Spatial and Contextual Models in Political Research

Edited by

Munroe Eagles

National Center for Geographic Information and Analysis,
State University of New York, Buffalo, New York

7

New methods for valid ecological inference

Bernard Grofman

Introduction

There are numerous circumstances, such as those involving the analysis of historical data, where political geographers and other social scientists must seek to understand voting (or other types of individual behavior) in situations where reliable surveys at the individual level are not available.[1] In coping with the absence of survey data, social scientists have developed a range of tools for making inferences using aggregate data based on geographical or other units.[2] Much early work made use of thematic maps to represent visually relationships between the characteristics of geographic units and the behavior of those who lived in them; this methodology remains important because it provides insights about spatial patterns (see, e.g., Rokkan and Valen, 1970; Ryssevik, 1990). Work prior to World War II focused on maps and/or on simple correlational or cross-tabular analyses (see, e.g., Durkheim, 1897; Siegfried, 1913); a review of bivariate correlations remains a starting point for many contemporary analysts (see, e.g., Ryssevik, 1990).

Most recent work on ecological data derives from two now-classic approaches, the ecological regression approach of Goodman (1953) and the method of bounds of Duncan and Davis (1953). These approaches were developed in part in response to Robinson's (1950) classic essay showing that changes in the magnitude or even the direction of the correlations between variables could occur, depending upon exactly how data were grouped into ecological units, and warning of the risks of using correlations among variables measured at some level of aggregation as evidence for what was true for individuals.

We discuss two important extensions of the Goodman approach to the analysis of voting behavior: one, the so-called double-equation approach (Loewen, 1982; Grofman *et al.*, 1985; Grofman and Migalski, 1988), designed to cope with the problem caused by the absence of data on the composition of the actual (as opposed to the potential) electorate; and one, a context effects model attributable to Boudon (1963), designed to compensate for the problem of excluded variables by positing linear contextual effects that give rise to a quadratic relationship between group population proportion and the vote. In addition, following the lead of authors such as Shively (1969, 1975) in drawing on information external to the model to help us decide what is plausible, we

consider an extension of the method of bounds proposed by Duncan and Davis (1953).[3]

There are eight important distinctions we wish to draw, the importance of which sometimes fail to be appreciated: We wish to distinguish (1) between aggregate data analyses that can properly be labeled as involving ecological inference and those that cannot;[4] (2) between models that are intended simply as descriptions of patterns of relationships and those that purport to 'explain' the relationships in causal terms;[5] (3) between cross-sectional applications of ecological inference and cross-temporal applications;[6] (4) between ecological regressions involving proportions where the dependent and independent variables can be thought of as having the same denominator, and those where they cannot;[7] (5) between bivariate and multivariate approaches;[8] (6) between linear and non-linear models;[9] (7) between models that make use of information external to the model to improve estimates/set bounds on what is possible and those that make use of no exogenous information; and (8) between approaches that focus on finding upper and lower bounds on proportions to be estimated by ruling out estimates that are logically/mathematically impossible, and those that are trying to find 'best' estimates under plausible assumptions about the nature of the underlying relationships.[10]

Basic ecological regression

When is a regression on ecological units not an ecological regression?

A regression on ecological units is one where the values of the independent variables being regressed are characteristics of some unit of aggregation (most commonly a type of geographic area, such as voting precinct, city, county or state).[11] Under this definition, many regressions commonly done by political geographers and other social scientists can be thought of as ecological in nature: for example, a regression of levels of black representation in city councils on a variety of other city attributes treated as independent variables, such as form of government (ward versus at-large, strong mayor versus council/city manager, partisan versus non-partisan elections), location (south versus non-south), and population characteristics (e.g. minority population share).[12] Often, ecological data are used to show how particular characteristics of ecological units are *correlated* with some particular behaviors.[13] Sometimes the concern is merely with the directionality of the effect. In this chapter, however, we reserve the term 'ecological regression' for attempts to produce *direct* estimates of individual-level behavior (e.g. the proportion of members of a given race who vote for a candidate of that race in a particular election) from information about the voting behavior and other characteristics of aggregate-level units, such as voting precincts.[14] The classic development of basic ecological regression methodology is due to Goodman (1953, 1959), although similar

methods were used by earlier scholars (see especially the discussion of the work of the 1930s scholar Fritz Bernstein in Lohmoller and Falter, 1986).

Goodman's ecological regression method

We will illustrate the approach taken by Goodman (1953, 1959) in terms of two dichotomous and mutually exclusive groups (x and not-x), some proportion of each of whose members engages in one of two mutually exclusive and exhaustive types of (voting) behavior (v and not-v).

Let

x = the proportion of the electorate that is in group x
$1 - x$ = the proportion of the electorate that is in group not-x
P'_{v} = the proportion of the electorate that engages in behavior v
$P'_{\text{not-v}}$ = the proportion of the electorate that does not engage in behavior v ($= 1 - P'_{v}$)
$P'_{x,v}$ = the proportion of members of group x who engage in behavior v
$P'_{\text{not-x,v}}$ = the proportion of members of group not-x who engage in behavior v
$P'_{x,\text{not-v}}$ = the proportion of members of group x who do not engage in behavior v ($= 1 - P'_{x,v}$)
$P'_{\text{not-x,not-v}}$ = the proportion of members of group not-x who do not engage in behavior v ($= 1 - P'_{\text{not-x,v}}$)

Goodman's approach can usefully be thought of as beginning with a very simple 'bookkeeping' identity:

$$P'_{v} = P'_{xv}\,x + P'_{\text{not-x,v}}(1 - x) = (P'_{xv} - P'_{\text{not-x,v}})x + P'_{\text{not-x,v}} \qquad \textbf{(7.1)}$$

In other words, the proportion of the electorate that engages in behavior v equals (the proportion of the members of group x who engage in that behavior multiplied by the proportion of members of the electorate who are in group x) plus (the proportion of the members of group not-x who engage in that behavior multiplied by the proportion of members of the electorate who are in that group).

While equation 7.1 is a tautology, it applies only to the electorate as a whole. The 'trick' of ecological regression is to posit that the same relationship holds approximately across all ecological units.[15]

By rearranging terms, we may rewrite equation 7.1 as a linear equation in x, where x is now the proportion of those in a given ecological unit (e.g. voting precinct) who belong to group x.

$$P'_{v} = (P'_{xv} - P'_{\text{not-x,v}})x + P'_{\text{not-x,v}} \qquad \textbf{(7.1)}'$$

Now we have a linear relationship, $y = sx + r$, with

$$s = P'_{xv} - P'_{\text{not-x,v}}$$

$$r = P'_{\text{not-x,v}}$$

and

$$y = P'_{v}$$

We may rewrite this to solve for P'_{xv} and $P'_{\text{not-x,v}}$ to obtain

$$P'_{xv} = r + s \quad \textbf{(7.2a)}$$

$$P'_{\text{not-x,v}} = r \quad \textbf{(7.2b)}$$

Now we may plug in the values of r and s obtained from ecological regression of v on x to estimate the values of P'_{xv} and P'_{BW} as the *mean* values for those individual-level parameters – parameters that are unobservable directly, absent survey data.[16]

The reader may not immediately recognize the connection between equations 7.1 and 7.2 and the analyses in Goodman (1953, 1959) because Goodman illustrates his modeling in terms of cross-temporal voting behavior. In his example, in our notation, x is the proportion of voters for a given party (call it party D) at time t, $1 - x$ is the proportion of voters for the opposite party at time t, and v is the behavior 'vote for party D at time $t + 1$'.

Thus, the bookkeeping equality used by Goodman is that the proportion of the electorate that voted for party D at time $t + 1$ equals (the proportion of those who voted for party D at time t who also voted for party D at time $t + 1$ multiplied by the proportion of the electorate who voted for party D at time t) plus (the proportion of those who did not vote for party D at time t who switched to vote for party D at time $t + 1$ multiplied by the proportion of the electorate who did not vote for party D at time t). Hence, when we regress support for party D at time $t + 1$ on support for party D at time t, the estimate of the proportion of voters for party D at time t who continued to vote for party D subsequently is given by $r + s$, while the estimate of the proportion of voters who failed to vote party D at time t but then switched to party D in the next election is given by r.

Potential problems with the Goodman methodology

There are a number of obvious potential problems with applying the Goodman linear regression methodology to actual situations of interest. These problems suggest the usefulness of modifying this model.

First, if we apply the basic Goodman model cross-temporally, then we are implicitly assuming that the set of voters at time $t + 1$ (or perhaps even at time $t + k$) is the same as at time t. Clearly, voters die and new voters come of age; moreover, over time, there will be movement of voters across ecological units.[17]

Second, even if the application of ecological regression is to a single election, it may not be the case that all potential voters vote. When the single-equation Goodman methodology is used, if the data used to measure the size of each group are based not on the actual electorate but on the number of *potential*

voters the group contains, then differences in the turnout rates of different groups relative to that potential electorate can create measurement error and thus estimation bias.[18]

Third, the model as explicated above posits only two groups. If there are n groups ($n > 2$) whose behavior we need to estimate, then we may need to extend the bivariate approach above to one which is multivariate – in terms of treating the proportion of the ecological unit made up by members of $n - 1$ such groups as independent variables.[19]

Fourth, even if there are only two groups, it may be the case that the behavior of members of one or both of the groups varies in a systematic way with other attributes of the group's members,[20] i.e. we may view the group as being 'non-homogeneous' (Lupia and McCue, 1990). Thus, it might be the case that we can improve estimates of the behavior of group members by introducing in some fashion additional variables that are related to the behavior(s) in question, or by partitioning the group into subgroups that are more nearly homogeneous with respect to the behavior in question. This again leads us to some type of multivariate approach.

Fifth, the model as explicated above posits only two behaviors. If there are multiple behaviors we need to estimate, then again we may need to extend the bivariate approach in the section above to one which is multivariate.[21]

Sixth, the Goodman methodology may produce estimates that conflict with our a priori knowledge. For example, we know that group proportions must lie between 0 and 1. Estimates outside that range must somehow be 'corrected'.[22] Or, we may have good reason to believe that the proportion of members of group x who engage in behavior v is never greater than the proportion of members of group non-x who engage in that same behavior.

In the remainder of the chapter we shall focus on a few techniques that we believe to be particularly important in addressing the problems discussed above, beginning with the use of double-equation regression to cope with measurement errors in the independent variable caused by differences in the turnout levels of the various groups.[23] The modifications/improvements to the basic single-equation bivariate Goodman methodology that we describe in the next three sections either add dependent variables and thus require use of more than one regression equation, and/or add additional independent variables and thus require a shift from a bivariate to a multivariate approach, perhaps by including polynomial terms. We also briefly consider how to 'add' exogenously derived knowledge so as to improve estimates, drawing on ideas suggested by the method of bounds proposed by Duncan and Davis (1953).[24]

To simplify exposition, in the remainder of this chapter, except where otherwise clearly indicated, we will draw all our examples from the analysis of patterns of racial voting. In the United States, this is the domain in which the greatest practical use of aggregate-level ecological analysis is now being made – in legal challenges to election system practices on the grounds that they illegally dilute the voting strength of racial or linguistic minorities. For nearly two decades, voting rights cases have almost invariably involved testimony

where expert witnesses have used ecological regression (coupled with descriptions of voting in racially homogeneous precincts) to infer patterns of voting by race, since there are rarely reliable survey data available on voting behavior in local elections, which are the most common subject of voting rights lawsuits. Assessing the extent of racially polarized voting is arguably now the most important of the empirical questions investigated in most voting rights cases in the United States.[25]

Double-equation regression

For convenience, we initially assume that there are two mutually exclusive and exhaustive groups, which we refer to as white voters and black voters. For simplicity of discussion, we also assume below that there is a single black and a single white candidate.[26] We wish to understand what proportion of each group's votes go to a candidate (or, more generally, candidates) identified with their own group. We introduce a notation that is consistent with that used in the previous section – a notation which has been used in the literature on voting rights issues (see, e.g., Grofman *et al.*, 1985; Grofman and Migalski, 1988; Grofman, *et al.*, 1992).

Let

x = the proportion of *eligible* voters who are white[27]
$1 - x$ = the proportion of *eligible* voters who are black
P_W = the proportion of total *eligible* voters who vote for the white candidate
P_B = the proportion of total *eligible* voters who vote for the black candidate
T = the proportion of *eligible* voters who vote

The values of these four variables are, in principle, directly observed.[28]

Let

P_{WW} = the proportion of white *eligibles* who vote for the white candidate(s)
P_{BW} = the proportion of black *eligibles* who vote for the white candidate(s)
P_{BB} = the proportion of black *eligibles* who vote for the black candidate(s)
P_{WB} = the proportion of white *eligibles* who vote for the black candidate(s)

and let

P'_B = the proportion of the *electorate* that votes for the black candidate
P'_W = the proportion of the *electorate* that votes for the white candidate
P'_{WW} = the proportion of white *voters* who vote for the white candidate
P'_{BW} = the proportion of black *voters* who vote for the white candidate
P'_{BB} = the proportion of black *voters* who vote for the black candidate
P'_{WB} = the proportion of white *voters* who vote for the black candidate

It is the latter four variables whose values we are really interested in estimating. However, if we know the values of the four 'unprimed' variables, we can readily obtain their primed equivalents by using the identities below.

$$P'_{WW} = P_{WW}/(P_{WW} + P_{WB}) \quad \textbf{(7.3a)}$$

$$P'_{BB} = P_{BB}/(P_{BB} + P_{BW}) \quad \textbf{(7.3b)}$$

$$P'_{WB} = P_{WB}/(P_{WW} + P_{WB}) = 1 - P'_{WW} \quad \textbf{(7.3c)}$$

$$P'_{BW} = P_{BW}/(P_{BB} + P_{BW}) = 1 - P'_{BB} \quad \textbf{(7.3d)}$$

To estimate P_{WW} and P_{BW}, following the same logic as for equation 7.1 discussed above, we make use of the bookkeeping identity

$$P_W = x(P_{WW}) + (1 - x)(P_{BW}) \quad \textbf{(7.4a)}$$

which can be rewritten as

$$P_W = (P_{WW} - P_{BW})x + P_{BW} \quad \textbf{(7.4a)'}$$

That equation is then used as the basis for a linear regression of

$$P_W \text{ on } x,\ P_W = m_1 x + b_1$$

to obtain

$$P_{BW} = b_1$$

$$P_{WW} = m_1 + b_1$$

In like manner, we make use of the bookkeeping identity

$$P_B = x(P_{WB}) + (1 - x)(P_{BB}) \quad \textbf{(7.4b)}$$

which can be rewritten as

$$P_B = (P_{WB} - P_{BB})x + P_{BB} \quad \textbf{(7.4b)'}$$

as the basis for a linear regression of P_B on x, $P_B = m_2 x + b_2$.

Thus, we obtain

$$P_{BB} = b_2$$

$$P_{WB} = m_2 + b_2$$

Now it easy to see that, under the assumptions previously discussed

$$P'_{WW} = (m_1 + b_1)/(m_1 + b_1 + m_2 + b_2) \quad \textbf{(7.5a)}$$

$$P'_{BB} = b_1/(b_1 + b_2) \quad \textbf{(7.5b)}$$

Let

T_B = the proportion of *eligible* black voters who vote
T_W = the proportion of *eligible* white voters who vote

It should also be clear that

$$T_B = 1 - P_{BW} - P_{BB} = 1 - b_1 - b_2 \quad \textbf{(7.6a)}$$

$$T_W = 1 - P_{WW} - P_{WB} = 1 - m_1 - b_1 - m_2 - b_2 \quad \textbf{(7.6b)}$$

Thus, the two-equation approach can not only compensate for turnout differences among each group's eligible votes, but yield a direct estimate of the turnout proportions of each group's eligible electorate.

If we regress P'_B (rather than P_B) on the proportion of members of group x who are in the eligible electorate, i.e. if we use the basic single-equation Goodman model in a situation where group size is measured in terms of potential rather than actual electorate, we get what can be thought of as a problem of unequal denominators. Since

P'_B = votes for the black candidate/total votes cast

and

x = number of members of the eligible electorate who are black/total number of eligible voters

while

P_B = votes for the black candidate/total number of eligible voters

if we use the first and second of these as, respectively, dependent and independent variables in the same equation, the variable on the right-hand side of the equation will have a different denominator from the variable on the left-hand side of the equation ('total number of eligible voters' vs. 'total votes cast'); in contrast, when the second and third of these are used as the independent and dependent variables, respectively, the same denominator, 'total number of eligible voters', appears on both sides of the equation.

If it were the case that k fraction of the eligible whites voted and j fraction of the eligible blacks voted, then

$$\text{Black share of turnout} = (j(1 - x))/(j(1 - x) + kx)$$

$$= (j - jx)/(j + (k - j)x)$$

This is a non-linear function in x. As a continuous fraction, it may be approximated by a polynomial in x of order m. Only if j is very close to k is it likely that approximating black share of turnout simply by x will yield a good fit, at least if there are a substantial number of racially mixed precincts (cf. the discussion of linear approximations to quadratic context effects in Owen and Grofman, 1994 forthcoming).

Because of the non-linearity caused by turnout differences between groups, any single-equation ecological regression that uses black share of eligibles as a proxy for black share of turnout is potentially suspect. However, the problem becomes really serious only if the difference between the turnout levels of white

and black eligibles is large and if there are a substantial number of racially mixed precincts.

Consider an extreme hypothetical, where 90 percent of black voters vote for the black candidate and 90 percent of white voters vote for the white candidate, but black turnout is only 50 percent of black eligibles, while white turnout is 100 percent of eligibles. For these assumptions, if we assume a uniform distribution of black eligibles across ecological units, it can be shown (Grofman, 1993a) that if we regress the black candidate's share of the vote on the black share of the eligibles, we will estimate black vote for the black candidate as roughly 80 percent (too low) and white vote for the black candidate as 3 percent (also too low). However, even for this extreme case, the single-equation estimates are not that far from the actual values. Of course, if we were to use the double-equation approach of equations 4a′ and 4b′, using the bookkeeping identities of equations 3a–3d, for the same assumption of a uniform distribution, it can be shown (Grofman, 1993a) that we would recover the correct percentages: $P'_{WW} = 0.9$ and $P'_{BB} = 0.9$, from the fitted equations

$$P_W = m_1 x + b_1 = 0.35x + 0.1$$

and

$$P_B = m_2 x + b_2 = -0.85x + 0.9$$

Coping with non-homogeneous groups and correlated errors

As noted earlier, one potential problem with the simple two-group Goodman method is that it assumes that errors in estimating the behavior of the members of each group are not correlated with the independent variable (the group proportion in the ecological unit).[29] If they are, then heteroskedasticity will result and estimates will not be unbiased. Relatedly, when groups are non-homogeneous, it may be possible to improve estimates by taking additional variables into account or partitioning the group into more nearly homogeneous subgroups. As defined by Lupia and McCue (1990: 365), homogeneity is the assumption that 'all members of a group make similar choices', i.e. that each may be thought of as being characterized by some probability of engaging in the behavior in question, which gives rise to a binomial (or other) distribution as we 'sample' members.[30]

A variety of techniques have been suggested to deal with correlated errors and non-homogeneity of groups. One we have already considered, the double-equation technique. It compensates for the heteroskedasticity caused by the systematic measurement error in the independent variable that is related to differences in turnout. Below we discuss three others, the partitioning into

subgroups technique proposed by Lupia and McCue (1990), the Miller (1977) quadratic context effects model, and multivariate modeling with dummy control variables.

Lupia–McCue partitioning approach

If one or both of the groups with which we begin can be thought of as non-homogeneous (in the technical sense of that term used by Lupia and McCue, 1990), then it would seem to make sense to partition the non-homogeneous group(s) into subgroups for which the homogeneity assumption is more plausible, and then run a multivariate regression with subgroup (rather than group) proportions as the independent variables – recovering a group's estimated behavior as the (population) weighted average of the behavior of the subgroups which comprise that group. This is the methodology recommended by Lupia and McCue (1990).

Using a multivariate ecological regression approach that looks at homogeneous subgroups is an excellent idea in theory, but may not be feasible in practice because of multicollinearity problems, or limited numbers of observations, or difficulty in deciding exactly how to partition,[31] or simple unavailability of the data.

For example, we may believe that, say, Cubans and Mexican-Americans vote differently, and thus that we should estimate the voting patterns of each subgroup separately rather than estimating data for Hispanics as a whole. Yet this may not feasible in practice.

Statisticians who testified as expert witnesses for the defendant jurisdiction, Los Angeles County, in *Garza* v. *Los Angeles Board of Supervisors* (D. Cal. 1990), 90 CDOS 8138 (9th Cir. 1990) *cert denied* January 1990, a successful voting rights challenge by minority plaintiffs and the Department of Justice to the 1981 redistricting plan for the County, found it impossible to derive meaningful separate estimates for the voting behavior of Cubans and Mexican-Americans. In my testimony in that case as an expert witness for the US Department of Justice, I made the commonsense point that one should not expect to be able to estimate separately Cuban and Mexican-American voting behavior simply because there were not that many Cubans in Los Angeles County (only a few thousand out of well over two million Hispanics) and because Cuban population concentrations were highly correlated with those of Mexican-Americans.[32]

Similarly, we may quite reasonably believe that white voters who are registered as Democrats will vote differently from those who are registered as Republicans, and thus that we should estimate the voting patterns of each subgroup separately rather than estimating data for whites as a whole. Yet doing so may not be possible. When I was analyzing data on black and white voting patterns in North Carolina legislative elections in my testimony for minority plaintiffs in *Gingles* v. *Edmisten* 590 F. Supp. 345 (EDNC 1984),[33] I

sought to analyze separately the voting behavior of white Democrats and white Republicans. However, since the percentage of white registered voters who were Democrats was highly correlated with the percentage of blacks in the precinct,[34] I found myself unable to estimate reliably a multigroup equation.[35]

Also, even if possible, a partitioning into subgroups may not actually 'improve' estimates sufficiently to be worth the bother. For example, for the data on the 1983 Los Angeles City Council election in council district 14, which is analyzed by Lupia and McCue (1990), in which the leading candidates were an Anglo incumbent (Snyder) and an Hispanic challenger (Rodriguez), their partitioning approach yields an estimated (absolute value) difference in Hispanic and non-Hispanic support for Rodriguez of 50 percentage points, and for Snyder of 39 percentage points. Applying the simplest possible single-equation ecological regression approach to the data they report, I estimate those difference as 47 percentage points and 37 percentage points, respectively. Such minor differences between the results they report and those obtained from the standard methods do not give much to argue about. Moreover, their estimate of the degree of bloc voting is actually marginally higher than what is obtained from the standard method. But the usual complaint by statisticians against the standard methods is that they *over*estimate bloc voting.

However, there may be circumstances where partitioning may be very important. Consider estimating the behavior of Hispanic and non-Hispanic voters in a 1984 or 1988 presidential primary in which Jesse Jackson was one of the candidates. As politically knowledgeable people, we know that non-Hispanic white and non-Hispanic black voters are apt to be radically different with respect to their levels of support for Reverend Jackson. Thus, as Lupia and McCue (1990) suggest, it might be desirable to partition the set of non-Hispanic voters into white/Anglo voters and black voters, especially if black voters were not randomly distributed with respect to Hispanic and non-Hispanic voters in terms of geography.[36]

In particular, if non-Hispanic black voters were disproportionately found in large numbers in the same neighborhoods as Hispanics, then we might easily overestimate the voting support of Hispanics for Jesse Jackson.[37]

Quadratic context effects model

A second way to deal with excluded variables/non-homogeneity of groups is to posit that the factors that affect group voting behavior are highly correlated with group density, i.e. to posit some form of context effect. Here, as we show below, for a particularly simple form of context effect, we may be able to use a bivariate quadratic instead of a bivariate linear model. However, contextual models may give rise to coefficients that cannot be directly related to the parameters of group behavior for which we are trying to solve.[38]

sought to analyze separately the voting behavior of white Democrats and white Republicans. However, since the percentage of white registered voters who were Democrats was highly correlated with the percentage of blacks in the precinct,[34] I found myself unable to estimate reliably a multigroup equation.[35]

Also, even if possible, a partitioning into subgroups may not actually 'improve' estimates sufficiently to be worth the bother. For example, for the data on the 1983 Los Angeles City Council election in council district 14, which is analyzed by Lupia and McCue (1990), in which the leading candidates were an Anglo incumbent (Snyder) and an Hispanic challenger (Rodriguez), their partitioning approach yields an estimated (absolute value) difference in Hispanic and non-Hispanic support for Rodriguez of 50 percentage points, and for Snyder of 39 percentage points. Applying the simplest possible single-equation ecological regression approach to the data they report, I estimate those difference as 47 percentage points and 37 percentage points, respectively. Such minor differences between the results they report and those obtained from the standard methods do not give much to argue about. Moreover, their estimate of the degree of bloc voting is actually marginally higher than what is obtained from the standard method. But the usual complaint by statisticians against the standard methods is that they *over*estimate bloc voting.

However, there may be circumstances where partitioning may be very important. Consider estimating the behavior of Hispanic and non-Hispanic voters in a 1984 or 1988 presidential primary in which Jesse Jackson was one of the candidates. As politically knowledgeable people, we know that non-Hispanic white and non-Hispanic black voters are apt to be radically different with respect to their levels of support for Reverend Jackson. Thus, as Lupia and McCue (1990) suggest, it might be desirable to partition the set of non-Hispanic voters into white/Anglo voters and black voters, especially if black voters were not randomly distributed with respect to Hispanic and non-Hispanic voters in terms of geography.[36]

In particular, if non-Hispanic black voters were disproportionately found in large numbers in the same neighborhoods as Hispanics, then we might easily overestimate the voting support of Hispanics for Jesse Jackson.[37]

Quadratic context effects model

A second way to deal with excluded variables/non-homogeneity of groups is to posit that the factors that affect group voting behavior are highly correlated with group density, i.e. to posit some form of context effect. Here, as we show below, for a particularly simple form of context effect, we may be able to use a bivariate quadratic instead of a bivariate linear model. However, contextual models may give rise to coefficients that cannot be directly related to the parameters of group behavior for which we are trying to solve.[38]

Let us modify the bloc voting model of a previous section by considering what happens if we posit a linear context effect of the following form:[39]

$$P'_{WW} = a_1(1 - x) + c_1 \tag{7.7a}$$

$$P'_{BW} = a_2(1 - x) + c_2 \tag{7.7b}$$

Here we are positing that the extent of white or black support for white candidates is contingent on the racial composition of the precinct. It is apparent from these expressions that the greater the (absolute) value of the parameters a_1 and a_2 relative to c_1 and c_2, the greater the contextual effect. The context effect may work in one of two directions: for example, either white support for the black candidate could increase with percentage black or it could decrease.

Using our earlier notation:

$$\begin{aligned} P'_W &= (a_1(1 - x) + c_1)x + (a_2(1 - x) + c_2)(1 - x) \\ &= (a_1 + c_1)x - a_1x^2 + (a_2 + c_2) - a_2x - (a_2 + c_2)x + a_2x^2 \\ &= (a_2 - a_1)x^2 + (a_1 + c_1 - 2a_2 - c_2)x + (a_2 + c_2) \end{aligned} \tag{7.8}$$

Thus, if there is a (linear) context effect, the vote for the white candidate is a *quadratic* function of the proportion black (minority) in the electorate,[40] which we may represent as

$$y = Cx^2 + Bx + A$$

Unfortunately, unless we make further simplifying assumptions, fitting a quadratic rather than a linear model to data on voting patterns by racial composition of precincts does not improve our ability to estimate white and black voting patterns because the set of equations to derive the four needed parameters from the three fitted coefficients of the quadratic regression is underdetermined.

Sometimes, however, it may be reasonable to assume that the context effect is present for only one of the groups: for example, black voters may be assumed to have a constant probability of supporting a black candidate of choice in a biracial contest.[41] This is equivalent to assuming that $a_2 = 0$. Under this assumption,

$$P'_W = -a_1x^2 + (a_1 + c_1 - c_2)x + c_2 \tag{7.9}$$

and we may solve for the remaining coefficients of interest, a_1, c_1 and c_2, from the coefficients of the fitted quadratic

$$P'_W = Cx^2 + Bx + A$$

as

$$\begin{aligned} a_1 &= C \\ c_1 &= B - C + A \\ c_2 &= A \end{aligned} \tag{7.10}$$

For any given distribution of x, we can solve for the mean values of P'_{WW} and P'_{BW} by integrating over the formulae in equations 7.7a and 7.7b.[42]

Multivariate approach with control variables

Still a third approach is to deal with excluded variables directly, by introducing additional variables as controls in a multivariate regression in which group proportion is one of the variables. Now we would estimate the 'average' behavior of a group as the behavior that results when we substitute the values of the set of traits that is characteristic of the average member of the group. However, multivariate ecological modeling which introduces independent variables that are other than simple group proportions brings with it a host of special problems in interpreting results in terms of the group behavioral proportions we wish to estimate,[43] and increases the complexity of the models, the range of possible models, and the needs for reliable data measured at the relevant ecological level.[44]

Making use of external data and mathematical bounds

The method of constrained percentages

Information on voting precincts that are homogeneous (or nearly so) in their racial composition can provide us with reliable information about the voting behavior of the preponderant group in the precinct,[45] especially if we are prepared to take into account insights derived from political science to improve the plausibility of our assumptions about voting behavior. Consider the ith voting precinct, denoted with a superscripted (i). Let x again refer to the actual electorate.

If we posit that $0.5 \leq P'^{(i)}_{W} \leq x^{(i)}$, then it must be the case that

$$P'_{WW} \leq (P'^{(i)}_{W} - (1 - x^{(i)}))/x^{(i)}$$

since, even if all black voters voted for the white candidate, that would still leave a remaining vote proportion of $P'^{(i)}_{W} - (1 - x^{(i)})$ which has to come from white voters. Similarly, if we posit that $0.5 \leq P'^{(i)}_{B} \leq 1 - x^{(i)}$, then it must be the case that

$$P'_{BB} \leq (P'^{(i)}_{B} - x^{(i)})/(1 - x^{(i)})$$

The results above are minimal bounds determined by the Duncan–Davis method of overlapping percentages (Duncan and Davis, 1953). The greater the level of political cohesion (similarity of voting) in the precinct, the tighter are the Duncan–Davis bounds we can set on electoral behavior; and the more nearly homogeneous the precinct, the clearer will be the evidence for racial bloc voting (or the lack thereof) within it.

For example, if $P'^{(i)}_B = 0.7$ and $1 - x = 0.7$, then we know that $P'_{BB} \geq 0.57$; while if $P'^{(i)}_B = 0.7$ and $1 - x = 0.8$, then we know that $P'_{BB} \geq 0.62$; and if $P'^{(i)}_B = 0.7$ and $1 - x = 0.9$, then we know that $P'_{BB} \geq 0.67$. Similarly, if $P'^{(i)}_B = 0.9$ rather than 0.7 and again $1 - x = 0.7$, then we know that $P'_{BB} \geq 0.85$; while if $P'^{(i)}_B = 0.9$ and $1 - x = 0.8$, then we know that $P'_{BB} \geq 0.87$; and if $P'^{(i)}_B = 0.9$ and $1 - x = 0.9$, then we know that $P'_{BB} \geq 0.88$. As the precinct becomes more exclusively of one racial group, of course, then the voting behavior of the precinct as a whole becomes more nearly the voting behavior of that group; similarly, the more nearly everyone in a precinct votes the same way, the more can we be confident that that behavior is characteristic of both (all) groups/racial groups in the precinct.

Useful as is the method of overlapping percentages, it is important to recognize that it provides a conservative estimate of the amount of racial bloc voting, since, in most circumstances, it would be unreasonable to posit (as we did above) that, in an interracial contest, a full 100 percent of the white voters in a precinct vote for the black candidate and only $100(P'^{(i)}_W - (1 - x^{(i)}))/x^{(i)}$ percent of the black voters support that candidate over his/her white opponent. Clearly, it is more reasonable to posit that, *ceteris paribus*, in an interracial contest, the proportion of voters of a given race who vote for candidates of the same race should, *ceteris paribus*, be at least as great as the proportion of voters of the opposite race who so vote.[46]

This latter assumption may be represented as

$$P'_{BB} \geq P'_{WB} \tag{7.11a}$$

$$P'_{WW} \geq P'_{BW} \tag{7.11b}$$

Given equation 7.11a if $0.5 < P'^{(i)}_B < 1 - x^{(i)}$, we can improve on the method of overlapping percentages and calculate the minimal possible value of P'_{BB} as

$$P'_B \geq P'_{BB} \tag{7.12a}$$

Similarly, given equation 7.11b, if $0.5 < P'^{(i)}_W < x^{(i)}$, we can improve on the method of overlapping percentages and calculate the minimal possible value of P'_{WW} as

$$P'_W \geq P'_{WW}{}^{47} \tag{7.12b}$$

We will refer to the bounds given in equations 7.12a and 7.12b as the method of 'simple percentages' (Loewen and Grofman, 1989).[48] In other words, if we assume that blacks (whites) are at least as likely as whites (blacks) to vote for a black (white) candidate, regardless of whether that is a very high or a very low probability, then, in reasonably homogeneous precincts, we can take behavior of the precinct as a whole as a lower bound for our estimate of the choices being made by the preponderant group in the precinct. Thus, if a precinct that is 70 percent black gives 70 percent of its vote to the black candidate, we would posit that at least 70 percent of the black voters supported the black candidate. In contrast, using the Davis–Duncan approach,

Table 7.1 Estimated raw counts with one negative cell

	Number of votes for the black candidate	Number of votes for the white candidate
Number of black voters 400	N_{BB} 416	N_{BW} −16
Number of white voters 600	N_{WB} 84	N_{WW} 516

we could only say that at least 57 percent of the black vote in that precinct went to the black candidate.

Estimates of proportions outside the feasible 0–1 range

In the racial bloc voting context, Alan Lichtman (personal communication, 1990) has proposed a simple way to cope with estimates that are outside the [0, 1] range. We set those values to 0 or 1, and estimate how many votes we need to account for. Then we shift raw votes accordingly in the cells of a table of estimated raw votes, in such a fashion as to preserve row and column marginals. Thus if we estimated P'_{BB} as 104 percent, and we also estimated that there were a total of 400 black voters and 600 white voters, then this would mean that we were assigning the black candidate 416 votes from black voters (see Table 7.1).

To correct for the estimates over 100 percent and below zero, we would add 16 votes for the black candidate to white voters, raising N_{WB} from 84 to 100 (and thus raising P'_{WB} from 0.14 to 0.16), and to compensate for the marginals, we would also subtract 16 votes from white support for the white candidate, thus lowering N_{WW} from 516 to 500 (and, correspondingly, lowering P'_{WW} from 0.86 to 0.83) (see Table 7.2).

Of course, if cell values are only slightly outside the feasible range, the effect of controlling for marginals is trivial, and thus we might simply set values to 0 or 1 and not bother to make further adjustments in cell values.[49]

Table 7.2 Corrected estimated raw counts with no negative cells

	Number of votes for the black candidate	Number of votes for the white candidate
Number of black voters 400	N_{BB} 400	N_{BW} 0
Number of white voters 600	N_{WB} 100	N_{WW} 500

Reliability of ecological inference

In my view, there are conditions under which valid ecological inference is possible, especially when various modifications to the basic Goodman approach are used, and when the Goodman approach is combined with ideas inspired by the work of Duncan and Davis (1953) and with applications of external knowledge about plausible constraints on relationships among parameters, such as those made use of by Shively (1969) and Claggett and Van Wingen (1993). For example, we have shown how the double-equation linear regression method developed for use in voting rights litigation (Grofman *et al.*, 1985) can improve the estimates of bloc voting in situations where data on the racial proportions in the *actual* electorate is unknown, but where data on the racial composition of the *eligible* electorate within each voting unit is known. We have also shown how data on homogeneous precincts may be used to provide supplementary information about the extent of polarized voting, if we make the plausible assumption that a group's level of support for a candidate of choice who is a member of the group is at least as high as the support levels for that candidate of those who are not in the group. And we have shown that, at least in the racial bloc voting context, estimates outside the [0, 1] range, if not too far outside that range, may be dealt with in a relatively straightforward way.

Our discussion is relevant to the general debate among social scientists and statisticians about the circumstances under which reliable inference about individual-level behavior is possible using methods of aggregate data analysis. Because I have elsewhere written extensively about this topic,[50] here I focus on three points. First, cross-sectional applications of ecological inference are inherently less problematic than cross-temporal applications. Second, linear ecological regression can often still work quite well even when one or more of the groups whose behavior is being analyzed is arguably non-homogeneous. Third, linear ecological regression to analyze racial bloc voting patterns can, under not implausible circumstances, yield sufficiently reliable estimates even when we posit the 'true' model to be one where there are context effects giving rise to heteroskedasticity.

Greater reliability of cross-sectional as compared to cross-temporal models

It is my contention that cross-sectional applications of Goodman's technique to election data such as those used in recent voting rights cases do not give rise to as many problems as the usual cross-temporal voting applications because (a) we do not have to worry about the extent to which the pool of potential voters has changed over time; (b) the restriction to a single election makes it more plausible that similar factors are affecting voters of each group; (c) the cross-sectional applications of the Goodman model in the voting rights literature are almost invariably to narrowly delimited geographic areas, again

making it more plausible to posit that similar factors are affecting voters of each group; and (d) in the usual voting rights applications of the Goodman model, data are available on very small units of aggregation in which we normally will have sufficient range in both the dependent and independent variables to develop reliable estimates of the underlying relationships.[51]

Reasonable accuracy even when groups are technically non-homogeneous

It has been asserted that a necessary condition for the use of ecological regression is that groups be homogeneous, in the technical sense of that term (Lupia and McCue, 1990: 365). I am skeptical of this claim. In their own example of voting in the 14th Councilmanic District in Los Angeles, as I showed above, the partitioning methodology they used did not demonstrate any real improvement over the Goodman approach.[52] I believe the title of Lupia and McCue's 1990 article, 'Why the 1980s Measures of Racially Polarized Voting are Inadequate for the 1990s' is unjustified. Given its data requirements, the potential for practical use of the partitioning methodology is suspect in courtroom situations where minority plaintiffs do not have resources to fund complicated computer analyses.[53] Moreover, as the authors recognize, the partitioning methodology is still not fully developed or tested. That makes it particularly vulnerable in courtroom situations to obfuscations by competing expert witnesses as to whether it has been applied correctly and whether its findings are reliable (cf. the discussion of the flaws in multivariate methods for making causal claims about voting in Grofman, 1993b).[54]

Reliable estimates even when there are context effects giving rise to heteroskedasticity

In applying the linear ecological regression to estimate racial bloc voting, Owen and Grofman (1994 forthcoming) show that the accuracy of the linear model will depend on the nature of the distribution of minority voting strength and of the magnitude of the probable contextual effects. In particular, if most whites/Anglos live in areas geographically segregated from those of most minority members and/or if most members of the minority group live in geographically segregated neighborhoods and/or if the magnitude of the context effect is relatively small and can be well approximated by a quadratic relationship, then linear ecological methods will provide good estimates of averages levels of bloc voting of both the minority and the non-minority electorate, despite the fact that a linear model omits potentially important variables that would give rise to what appear to be contextual effects. This point is further reinforced by various empirical work matching ecological regression estimates in biracial contests with data from exit polls (see especially Grofman,

1991a; Lichtman, 1991; Loewen *et al.*, 1993). Similarly, the work of Kohfeld and Sprague (1992), using the Boudon quadratic contextual model, shows that contextual effects are relatively small in most of the biracial contests in St Louis that they examine.[55]

Acknowledgements

This research was supported by Ford Foundation Grant # 446740-47007 and also draws on previous research that was supported by NSF Grant SES # 88-09392. An earlier version of portions of this chapter was prepared for delivery at the National Center for Information and Geographic Analysis Conference on 'Spatial and Contextual Models of Political Behavior', State University of New York at Buffalo, 23–5 October 1992. The National Science Foundation, the NCGIA and the Ford Foundation are not in any way responsible for the opinions expressed in this essay. I am indebted to the Word Processing Center, School of Social Sciences, UCI, for manuscript typing of an earlier version of this chapter, and to Dorothy Gormick for bibliographic assistance.

Notes

1. In this chapter we will focus on the analysis of voting behavior; however, the methods we review can usually be readily adapted for use in other contexts.
2. The distinction between aggregated data and individual-level data is not the same as the distinction between survey and non-survey data. Survey data might be about the behavior/attitudes/characteristics of collective entities such as churches or PTAs. Also, the line between individual-level and aggregate data is not hard and fast: for example, sometimes survey data at the individual level are aggregated to provide insight about behavior in various groupings, such as particular political constituencies, or controls are introduced for geographic location to look for context effects (see, e.g., Wright, 1977).
3. See also Claggett and Van Wingen (1993).
4. We argue below that very few regressions on ecological units are really ecological regressions in the sense of being directed toward inferences about individual-level parameters.
5. For reasons of space, this topic will not be pursued here; see Grofman (1991b, 1991c).
6. We argue below that, in general, the former are less likely to give rise to fallacious inference than the latter.
7. Following Grofman (1987) and Loewen and Grofman (1989), we argue that the failure to use consistent denominators can give rise to important error bias. See below.
8 Here, *inter alia*, we suggest that there are circumstances where, because of multicollinearity or identification problems, simple models may do better for purposes of ecological inference than more complicated ones.

9. Here we focus on one particular quadratic context model, found in Boudon (1963), used by Sprague (1976), Miller (1977) and Grofman (1987).
10. Here we show how use of general political science knowledge about racial bloc voting can help specify what is most probable given the observed data.
11. Ecological units need not be geographic in nature: for example, they might reflect institutional types. However, for simplicity of exposition, we shall use examples where the unit of analysis is geographically rooted.
12. See, e.g., literature reviewed on the effects of election type on minority representation in municipal government, in Grofman and Davidson (1994).
13. This is true, for example, of most of the essays in Berglund and Thomsen (1990).
14. When individual-level inference is our concern, we must be especially cautious to minimize the risk of committing ecological fallacies of the sort warned about by Robinson (1950). We consider this topic in the concluding section of this paper.
15. That is, that errors are essentially uncorrelated with the independent variable x.
16. We omit discussion of how to calculate a confidence range around each of these parameter estimates (see Grofman and Migalski, 1988).
17. A variety of techniques have been devised to cope with this problem (see, e.g., Brown, 1982; Falter and Zintl, 1988).
18. See discussion below.
19. Note that we use $n - 1$ groups as independent variables rather than n groups, just as in the case of two groups we used one group's proportion as the sole independent variable. Because the total population is fixed, we may use a bookkeeping equation to solve for the residual category. Alternatively, we might be able to study multiple groups by considering each group and its complement in a series of bivariate regressions, and then solving a set of simultaneous equations based on the estimates so derived. Because of systematic patterns in the way the various groups are spatially distributed relative to one another, reliable estimates from this latter approach may not always be possible.
20. In addition to differences in socio-demographic attributes and life histories, individuals may differ simply in terms of their geographical context (defined in terms of characteristics of the ecological unit in which the individual finds himself or herself).
21. Adding non-voting as an option can be thought of as adding a third choice. As an alternative to modeling trichotomous behavior (vote for D, vote for not-D, don't vote) in multivariate terms, we show below that it is possible to estimate each of the first two behaviors from a bivariate regression equation, and then use a bookkeeping equality to solve for the third type of behavior as a residual category.
22. Lupia and McCue (1990: 376 at n. 28) quote Shively (1969) as saying that 'The lack of interest in ecological regression is probably due to the fact that errors in estimation are likely to turn up either as negative percentage, or as percentages which are greater than one hundred. This is disheartening to the researcher, and is difficult to present to his colleagues'.
23. Of course, most such 'improvements' also add complexity and almost invariably create new measurement and data problems.
24. Because this topic has been dealt with much more extensively by Philip Shively (1969, 1975, 1991) and other authors, such as Claggett and Van Wingen (1993), we focus on one simple modification of the Duncan and Davis technique, the method of 'simple' percentages (Loewen and Grofman, 1989), which we apply to the study of racial voting patterns.
25. In *Thornburg* v. *Gingles*, 478 US 30 (1986), the Supreme Court in 1986 identified a requisite level of racial bloc voting as one of the basic factors in the proof of a minority vote dilution challenge to an at-large or multimember district system under the 1982 amendments of the Voting Rights Act of 1965. Subsequent cases made racial bloc voting a crucial element in challenges to single-member-district

plans as well. For a discussion of legal issues in voting rights, see Grofman and Handley (1992); Grofman, *et al.* (1992).

26. Whether the black candidate is also the candidate of choice of the (majority of) black voters is a matter for empirical determination. For discussion of legal issues in defining 'candidate of choice' see Grofman, *et al.* (1992).
27. Note that x no longer has the same meaning as in the previous section, where x was defined in terms of the actual electorate.
28. In the USA, while registration data by race are available only for a few states, data on voting age population (or citizen voting age population) by race are available for census units, and these census units can then be matched to electoral precincts.
29. See Goodman (1953).
30. The homogeneity assumption, which is probabilistic in nature, is sometimes referred to, rather misleadingly, as the 'constancy' assumption (Freedman *et al.*, 1991).
31. There are many different subgroupings possible, and no good way is known as yet to determine which partitioning into subgroups should be used. However, as Lupia and McCue (1990: 369) observe, 'research on this [partitioning] problem is still in its infancy'.
32. In contrast, Professors Jerome Sacks and David Freedman, who were expert witnesses for the County, as part of their unsuccessful general attack in the *Garza* trial on the validity of ecological methods (see Freedman *et al.*, 1991) argued that the failure to obtain reliable estimates for subgroup voting showed the general unreliability of ecological regression. For rebuttals to the Freedman *et al.* nihilistic views of the impossibility of valid ecological inference, see Grofman (1991a, 1993a), Lichtman (1991), Loewen *et al.* (1993).
33. Heard sub nom *Thornburg* v. *Gingles*, 478 US 30 (1986).
34. While I had data only on the marginals (total number of registered Democrats and Republicans, total numbers of white and black registrants), I could estimate the number of white Democrats by positing that 95 percent of the black registrants were Democrats.
35. Of course, because of stark patterns of residential segregation in most of the North Carolina jurisdictions I examined, estimating average white and average black voting behavior could be reliably done simply by looking at voting in the racially homogeneous precincts of each group (see Table 1 in Grofman, *et al.*, 1985).
36. It is important to remember that white and black are racial categories; Hispanics may be of any race.
37. This is the type of situation that tripped up expert witness Race Davies in his testimony for black plaintiffs in *Badillo* v. *City of Stockton* (D. Cal. 1989) 956 F 2nd 884. The bivariate analyses he performed in that case of the Jackson–Dukakis vote in Stockton, using Hispanic voting age proportion as the dependent variable, went awry because the black and Hispanic populations were commingled and there were no homogeneous black or homogeneous Hispanic precincts to use as a check. He estimated Hispanic support for Jackson at around 90 percent; the figure derived from exit poll data was closer to 30 percent. However, if we were to switch to multivariate three-group regression, estimates of Jackson support from Hispanic voters would match up reasonably well with those from exit polls. (The analyses leading to this conclusion were completed after a dispute between the plaintiffs' attorneys led to my being withdrawn as an expert witness in the Stockton case.)
38. This is rather like curing the disease (heteroskedasticity) by killing the patient!
39. This model has been proposed by a number of different authors. I gained my familiarity with it from Miller (1977). John Sprague (Sprague, 1976; Kohfeld and Sprague, 1992) attributes it to Boudon (1963).
40. The neighborhood model of Freedman *et al.* (1991) can be thought of as a special case of this model, one where $a_2 = a_1$. Not only is this a restrictive condition in

general, but it can never happen when a_2 and a_1 are of opposite sign. These two parameters will be of opposite sign if there is a context effect in which the willingness of blacks to vote for black candidates increases among blacks who live in blacker precincts, but white solidarity behind white candidates also increases the blacker is the precinct population (see below). For a further discusson of the Freedman *et al.* (1991) model, see Grofman (1991a), Lichtman (1991) and Loewen *et al.* (1993).

41. See Miller (1977) and discussion of this special case of the general model in Grofman (1987).
42. See Owen and Grofman (1994 forthcoming).
43. For example, we may well derive coefficients that make no sense in terms of the [0, 1] bounds.
44. For an excellent example of this approach, see Kohfeld and Sprague (1992).
45. Of course, there may still be an issue as to how well that behavior generalizes to other precincts where the racial composition is more mixed. That is why most experts in voting rights cases draw on both ecological regression and evidence from homogeneous precincts (see Grofman, *et al.*, 1992, chapter 4).
46. Of course, the *ceteris paribus* assumption might fail if there were a partisan contest between candidates of different races in which each was from a party opposite to the party registration held by a majority of the members of his/her race, e.g. a black Republican facing a white Democrat.
47. We leave the proof as an exercise for the reader.
48. In my testimony in *Gingles* v. *Edminsten*, 590 F. Supp. 345 (EDNC 1984), heard sub nom *Thornburg* v. *Gingles*, 478 US 30 (1986), I referred to it as the method of 'constrained percentages'.
49. When testifying as an expert witness in voting rights cases, this is the strategy I would follow as long as substantive conclusions would be completely unaffected.
50. See Grofman (1991a, 1991b, 1991c, 1993a, 1993b); Grofman and Migalski (1988); Owen and Grofman (1994 forthcoming).
51. In contrast, many cross-temporal applications of the Goodman methodology use relatively large units (such as US counties) as data points. As I have argued elsewhere, the Goodman methodology is especially likely to fail when we are estimating parameters over very large units of aggregation in which the range of variation is limited in both the dependent and independent variable. Thus, for example, ecological regressions to infer racial bloc voting patterns in presidential voting that use data aggregated at the level of states will produce nonsensical estimates (see Grofman, 1993b).
52. I do not mean to say that the partitioning approach is never useful. As noted earlier, in the racial bloc voting context, it may sometimes be necessary to examine separately the behavior of Anglo, black and Hispanic voters in a multivariate multigroup model (see Engstrom and McDonald, 1987).
53. The data analysis reported in Lupia and McCue (1990) was funded by the City of Los Angeles, which was a defendant in a voting rights lawsuit brought by the Mexican-American Legal Defense and Educational Fund (MALDEF) and the Department of Justice to overturn the city's 1981 redistricting plan.
54. I should also note that, while I agree with the Lupia and McCue's (1990) criticisms of the almost exclusive reliance on correlation coefficients by some expert witnesses in early voting rights cases, all the major errors in interpreting correlations they refer to occurred in 1984 or earlier, prior to the influential discussion of the differences between statistical and substantive significance in Grofman *et al.* (1985) and the Supreme Court's opinings on that topic in *Thornburg* v. *Gingles* 478 US 30 (1986) (which draw on that article and other work by social scientists as well as on my courtroom testimony in the case: see Grofman, 1992; Grofman *et al.*, 1992, chapter 4).

55. Given the striking degree of racial segregation in St Louis, the existence of relatively small contextual effects does not greatly change estimates of the propensities of the average black or the average white voter to support black candidates.

References

Berglund, S. and Thomsen S. R. (Eds.), 1990, *Modern Political Ecological Analysis*, Finland Abo: Abo Academy Press.

Boudon, R., 1963, 'Propriétés individuelle et propriétés collective: Une problème d'analyse écologique', *Revue Franaise de Sociologie*, 7, 275–99.

Brown, C., 1982, 'The Nazi Vote: A National Ecological Study', *American Political Science Review*, **76**, 285–302.

Claggett, W. and Wingen J. V., 1993, 'An Application of Linear Programming to Ecological Inference: An Extension of an Old Procedure', *American Journal of Political Science*, **37**, 2, 633–61.

Duncan, D. and Davis B., 1953, 'An Alternative to Ecological Correlation', *American Sociological Review*, **18**, 665–66.

Durkheim, E., 1897, *Suicide*, Paris: Quadriges/Presses Universitaires de France.

Engstrom, R. L. and McDonald M. D., 1987, 'Quantitative Evidence in Vote Dilution Litigation, Part II: Minority Coalitions and Multivariate Analysis', *Urban Lawyer*, **19**, 65–76.

Falter, J. and Zintl R., 1988, 'The Economic Crisis of the 1930s and the Nazi Vote', *Journal of Interdisciplinary History*, **19**, 1, 55–85.

Freedman, D. A., Klein S. P., Sacks J., Smyth C. T. and Everett C. G., 1991, 'Ecological Regression and Voting Rights', *Evaluation Review*, **15**, 6, 673–713.

Goodman, L., 1953, 'Ecological Regression and the Behavior of Individuals', *American Sociological Review*, **18**, pp. 663–4.

Goodman, L., 1959, 'Some Alternatives to Ecological Correlation', *American Journal of Sociology*, **64**, 610–25.

Grofman, B., 1987, 'Models of Voting', in Long S. (Ed.), *Micropolitics Annual*, Greenwich, CT: JAI Press, pp. 31–61.

Grofman, B., 1991a, 'Statistics without Substance: A Critique of Freedman *et al.* and of Clark and Morrison', *Evaluation Review*, **125**, 6, 746–69.

Grofman, B., 1991b, 'Multivariate Methods and the Analysis of the Racially Polarized Voting: Pitfalls in the Use of Social Science by the Courts', *Social Science Quarterly*, **72**, 4, 826–33.

Grofman, B., 1991c, 'Rejoinder: Straw Men and Stray Bullets, A Reply to Bullock', *Social Science Quarterly*, **72**, 4, 840–3.

Grofman, B., 1992, 'Expert Witness Testimony and the Evolution of Voting Rights Case Law', in Grofman B. and Davidson C. (Eds.), *Controversies in Minority Voting: The Voting Rights Act in Perspective*, Washington, DC: Brookings Institution.

Grofman, B., 1993a, 'Throwing Darts at Double Regression: A Rejoinder to Wildgen', *Social Science Quarterly*, **74**, 3, pp. 480–7.

Grofman, B., 1993b, 'The Use of Ecological Regression to Estimate Racial Bloc Voting', *University of San Francisco Law Review*, **27**, 3, 593–625.

Grofman, B. and Davidson, C., 1994, 'The Voting Rights Act and the Second Reconstruction', in Davidson C. and Grofman B. (Eds.), *Quiet Revolution in the South: The Impact of the Voting Rights Act, 1965–1990*, Princeton, NJ: Princeton University Press.

Grofman, B. and Handley, L., 1992 ,'Identifying and Remedying Racial Gerrymandering in Single Member Districts', *Journal of Law and Politics*, **8**, 21, 746–69.

Grofman, B. and Migalski, M., 1988, 'Estimating the Extent of Racially Polarized Voting in Multicandidate Elections', *Sociological Methods and Research*, **16**, 4, 427–54.

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

Grofman, B., Migalski, M. and Noviello. N., 1985, 'The "Totality of Circumstances" Test in Section 2 of the 1982 Extension of the Voting Rights Act: A Social Science Perspective', *Law and Policy*, **7**, 2, 209–23.

Kohfeld, C. W. and Sprague. J., 1992, 'The Effects of Racial Context on Voting Behavior in One Party Urban Political Systems: Consequences of Racial Homogeneity and Heterogeneity', prepared for delivery at the National Center for Information and Geographic Analysis Conference on Spatial and Contextual Models of Political Behavior, State University of New York at Buffalo, 23–5 October.

Lichtman, A., 1991, 'Passing the Test: Ecological Regression Analysis in the Los Angeles County Case and Beyond', *Evaluation Review*, **15**, 6, 770–99.

Loewen, J., 1982, *Social Science in the Courtroom*, Lexington, MA: Lexington Books.

Loewen, J. and Grofman. B., 1989, 'Comment: Recent Developments in Methods Used in Voting Rights Litigation', *Urban Lawyer*, **21**, 3, 589–604.

Loewen, J., Burton, O. V., Brischetto, R. R. and Finnegan. T., 1993, 'It Ain't Broke So Don't Fix It: The Legal and Factual Importance of Recent Attacks on Methods Used in Vote Dilution Litigation', *University of San Francisco Law Review*, **27** Summer, 737–80.

Lohmoller, J. B. and Falter. J. W., 1986, 'Some Further Aspects of Ecological Regression Analysis', *Quality and Quantity*, **20**, 109–25.

Lupia, A. and McCue. K., 1990, 'Why the 1980s Measures of Racially Polarized Voting are Inadequate for the 1990s', *Law and Policy*, **12**, 4, 355.

Miller, W. L., 1977, *Electoral Dynamics in Britain since 1918*, London: Macmillan.

Owen, G. and Grofman. B., 1994 forthcoming, 'Estimating the Likelihood of Fallacious Ecological Inference: Linear Ecological Regression in the Presence of Context Effects', *Political Geography*.

Robinson, W. S. 1950, 'Ecological Correlations and the Behavior of Individuals', *American Sociological Review*, **5**, 351–7.

Rokkan, S. and Valen. H., 1970, 'Regional Contrasts in Norwegian Politics', in Allardt E. and Rokkan S. (Eds.), *Mass Politics: Studies in Political Sociology*, New York: Free Press.

Ryssevik, J., 1990, 'Regional Contrasts Revisited: A Study of the Relative Impact of Class and Culture on Norwegian Political Alignments 1909–1936', in Berglund S. and Thomsen S. R. (Eds.), *Modern Political Ecological Analysis*, Abo, Finland: Abo Academy Press, pp. 178–237.

Siegfried, A., 1913, *Tableau Politique de la France de l'Ouest*, Paris: Colin.

Shively, W. P., 1969, 'Ecological Inference: The Use of Aggregate Data to Study Individual', *American Political Science Review*, **63**, 1183–96.

Shively, W. P., 1975, 'Using External Evidence in Cross-Level Inference', *Political Methodology*, **1**, 61–73.

Shively, W. P., 1991, 'A General Extension of the Method of Bounds, with Special Application to Studies of Electoral Transition', *Historical Methods*, **24**, 81–94.

Sprague, J., 1976, 'Estimating a Boudon Type Contextual Model: Some Practical and Theoretical Problems of Measurement', *Political Methodology*. **3**, 3, 333–54.

Wright, G. C., 1977, 'Community Structure and Voting in the South', *Public Opinion Quarterly*, **40**, Summer, 201–15.