and to Robert Parrot for his assistance in teaching us how to use the grid; to Gary King, Lewis Kaplow, Jeffrey Lewis, Andrew Martin, and Kathy Spier for helpful comments; and to the JEHT Foundation, the Rappaport Institute for Greater Boston, the Boston Foundation, Sheldon Seevak Fund, and the Harvard University Office of the Provost for financial support. In addition, this paper benefited greatly from feedback received during presentations at the Applied Statistics Workshop at the Institute for Quantitative Social Science at Harvard University; at the Law, Economics, and Organization seminar at Harvard Law School; and at the 2009 Conference on Empirical Legal Studies.

*AMS 2000 subject classifications:* Primary 46N30, 46N30; secondary 46N30

*Keywords and phrases:* ecological inference, Bayesian inference, voting rights litigation, exit polls, survey sampling

Cross-level or ecological inference is the attempt to draw conclusions about statistical relationships at one level from data aggregated to a higher level. Frequently, ecological inference is conceptualized as the attempt to infer individual-level relationships from a set of contingency tables when only the row and column totals are observed. One important application of ecological inference is in United States redistricting litigation, in which a critical issue is whether the voting patterns of racial groups differ. Because the secret ballot prevents direct observation of voter races and voter choices, redistricting litigants and their experts are ordinarily required to attempt to infer racial voting patterns by examining election returns (reported at the precinct or perhaps the “vote tabulation district” level) as married to demographic information from the Decennial Census. In this paper, we explore issues associated with incorporating individual level information, in the form of responses to an exit poll we administered in the City of Boston, into an  $R \times C$  ecological model.

Speaking broadly, the lack of identification in ecological models was famously discussed in [Robinson \(1950\)](#). Since then, most to consider the question have agreed that, if ecological inference is to be attempted, the best way to proceed is to incorporate additional, preferably individual-level (from a survey), information into the model. In the sampling literature, combining survey results with population-level information has a long and rich history, dating back at least to [Deming and Stephan \(1940\)](#). [Bishop, Fienberg and Holland \(1975\)](#) (see pp. 97-102) discuss what they call the “classical” use of iterative proportional fitting, also called “raking” to known marginal totals or “incomplete post-stratification” (see [Deville, Sarndal and Sautory \(1993\)](#)). Additional examples include [Belin \*et al.\* \(1993\)](#), [Little \(1993\)](#), and [Zaslavsky \(1993\)](#). As relevant to this paper, the idea is that two categorical measurements are made on each of  $K$  in-sample units so that estimated counts are generated for each of the cells of an  $R \times C$  two-way contingency table. These estimated internal cell counts are adjusted so as to conform to row and column sums known from another source. The quirk in ecological data is that there are many such contingency tables (precincts, in our application), and as to some of those, the row sums disclose that one category so dominates the table count as to make bounding information (see [Duncan and Davis \(1953\)](#)) informative there. That in turn renders sampling in such single-category-dominated precincts a potential waste of resources. (Were the column sums known in advance, the same principle might apply, but in our application the columns sums are vote totals that are unknown in advance.) We explore such issues of sample allocation in this paper.

Meanwhile, on the social science side, the past decade or so has seen

several papers (Steel, Tranmer and Holt (2003), Raghunathan, Diehr and Cheadle (2003), Glynn *et al.* (2008), Haneuse and Wakefield (2008), Glynn *et al.* (2009)) addressing how best to combine ecological data with limited individual-level information. As is true in ecological inference more generally, most papers addressing incorporation of additional information into ecological data have focused on sets of  $2 \times 2$  contingency tables, which (after conditioning on the row and column totals) involve one missing quantity per table.

In this paper, we address the  $R \times C$  case,<sup>1</sup> building on earlier work of our own (Greiner and Quinn (2009)), which in turn built on Brown and Payne (1986) and Wakefield (2004). We do so because of the importance of  $R \times C$  ecological inference to many fields of inquiry; a particular interest of ours is in United States voting rights litigation. Sections 2 and 5 of the Voting Rights Act prevent dilution and retrogression of the voting strength of racial (“racial” means racial or ethnic) minorities via gerrymandering of districts. In both settings, proof that members of different racial groups vote similarly within-group and differently between-group constitutes part of what is called “racially polarized” or “racial bloc” voting, which is the “keystone” to litigation (11th Cir. (1984)) and the “undisputed and unchallenged center” (Issacharoff (1992)) to the area of law. This law can in turn decide the composition of Congress as well as of local legislative bodies (Lublin (1995)). Thus, the need for accurate information regarding the voting preferences of different racial groups is acute. As mentioned above, because the secret ballot prevents direct observation of voter decisions, proof of racial bloc voting is most frequently made via  $R \times C$  ecological inference methods, as follows. The Census provides voting-age-population figures for each racial group, which are arranged along the rows of contingency tables. Table columns are official vote counts for each candidate (Democrat, Republican, etc.), along with an additional “Abstain” column to account for persons declining to exercise the franchise. There is one such contingency table for each voting precinct, and the goal of the inference is to calculate, say, the percentage of Hispanic voters who voted for the Democrat (see below for more technical definitions). That requires filling in the missing internal cell counts of the contingency tables, subject to the constraints imposed by the row and column totals.

The need for better and more accurate techniques in this area has grown in recent years. As the number of relevant racial and ethnic groups in the United States polity increases (from, say, black versus white to include His-

---

<sup>1</sup>Software to implement the methods we propose, including those used in this paper, is available via the R package “RxCcolInf”; access CRAN from <http://www.r-project.org/>.

panics and Asians), inference becomes more complicated. Additional races represent additional rows in the contingency tables, requiring more parameters in a model and imposing greater challenges at the model-fitting stage (Greiner (2007)). Incorporating individual-level information from a survey into the  $R \times C$  ecological inference model represents one promising avenue in this area.

We accordingly subject the task of combining individual-level and  $R \times C$  ecological data to a stress test in the form of an effort to draw inferences about the voting behavior of  $R$  racial groups using data aggregated to the level of the precinct together with an exit poll in which not all precincts were in-sample. Specifically, we discuss the challenges, choices, and results of a 400-pollster, 11-university, 39-polling-place exit poll we administered in the City of Boston on the November 4, 2008 election. Combining ecological data with an exit poll constitutes a stress test for a hybrid model because (i) the nature of exit polling prevents us from implementing optimal subsampling techniques recently explored in the literature, (ii) survey nonresponse is ever-present, and (iii) the fact that several precincts may be combined within a single voting location requires additional assumptions regarding the aggregation process, as we explain below. In our view, our hybrid model passes this stress test by supporting substantive conclusions, particularly regarding voting behavior of hard-to-estimate groups such as Asian- and Hispanic-Americans, that could not be reached without its use (all of this assuming the reasonableness of the model).

We organize this paper as follows: we clarify notation before presenting a brief taxonomy of  $R \times C$  ecological techniques that focuses on the advantages and disadvantages of fraction versus count models. We articulate the details of our hybrid ecological/survey proposal and use simulation to study its behavior, focusing in particular on its performance in the presence of aggregation bias, defined immediately below. On the basis of these simulations, we offer guidance for practitioners confronted with a choice of three classes of estimators: an ecological model alone, a survey sample alone, and a hybrid. We demonstrate that (i) the hybrid is always preferable to the ecological model; (ii) in the absence of severe aggregation bias, the hybrid dominates the survey sample estimator; (iii) in the presence of severe aggregation bias, the hybrid is still probably preferable, although the researcher's choice of estimator depends on, among other things, whether the contingency tables tend to be dominated by one row (in voting applications, this corresponds to a high level of housing segregation), and whether interest lies primarily in the point estimate or valid intervals.

We then present the process leading to and the results of our City of

Boston exit poll, focusing on voting behavior by race in a Massachusetts ballot initiatives regarding marijuana (other results from this exit poll are available from the authors on request). We demonstrate that our hybrid estimator allows inferences unavailable from either the exit poll or the ecological inference model alone. Without the hybrid estimator, for example, little can be said regarding Asian-American voting preferences in Boston, nor can one easily distinguish between Hispanic and white preferences. We also find little evidence of aggregation bias in the Boston data.

Regarding the definition of aggregation bias, the critical assumption of most ecological inference techniques is the absence of contextual effects. Contextual effects can occur when the distribution of the internal cell counts varies with the distribution of the allocation of the counts by row. In voting parlance, if white voting behavior varies with the fraction of whites in the precinct, this contextual effect will cause the aggregation process to induce bias in almost any ecological estimator, unless a covariate/predictor can be included in the model to remove this effect.

Regarding notation, any quantity with the subscript  $rc_i$  refers to that quantity in the  $i$ th contingency table's (precinct's)  $r$ th row,  $c$ th column. In our application,  $r$  can be  $b$  for black,  $w$  for white,  $h$  for Hispanic, or  $a$  for Asian;  $c$  can be  $D$  for Democrat,  $R$  for Republican, or  $A$  for Abstain (meaning choosing not to vote).  $N$ 's,  $M$ 's, and  $K$ 's refer to counts, as follows:  $N$ 's are the unobserved, true internal cell counts;  $K$ 's are the counts as observed in the survey; and  $M_{rc_i} = N_{rc_i} - K_{rc_i}$ . We italicize unobserved counts but leave observed quantities in ordinary typescript. Table 1 clarifies our representations for the case of  $3 \times 3$  precinct tables involving African-American, Caucasian, and Hispanic groups in a Democrat versus Republican contest.

TABLE 1  
*3 × 3 Table of Voting By Race*

	Dem	Rep	Abstain	
black	$N_{bD_i}$	$N_{bR_i}$	$N_{bA_i}$	$N_{b_i}$
white	$N_{wD_i}$	$N_{wR_i}$	$N_{wA_i}$	$N_{w_i}$
Hispanic	$N_{hD_i}$	$N_{hR_i}$	$N_{hA_i}$	$N_{h_i}$
	$N_{D_i}$	$N_{R_i}$	$N_{A_i}$	$N_i$

We further suppose that a survey or exit poll is implemented in a subset  $S$  of the  $I$  precincts in the jurisdiction and contest of interest. In precinct  $i \in S$ ,  $\mathbf{K}_i$  is a random matrix of dimension  $J_i \times (R \times C)$ , where  $J_i$  is the number of individuals surveyed in this precinct. Each row of  $\mathbf{K}_i$  is a vector of 0s except for a 1 corresponding to the cell of the precinct contingency table in

which the surveyed individual belongs, where the cells are vectorized row major. In the Table 1 example, a vector  $(0, 0, 0, 0, 0, 1, 0, 0, 0)$  would indicate a white person who abstained from voting. Let  $\mathbf{K}$  represent a matrix of all of the  $\mathbf{K}_i$ s (organized in any coherent manner).

Let  $\underline{N}_{\text{row}_i}$  ( $\underline{N}_{\text{col}_i}$ ) represent the vector of observed row (column) totals in the  $i$ th precinct, with  $\mathbf{N}_{\text{row}}$  ( $\mathbf{N}_{\text{col}}$ ) a matrix of all  $\underline{N}_{\text{row}_i}$ 's ( $\underline{N}_{\text{col}_i}$ 's), and  $\mathbf{N}_{\text{obs}} = [\mathbf{N}_{\text{row}} \ \mathbf{N}_{\text{col}}]$ . Let  $\mathbf{N}_{\text{comp}_i}$  equal the (unobserved) full set of internal cell counts in the  $i$ th precinct. Finally, let  $\mathbf{N}_{\text{miss}_i}$  denote any set of  $(R-1)*(C-1)$  counts for the  $i$ th precinct which, had they been observed in conjunction with  $\underline{N}_{\text{row}_i}$  and  $\underline{N}_{\text{col}_i}$ , would have been sufficient to determine all table counts.

In Table 1, for example,  $\mathbf{N}_{\text{miss}_i}$  could equal  $\begin{bmatrix} N_{\text{bD}_i} & N_{\text{bR}_i} \\ N_{\text{wD}_i} & N_{\text{wR}_i} \end{bmatrix}$ . Note that  $\mathbf{N}_{\text{comp}_i}$  and  $\mathbf{N}_{\text{miss}_i}$  are used in the missing data sense (*e.g.*, Little and Rubin (2002)).

Finally, because our interest is primarily in ecological inference as opposed to survey methods, we do not investigate potential biases in surveys or exit polls, except to compare the predictions of our City of Boston exit poll to the observed results. That comparison suggests an encouraging absence of systematic biases, including the absence of a ‘‘Bradley’’ effect for Obama versus McCain, a result in accord with recent findings (Hopkins (2008)). Moreover, while we acknowledge the potential for a variety of sources of bias in ecological studies (see Salway and Wakefield (2005) for a review), we focus our attention on aggregation bias, which we believe to be potentially most problematic in this area (Rivers (1998)).

**1. Fraction Versus Count Models.** We discuss briefly some advantages and disadvantages of modeling unobserved internal cell counts as opposed to the fractions produced when a researcher divides these unobserved counts by their corresponding row totals.

Apart from the approach we advocate, a variety of  $R \times C$  ecological models have been proposed: for example, the unconstrained (see Achen and Shively (1995)) or constrained (Gelman *et al.* (2001)) linear model, the truncated multivariate normal proposal in King (1997), the Dirichlet-based method in Rosen *et al.* (2001), and the information theoretic proposal in Judge, Miller and Cho (2004). These other proposals all share the feature that they model (at various levels), not the internal cell counts themselves, but rather the fractions produced when the unobserved internal cell counts are divided by their row totals. In contrast, we model internal cell counts. There are strengths and weaknesses to each approach.

Formally, let  $\beta$ s refer to the (unobserved) internal cell fractions so  $\beta_{\text{bD}_i} = \frac{N_{\text{bD}_i}}{N_{\text{b}_i}}$ , and  $\underline{\beta}_i$  refer to the vector of the  $\beta$ s in the  $i$ th precinct. If modeling

fractions and proceeding in a Bayesian fashion, a researcher might put a prior on the  $\underline{\beta}_i$ 's with parameter  $\underline{\zeta}$ , in which case one representation of this class of models is as follows:

$$(1) \quad p(\underline{\zeta} | \mathbf{N}_{\text{col}}, \mathbf{N}_{\text{row}}) \propto p(\underline{\zeta}) \prod_{i=1}^I \left[ \int p(\underline{N}_{\text{col}_i} | \underline{\beta}_i, \underline{N}_{\text{row}_i}) \times p(\underline{\beta}_i | \underline{\zeta}) d\underline{\beta}_i \right]$$

For example, in the simplest version of linear model,  $p(\underline{\beta}_i | \underline{\zeta})$  can be conceptualized as a multivariate normal with mean vector  $\underline{\beta}$  and null variance. In [Rosen \*et al.\* \(2001\)](#),  $p(\underline{\zeta})$  is a set of mutually independent univariate gamma distributions,  $p(\underline{\beta}_i | \underline{\zeta})$  a product Dirichlet, and  $p(\underline{N}_{\text{col}_i} | \underline{\beta}_i, \underline{N}_{\text{row}_i})$  a multinomial parameterized by a mixture of  $\beta$ 's and the fractions produced when  $\underline{N}_{\text{row}_i}$  is divided by its sum. Particularly important is the fact that in Equation 1, because there is no distribution posited for the unobserved internal cell counts, there is no summation needed to eliminate them. (Note that throughout this paper, including in Equation 1, we have written the models we fit in terms of posterior distributions for the hyperparameters. We have done so because, as we will explain, interest sometimes centers on these population-level hyperparameters. As a practical matter, the Markov chain Monte Carlo (MCMC) algorithms used to fit these models typically work on the full joint posterior of all model parameters. For more detail on model fitting, see appendix A.2 of [Greiner and Quinn \(2009\)](#).)

In contrast, consider a class of techniques that models the unobserved internal cell counts. A researcher proceeding in a manner analogous to Equation 1 might specify a distribution for each precinct's internal cell counts given some precinct-level intermediate parameters (call these intermediate parameters  $\underline{\Upsilon}_i$ ), might specify a prior on the  $\underline{\Upsilon}_i$ 's (call the parameters in this prior  $\underline{\Xi}$ ), and might sum out the unobserved internal cell counts. Thus, the proportionality corresponding to 1, above, is

$$(2) \quad p(\underline{\Xi} | \mathbf{N}_{\text{col}}, \mathbf{N}_{\text{row}}) \propto p(\underline{\Xi}) \prod_{i=1}^I \left[ \int \sum_{\mathbf{N}_{\text{miss}_i}} p(\underline{N}_{\text{col}_i} | \mathbf{N}_{\text{comp}_i}) \right. \\ \left. \times p(\mathbf{N}_{\text{comp}_i} | \underline{\Upsilon}_i, \underline{N}_{\text{row}_i}) \times p(\underline{\Upsilon}_i | \underline{\Xi}) d\underline{\Upsilon}_i \right]$$

$p(\underline{N}_{\text{col}_i} | \mathbf{N}_{\text{comp}_i})$  appears to make the relationship between the left- and right-hand sides of the  $\propto$  symbol more transparent; in fact,  $\mathbf{N}_{\text{comp}_i}$  determines  $\underline{N}_{\text{col}_i}$ , rendering  $p(\underline{N}_{\text{col}_i} | \mathbf{N}_{\text{comp}_i})$  degenerate. Note in this formulation, there is

an explicit model for the internal cell counts ( $p(\mathbf{N}_{\text{comp}_i} | \Upsilon_i, \mathbb{N}_{\text{row}_i})$ ), which in turn requires a summation over  $\mathbf{N}_{\text{miss}_i}$  to produce the observed-data likelihood. But the distribution of  $\mathbf{N}_{\text{miss}_i}$  is complicated; the permissible support of each element of  $\mathbf{N}_{\text{miss}_i}$  depends on the value of the other elements. Further, in voting applications, the number of voters involved is typically large enough to render infeasible full computation of the posterior probabilities associated with every permissible count.

Thus, Equations 1 and 2 make explicit the benefits of each approach. By avoiding the need for a summation over a complicated discrete distribution, Equation 1 makes fitting easier. This benefit should not be understated. As we will discuss below, the lack of information in ecological data can make model fitting, even via MCMC, slow and cumbersome. The model we advocate requires drawing from two multivariate distributions (one for the internal cell counts, one for  $\Upsilon_i$ ) for each precinct for each of a minimum of several hundred thousand iterations of an overall Gibbs sampler. In contrast, the proposal in Rosen *et al.* (2001), for example, requires only one draw per precinct from a more standard distribution, resulting in substantially less time to analyze a dataset.

The speed gain has tradeoffs. For the purposes of this paper, the primary down side is the lack of an easily conceptualized way of incorporating individual-level information into the model due to the lack of an explicit distribution  $p(\mathbf{N}_{\text{comp}_i} | \Upsilon_i, \mathbb{N}_{\text{row}_i})$ . In contrast to Equation 1, Equation 2 can be modified in a simple way to incorporate data from a sample, as follows:

$$(3) \quad p(\Xi | \mathbf{K}, \mathbf{N}_{\text{col}}, \mathbf{N}_{\text{row}}) \propto p(\Xi) \prod_{i=1}^I \left[ \int \sum_{\mathbf{N}_{\text{miss}_i}} p(\mathbf{K}_i | \mathbf{N}_{\text{comp}_i})^{(i \in S)} \right. \\ \times p(\mathbb{N}_{\text{col}_i} | \mathbf{N}_{\text{comp}_i}) \times p(\mathbf{N}_{\text{comp}_i} | \Upsilon_i, \mathbb{N}_{\text{row}_i}) \\ \left. \times p(\Upsilon_i | \Xi) d\Upsilon_i \right]$$

Additional costs, discussed in Greiner and Quinn (2009), to the approach in Equation 1 are the difficulty in articulating an individual-level (voter) conceptualization of the underlying data-generating process (assuming one is desirable, see King (1997) for a different view) and the fact that most such models weight contingency tables equally regardless of size.

Equation 3 further demonstrate that this formulation allows for any within-contingency-table sampling scheme to be implemented, so long as one can write down  $p(\mathbf{K}_i | \mathbf{N}_{\text{comp}_i})$ . Note, however, that the exchangeability assumption (reflected in the product over  $i$ ) prevents incorporation of contingency-

table-level sample weights into the likelihood. In other words, Equation 3 does not take into account whether the contingency tables in  $\mathcal{S}$  are selected via simple random sampling, sampling in proportion to size, etc. As we explain below, this fact can be a strength or a weakness, but whichever it is, it does not mean that all contingency-table sampling schemes are equally beneficial.

Finally, Equation 3 demonstrates that a variety of choices of likelihoods, priors, and hyperpriors for count models are available. We next discuss our choices.

**2. Our Proposal.** In the language of Equation 3, our proposal consists of the following. For  $\mathbf{N}_{\text{comp}_i} | \mathcal{Y}_i, \mathbf{N}_{\text{row}_i}$ , we assume that the counts in each contingency table row follow an (independent) multinomial distribution with count parameter  $N_{r_i}$  and probability parameter  $\underline{\theta}_{r_i}$ . We choose the multinomial because it corresponds to an individual-level account of voting behavior (each potential voter of race  $r$  in precinct  $i$  independently behaves according to the same vector  $\underline{\theta}_{r_i}$ ) and because once one conditions on the row totals (as is customary in voting applications), few other tractable multivariate count distributions are available.

For  $\mathcal{Y}_i | \Xi$ , we apply a multidimensional additive logistic transformation (Aitchison (2003)) to each row’s  $\underline{\theta}_{r_i}$ , resulting in  $R$  vectors of dimension  $(C-1)$ , which we stack to form a single vector  $\omega_i$  of dimension  $R*(C-1)$  for each precinct. We then assume  $\omega_i \stackrel{i.i.d.}{\sim} N(\underline{\mu}, \underline{\Sigma})$ . We prefer the multidimensional additive logistic to, say, a Dirichlet or a different transformation because of the additive logistic’s greater flexibility relative to the Dirichlet (Aitchison (2003)) and because of the intuitive choice of a “reference category” in voting applications, namely, the Abstain column. The stacking of the transformed  $\underline{\theta}_{r_i}$ ’s into a single vector allows for exploration of within- and between-row relationships; as we demonstrate in our application (see Figure 4), capacity to model between-row relationships can be important to inference.

For the hyperprior ( $p(\Xi)$ ), we use semi-conjugate multivariate normal and inverse Wishart forms, specifically  $\underline{\mu} \sim N(\underline{\mu}_0, \underline{\kappa}_0)$  and  $\underline{\Sigma} \sim InvWish_{\nu_0}(\underline{\Psi}_0)$ . We do so both for computational convenience and because, after extensive simulations, we have found these distributions rich enough to express most reasonable prior beliefs regarding the content of the contingency tables.

For  $p(\mathbf{K}_i | \mathbf{N}_{\text{comp}_i})$ , we assume a simple random sample, out of necessity. Several recent papers (e.g., Glynn *et al.* (2008), Haneuse and Wakefield (2008), and Glynn *et al.* (2009)) have discussed optimal within-contingency-table sampling designs, with the optimal scheme varying according to the process assumed to generate the data and to whether one of the rows or

columns corresponds to a relatively rare event (often true in epidemiology applications). All of these schemes depend on the assumption that the researcher can observe some characteristic of an individual unit before deciding whether to include it in the sample. This is not always possible in exit polls because voters exit polling locations rapidly, and for this reason, exit polls are often interval samples, with the assumption that the interval produces a random sample made plausible by keeping the interval at reasonable length.

If the exit poll constitutes a simple random sample in each  $i \in S$ , we can work with the  $R \times C$ -dimension vector  $\underline{K}_i$  formed by summing  $\mathbf{K}_i$ 's columns; this results in a vector of counts of the number of sampled potential voters in each contingency table cell, with the contingency table vectorized row major. Denote the elements of  $\underline{K}_i$  as  $K_{rci}$ ,  $K_i = \sum_{r,c} K_{rci}$  and for each  $i \in S$ , recall  $M_{rci} = N_{rci} - K_{rci}$ . Accordingly, the probability of observing a

particular vector  $\underline{K}_i$  is the familiar 
$$\frac{\prod_{r,c} \binom{M_{rci} + K_{rci}}{K_{rci}}}{\binom{N_i}{K_i}}$$
 (see [McCullagh and Nelder \(1989\)](#)).

Upon discarding terms for  $i \in S$  that do not involve unobserved quantities, combining terms, canceling, and including the Jacobian of the transformation from  $\theta$  to  $\omega$  space, our proposal has the following observed-data posterior.

(4)

$$\begin{aligned} p(\underline{\mu}, \underline{\Sigma} | \mathbf{K}, \mathbf{N}_{\text{col}}, \mathbf{N}_{\text{row}}) &\propto N(\underline{\mu} | \underline{\mu}_0, \underline{\kappa}_0) \times \text{Inv-Wish}_{\nu_0}(\underline{\Sigma} | \underline{\Psi}_0) \\ &\times \prod_i \left[ \int \left( \sum_{\mathbf{M}_{\text{miss}_i}} \left( \prod_{r,c} \frac{\theta_{rci}^{M_{rci}}}{M_{rci}!} \right) \right)^{i \in S} \right. \\ &\times \left. \left( \sum_{\mathbf{N}_{\text{miss}_i}} \left( \prod_{r,c} \frac{\theta_{rci}^{N_{rci}}}{N_{rci}!} \right) \right)^{i \notin S} \right. \\ &\times \left. \left( |\underline{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\underline{\omega}_i - \underline{\mu})^T \underline{\Sigma}^{-1} (\underline{\omega}_i - \underline{\mu}) \right\} \right) d\theta_i \right] \end{aligned}$$

Proportionality 4 can be understood as follows: the first line represents the hyperprior. The second and third lines correspond to the multinomial assumptions for the internal cell counts, with the second line demonstrating one of the contributions that the survey makes to the information in the

posterior. As  $K_{rc_i}$  gets large,  $M_{rc_i} = N_{rc_i} - K_{rc_i}$  decreases, reducing the uncertainty in the exponent of the numerator of  $\frac{\theta_{rc_i}^{M_{rc_i} + K_{rc_i} - 1}}{M_{rc_i}!}$  and driving the denominator to 1. If  $K_i = N_i$  (meaning that all voters in precinct  $i$  were sampled), then this portion of the posterior corresponds to the non-constant portion of the likelihood of the probability vector of a multinomial distribution. The fourth line is the multivariate normal. Note that a fair amount of structure is contained within the summations over  $M_{miss_i}$  and  $N_{miss_i}$  as well as the integral over  $\underline{\theta}_i$ . In each precinct  $i$ , the missing internal cell counts must sum to their row and column totals, and each contingency table row's  $\theta$ s must stay within a simplex. A more complex version of Proportionality 4, which demonstrates more explicitly the constraints involved, appears in [Greiner and Quinn \(2010\)](#).

In many voting applications, particularly in redistricting, quantities represented above by Greek letters are of limited interest. Instead, interest lies in functions of the counts produced upon summation of the contingency tables over  $i$ . These functions include  $\Lambda_{rc} = \frac{\sum_i N_{rc_i}}{\sum_i (N_{r_i} - N_{rA_i})}$ ,  $\Gamma_r = \frac{\sum_i (N_{r_i} - N_{rA_i})}{\sum_i (N_i - N_{A_i})}$ , and  $TO_{rc} = \frac{\sum_i (N_{r_i} - N_{rA_i})}{\sum_i N_{r_i}}$  representing, respectively, the fraction of actual (as opposed to potential) voters of race  $r$  supporting candidate  $c$ , the fraction of actual voters who are of race  $r$ , and the turnout of race  $r$ 's potential voters. The interest in these (and other) functions of the internal cell counts leads us to fit our proposal via a three-part Gibbs sampler; details appear in [Greiner and Quinn \(2010\)](#).

Speed is a serious concern here. [Greiner and Quinn \(2010\)](#) has some details, but depending on the constraints imposed by the bounds, ecological data can have little information in them, resulting in slow mixing. At present, after experimenting with several choices of proposal distributions (see [Metropolis \*et al.\* \(1953\)](#), [Tanner and Wong \(1987\)](#)) and fitting algorithms, our software run on a reasonable laptop can ordinarily analyze a dataset of the approximate size of a typical United States congressional district in a few hours. As of now, then, analyzing multiple datasets in a short period of time, a feature of some modern United States voting rights litigation, may require special computational tools. We continue to work to address this situation.

**3. A Comparison of Estimators.** We present the results of simulation studies primarily addressing two broad questions. First, in the  $R \times C$  context, what is the relative performance of an ecological model alone, a survey estimator alone, and our hybrid technique? In particular, we are interested in the relative performance of these three classes of estimators (i) in

the presence or absence of aggregation bias, and (ii) when contingency tables have relatively even distribution of counts among rows versus a moderate tendency for counts to be concentrated in one or another row. Note that if counts in contingency tables tend to be distributed relatively evenly among the rows, the bounds (Duncan and Davis (1953)) constrain the posterior less. In voting parlance, segregated housing patterns tend to lead to better performance of an ecological model.

Our second question of interest is whether the method of selecting the contingency tables (precincts) for inclusion in the sample  $S$  affects estimation. The advantages of probability weighting according to some observed criteria, such as size, are well understood in the survey literature. In the context of ecological data, however, we are interested in whether any benefits accrue to weighting contingency tables according to whether their bounds were likely to constrain, *i.e.*, whether a particular table’s counts were mostly in one row. In voting parlance, is there an advantage to weighting racially uniform precincts differently from racially mixed precincts?

*3.1. Simulation Methods.* We simulated blocks of 100 voting jurisdictions, producing datasets that generally resembled a United States congressional district in which a court might look for racial bloc voting. We assumed three racial groups (black, white, Hispanic) and two candidates (Democrat, Republican), producing precinct-level tables as per Table 1. For each jurisdiction, we applied seven estimation techniques: an ecological model alone; three two-stage sampling estimators in which sampled precincts were selected using different weighting schemes, after which a simple random sample was taken of potential voters within each precinct; and three hybrid estimators, in which the ecological model was combined with the data from each of the two-stage samples. With respect to the three survey samples, one (“Sampling Scheme 1”) assigned much heavier weights to racially integrated precincts, the second (“Sampling Scheme 2”) applied moderately greater weights to racially integrated precincts, and the third (“Sampling Scheme 3”) applied much heavier weights to racially uniform precincts. Population fractions, turnout levels, and party preferences of blacks, whites, and Hispanics were set at levels approximating behavior we have observed in United States congressional districts.

We present the results for six simulated blocks of jurisdictions: integrated (less integrated) without aggregation bias; integrated (less integrated) with aggregation bias; and integrated (less integrated) with severe aggregation bias. To induce aggregation bias, we turned the top-level (normal distribution) location parameters for whites ( $\mu_{wD}$  and  $\mu_{wR}$ ) into linear functions of

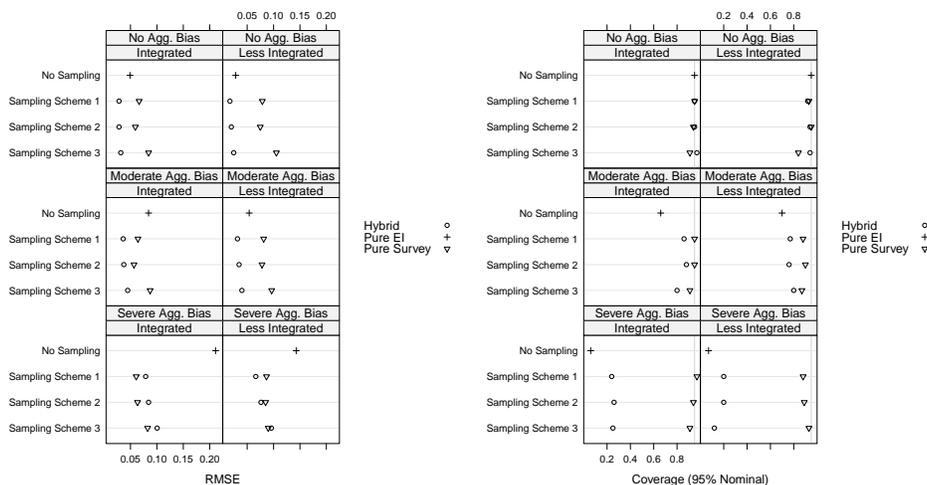


FIG 1. *Summary of results from simulations. The left panels display RMSE while the right panels display the coverage of nominal 95% credible intervals. Sampling Scheme 1 heavily overweights racially mixed precincts, Sampling Scheme 2 mildly overweights racially mixed precincts, and Sampling Scheme 3 heavily overweights racially uniform precincts. Note that “Integrated” datasets have less information in the bounds. The results show that the hybrid estimator generally outperforms the pure survey and pure ecological inference estimators, and offers substantial RMSE reductions in many circumstances. In the absence of severe aggregation bias, the hybrid estimator’s coverage is typically less than but comparable to that of the pure survey estimator.*

the fraction Hispanic,  $X_{h_i} = \frac{N_{h_i}}{N_i}$ . After speaking to a few persons knowledgeable in the field of voting rights and racial bloc voting regarding what might be realistic, we chose figures for aggregation bias that, in expectation, would induce white voters in an 20% Hispanic precinct to vote approximately 70% for the Republican, while white voters in an 80% Hispanic precinct would vote 55% for the Republican. The corresponding figures for the severe aggregation bias (an approximately 90% to 30% swing in white Republican support) were designed to be unrealistically harsh and to test the outer limits of the method. We present here the results for the quantity  $\Lambda_{hD}$  because, in voting applications, it is often difficult to estimate and particularly vulnerable to aggregation bias, with both of these factors due to Hispanics’ lower turnout rates and greater tendency (relative to blacks) to vote in non-uniform patterns.

When comparing estimators, we proceed on several of the usual fronts, examining coverage of 95% intervals, 95% interval length, and root mean

squared error (“RMSE”). In addition, because we apply the same seven estimation techniques to each simulated dataset, we examine how often estimators outperform one another in each simulation block in terms of squared error by calculating a binomial p-value under a null hypothesis of that the two estimators compared are the same. When we report a p-value, we mean this value unless we state otherwise.

Additional details of our simulations appear [Greiner and Quinn \(2010\)](#).

*3.2. Simulation Results.* Basic results are summarized in Figure 1. We draw the following conclusions. First, hybrid estimators trounce the pure ecological inference estimator under all circumstances. While we do not find this result surprising in the abstract, the magnitude of the improvement is worthy of note. In the absence of aggregation bias, the hybrid estimators offer greater precision, producing posterior intervals that are narrower but that still provide stochastically nominal coverage. The best-performing hybrid (Sampling Scheme 1) results in a reduction of posterior interval length of approximately 30-50%, depending on the level of integration in housing patterns. With aggregation bias, the hybrid raises the coverage of the 95% intervals from poor (roughly .68) to a level that, while less than nominal, might approach tolerability (roughly .85). Meanwhile, the RMSE reductions are on the order of 30-60%. With severe aggregation bias, any estimator that uses the ecological data fails to achieve nominal coverage. Nevertheless, all hybrids substantially outperform the ecological estimator alone. The reduction in RMSE, on the order of 55%, is substantial, with this result stemming from both a noticeable decrease in bias and a noticeable increase in precision. In comparing any hybrid to the ecological inference estimator, all p-values from our simulations are 0. From this, we provide the following recommendation: always include the survey.

Second, comparing hybrids to one another, there are advantages to avoiding a sampling scheme that oversamples contingency tables in which one row dominates, *i.e.*, racially homogenous precincts. Without aggregation bias, the difference between the hybrid that oversamples racially homogenous precincts (Sampling Scheme 3) versus the other two (Sampling Schemes 1 and 2, which overweight racially mixed precincts) is noticeable but modest; the latter offer 10-20% reductions in 95% interval length (all p-values less than .01). With aggregation bias or severe aggregation bias, the improvement is larger. The lack of nominal coverage makes 95% interval length less informative. But regarding RMSE, Sampling Scheme 1, which oversamples racially mixed precincts, achieves 20-30% reduction as compared to Sampling Scheme 3, which oversamples racially uniform precincts (all p-values

are 0).

The most difficult comparison is the hybrid estimators versus the pure survey estimators. In the absence of aggregation bias, the conclusion is simple, with any hybrid estimator constituting an enormous improvement. The greater precision of the hybrid estimators is reflected in both the length of the 95% intervals, which can be as much as 70% narrower, as well as RMSE comparisons. Any hybrid outperforms any pure survey estimator (all p-values are 0).

With aggregation bias, we again recommend the hybrid over the pure survey estimator, but we do so more cautiously. Although the pure survey estimators' intervals come closer than the hybrids to achieving nominal coverage, the coverage gains are modest (around 7%). Meanwhile, the RMSE gains from the hybrids, on the order of 35-60%, are substantial. On average, the bias of the hybrid estimates is modest, roughly two or three percentage points (*i.e.*, a point estimate of .53 when the truth is .51). Thus, even in the presence of aggregation bias, the hybrids offer substantial benefits over the pure survey estimators.

In the presence of severe aggregation bias, the results are mixed. With integrated housing patterns and in the presence of severe aggregation bias, the combination of bias and lack of bounding information renders the pure survey estimators superior, with hybrid RMSEs approximately 10-20% larger than their pure survey counterparts. With severe aggregation bias and with less integrated housing patterns, interval coverage for both types of estimators was less than nominal (and worse for the hybrids). With respect to RMSE, however, on average, the hybrids usually outperform their specific pure survey counterparts, and the reductions are on the order of 10% to as high as 25%. Average does not mean always, however. And on a simulation-by-simulation basis, the comparison of some pure survey estimators to the hybrids results in p-values near 0 in favor of the pure survey estimators (recall that our p-values represent which method prevails simulation-by-simulation, a 0-1 outcome.) The reason for this is that the higher variances associated with the pure survey estimators mean that when these estimators miss the target, they can miss badly, raising the RMSE, which as a function of an average is sensitive to large misses. In the presence of contextual effects, the lower-variance hybrid estimators reduce the risk of a point estimate that is badly wide of the mark, at the cost of some bias.

*3.3. Simulation Conclusions.* Thus, as between hybrid versus survey estimators, which estimator should a researcher prefer? In our view, the answer depends primarily on three factors: the extent to which contingency

tables tend to be dominated by one row (*i.e.*, the extent of racial segregation in housing patterns), the magnitude of aggregation bias in the data, and whether the ultimate user cares more about an accurate point estimate or a valid interval. The first factor is observable. The second is not observable, and it may or may not be that in some instances, a researcher or expert witness will have some information about aggregation bias from external sources. Regarding the third, some users pay attention primarily to point estimates. Courts, for example, who in voting rights litigation may examine results from dozens of elections, typically do not incorporate uncertainty estimates into their opinions, despite exhortations from social scientists to the contrary. Other users make what we suspect for statisticians is the more traditional choice. In general, however, our recommendation is that unless the researcher has reason to fear extremely strong (“severe” really means “brutal”) aggregation bias, the hybrid estimator is preferable.

**4. Boston Area Colleges Exit Poll.** Did Asian-American voters in the City of Boston support a Massachusetts ballot initiatives repealing criminal penalties for possession of small amounts of marijuana and banning gambling on greyhound racing? Were support rates for the marijuana initiative different as between Caucasian versus Hispanic voters? To test the methods we propose, we conducted an exit poll in the City of Boston on November 4, 2008. Because our interest is in both the operational feasibility as well as the comparative technical advantages or disadvantages of hybrid estimators, we briefly describe the running of the poll and the necessary preprocessing of the data before articulating required assumptions and providing results. We demonstrate that the two questions articulated above are difficult to answer with either the exit poll or the ecological estimator standing alone, but that the hybrid permits reasonable inferences as to both.

4.1. *Mechanics And Initial Results.* We recruited law, graduate, and undergraduate students from 11 Boston area colleges and universities to participate in an exit poll. Our recruiting efforts yielded over 400 pollsters, which we organized into teams captained by a law or graduate student. There were two election day shifts lasting seven hours each, which covered the whole of the election day. Captains attended one of several 90-minute training sessions, while training for non-captain pollsters lasted an hour. All sessions were live and covered essential survey/exit polling techniques. For example, pollsters were instructed to step away from voters after making a successful approach and to request that voters themselves place completed questionnaires in a visibly closed box (see [Bishop and Fisher \(1995\)](#)). Five specially trained, two-person roving quality control teams circulated in cars, visiting

each polling location multiple times throughout election day and monitoring compliance with the required techniques. We attempted to deploy multilingual pollsters to locations in which a comparatively high percentage of voters spoke languages other than English.

Pollsters approached every eighth voter but alternated between a “voter choices” questionnaire, which generated the data used in this paper, and a “voter experience” form, which was used for other purposes. Effectively, this meant a targeted  $\frac{1}{16}$  sampling interval for the race-and-voter-choices exit poll. Prior coordination with the City of Boston Election Department, together with the absence of a law in Massachusetts regulating exit polls, enabled pollsters to stand immediately outside the exits to the buildings in which voting occurred, and teams were large enough to cover all exits.

The poll covered 39 of Boston’s 160-odd polling locations. 26 of the 39 locations were selected in a non-random manner due to the research design associated with the voter experience questionnaire; the other 13 were randomly selected using inverted Herfindahl index weights that resulted in a higher probability of selecting polling locations in which several racial groups were represented (see [Greiner and Quinn \(2010\)](#) for details).

Overall, Boston Area Colleges Exit Poll pollsters approached approximately 4300 voters with voter choice questionnaires and achieved approximately a 57% response rate. Voter choice data were collected for United States president and for three Massachusetts ballot initiatives, one repealing the state income tax, one eliminating criminal penalties for possession of small amounts of marijuana, and one banning gambling on dog racing. After multiply imputing for nonresponse (see below), we applied a stratified (to reflect the separate deterministic versus random precinct-selection schemes), two-stage (cluster followed by simple random sample) estimator to the results to check our predictions against the known truth. As [Figure 2](#) demonstrates, we found that our projections closely approximated the overall true two-party vote fractions, where “two-party” means the percentage of Obama supporters out of those who voted for either Obama or McCain, or the percentage of Yes votes out of those who voted Yes or No on the ballot initiatives. We did find, however, a curious (see [Silver, Anderson and Abramson \(1986\)](#)) tendency among poll respondents to overreport non-voting behavior, and the prior in our multiple imputation algorithm may have exaggerated this aspect of the data. For these reasons, we compare estimators for the marijuana initiative, where our two-party projection was accurate, where non-voting overreport was comparatively low, and where the two-party vote was reasonably close. Results for the dog racing ballot initiative, which share these characteristics, were similar, and are available

from the authors.

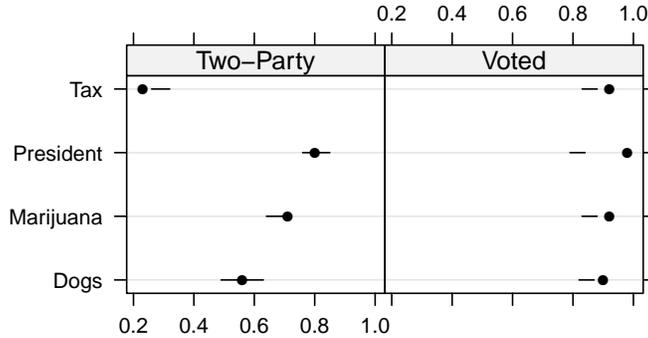


FIG 2. Results of Boston Area Colleges Exit Poll (pure survey estimators). “Two-Party” refers to the percentage of actual voters voting for Obama (Presidential) or Yes (income tax, marijuana, and dog racing ballot initiatives), while “Voted” refers to the percentage of persons entering the ballot who cast ballots in the relevant contest. True values are represented by solid dots, 95% confidence intervals are represented by the dark lines. Two-party point estimates are generally accurate, but non-voting behavior is overestimated.

4.2. *Data Processing and Critical Assumptions.* We detail in this section the critical assumptions underlying our various estimators. First, to account for nonresponse, we created 10 completed datasets via multiple imputation. The imputation model was a loglinear model for categorical data as implemented in Joe Shafer’s `cat` package.<sup>2</sup> Computational challenges arose because of the fairly large number of variables to impute and our desire to allow for more complicated associations than would be possible under a multivariate normal model or a 2-way loglinear model. To overcome these challenges we made use of a parametric bootstrap approach (Honaker and King, 2009) along with a factorization of the full data distribution that allowed us to work with the data in moderately-sized chunks.

Our procedure was the following. First, we created 10 bootstrap datasets by sampling rows with replacement from the observed data matrix. We partitioned the variables in each of these bootstrap datasets into three sets—pollster-specific attributes, voter demographics, and voter choice variables. Then, for each of the bootstrap datasets, we imputed pollster-specific at-

<sup>2</sup> <http://cran.r-project.org/web/packages/cat/index.html>

tributes, voter demographics given the imputed pollster attributes, and finally voter choice data given the imputed voter demographics and a subset of the imputed pollster characteristics.

Each imputation step worked as follows. Given a particular bootstrap dataset we calculated the posterior mode of the cell probabilities using the ECM algorithm. We then sampled the missing data from the appropriate multinomial distribution with probabilities given by the maximum a posteriori estimates. For the pollster-specific data (which had very little missingness) and the voter demographic data we employed a loglinear model with all 3-way interactions and a Dirichlet prior for the cell probabilities with parameters all equal to 1.0001. For the voter choice data (which had more missingness) we used a loglinear model with all 2-way interactions and a Dirichlet prior on the cell probabilities with parameters equal to 1.001.

The assumptions underlying the multiple imputations are the primary ones needed to render the pure survey estimators discussed below valid. Another assumption is that the interval sample produced a simple random sample of voters in the in-sample precincts. We deem this assumption plausible in light of the  $\frac{1}{16}$  target interval. Overall, in assessing these assumptions, for the two electoral contests presented in this paper, we are encouraged by the exit poll's ability to project closely the two-party vote and to approximate the amount of non-voting observed in the ballot initiatives.

For the hybrid and pure ecological estimators, the most important assumption is lack of contextual effects. With respect to this dataset, however, the no-contextual-effect assumption is slightly stronger for the data as used by the hybrid estimator than for the data as used by the pure ecological counterpart because the hybrid operates on more aggregated data, as follows. Exit polls survey voters by polling location, not by precinct, and in the City of Boston, many polling locations host voters of more than one precinct such that pollsters standing outside of a polling location's building are unable to distinguish voters from the various precincts housed there. Thus, the data used by the hybrid estimator had to consist of figures at the level of the polling location (for in-sample polling locations), *i.e.*, further aggregated. The ecological estimator could operate at the level of the individual precinct. Note that for the data used by the hybrid, for out-of-sample polling locations, we used precinct-level (as opposed to polling-location-level) figures. Note also that, at least in Boston, precincts are never split into two or more polling locations; that is, each precinct is contained wholly within one polling location.

According to figures based on Census 2000 and provided by the Boston Redevelopment Authority, the City of Boston's voting age percentages by

race are as follows: 55% white, 20% black, 12% Hispanic, 9% Asian, and the rest of “other” race.<sup>3</sup> We investigated whether the various estimators under consideration could say anything useful about Boston’s four most populous racial groups.

4.3. *Results of Various Estimators: Voting Preferences by Race.* Our results are encapsulated in Figures 3 and 4. We draw the following conclusions. First, there is little evidence to contradict the critical no-aggregation-bias assumption needed for the ecological and hybrid estimators. The point estimates from the survey estimator generally align with those from the other two. This fact does not provide total security, given the high variance of the survey estimator, but total security is rarely available when analyzing ecological data.

Second, even after accounting for nonresponse via multiple imputation, which necessarily involves higher variances than would be present for a survey without nonresponse, the hybrid estimator provides substantial variance reduction in a way that makes a substantive difference. For example, in the marijuana ballot initiative, the 95% interval for the Asian support rate was (.03, .99) for the pure ecological inference estimator and was (.34, .68) for the pure survey, but the hybrid interval was (.54, .73). Thus, only via the hybrid estimator would a researcher or an expert witness be able to conclude that Asian voters in the City of Boston supported the marijuana initiative. The same phenomenon occurs in the Asian vote on the initiative to ban gambling on greyhound racing (results not shown). Further, the pure survey and the pure ecological estimators are less able to distinguish Hispanic versus white preferences regarding the marijuana initiative. For the hybrid estimator, in contrast, these 95% confidence intervals intersect by only a hair’s breadth.

These results are substantively interesting in their own right, but we are encouraged by the fact that the hybrid estimator appears to help where help is most needed. The variance reduction available for the estimates of Asian and Hispanic voting behavior is substantial. As the two racial groups with the lowest VAP and lowest turnout, Hispanics and Asians represent the most difficult challenge to inference about voting behavior by race, and the performance of the hybrid estimator here is encouraging.

A question arises: how could this happen? How could the combination of information from a survey and from ecological data, neither of which

---

<sup>3</sup>Recalling that Census 2000 allowed respondents to mark more than one race box, these categories are in fact shorthand for the following: “Hispanic” means Hispanic (regardless of any other race box checked), “Asian” means non-Hispanic any part Asian, “black” means non-Asian non-Hispanic any part black, and “white” means anyone left not who was not in the other race category.

alone provides useful results, reduce variance enough to allow for meaningful substantive inference? We offer the hypothesis that the answer lies in the better estimation of parameters governing between-contingency-table-row (as opposed to within-contingency-table-row) relationships. An example to clarify this distinction: a within-contingency-table-row relationship would be an tendency for precincts that have high counts of Asians voting Democrat to also have high counts of Asians who abstain from voting. A between-contingency-table-row relationship with be a tendency for precincts that have high counts of Asians voting Democrat to also have high counts of blacks who vote Democrat.

Several commentators (*e.g.*, [King \(1997\)](#)) have noted the difficulty in estimating model parameters that govern behavior between (as opposed to within) contingency table rows. We explored the relative paucity of information about between-row relationships in [Greiner and Quinn \(2009\)](#). It appears, however, that individual-level data can stabilize estimates of between-row parameters in an important way. Recall that in our model, we stack the logistic-transformed probability vectors from each contingency table's row multinomial to form a single vector of dimension  $R^*(C-1)$ , which we then assume follows a multivariate normal. Accordingly, the covariance matrix of this normal ( $\Sigma$ ) can be decomposed into block diagonal elements, which govern within-contingency-table-row relationships, and block off-diagonal elements, which govern between-contingency-table-row relationships. As applied to the City of Boston, with black, white, Hispanic, and Asian racial groups, the matrix is as follows:

$$\Sigma = \begin{bmatrix} \Sigma_b & \Sigma_{bw} & \Sigma_{bh} & \Sigma_{ba} \\ \Sigma_{bw} & \Sigma_w & \Sigma_{wh} & \Sigma_{wa} \\ \Sigma_{bh} & \Sigma_{wh} & \Sigma_h & \Sigma_{ha} \\ \Sigma_{ba} & \Sigma_{wa} & \Sigma_{ha} & \Sigma_a \end{bmatrix}$$

Note that each of  $\Sigma_{ba}$ ,  $\Sigma_{wa}$ , and  $\Sigma_{ha}$  is of dimension  $2 \times 2$ , and because each is off the main diagonal, each has four correlations within it.

It appears that the introduction of individual-level information allows estimation of Asian voting behavior to borrow strength from estimates of white, black, and Hispanic voting behavior by way of better and more precise estimation of the correlations in  $\Sigma_{ba}$ ,  $\Sigma_{wa}$ , and  $\Sigma_{ha}$ . [Figure 4](#) compares the posterior intervals of these correlations in the marijuana ballot initiative in the pure EI model versus the hybrid. The narrower intervals of the correlations from the hybrid, together with the fact that most of the distributions from the hybrid have most of their mass above 0, appear to enable better

modeling of between-contingency-table-row relationships. In other words, the correlations represented suggest that within a precinct, Asian voting behavior is similar to that of other racial groups, particularly that of whites. We hypothesize that this similarity, together with the substantial information about white voting behavior (from the bounds), in turn allows non-Asian voting behavior to inform estimation of Asian preferences. If we are right, this fact highlights the importance of using a model flexible enough to allow estimation of between-contingency-table-row relationships, something few other  $R \times C$  models do.

**5. Conclusion.** In this paper, we have proposed a hybrid count ecological inference model capable of handling datasets with contingency tables of any size and shape. We have briefly explored the benefits of count versus fraction models in the  $R \times C$  context as well as the implications of different contingency-table-level sampling schemes. We have met the challenge of operationalizing the use of our hybrid to voting data by conducting an exit poll in the City of Boston, and in doing so have confronted a difficult scenario for a hybrid estimator because of (i) the impossibility of using optimal within-table sampling schemes, (ii) the problem of nonresponse, (iii) the additional level of aggregation occurring when more than one precinct share the same polling location, and (iv) the desire to estimate behavior of groups with low VAP and turnout. Our operationalization demonstrates that the hybrid model offers benefits to those who seek inferences regarding racial voting patterns.

#### SUPPLEMENTARY MATERIAL

**Supplement A: Supplement to “Exit Polling and Racial Bloc Voting: Combining Individual-Level and  $R \times C$  Ecological Data”** (doi: <http://lib.stat.cmu.edu/aoas/???/???>). This supplement describes the algorithms used to fit the models described in “Exit Polling and Racial Bloc Voting: Combining Individual-Level and  $R \times C$  Ecological Data”.

**Supplement B: Replication Materials for “Exit Polling and Racial Bloc Voting: Combining Individual-Level and  $R \times C$  Ecological Data”** (doi: <http://lib.stat.cmu.edu/aoas/???/???>). This supplement provides data and computer code that can be used to replicate the results in “Exit Polling and Racial Bloc Voting: Combining Individual-Level and  $R \times C$  Ecological Data”.

## References.

- 11TH CIR., (1984). United States v Marengo County Commission. *Federal Reporter, Second Series* **731** 1546.
- ACHEN, C. H. and SHIVELY, W. P. (1995). *Cross-Level Inference*. University of Chicago Press, Chicago.
- AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*, 2nd ed. The Blackburn Press, Caldwell, NJ.
- BELIN, T. R., DIFFENDAL, G. J., MACK, S., RUBIN, D. B., SCHAFER, J. L. and ZASLAVSKY, A. M. (1993). Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation. *Journal of the American Statistical Association* **88** 1149-1159.
- BISHOP, Y. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.
- BISHOP, G. F. and FISHER, B. S. (1995). Secret Ballots And Self-Reports in an Exit-Poll Experiment. *Public Opinion Quarterly* **59** 568-588.
- BROWN, P. J. and PAYNE, C. D. (1986). Aggregate Data, Ecological Regression, and Voting Transitions. *Journal of the American Statistical Association* **81** 452-460.
- DEMING, W. and STEPHAN, F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics* **11** 427-444.
- DEVILLE, J.-C., SARNDAL, C.-E. and SAUTORY, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* **88** 1013-1020.
- DUNCAN, O. D. and DAVIS, B. (1953). An Alternative to Ecological Correlation. *American Sociological Review* **18** 665-666.
- GELMAN, A., ANSOLABEHERE, S., PRICE, P. N., PARK, D. K. and MINNITE, L. C. (2001). Models, Assumptions, And Model Checking in Ecological Regressions. *Journal of the Royal Statistical Society A, Part 1* **164** 101-118.
- GLYNN, A. N., WAKEFIELD, J., HANDCOCK, M. S. and RICHARDSON, T. S. (2008). Alleviating Linear Ecological Bias And Optimal Design with Subsample Data. *Journal of the Royal Statistical Society, Series A* **171** 179-202.
- GLYNN, A. N., WAKEFIELD, J., HANDCOCK, M. S. and RICHARDSON, T. S. (2009). Alleviating Ecological Bias in Generalized Linear Models with Optimal Subsample Design. On file with authors.
- GREINER, D. J. (2007). Ecological Inference in Voting Rights Act Disputes. *Jurimetrics Journal* **47** 115-167.
- GREINER, D. J. and QUINN, K. M. (2009). R x C Ecological Inference: Bounds, Correlations, Flexibility, and Transparency of Assumptions. *Journal of the Royal Statistical Society, Series A* **172** 67-81.
- GREINER, D. J. and QUINN, K. M. (2010). Supplement to "Exit Polling and Racial Bloc Voting: Combining Individual-Level and R x C Ecological Data". DOI: ???
- HANEUSE, S. J. and WAKEFIELD, J. C. (2008). The Combination of Ecological And Case-Control Data. *Journal of the Royal Statistical Society, Series B* **70** 73-93.
- HONAKER, J. and KING, G. (2009). What to do About Missing Values in Times Series Cross-Section Data. Harvard University Working Paper.
- HOPKINS, D. J. (2008). No More Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead about Black and Female Candidates. available at <http://people.iq.harvard.edu/~dhopkins/wilder13.pdf>.
- ISSACHAROFF, S. (1992). Polarized Voting and the Political Process: The Transformation

- of Voting Rights Jurisprudence. *Michigan Law Review* **93** 1833.
- JUDGE, G., MILLER, D. J. and CHO, W. K. T. (2004). An Information Theoretic Approach to Ecological Estimation and Inference. In *Ecological Inference: New Methodological Strategies* (G. King, O. Rosen and M. A. Tanner, eds.) Cambridge University Press, Cambridge.
- KING, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton University Press.
- LITTLE, R. J. (1993). Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association* **88** 1001-1012.
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience.
- LUBLIN, D. I. (1995). *Classifying by Race* Race, Representation, and Redistricting 111-128. Princeton University Press.
- MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*, 2 ed. Chapman & Hall/CRC.
- METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21** 1087-1092.
- RAGHUNATHAN, T. E., DIEHR, P. K. and CHEADLE, A. D. (2003). Combining Aggregate And Individual Level Data To Estimate an Individual Level Correlation Coefficient. *Journal of Education and Behavioral Statistics* **28** 1-19.
- RIVERS, D. (1998). Review of "A Solution to the Ecological Inference Problem". *The American Political Science Review* **92** 442-443.
- ROBINSON, W. (1950). Ecological Correlations And the Behavior of Individuals. *American Sociological Review* **15** 351-357.
- ROSEN, O., JIANG, W., KING, G. and TANNER, M. A. (2001). Bayesian And Frequentist Inference for Ecological Inference: the R x C Case. *Statistica Neerlandica* **55** 134-156.
- SALWAY, R. and WAKEFIELD, J. (2005). Sources of Bias in Ecological Studies of Non-Rare Events. *Environmental And Ecological Studies* **12** 321-347.
- SILVER, B. D., ANDERSON, B. and ABRAMSON, P. R. (1986). Who Overreports Voting. *American Political Science Review* **80** 613-624.
- STEEL, D., TRANMER, M. and HOLT, D. (2003). *Analysis of Survey Data* Analysis Combining Survey and Geographically Aggregated Data. Wiley.
- TANNER, M. A. and WONG, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82** 528-540.
- WAKEFIELD, J. (2004). Ecological Inference for 2 x 2 Tables. *Journal of the Royal Statistical Society, Series A* **167** 385-445.
- ZASLAVSKY, A. M. (1993). Combining Census, Dual-System, and Evaluation Study Data to Estimate Population Shares. *Journal of the American Statistical Association* **88** 1092-1105.

D. JAMES GREINER  
 HARVARD LAW SCHOOL  
 504 GRISWOLD HALL  
 CAMBRIDGE MA, 02138  
 E-MAIL: [jgreiner@law.harvard.edu](mailto:jgreiner@law.harvard.edu)

KEVIN M. QUINN  
 UC BERKELEY SCHOOL OF LAW  
 490 SIMON HALL  
 BERKELEY, CA 94720-7200  
 E-MAIL: [kquinn@law.berkeley.edu](mailto:kquinn@law.berkeley.edu)

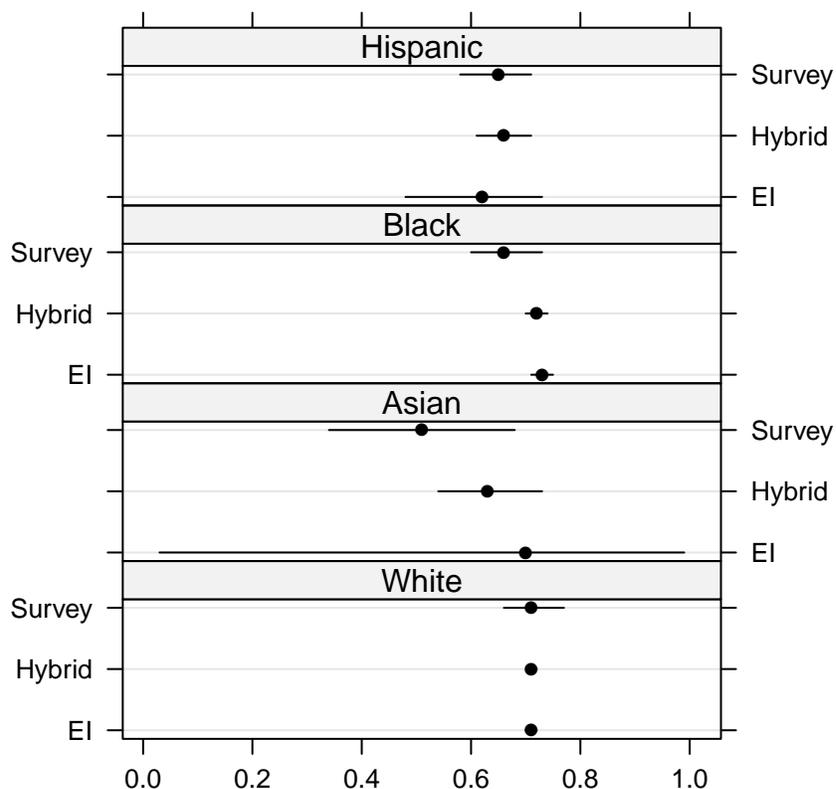


FIG 3. *Estimated fractions of support for the marijuana decriminalization ballot initiative among the four most numerous racial groups in the City of Boston. Filled circles represent point estimates and dark lines represent 95% credible intervals. “EI” refers to ecological inference estimator alone, “Survey” is the exit poll alone, and “Hybrid” is the hybrid estimator. Survey and Hybrid estimates come from multiple imputation. Only the hybrid estimator offers enough precision in the Asian category to allow substantive inference. The hybrid estimator also best differentiates Hispanic from white preferences.*

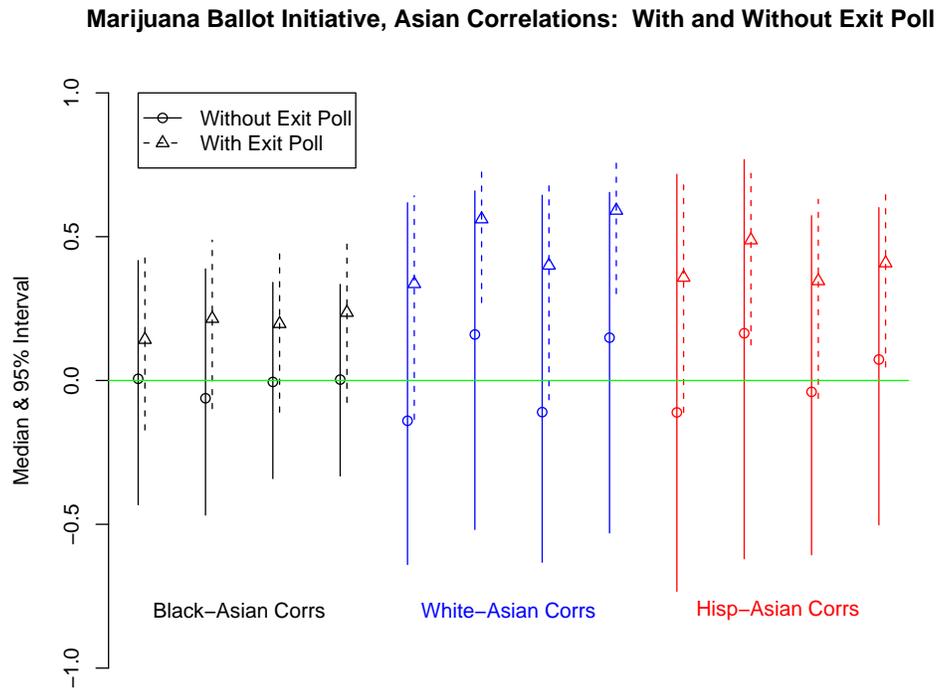


FIG 4. Comparison of posterior distributions from ecological inference, with and without exit poll, of between-contingency-table-row correlations governing the relationship of black, white, and Hispanic voters to Asian voters. (The with-exit-poll figures are averages of ten multiple imputations.) The narrower posterior intervals, and the greater density above zero, in the with-exit-poll correlations suggest that the with-exit-poll model is taking advantage of a tendency of various racial groups to vote similarly within a precinct to provide better estimates of Asian voting behavior. The without-exit-poll model is unable to take advantage of this tendency.