Title

Some Empirical Evidence on the Impact of Measurement Errors in Making Ecological Inferences

Authors

Cho, Wendy K. T.
Judge, George G.

Publication Date

2003

Some Empirical Evidence on the Impact of Measurement Errors in Making Ecological Inferences

Wendy K. Tam Cho
University of Illinois at Urbana-Champaign

George G. Judge*
University of California at Berkeley

Abstract

We seek to identify the impact of data measurement error problems in the context of ecological inference applications. We explore the statistical and substantive implications of using inaccurate proxy variables in the estimation and inference process. The focus of our analysis is on applications of ecological inference in cases involving the Voting Rights Act. We demonstrate our findings with a unique data set on racial registration and turnout in Louisiana and South Carolina.

Keywords: Ecological Inference, Maximum Entropy, Voting Rights.

* *Address for correspondence*: George G. Judge, 207 Giannini Hall, University of California, Berkeley, CA 94720–3310. E-mail: *judge@are.berkeley.edu*.

The authors owe a debt of gratitude to Michael Tomz and Robert van Houewling for sharing their Louisiana and South Carolina data, to Geoffrey Brewster for research assistance in merging the census data to the electoral counterparts for the Louisiana data, and to David Lublin and Stephen Voss for supplying some additional data. Without implicating them, we also want to thank Bernard Grofman, Morgan Kousser, Ben Pelzer, Stephen Voss and Christina Wolbrecht for their helpful and substantive comments.

1 Introduction

In the ecological inference chain there are four important links. One link is related to the statistical model that describes a plausible sampling process. A second link involves the estimation method used to process and recover estimates from the observable sample data. The third link is connected to the observable data that form the basis for the estimation and inference. Finally, the fourth link is the conceptual framework that ties the first three links together. In this paper, our focus is on the data link in this chain, its impact on the other described links, and the corresponding statistical and substantive political conclusions.

In the social sciences, non-experimental data are the primary sample information used for estimation and inference and the corresponding parameter estimates that are used for understanding, prediction, choice, and decision making. This data restriction, while important across the social sciences, is especially acute in certain applications in political science where individual-level information on vote choice is sought, but the secret ballot precludes the availability of these voter response data. Because these questions are the basis for many social scientific theories but these types of data are not observable, a good deal of effort has been expended over the last five decades toward developing estimation and inference tools that would yield the missing micro information from aggregate macro data. The seminal works of Robinson (1950) and Goodman (1953, 1959), and the more recent works of Achen and Shively (1995) and King (1997) are but a few examples. Building on these productive efforts, Judge, Miller, and Cho (2002) formulated the ecological inference problem as an ill-posed inverse problem and suggested information-theoretic procedures as a basis for recovering micro responses from aggregate data. In this paper, we build on this work and focus specifically on the impact of data measurement error.

In particular ecological inference applications, the observed data and the unknown and unobservable parameters for the problem are often summarized in the form of a contingency table where the observable data are reflected in row and column sums and the unknown or unobservable data are the conditional probabilities in the interior or cells of the table. In a particular application, the row sum group data may contain a significant measurement error component.

In this paper, without loss of generality, we focus our discussion on specific applications to the voting rights literature. In these cases, one might be interested, for example, in how different racial groups cast their ballots. Since ballots are secret, we have no direct measure of individual vote choice. One may construct data set aggregated at the election precinct level to examine this phenomenon. For instance, it is possible to merge reasonably the census racial data to the election geography. One may then use another procedure (e.g., name identification) to place certain registrants into certain racial groups. However, the variable of interest, turnout (not registration) data by race, is unattainable. Complicating

matters, the racial registration data is injected with error in the data processing stage. Indeed, there are multiple sources of possible additional error, including, for example, the documented undercount of minorities by the U.S. Census, or the discrepancies that arise in the quality of data across time and/or place. If this measurement error is not taken into account, the assumptions underlying the statistical model may be violated. When these assumptions are violated, our ability to recover the unknown conditional probabilities in the table cells and to trace out the corresponding political implications may be severely restricted. In this context, we also face the ecological inference problem where based on the available aggregate data, the parameters that we seek to recover from the precinct-level data are the propensities, proportions, or conditional probabilities of individuals from the various groups of interest to vote for a particular candidate.

This paper begins with a presentation of the basic voter behavior problem in this ecological inference application. We model this problem as a pure inverse problem and suggest an information-theoretic formulation and solution. Next, we provide an explanation of the measurement error problem. Using information-theoretic procedures, we then present some empirical results from some unique data on the 2000 elections in Louisiana and South Carolina (where registration data *and* turnout data are available by race) to demonstrate the problems associated with using an imperfect measure of who voted. Finally, we conclude with some general remarks concerning the implications of this approach to the measurement error problem and proffer some suggestions for mitigating its impact.

2 Notation and Basic Ecological Inference Problem

In this section, we formulate inverse models that one might develop based solely on a contingency-type table containing known and unknown quantities. At this stage, our concern is with the entries in the table, and little consideration is given to a theory that might have generated these data. We will develop and utilize this theory at a later point.

To develop a model that will reflect the characteristics of voter response, consider the observed outcomes for a particular election across $i = 1, \ldots, m$ electoral units (e.g., precincts or districts). Each unit has $j = 1, \ldots, g$ types of individual voters and $k = 1, \ldots, c$ candidates for office, including perhaps an abstention or no-vote category). Assume without loss of generality that the election units are precincts. For each precinct, the observed information is the number of votes for each candidate, $N_{i\cdot k} = \sum_{j=1}^{g} N_{ijk}$, and the number of voters in each group, $N_{ij\cdot} = \sum_{k=1}^{c} N_{ijk}$. The total number of ballots cast in the precinct is $N_i = \sum_{j=1}^{g} \sum_{k=1}^{c} N_{ijk}$. Because of the secret ballot, the total number of votes cast by each group for particular candidates in the election is unknown and unobserved. Given the observed data, our initial objective is to formulate an inverse model that will permit us to estimate

Table 1: Known and Unknown Components in an Ecological Inference Problem

	Candidate				
Group	1	2	3	4	Count
1	$\beta_{11}N_{1\cdot}$	$\beta_{12}N_{1\cdot}$	$\beta_{13}N_{1\cdot}$	$\beta_{14}N_{1\cdot}$	$N_{1\cdot}$
2	$\beta_{21}N_{2\cdot}$	$\beta_{22}N_{2\cdot}$	$\beta_{23}N_{2\cdot}$	$\beta_{24}N_{2\cdot}$	$N_{2\cdot}$
3	$\beta_{31}N_{3\cdot}$	$\beta_{32}N_{3\cdot}$	$\beta_{33}N_{3\cdot}$	$\beta_{34}N_{3\cdot}$	$N_{3\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	$N_{\cdot 4}$	N

N_{ijk}, the unobserved number of votes cast in precinct i by voters of type j for candidate k, from the sample of voters who voted in the election.

For the purposes of formulating the basic model, the data may be stated in terms of the observed row or column proportions, i.e., for precinct i, $n_{i\cdot k} = N_{i\cdot k}/N_i$ or $n_{ij\cdot} = N_{ij\cdot}/N_i$. The inverse problem may be equivalently stated in terms of the proportion of voters in each category, $\beta_{ijk} = N_{ijk}/N_{ij\cdot} = n_{ijk}/n_{ij\cdot}$, where $\sum_{k=1}^{c} \beta_{ijk} = 1$ for each i and j. In this context, β_{ijk} may be interpreted as the conditional probability that voters in precinct i and group j voted for candidate k, where the conditioning indices are i and j. In the Voting Rights arena, the index j represents race, and the index k represents a set of candidates. The objective in this case is to estimate the conditional probability that voters selected candidate k given that they are a member of racial group j.

2.1 Modeling Voting Behavior as an Inverse Problem

The components of this information recovery problem for a particular precinct (i suppressed) are summarized in Table 1. The observed number of ballots cast by registered voters in each group ($N_{j\cdot}$) are the row sums, and the observed number of votes received by each candidate ($N_{\cdot k}$) are the column sums. What we do not know and cannot observe is the number of votes cast by each group, N_{jk}, or the proportion of votes cast by each group for each candidate, n_{jk}. If the conditional probabilities β_{jk} were known, we could derive the unknown number of voters as $N_{jk} = \beta_{jk}N_{j\cdot}$. However, the conditional probabilities are unobserved and not accessible by direct measurement. Thus, we are faced with an inverse problem where we must use indirect, partial, and incomplete macro measurements as a basis for recovering the unknown conditional probabilities.

One bit of structure is provided by the realization that the conditional probabilities β_{jk} must satisfy the row sum, $\sum_{k=1}^{c} \beta_{jk} = 1$, and column sum, $\sum_{j=1}^{g} \beta_{jk} N_{j\cdot} = N_{\cdot k}$, conditions. If we make use of the

column sum conditions, we have the relationship

$$n_{i\cdot k} = \sum_{j=1}^{g} n_{ij\cdot}\beta_{ijk} \ , \tag{1}$$

for $i = 1, \ldots, m$ and $k = 1, \ldots, c$. To formalize our notation, we let $\mathbf{x}(i) = (n_{i1\cdot}\ n_{i2\cdot}\ \cdots\ n_{ig\cdot})'$ represent the $(g \times 1)$ vector of proportions for each of the groups $j = 1, \ldots, g$ in precinct i, and let $\boldsymbol{y}(i) = (n_{i\cdot 1}\ n_{i\cdot 2}\ \cdots\ n_{i\cdot c})'$ represent the $(c \times 1)$ sample outcome vector of vote proportions for each candidate $k = 1, \ldots, c$ in precinct i. Then, the relationship among the observed marginal proportions and unknown conditional probabilities may be written as

$$\boldsymbol{y}'(i) = \mathbf{x}'(i)\mathbf{B}(i) \ . \tag{2}$$

The component $\mathbf{B}(i) = (\boldsymbol{\beta}_{i1}\ \boldsymbol{\beta}_{i2}\ \cdots\ \boldsymbol{\beta}_{ic})$ is an unknown and unobservable $(g \times c)$ matrix of conditional probabilities and $\boldsymbol{\beta}_{ik} = (\beta_{i1k}\ \beta_{i2k}\ \cdots\ \beta_{igk})'$ is the $(g \times 1)$ vector of conditional probabilities associated with precinct i and candidate k. If we rewrite $\mathbf{B}(i)$ in $(gc \times 1)$ vectorized form as $\boldsymbol{\beta}(i) = \text{vec}(\boldsymbol{\beta}(i)) = \left(\boldsymbol{\beta}'_{i1}\ \boldsymbol{\beta}'_{i2}\ \cdots\ \boldsymbol{\beta}'_{ic}\right)'$, then we may rewrite (2) as

$$\begin{bmatrix} y_1(i) \\ y_2(i) \\ \vdots \\ y_c(i) \end{bmatrix} = \begin{bmatrix} \mathbf{x}'(i) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'(i) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'(i) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{i1} \\ \boldsymbol{\beta}_{i2} \\ \vdots \\ \boldsymbol{\beta}_{ic} \end{bmatrix} \tag{3}$$

or in compact form as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} \tag{4}$$

thus providing a basis for interpreting (3) as $m \geq 2$ precincts.

Under this form, the absence of sampling errors and other stochastic noise components in Equations (2)–(4) implies that the problem of recovering $\boldsymbol{\beta}$ from observed $\boldsymbol{y}$ and $\mathbf{X}$ is a pure inverse problem. For each precinct-specific problem (3), note that the matrix $\mathbf{X}(i)$ has dimension $(c \times gc)$ and is underdetermined and generally not invertible. Thus, under traditional mathematical inversion procedures, the voter pure inverse (VPI) problem is said to be ill-posed, and the solution space for the problem contains arbitrary parameters.

3 Information Theoretic Formulation and Solution

This general formulation expressed in Equations (1)–(4) captures a frequently occurring problem in political analysis where a function must be inferred from insufficient information that specifies only a feasible or plausible set of functions of solutions. In other words, this is an ill-posed inverse problem that is fundamentally underdetermined and indeterminate because there are more unknown and unobservable parameters than data points on which to base a solution. Consequently, prima facie, using traditional logic, insufficient sample information exists to solve the problem uniquely. In order to provide a basis for proceeding in these ill-posed situations, one may redefine the inverse problem by selecting an element of the feasible set by some ad hoc rule such as minimizing some function by L_2–norm minimization or the Kullback-Leibler (K-L) minimum distance measure.

3.1 Choosing the Criterion Function

As with the likelihood and least squares solution for the general linear statistical model, the long journey in defining a solution begins with the selection of a goodness-of-fit criterion. If we recognize and maintain the distinction that the unknown elements β_{ijk} are conditional probabilities rather than joint probabilities, then our pure voting inverse model is similar to allocating values to each of the cells in a contingency table. Consequently, the Cressie-Read power-divergence (CRPD) statistic (Cressie and Read 1984, Read and Cressie 1988; Baggerly 1998)

$$I(\mathbf{w}, \boldsymbol{q}, \lambda) = \frac{2}{\lambda(1+\lambda)} \sum_i w_i \left[\left(\frac{w_i}{q_i} \right)^{\lambda} - 1 \right] , \tag{5}$$

provides a pseudo-distance measure between $\mathbf{w}$ (i.e., conditional probabilities in the VPI problem) and a set of reference weights $\boldsymbol{q}$. The discrete weights must satisfy $(w_i, q_i) \in (0,1) \times (0,1) \; \forall \; i$ and $\sum_i w_i = \sum_i q_i = 1$. The distance measure (5) encompasses a family of empirical likelihood estimation objective functions that includes the Kullback-Leibler (Kullback 1959, Gokhale and Kullback 1978) entropy (Jaynes 1957) estimating criterion and the empirical likelihood criterion (Owen 1988, 1990). In these cases, the minimum distance estimation problem is solved by maximizing the criterion function with respect to $\mathbf{w}$.

3.2 Solution to a Pure Inverse Problem

Under the minimum CRPD estimation criterion, an estimator for the VPI problem may be formulated as

$$\arg\min_{\beta_{ijk}} \frac{2}{\lambda(1+\lambda)} \sum_{i=1}^{m} \sum_{j=1}^{g} \sum_{k=1}^{c} \beta_{ijk} \left[\left(\frac{\beta_{ijk}}{q_{ijk}} \right)^{\lambda} - 1 \right] , \tag{6}$$

subject to

$$n_{i \cdot k} = \sum_{j=1}^{g} n_{ij \cdot} \beta_{ijk} \tag{7}$$

and the row-sum or additivity condition

$$\sum_{k=1}^{c} \beta_{ijk} = 1 \qquad \forall \; i, j \; . \tag{8}$$

The solution for the conditional probabilities in this constrained minimization problem is

$$\widehat{\beta}_{ijk} = q_{ijk} \left[\frac{1}{1+\lambda} + \frac{\lambda}{2} \left(\widehat{\alpha}_{ik} n_{ij \cdot} + \widehat{\gamma}_{ij} \right) \right]^{1/\lambda} \; . \tag{9}$$

In general, the solution does not have a closed-form expression and and the optimal values of the unknown parameters must be numerically determined. As $\lambda \to 0$ in (6), the estimating criterion is

$$\arg\min_{\beta_{ijk}} \sum_{i=1}^{m} \sum_{j=1}^{g} \sum_{k=1}^{c} \beta_{ijk} \ln \left(\frac{\beta_{ijk}}{q_{ijk}} \right) \; . \tag{10}$$

and the intermediate solution for the constrained optimal β_{ijk} may be expressed as

$$\widehat{\beta}_{ijk} = \frac{q_{ijk} \exp \left(\widehat{\alpha}_{ik} n_{ij \cdot} \right)}{\sum_{k=1}^{c} q_{ijk} \exp \left(\widehat{\alpha}_{ik} n_{ij \cdot} \right)} \; . \tag{11}$$

The elements $\widehat{\alpha}_{ik}$ are the optimal values of the Lagrange multipliers for constraints (2)–(3). If one is willing to make the additional assumptions necessary to formulate (4) as an inverse problem with noise, then the procedures such as those proposed by Goodman (1953, 1959) or King (1997) may be used as a basis for estimating the β_{ijk}.

4 Basic Idea and Approach

Given the possibility of formulating voter response as an inverse problem and the possibility of using the CRPD estimation criterion as a basis for a solution, we now focus on the input variable, $\mathbf{X}$, in (2)

or (4). In the context of Voting Rights cases and this input variable, sample information in the form of census or registration data are usually used. However, the usually unobserved voter turnout data would appear to be the correct input variable in our ecological inference formulations. To explore the implications of this measurement error in the context of the pure inverse problem (2) or (4), if $\mathbf{X}^*$ is the observed voter racial registration group shares and $\mathbf{X}$ is the true unobservable voter racial turnout group shares, then we may model $\mathbf{X}^*$ as

$$\mathbf{X}^* = \mathbf{X} + \mathbf{u} \ , \tag{12}$$

where $\mathbf{u}$ is an unobserved noise vector. Therefore, in the context of (2), the underlying inverse model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} \ , \tag{13}$$

but the observable version is an inverse problem with noise

$$\boldsymbol{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \ , \tag{14}$$

where

$$\boldsymbol{\epsilon}^* = \mathbf{u}\boldsymbol{\beta} \ . \tag{15}$$

In the inverse model (14) that is based on the observable data, $\mathbf{X}^*$ is correlated with the random noise $\boldsymbol{\epsilon}^*$. Thus, if one makes use of the traditional L_2–norm, the usual condition in the linear model that states that the right-hand-side explanatory variables are orthogonal in expectation to the error process is violated. Since this condition does not hold, traditional estimation rules based on $\mathrm{E}(\mathbf{x}^{*\prime}\boldsymbol{\epsilon}^*) = \mathbf{0}$ will not have the property of consistency and, in finite samples, will have questionable statistical performance.

While the statistical implications of errors in the $\mathbf{X}$ variable are important for traditional ecological inference estimation methods, our emphasis is on the substantive implications of this measurement error when estimating the conditional probabilities associated with voter response and one uses the pure inverse model $y = \mathbf{X}^*\boldsymbol{\beta}$ instead of $y = \mathbf{X}\boldsymbol{\beta}$. By substantive implications, we mean the impact on court decisions in Voting Rights cases from using estimates based on the incorrect independent variable, $\mathbf{X}^*$. To sort out this impact, we use 2000 election data from Louisiana and South Carolina along with the information-theoretic pure inverse model and solution basis suggested in Sections 2 and 3 to compare and contrast estimates of $\boldsymbol{\beta}$ that result when $\mathbf{X}$ or $\mathbf{X}^*$ is used. If one is willing to make the additional assumptions necessary, the traditional ecological inference models and the corresponding

estimators could be used to make similar measurement error comparisons as they relate to estimation and inference.

5 The Case of Louisiana and South Carolina

Using the pure inverse formulations and solutions developed in Sections 2–3, we demonstrate the impact of a realization of measurement error in two sets of real data. The first data set is composed of election returns from the 2000 election in Louisiana. The second data set has similar data for the 2000 elections in South Carolina. Both data sets are unique and ideal for our purposes here because these two southern states provide not only registration data by race, but also turnout data by race. Hence, we have a true measure of our independent variable (turnout by race) and an imperfect proxy of our independent variable (voter registration by race). With these data, we examine the empirical differences that arise from using a proxy variable as the independent variable, $\mathbf{X}$, while employing a pure inverse model.

In Table 2, some data from the first congressional districts in Louisiana and South Carolina are presented. For each district, data from five actual precincts are listed to give an indication of the types of differences that exist between registration and turnout data. As we can see from the table, turnout rates vary quite a bit from precinct to precinct. In Louisiana, the mean turnout rate for whites is 67.4% (median is 68.2%, standard deviation 9.2%) while the mean turnout rate for blacks is 55.4% (median is 57.1%, standard deviation is 22.4%). In South Carolina, the numbers are similar. Mean turnout rate for whites is 64.1% (median is 65.6% and standard deviation is 10.6%) while mean turnout rate for blacks is 55.8% (median is 57.8% and standard deviation is 16.7%). In general, it appears that turnout rates in these areas for whites are more consistent and generally higher than those for blacks. We will use these raw data in our pure inverse model to compute estimates of group propensities to support certain candidates, and to compare the results from using these two different sources of data.

The differences that we observed here already appear to be the type of differences that would have a substantive impact on our results. Moreover, this impact is likely to be lopsided in two ways. First, effects on estimates of white behavior will be affected by the larger voter base. Because we must allocate each voter into the support group for one candidate, and there is uncertainty with each allocation, the more voters we must allocate, the greater the effect that the differences between the data sets is likely to have. Second, to the extent that black turnout rates are lower than white turnout rates, there is more measurement error for the black data than the white data.

Table 2: Difference between Registration and Turnout Numbers (Five Example Precincts)

LA–1	White			Black		
	Registration	Turnout	Rate	Registration	Turnout	Rate
1	1319	1065	80.7%	5	3	60.0%
2	759	615	81.0%	2	2	100.0%
3	741	318	42.9%	899	383	42.6%
4	663	425	64.1%	280	172	61.4%
5	26	13	50.0%	701	418	59.6%
SC–1	White			Black		
	Registration	Turnout	Rate	Registration	Turnout	Rate
1	1027	617	60.1%	362	234	64.6%
2	1404	938	66.8%	458	279	60.9%
3	2975	1695	57.0%	165	105	63.6%
4	1202	729	60.6%	26	7	26.9%
5	266	166	62.4%	1232	778	63.1%

Table 3: Actual Estimates of Support for Republican Candidate using Registration and Turnout Data

LA–1	White		Black	
	Registration	Turnout	Registration	Turnout
1	0.5997	0.7522	0.1748	0.1713
2	0.6221	0.7727	0.1699	0.1703
3	0.1283	0.3040	0.0996	0.3134
4	0.2160	0.3937	0.2734	0.3073
5	0.1644	0.1653	0.0314	0.0590
SC–1	White		Black	
	Registration	Turnout	Registration	Turnout
1	0.3467	0.6108	0.3303	0.4171
2	0.3827	0.6228	0.3427	0.3849
3	0.3733	0.6111	0.2590	0.2330
4	0.2781	0.4596	0.1811	0.1702
5	0.2405	0.1860	0.1707	0.1686

5.1 X's Measured by Voter Registration and Voter Turnout by Race

As we see from Table 2, registration data are not particularly good approximations of turnout data for this particular election. Turnout rates can be relatively high, but they can also be quite low. A casual perusal of these data, then, already suggests that using registration data rather than turnout data will have quite a substantial impact on the resulting voter response estimates. Some indication of what this impact can be in certain cases is demonstrated in Table 3, where results from the same five precincts listed in Table 2 are reported. The numbers indicate the support rate (as estimated from our information-theoretic technique) for the Republican candidate in the first congressional district in Louisiana and South Carolina. So, in precinct "1" in Louisiana, if we use registration data and our pure inverse model, we would estimate that 60% of whites supported the Republican candidate while 17% of blacks supported the Republican candidates. However, if we use the turnout data, our estimates would change substantially, and we would estimate that over 75% of whites supported the Republican candidate while black support remains near 17%.

5.1.1 Conditional Probability Voter Response Estimates

Consider now the results presented in Table 4 where the combined results from every precinct in the first Congressional district in Louisiana and the first Congressional district in South Carolina in the 2000 elections are listed. (A complete listing of all 2000 congressional districts for Louisiana and South Carolina is available in Appendix A). The numbers in Table 4 represent the difference between the voter response estimate obtained from the registration data and the estimate derived from the turnout data. Using our pure inverse model, one estimate is obtained for each precinct. The number reported in the table is the mean of the differences for the entire district. The numbers in parentheses are the standard deviations for the differences across the precincts. A large standard deviation thus implies that there is wide variance in the differences that we observe from precinct to precinct.

In general, it appears that the measurement error problem has a bigger impact on support estimates for the white group than for the black group or the "other" group. This pattern is consistent with the one we might expect from the numbers we observed in Table 2. It appears, in these districts, that there are generally more whites than blacks, and so the difference between the registration and turnout numbers and proportions are larger on a pure scale, resulting in greater uncertainty in how to allocate the larger number of voters into the proper candidate camps. In addition, if the black vote is largely homogeneous and blacks reside in precincts where support for a candidate is especially low or especially high, then there is again more certainty in allocating the black vote than the white vote. In addition, perhaps a "neighborhood effect" might be at play where support rates for a certain candidate tend to

Table 4: Difference between Estimates Obtained Using Registration versus Turnout Proportions

		Registration Estimate − Turnout Estimate					
District		Republican	Democrat 1	Democrat 2	Independent 1	Independent 2	Abstention
LA–1	Overall	-0.0648	-0.0140	-0.0122	-0.0142	-0.0145	0.1198
		(0.1123)	(0.0366)	(0.0222)	(0.0232)	(0.0254)	(0.1290)
	White	-0.2204	-0.0450	-0.0127	-0.0024	-0.0016	0.2821
		(0.0550)	(0.0375)	(0.0105)	(0.0036)	(0.0024)	(0.0791)
	Black	0.0056	-0.0018	-0.0121	-0.0153	-0.0137	0.0373
		(0.0317)	(0.0341)	(0.0253)	(0.0226)	(0.0246)	(0.0704)
	Other	0.0242	0.0121	-0.0103	-0.0348	-0.0332	0.0420)
		(0.0160)	(0.0112)	(0.0247)	(0.0286)	(0.0325)	(0.0280)
District		Republican	Democrat	Libertarian	Natural Law	Reform	Abstention
SC–1	Overall	-0.0669	-0.0523	-0.0106	-0.0103	-0.0097	0.1498
		(0.1011)	(0.0889)	(0.0249)	(0.0293)	(0.0250)	(0.1578)
	White	-0.1970	-0.1199	-0.0117	-0.0029	-0.0015	0.3331
		(0.0707)	(0.0754)	(0.0191)	(0.0206)	(0.0131)	(0.0926)
	Black	-0.0285	-0.0470	-0.0207	-0.0115	-0.0147	0.1224
		(0.0551)	(0.0848)	(0.0295)	(0.0309)	(0.0314)	(0.1174)
	Other	0.0064	0.0093	-0.0009	-0.0184	-0.0135	0.0171
		(0.0211)	(0.0205)	(0.0246)	(0.0384)	(0.0381)	(0.0236)

be high among all voters in that precinct regardless of race. This effect may be due to other factors (e.g. socioeconomic characteristics) that may link the voters from various racial groups. Race is often a large factor, but few would argue that it is the only factor, and whites who live in predominantly black neighborhoods do not uncommonly exhibit preferences that are similar, at least in some ways, to blacks. All of these factors are consistent with the larger variance and greater mean difference that we observe for the white vote.

The other striking pattern in Table 4 is seen in the large differences that characterize the Abstention category. Indeed, these numbers are generally the largest within a group. While these differences are quite large and the magnitude of the numbers is somewhat surprising, the pattern, again, is not particularly surprising given our substantive priors. After all, if one has turnout data, the abstention rate is known. However, if one has access to the registration data only, then the abstention rate must be estimated or assumed to be much larger than it is. One is prone to assume that many people "abstained" from voting in every race while the truth is that these people did not even enter the polling booth. Hence, it is a foregone conclusion that "abstention rates" will be much higher if one is using registration data rather than turnout data. Because there is a fixed number of voters to allocate into a fixed number of categories, more voters in the abstention category necessarily means fewer supporters available for each of the candidates. In other words, the high abstention rates lower the support rates for the candidates.

These results are, of course, specific to these two data sets. However, the general lessons that we derive are more broadly applicable. For instance, we note in these data that the measurement error problem poses more difficulties for estimating white voting tendencies. However, this may not be the case for every Voting Rights claim. The black estimates benefit from a number of factors that are more representative of the Southern region of the US. In particular, quite a benefit seems to be derived from the relatively homogeneous geographic patterns in this area. In states that are more geographically diverse, such a benefit will not be realized. The rate of abstention will also be a component of the quality of estimates. These factors, along with some others previously mentioned, will determine the extent of the measurement error impact in any given instance.

5.2 An Artifact of Modeling?

As we expected just from a casual perusal of the data, and as borne out by the data analysis, the difference between the estimates obtained from the various data sources is considerable. For these data, these differences are especially notable for the white group in the Republican and Abstention categories. However, this same pattern is also prominent among the other electoral groups. Moreover,

Table 5: Difference between Estimates Obtained Using Registration versus Turnout Proportions (OLS and EI)

		Registration	Turnout
LA–1	OLS Estimates		
	White	49.2%	69.8%
	Non-white	-2.7%	5.3%
	EI Estimates		
	White	48.1%	69.6%
	Non-white	2.7%	6.7%
SC–1	OLS Estimates		
	White	40.2%	60.2%
	Non-white	-0.01%	3.1%
	EI Estimates		
	White	39.8%	59.6%
	Non-white	0.7%	5.4%

these differences are of such a magnitude that the substantive impact is indisputable. In both Louisiana's and South Carolina's first congressional district, if one had the racial registration numbers/proportions only, white support rates for the Republican candidate would be underestimated by about 20%. In a Voting Rights case, such a range could easily swing the decision to a finding of no racial polarization while the truth may be quite the contrary.

It is important to note that the discrepancy between the registration and turnout numbers here is not an artifact of the information theoretic model that we employ. When one endeavors to make ecological inferences, the choice of model and estimator certainty has some influence on voter response estimates. Here, however, the measurement error problems in the data sources clearly play a large part in driving the results, *and* is independent of the estimation method. In this sense, we do not and have no need to visit the dispute concerning how to best make individual-level inferences from aggregate data (see e.g., Cho 1998 and Herron and Shotts 2002). If the error-in-the-equations statistical model suggested by Goodman could be assumed, the error in the $\mathbf{X}$ variable would mean in terms of the least squares estimator that the assumed orthogonality between the observed $\mathbf{X}^*$, and the equation error would not hold. Consequently, not only would the measurement error problems noted in Section 5 hold, but in addition, the least squares estimator would be inconsistent and asymptotically biased. In addition, the size of the standard deviations implies that the difference in voter responses among precincts can be quite large and thus places emphasis for decision purposes on disaggregated district data.

The results from employing Goodman's LS regression and King's EI model on the data is displayed in Table reftable:dolsei. As we see, the discrepancies remain despite the change in estimator, thus strongly implicating the measurement error as the source of the problems. Confirming previous research, the

estimates from LS and EI are similar to one another as one might expect (Cho and Yoon 2001, Anselin and Cho 2002). Some effects are noticeable between the estimates from our information theoretic model and those from either LS or EI, but the contribution to the difference in estimates from the different sources of data (turnout versus registration versus VAP, etc.) is much larger and indisputable. Notably, we see the same pattern for these two alternative estimators as we saw with our pure inverse model. In particular, the magnitude of the differences in the estimates for whites are greater, and the size of the effect is roughly the same.[1]

Finally, we emphasize that the model we employ is able to provide estimates for $r \times c$ contingency tables while some of the EI software packages are limited to 2×2 cases. This is an important distinction because while vote decisions are sometimes limited to two choices and made by only two groups in the electorate, this is often not the case. Because of this restriction with the EI software, researchers who employ EI often make major assumptions to transform their substantive problem into a 2×2 problem that is tractable for the software. Indeed, they may (though not for theoretical or substantive reasons) dismiss entirely "Other" voters and the "Abstention" category or employ a two-stage procedure to estimate abstention rates and then use these estimates in a second stage estimation (for an example of both, see Burden and Kimball 1998). These assumptions have major substantive consequences and are not warranted by theory, but by the software. In this sense, it is not possible to compare directly the results from our model to those from EI. Importantly, however, such a comparison is neither necessary nor helpful for our main argument. Nonetheless, what comparisons can be made indicate that the estimator choice is inconsequential. The large difference in estimates remain and the measurement error problem is of sufficient magnitude that it eclipses the discrepancies that we may encounter from employing different estimating procedures.

6 Concluding Remarks and Future Direction

Across the social sciences, because of incomplete theories and non-experimental data, the functional relationships pursued contain errors in the equations and errors in the variables. In our context, because of the secret ballot, racial turnout data is desired but unobservable, and so researchers usually have to work with registration or census data. Thus, it is not surprising, in ecological inference, that the quality of the aggregate data does matter. As we have attempted to document in this study, there are important and vast consequences associated with data availability that limit progress in the ecological estimation and inference area. In terms of importance relative to the quality of conditional probability

[1] As previously noted, the Goodman LS estimates are available for the entire district only (the model assumes that the behavior across precincts is the same), while our information theoretic model provides an estimate for each individual precinct.

estimates and the corresponding substantive implications, the measurement error issue dominates the debate over choice of statistical model and estimators, a debate that has occupied center stage over the years.

Clearly, in ecological inference applications in the Voting Rights arena, the independent variable of interest is turnout data. That is, we would like to determine how the actual voters (minority and otherwise) in a given election behaved. However, data on voter turnout are rarely available by race. In a handful of southern states, voter registration data by race is available. It would appear that using registration data instead of turnout data, is problematic insofar as the registration data are imperfect proxies. For instance, if the proportion of voters who vote for the office in question is not identical, and more commonly, these proportions are not identical, then using these data creates a measurement error problem that may well dominate the results from traditional ecological inference statistical models and estimators.

As we have demonstrated in our empirical examples above, the magnitude of the measurement error problem can be great, in general, and is likely to have an substantial impact on Voting Rights cases, in particular. In Louisiana's first congressional district, the estimates of white support for the Republican candidate differs by 0.21 depending on the data choice (or perhaps data availability). A difference of such a magnitude could be pivotal in the outcome of a case where these estimates are the sole determinants used by a judge in determining whether racially polarized voting exists. The effect touches individual cases as well as having broader implications for minority voting rights and representation.

In order to mitigate the problems that arise from using registration data rather than turnout data as our independent variable, it would be helpful to understand how the transition between these two entities occurs. Developing a framework that leads to the identification of variables that determine or condition who votes is our next goal. We plan to use the additional information gleaned from the empirical research in the voter turnout literature and identify them in the form of instrumental variables. This source of information incorporates the large voter turnout literature that examines the variables that condition the transition from registered voters to voters who turn out to vote. To reflect this potential heterogeneity in the micro behavior, we assume that the β_{ijk}'s are conditional on a set of explanatory-instrumental variables, and that these covariates reflect the individual, spatial, or temporal differences in voter decisions. As such, the instrumental variable (IV) approach provides a method for estimating causal effects in a measurement error or simultaneous equation model framework. Using this information, along with the observed macro data, it is possible to form a set of estimating equations as a basis for recovering the unknown conditional probabilities and identifying the impact of the explanatory variables on the corresponding conditional probabilities. The ultimate success of

the moment-based specification depends on a plausible theory of micro voter behavior that helps to identify the important behavior conditioning factors. If a reliable procedure can be developed for making ecological inferences when measurement errors exist in the data, then we could relieve the courts from their oversimplified decision making, and voting rights laws could be much more effective and efficient.

References

Achen, C. H. and W. P. Shively (1995). *Cross-Level Inference*. University of Chicago Press.

Anselin, L. and W. K. T. Cho (2002, Summer). Spatial effects and ecological inference. *Political Analysis 10*(3), 276–297.

Baggerly, K. (1998). Empirical likelihood as a goodness of fit measure. *Biometrika 85*(3), 535–47.

Cameron, C., D. Epstein, and S. O'Halloran (1996, December). Do majority-minority districts maximize substantive black representation in congress? *American Political Science Review 90*(4), 794–812.

Cho, W. K. T. (1998). Iff the assumption fits:... a comment on the king ecological inference solution. *Political Analysis 7*, 143–163.

Cressie, N. and T. R. Read (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B 46*, 440–64.

Epstein, D. and S. O'Halloran (1999, April). Measuring the electoral and policy impact of majority-minority voting districts. *American Journal of Political Science 43*(2), 367–395.

Gokhale, D. and S. Kullback (1978). *The Information in Contingency Tables*. Marcel Dekker.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review 18*, 663–669.

Goodman, L. A. (1959, May). Some alternatives to ecological correlation. *American Journal of Sociology 64*(6), 610–625.

Grofman, B., L. Handley, and R. G. Niemi (1992). *Minority Representation and the Quest for Voting Equality*. Cambridge University Press.

Grofman, B. and M. Migalski (1988). Estimating the extent of racially polarized voting in multicandidate elections. *Sociological Methods and Research 16*(4), 427–454.

Grofman, B. N. and N. Noviello (1985, June). Jai-alai outcomes as a function of player position and player skill level. *Simulation and Games 16*(2), 211–223.

Herron, M. C. and K. W. Shotts (2003, Autumn). Using ecological inference point estimates in second-stage linear regressions. *Political Analysis 11*(3), 44–64.

Hill, K. (1995). Does the creation of majority black districts aid republicans? *Journal of Politics 57*, 384–401.

Jaynes, E. (1957). Information theory and statistical mechanics i. *Physics Review 106*, 620–630.

Judge, G. G., D. J. Miller, and W. K. T. Cho (Forthcoming). *An Information Theoretic Approach to Ecological Estimation and Inference*.

Kimball, B., B. Grofman, and L. Handley (1987). Does redistricting aimed to help blacks necessarily help republicans? *Journal of Politics 49*, 169–185.

King, G. (1997). *Reconstructing Individual Behavior from Aggregate Data: A Solution to the Ecological Inference Problem*. Princeton Uiversity Press.

Kousser, J. M. (1973, Autumn). Ecological regression and the analysis of past politics. *The Journal of Interdisciplinary History 4*(2), 237–262.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley and Sons.

Lublin, D. (1997). *The Paradox of Representation*. Princeton, NJ: Princeton University Press.

Mittelhammer, R., G. G. Judge, and D. J. Miller (2000). *Econometric Foundations*. Cambridge University Press.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics 18*, 90–120.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*, 237–49.

R.C.Read, T. and N. Cressie (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 351–357.

Shotts, K. W. (2002). Gerrymandering, legislative composition, and national policy outcomes. *American Journal of Political Science 46*, 398–414.

Swain, C. (1993). *Black Faces, Black Interest: The Representation of African Americans in Congress.* Cambridge: Harvard University Press.

Thernstrom, S. and A. Thernstrom (1997). *American in Black and White: One Nation, Indivisible.* New York: Simon and Schuster.

Appendix

Table A–1. Registration Estimate – Turnout Estimate (Louisiana).

District		Republican	Democrat1	Democrat2	Indep1	Indep2	Indep3	Abstention
LA–1	Overall	-0.0648	-0.0140	-0.0122	-0.0142	-0.0145		0.1198
		(0.1123)	(0.0366)	(0.0222)	(0.0232)	(0.0254)		(0.1290)
	White	-0.2204	-0.0450	-0.0127	-0.0024	-0.0016		0.2821
		(0.0550)	(0.0375)	(0.0105)	(0.0036)	(0.0024)		(0.0791)
	Black	0.0056	-0.0018	-0.0121	-0.0153	-0.0137		0.0373
		(0.0317)	(0.0341)	(0.0253)	(0.0226)	(0.0246)		(0.0704)
	Other	0.0242	0.0121	-0.0103	-0.0348	-0.0332		0.0420)
		(0.0160)	(0.0112)	(0.0247)	(0.0286)	(0.0325)		(0.0280)
LA–3	Overall	-0.0679			-0.0164	-0.0153	-0.0140	0.1136
		(0.1036)			(0.0204)	(0.0192)	(0.0176)	(0.1199)
	White	-0.1692			-0.0201	-0.0166	-0.0168	0.2227
		(0.0799)			(0.0115)	(0.0126)	(0.0096)	(0.0892)
	Black	-0.0357			-0.0235	-0.0179	-0.0192	0.0963
		(0.0720)			(0.0235)	(0.0174)	(0.0179)	(0.1050)
	Other	0.0034			-0.0031	-0.0052	-0.0039	0.0087
		(0.0049)			(0.0056)	(0.0089)	(0.0059)	(0.0143)
LA–4	Overall	-0.0703	-0.0311	-0.0163		-0.0162		0.1339
		(0.1078)	(0.0580)	(0.0258)		(0.0271)		(0.1349)
	White	-0.1809	-0.0545	-0.0062		-0.0069		0.2485
		(0.0900)	(0.0350)	(0.0109)		(0.0124)		(0.1046)
	Black	-0.0277	-0.0368	-0.0247		-0.0248		0.1139
		(0.0705)	(0.0689)	(0.0275)		(0.0309)		(0.1183)
	Other	0.0151	0.0087	-0.0226		-0.0263		0.0250
		(0.0166)	(0.0112)	(0.0262)		(0.0327)		(0.0249)
LA–5	Overall	-0.0673	-0.0303		-0.0150	-0.0139		0.1266
		(0.1034)	(0.0541)		(0.0236)	(0.0237)		(0.1406)
	White	-0.1748	-0.0641		-0.0081	-0.0080		0.2550
		(0.0762)	(0.0352)		(0.0101)	(0.0087)		(0.0970)
	Black	-0.0232	-0.0361		-0.0236	-0.0198		0.1027
		(0.0585)	(0.0648)		(0.0272)	(0.0260)		(0.1133)
	Other	0.0037	0.0031		-0.0040	-0.0088		0.0060
		(0.0141)	(0.0145)		(0.0208)	(0.0325)		(0.0154)
LA–6	Overall	-0.0651	-0.0428		-0.0172			0.1252
		(0.0897)	(0.0668)		(0.0345)			(0.1310)
	White	-0.1271	-0.0569		-0.0276			0.2116
		(0.1115)	(0.0611)		(0.0387)			(0.1406)
	Black	-0.0765	-0.1285		-0.0165			0.2215
		(0.0652)	(0.1113)		(0.0228)			(0.1578)
	Other	0.0055	0.0061		-0.0438			0.0321
		(0.0132)	(0.0161)		(0.0368)			(0.0192)
LA–7	Overall		-0.0844		-0.0234			0.1077
			(0.1074)		(0.0223)			(0.1213)
	White		-0.2082		-0.0420			0.2502
			(0.0782)		(0.0208)			(0.0833)
	Black		-0.0497		-0.0209			0.0706
			(0.0987)		(0.0239)			(0.1086)
	Other		-0.0006		-0.0044			0.0050
			(0.0067)		(0.0092)			(0.0147)

Table A–2. Registration Estimate – Turnout Estimate (South Carolina).

District		Republican	Democrat	Libertarian	Natural Law	Reform	Constitution	United Citizen	Abstention
SC–1	Overall	-0.0669	-0.0523	-0.0106	-0.0103	-0.0097			0.1498
		(0.1011)	(0.0889)	(0.0249)	(0.0293)	(0.0250)			(0.1578)
	White	-0.1970	-0.1199	-0.0117	-0.0029	-0.0015			0.3331
		(0.0707)	(0.0754)	(0.0191)	(0.0206)	(0.0131)			(0.0926)
	Black	-0.0285	-0.0470	-0.0207	-0.0115	-0.0147			0.1224
		(0.0551)	(0.0848)	(0.0295)	(0.0309)	(0.0314)			(0.1174)
	Other	0.0064	0.0093	-0.0009	-0.0184	-0.0135			0.0171
		(0.0211)	(0.0205)	(0.0246)	(0.0384)	(0.0381)			(0.0236)
SC–2	Overall	-0.0670	-0.0682	-0.0102	-0.0116				0.1569
		(0.0870)	(0.0955)	(0.0211)	(0.0212)				(0.1548)
	White	-0.1586	-0.1283	-0.0034	-0.0022				0.2926
		(0.0650)	(0.0732)	(0.0068)	(0.0042)				(0.1117)
	Black	-0.0443	-0.0855	-0.0136	-0.0128				0.1561
		(0.0557)	(0.1190)	(0.0213)	(0.0176)				(0.1515)
	Other	0.0054	0.0054	-0.0096	-0.0121				0.0110
		(0.0099)	(0.0100)	(0.0221)	(0.0252)				(0.0149)
SC–3	Overall	-0.0808	-0.0500	-0.0095	-0.0068			-0.0101	0.1571
		(0.1144)	(0.0850)	(0.0219)	(0.0225)			(0.0230)	(0.1627)
	White	-0.2248	-0.1160	-0.0040	-0.0007			-0.0030	0.3486
		(0.0657)	(0.0645)	(0.0071)	(0.0078)			(0.0096)	(0.0837)
	Black	-0.0328	-0.0496	-0.0228	-0.0130			-0.0201	0.1383
		(0.0615)	(0.0894)	(0.0259)	(0.0260)			(0.0283)	(0.1215)
	Other	0.0038	0.0034	-0.0042	-0.0047			-0.0046	0.0063
		(0.0083)	(0.0079)	(0.0119)	(0.0232)			(0.0161)	(0.0094)
SC–4	Overall	-0.0693	-0.0090	-0.0085	-0.0081	-0.0117	-0.0103	0.1169	
		(0.1093)	0.0191	(0.0179)	(0.0182)	(0.0234)	(0.0189)	(0.1266)	
	White	-0.1945	-0.0258	-0.0024	-0.0018	-0.0341	-0.0076	(0.2662	
		(0.0719)	0.0149	(0.0070)	(0.0061)	(0.0182)	(0.0100)	(0.0906)	
	Black	-0.0084	-0.0070	-0.0168	-0.0170	-0.0067	-0.0184	0.0743	
		(0.0559)	0.0222	(0.0248)	(0.0273)	(0.0241)	(0.0266)	(0.0845)	
	Other	0.0043	0.0017	-0.0057	-0.0060	0.0025	-0.0035	0.0067	
		(0.0066)	0.0049	(0.0125)	(0.0112)	(0.0050)	(0.0091)	(0.0083)	
SC–5	Overall	-0.0613	-0.1088	-0.0145					0.1847
		(0.0720)	(0.1320)	(0.0265)					(0.1768)
	White	-0.1441	-0.2286	-0.0043					0.3770
		(0.0693)	(0.0946)	(0.0073)					(0.0921)
	Black	-0.0508	-0.0962	-0.0282					0.1753
		(0.0495)	(0.1179)	(0.0308)					(0.1398)
	Other	-0.0004	0.0000	-0.0063					0.0066
		(0.0248)	(0.0250)	(0.0393)					(0.0252)
SC–6	Overall	-0.0476	-0.1249	-0.0092	-0.0113				0.1930
		(0.0612)	(0.1498)	(0.0239)	(0.0262)				(0.1739)
	White	-0.0864	-0.1284	-0.0094	-0.0073				0.2315
		(0.0515)	(0.0873)	(0.0205)	(0.0135)				(0.0982)
	Black	-0.0566	-0.2359	-0.0057	-0.0073				0.3055
		(0.0369)	(0.1436)	(0.0102)	(0.0139)				(0.1278)
	Other	0.0053	0.0070	-0.0081	-0.0143				0.0100
		(0.0198)	(0.0209)	(0.0306)	(0.0396)				(0.0218)