The Role of Confounders and Linearity in Ecological Inference: A Reassessment

Shiro Kuriwaki[*]
Department of Political Science
Yale University

Cory McCartan
Department of Statistics
Pennsylvania State University

September 1, 2025

Abstract

Estimating conditional means only with aggregate data is commonly known as the ecological inference problem (EI). We provide a reassessment of ecological inference: a formalization of the problem that differs from existing work, identification conditions, and an empirical characterization of how these conditions fail in common cases. In particular, the quantity of interest in EI is governed by a conditional expectation function, and identification requires estimating this function controlling for confounders. In this way, EI is similar to causal inference with observational data, but with aggregation contributing additional structure to assist in the estimation problem. Using this perspective, we clarify the differences between the EI methods commonly used in the literature, and explain when they lead to ecological fallacies. Using datasets for common EI problems in which the ground truth is fortuitously observed, we show how all methods are prone to overestimating racial polarization and underestimating ticket splitting, but covariates can help.

1 Introduction

Estimating conditional means with only aggregate data is commonly known as the *ecological inference* problem (EI). These estimation challenges occur in political science, where for example voter's choices are aggregated to geographical districts, and in economics (where consumer's purchasing choices for goods are aggregated into regional markets), public health (where resident's health outcomes are aggregated into census areas to preserve privacy), and to other swaths of research that use census statistics. Researchers since Robinson (1950) have been aware that such relationships between aggregate data may not correspond to the underlying relationship between individuals, calling incorrect inferences an *ecological fallacy*. They have produced a long literature with statistical methods to make valid inferences from aggregate data. Even after the publication of King's (1997) *A Solution to the Ecological Inference Problem*, 20 distinct sets of authors have proposed methods and adjustments proposed for ecological inference (Appendix A).

However, by treating EI almost as a unique problem requiring unique solutions, the methodological literature may have obscured its similarities to more general statistical problems. And despite the

[*]To whom correspondence should be addressed.

range of existing methods, the concern that data from aggregates could result in an ecological fallacy persists. Some practitioners take a cautious stance, refusing to make any inferences with ecological data. A wider community (including those beyond academia) uses the existing EI methods regularly, and sometimes without interrogating the possibility of an ecological fallacy. As a result, while students may learn that inferences with aggregate data always carry the risk of a ecological fallacy, many do not learn the conditions under which these fallacies can occur, or why the occur in the first place.

In this paper, we provide a reassessment of ecological inference. We go beyond a mere review of past work, and provide both intuition and formal identification conditions that differ from the existing literature. Our reformulation casts the ecological inference problem in terms of causal inference and regression modeling. This move, we demonstrate, addresses most directly when and how ecological fallacies occur. Our reassessment is also empirical. We evaluate the accuracy of the three most common methods as well as our own proposed alternative estimator. We evaluate the four methods on two common examples in political science, and show why estimates tend to underestimate or overestimate the quantity of interest.

Our paper consists of three major parts. After an illustrative example of the ecological fallacy, we first define our quantities of interest and the estimation challenge. The goal is to identify a conditional expectation of an outcome Y conditional on a categorical predictor variable X, using data that has been coarsened into groups that are a mix of various levels of X. For example, we are interested in the population average of vote choice conditional on racial group identification, but electoral districts coarsen observed data into groups that are each a mixture of White voters, Black voters, and Hispanic voters.

We show that a sufficient condition for identification is that this coarsening be unrelated to the racial composition of each district, after controlling for a set of observed covariates. Formally, this amounts to a *coarsening at random* or CAR condition (Heitjan and Rubin 1991). Our particular use of CAR for ecological inference is akin to selection on observables in causal inference. Although many researchers have referred to a similar identification condition over time (Robinson 1950; Hanushek, Jackson, and Kain 1974; Blalock 1984; Glynn et al. 2008), none to our knowledge have stated the identification condition as formally as we do here.[2] All methods of ecological inference fail to consistently estimate the quantity of interest when this condition does not hold.

We further show in this reformulation that aggregation itself, inherent in ecological (aggregate) data, provides additional valuable structure when incorporating covariates: the outcome is always partially linear in the groups of interest under CAR. This makes estimation for the practitioner straightforward: interact each covariate with the predictor variable and run a linear regression of the outcome on these interactions. The estimate is then not any particular coefficient, but the weighted average across predicted values under a hypothetical dataset where all observations are purely of one predictor group. Even when the identification condition is not met, the intuition of a linear function helps diagnose the presence and direction of ecological fallacies. We show that the familiar logic of leverage points and outliers in estimating a linear regression carries through into ecological inference estimates.

In the second part of the paper, we re-characterize existing methods for ecological inference in this framework. Although prominent existing methods do not explain their models in this way, their models can be represented by a certain linear regression as well. Perhaps the most prominent methodology

[2]King (1997), Chambers and Steel (2001), and Imai, Lu, and Strauss (2008) come close, as we distinguish later in the paper.

known in the political science literature is that of King (1997), which specifies that the error term in the regression follows a truncated distribution so that predictions never fall outside the feasible parameter space. On the other hand, count models that enabled researchers to handle more than two predictors, commonly known as $R \times C$ EI methods, impose a markedly different regression model with a different set of conditions.

In this part of the paper, we also discuss how the linear regression framework leads to novel ecological inference methods that more flexibly include confounders. Our setup of a CEF with some known features (partial linearity) and other unknown features (the interactive form of confounders in the CEF) naturally motivates a semiparametric modeling approach. We explain new methodology we have developed (McCartan and Kuriwaki 2025a; 2025b) that leverages recent advances in semiparametric machine learning to construct an estimator that is doubly robust to functional form misspecification of the confounders, and discuss its connections to existing approaches to EI.

In the third part of the paper, we explore how these methods perform in practice. We use real datasets from two common use cases of ecological inference in political science. The datasets we select also fortuitously observe the ground truth quantity of interest $\mathbf{B}$, so that we can compare methods on their actual accuracy. In the example of estimating vote choice by racial group, we show that ecological inference estimates tend to *overestimate* the degree of racially polarized voting. This is partly because the types of Black voters who reside in neighborhoods with higher proportions of Black residents disproportionately affect the estimates for Black voters overall. In the example of estimating ticket splitting between two offices, such as the vote for President and the vote for a U.S. House representative, we show that ecological inference tends to *underestimate* the degree of ticket splitting. This is partly because that partisans are less likely to split their ticket in precisely the areas that have high leverage in linear regression. These two cases feature different patterns, but the regression framework we advance helps practioners make sense of why these biases arise.

Together, our reassessment of ecological inference generates theoretical and empirical results that are more succinct, general, and intuitive than the current state of the literature. This framework is also fruitful in leading to new approaches that flexibly control for covariates and squarely tackle the ecological fallacy.

2 An Example of the Fallacy

Interpreting associations from aggregated data as individual associations is often admonished as an ecological fallacy, in which aggregate associations can be completely misleading. However, not all inferences from aggregated data are ecological fallacies. An illustrative example will help foreground our formal treatment of the problem.

A classic ecological fallacy occurs in analyzing the U.S. Presidential Election of 1968. George Wallace, the former Governor of Alabama-turned third-party candidate who had the most segregationist policy platform in this election, won over a third of the vote in the former Confederate Southern states. Understanding the source of voter support for Wallace in this election is of interest to research on American Politics, especially in the context of 1968, when the New Deal coalition is thought to have fallen apart.

In particular, in the absence of comprehensive survey data, what do aggregate election results tell us about how many Black voters voted for Wallace in the South? From substantive context, we would

expect the answer is close to zero. Newly enfranchised Black voters would have little reason to vote for Wallace, a candidate explicitly running against the Civil Rights movement.[3] Estimating this quantity would be trivial if a tally of votes cast are reported by racial group. But elections report results by geography, each of which contain a mix of Black and non-Black voters. Absent a representative survey, the second best alternative is to compare the Wallace voteshare in counties predominantly with Black voters to the Wallace voteshare in counties with fewer Blacks. If the Wallace voteshare is 1 point lower in a county that is 61% Black vs. an otherwise equivalent county that is 60% Black, we might infer that Black voters do not vote for Wallace at all.

This reasoning suggests a simple (and as we will show, surprisingly sound) intuition for ecological inference: plot the Wallace vote against the racial composition, and extend the slope of the line to a hypothetical, 100% Black county. Figure 1 shows exactly such a scatterplot in two states: South Carolina and North Carolina.[4] In South Carolina, counties with a higher Black population were less likely to vote for Wallace. This loosely implies that very few Black voters overall voted for Wallace. However, in neighboring North Carolina, the relationship flips. Counties with a higher Black population, covering the coastal Piedmont region, is where Wallace wins the most votes. A naive observer might conclude that Black voters were therefore more likely to vote for Wallace than non-White voters in North Carolina.

Observers of these elections knew that these results could not be meant to indicate Black voters preferred Wallace. Schoenberger and Segal (1971) understood that "[i]t would be a fallacy—ecological, logical, sociological and political—to infer from these data that blacks in the South provided a major source of Wallace support." Instead, they reasoned that White voters reacted to the Civil Rights movement of 1964 and the Voting Rights Act of 1965, turned against the incumbent Democratic administration and cast their votes for the Southern candidate, Wallace (Wright 1977; Phillips 2014;

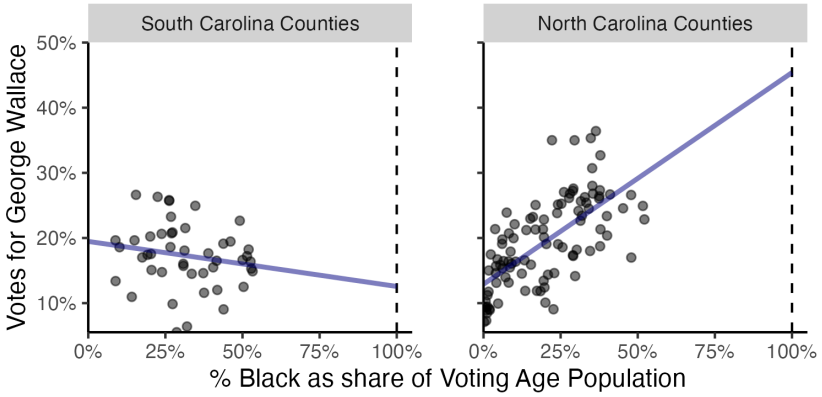


Figure 1: **Examples of the Ecological Fallacy**. X-axis shows the Black population as a share of the overall voting-age population (VAP). The y-axis shows the total number of votes cast for George Wallace as a share of VAP. The estimated VAP turnout in the two states was 54%, which was in turn split across Nixon 21%, Wallace 17%, and Humphrey 16%.

[3]Indeed, a academic survey of this election showed that only 0.3% of Black self-reported voters in the former confederate states reported for Wallace. See Wright (1977) and ICPSR 07508.

[4]We use the 1870 decennial Census and historical county-level election results, showing the vote for George Wallace against the Black population, both as a share of the voting age population.

Schoenberger and Segal 1971).[5] This dynamic leads to an ecological fallacy in North Carolina: the aggregate correlation indicates Black voters being one of the leading voting blocs for Wallace, while in reality the opposite is true.

The Wallace example is instructive in several ways. First, the fact that the relationship between Wallace vote and race flips in two neighboring states suggests that the presence of ecological fallacy depends on particular contexts not easily observed. The use of linear regression to make extrapolation, too, also turns out to be theoretically justified under particular conditions. As we will show, a certain linear regression that accounts for potential confounders is the correct approach to infer individual means from aggregate data.

3 Confounding and Linearity in Ecological Inference

This section formalizes the ecological inference problem. We clarify how certain variables confound the identification of the estimand, and show how a linear model structure arises naturally from the necessary identification assumption. We then discuss how intuition from linear models can be used to understand the inference problem. We conclude by putting our identification result in context of the existing literature.

3.1 The core estimation challenge and the accounting identity

The ecological inference problem is estimating the expectation of a variable Y given a discrete variable X when individual observations are averaged according to a grouping variable G. Here we will define this quantity of interest from the bottom up, starting from individual data.

Suppose individuals belong to one of K categories, so X_{ik} is a binary variable indicating whether individual i belongs to category k. Across n observations, define the (finite sample) mean of Y among individuals for whom $X_k = 1$ as capital B_k. To indicate these subpopulations, we use the notation I to indicate a set of individuals, so, for example, I_k indicates all individuals who are of category k. Similarly, we use N to count these individuals, so N_k is the number of individuals in the set I_k. Using this notation,

$$B_k := \frac{1}{N_k} \sum_{i \in I_k} Y_i. \tag{1}$$

We call the superpopulation mean of these quantities, $\beta := \mathbb{E}[\mathbf{B}]$ (where the boldface indicates a vector quantity) the *global estimand* (or global mean). This is the usual quantity of interest in ecological inference. Next, we let B_{gk} indicate the local (again, finite sample) conditional averages for each geography (or group) g. That is,

[5] One wrinkle in this particular example is that turnout may have differed by race. This was the first general election since the Voting Rights Act and parts of the Black vote may have been suppressed or lagged behind. If there were no Black voters who actually turned out in moderately Black North Carolina counties, the premise of our fallacy disappears. In our scatterplot, we implicitly assume that turnout rates are equal across races. That said, a 1969 U.S. Census report estimates that turnout among Black registrants in this election lagged White voters only by 10 points, even in the South (see http://bit.ly/4lsvOKu). Additionally, an analysis that uses abstention as an additional vote category finds a similarly counterintuitive result.

$$B_{gk} := \frac{1}{N_{gk}} \sum_{i \in I_{gk}} Y_i, \tag{2}$$

where I_{gk} again refers to individuals who are in both category k and geography g. To disambiguate between $\mathbf{X}$ and G, we refer to levels of $\mathbf{X}$ as *categories* (indexed by k) and G as geographies (indexed by g), although in practice G can represent non-geographic categories as well. These *local means* can also be of interest to researchers. Many EI methods produce estimates of local means, though these are more sensitive to modeling assumptions.

Because the category $\mathbf{X}$ is discrete, the sum of outcomes Y in the entire geography can be partitioned into K terms:

$$\sum_{i \in I_g} Y_i = \sum_{i \in I_{g1}} Y_i + \cdots + \sum_{i \in I_{gK}} Y_i.$$

The following modifications show the key identity: overall mean of Y in geography, $\overline{Y}_g$, is exactly a function of the sample means, $\overline{\mathbf{X}}$, and the local conditional means, $\mathbf{B}_g$.

$$\begin{aligned}
\overline{Y}_g &= \frac{1}{N_g} \sum_{i \in I_g} Y_i \\
&= \frac{1}{N_g} \frac{N_{g1}}{N_{g1}} \sum_{i \in I_{g1}} Y_i + \cdots + \frac{1}{N_g} \frac{N_{gK}}{N_{gK}} \sum_{i \in I_{gK}} Y_i \\
&= \frac{N_{g1}}{N_g} B_{g1} + \cdots + \frac{N_{gK}}{N_g} B_{gK} \\
&= \overline{X}_{g1} B_{g1} + \cdots + \overline{X}_{gK} B_{gK}.
\end{aligned} \tag{3}$$

The third line follows from Eq. 2 by substitution, and the final line holds because the proportion of the geography g that is also of category k is exactly the sample mean of X_k. Finally, we can write Eq. 3 compactly with vector notation as

$$\overline{Y}_g = \mathbf{B}_g^\top \overline{\mathbf{X}}_g \tag{4}$$

This expression is known as the *accounting identity*: aggregation enforces on the data a requirement that the a linear combination of the $\overline{X}_{gk}$ with the B_{gk} add up exactly to $\overline{Y}_g$.

The local means are connected to the global mean in a similar way as the individual data are connected to the local parameters. Specifically, we can write

$$B_k := \frac{1}{N_k} \sum_{i \in I_k} Y_i = \frac{1}{N_k} \sum_g \sum_{i \in I_{gk}} Y_i = \frac{1}{N_k} \sum_g \underbrace{N_{gk} \frac{1}{N_{gk}}}_{=1} \sum_{i \in I_{gk}} Y_i = \frac{\sum_g N_{gk} B_{gk}}{\sum_g N_{gk}};$$

the last equality follows from Eq. 2 and because $N_k = \sum_g N_{gk}$. Thus B_k can be expressed as a weighted average of the local parameters B_{gk}, where the weights are N_{gk} (or equivalently $\overline{X}_{gk} N_g$).

A natural approach to estimating β would therefore be to estimate the local estimates $\mathbf{B}_g$, and then take the weighted average of these estimates. However, as Eq. 4 shows, each observation contributes

K unknown parameters (the entries of $\mathbf{B}_g$), so there are K times as many parameters as observations. This makes estimating $\mathbf{B}_g$, and thus β, impossible without further assumptions.

3.2 Identifying the global estimand

Perhaps the strongest assumption one could make is to set local parameters $\mathbf{B}_g$ be equal across geographies g, so there are only K total unknown parameters: the entries of β. Eq. 4 would then simplify to $\overline{Y}_g = \beta^\top \overline{\mathbf{X}}_g$, which is a linear regression model with no error term. The entries of β could then be read off a linear regression fit on data $\overline{Y}_g$ against $\overline{\mathbf{X}}_g$. Of course, no ecological data actually forms a line without deviation when plotted, so this assumption is never practically tenable.[6]

We need not assume that the $\mathbf{B}_g$ are all equal for the purpose of estimating the global estimand β. Rather, it is sufficient to assume that any variation in $\mathbf{B}_g$ around their mean is unrelated to the $\overline{\mathbf{X}}_g$ and N_g. Formally, the assumption of *coarsening completely at random* (CCAR) is that $\mathbb{E}\left[\mathbf{B}_g \mid \overline{\mathbf{X}}_g, N_g\right] = \beta$ for all g. The following proposition shows that this assumption is sufficient to identify β. Proofs for all propositions in this paper are in Appendix B.

Proposition 1 (Identification under CCAR): If CCAR holds so that $\mathbb{E}\left[\mathbf{B}_g \mid \overline{\mathbf{X}}_g, N_g\right] = \beta$ for all g, then β is identified as

$$\beta = \mathbb{E}\left[\overline{\mathbf{X}}\overline{\mathbf{X}}^\top\right]^{-1}\mathbb{E}\left[\overline{\mathbf{X}}\overline{Y}\right],$$

the (population) regression coefficients of $\overline{Y}$ on $\overline{\mathbf{X}}$.

In other words, a regression of $\overline{Y}_g$ on $\overline{\mathbf{X}}_g$ can consistently estimate β, as long as one believes that CCAR holds. This estimator is known as *Goodman regression* or *ecological regression* (Goodman 1953).

Is coarsening completely at random plausible in practice? Consider again the 1968 election example. In this case, CCAR means that the preference of White voters for Wallace is unrelated to the proportion of Black voters in a county or the population of the county. For instance, this means that all of the following groups support Wallace to the same extent in North Carolina in expectation: White voters in rural lowland counties with a higher proportion of Black voters, White voters in urban counties, and White voters in mountainous counties with few racial minorities. This is clearly not the case, and helps explain why the simple linear regression shown in Figure 1 for North Carolina has the wrong slope.

A more general understanding of the CCAR assumption can be gained by analogy to causal inference. With two racial categories, the accounting identity is

$$\overline{Y}_g = B_{g1}\overline{X}_{g1} + B_{g2}\overline{X}_{g2},$$

where B_{g1} and B_{g2} are the unobserved local means which are coarsened into $\overline{Y}_g$ by aggregation. In causal inference with a binary treatment T, we have

$$Y_i = Y_i(0)(1 - T_i) + Y_i(1)T_i,$$

where the potential outcomes $Y(0)_i$ and $Y(1)_i$ are unobserved and are "coarsened" into Y_i by the treatment assignment T_i. The only difference is that in the causal setting, T_i is binary, so exactly one

[6] For example, there points are dispersed around the regression line in Figure 1.

potential outcome is observed for each individual, while for EI, $\overline{\mathbf{X}}_g$ is continuous, and we observe a mixture of the two potential outcomes $\mathbf{B}_g$.[7]

The fundamental problem of causal inference is that there are twice as many unobserved potential outcomes as there are individuals, so some identification assumption must be imposed. The simplest assumption is that the potential outcomes are independent of the treatment assignment. In randomized experiments where this assumption applies, the average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$ can be consistently estimated by the difference in sample means between the treated and control groups.

The assumption of a fully randomized experiment is exactly analogous to the CCAR assumption: the missing data ($\mathbf{B}_g$ in EI, the potential outcomes in causal inference) are independent of the coarsening variable ($\overline{\mathbf{X}}_g$ in EI, T_i in causal inference).[8] Thus Goodman's regression can reasonably be thought of as the analog to the difference-in-means estimator in causal inference. It can only be expected to produce a reasonable estimate if $\overline{\mathbf{X}}_g$ is exogenous or as-if randomly assigned.

If CCAR is implausible, what is to be done? In causal inference, an assumption of complete randomization can be weakened to hold conditional on covariates, which is known as a *selection-on-observables* assumption. The same idea applies to ecological inference: covariates can solve the ecological fallacy under certain conditions.

Suppose there exists a variable $\mathbf{Z}_g$ (in general, a vector) that are observed at the geography level. For example, Z can be the income in a geography, or a binary variable indicating whether the geography is in the South part of the state. Then, if

$$\mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g, N_g\big] = \mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g\big], \tag{5}$$

we say that *coarsening at random* (CAR) holds.[9] The intuition of this equation is that among geographies with a particular set of features $\mathbf{Z}_g = \mathbf{z}$, then knowing $\overline{\mathbf{X}}$ and N does not change the expected value of $\mathbf{B}_g$. CCAR is a special case of CAR where $\mathbf{Z}_g$ contains no covariates at all. Versions of this assumption have been discussed, often informally, in the literature, a history we review in Section 3.5 below. This assumption is sufficient to identify the quantity of interest β, as the following proposition formalizes.

Proposition 2 (Identification under CAR): For all k, under coarsening at random, β is identified as

$$\beta_k = \frac{\mathbb{E}\big[N_k \mathbb{E}\big[\overline{Y} \mid \mathbf{Z}, \overline{X}_k = 1\big]\big]}{\mathbb{E}[N_k]}.$$

The natural plug-in estimator implied by Proposition 2 is as follows: - First, fit a regression model of $\overline{Y}$ on $\overline{\mathbf{X}}$ and $\mathbf{Z}$. - To estimate β_k, first generate fitted values but on a hypothetical dataset that sets $\overline{X}_{gk} = 1$ for all observations and sets $\overline{X}_{gk'} = 0$ for all other $k' \neq k$ - Then average over these fitted

[7] Ecological inference is also different from causal inference with a continuous-valued treatment. In causal inference with continuous treatment, there is a different potential outcome for each of the infinitely-many treatment values.

[8] The additional portion of the CCAR assumption involving N_g is necessary to avoid having to weight by N_g in the regression, as McCartan and Kuriwaki (2025a) discuss in more detail.

[9] Note that compared to Imai, Lu, and Strauss (2008), who discuss similar assumptions, our acronyms are reversed: there, CAR is the stronger assumption (our CCAR), and CCAR (where the first C stands for "conditional") is the weaker assumption. We have opted for CCAR/CAR here to match the existing MCAR/MAR terminology in the missing data literature.

values, weighted by N_{gk}, the number of individuals in category k in each geography g We discuss the practical and statistical aspects of this approach in Section 4.

At first glance, Proposition 2 may appear very different from Proposition 1. However, when CCAR holds and $\mathbf{Z}$ is empty, $\mathbb{E}\big[\overline{Y} \mid \mathbf{Z}, \overline{X}_k = 1\big] = \mathbb{E}\big[\overline{Y} \mid \overline{X}_k = 1\big]$ is constant, and so the identification expression simplifies to $\beta_k = \mathbb{E}\big[\overline{Y} \mid \overline{X}_k = 1\big]$. The value of the Goodman regression when $\overline{X}_k = 1$ and all other $\overline{X}_{k'} = 0$ is exactly the coefficient on $\overline{X}_k$, which is the identification result in Proposition 1.

3.3 The role of linearity

Proposition 2 shows that it is possible in principle to estimate $\boldsymbol{\beta}$, as long as one collects enough covariates $\mathbf{Z}$ so that the CAR assumption is plausible. But CAR provides benefits beyond identification when combined with the accounting identity Eq. 4: it implies that the true regression function of $\overline{Y}$ on $\mathbf{Z}$ and $\overline{\mathbf{X}}$ is *partially linear* in $\overline{\mathbf{X}}$. This partial linearity proves to be helpful both in estimating $\boldsymbol{\beta}$ in practice, and in understanding the statistical challenges of ecological inference.

Why does partial linearity arise in ecological inference particularly? Note that the CAR assumption implies the following structure for B_g:

$$\mathbf{B}_g = f(\mathbf{Z}_g) + \boldsymbol{\varepsilon}_g, \quad \text{with} \quad \mathbb{E}\big[\boldsymbol{\varepsilon}_g \mid \overline{\mathbf{X}}_g\big] = 0 \tag{6}$$

where $f(\mathbf{Z}_g) = \mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g\big] = \mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g\big]$ (by CAR) is a vector-valued function of $\mathbf{Z}_g$. The crucial feature of this statement is that $\boldsymbol{\varepsilon}_g$ is conditionally mean-zero, because the residual from the conditional expectation is orthogonal to the conditioning variables. Eq. 6 merely re-expresses the CAR assumption in a form that will prove more convenient; it makes no new assumptions.

Now, substituting Eq. 6 into Eq. 4, we have that

$$\overline{Y}_g = \big(f(\mathbf{Z}_g) + \boldsymbol{\varepsilon}_g\big)^\top \overline{\mathbf{X}}_g = f(\mathbf{Z}_g)^\top \overline{\mathbf{X}}_g + \boldsymbol{\varepsilon}_g^\top \overline{\mathbf{X}}_g.$$

Taking conditional expectations of both sides, we find

$$\begin{aligned} \mathbb{E}\big[\overline{Y}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g\big] &= f(\mathbf{Z}_g)^\top \overline{\mathbf{X}}_g \\ &= f_1(\mathbf{Z}_g)\overline{X}_{g1} + \cdots + f_K(\mathbf{Z}_g)\overline{X}_{gK}, \end{aligned} \tag{7}$$

because the residual term $\boldsymbol{\varepsilon}_g$ is mean-zero from above. The left-hand side of Eq. 7 is the conditional expectation function (CEF) of $\overline{Y}_g$ on $\mathbf{Z}_g$ and $\overline{\mathbf{X}}_g$, which appears in Proposition 2. What the right-hand side of Eq. 7 shows is that the CEF has a partially linear structure: it is linear in $\overline{\mathbf{X}}_g$ with coefficients that depend (possibly nonlinearly) on the covariates $\mathbf{Z}_g$. This type of model is also known as a *varying coefficient* model (Hastie and Tibshirani 1993; Fan and Zhang 1999).

We can take Eq. 7 one step farther in the case where we are willing to model $\mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g\big] = f(\mathbf{Z}_g)$ as a linear function of $\mathbf{Z}_g$. Let p be the number of covariates in $\mathbf{Z}_g$, then we can write such a model as

$$f_k(\mathbf{Z}_g) = \gamma_{k0} + \gamma_{k1} Z_{g1} + \cdots + \gamma_{kp} Z_{gp}.$$

Substituting this expression into Eq. 7, we have

$$\mathbb{E}\left[\overline{Y}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g\right] = \gamma_{10}\overline{X}_{g1} + \gamma_{11}Z_{g1}\overline{X}_{g1} + \cdots + \gamma_{1p}Z_{gp}\overline{X}_{g1} + \cdots + \gamma_{K0}\overline{X}_{gK} + \gamma_{K1}Z_{g1}\overline{X}_{gK} + \cdots + \gamma_{Kp}Z_{gp}\overline{X}_{gK}. \tag{8}$$

This CEF is fully linear and so can be estimated by ordinary least squares. That is, practioners should consider running a linear regression of the outcome on a pairwise interaction of each of the p covariates $\mathbf{Z}_g$ with the K categories $\mathbf{X}$. We consider Eq. 8 to be a generalization of Goodman's regression to the case where covariates are included.[10] The additional covariates must each be interacted with every $\overline{X}_{gk}$, rather than just included as additional linear terms, to obtain this proposition.

3.4 Implications of linearity

The connection between aggregation and linear regression under CAR, is useful not only for justifying the use of linear regression. The connection allows us to use familiar intuition about linear regression to understand tradeoffs in ecological modeling and anticipate when ecological inference will fail.

To demonstrate these connections, we introduce in Figure 2 (a) a simple simulated example. We generated 20 synthetic precincts with two racial groups that satisfy the CCAR assumption, with the global parameter $\boldsymbol{\beta}$ set to $(0.5, 0.2)$.[11] In this sample, B_1 was 0.508. We then fit an OLS regression of the implied $\overline{Y}$ on $\overline{\mathbf{X}}$. The line evaluated at $\overline{X}_1$ in this particular example was 0.509, quite close β_1 and even closer to the finite-sample B_1. We also ran King's (1997) model on the same 20 data points, as a preliminary illustration of our point that King's estimator can be represented by a particular Goodman regression. Its estimate for B_1 was 0.511, also quite similar. This example neatly illustrates our case (Proposition 2) that, under CCAR, estimates of $\boldsymbol{\beta}$ can be obtained by plugging in $\overline{X}_1 = 1$ and $\overline{X}_1 = 0$ into a simple regression.

The near-perfect accuracy of the ecological inference in Figure 2 (a), is, however, somewhat fragile. There are at least three major ways in which we can characterize the direction of potential errors.

Fragility in extrapolation

In many social science examples of $\overline{X}$ and geographical groupings g cases of pure sorting ($\overline{X}_1 = 1$ or 0) is rare, so there is usually a degree of extrapolation from the observed data to produce EI estimates. This is clearly seen in Figure 2 (a), where the observed $\overline{X}_1$ are clustered near zero. As a result, the Goodman estimate for β_1 (plugging in $\overline{X}_1 = 1$) will be more variable than the estimate for β_2 (plugging in $\overline{X}_1 = 0$), because more extrapolation is required for the former.

This intuition helps explain why EI estimates for smaller minority groups are typically poor. Extrapolation in these cases can lead to highly variable or even negative estimates. Negative estimates have occasionally been discussed as an observable refutation of the CCAR assumption (Gelman et al. 2001), but we point out that finite-sample variation combined with inevitable extrapolation can also produce negative estimates *even when CCAR holds*. A similar issue arises when EI is applied to estimate voting behavior by gender with geographic data, because gender ratios in most geographies are near parity.

[10] Notice that if the stronger CCAR assumption holds, so $\mathbf{Z}_g$ is empty, then $f(\mathbf{Z}_g) = \boldsymbol{\beta}$ and Eq. 7 exactly expresses Goodman's regression.

[11] The racial group proportions $\overline{X}_1$ were drawn from a Normal distribution centered at 0.2. Thus, we see that the points in panel (a) all have average outcomes around 0.25, because $0.5 \times 0.2 + 0.2 \times (1 - 0.2) = 0.26$. Each precinct's $\mathbf{B}_g$ were drawn from a tightly distributed bivariate truncated normal distribution as in the model of King (1997).

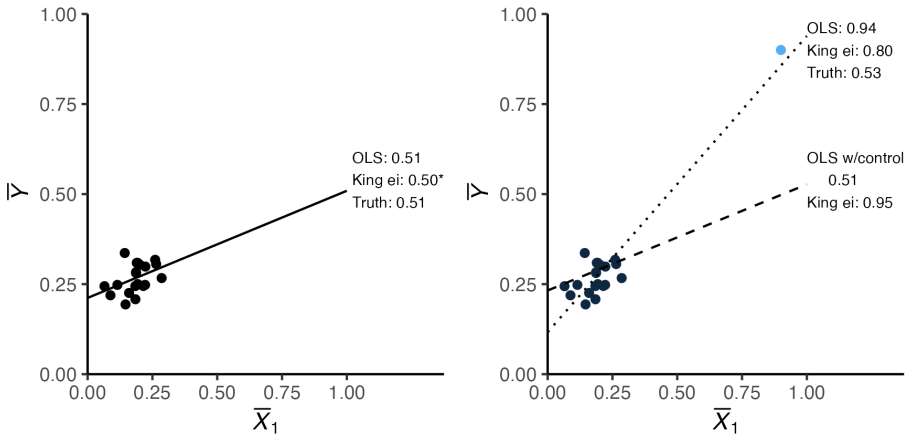


Figure 2: **Intuition for EI as linear regression.** A simulated example where the quantity of interest is $\beta_1 = 0.5$. In panel **(a)**, $n = 20$ data points are simulated from the model of King (1997). A simple OLS fit evaluated at $\overline{X}_1 = 1$ provides an estimate that agrees with both King's EI algorithm and the ground truth. In panel **(b)**, an outlier with high leverage is added to the dataset. The resulting OLS fit severely overestimates the truth, as does, to a lesser extent, King's EI. However, if the outlier differs with the other $n = 20$ points on some covariate Z, controlling for Z in the OLS reduces the bias significantly.

There is significant extrapolation when predicting vote share in a hypothetical all-female or all-male geography.

The issue of extrapolation also helps explain why EI with fine-grained geographies (e.g. precincts) tends to perform better in practice than coarser geographies (e.g. counties). Smaller geographies often exhibit more racial homogeneity: it is often possible to find several precincts where even a minority group is a local majority. Thus, even though nothing in Proposition 2 suggests more plausible identification on more granular data, it happens to be that more granular data reduces the extent to which the model relies on extrapolation.

Influence and high-leverage points

Intuition about high-leverage or influential points in linear regression also carries over to ecological inference. An influential point is an observation that has a disproportionate effect on the slope of a regression (Blackwell 2025; Chatterjee and Hadi 1986). Specifically, influence is computed by the leverage of a point multiplied by its outlier value from the leave one out regression, so high-influence points tend to be outliers. In Figure 2 (b) we illustrate the influence of such a point by adding a single, high-influence observation to the sample. This precinct has $\overline{X}_1 = 0.9$. More importantly, it has a markedly different outcome than the other distribution: $B_{g1} = 1$, $B_{g2} = 0$, resulting in $\overline{Y}_g = 0.9$. The Cook's distance of this observation (a standard measure of influence) is 34.6, compared to less than 0.2 for all other observations. As a result, the slope of the OLS line from the 21 points changes drastically to 0.94, and the extrapolated estimate at $\overline{Y} = 1$ is now off by over 40 percentage points, even though the global parameter B_1 only increased to 0.53. King (1997)'s estimate of the same data is less drastically off, but not much better. Both methods fail because of the violation of the constancy assumption. While the assumption itself is untestable from observed data, the lesson here is that the connection to regression serves as a reliable diagnostic to anticipate or identify possible violations.

Fortunately, the use of control variables can reduce the undue impact of influence points. Consider again the example in Figure 2 (b), where the linear regression is disproportionately influenced by an influence point that has a different conditional outcome. However, if there is a third variable Z that distinguishes this influence point from the remaining 20 variables, the regression may be corrected.[12] Applying the extended Goodman regression to the data with Z included generates a linear fit that is 0.51 when evaluated at $\overline{X}_1 = 1$ and $Z = 0$, as shown by the dashed line in the figure. Here, then, Z helps satisfy the CAR assumption while decreasing the residual variance of the regression, and the overall estimate is much improved.

Positivity violations

Linear regression intuition is also helpful in the potential risks of adding certain control variables. This is essentially a tradeoff between the CAR identification assumption, and the need for *positivity*: sufficient variation in $\overline{\mathbf{X}}$, even after accounting for covariates. To understand this tradeoff, recall that the variance of linear regression estimates is affected both by the amount of residual variance and the amount of variance in the predictors. Smaller residual variance leads to more precise estimates, but smaller variance in the predictors leads to less precise estimates. In the extreme case where there is no variance in the predictors, it is impossible to estimate the regression slope at all.

The Frisch-Waugh-Lovell theorem relates this aspect of linear regression to the inclusion of covariates. It states that, in the expanded Goodman regression in Eq. 8, the coefficients γ_0 on $\overline{\mathbf{X}}$ alone can be estimated in two steps: first, by partialling out the effect of $\mathbf{Z}$ on $\overline{Y}$ and $\overline{\mathbf{X}}$, and then regressing the residuals of these two variables after partialling out $\mathbf{Z}$ on each other. If $\mathbf{Z}$ explains much variation in $\overline{Y}$, and little in $\overline{\mathbf{X}}$, then the inclusion of $\mathbf{Z}$ will lead to a more precise estimate of β. Conversely, if $\mathbf{Z}$ explains much variation in $\overline{\mathbf{X}}$, but little in $\overline{Y}$, then the inclusion of $\mathbf{Z}$ will lead to a less precise estimate of β. In the extreme case where $\mathbf{Z}$ is collinear with $\overline{\mathbf{X}}$ and thus explains all of its variance, then the regression model is unidentified.[13] The need for variation in $\overline{\mathbf{X}}$ left over after controlling for $\mathbf{Z}$ is closely related to the *overlap* assumption in causal inference.

To illustrate this tradeoff, return to the 1968 election and imagine controlling for the proportion of the population in each county that was enslaved in 1860. This variable is an extremely strong predictor of the proportion of Black voters in a county and also of the vote share for Gov. Wallace. It is more plausible to believe that, among counties with a similar history of slavery, White voters' support for Wallace is unrelated to the proportion of Black voters. However, the near-collinearity of the covariate with $\overline{X}_{g1}$ means that the regression will be highly unstable. This creates the tradeoff: the effect of including $\mathbf{Z}$ on the variance of the estimates must be weighed against the increased plausibility of the CAR assumption when more covariates are included.

Fundamentally, ecological inference requires essentially exogenous variation in $\overline{X}_{g1}$, from which estimates of Y for each category can be extrapolated. When the covariates needed to satisfy CAR also mop up all of the variation in $\overline{X}_{g1}$, then there is simply not enough information left in the data to perform ecological inference at all.

[12] In this example, we suppose that $Z = 1$ for the influence point and it is a small random Normal error centered at 0 and with a standard deviation of 0.001 for the other 20 observations.

[13] Notice that CAR trivially holds if we let $\mathbf{Z} = \overline{\mathbf{X}}$. But the resulting collinearity rules out this trick as an approach to estimating β. See McCartan and Kuriwaki (2025a) for a more detailed discussion of this positivity assumption.

3.5 On the history of controlling for confounders

We suspect many methodologists developing ecological inference or avoiding ecological inference methods have been aware of the sort of conditional identification result we show here. Hanushek, Jackson, and Kain (1974) was an early outlier that proposed a linear regression with covariates, but included no interactions and provided no identification assumption or result. Literature since Hanushek et al. has formalized the intuition even less. This literature of the last 50 years has not established the identification result of the plug-in estimator. In general, the political methodology literature writes less about this aspect of ecological inference. We make several observations on this front.

King (1997, 170–171) presents a conditional independence assumption briefly in its chapter on linear contextual effects and avoiding aggregation bias, but does not formalize an identification result. Imai, Lu, and Strauss (2008) identifies the need for a coarsening at random assumption, but focuses on likelihood identification within a particular model. Neither of these sets of authors considers the role of N_g in the identification assumption. In an unpublished manuscript, Ansolabehere and Rivers (1995) derives a closed form expression of the in-sample estimation error in the 2×2 regression case, which leads to a weaker result but which less directly implicates nonparametric identification in the statistical population. This lack of focus in the literature on identification is notable since, as Lewis (2001, 175) observes, "many [applied] scholars have considered aggregation bias (the violation of [the constancy] assumption) to be *the* problem in making ecological inferences" (emphasis ours).

Controlling for covariates may have sat uneasily with the urge to not "control out" important factors that explain $\overline{\mathbf{X}}$. This urge may have been reinforced by the U.S. Supreme Court Case *Thornburg v. Gingles* (1986). The *Gingles* ruling made more specific the conditions under which racial minorities were essentially entitled to descriptive representation. The opinion set out a *results*-test, rather than an *intent*- test for a racial minority meeting this threshold. Under *Gingles*, the mere result of racially polarized voting was considered a valid piece of evidence regardless of whether it was racial animus (or even race itself) that caused this difference in voting behavior. Some scholars have taken this to mean that one *should not* control for covariates in ecological inference (King 1997, 171). However, our results clarify that controlling for covariates is often necessary to estimate β correctly (that is, irrespective of whether the goal is to isolate any *causal* effect of race).

The use of covariates also goes against a forceful argument made in an earlier, well-cited book by Achen and Shively (1995). The biases of ecological regression, they argued, "are not curable by … controlling for demographic variables" (92). Their reasoning appears premised on the setup that these controls only applied to individual-level, or micro models. Even if a researcher could collect covariates that would explain the outcome at the individual-level, Achen and Shivley argued, this does not mean that inference using aggregate means would be unbiased (94). We agree with this particular caution. Even with a perfect model of the outcome at the individual level, geographic aggregation can induce a violation in CCAR. However, our identification results under CAR satisfy Achen and Schivley's skepticism by specifying that controls should remove the confounders present in aggregate data.

Unlike causal inference where the quantity of interest is the *ceteris paribus* difference in the outcome, the quantity of interest in EI is the total expectation of the outcome for all individuals that are in group k. However, this does not invalidate the use of covariates. The correct lesson to take away is that these

control variables should be accounted for in the model, and quantities of interest should be computed from plug-in fitted values rather than simply reading off coefficient values on $\overline{\mathbf{X}}$.

4 Comparison of Methods for Ecological Inference

We now re-assess the differences across the EI methods commonly used by practitioners relative to our regression framework. Critiques, reviews, and user guides to King's (1997) method already abound. However, little work, if any, compare King's original method to the commonly used *R×C* count models, provide an intuition of these differences based on statistical theory, or evaluate the existing methods' computational strategies. We provide all of these comparisons here.

4.1 King's 2×2 ecological inference model

The most common exposition of King's model is as a random coefficient model with zero error. The aggregate outcome is modeled (conditional on $\overline{\mathbf{X}}$) as

$$\overline{Y}_g = B_{g1}\overline{X}_{g1} + B_{g2}\overline{X}_{g2},$$

with the coefficients $\mathbf{B}_g$ jointly drawn from a bivariate Normal distribution truncated to the unit square:

$$\mathbf{B}_g \overset{\text{iid}}{\sim} \mathcal{N}_{[0,1]^2}(\boldsymbol{\mu}, \Sigma),$$

where $[0,1]^2$ denotes truncation to a unit square, $\boldsymbol{\mu}$ is the location parameter, and Σ is the covariance matrix. The truncation is applied so that the B_g remain between 0 and 1, since King's model applies only to binary Y.

Re-expression

The immediate difference with Goodman's regression and King's 1997 model, then, is that the former targets the global mean $\mathbf{B}$ while the latter attempts to estimate the local means $\mathbf{B}_g$. This does not seem like a substantively large difference since the global β is readily calculable as the mean of the estimated truncated Normal.[14]

Indeed, the implications for identification remain the same in Goodman and King's models. We can rewrite King's model for $\mathbf{B}_g$ in terms of the global parameter $\mathbf{B} = \mathbb{E}\big[B_g\big]$ and a residual term $\boldsymbol{\varepsilon}_g$. To do so, define $\mathbf{m}_g = \boldsymbol{\mu} - \mathbf{B}$. Unlike with an untruncated Normal, $\mathbf{m}_g$ is not always 0, because the truncation can shift the mean of the distribution away from its location parameter $\boldsymbol{\mu}$. Then we can write

[14] Because the form of King's model and Goodman's regression are the same, they will produce the same estimates for β asymptotically. The parametrization of the model that King uses does not directly produce β, but it is readily calculable as the mean of the $\mathcal{N}_{[0,1]^2}(\boldsymbol{\mu}, \Sigma)$ distribution (our $\boldsymbol{\mu}$ corresponds to King's $\mathfrak{B}$). In the causal inference analogy, the mean of $\mathcal{N}_{[0,1]^2}(\boldsymbol{\mu}, \Sigma)$ estimates the superpopulation effect β, while the mean of the posterior distribution of $\mathbf{B}_g$ estimates the finite-sample effect $\mathbf{B}$. Goodman's regression does not estimate the $\mathbf{B}_g$ and can only estimate the superpopulation parameter. However, King does not advocate using this as the estimate of β. Rather, because he works in the Bayesian framework, he recommends estimating $\mathbf{B}$ itself from a weighted mean of the estimated $\mathbf{B}_g$. The posterior distribution of these local parameters is the original $\mathcal{N}_{[0,1]^2}(\boldsymbol{\mu}, \Sigma)$ distribution restricted to the tomography line defined by the accounting identity, $\overline{Y}_g = B_{g1}\overline{X}_{g1} + B_{g2}\overline{X}_{g2}$. Beyond very small samples, the difference between $\mathbf{B}$ and β is often negligible.

$$\mathbf{B}_g = \mathbf{B} + \boldsymbol{\varepsilon}_g, \quad \boldsymbol{\varepsilon}_g \overset{\text{iid}}{\sim} \mathcal{N}_{[0,1]^2-\mathbf{B}}(\mathbf{m}_g, \Sigma),$$

where $[0,1]^2 - \mathbf{B}$ is the unit square shifted by the vector $\mathbf{B}$. Since $\boldsymbol{\varepsilon}_g$ is drawn independent of $\overline{\mathbf{X}}$, by construction, we have $\mathbb{E}\left[\boldsymbol{\varepsilon}_g \mid \overline{\mathbf{X}}_g\right] = 0$, so $\mathbb{E}\left[\mathbf{B}_g \mid \overline{\mathbf{X}}_g\right] = B$. This is exactly the CCAR assumption. Thus in contrast to some suggestions by King (1997), his model makes the same strong assumption as Goodman's regression.[15] Since the model does not involve covariate adjustment, it does nothing to ameliorate aggregation bias due to confounders (Rivers 1998; Lewis 2001).[16]

Substituting the reexpression into the outcome model, we have

$$\overline{Y}_g = \left(B_1\overline{X}_{g1} + B_2\overline{X}_{g2}\right) + \left(\varepsilon_{g1}\overline{X}_{g1} + \varepsilon_{g2}\overline{X}_{g2}\right),$$

or, using vector notation, and letting $\overline{\varepsilon}_g = \boldsymbol{\varepsilon}_g^\top \overline{X}$,

$$\overline{Y}_g = \mathbf{B}^\top \overline{\mathbf{X}}_g + \overline{\varepsilon}_g. \tag{9}$$

This is exactly the form of Goodman's regression. The difference is that Goodman's regression makes only the assumption that $\mathbb{E}\left[\overline{\varepsilon}_g \mid \overline{\mathbf{X}_g}\right] = 0$, whereas King's model assumes a specific distribution for the error term $\overline{\varepsilon}_g$. This distribution is not the same for every observation, and in fact depends on the coefficients $\mathbf{B}_g$, the same way that the error term in a generalized linear model depends on the linear predictor. This makes fitting the model computationally much more challenging, but also allows for improved estimation in finite samples white the truncated Normal assumption is appropriate, since information on the bounds is incorporated.

Innovations on computation

King's model Eq. 9 requires more advanced computational techniques than linear regression, because the residual $\overline{\varepsilon}_g$ is non-Normal and has distribution which depends on $\mathbf{B}$ and $\overline{\mathbf{X}}_g$. In fact, because $\overline{\varepsilon}_g$ is a linear combination of (correlated) truncated Normal variables, it is not even truncated Normal itself. There is no closed-form expression for its distribution and so direct maximum likelihood estimation of Eq. 9 is not feasible. Instead, King reparametrizes the model so that the $\mathbf{B}_g$ can be analytically integrated out. The result is a heteroskedastic regression model with an additional ratio of normalizing constants related to the truncation; the detailed likelihood is derived in Appendix D of King (1997). Because of the specialized form of the likelihood, King's original computational proposal remains a good strategy, even as more advanced generic Markov chain Monte Carlo (MCMC) methods have been developed.

That said, recent advances in sampling from truncated multivariate Normal distributions and estimating their moments allows for more efficient sampling of derived quantities. These methods also open up the possibility of a generalization of King's model to more than two categories. Our software

[15] See, e.g., "even when the process of aggregation causes existing methods to give answers that bear no relationship to the truth, the method proposed here still usually gives accurate answers" (King 1997, 20).

[16] King (1997) defines aggregation bias in a looser way. Chapter 9 of the book discusses covariate adjustment, but also appears to include out of bound estimates as a part of aggregation bias. It is true that the modeling specification imposed here ultimately reduces estimation error in likely many cases, as cited in Lewis (2001). However, King's main method as stated here still requires an untested assumption of conditional independence, and so we agree with Lewis' 2001 (175) interpretation that King's main method "should be thought of as a 'solution' in the sense that, assuming its assumptions hold, it allows the user to make more efficient estimates [of the global parameter] than can be made using conventional regression techniques and also allows the estimation of [local parameters]."

(McCartan and Kuriwaki 2025b) implements both of these techniques in efficient C++ code, which in the future will enable faster inference of King's model in the 2×2 case by as much as an order of magnitude.

Unlike univariate truncated Normal distributions, which have a closed-form CDF and normalizing constant, higher-dimensional truncated Normal distributions involve intractable normalizing constants. Recent expectation propagation (EP) methods can quickly and accurately approximate the normalizing constants of truncated multivariate Normal distributsions (Cunningham, Hennig, and Lacoste-Julien 2011). These innovations also provide faster approximations to the moments of the truncated distribution. These advances have, for instance, enabled large-scale estimation of multivariate probit models (Ding et al. 2024).

Progress has also been made in sampling quickly from truncated multivariate Normal distributions, using elliptical slice sampling (ESS) (Wu and Gardner 2024). ESS is a MCMC method that is particularly efficient by design, with an acceptance rate of 1. The ESS sampler for truncated multivariate Normals can be used in the 2×2 case for simulating $\mathbf{B}_g$ from the model, and in the general multi-category case for drawing $\mathbf{B}_g$ from their posterior distributions.

Random coefficient models

Inference in the untruncated version of King's model is much simpler, because of properties of the multivariate Normal distribution which do not apply under truncation. The untruncated version of King's model is also known as a random coefficient model, which have been well-studied in the statistics literature. In this case, the MLE for the local means, conditional on Σ, is given by

$$\hat{\mathbf{B}}_g = \underbrace{\hat{\mathbf{B}}}_{\text{global estimate}} + \Sigma\overline{\mathbf{X}}_g\left(\overline{\mathbf{X}}_g^\top\Sigma\overline{\mathbf{X}}_g\right)^{-1}\underbrace{\left(\overline{Y}_g - \overline{\mathbf{X}}_g^\top\hat{\mathbf{B}}\right)}_{\text{residuals}}. \tag{10}$$

This result has been known since Griffiths (1972), as cited in Anselin and Cho (2002), and a version in the 2×2 case is also discussed by King (1997). That is, the estimates are the combination of the global estimate and the residual of the linear regression reweighted by the covariances of the two coefficients. This result shows that, intuitively, King's model produces unit-level estimates by allocating the residuals from a certain weighted Goodman regression in accordance with an estimate of the covariance m atrix Σ.[17]

Lest Eq. 10 and our above discussion overstate the similarity of King's model with Goodman's regression in practice, we stress the importance of the truncation in practice. King's innovation was to sample from a truncated distribution to have estimates fit into bounds, a highly desirable property for EI estimates, especially when they are consumed by non-statisticians. The naive move to use a non-linear regression like a logit regression would have gone against the linearity of the CEF (Eq. 7) and created issues with computing the posterior distribution of the local parameters. In finite samples, the truncation also often helps improve the global estimates themselves.

[17]Since in general Σ is unknown, full inference for the untruncated model requires more than Eq. 10. A simple expectation-maximization (EM) scheme is likely to perform well, and be quite computationally efficient, especially compared to the truncated version. However, variance quantification under vanilla EM is difficult, and other methods such as full Bayesian inference may be preferable.

4.2 Count models

Inference in the multivariate version of King's model is much more computationally difficult than inference in the 2×2 model. However, researchers are interested in estimating conditional means for more than two categories, or modeling discrete choice behavior across more than two choices. Both of our empirical evaluations are of this case: four racial groups or four types of candidates. Soon after King (1997), several researchers developed methods that can account for such settings, and came to be known as *R×C* EI methods, building on Brown and Payne (1986).

King's 1997 model is expressed in terms of the (unobserved) local means $\mathbf{B}_g$, which must strictly satisfy the accounting identity (Eq. 4). This restricts the parameters to a *tomography line*, which in general is a hyperplane in the parameter space of dimension $K-1$. The truncated Normal distribution used in King's model has the advantage of being able to be projected onto this hyperplane, in the sense that the restriction of the truncated Normal distribution to the tomography line is still a (degenerate) truncated Normal distribution. Unfortunately, the normalizing constant for the truncated Normal is not available in closed form in two or more dimensions.

The *R×C* models, therefore, take a different approach, parametrizing the model in terms of the superpopulation parameters β_g rather than the finite-sample $\mathbf{B}_g$. The observed data are connected to these parameters through a count model, which we describe next. The key implication is that β_g are not required to exactly satisfy the accounting identity, due to the uncertainty introduced by the count model. This additional wiggle room greatly aids estimation. However, to make inference for the count models tractable, distributions for β_g other than the truncated multivariate Normal are used. Some of these distributions are limited in their expressive power, and these limitations can have a severe impact on the estimates.

Rosen, Jiang, King, and Tanner (2001)'s model

To describe the model, we have to first generalize Y to be multivariate. Let $M_{g1}, ..., M_{gJ}$ be the *counts* of the number of individuals in each level of Y, such as the number of votes for each party, in geography g. The parameter β_g is now a matrix, with J rows and K columns. For example, $\beta_{g,\,\text{dem},\,\text{white}}$ is the proportion of White voters in geography g who vote for the Democratic candidate. Rosen et al. (2001) propose a simple count model for $\mathbf{M}_g$:

$$\mathbf{M}_g \overset{\text{iid}}{\sim} \text{Multinom}\left(N_g, \beta_g^\top \overline{\mathbf{X}}_g\right). \tag{11}$$

This model does not involve covariates Z, and model β_g independent of $\overline{\mathbf{X}}_g$, and so, like King (1997), relies on the CCAR assumption.

How does the move to model counts affect our argument that ecological inference should be thought of as a variant of linear regression? Because of the Multinomial distribution, it is not the case that the outcome counts satisfy

$$\mathbf{M}_g = N_g \beta_g^\top \overline{\mathbf{X}}_g$$

the way that the aggregate outcome $\overline{Y}$ does in King's model. This is because β_g are superpopulation parameters, which may be different from the unobserved data $\mathbf{B}_g$, which does have to satisfy the accounting identity. Because β_g no longer is restricted to a lower-dimensional subspace, estimation

becomes much easier: in principle, any values of β_g within bounds are possible, and the accounting identity is satisfied only in expectation.

The choice of Multinomial distribution, while common for count data, also implies several substantive assumptions about voting behavior. A Multinomial distribution is appropriate when each voter's choice can be considered a coin toss with the same constant probability, independent of other voters. That cannot be the case in ecological inference, where the goal is to estimate differences in voting behavior across groups. For example, suppose in truth $\beta_{g,\text{ dem, white}} = 0$ and $\beta_{g,\text{ dem, black}} = 1$, so all White voters in precinct g prefer Republicans and all Black voters prefer Democrats. Suppose further that there are equal numbers of White and Black voters. Then Eq. 11 assumes that *all* the voters in the precinct, White and Black, vote for Democrats with probability $1 \cdot 0.5 + 0 \cdot 0.5 = 0.5$. While on average this produces the correct number of Democratic votes in the precinct, it cannot reflect the underlying voting behavior being modeled.

Beyond this inconsistency, even within a racial group, some voters will almost always vote for one party, while others will almost never vote for that party, and voters' choices are highly correlated. The Multinomial distribution rules out both voter heterogeneity and correlation. It is unclear the extent to which this modeling choice affects the location of the estimates of β, but it certainly affects the uncertainty of those estimates: alternative count models which allow for more variation due to heterogeneity and correlation would constrain β less, leading to more uncertainty in the estimates.

The choice of distribution for $\beta_g^\top$ is also consequential. Rosen et al. (2001) adopt a simple Dirichlet distribution for each *row* of β_g, with the entries within each column independent. In other words, $\beta_{g,\text{ dem, white}}$ and $\beta_{g,\text{ rep, white}}$ are together drawn as components from the same Dirichlet distribution, but $\beta_{g,\text{ dem, white}}$ and $\beta_{g,\text{ dem, black}}$ are completely independent. This is a marked departure from King's model, where $\beta_{g,\text{ white}}$ and $\beta_{g,\text{ black}}$ are assumed to be correlated according to Σ. Even within a row, the Dirichlet distribution (with J parameters) is far less flexible than a Normal distribution (with $J + J(J-1)/2$ parameters). Specifically, under Dirichlet, the correlation between any two entries in a row is determined completely by the mean of each entry and the overall dispersion of the distribution. A Normal distribution, in contrast, allows for the correlation between any two entries to be much more freely specified.

Greiner and Quinn (2009)'s model

The independence within columns (across racial groups) and the lack of separate parameters to control the correlation between even within a row (across candidates) makes the Rosen et al. (2001) model much less flexible than King's 1997 model. As the ticket splitting example in Section 5.2 below shows, this can lead to strange behavior, where the choice to consolidate several small outcome categories can have dramatic impacts on the estimates of the other parameters.

Greiner and Quinn (2009) directly addressed this drawback of the Rosen et al. (2001) model by proposing a more flexible model which allows for more correlation within β_g. They use a slightly different count model which models the local counts $N_{gk}\mathbf{B}_{gk}$ individually rather than the totals $\mathbf{M}_g$:

$$N_{gk}\mathbf{B}_{gk} \overset{\text{iid}}{\sim} \text{Multinom}\big(N_{gk}, \beta_{gk}\big),$$

where B_{gk} and β_{gk} are vectors describing the vote preferences across the J choices for voters in category k and geography g. The observed outcome counts $\mathbf{M}_g$ are simply the sum of the unobserved counts for each group:

$$\mathbf{M}_g = \sum_{k=1}^{K} N_{gk}\mathbf{B}_{gk}.$$

This setup at least reflects that voters in different racial groups vote differently, in contrast with Rosen et al. (2001), but still assumes that votes within each racial group are independent with the same probability.

To allow for more correlation within β_g, Greiner and Quinn (2009) transform each β_{gk} to a vector of $J-1$ logits, picking one of the J choices as the reference category. These logit preferences are then modeled as a multivariate Normal distribution, which allows for correlation.

Both *R×C* approaches, then, make strong assumptions about voting behavior in their count models, but quite different assumptions about the correlations between the β_g. King's (1997) model does not place restrictions on these correlations, nor does Goodman's regression, which targets the global β and does not model the local $\mathbf{B}_g$ or β_g at all. Despite these differences, and the outwardly different appearances, the CEF of the outcome is still linear for these models, just as it is for King's model and Goodman's regression. For both models, we have

$$\mathbb{E}\left[\mathbf{M}_g \mid N_g, \overline{\mathbf{X}}_g\right] = N_g \beta_g^{\top} \overline{\mathbf{X}}_g.$$

Thus both models can still be viewed as (multivariate) linear regressions, with residuals distributed as a discrete mixture of Multinomials.

Computation in count models

The move to count models makes computation much easier than King's 1997 model, since there is no intractable normalizing constant and no restriction to a lower-dimensional subspace. Both Rosen et al. (2001) and Greiner and Quinn (2009) derive Gibbs samplers for their models. Because of the often strong correlations in the posterior between certain parameters in these models, modern general-purpose MCMC methods like the No-U-Turn Sampler (NUTS) implemented in the Stan software (Carpenter et al. 2017) are usually less efficient. Unfortunately, despite the superior choice of distribution for β_g in Greiner and Quinn (2009), their R implementation is no longer publicly available in CRAN, while the Rosen et al. (2001) model is widely used as part of the `eiPack` software.

4.3 Semiparameteric modeling

So far, we have shown the similarity between Goodman's regression and King's model: how the latter makes a parametric assumption about the residuals in order to improve finite-sample performance and produce in-bounds estimates. Models that generalized to multivariate outcome added additional distributional assumptions. Neither of thse approaches—King's model and the count models—focus on the inclusion of covariates, though it is possible to incorporate covariates in both.[18]

[18]For including covariates in the King's 1997 model, see discussion under Section 3.5. The two count models discussed allow β_g (or its logit reparametrization) to be further modeled as a function of covariates, so that the CCAR

Here, we sketch an alternative approach introduced by McCartan and Kuriwaki (2025a) that instead makes fewer assumptions and leverages recent statistical advances in estimating a low-dimensional parameter of interest (β) in the presence of high-dimensional nuisance parameters. This approach follows naturally from the linearity of the CEF under the CAR assumption.

Recall that Eq. 7, reproduced below, makes no assumptions beyond CAR:

$$\mathbb{E}\left[\overline{Y}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g\right] = f(\mathbf{Z}_g)^\top \overline{\mathbf{X}}_g.$$

Thus, if $f(\mathbf{Z}_g)$ could be estimated, then following the identification result in Proposition 2, we could estimate β without any additional assumptions. We have already discussed modeling $f(\mathbf{Z}_g)$ as linear in $\mathbf{Z}_g$, but this is a strong and likely unrealistic assumption. Instead, we can expand $\mathbf{Z}_g$ into a rich basis $\Phi(\mathbf{Z}_g)$, such as splines, interactions, or trees. Well-developed statistical theory establishes that if the basis expansion Φ is rich enough, and grows richer as the sample size increases, then $f(\mathbf{Z}_g)$ can be consistently estimated *nonparametrically* (Shen and Wong 1994).

In practice, this involves picking a specific basis expansion Φ, interacting the expanded $\Phi(\mathbf{Z}_g)$ with $\overline{\mathbf{X}}_g$, as in Eq. 8, and then regressing $\overline{Y}_g$ on these terms. The large number of terms in the basis expansion Φ means that ordinary least squares will likely overfit, or be unable to be fit at all. A ridge penalty is therefore recommended. The value of the penalty can be automatically selected by leave-one-out cross-validation, for which a closed-form expression exists for ridge regression.

One risk of this approach is that, when there are many terms in $\Phi(\mathbf{Z}_g)$, the ridge penalty will be large, and the resulting regularization bias will lead to bias in the primary estimate of β. This is a well-known problem when estimating a high-dimensional nuisance function (here, f). New *double/debiased machine learning* (DML) methods (Chernozhukov, Newey, and Robins 2018) address this by learning a *second* nuisance function known as the Riesz representer. For our purposes here, the Riesz representer is a set of weights for each geography g, such that weighted averages of $\overline{Y}_g$ with these weights can estimate β. The Riesz representer is a generalization of the inverse propensity score used in causal inference, and the DML estimator that combines the two nuisance functions is a generalization of the augmented inverse propensity weighting (AIPW) estimator that is commonly used in causal inference (Robins, Rotnitzky, and Zhao 1995). Like AIPW, the DML approach is doubly robust, and so tolerates misspecification of either nuisance function, as long as the other is correctly specified.

In accompanying work, we develop this approach for the EI context, and provide public software, `seine`, which stands for sempiparametric ecological inference (McCartan and Kuriwaki 2025b).[19] In brief, we produce the DML estimate of β_k as follows.

1. Fit the semiparametric linear regression model of $\overline{Y}$ on $\Phi(\mathbf{Z}_g)$ interacted with $\overline{\mathbf{X}}_g$.
2. Estimate the Riesz representer α_k for each category k, using the same basis expansion $\Phi(\mathbf{Z}_g)$.
3. Plug in $\overline{X}_{gk} = 1$ (and $\overline{X}_{gk'} = 0$ for all $k' \neq k$) into the fitted regression model, which produces predictions $\hat{f}_k(\mathbf{Z}_g)$.
4. For each geography g, calculate the DML score:

assumption may be used. Rosen et al. (2001) discuss this explicitly, and the widely-used `eiPack` software (Lau, Moore, and Kellermann 2007) that implements their model allows covariates to be provided.

[19] Available at `www.corymccartan.com/seine`.

$$s_g = \hat{f}_k(\mathbf{Z}_g)\frac{N_{gk}}{N_k} + \alpha_k\left(\overline{Y} - \hat{\overline{Y}}\right),$$

where $\hat{\overline{Y}} = \sum_k \hat{f}_k(\mathbf{Z}_g)\overline{X}_{gk}$ are the fitted values from the regression model.

5. Estimate β_k with the mean of s_g. The standard error of this estimate can be estimated as $\text{sd}(s_g)/\sqrt{|G|}$, where $|G|$ is the number of geographies.

Compared to existing EI methods, the approach summarized here has several advantages. Computationally, it is much faster to fit, since both the regression and the Riesz representer have closed-form solutions. By construction, the regression respects the partially linear form of Eq. 7 while allowing for nonlinearities through the analyst's use of a basis expansion $\Phi(\mathbf{Z})$. Finally, it efficiently estimates the global parameter of interest, β, without making parametric assumptions about the form of the error term. However, as we have discussed, often the assumptions on the error term can prove helpful in finite-sample estimation. This is especially true when the β are near the bounds of Y, as is true for our ticket splitting example below in Section 5.2. We next turn to two empirical applications of these ecological inference methods, which illustrate the perennial challenges of confounding, and some of the tradeoffs between methods discussed here.

5 Empirical performance on political science applications

The degree of aggregation bias is clearly application-specific. The patterns of bias that arise from racial segregation in residential choices likely differ from bias that arise in the correlates of vote choice transitions. While the biases are not observed in typical applications, there exist several instances where the true conditional means are observed. We discuss two of the most common applicaitons in political science. We organize the discussions of both applications by (i) the patterns of the observable data, (ii) the performance of methods in estimating the unobserved quantities of interest, and (iii) patterns of confounding.

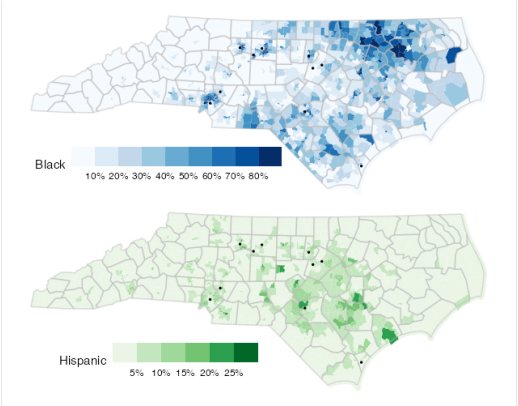
5.1 Vote choice by race

The most common application of ecological inference has been in the study of vote choice by race/ethnicity (Greiner and Quinn 2010; Freedman et al. 1991; Kuriwaki et al. 2024a). The outcome Y is an individual's vote choice for a candidate or party, and X is the race/ethnic membership of the voter. When a White majority votes predominantly for Republicans while a sizable Black and Hispanic population votes predominantly for Democrats, for example, U.S. jurisprudence of the Voting Rights Act has recognized a need for states to redraw their districts in essentially a way that Democrats can win office. Determining if these conditions are met, however, is an ecological inference challenge, because election data are reported only by geography, containing a mix of White and non-White voters.

Here we use North Carolina voter's party registration on the voter file as a proxy for their partisan voting behavior, and evaluate how well our key ecological inference methods can recover vote choice by race. Voter files in North Carolina and Florida are public, individual-level administrative datasets that include a registrant's self-identified racial group, party registration, and the precinct which they are assigned to. Although party registration is not vote choice, surveys indicate the two are heavily correlated (Kuriwaki et al. 2024a), so EI biases in voter registration are informative of biases in evaluating EI applications of vote choice as well. Of the two states, we use the North Carolina voter file

	Party Registration (%)				
Race	Democrat	Republican	Unaffiliated	Lbt.	N
White	19.5	42.2	37.7	0.6	3,619,793
Black	73.8	2.8	23.1	0.3	1,030,788
Hispanic	33.2	18.0	47.9	0.8	266,577
Other	32.3	18.7	48.4	0.6	228,503

(a) Ground Truth Joint Distribution



(b) Geography of Racial Minorities

Figure 3: **Voter's Party and Race Registration in North Carolina.** We use the North Carolina Voter file as a proxy validation for measuring racially polarized voting. Panel (a) shows the joint distribution of party and race as recorded in the public state public voter file. Cells show row percentages within each race, and are the quantities of interest. Panel (b) shows the geographic distribution of the two racial minorities. Black registrants live in the Northeast part of the state (Piedmont) and around large cities. Hispanic registrants constitute only 5% of the entire dataset and are concentrated in rural parts of the state.

because it is more up-to-date. All datasets were downloaded from the North Carolina State Board of Elections (NCSBE) in July 2025. We then standardized precinct labels so that it would match exactly to a precinct in the NCSBE's precinct 2025 shapefiles.

The North Carolina dataset covers 2,465 precincts, with each precinct averaging about 1,600 voters.[20]. The state is evenly divided between registered Democrats (31%), Republicans (31%), and non-affiliated voters, but racial groups sort into distinct party patterns. Only about 19% of White voters register as Democrats, while 74% of Black voters do the same (Figure 3a). We test four methods: (1) our new estimator `seine` with covariates entering through a BART expansion basis, (2) an OLS regression with no covariates, (3) Rosen et al.'s count model, as implemented in `eiPack`, with no covariates modeling the 4-by-4 matrix at once, and (4) King's 2×2 method applied to each racial group and each outcome separately. The covariates we included in `seine` were generated by obtaining block-group level ACS 2023 estimates of education, age, and income, and aggregating them up to the corresponding precinct.

The difficulty of ecological inference is shaped by the distribution of the racial group. Racial minorities are concentrated in certain parts of the state. Black voters are about 20% of the dataset, and concentrated around a few of the major cities—Greensboro and Charlotte—but especially concentrated in the Black Belt of the northeast coastal plains (Figure 3b). A few precincts are almost completely composed of Black voters: 3 out of the 2,465 precincts are over 95% Black and 17 are over 90% Black. In contrast, Hispanic voters are only 5% statewide, and are much more dispersed. No precinct is more than 27% Hispanic, and about 200 precincts are between 10% and 20% Hispanic.

Hispanic voters are also concentrated in different locations as Black voters. They tend to be concentrated in suburbs around the center of the state, and residential patterns are correlated with manufacturing, agriculture, and food processing (Figure 3b). The key pattern to note for EI is that

[20] Although some of these include non-voting active registrants, we refer to all individuals as voters for simplicity.

inference about Hispanic voters are made from precinct-level aggregates in which no observation comes close to majority Hispanic.

Past methodological studies have examined the performance of ecological inference methods on similar examples, but none have interpreted the results across different EI methods. de Benedictis-Kessner (2015) evaluates five EI methods on similar voter file data and finds that King (1997)'s 2×2 and Rosen et al. (2001)'s count model has lower estimation error than Goodman's Regression, but only explores where and why errors differ across methods in a cursory fashion.[21] Barreto et al. (2022) compares differences between King (1997) and Rosen et al. (2001) methods on election results and finds similar estimates, but does not diagnose which one more correctly captures the ground truth. Kuriwaki et al. (2024a) studies the accuracy of Rosen et al. (2001)'s method on the Florida voter file and finds several errors with confidence intervals that are too tight, and speculates that these may be due to the undue influence of homogeneous precincts. None of the three test how including covariates affects the accuracy of the estimates.

Our proposed estimator, `seine`, performs well for White and Hispanic voters, but less so on Black voters. See Figure 4. `seine` estimates 18% of White voters are registered Democrats with a 95% CI of $[0.163, 0.197]$, where the true value is 19.5%, and it estimates that 44% of White voters are registered Republicans with a 95% CI of $[0.418, 0.462]$, where the true value is 42.2%. While the OLS with no covariates and King's EI come close as well, neither has confidence intervals that cover the truth and the estimates are further away.

Among Hispanic voters, shown in the bottom row, the estimates are less accurate across all methods. The linear regression methods (`seine` and Goodman) underestimate the Democratic percentage among Hispanics by double digits; the King and Rosen estimates overestimate it by more. `seine` correctly

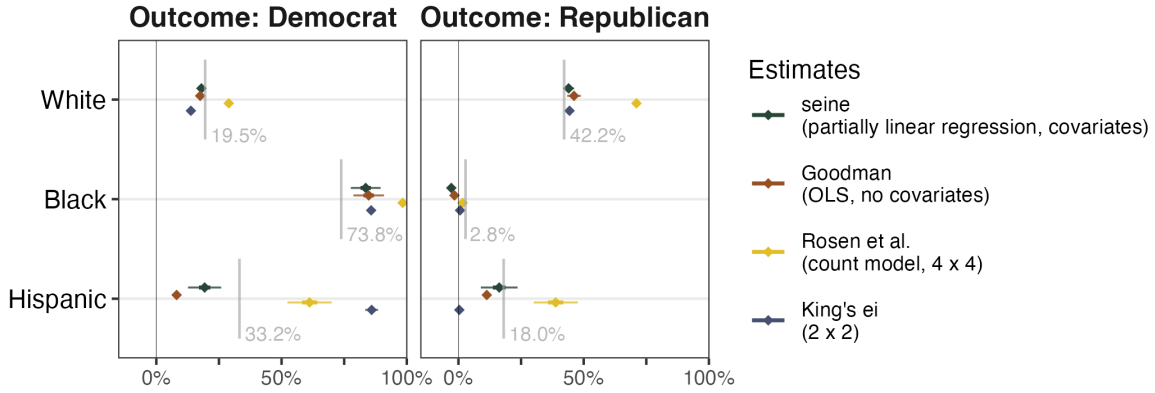


Figure 4: **Accuracy of EI Methods in Uncovering Vote Choice among Racial Groups.** Ecological inference estimates with 95% and 50% confidence intervals, in color. The true values of the registrations are shown in gray vertical lines and labelled. `seine` includes the covariates education, median income, median age, density, and distance to a city/university. Rosen et al.'s model refers to the multinomial Dirichlet model available in `eiPack`.

[21] de Benedictis-Kessner (2015) computes the average precinct-level error in the estimates of percentage White, Black, and Hispanic voters registering Democrat. He finds that Goodman regression's error is over 68 points, Goodman regression with post-hoc truncation has an average error of 25 points, and King (1997) and Rosen et al. (2001) methods having an error of both around 15 points. Even the smallest of these errors is substantially large. We suspect that estimating the global parameter (district or statewide quantities) rather than precinct-level local parameters is much less error-prone. de Benedictis-Kessner (2015) discusses differences in errors may be explained by the racial composition of the five states he studies (373) or the racial composition of the county (377) but does not lay out a specific hypothesis.

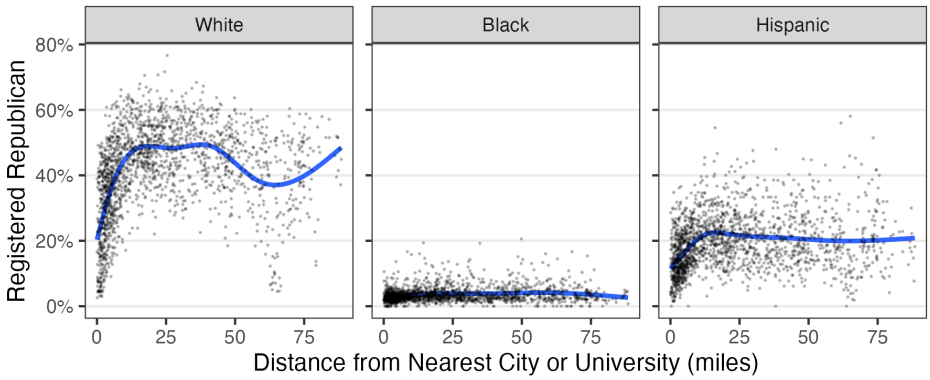


Figure 5: **Potential Confounders for Racially Polarized Voting.** Each point is a precinct with at least 20 voters in all three racial groups, sorted by distance to a nearest city or university on the horizontal axis. The vertical axis shows the ground truth level of Republican registration in those racial groups in those precincts (the local parameter b of interest). Distance to a city is a potential confounder that shapes the parameters nonlinearly. A GAM regression showing the line of best fit is shown in blue.

estimates the Hispanic degree of Republican registration, while all other methods without covariates fail to do so.

Black voters have more stubborn estimation challenges. All methods overestimate the Democratic percentage among Black voters by about 10–20 points. The `seine` methods come the closest but the confidence interval does not cover the true values. Notably, all estimates miss in the direction of *overestimating* racially polarized voting. If Black voters in heavily Black precincts tend to be more Democratic than Black voters in more racially diverse precincts, the intuition of linear regression is consistent with this pattern. The overestimation of Democratic leaning among Black voters also goes hand-in-hand with the underestimation of the Republican leaning of Black voters. The linear regression-based model gives negative estimates of -2.8% and -1.6%. King and Rosen's models, which constrain estimates to be valid proportions, naturally do better, with Rosen's model's estimates covering the true value of 2.8%. We return to the instability of Rosen's model for small groups in the next section.

What sort of confounding does the inclusion of covariates help solve? One potential confounder is the urbanization of an area. Voters of some or all racial groups in urban areas may differ in their Republican leanings than those in rural areas, which would then be reflected in a correlation between the parameter of interest and urbanicity. If urbanicity is also correlated with the prevalence of a racial group (which is apparent in Figure 3a), then this will induce a bias in overall estimates. In Figure 5, we operationalize urbanicity as the distance from the center of a large city (as in Rodden (2019)) or a R1 university, and show its relationship with the parameter of interest. The Republican preference of racial groups in this data indeed differs systematically by the urbanicity of the precinct. Voters near urban areas are systematically less Republican than distant areas, consistent with Rodden (2019) and Kuriwaki et al. (2024a). However, the degree of Republican registration among Black voters is constant. Moreover, the relationship between distance and the parameter appears nonlinear for White and Hispanic voters. This suggests that allowing covariates to enter the model in a flexible functional form as we do with `seine`'s bases is important.

A takeaway from these analyses is that it is easier to estimate the outcomes among dominant racial groups than racial minorities. Linear regression estimators with no truncation struggle to estimate very

small condition means, such as the Republican registration of Black voters. Finally, models that do not incorporate covariates tend to do worse than those that do.

5.2 Ticket splitting

The measurement of ticket splitting is another application of ecological inference. Voters who vote for party *A*'s candidate in one office but vote for party *B*'s candidate in another office are called ticket splitters. The prevalence of these ticket splitters are of interest because they shape the degree of divided government. Each geographic unit provides the vote shares of candidates in their respective contests, with potentially multiple contests (or offices).

Here we evaluate the ability of EI to recover degrees of ticket splitting that occurred in the U.S. House of Representatives elections in 2020. We use a tranche of anonymous ballot records (*cast vote records*) by Kuriwaki et al. (2024b) to simultaneously observe the ground truth rates of ticket splitting exactly. Each cast vote record records a ballot's vote in the U.S. House race, the race for President, which coincided on the same ballot, and a host of other votes for other offices. It also records the precinct in which the vote was cast. An overwhelming majority of the voters—more than 9 in every 10—voted for the same party as they did for President, while the remaining voters either split their ticket or chose not vote for any candidate (undervote). See Section C for details.

Our quantity of interest is the proportion of Biden voters and Trump voters in each congressional district who each split their ticket for a different party in the U.S. House race. Consider ballots from a part of Wisconsin's 8th congressional district, which includes Green Bay and its suburbs, represented by Republican Mike Gallagher from 2017 to 2024. Figure 6 (a) shows the distribution of votes in this district. While only 2% of Trump voters voted for the Democratic House candidate, over 16% of Biden voters voted for Gallagher, the incumbent. Gallagher indeed won the entire district handily, winning more votes than Trump. This may be due to Gallagher's name recognition as an incumbent, his candidate quality, or his relatively more moderate stances (Kuriwaki 2023). Figure 6 (b) shows the aggregate voteshares at the precinct level in this district. Typically, analysts are only given this set of data to infer the level of ticket splitting. In all precincts, the Democratic House candidate trails Biden's voteshare, meaning that Gallagher does better than Trump in all precincts. The gap is relatively uniform across all ranges of the horizontal axis. Indeed, a linear regression line fit to this scatterplot, shown in the blue line has a slope of 0.97.

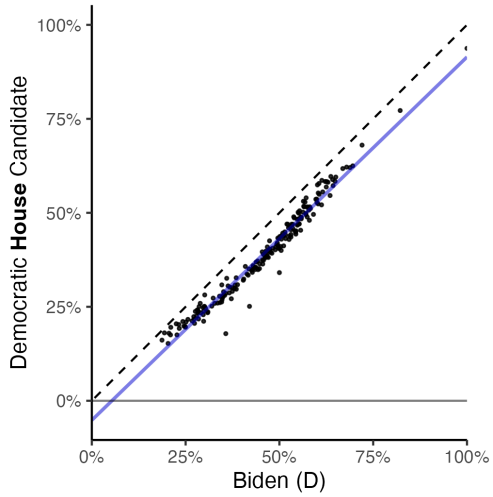
Because of the uniform overperformance of Gallagher, the intercept of the best fit line is negative, at -0.05, implying an impossible negative value of Trump voters who vote for the Democratic candidate. The problem in estimation here is not the presence of outliers. The pattern in the existing data points are strong enough that a few leverage points are unlikely to nudge the line towards a non-negative intercept. Rather, the problem is the lack of sufficient coverage for the predictor variable.

We limit our EI estimation to 63 congressional districts in Kuriwaki et al. (2024b) that are contested by a Democratic and Republican House candidate, and the CVR data contain at least 3 different counties in the district. We impose the latter condition because the covariates we will use are measured at the county level, so a district contained wholly within one county has no variation in the covariate. The

[22] Ballot records are collected county by county. Each district's collection of ballots are only a partial set of the district's counties. The districts of the final data are in AZ, CA, CO, FL, GA, IL, MD, MI, NJ, NV, OH, OR, TX,and WI.

	U.S. House Vote (%)			
Presidential Vote	Democrat	Republican	Other	N
Biden (D)	81.7	16.4	1.8	124,159
Trump (R)	2.1	96.8	1.1	150,998

(a) Joint distribution in WI-08



(b) Precinct aggregates

Figure 6: **Typical Data Aggregation Problem in Ticket Splitting in Wisconsin's 8th Congressional District.** EI problem in precincts from a portion of Wisconsin's 8th Congressional District. The Republican U.S. House Representative, Michael Gallagher, outperformed Donald Trump's performance in this area. (a) Over 16% of Biden voters voted for the Republican, Gallagher, rather than the Democrat. (b) In all precincts, Gallagher ran ahead of Trump, making it difficult to estimate how many, if any, Trump voters voted for the Democratic candidate. The OLS line, shown in blue, has a negative estimate at Biden = 0%.

median congressional district fragment contains 177 precincts.[22] Our validation of ticket splitting is the most comprehensive to date in the literature.[23]

We find that all methods we test tend to underestimate the degree of ticket splitting. Figure 7 compares the estimates on the vertical axis with the CD-level ground truth on the horizontal axis. The first count model estimates a 2 (Biden and Trump vote) by 4 (House Republican, House Democrat, House undervote, and House other vote) problem overestimates the prevalence of ticket splitting by 18 points. Its estimate that 20% of Trump voters split their ticket is highly implausible, and an outlier from the other methods. When we coarsen the setup so that the same algorithm solves a 2×2 problem, with the House vote coarsened to, e.g., House Republican and everything else, the error reduces about three-fold and flips to under-estimating ticket splitting. The severe misestimation may be due to the pattern of the quantities of interest (Figure 9) may be due to the issues with the `eiPack` model as discussed above.

King's 2×2 model of proportion, shown in the third plot, generates estimates that have a lower absolute error than the count models: a mean absolute error of about 3 percentage points instead of 5.5. All but a few estimates still underestimate the true levels of ticket splitting by about the same amount.

[23]Burden and Kimball (1998) and Burden and Kimball (2009) were the first to use ecological inference methods, applying King's (1997) algorithm on congressional district level aggregate data. Cho and Gaines (2004) critiqued the uninformativeness of aggregate data in the case of ticket splitting. Neither work had access to ground truth levels of ticket splitting. Other countries such as Austria and New Zealand reported vote results in a way that allowed ticket splitting rates to be computed from ballots (Klima et al. 2016; Pavía and Romero 2024). By the 2010s, as some U.S. states started to release cast vote records, Park, Hanmer, and Biggers (2014) evaluated King's EI model and Thomsen's nonlinear regression model in ten counties. They did not test other models or include covariates.

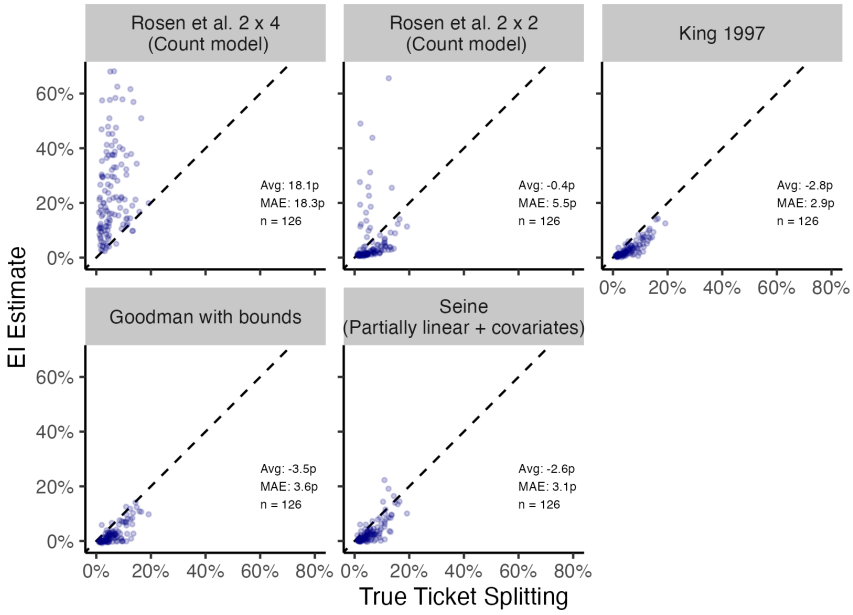


Figure 7: **Predictive Performance of Ticket Splitting.** Each facet shows estimates of congressional district (CD)-level ticket splitting for a given method. Data contain the same districts. For each district, there are two ticket splitting estimates: the one for Biden voters and the one for Trump voters. Seine models use ridge-only, with a bound set to [0, 1]. Covariates are county-level ACS estimates and precinct-level Biden voteshare coarsened into five equally sized bins. Avg: Mean Error. MAE: Mean Absolute Error.

Our linear regression yields similar, if somewhat noisier, patterns. In contrast to the racially polarized voting example, we estimated the regression by including only the regression component and specifying the ridge regression to estimate a solution that is strictly within $[0, 1]$.[24] The model thus loses its double robustness property, but it avoids producing negative, infeasible estimates that linear regression is prone to give in this case. Without covariates, the mean absolute error is 3.6 points, with estimates again under-estimating ticket splitting by the same amount. King's assumption of a truncated normal distribution of the local means turns out to produce estimates that are 0.5 points closer to the truth in this case.

Finally, including covariates in the regression improves estimates. We include the following covariates: county-level median income, county-level proportion White, county population density, county-level proportion of elderly residents, and indicators for five equally populated bins of the precinct-level Presidential vote. The last covariate is a coarsened version of the variable $\overline{\mathbf{X}}$ that serves as an attempt to adjust for contextual effects, which we discuss further below. The other parts of the regression model remain the same: a ridge regression with all covariates interacted with the predictors, and fit to have predictions between 0 and 1. The resulting estimates are nudged upwards, resulting in a mean absolute error of 3 points, comparable to King's model.

[24] Misspecifications in flexibly modeling $f(\mathbf{Z}_g)$ could produce estimates that are implausible, out of the possible range of Y. In these cases, one can simply enforce that the predicted values of the regression when each $\overline{X}_{gk}$ is set to 1 must lie within the range of Y (e.g., 0 to 1). These constraints are linear, since the regression itself is linear. Combined with the least-squares loss and ridge penalty, the overall optimization problem is one of quadratic programming, which can be solved efficiently with standard software. Our `seine` software package (McCartan and Kuriwaki 2025b) implements this approach with and without bounds, and allows for the use of any basis expansion Φ, many of which can be generated with the `bases` package (McCartan 2025).

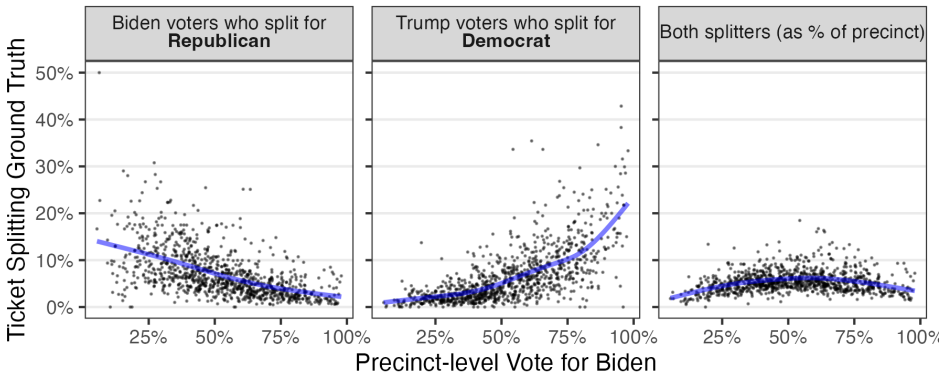


Figure 8: **Variation in Ticket Splitting Rates** The relationship between $\overline{X}$ and quantities of interest across a random sample of 1,000 precincts with more than 100 voters for visual clarity. Each point is a sampled precinct, sorted on the horizontal axis by the percentage of total votes for the Democratic Presidential candidate. Each facet represents a different measure of ticket splitting on the vertical axis. All three plots show that each measure of ticket splitting varies with presidential voteshare.

In summary, King's model (without covariates) and our linear regression with covariates are similarly accurate, with King's model leading to slightly smaller errors in this case. King's model benefits from the choice of the truncated normal distribution, while the linear regression models, though disadvantaged by their tendency to predict towards negative estimates, benefit from contextual covariates that account for the heterogeneity in the local levels of ticket splitting. A regression model that combines both is a potential area of innovation.

Why do all models underestimate rather than overestimate ticket splitting? One reason is that the correlation of votes on votes in this setting is close to 1, and so most lines of best fit with even a slightly non-zero intercept will lead to one of its extremes evaluated out of bounds (Figure 6 (b)).

The second reason is that ticket splitting is somewhat lower in precincts that almost completely favor one Presidential candidate. In other words, the constancy assumption is violated in a way that is prone to underestimation. Figure 8 shows how local ticket splitting varies by precinct-level context. We plot the exact amount of ticket splitting against the two-party presidential voteshare. All three measures show a nonlinear relationship. Biden voters are least likely to split their ticket when they are in nearly unanimous Biden precincts (left panel). Similarly, Trump voters are least likely to split their ticket when they are in nearly unanimous Trump precincts (center panel). These voters in unanimous precincts also disproportionately influence the global estimate, because they are at the extremes and there are more of them in those extremes. Put together, patterns of ticket splitting follow an inverse-U shaped pattern (right panel), where it is highest in mixed precincts where ecological inference is most difficult.

6 Conclusion

There has been much specialized methodological development into techniques for inferring conditional means from aggregate data. Our theoretical approach in this paper offered a new perspective on understanding the statistical problem: a view that highlights the connections between ecological inference and more general concepts in statistical inference that are more familiar to practical users of

ecological inference. We then illustrated how observable patterns in common empirical applications correctly foreshadow the direction of bias.

Existing literature on ecological inference often treats linear regression as a simplistic approximation to a more complicated data structure. A large focus in political methodology in the decade since King (1997) has focused on tailoring software to obey this structure. However, our paper suggests that this move in the literature, while not without its benefits, may have obscured that ecological inference of all sorts are free from ecological fallacy risks unless all confounders are appropriately controlled for. We demonstrated this point with a formalization, a reassessment of the existing computational methods, and for the first time showing how the linear nature of aggregation aids in the estimation the quantity of interest.

Using this framework, we evaluated two of the most common applications of ecological inference in political science on real data, and showed that intuitions from causal inference and regression diagnostics carry to the typical nature of ecological fallacies. Past validations and evaluations have been limited by their typical omission of control variables, and their implications have focused on ranking different ecological inference methods. Our framework correctly anticipated the typical pattern of mis-identification in applications. EI estimates of racially polarized voting tend to be overestimated because racial minorities are typically more left-wing in concentrated areas than in diverse ones. EI estimates of ticket splitting tend to be underestimated in recent elections because the correlation of voteshares between offices is strong, and a party's supporters tend to split their ticket less in concentrated areas than in diverse ones. In both examples, however, the inclusion of covariates reduces the impact of ecological fallacies.

In these respects, the intuition and challenges of observational causal inference carry on to ecological inference. The parameter of interest differs between the two traditions, but we emphasize that the underlying challenge is the same. Ecological inference targets a population conditional average, while causal inference estimates a *ceteris paribus* difference in means. However, to identify population conditional averages from aggregate data, the regression still needs to estimate coefficients accounting for confounders. By the same token, while some users may consider the CAR condition too unrealistic to warrant any ecological inference, the condition is no less reasonable than that of the selection on observable condition in causal inference.

A Methodological Developments in Ecological Inference

We summarize the proposals made in the literature for ecological inference. The list below captures, to our knowledge, most of the major work proposing a new method or methodological adjustment to ecological inference. We omit work whose main purpose is to evaluate the numerical performance of an existing method.

Ecological Inference before 1997:

1. Summary of literature and applications to political science: Achen and Shively (1995)
2. Inclusion of variables: Hanushek, Jackson, and Kain (1974)
3. Neighborhood model: Freedman et al. (1991)
4. Latent utility model: Thomsen (1987)
5. Multinomial models for ($R \times C$) elections: Brown and Payne (1986)

Literature after or adjacent to King (1997), in political science and statistics:

1. Extending 2×2 to $R \times C$:
   - Iterative methods: Rosen et al. (2001)
   - Multinomial Dirichlet count models: Rosen et al. (2001)
   - Beta-binomial models: King, Rosen, and Tanner (1999)
   - Count models allowing for correlation across precincts: Greiner and Quinn (2009)
2. Hierarchical modeling for contextual effects: Wakefield (2004)
3. Relaxing parametric assumptions on the error term: Imai, Lu, and Strauss (2008)
4. Geographic adjacency: Calvo and Escolar (2003), Anselin and Cho (2002)
5. Local smoothing: Chambers and Steel (2001)
6. Narrowing Duncan-Davis bounds: Jiang et al. (2020)
7. Derivation adjustments: McCue (2001)
8. Diagnostic tools for Goodman regression: Gelman et al. (2001), Ansolabehere and Rivers (1995)
9. Integrating survey data: Greiner and Quinn (2010), Glynn et al. (2008)
10. Projection to tomography line: Grofman and Merrill (2004)
11. Temporal dependency across precincts: Quinn (2004), Lewis (2004)
12. Using ecological inference output as an independent variable: Herron and Shotts (2003)

In other disciplinary traditions, work not related or not responding to King's model:

1. Extensions of Thomsen's regression: Park (2008), Pavía and Thomsen (2024)
2. Linear programming with constraints: Pavía and Romero (2024)
3. Inferences on bounds in generalized regression problem: Cross and Manski (2002)
4. Ecological inferences with continuous predictor (price) and instruments for predictor: Berry, Levinsohn, and Pakes (2004) (BLP); flexible functional forms with microdata: Berry and Haile (2024)
5. Inference on bounds with causal inference framework: Fan, Sherman, and Shum (2016)
6. Modeling with individual data and aggregate outcomes: Flaxman, Wang, and Smola (2015), Rosenman and Viswanathan (2018), Fishman and Rosenman (2024)

B Proofs of propositions

B.1 Proof of Proposition 1

Proof. When $\mathbf{Z}$ is empty, CAR and CCAR coincide. In this case, Proposition 1 follows immediately from Proposition 2, because as shown in the main text, Proposition 2 implies that the true CEF is linear in $\overline{\mathbf{X}}_g$.

B.2 Proof of Proposition 2

Proof. The CAR assumption implies $\mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g\big] = \mathbb{E}\big[\mathbf{B}_g \mid \mathbf{Z}_g\big]$ by the law of total expectation (integrating out N_g). Applying this property once to drop the conditioning on $\overline{X}_{gk} = 1$ (which, by the sum-to-1 constraint, fixes $\overline{\mathbf{X}}_g$) and applying CAR to condition on $\overline{\mathbf{X}}_g$ and N_g,

$$\begin{aligned}
\mathbb{E}\big[N_{gk}\mathbb{E}\big[\overline{Y}_g \mid \mathbf{Z}_g, \overline{X}_{gk} = 1\big]\big] &= \mathbb{E}\Big[N_{gk}\mathbb{E}\Big[B_g^\top \overline{\mathbf{X}}_g \mid \mathbf{Z}_g, \overline{X}_{gk} = 1\Big]\Big] \\
&= \mathbb{E}\big[N_{gk}\mathbb{E}\big[B_{gk} \mid \mathbf{Z}_g, \overline{X}_{gk} = 1\big]\big] \\
&= \mathbb{E}\big[N_{gk}\mathbb{E}\big[B_{gk} \mid \mathbf{Z}_g\big]\big] \\
&= \mathbb{E}\Big[N_{gk}\mathbb{E}\Big[B_{gk} \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g, N_g\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[N_{gk}B_{gk} \mid \mathbf{Z}_g, \overline{\mathbf{X}}_g, N_g\Big]\Big] \\
&= \mathbb{E}\big[N_{gk}B_{gk}\big] = E\big[N_{gk}\big]\beta_k.
\end{aligned}$$

Dividing by $\mathbb{E}\big[N_{gk}\big]$ yields the result.

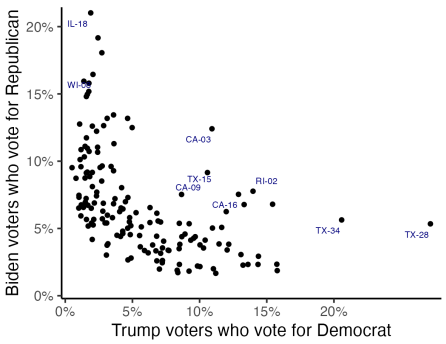
C Details on Empirical Validations

C.1 Cast Vote Record Data

Figure 9 shows an overview of this data. The table in panel (a) shows row percentages that sum to 100% in each row. Among the 20 million or so ballots cast for Joseph Biden, the Democratic candidate for President, 5.6% voted for a Republican candidate, and 4% either skipped the House race (undervote) or voted for a third-party candidate. In panel (b) we show the level of ticket splitting among Trump voters (on the horizontal axis) against the level of ticket splitting among Biden voters in the same district (on the vertical axis). The two quantities of interest are negatively correlated. Some districts exhibit substantial ticket splitting for the Republican, and others exhibit splitting for the Democrat. These two variables cancel out when added to the entire population in (a).

	U.S. House Vote (%)			
Presidential Vote	Democrat	Republican	Other	N
Biden (D)	90.3	5.6	4.0	20,627,643
Trump (R)	4.8	91.4	3.7	17,390,960

(a) All available records



(b) Estimands, by district

Figure 9: **Ticket Splitting Patterns from Election Data.** The distribution of vote choices in the 2020 election data from Kuriwaki et al. (2024b). (a) Across all districts available in the dataset, approximately 5% of Biden voters and Trump voters vote for a different party's candidates in the U.S. House race. These quantities of interest are off-diagonal entries and cancel out when aggregated. (b) Quantities of interest, separated for each congressional district. In typical districts ticket splitting is asymmetric and benefits one candidate more than their opponent.

Bibliography

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

Anselin, Luc, and Wendy K Tam Cho. 2002. "Spatial Effects and Ecological Inference". *Political Analysis* 10 (3): 276–97.

Ansolabehere, Stephen, and Douglas Rivers. 1995. "Bias in Ecological Regression". *Massachusetts: Department of Political*.

Barreto, Matt, Loren Collingwood, Sergio Garcia-Rios, and Kassra AR Oskooii. 2022. "Estimating Candidate Support in Voting Rights Act Cases: Comparing Iterative EI and EI - R x C Methods". *Sociological Methods & Research* 51 (1): 271–304.

Benedictis-Kessner, Justin de. 2015. "Evidence in Voting Rights Act Litigation: Producing Accurate Estimates of Racial Voting Patterns". *Election Law Journal* 14 (4): 361–81.

Berry, Steven, and Philip Haile. 2024. "Nonparametric Identification of Differentiated Products Demand Using Micro Data". *Econometrica* 92 (4): 1135–62.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market". *Journal of Political Economy* 112 (1): 68–105.

Blackwell, Matthew. 2025. "A User's Guide to Statistical Inference and Regression."

Blalock, Hubert M. 1984. "Contextual Effects Models: Theoretical and Methodological Issues". *Annual Review of Sociology*, 353–72.

Brown, Philip J, and Clive D Payne. 1986. "Aggregate Data, Ecological Regression, And Voting Transitions". *Journal of the American Statistical Association* 81 (394): 452–60.

Burden, Barry C, and David C Kimball. 1998. "A New Approach to the Study of Ticket Splitting". *American Political Science Review* 92 (3): 533–44.

Burden, Barry C, and David C Kimball. 2009. *Why Americans split their tickets: Campaigns, competition, and divided government*. University of Michigan Press.

Calvo, Ernesto, and Marcelo Escolar. 2003. "The Local Voter: A Geographically Weighted Approach to Ecological Inference". *American Journal of Political Science* 47 (1): 189–204.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language". *Journal of Statistical Software* 76:1–32.

Chambers, Raymond L, and David G Steel. 2001. "Simple Methods for Ecological Inference in 2 x 2 Tables". *Journal of the Royal Statistical Society: Series a (Statistics in Society)* 164 (1): 175–92.

Chatterjee, Samprit, and Ali S Hadi. 1986. "Influential Observations, High Leverage Points, And Outliers in Linear Regression". *Statistical Science*, 379–93.

Chernozhukov, Victor, Whitney K Newey, and James Robins. 2018. "Double/de-Biased Machine Learning Using Regularized Riesz Representers."

Cho, Wendy K Tam, and Brian J Gaines. 2004. "The Limits of Ecological Inference: The Case of Split-Ticket Voting". *American Journal of Political Science* 48 (1): 152–71.

Cross, Philip J, and Charles F Manski. 2002. "Regressions, Short and Long". *Econometrica* 70 (1): 357–68.

Cunningham, John P, Philipp Hennig, and Simon Lacoste-Julien. 2011. "Gaussian Probabilities and Expectation Propagation". *Arxiv Preprint Arxiv:1111.6832*.

Ding, Patrick, Guido Imbens, Zhaonan Qu, and Yinyu Ye. 2024. "Computationally Efficient Estimation of Large Probit Models". *Arxiv Preprint Arxiv:2407.09371*.

Fan, Jianqing, and Wenyang Zhang. 1999. "Statistical Estimation in Varying Coefficient Models". *The Annals of Statistics* 27 (5): 1491–1518.

Fan, Yanqin, Robert Sherman, and Matthew Shum. 2016. "Estimation and Inference in an Ecological Inference Model". *Journal of Econometric Methods* 5 (1): 17–48.

Fishman, Nic, and Evan Rosenman. 2024. "Estimating Vote Choice in US Elections with Approximate Poisson-Binomial Logistic Regression". In *OPT 2024: Optimization for Machine Learning*.

Flaxman, Seth R, Yu-Xiang Wang, and Alexander J Smola. 2015. "Who Supported Obama in 2012? Ecological Inference Through Distribution Regression". In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 289–98.

Freedman, David A, Stephen P Klein, Jerome Sacks, Charles A Smyth, and Charles G Everett. 1991. "Ecological Regression and Voting Rights". *Evaluation Review* 15 (6): 673–711.

Gelman, Andrew, David K Park, Stephen Ansolabehere, Phillip N Price, and Lorraine C Minnite. 2001. "Models, Assumptions and Model Checking in Ecological Regressions". *Journal of the Royal Statistical Society Series A: Statistics in Society* 164 (1): 101–18.

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: Statistics in Society* 171 (1): 179–202.

Goodman, Leo A. 1953. "Ecological Regressions and Behavior of Individuals.". *American Sociological Review* 18 (6).

Greiner, D James, and Kevin M Quinn. 2010. "Exit Polling and Racial Bloc Voting: Combining Individual-Level and R x C Ecological Data". *The Annals of Applied Statistics*, 1774–96.

Greiner, James D, 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: Statistics in Society* 172 (1): 67–81.

Griffiths, William E. 1972. "Estimation of Actual Response Coefficients in the Hildreth-Houck Random Coefficient Model". *Journal of the American Statistical Association* 67 (339): 633–35.

Grofman, Bernard, and Samuel Merrill. 2004. "Ecological Regression and Ecological Inference". Edited by Gary King, Martin A. Tanner, and Ori Rosen. *Ecological Inference: New Methodological Strategies*, Ch.5.

Hanushek, Eric A, John E Jackson, and John F Kain. 1974. "Model Specification, Use of Aggregate Data, And the Ecological Correlation Fallacy". *Political Methodology*, 89–107.

Hastie, Trevor, and Robert Tibshirani. 1993. "Varying-Coefficient Models". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 55 (4): 757–79.

Heitjan, Daniel F, and Donald B Rubin. 1991. "Ignorability and Coarse Data". *The Annals of Statistics*, 2244–53.

Herron, Michael C, and Kenneth W Shotts. 2003. "Using Ecological Inference Point Estimates as Dependent Variables in Second-Stage Linear Regressions". *Political Analysis* 11 (1): 44–64.

Imai, Kosuke, Ying Lu, and Aaron Strauss. 2008. "Bayesian and Likelihood Inference for 2x2 Ecological Tables: An Incomplete-Data Approach". *Political Analysis* 16 (1): 41–69.

Jiang, Wenxin, Gary King, Allen Schmaltz, and Martin A Tanner. 2020. "Ecological Regression with Partial Identification". *Political Analysis* 28 (1): 65–86.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

King, Gary, Ori Rosen, and Martin A Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference". *Sociological Methods & Research* 28 (1): 61–90.

Klima, André, Paul W Thurner, Christoph Molnar, Thomas Schlesinger, and Helmut Küchenhoff. 2016. "Estimation of Voter Transitions Based on Ecological Inference: An Empirical Assessment of Different Approaches". *Asta Advances in Statistical Analysis* 100 (2): 133–59.

Kuriwaki, Shiro. 2023. "Ticket Splitting in a Nationalized Era."

Kuriwaki, Shiro, Stephen Ansolabehere, Angelo Dagonel, and Soichiro Yamauchi. 2024a. "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level". *American Political Science Review* 118 (2): 922–39.

Kuriwaki, Shiro, Mason Reece, Samuel Baltz, Aleksandra Conevska, Joseph R Loffredo, Can Mutlu, Taran Samarth, et al. 2024b. "Cast Vote Records: A Database of Ballots from the 2020 US Election". *Scientific Data* 11 (1): 1304.

Lau, Olivia, Ryan T Moore, and Michael Kellermann. 2007. "eiPack: R x C Ecological Inference and Higher-Dimension Data Management". *New Functions for Multivariate Analysis* 7 (1): 43.

Lewis, Jeffrey B. 2001. "Understanding King's Ecological Inference Model: A Method-of-Moments Approach". *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 34 (4): 170–88.

Lewis, Jeffrey B. 2004. "Extending King's Ecological Inference Model to Multiple Elections Using Markov Chain Monte Carlo". Edited by Gary King, Martin A. Tanner, and Ori Rosen. *Ecological Inference: New Methodological Strategies*, Ch.4.

McCartan, Cory. 2025. "Bases: Basis Expansions for Regression Modeling". https://corymccartan.com/bases/.

McCartan, Cory, and Shiro Kuriwaki. 2025a. "Identification and Estimation of Conditional Means from Aggregate Data". *In Preparation*.

McCartan, Cory, and Shiro Kuriwaki. 2025b. "Seine: Semiparametric Ecological Inference". 2025. https://corymccartan.com/seine/.

McCue, Kenneth F. 2001. "The Statistical Foundations of the EI Method". *The American Statistician* 55 (2): 106–10.

Park, Won-ho. 2008. "Ecological Inference and Aggregate Analysis of Elections.."

Park, Won-ho, Michael J Hanmer, and Daniel R Biggers. 2014. "Ecological Inference under Unfavorable Conditions: Straight and Split-Ticket Voting in Diverse Settings and Small Samples". *Electoral Studies* 36:192–203.

Pavía, Jose M, and Rafael Romero. 2024. "Improving Estimates Accuracy of Voter Transitions. Two New Algorithms for Ecological Inference Based on Linear Programming". *Sociological Methods & Research* 53 (3): 1491–1533.

Pavía, Jose M, and Søren Risbjerg Thomsen. 2024. "ecolRxC: Ecological Inference Estimation of R x C Tables Using Latent Structure Approaches". *Political Science Research and Methods*, 1–19.

Phillips, Kevin P. 2014. *The Emerging Republican Majority: Updated Edition*. Princeton University Press.

Quinn, Kevin. 2004. "Ecological Inference in the Presence of Temporal Dependence". Edited by Gary King, Martin A. Tanner, and Ori Rosen. *Ecological Inference: New Methodological Strategies*, Ch.9.

Rivers, Douglas. 1998. "Review of: A Solution to the Ecological Inference Problem by Gary King". *American Political Science Review* 92 (2): 442–43.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data". *Journal of the American Statistical Association* 90 (429): 106–21.

Robinson, WS. 1950. "Ecological Correlations and the Behavior of Individuals". *American Sociological Review* 15 (3): 351–57.

Rodden, Jonathan A. 2019. *Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide*. Basic Books.

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): 134–56.

Rosenman, Evan, and Nitin Viswanathan. 2018. "Using Poisson Binomial GLMs to Reveal Voter Preferences". *Arxiv Preprint Arxiv:1802.01053*.

Schoenberger, Robert A, and David R Segal. 1971. "The Ecology of Dissent: The Southern Wallace Vote in 1968". *Midwest Journal of Political Science* 15 (3): 583–86.

Shen, Xiaotong, and Wing Hung Wong. 1994. "Convergence Rate of Sieve Estimates". *The Annals of Statistics*, 580–615.

Thomsen, Søren Risbjerg. 1987. "Danish Elections 1920-79. A Logit Approach to Ecological Analysis and Inference.". *Politica*.

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

Wright, Gerald C. 1977. "Contextual Models of Electoral Behavior: The Southern Wallace Vote". *American Political Science Review* 71 (2): 497–508.

Wu, Kaiwen, and Jacob R Gardner. 2024. "A Fast, Robust Elliptical Slice Sampling Implementation for Linearly Truncated Multivariate Normal Distributions". *Arxiv Preprint Arxiv:2407.10449*.