# Exit Polling and Racial Bloc Voting: Combining Individual Level and R x C Ecological Data

*By D. James Greiner and Kevin M. Quinn*
*Harvard Law School and Harvard University*

May 2009

HARVARD Kennedy School
RAPPAPORT INSTITUTE
for Greater Boston

# Exit Polling and Racial Bloc Voting: Combining Individual-Level and R × C Ecological Data[*]

D. James Greiner[†] & Kevin M. Quinn[‡]

Draft of May 12, 2009

**Abstract**

Cross-level or 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. To do so, we extend and study a hybrid model that addresses two-way tables of arbitrary dimension. We apply the hybrid model to our 400-pollster exit poll, taken in 39 polling locations in the City of Boston and covering four contests in November, 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 in realistic settings.

**Keywords:** ECOLOGICAL INFERENCE, BAYESIAN INFERENCE, VOTING RIGHTS LITIGATION, EXIT POLLS, SURVEY SAMPLING

[*] The authors extend great thanks to Jayanta Sircar at the Harvard School of Engineering for his assistance in providing computing power; to Andrew Martin 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.

[†] Assistant Professor, Harvard Law School, 504 Griswold Hall, Cambridge, MA, USA 02138; jgreiner@law.harvard.edu.

[‡] Associate Professor, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA, USA 02138; kquinn@fas.harvard.edu.

Cross-level or ecological inference is the attempt to draw conclusions about statistical relationships at one level from data aggregated to a higher level. Most frequently, ecological inference is conceptualized as an effort to infer individual-level relationships from a set of contingency tables when only the row and column totals are observed. 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, information into the model. The past decade or so has seen several papers (Steel et al. (2003), Raghunathan et al. (2003), Glynn et al. (2008), Haneuse and Wakefield (2008), Glynn et al. (2009)) addressing how best to do so. As is true of ecological inference more generally, most papers addressing incorporation of additional information have focused on sets of 2 × 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 × C case, building on earlier work of our own (Greiner and Quinn (2009)), which in turn built on Brown and Payne (1986) and Wakefield (2004). We subject the task of combining individual-level and R × C ecological data to a stress test in the form of an effort to draw inferences about the voting behavior of R racial ("racial" means racial or ethnic) 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-precinct exit poll we administered in the City of Boston on the November 4, 2008 election. Combining ecological data and an exit poll constitutes a stress test for a hybrid model because (i) the nature of exit polling prevented 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 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 × 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. 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 the 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 Massachusetts ballot initiatives regarding marijuana and dog racing. We use our exit poll as a springboard to discuss challenges in this area unique to inferences regarding voting behavior, challenges that render some suggestions in the literature regarding optimal sampling schemes difficult to implement. 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.

Regarding notation, any quantity with the subscript $_{row\ COLUMN_i}$ refers to that quantity in the ith contingency table's (precinct's) rth row, cth column; r runs from 1 to R, c from 1 to C, i from 1 to I, although in our examples we give more substantive content to the row and column subscripts. $N$'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 $M_{rc_i} = N_{rc_i} - \mathrm{K}_{\mathrm{rc_i}}$. We italicize unobserved counts but leave observed quantities in ordinary typescript. Table 1 clarifies our representations for the case of 3 × 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 | $N_{bD_i}$ | $N_{bR_i}$ | $N_{bA_i}$ | $\mathrm{N_{b_i}}$ |
| white | $N_{wD_i}$ | $N_{wR_i}$ | $N_{wA_i}$ | $\mathrm{N_{w_i}}$ |
| Hispanic | $N_{hD_i}$ | $N_{hR_i}$ | $N_{hA_i}$ | $\mathrm{N_{h_i}}$ |
| | $\mathrm{N_{D_i}}$ | $\mathrm{N_{R_i}}$ | $\mathrm{N_{A_i}}$ | $\mathrm{N}_i$ |

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 ∈ S, $\mathbf{K}_i$ is a random matrix of dimension $\mathrm{J_i}\times$ (R*C), where $\mathrm{J_i}$ 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.

Let $\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}$ ($\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}$) represent the vector of observed row (column) totals in the ith precinct, with $\mathbf{N}_{\mathrm{row}}$ ($\mathbf{N}_{\mathrm{col}}$) a matrix of all $\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}$'s ($\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}$'s), and $\mathbf{N}_{obs} = [\mathbf{N}_{\mathrm{row}}\ \mathbf{N}_{\mathrm{col}}]$. Let $\boldsymbol{N}_{comp_i}$ equal the (unobserved) full set of internal cell counts in the ith precinct. Finally, let $\boldsymbol{N}_{miss_i}$ denote any set of (R-1)*(C-1) counts for the ith precinct which, had they been observed in conjunction with $\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}$ and $\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}$, would have been sufficient to determine all table counts. In Table 1, for example, $\boldsymbol{N}_{miss_i}$ could equal $\begin{bmatrix} N_{bD_i} & N_{bR_i} \\ N_{wD_i} & N_{wR_i} \end{bmatrix}$. Note that $\boldsymbol{N}_{comp_i}$ and $\boldsymbol{N}_{miss_i}$ are used in the missing data sense (*e.g.*, Little and Rubin (2002)).

# 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 counts by their corresponding row totals.

Apart from the approach we advocate, a variety of R × 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 et al. (2004). These proposals all share the feature that they model (at various levels), not the internal cell counts themselves, but rather the fractions produced when the 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 $\beta_{bD_i} = \frac{N_{bD_i}}{\mathrm{N}_{\mathrm{b_i}}}$, and $\underset{\sim}{\beta}_i$ refer to the vector of the $\beta$'s in the ith precinct. If modeling fractions and proceeding in a Bayesian fashion, a researcher might put a prior on the $\underset{\sim}{\beta}_i$'s with parameter $\zeta$, in which case one representation of this class of models is as follows:

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

For example, in the simplest version of linear model, $p(\underset{\sim}{\beta}_i|\zeta)$ can be conceptualized as a multivariate normal with mean vector $\underset{\sim}{\beta}$ and null variance. In Rosen et al. (2001), $p(\zeta)$ is a set of mutually independent gamma distributions, $p(\underset{\sim}{\beta}_i|\zeta)$ a product Dirichlet, and $p(\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}|\underset{\sim}{\beta}_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}})$ a multinomial parameterized by a mixture of $\beta$'s and the fractions produced when $\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}$ is divided by its sum. Note that in equation (1), because there is no distribution posited for the unobserved internal cell counts, there is no summation needed to eliminate them.

In contrast, consider a class of techniques that models the unobserved internal cell counts. A researcher proceeding an 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 $\Upsilon_i$), might specify a prior on the $\Upsilon_i$'s (call the parameters in this prior $\Xi$), and might sum out the unobserved internal cell counts. Thus, the equation corresponding to

(1), above, is

$$p(\Xi|\mathbf{N}_{\mathrm{col}},\mathbf{N}_{\mathrm{row}}) \;\propto\; p(\Xi)\prod_{i=1}^{\mathrm{I}}\left[\int\sum_{\boldsymbol{N}_{miss_i}} p(\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}|\boldsymbol{N}_{comp_i})\times\; p(\boldsymbol{N}_{comp_i}|\Upsilon_i,\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}})\times p(\Upsilon_i|\Xi)d\Upsilon_i\right] \tag{2}$$

$p(\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}|\boldsymbol{N}_{comp_i})$ appears to make the relationship between the left- and right-hand sides of the $\propto$ symbol more transparent; in fact, $\boldsymbol{N}_{comp_i}$ determines $\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}$, rendering $p(\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}|\boldsymbol{N}_{comp_i})$ degenerate. Note in this formulation, there is an explicit model for the internal cell counts ($p(\boldsymbol{N}_{comp_i}|\Upsilon_i,\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}})$), which in turn requires a summation over $\boldsymbol{N}_{miss_i}$ to produce the observed-data likelihood. But the distribution of $\boldsymbol{N}_{miss_i}$ is complicated; the permissible support of each element of $\boldsymbol{N}_{miss_i}$ depends on the value of the others. 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, particularly via markov chain monte carlo, 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 $p(\boldsymbol{N}_{comp_i}|\Upsilon_i,\underset{\sim}{\mathrm{N}}_{\mathrm{row_i}})$. In contrast to equation (1), equation (2) can be modified in a simple way to incorporate data from a sample, as follows:

$$p(\Xi|\mathbf{K},\ \mathbf{N}_{\mathrm{col}},\ \mathbf{N}_{\mathrm{row}}) \ \propto\ p(\Xi)\prod_{i=1}^{\mathrm{I}}\Bigg[\int\sum_{\boldsymbol{N}_{miss_i}} p(\mathbf{K}_i|\boldsymbol{N}_{comp_i})^{(i\in\mathrm{S})} \ \times\ p(\underset{\sim}{\mathrm{N}}_{\mathrm{col_i}}|\boldsymbol{N}_{comp_i}) \tag{3}$$

$$\times\ p(\boldsymbol{N}_{comp_i}|\Upsilon_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}) \ \times\ p(\Upsilon_i|\Xi) \quad d\Upsilon_i] \tag{4}$$

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.

Equations (3-4) further demonstrate that this formulation allows for any within-contingency-table sampling scheme to be implemented, so long as one can write down $p(\mathbf{K}_i|\boldsymbol{N}_{comp_i})$. Note, however, that the exchangeability assumption (reflected in the $\prod_{i=1}^{\mathrm{I}}$) prevents incorporation of contingency-table-level sample weights into the likelihood. In other words, equation (3-4) does not take into account whether the contingency tables 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, equations (3-4) demonstrates a variety of choices of likelihoods, priors, and hyperpriors for count models are available. We next discuss our choices.

## 2 Our Proposal

We provide the specifics of our proposal, which extends that in (Greiner and Quinn (2009)); in doing so, we demonstrate how our ideas fit within the frameworks of equations (3-4) and we justify our distributional assumptions in the context of voting applications.

In the language of equations (3-4), our proposal consists of the following. For $\boldsymbol{N}_{comp_i}|\Upsilon_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}}$, we assume that the counts in each contingency table row follow an (independent) multinomial distribution with count parameter $\mathrm{N}_{r_i}$ and probability parameter $\underset{\sim}{\theta}_{r_i}$. We choose the multinomial for the following reasons. First, the multinomial corresponds to an individual-level account of voting behavior. If the vote choice of each potential voter of race r in precinct i independently follows from the same vector $\underset{\sim}{\theta}_{r_i}$, the sum of the choices of voters of race r in precinct i will follow a multinomial as specified. Second, once one conditions on the row totals (as is customary in voting applications), few other tractable multivariate count distributions are available.

Next, for $\Upsilon_i|\Xi$, we apply a multidimensional additive logistic transformation (Aitchison (2003)) to each row's $\underset{\sim}{\theta}_{r_i}$, resulting in R vectors of dimension (C-1), which we stack to form a single vector $\underset{\sim}{\omega}_i$ of dimension R*(C-1) for each precinct. We then assume $\underset{\sim}{\omega}_i \overset{i.i.d.}{\sim} N(\underset{\sim}{\mu}, \mathbf{\Sigma})$. 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 easy choice of a "reference category" in voting applications, namely, the Abstain column. The stacking of the transformed $\underset{\sim}{\theta}_{r_i}$'s into a single vector allows for exploration of within- and between-row relationships; as we demonstrate in our application (see Table 3), capacity to model between-row relationships can be important to inference. The multivariate normal form facilitates various extensions, including additional robustness via a switch to a multivariate t, additional hierarchical modeling (corresponding to simultaneous analysis of multiple elections), and incorporation of aggregate-level variables (by changing individual elements of $\underset{\sim}{\mu}$ into regressions).

For the hyperprior ($p(\Xi)$), we use semi-conjugate multivariate normal and inverse Wishart forms, specifically $\underset{\sim}{\mu} \sim N(\underset{\sim}{\mu}_0, \boldsymbol{\kappa}_0)$ and $\mathbf{\Sigma} \sim InvWish_{\nu_0}(\mathbf{\Psi}_0)$. We do so both for computational convenience and because, after extensive simulations, we have found that we can express most reasonable prior beliefs regarding the content of the contingency tables (or functions thereof) by manipulating the parameters of these hyperpriors.

For $p(\mathbf{K}_i|\boldsymbol{N}_{comp_i})$, we assume a simple random sample. We do so not out of desire but 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 a relatively rare event (often true in the disease context). All of these schemes have one thing in common, however: they depend on the assumption that the researcher can observe some characteristic of an individual unit before deciding whether to include it in the sample. Unfortunately, such is not always the case in many voting applications, particularly exit polls. Voters exit polling locations rapidly, and operationally, 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. To use one of the optimization schemes, exit pollsters need to discern, quickly and accurately, the race of persons leaving the polling area. Because one of the purposes of our City of Boston exit poll was to find out precisely whether pollsters can make such judgments accurately, we could not attempt to implement a subsampling scheme in our poll.

If the exit poll constitutes a simple random sample in each $\mathrm{i} \in \mathrm{S}$, we can simplify the corresponding notation. In the ith precinct, we can work with the R*C-dimension vector $\underset{\sim}{\mathrm{K}}_{\mathrm{i}}$ 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 $\underset{\sim}{\mathrm{K}}_{\mathrm{i}}$ as $\mathrm{K}_{\mathrm{rc_i}}$, $\mathrm{K}_{\mathrm{i}} = \sum_{\mathrm{r,c}} \mathrm{K}_{\mathrm{rc_i}}$ and for each $\mathrm{i} \in \mathrm{S}$, recall $M_{rc_i} = N_{rc_i} - \mathrm{K}_{\mathrm{rc_i}}$. Accordingly, the probability of observing a particular vector $\underset{\sim}{\mathrm{K}}_{\mathrm{i}}$ is the familiar $\frac{\prod_{r,c} \binom{M_{rc_i} + \mathrm{K}_{\mathrm{rc_i}}}{\mathrm{K}_{\mathrm{rc_i}}}}{\binom{\mathrm{N_i}}{\mathrm{K_i}}}$ (see McCullagh and Nelder (1989)).

Upon discarding terms for $\mathrm{i} \in \mathrm{S}$ 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.

$$p(\underset{\sim}{\mu}, \mathbf{\Sigma}|\mathbf{K},\ \mathbf{N}_{\text{col}},\ \mathbf{N}_{\text{row}})) \propto N(\underset{\sim}{\mu}|\underset{\sim}{\mu}_0, \boldsymbol{\kappa}_0) \times Inv - Wish_{\nu_0}(\mathbf{\Sigma}|\mathbf{\Psi}_0) \tag{5}$$

$$\prod_{\text{i}\in\text{S}} \Bigg[ \int \sum_{\boldsymbol{M}_{miss_i}} \left( \prod_{r,c} \frac{\theta_{rc_i}^{M_{rc_i}+\text{K}_{\text{rc}_\text{i}}-1}}{M_{rc_i}!} \right) \times \tag{6}$$

$$\prod_{c} I \left( \sum_{r} M_{rc_i} = \text{N}_{\text{c}_\text{i}} - \sum_{r} \text{K}_{\text{c}_\text{i}} \right) \times \tag{7}$$

$$\prod_{r} I \left( \sum_{c} M_{rc_i} = \text{N}_{\text{r}_\text{i}} - \sum_{c} \text{K}_{\text{r}_\text{i}} \right) \tag{8}$$

$$\left( |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\underset{\sim}{\omega}_i - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{\omega}_i - \underset{\sim}{\mu}) \right\} \right) \times \tag{9}$$

$$\prod_{r} I \left( \sum_{c} \theta_{rc_i} = 1 \right) d\underset{\sim}{\theta}_i \Bigg] \tag{10}$$

$$\prod_{\text{i}\notin\text{S}} \Bigg[ \int \sum_{\boldsymbol{N}_{miss_i}} \left( \prod_{r,c} \frac{\theta_{rc_i}^{N_{rc_i}-1}}{N_{rc_i}!} \right) \times \tag{11}$$

$$\prod_{c} I \left( \sum_{r} N_{rc_i} = \text{N}_{\text{c}_\text{i}} \right) \times \tag{12}$$

$$\prod_{r} I \left( \sum_{c} N_{rc_i} = \text{N}_{\text{r}_\text{i}} \right) \times \tag{13}$$

$$\left( |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\underset{\sim}{\omega}_i - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{\omega}_i - \underset{\sim}{\mu}) \right\} \right) \times \tag{14}$$

$$\prod_{r} I \left( \sum_{c} \theta_{rc_i} = 1 \right) d\underset{\sim}{\theta}_i \Bigg] \tag{15}$$

Lines 7, 8, 10, 12, 13, and 15 impose the constraints that the internal cell counts sum to their observed row and column totals and that the probability parameters for each row sum to one. Lines 7 and 8 reflect the fact that for i $\in$ S, some of the internal cell counts are observed in the survey, requiring a corresponding adjustment to the bounds. Lines 6 and 14 are the prior, and line 5 the hyperprior, for the internal cell counts. Lines 6 and 11 correspond to the multinomial assumption for the internal cell counts. The expressions in lines 6-8 ease understanding of the primary contribution that the survey makes to the information in the posterior. As $\text{K}_{\text{rc}_\text{i}}$ gets large, $M_{rc_i} = N_{rc_i} - \text{K}_{\text{rc}_\text{i}}$ decreases, reducing the uncertainty in the exponent of the numerator of $\frac{\theta_{rc_i}^{M_{rc_i}+\text{K}_{\text{rc}_\text{i}}-1}}{M_{rc_i}!}$ and driving the denominator to 1. If $\text{K}_\text{i} = \text{N}_\text{i}$, then this portion of the posterior corresponds to the non-constant portion of the likelihood of the probability vector of a multinomial distribution. Meanwhile, the larger each $\mathrm{K}_{\mathrm{rc}_\mathrm{i}}$, the tighter the bounds induced on $M_{rc_i}$, as is clear from lines 7-8.

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 $\Lambda_{rc} = \frac{\sum_i N_{rc_i}}{\sum_i (\mathrm{N}_{\mathrm{r}_\mathrm{i}} - N_{rA_i})}$, $\Gamma_r = \frac{\sum_i (\mathrm{N}_{\mathrm{r}_\mathrm{i}} - N_{rA_i})}{\sum_i (\mathrm{N}_\mathrm{i} - \mathrm{N}_{\mathrm{A}_\mathrm{i}})}$, and $TO_{rc} = \frac{\sum_i (\mathrm{N}_{\mathrm{r}_\mathrm{i}} - N_{rA_i})}{\sum_i \mathrm{N}_{\mathrm{r}_\mathrm{i}}}$ 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 that draws the internal cell counts ($N$'s and $M$'s), then the intermediate level parameters ($\theta$'s or $\omega$'s), then the upper level parameters ($\underset{\sim}{\mu}$ and $\mathbf{\Sigma}$). The draws of the internal cell counts allow calculation of functions of interest at each iteration, although in practice we thin the draws considerably.

Speed is a serious concern here. The appendix has some details, but depending on the constraints imposed by 7-8 and 11-12, ecological data can have little information in them. 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 requirement of modern United States voting rights litigation, may require special computation 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 × 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. By way of explanation, aggregation bias, also called a contextual effect, 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, the aggregation process will induce bias in almost any ecological estimator. Further, 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 (if at all) 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 an analyst or an expert witness might be interested in whether voting is "racially polarized" (meaning voting patterns of racial groups are different). 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 heavier weights to racially integrated precincts, and the third ("Sampling Scheme 3") applied much heavier weights to racially uniform precincts. The potential black, white, or Hispanic voters had population fractions (in expectation) of 35%, 45%, and 20%. Turnout was in expectation approximately 45%, 50%, and 40%, respectively. Potential voters could vote Democrat, vote Republican, or Abstain, and Democrat support rates among actual voters ($\Lambda_{rc}$'s) were (in expectation) 80%, 35%, and 50% for the three racial groups. Parameters were chosen to make the simulations noisy, so for example $\Lambda_{hD}$ varied from below 20% to above 80%. The appendix has additional details.

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. When the results surprised us (as explained below), we conducted additional simulations. To induce aggregation bias, we turned the top-level (normal distribution) location parameters for whites ($\mu_{wD}$ and $\mu_{wR}$) into linear functions of the fraction Hispanic, $\mathrm{X}_{h_i} = \frac{\mathrm{N_{h_i}}}{\mathrm{N_i}}$. For aggregation bias (severe aggregation bias), we chose coefficients such that, in a 20% Hispanic precinct, roughly $\frac{1}{4}$ $(\frac{1}{3})$ of whites voters in expectation would vote for the Democrat, but in a 80% Hispanic precinct, roughly $\frac{2}{5}$ $(\frac{9}{10})$ of white voters in expectation would vote for the Democrat. We chose the figures for the aggregation bias simulations based on an informal survey of persons knowledgeable in the field of voting rights and racial bloc voting. In contrast, we believe the figures for the severe aggregation bias to be unrealistically harsh, and we include the results to demonstrate the performance of the hybrid at an extreme outer limit.

We present here the results for the quantity $\Lambda_{hD}$. We focus on this quantity for several reasons. First, voting patterns for the racial group with the lowest population and the lowest turnout have proven in past to be the most difficult to estimate. Second, estimates of $\Lambda_{hD}$ were likely to be particularly vulnerable to the aggregation bias we induced because the data would associate an increase in fraction Hispanic with an increase in Democratic votes, even though this association did not in expectation stem from the voting of Hispanics. Third, prior experience with our ecological model taught us that the $\Lambda$ quantities, as non-linear functions of the counts that depend solely on within-contingency-table-row behavior, were among the hardest to estimate well.

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.

### 3.2 Simulation Results

Basic results are summarized in Figures 1 and 2. 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 30-50% 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 overweights contingency tables in which one row dominates, *i.e.*, racially homogenous precincts. Without aggregation bias, the difference between the hybrid that overweights 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 overweights racially mixed precincts, achieves 20-30% reduction as compared to Sampling Scheme 3, which overweights racially uniform precincts (all p-values are 0).

As an aside, we note a somewhat disturbing lack of coverage in the 95% intervals. The scale of Figure 3 makes this difficult to see, but interval coverage for the pure sampling estimators ranged from .84 to .97 with an unusual number of values at or below .90. It is not clear that the t-based interval is wide enough to assure nominal coverage.

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 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 quite modest, roughly two or three percentage points (*i.e.*, meaning 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. 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; this number so exceeded our targets as to cause coordination issues, which we addressed by requiring a law or graduate student shift captain to oversee each location's team. 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 visible box (see Bishop and Fisher (1995)). Five specially trained, two-person roving quality control teams, which circulated in cars visiting each polling location multiple times throughout election day, monitored 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.

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 appendix 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 was 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 Table 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 et al. (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 and dog racing ballot initiatives, where our two-party projections were accurate, where non-voting overreport was comparatively low, and where the two-party vote was closest.

| | Two-Party | | | | | Voted | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Contest | PtEst | SE | .025 | .975 | Truth | PtEst | SE | .025 | .975 | Truth |
| Pres | .80 | .023 | .76 | .85 | .80 | .81 | .013 | .79 | .84 | .98 |
| Tax | .29 | .015 | .26 | .32 | .23 | .85 | .014 | .83 | .88 | .92 |
| Pot | .68 | .021 | .64 | .72 | .71 | .86 | .013 | .83 | .88 | .92 |
| Dogs | .56 | .013 | .49 | .63 | .56 | .84 | .012 | .82 | .87 | .90 |

Table 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. Point estimates and standard errors stem from multiple imputation, quantiles stem from t intervals. 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 (`http://cran.r-project.org/web/packages/cat/index.html`). 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 attributes, 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 aggregation bias. With respect to this dataset, however, the no-aggregation-bias assumption is slightly stronger for the hybrid estimator than for the pure ecological counterpart. The reason is that exit polls survey voters by polling location, not by precinct, and that in the City of Boston, many polling locations host voters more than one precinct. Precincts housed in the same voting location are separate in that they use separate check-in tables and equipment in the room in which voting occurs, but pollsters standing outside of buildings are unable to distinguish voters from different precincts. The 39 polling locations included in the Boston Area Colleges Exit Poll represented 69 precincts. Although technically what matters is the aggregation process and not the number of individuals (voters) aggregated, our sense is that the no-aggregation-bias assumption is generally more plausible when the level of aggregation is lower. Therefore, the hybrid estimators presented below depend on the assumption that, for the voters that voted in the in-sample polling locations, the no-aggregation-bias assumption holds approximately at the level of the polling location, which is a typically higher level of aggregation than the precinct. Recall, however, that as detailed above, the hybrid's incorporation of the exit poll provides substantial protection on this score. Note that out-of-sample precincts were not aggregated further.

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.[1] We decided to see 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 the two tables immediately below and in Figure 3. 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.

[1] Recalling that Census 2000 allowed respondents to mark more than one race box, these categories are in fact shorthand for the following: "Asian" means any part Asian, "Hispanic" means non-Asian but Hispanic (regardless of any other race box checked), "black" means non-Asian non-Hispanic any part black, and "white" means anyone left not who was not in the other race category.

2008 MA Marijuana Ballot Initiative: Comparison of Estimators

| Race | Est. | Mean | SE | .025 | .975 |
|---|---|---|---|---|---|
| bla | EI | .73 | .010 | .71 | .75 |
| bla | S | .66 | .033 | .60 | .73 |
| bla | H | .72 | .009 | .70 | .74 |
| whi | EI | .71 | .005 | .70 | .72 |
| whi | S | .71 | .029 | .66 | .77 |
| whi | H | .71 | .004 | ***.71*** | ***.72*** |
| his | EI | .62 | .065 | .48 | .73 |
| his | S | .65 | .034 | .58 | .71 |
| his | H | .66 | .026 | ***.61*** | ***.71*** |
| asa | EI | .70 | .290 | .03 | .99 |
| asa | S | .51 | .084 | .34 | .68 |
| asa | H | .63 | .047 | ***.54*** | ***.73*** |

*Estimated preferences for four most numerous racial groups in the City of Boston. "EI" refers to ecological inference estimator alone, "S" is the exit poll alone, "H" is the hybrid estimator. For "S" and "H", the figures come from multiple imputation. Note that . . .*

2008 MA Dog Racing Ballot Initiative: Comparison of Estimators

| Race | Est. | Mean | SE | .025 | .975 |
|---|---|---|---|---|---|
| bla | EI | .43 | .013 | .41 | .46 |
| bla | S | .49 | .041 | .41 | .57 |
| bla | H | .45 | .009 | .43 | .47 |
| whi | EI | .59 | .006 | .58 | .60 |
| whi | S | .60 | .050 | .50 | .70 |
| whi | H | .60 | .004 | .59 | .61 |
| his | EI | .67 | .072 | .53 | .81 |
| his | S | .56 | .080 | .39 | .72 |
| his | H | .56 | .027 | .51 | .61 |
| asa | EI | .80 | .242 | .12 | .99 |
| asa | S | .56 | .083 | .39 | .72 |
| asa | H | .62 | .057 | ***.51*** | ***.73*** |

*in both the marijuana the dog racing ballot initiatives, only the hybrid estimator offers enough precision in the Asian estimates to allow substantive inference. The hybrid also better differentiates Hispanic from white preferences on the marijuana initiative.*

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. 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 the help is most needed. The variance reduction available for the estimates of Asian and Hispanic voting behavior in the greyhound ballot initiative is substantial. As the two racial groups with the lowest VAP and lowest turnout, Hispanics and Asians represent the most difficult challenge to inferences 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 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 relationship.

Several commentators have noted the difficulty in estimating model parameters that govern behavior between (as opposed to within) contingency table rows. For example, King (1997) notes that the $\rho$ parameter relating African-American to Caucasian voting behavior in the truncated bivariate normal model can be poorly estimated, while we have previously documented the greater uncertainty associated with the estimation of between-contingency-table-row correlations as opposed to within-row correlations (see 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 mulinomial 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 ($\mathbf{\Sigma}$) 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:

$$\mathbf{\Sigma} \quad = \quad \begin{bmatrix} \mathbf{\Sigma}_b & \mathbf{\Sigma}_{bw} & \mathbf{\Sigma}_{bh} & \mathbf{\Sigma}_{ba} \\ \mathbf{\Sigma}_{bw} & \mathbf{\Sigma}_w & \mathbf{\Sigma}_{wh} & \mathbf{\Sigma}_{wa} \\ \mathbf{\Sigma}_{bh} & \mathbf{\Sigma}_{wh} & \mathbf{\Sigma}_h & \mathbf{\Sigma}_{ha} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{wa} & \mathbf{\Sigma}_{ha} & \mathbf{\Sigma}_a \end{bmatrix}$$

See also Greiner and Quinn (2009) (page 71) for a similar representation. Note that each of $\mathbf{\Sigma}_{ba}$, $\mathbf{\Sigma}_{wa}$, and $\mathbf{\Sigma}_{ha}$ is of dimension 2 × 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 $\mathbf{\Sigma}_{ba}$, $\mathbf{\Sigma}_{wa}$, and $\mathbf{\Sigma}_{ha}$. Figure 3 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; we hypothesize that this 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 × 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 × C context as well as the implications of different contingency-table-level sampling schemes. We have confronted the challenges in 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.

# A Technical Appendix

Our software fits the model articulated in lines (5-13) using data augmentation (see Tanner and Wong (1987)). We draw iteratively from conditional posteriors $p(\mathbf{\Sigma}|\underset{\sim}{\mu},\ \{\underset{\sim}{\omega}_i\},\ \nu_0,\ \mathbf{\Psi}_0)$; $p(\underset{\sim}{\mu}|\{\underset{\sim}{\omega}_i\},\ \mathbf{\Sigma},\ \underset{\sim}{\mu}_0,\ \boldsymbol{\kappa}_0)$; $p(\underset{\sim}{\theta}_i|\boldsymbol{N}_{comp_i},\ \underset{\sim}{\mu},\ \mathbf{\Sigma})$; $p(\boldsymbol{N}_{miss_i}|\underset{\sim}{\theta}_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}},\ \underset{\sim}{\mathrm{N}}_{\mathrm{col_i}})$ for i $\notin$ S; and $p(\boldsymbol{M}_{miss_i}|\underset{\sim}{\theta}_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}},\ \underset{\sim}{\mathrm{N}}_{\mathrm{col_i}},\ \underset{\sim}{\mathrm{K}}_{\mathrm{i}})$ for i $\in$ S. The first two distributions are in standard form. We draw from $p(\underset{\sim}{\theta}_i|\boldsymbol{N}_{comp_i},\ \underset{\sim}{\mu},\ \mathbf{\Sigma})$ using a Metropolis-Hastings (Metropolis et al. (1953)) step, generating a proposal in $\underset{\sim}{\omega}$ space from a multivariate $t_4(\underset{\sim}{\mu}^{(t)}, \gamma_i \mathbf{\Sigma}^{(t)})$ and transforming back to $\underset{\sim}{\theta}$. $\gamma_i$ is a (constant) tuning parameter set in initial runs, and $^{(t)}$ denotes the iteration.

We experimented with several methods of drawing from $p(\boldsymbol{N}_{miss_i}|\underset{\sim}{\theta}_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}},\ \underset{\sim}{\mathrm{N}}_{\mathrm{col_i}})$ (drawing from $p(\boldsymbol{M}_{miss_i}|\underset{\sim}{\theta}_i,\ \underset{\sim}{\mathrm{N}}_{\mathrm{row_i}},\ \underset{\sim}{\mathrm{N}}_{\mathrm{col_i}},\ \underset{\sim}{\mathrm{K}}_{\mathrm{i}})$ is a nearly identical process). We ultimately settled on a Metropolis-Hastings step in which we choose a row r*, generate R-1 independent proposals from $Multinom(\mathrm{N}_{\mathrm{r_i}},\ \underset{\sim}{\theta}_{r_i})$ for r $\neq$ r*, and calculate the cell counts for the r* row deterministically. A key concern here is the choice of r*, because this method of generating proposals can lead to values for $\boldsymbol{N}_{miss_i}$ that violate the bounds. We have had good luck by choosing r* randomly, but making rows with large counts bear a higher probability of being chosen as r* because variation in lower-count rows is less likely to generate an impossible proposal. An advantage of this overall method is that, due to cancellation of terms, the Metropolis-Hastings ratio depends solely on the r*th row and has the form (on the log scale) $\sum_c\ (N^p_{\mathrm{r^*}c_i} - N^{(t-1)}_{\mathrm{r^*}c_i}) log\theta^{(t-1)}_{\mathrm{r^*}c_i}\ +\ logN^{(t-1)}_{\mathrm{r^*}c_i}!\ -\ logN^p_{\mathrm{r^*}c_i}!$, where the superscript p denotes "proposal" and the superscript t refers to the iteration. This quantity can be rapidly calculated using a log-factorial lookup table.

Regarding our simulations, we began by drawing $(\mathrm{X}_{b_i}, \mathrm{X}_{w_i}, \mathrm{X}_{h_i}) \sim Diri(\alpha * (.35, .45, .2))$, where $\mathrm{X}_{b_i} = \frac{\mathrm{N_{b_i}}}{\mathrm{N_i}}$ and so on. We set $\alpha$ to 3 (1.3) for integrated (less integrated) datasets; this value corresponds to approximately 4% (20%) of precincts in which one row contains 90% or more of the contingency table's counts. We have encountered values in this range in analyzing redistricting data. We drew precinct sizes from a $Poi(Q)$, varying $Q$ from 700 to 1700 uniformly, and calculated

$\underset{\sim}{\mathrm{N}}_{\mathrm{row}_\mathrm{i}}$ deterministically (rounding to whole numbers). We drew $\underset{\sim}{\mu} \sim N(\underset{\sim}{\mu}_0, \kappa_0 * \boldsymbol{I}_6)$, initially setting $\underset{\sim}{\mu}_0 = (-.6, -2.05, -1.7, -.2, -1.45, -1.45)$ and $\kappa_0 = .3$. When varying expected $\Lambda_{hD}$, we set $\underset{\sim}{\mu}_0$ alternately to $(-.6, -2.05, -1.7, -.2, -1, -2.55)$ and $(-.6, -2.05, -1.7, -.2, -2.55, -1)$ to obtain $\Lambda_{hD}$ approximately .75 and .25. When inducing severe aggregation bias, we made the the middle two elements of $\underset{\sim}{\mu}_0$ equal to $(-1.7 + 1.7\mathrm{X}_{h_i}, .2 + (-3)\mathrm{X}_{h_i})$. These values induced an expected $\Lambda_{wD}$ in a 20% Hispanic precinct of .33 as compared to .90 for an 80% Hispanic precinct. We simulated $\underset{\sim}{\omega}_i$ from the resulting normal, transformed to $\theta$ space, drew $\boldsymbol{N}_{comp_i}$ from the relevant multinomials, and calculated $\underset{\sim}{\mathrm{N}}_{\mathrm{col}_\mathrm{i}}$ via addition.

To generate our surveys, we sampled 40 precincts with unequal sampling weights. The weights were calculated using a function of the quantity $g(herf) = \mathrm{X}_{b_i}^{herf} + \mathrm{X}_{w_i}^{herf} + \mathrm{X}_{h_i}^{herf}$. For S1, the weight for the ith precinct was $\frac{1}{g(3.5)}$; for S2, $\frac{1}{g(2)}$; for S3, $g(3.5)$. The inversion in S1 and S2 results in overweighting of racially mixed precincts. For in-sample precincts, we draw a simple random sample of approximately $\frac{\mathrm{N_i}}{14}$ voters from $\boldsymbol{N}_{comp_i}$. For our City of Boston poll, we used weights equal to $\frac{1}{g(2)}$.

For the simulations, for the pure ecological and hybrid estimators, we ran three chains of 2,000,000 iterations each. For our substantive analysis on the City of Boston data, we ran three chains of 30,000,000 iterations each. Convergence diagnostics (Heidelberger and Welch (1983) and Geweke (1992)) were generally unremarkable in both the simulations and the City of Boston exit poll.

For the City of Boston data, we randomly allocated the “other” race population to the four principle groups (black, white, Hispanic, and Asian) in rough proportion to their numbers. In addition, population shifts since 2000 had rendered one precinct’s VAP total obviously incorrect because it was less than the number of votes cast in that precinct. For this precinct, we kept the racial proportions from Census 2000 but assigned this precinct a total VAP according to the average voter-to-VAP ratio from the City’s 252 other precincts.

Speed remains a concern. For a 175-precinct simulated dataset, a single chain of 2,000,000 iterations takes just under an hour to run. Such speed may serve for research purposes; in litigation, an expert witness may need to analyze dozens of elections over the space of one or two weeks, a task which would require special computational tools. The addition of survey information can save computer time by speeding chain convergence and thus allowing shorter chains (convergence diagnostics for the hybrid estimators were generally better than those for the pure ecological inference estimator). The three chains of 30,000,000 iterations for the City of Boston data took several days; the number of iterations was designed to eliminate any possible concern regarding chain convergence and was excessive, but the low VAP and turnout of Asian-Americans did require long chains.

# References

Achen, Christopher H., and W. Phillips Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Aitchison, J. 2003. *The Statistical Analysis of Compositional Data*. Caldwell, NJ: The Blackburn Press, second edition.

Bishop, George F., and Bonnie S. Fisher. 1995. "Secret Ballots And Self-Reports in an Exit-Poll Experiment." *Public Opinion Quarterly* 59:568–588.

Brown, Philip J., and Clive D. Payne. 1986. "Aggregate Data, Ecological Regression, and Voting Transitions." *Journal of the American Statistical Association* 81:452–460.

Duncan, Otis D., and Beverly Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18(6):665–666.

Gelman, Andrew, Stephen Ansolabehere, Phillip N. Price, David K. Park, and Lorraine C. Minnite. 2001. "Models, Assumptions, And Model Checking in Ecological Regressions." *Journal of the Royal Statistical Society A, Part 1* 164:101–118.

Geweke, John. 1992. "Evaluating the Accuracy of Sampling-Based Approaches To Calculating Posterior Moments." In *Bayesian Statistic 4: Proceedings of the Fourth Valencia International Meeting* ( J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors), Oxford: Clarendon Press.

Glynn, Adam N., Jon Wakefield, Mark S. Handcock, and Thomas S. Richardson. 2008. "Alleviating Linear Ecological Bias And Optimal Design with Subsample Data." *Journal of the Royal Statistical Society, Series A* 171(1):179–202.

Glynn, Adam N., Jon Wakefield, Mark S. Handcock, and Thomas S. Richardson. 2009. "Alleviating Ecological Bias in Generalized Linear Models with Optimal Subsample Design." On file with authors.

Greiner, D. James, and Kevin 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.

Haneuse, Sebastien J., and Jonathan C. Wakefield. 2008. "The Combination of Ecological And Case-Control Data." *Journal of the Royal Statistical Society, Series B* 70(1):73–93.

Heidelberger, Philip, and Peter D. Welch. 1983. "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research* 31(6):1109–1144.

Honaker, James, and Gary King. 2009. "What to do About Missing Values in Times Series Cross-Section Data." Harvard University Working Paper.

Judge, George, Douglas J. Miller, and Wendy K. Tam Cho. 2004. "An Information Theoretic Approach to Ecological Estimation and Inference." In *Ecological Inference: New Methodological Strategies* ( Gary King, Ori Rosen, and Martin A. Tanner, editors), Cambridge: Cambridge University Press.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton University Press.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Wiley-Interscience, second edition.

McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models*. Chapman & Hall/CRC, second edition.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21(6):1087–1092.

Raghunathan, Trivellore E., Paula K. Diehr, and Allen D. Cheadle. 2003. "Combining Aggregate And Individual Level Data To Estimate an Individual Level Correlation Coefficient." *Journal of Education and Behavioral Statistics* 28(1):1–19.

Robinson, W.S. 1950. "Ecological Correlations And the Behavior of Individuals." *American Sociological Review* 15(3):351–357.

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A. Tanner. 2001. "Bayesian And Frequentist Inference for Ecological Inference: the R x C Case." *Statistica Neerlandica* 55(2):134–156.

Silver, Brian D., Barbara Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting." *American Political Science Review* 80(2):613–624.

Steel, D., M. Tranmer, and D. Holt. 2003. *Analysis of Survey Data*, Wiley, chapter Analysis Combining Survey and Geogrphically Aggregated Data.

Tanner, M. A., and W. H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82:528–540.

Wakefield, Jon. 2004. "Ecological Inference for 2 x 2 Tables." *Journal of the Royal Statistical Society, Series A* 167(3):385–445.

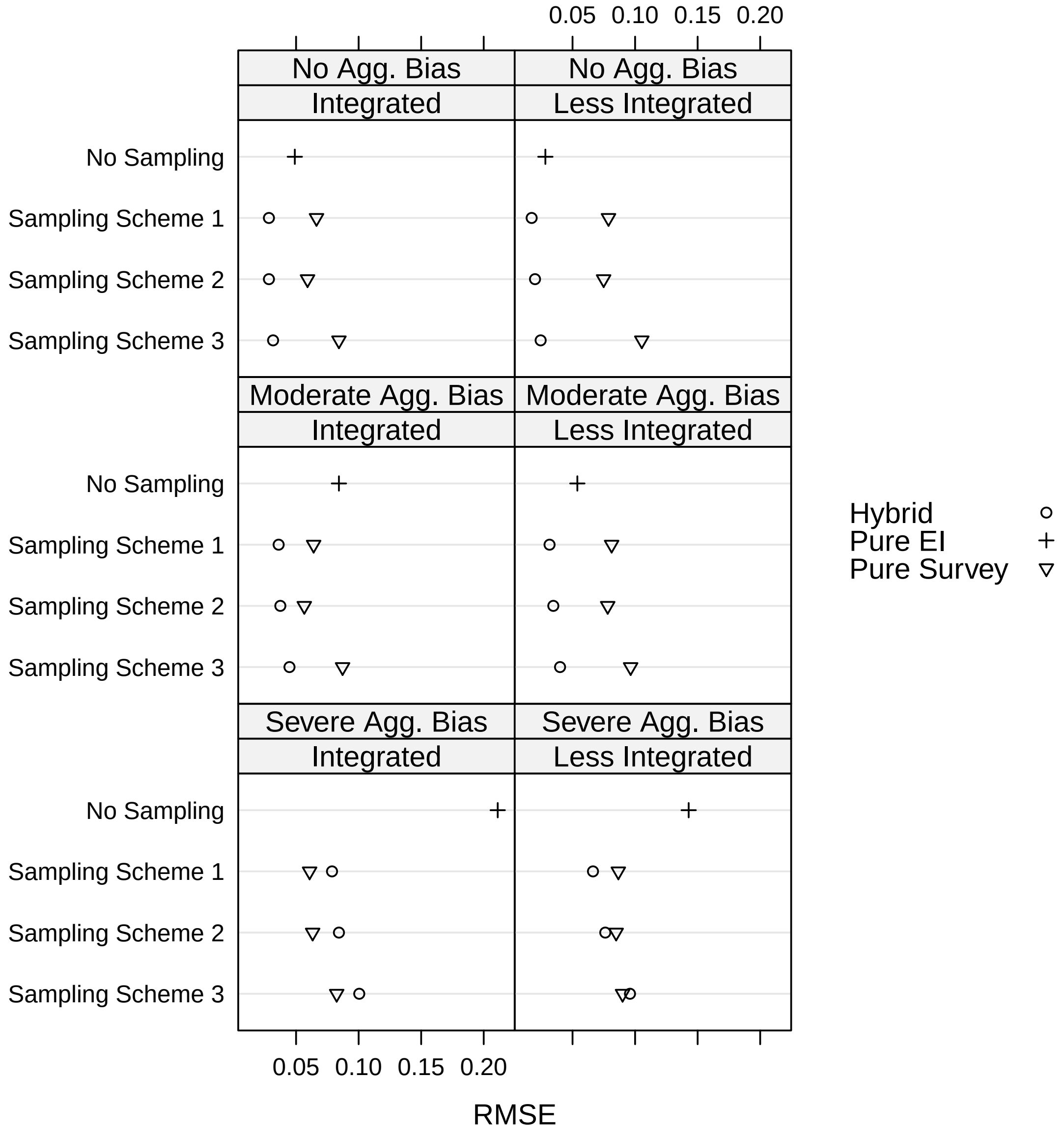

Figure 1: *Summary of RMSE results from simulations Sampling Scheme 1 heavily overweights racially mixed precincts, Sampling Scheme 2 mildly overweights racially mixed precincts, and Sampling Scheme 3 heavily overweights racially uniform precincts. Note that "Integrated" datasets have less information in the bounds (Duncan and Davis (1953)). The results show that the hybrid estimator generally outperforms the pure survey and pure ecological inference estimators, and offers substantial RMSE reductions in some circumstances.*

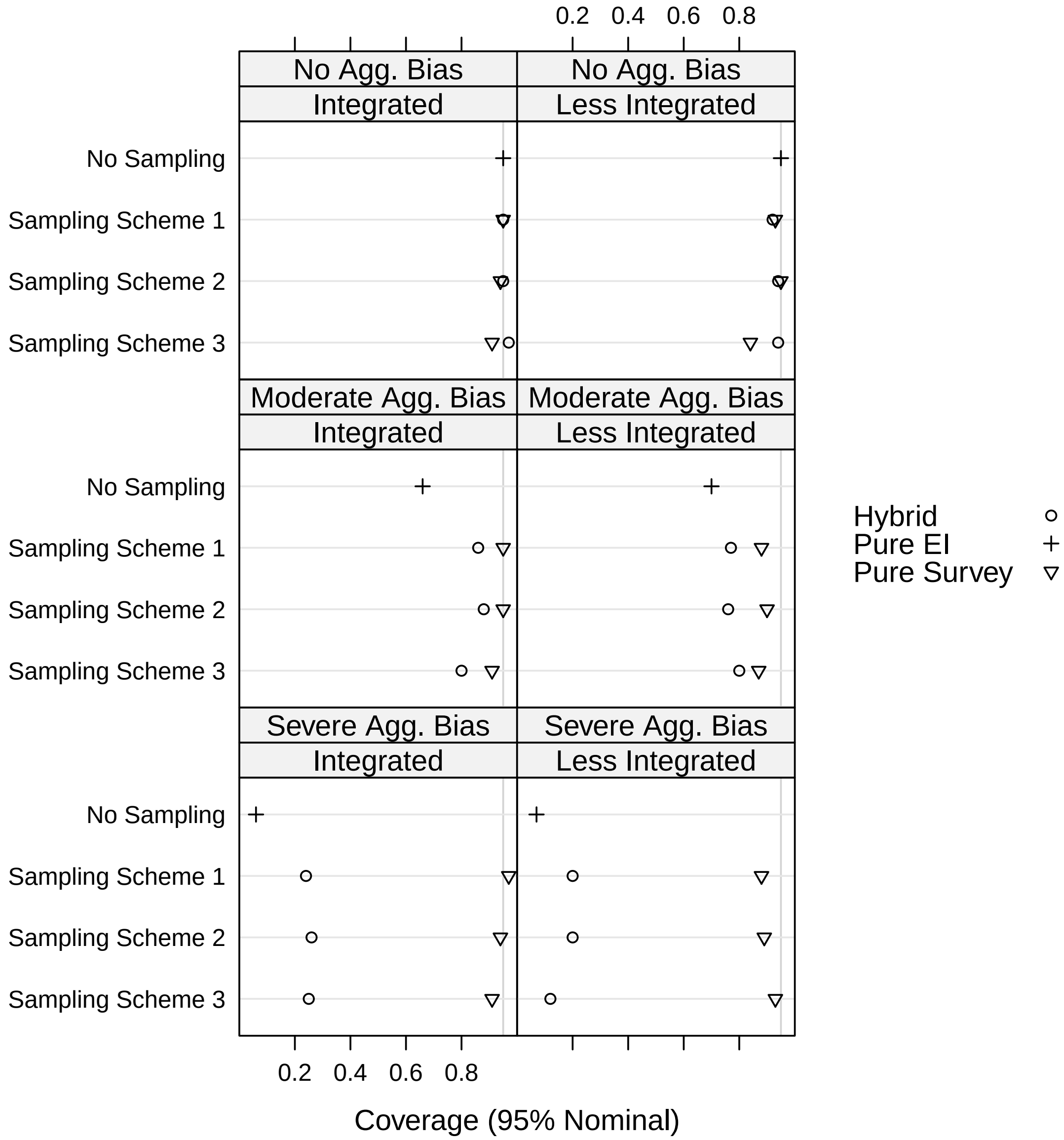

Figure 2: *Coverage of nominal 95% credible intervals from simulations. The gray vertical line in each panel is at the nominal 95% coverage probability. The Sampling Schemes are explained in the caption to Figure 2, and again,"Integrated" datasets have less information in the bounds. The results show that 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.*

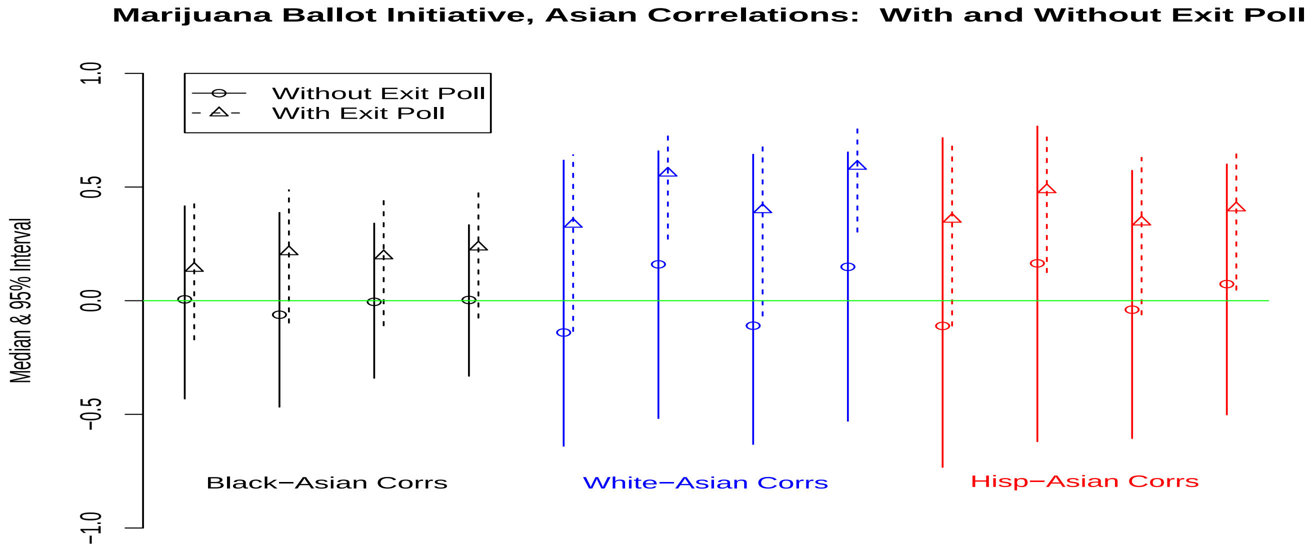

Figure 3: *Comparison of posterior distributions from ecological inference, with and without exit poll, of between-contingency-table-row correlations governing the relationship of black, white, and Hispanic voters to Asian voters. The with-exit-poll figures are averages of ten multiple imputations. The 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 the relationships among various racial groups to provide better estimates of Asian voting behavior, something the without-exit-poll model is unable to accomplish.*