THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

Research Paper No. 12

Did Illegally Counted Overseas Absentee Ballots Decide the 2000 Presidential Election?

Kosuke Imai
Gary King

December 2001
ISSN 0000 0000

HARVARD BURDEN OF DISEASE UNIT
NATIONAL INSTITUTE ON AGING GRANT 1-P01-AG17625

THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

This working paper series reports on research supported by the National Institute on Aging program grant entitled The Global Burden of Disease 2000 in Aging Populations (1-P01-AG17625). The purpose of the grant is to strengthen the methodological and empirical bases for undertaking comparative assessments of health problems, their determinants and consequences in aging populations.

Since the publication of the Global Burden of Disease Study 1990, there has been increasing interest in comparative analyses of health outcomes, determinants and consequences. A major revision of the Global Burden of Disease Study has been launched for the year 2000 with the full commitment of the World Health Organization (WHO). The Global Programme on Evidence for Health Policy at WHO has developed a Global Burden of Disease Network, which operates in parallel to the research conducted as part of the program project. The program project will strengthen the scientific basis for the large-scale undertaking led by WHO at the global, regional and national level.

The purpose of this working paper series is to present original research that emerges from the various project components of this program grant. The views expressed in these working papers are those of the author(s) and do not necessarily reflect the views of the Harvard Burden of Disease Unit, the World Health Organization nor the National Institute on Aging.

THE HARVARD BURDEN OF DISEASE UNIT

The Harvard Burden of Disease Unit was established to design, test, and implement methodologies to aid in the effective allocation of health resources. To achieve this end, the Unit conducts research in collaboration with national governments, international agencies and other researchers and policy-makers. The Unit's research has two main foci:

- to forge the theory, design, and implementation of approaches to the combined measurement of mortality and non-fatal health outcomes, in order to develop valid, reliable, comparable and comprehensive measures of population health and comparative assessments of the burden of diseases, injuries and risk factors; and
- to investigate the costs, efficacy and effectiveness of major health interventions applied in diverse settings, toward the goal of establishing a broad database on cost-effectiveness.

Harvard Burden of Disease Unit
Center for Population and Development Studies
9 Bow Street
Cambridge, MA 02138
www.hsph.harvard.edu/organizations/bdu

Did Illegally Counted Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?[1]

Kosuke Imai[2] and Gary King[3]

Abstract

Although not widely known until much later, Al Gore received 202 more votes than George W. Bush on election day in Florida. George W. Bush is president because he overcame his election day deficit with overseas absentee ballots that arrived and were counted after election day. In the final official tally, Bush received 537 more votes than Gore. These numbers are taken from the official results released by the Florida Secretary of State's office and so do not reflect overvotes, undervotes, unsuccessful litigation, butterfly ballot problems, recounts that might have been allowed but were not, or any other hypothetical divergence between voter preferences and counted votes. After the election, the *New York Times* conducted a six month long investigation and found that 680 of the overseas absentee ballots were illegally counted, and no partisan, pundit, or academic has publicly disagreed with their assessment. In this paper, we describe the statistical procedures we developed and implemented for the *Times* to ascertain whether disqualifying these 680 ballots would have changed the outcome of the election. The methods involve adding formal Bayesian model averaging procedures to King's (1997) ecological inference model. Formal Bayesian model averaging has not been used in political science but is especially useful when substantive conclusions depend heavily on apparently minor but indefensible model choices, when model generalization is not feasible, and when potential critics are more partisan than academic. We show how we derived the results for the *Times* so that other scholars can use these methods to make ecological inferences for other purposes. We also present a variety of new empirical results that delineate the precise conditions under which Al Gore would have been elected president, and offer new evidence of the striking effectiveness of the Republican effort to convince local election officials to count invalid ballots in Bush counties and not count them in Gore counties.

[1] This work has been supported by the National Institute on Aging Grant 1-P01-AG17625.

[2] Ph.D. Candidate, Department of Government, Harvard University

[3] Professor of Government, Harvard University and Senior Science Advisor, Evidence and Information for Policy Cluster, World Health Organization

	Gore	Bush	margin
Ballots cast/received by Nov. 7	2,911,417	2,911,215	Gore leads by 202
Overseas absentee ballots[a]	836	1,575	Bush leads by 739
Total	2,912,253	2,912,790	Bush leads by 537

Table 1: Official results of the 2000 presidential election in Florida. Source: Florida Secretary of State's office. [a]Ballots arriving from overseas after November 7th and before the 17th.

1 Introduction

Many aspects of the 2000 election were the subject of considerable media attention and litigation in the uncertain month that followed the voting, and, as such, considerable information about the process is on the public record. For example, we know how the butterfly ballot in Palm Beach County led to the disqualification of many votes that were apparently intended for Gore (Wand et al., 2001; Adams and Fastnow, 2000). By all accounts, this effect was inadvertent, as indicated by the failure of the Democrats to object to this ballot design prior to the election when they were permitted to do so, and it was not illegal. We also now know a great deal about hanging chads, overvotes, undervotes, and various other failures of punch card ballot designs.[1] Democrats and Republicans worked hard at convincing the Courts and local election officials to count and recount the ballots the way they wanted, but ultimately the system produced an explicit and final decision about each of these elements of the overall picture. Many decisions were tremendously controversial, but they were official, legal decisions made in the presence of full information and were eventually accepted by almost all concerned.

In contrast, the overseas absentee ballots were spared most of this attention and all litigation. Yet, they clearly determined the outcome of the election: If only the votes cast on election day were counted, Al Gore would have beat George W. Bush by 202 votes and become the next president. According to official results from the State of Florida, it took the overseas absentee ballots for Bush to outdistance Gore, which he did in the end by 537 votes (see Table 1). The extent to which the law was followed in this small but consequential part of the story escaped scrutiny for some time. After the election was decided, however, the *New York Times* conducted a six month long investigation during which they retrieved the envelopes in which the ballots were mailed, and searched for violations of the law (Barstow and Van Natta, Jr., 2001). In one of the longest set of articles ever published by the *Times*, they concluded that 680 of the overseas absentee ballots that had been counted by Florida counties unambiguously violated one or more aspects of Florida election law and, by any reasonable interpretation of the law, should have been discarded. Indeed, after the *Times* story appeared, commentators and partisans did not contradict these factual claims.

In comparison with other features of the election that have been studied, this problem was not caused by old machines or the inattention of local election officials or political party representatives. It is also not a theoretical problem, in the sense that it does not rely on comparing the intent of the voters with the official vote. Rather, according to the *Times*, the overseas ballot problem was due to blatantly illegal actions on the part of local

[1]See Herron and Sekhon (2001). The 2000 election produced literally dozens of papers, many written within days of the election and posted on the web. For links to many on this growing list, see `http://madison.hss.cmu.edu/`.

election officials that had not been previously noticed. The *Times* argued that local officials were influenced by the deliberate political strategies employed by the Bush campaign, and comparative neglect by the Democrats.[2] They concluded that "Under intense pressure from the Republicans, Florida officials accepted hundreds of overseas absentee ballots that failed to comply with the state laws" (Barstow and Van Natta, Jr., 2001).

Were these 680 inappropriately counted ballots enough to have thrown the election to the wrong candidate? The *Times* hired us to find out. Our conclusions were presented as part of the story (Barstow and Van Natta, Jr., 2001) and our methods were briefly described in a sidebar (Barbanel, 2001). In this paper, we discuss in detail the methods we developed for this project so that others might use them for similar problems. The problem is a straightforward ecological inference: we observe the number of bad ballots in each of Florida's counties and the number of ballots cast and counted for each of the candidates. From these variables, and a variety of other auxiliary information, we try to infer the total number of bad ballots that had been cast for each candidate and see whether this is enough to make up for Bush's 537 official vote margin.

Since the partisan atmosphere surrounding public discourse on this issue was so highly charged, we knew that our work would be subject to more than the usual academic scrutiny, and so we sought a method that was less vulnerable to criticism — perhaps even, from a scholarly perspective, somewhat unreasonable criticism. No such statistical method like this exists of course, especially for an area as uncertain as ecological inference, but much can be done. Our approach, in addition to focusing on the bounds, is to introduce formal Bayesian model averaging procedures to the ecological inference literature and to include a wide range of models that even partisans might have considered. The procedure then computes a combined estimate from these models based on their relative probability of being correct, as indicated by the data. So that others can use the methods we introduce here to analyze other problems, we have included all methods introduced here in the popular program *EI: A Program for Ecological Inference* (available at `http://GKing.Harvard.edu`).

Our results give the exact probability that Gore would have won the election if the law had been followed in this instance. This probability is small, but we know with mathematical certainty that it is greater than zero — a conclusion that can only be seen as remarkable for the world's premier democracy. Secondly, although our analyses show that it is unlikely that illegal overseas absentee ballots alone changed the outcome of the election, we show that Bush's margin of victory would likely have been much narrower if those flawed ballots had not been counted. This supports the argument made by *The New York Times* that the flawed ballots favored Bush much more than other candidates. We also present here a variety of results that did not appear in the *Times* article, including the probability that Gore would have won under various hypothetical scenarios, such as if Katherine Harris had accepted Palm Beach County's recount, which was submitted two hours late. In some plausible scenarios, the probability that Gore would have won is nearly 100%. Finally, we present evidence that the propensity of local election officials to violate the law and accept bad ballots was substantially greater in counties where Bush strategists believed there were more absentee ballot support for Bush and tried to convince election officials to accept bad ballots. This is consistent with the *Times*' thesis and evidence that

[2]The Democrats had planned to contest the absentee ballots, but Democratic Vice Presidential Candidate Joe Lieberman on Meet the Press ended this strategy when he explained that he "would give the benefit of the doubt to ballots coming in from military personnel generally.... Al Gore and I don't want to ever be part of anything that would put an extra burden on the military personnel abroad who want to vote." Gore then backed him up and the Democrats left the Republican strategy unchallenged. See Berke (2001).

these local election officials bent to the will of Republican lobbists.

Section 2 reviews the illegal overseas absentee ballots in Florida and their importance in the course of events following the election. It also provides the description of the data set we received from *The New York Times*. Section 3 introduces our statistical model, Bayesian modeling averaging, and our estimation procedure. Section 4 discusses our substantive results. Section 5 concludes.

2 Invalid Overseas Absentee Ballots in Florida

On July 15, 2001, *The New York Times* published an article, "How Bush Took Florida: Mining the Overseas Absentee Vote," as the result of its six-month investigation on the 2000 US election. The *Times* reporters describe the details of the Bush campaign effort to secure victory by pressuring selected local election officials to count invalid overseas absentee ballots in Florida. In particular, Republicans focused on military ballots and the counties where Bush had his strongest voting base. For example, in counties such as Escambia, Okaloosa, and Santa Rosa, Bush lawyers argued that every vote cast by Americans in uniform should be counted, regardless of the letter of the law. In counties more favorable to Democrats, Bush's lawyers argued exactly the opposite — that local election officials must follow the letter of the law and disqualify any ballot not meeting the rules.

According to the *Times*, this unequal pressure led to unequal treatment by local officials of citizens who cast their ballots from overseas. That partisans would pursue their interests creatively, relentlessly, and even inconsistently in different places is neither a novel claim nor remotely illegal. That local election officials would respond to this pressure by treating voters unequally is a more serious claim. The *Times'* view — "The result was unequal treatment of ballots with the same flaws." — contradicts statements by Florida Secretary of State, Katherine Harris, that the rules were applied uniformly. It also would seem to violate the Equal Protection Clause of the U.S. Constitution, which was part of the stated grounds under which the United States Supreme Court in *Bush v. Gore* stopped the manual recounts.

The 680 ballots that the *Times* judged as flawed fell into one or more of these categories (Barstow and Van Natta, Jr., 2001):[3]

- 344 ballots had late, illegible or missing postmarks (postmarks must indicate that the ballot was cast on or before election day).
- 183 ballots with United States postmarks.
- 169 ballots were received from voters who were not registered, who had failed to sign the envelope, or who had not requested a ballot.
- 96 ballots lacked the required signature or address of a witness
- 19 voters cast two ballots, both of which counted.
- 5 ballots were received after the Nov. 17 deadline but counted anyway.

If we knew for which candidate the illegal ballots were cast, we would immediately know their effect on the election. However, the secret ballot makes this impossible in most

[3] For pictures of some of the bad ballots, see `http://www.nytimes.com/images/2001/07/15/politics/absentee/nat_ABSENTEE_count_index.html`.

	Gore	Bush	Others	total
invalid ballots	?	?	?	680
valid ballots	?	?	?	1824
	836	1575	79	2504

Table 2: The Ecological Inference Problem in Florida. "?" indicates the unknown quantities to be estimated.

cases. The secret ballot was implemented in this case by separating the envelope, with all the information above, from the ballot contained inside the envelope once the latter was counted. Thus, we only have access to these envelopes, the county in which they were counted, and county-level data on the number of bad ballots and the number of ballots cast for Gore and Bush.

Table 2 illustrates the estimation problem at the state level. The question mark indicates the unknown quantities to be estimated. The table illustrates that while we know the aggregate number of invalid and valid ballots as well as the total number of votes each candidate obtained from overseas absentee voters, we do not know their composition, which is the goal of the analysis. Analogous contingency tables also exist for each of the 67 Florida counties, and the same ecological inference problem exists in each.

In addition to this contingency table for each county, we received three other kinds of data for each county. First, from voter registration records, we have data about each overseas absentee voter, including their sex, race, party registration, and whether they were military personnel or civilian. Second, for comparative purposes we also have data available for election-day voters in the 67 counties. Finally, the *Times* also us provided indicator variables for four regions in Florida and some other measures. We use this extra information in ways we describe below to improve our ecological inferences.

3 Ecological Inference for Flawed Ballots

Table 3 presents our formal notation. For each county i ($i = 1, \ldots, 67$), we denote the proportion of invalid ballots among all overseas absentee ballots as X_i, and the total number of overseas absentee ballots which were counted as N_i. We also let Gore's proportion of the vote be T_i. To simplify presentation, we combine the votes for Bush and the other minor candidates as Bush votes. (Although our presentation always involves only the Bush/Gore choice, our empirical results using deterministic bounds in Section 3.1 includes the possibility of bad ballots having been cast for minor party candidates. We handle minor parties in our statistical analyses in Section 3.2 by ignoring the problem at first and then conducting sensitivity analyses; since votes for minor party candidates only total 3 percent, we find, as expected, that they have a very small effect on the overall result. Other analyses (not shown) using more computationally intensive techniques designed to model these choices separately confirm these results (see Rosen et al., 2001).) Each of these quantities are observed. We denote unobserved quantities with Greek letters. Thus, the proportions of invalid and valid ballots cast for Gore are β_i^{bad} and β_i^{good}, respectively.

Although β_i^{bad} and β_i^{good} are used for the estimation, our ultimate quantity of interest is Bush's margin after dropping the invalid absentee ballots. To define this quantity, first define the statewide fraction of bad ballots that went to Gore as the weighted average of

	Gore	Bush	
invalid ballots	β_i^{bad}	$1-\beta_i^{\text{bad}}$	X_i
valid ballots	β_i^{good}	$1-\beta_i^{\text{good}}$	$1-X_i$
	T_i	$1-T_i$	

Table 3: Ecological inference for invalid overseas absentee ballots in Florida

the individual county quantities:

$$\beta^{\text{bad}} = \frac{\sum_{i=1}^{67} N_i \beta_i^{\text{bad}}}{\sum_{i=1}^{67} N_i}, \tag{1}$$

so that we then have

$$\begin{aligned} \text{Bush's margin} &= \text{official margin} - [\text{Bush's bad ballots} - \text{Gore's bad ballots}] \\ &= 537 - [(1-\beta^{\text{bad}})680 - \beta^{\text{bad}}680] \\ &= 1360\beta^{\text{bad}} - 143. \end{aligned} \tag{2}$$

Once we estimate this quantity, we can also estimate the probability of Gore's victory, Pr(Bush's margin < 0), which is another quantity of interest. (Note that β_i^{good} is not used in Equation 2 but is necessary as an ancillary parameter during estimation.)

3.1 Analysis Without Statistical Assumptions

The relationship among the parameters in Table 3 can be represented by the accounting identity

$$T_i = \beta_i^{\text{bad}} X_i + \beta_i^{\text{good}}(1 - X_i). \tag{3}$$

This equation is an identity generated by the aggregation process, and therefore always holds and has no stochastic term. Furthermore, we note that this accounting identity implies a deterministic linear relationship between the two unknown parameters,

$$\beta_i^{\text{good}} = \frac{T_i}{1-X_i} - \frac{X_i}{1-X_i}\beta_i^{\text{bad}}, \tag{4}$$

which traces out what King (1997) calls a tomography line. In addition, before we observe X_i and T_i in any county, we also know that $\beta_i^{\text{bad}} \in [0,1]$ and $\beta_i^{\text{good}} \in [0,1]$.

Once we observe X_i and T_i, we can narrow the bounds further (simply by projecting the line in Equation 4 to the two axes). Thus, without any statistical assumptions, we can derive the upper and lower bounds of β_i^{good} and β_i^{bad} for each county i, which in turn implies the bounds for our quantity of interest, Bush's margin after dropping flawed overseas absentee ballots.

Table 4 shows how the analysis of bounds can be very powerful in some situations. For example, Escambia is one of the counties where many invalid ballots were found. At the same time, this county is one of Bush's strongholds: about 76 percent of overseas absentee ballots for this county were counted for Bush. The analysis of bounds shows that the minimum number of invalid ballots cast for Bush was as high as 48 votes while that for Gore is zero. Thus we know that at least 19 percent of invalid ballots belong to Bush rather than Gore. Santa Rosa County is similar: Out of the total of 55 flawed absentee ballots, at least 37 votes were cast for Bush. Finally, Baker County illustrates

County	Total invalid ballots	Gore's votes official counts	Gore's votes invalid ballots minimum	Gore's votes invalid ballots maximum	Bush's votes official counts	Bush's votes invalid ballots minimum	Bush's votes invalid ballots maximum
Escambia	102	47	0	47	154	48	102
Santa Rosa	55	16	0	16	65	37	55
Baker	1	0	0	0	1	1	1
all counties	**680**	**836**	**5**	**527**	**1575**	**128**	**668**

Table 4: Analysis of bounds for the state and selected counties. Votes for minor parties are included in the calculation of these bounds.

why sometimes the "secret ballot" is not really secret. In this county, only one absentee ballot was cast and also was found to be invalid. Hence, we know from the total tally of absentee ballots in this county — one vote for Bush — that this person voted for Bush and that it was an invalid ballot but included in the official count. (In fact, the name, address, and individual vote cast of all people in counties, like Baker, that cast all their overseas absentee ballots for one candidate, are on the public record. This is because the bounds have zero width whenever either X_i or T_i is zero or one.)

From these county level bounds, we can derive the aggregate bounds for the total number of invalid ballots for each candidate at the state level. The result shows that at least 8 percent, or 128 votes of Bush's 1,575 absentee ballots, should not have been counted, whereas for Gore the minimum number of invalid ballots is only 5 out of his total 836 votes (0.6 percent). Furthermore, it is possible that Bush could have inappropriately benefited from up to 668 out of the 680 invalid ballots.

The most significant conclusion from this analysis is that we cannot exclude the possibility that Gore actually won the election. That is, without making any assumptions other than that the *Times* coding decisions were correct (and again, we saw no objection to them in the media discussion that followed their story), the 537 Bush margin now changes to somewhere from a 126 vote victory for Gore to a 936 vote victory for Bush. Once the ballots were removed from the envelopes, America forever gave up the possibility of knowing for certain who won the 2000 election.

3.2 Statistical Modeling

In order to learn more about who actually won the election — the margins of victory within the bounds that are most likely — we must add some statistical assumptions. With these assumptions, we can make precise probabilistic statements about our quantities of interest. Readers might not agree with us that the assumptions we chose are plausible, or even the best possible, but the price of the more precise conclusions that follow is the additional uncertainty owing to the conclusions being conditional on the assumptions. Of course this is the situation with almost any model-based statistical analysis, but we try to avoid some of these problems by using Bayesian model averaging.

While confidence intervals and standard errors in most statistical models include estimation uncertainty, they do not usually include uncertainty due to model choice. This is especially problematic since analysts rarely have a strong theory to justify their model choice. Bayesian model averaging is a formal statistical methodology that incorporates model uncertainty into the estimation.[4] It more naturally allows one to consider a wider

[4] See Hoeting et al. (1999) for an introduction to Bayesian model averaging in the general case. Draper

range of models, while still producing one set of results. It also overcomes some of the uncertainty due to model choice by averaging different models and weighting them according to how much support each model receives from the data.

Ecological inferences about bad ballots in Florida is an especially good application of Bayesian model averaging methodology for the following reasons. First, model assumptions can play a critical role in ecological inference when the bounds are wide. In addition to sensitivity to model specification, an inherent feature of the ecological inference problem is weak identification.[5] A consequence is that one cannot always use models with as many covariates as are available to analyze ecological data because such models are often not identified. On the other hand, if we have many covariates to predict individual behavior, it makes sense that we should be able to improve our inferences by using them. After all, more data should be better in this area as it is in every other. Our application of Bayesian model averaging reduces this problem by letting analysts combine many small models with the full range of covariates.

Bayesian model averaging was especially important here since political scientists have only rarely studied absentee ballots and we therefore have little prior theory with which to assist in model specification. The procedure thus enables us to conduct an analysis without having to defend one particular specification, or even a small set of specifications. We can, and did, include any specification in the analysis that we or anyone we discussed the matter with thought might be relevant.

3.2.1 Ecological Inference

We now give a brief qualitative overview of King's (1997) models of ecological inference and their assumptions in the context of our overseas absentee ballot analysis. These models have come to called EI, after the software that implements them. We then explain the Bayesian model averaging procedure and the logic behind it and show how to incorporate this methodology into EI. Appendix A gives technical details.

We begin by thinking about the data in terms of the *possible* values of the $(\beta_i^{\text{good}}, \beta_i^{\text{bad}})$ points for each county as line segments (defined by Equation 4) in the tomography plot in Figure 1. Any possible values of the two parameters for each county has to lie on the portion of its own tomography line that falls within the unit square. Without any further assumptions, this is all the information that exists about our quantities of interest. In a sense, then, Figure 1 presents the data in the form closest to the answers we seek.

We then add three assumptions, all conditional on X and a set of covariates Z. First, we assume that β_i^{bad} and β_i^{good} follow a truncated bivariate normal distribution, with truncation on the unit square. The idea here is that whatever the values of the unknown parameters from Florida's 67 counties on their respective tomography lines, they all have something in common, and that any systematic difference between them will be picked up by covariates. The main constraint of this assumption is that the bivariate density is unimodal and that it is constrained to the [0,1] interval in both directions.

(1995) discusses how Bayesian model averaging deals with the model uncertainty. Bartels (1997) is a political science application that uses an approximation to formal Bayesian model averaging procedures.

[5]Including multiple covariates uncorrelated with X_i cause no identification problems in ecological inference. Ecological inference models that do not incorporate information from the deterministic bounds (such as in Goodman (1953)) are not identified when including X_i or variables related to X_i as covariates (King, 1997: 42). Thus, to the extent that ecological inference models that incorporate the bounds, such as King's (1997), are estimable when including X_i or covariates that are related to X_i, the information that makes this possible is coming from the bounds. Hence, weak identification with multiple covariates in modern ecological inference models occurs primarily when both the data have relatively uninformative bounds and the model includes one or more covariates that are correlated with X_i.

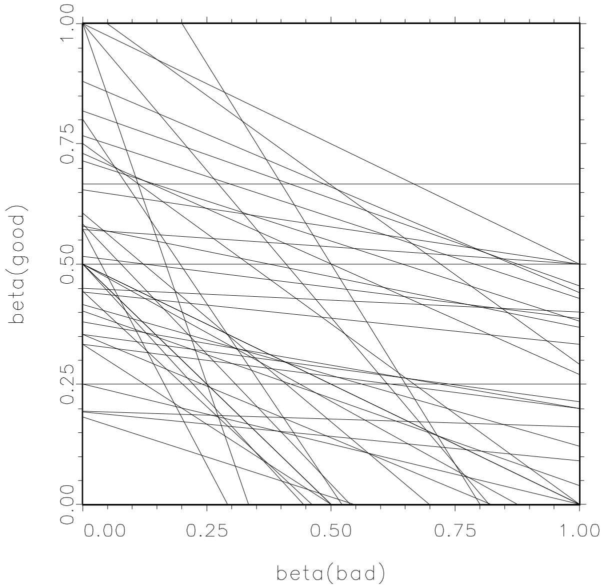

Figure 1: Tomography plot for invalid overseas absentee ballots. β^{bad} and β^{good} are the proportion of Gore's invalid and valid ballots, respectively. Each line traces out the possible values of the $\beta_i^{\text{bad}}, \beta_i^{\text{good}}$ point for each county i.

The second assumption is that the absence of residual spatial correlation in T after taking into account X and Z. King's ecological inference model has not been shown to be very sensitive to anything but extreme levels of spatial autocorrelation, but we make this assumption even more plausible by including tests with covariates that tap into Florida's regions and other spatial features.

The final assumption states that the two unknown quantities, β_i^{bad} and β_i^{good}, are independent of X_i, given Z. For example, if the *The New York Times* reporters are correct that bad ballots were cast disproportionately for Bush in Bush counties, then we should find a strong relationship between β_i^{bad} and variables that measure Bush support.

Each of these three assumptions can be modified to any degree by the inclusion of different covariates. We rely on this point to produce a diverse array of models to which we then apply Bayesian model averaging.

3.2.2 Bayesian Model Averaging

In Bayesian inference, we are interested in the posterior distribution which we obtain through the application of Bayes theorem:

$$\mathrm{P}(\Theta|T) \propto \mathrm{P}(\Theta)\mathrm{P}(T|\Theta). \tag{5}$$

where $\mathrm{P}(\Theta)$ is the prior probability distribution on some unknown parameter Θ, and $\mathrm{P}(T|\Theta)$ is the likelihood (which in our case is given in Equation 10). Everything is conditioned on X, N, and Z, which we observe. We use the standard independent prior on each parameter of Θ as described in section 7.4 of King (1997). This prior distribution and the likelihood function together define King's model in a standard Bayesian framework.

Along with T_i, X_i and N_i, we have 27 other variables including race, sex, and party registration for the overseas absentee voters as well as county level election and demographic data. We define 27 models, each using one of these 27 variables as the covariate. We also include a model with no covariates and 3 models with X_i as the covariate for the mean of β^{bad}, β^{good}, and for both. For each model, we use the same independent prior distribution as explained above. The idea of Bayesian model averaging is to produce one set of results combining these 31 models according to how much support each model receives from the data.[6]

Let M_k denote the kth model specification ($k = 1, \dots, 31$). Then we make an inference about some quantity of interest Δ by computing its posterior distribution via Bayesian model averaging. To do this, we first compute, for each of the 31 models, the posterior distribution of Δ (computed from Equation 5 with Δ being some known function of Θ): $\mathrm{P}(\Delta|M_k, T)$. Then we average over these models by weighting by the relative posterior probability that each model is correct given the data, $\Pr(M_k|T)$:

$$\begin{aligned}\Pr(\Delta|T) &= \sum_{k=1}^{31} \mathrm{P}(\Delta, M_k|T) \\ &= \sum_{k=1}^{31} \mathrm{P}(\Delta|M_k, T)\Pr(M_k|T) \qquad (6)\end{aligned}$$

where the posterior model probability is computed as

$$\Pr(M_k|T) = \frac{\mathrm{P}(T|M_k)\Pr(M_k)}{\sum_{j=1}^{31}\Pr(T|M_j)\Pr(M_j)}. \qquad (7)$$

This is the probability that model k is correct, given the set of models in the analysis; it should not be confused with R^2-like measures which typically reward models that overfit without distinguishing systematic from idiosyncratic features of the data.

To compute the posterior model probability, we need two elements. First is a prior probability that each model is correct, $\Pr(M_k)$. For simplicity, we assign an equal prior probability, 1/31, to each model. This at least does not given any a priori advantage to one model over another. (Section 4.4 discusses sensitivity to priors.)

The other quantity we need to compute for the posterior model probability is the marginal likelihood, $\mathrm{P}(T|M_k)$, which is

$$\mathrm{P}(T|M_k) = \int \mathrm{P}(T|\Theta_k, M_k)\mathrm{P}(\Theta_k|M_k)d\Theta_k, \qquad (8)$$

[6] A list of all 31 models follows: (1) no covariate, (2) X_i for β^{bad}, (3) X_i for β^{good}, (4) X_i for both β^{bad} and β^{good}, (5) military absentee voters, (6) registered Republican absentee voters, (7) registered Democratic absentee voters, (8) female absentee voters, (9) White absentee voters, (10) Black absentee voters, (11) Hispanic absentee voters, (12) rejected military absentee voters, (13) rejected registered Republican absentee voters, (14) rejected registered Democratic absentee voters, (15) rejected female absentee voters, (16) rejected White absentee voters, (17) rejected Black absentee voters, (18) rejected Hispanic absentee voters, (19) Democratic vote share among residents, (20) vote share of Republican candidates, (21) vote share of other candidates, (22) registered Democratic residents, (23) registered Republican residents, registered Black Democratic residents, (24) proportion of voting age population not registered, (25) Black registered Democrats, (26) Black registered Republican residents, (27) acceptance ratio of overall absentee ballots, (28) ratio of invalid absentee ballots, (29) Panhandl Florida regional indicator variable, (30) Southern Florida regional indicator variable, (31) corruption indicator. All the covariates except indicator variables are entered as a ratio varying from 0 to 1. Except the first three models, the covariate was used to model the conditional untruncted mean of both parameters, β^{bad} and β^{good}. Models (5) to (11) are based on information about the absentee ballots and so different variables were available regarding the invalid and valid ballots; we used the former group to predict β^{bad} and the latter to predict β^{good}.

where $P(\Theta_j|M_j)$ is the prior distribution for the parameter vector Θ in Model k. That is, the marginal likelihood is obtained by averaging the likelihood over the prior distribution. The marginal likelihood can be thought of as "the probability of seeing the data that actually were observed, calculated *before* any data became available" (Kass and Raftery, 1995, p.776). That is, instead of maximizing the likelihood with respect to the parameter given the data, as we would do to compute the maximum likelihood estimate, the marginal likelihood does not have a maximization step: it is the average value of the likelihood evaluated at parameter values drawn from their prior density. (Although this quantity could be computed by simulation in this way, such a method tends to be highly inefficient, especially for problems with relatively flat priors or high dimensional parameter vectors. We discuss these and other computational issues in the Appendix.)[7]

4 Empirical Results

4.1 The Probability That Gore Would Have Won Without Bad Absentee Ballots

Figure 2 portrays the posterior distribution from our analysis of Bush's margin of victory if the bad ballots had not been counted (the histogram of 1000 draws from the posterior). Note first that, as required by the procedure, all area for this distribution is contained with the bounds we found for this quantity of -126 to 936. The weight of the statistical evidence within these bounds clearly demonstrates that Bush benefited considerably by the bad ballots, and removing them thus takes away from his margin. This is evident in the figure because almost all of the area of the histogram of posterior probability falls to the left of the official margin of 537. The mean margin of victory for Bush without the bad ballots is only 251 votes. The figure also portrays the probability that Gore actually won the election by the area under the curve to the left of zero. This is only about 0.2 percent, indicating that Gore probably would not have won, even if the bad ballots had been discarded.

4.2 Other Counterfactuals

While our results indicate that it is unlikely that invalid overseas absentee ballots alone would have changed the election outcome, the illegally counted ballots could have had a much more a significant effect when combined with slight changes in decisions regarding the manual recounts. We show this result by first focusing on several scenarios about the two key counties where a manual recount was conducted, Miami Dade and Palm Beach. In Miami Dade County, election officials decided to stop the manual recount when they made the judgment that they could not meet the recount deadline, set by the Florida Supreme Court, 5p.m. Sunday, November 27. The partial manual recount gave a net gain of 157 votes to Gore. In Palm Beach county, they also could not finish the manual recount, but they submitted the result of the partial recount just before the deadline, which would have given Gore a net gain of 192 votes for Gore. Later that day, Palm Beach electoral officials reported the result of the complete recount to Harris. She rejected this complete recount

[7] Estimates obtained via Bayesian model averaging are also known to provide better predictions under certain widely applicable conditions than using any of the individual component models (Madigan and Raftery, 1994). This result is similar to insights from the closely related literature on committee methods (Bishop, 1995), although suprisingly the literatures have relatively few cross-citations. See also Rosen, Jiang, and Tanner (2000) and Robert (1996).

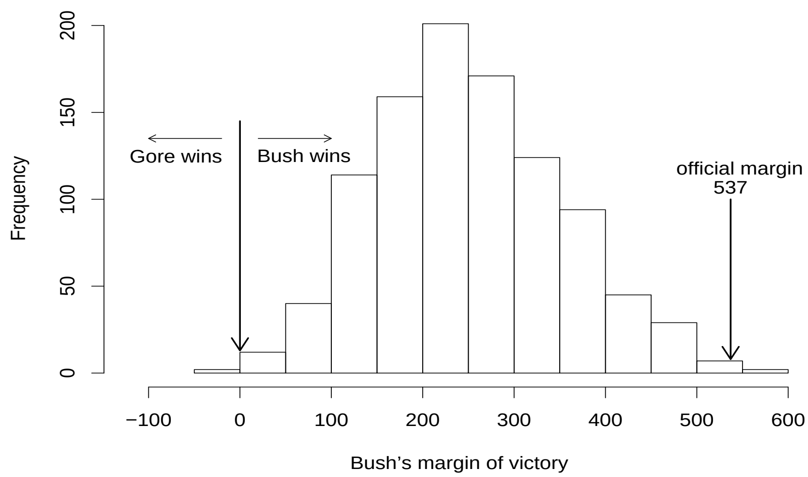

Figure 2: Posterior distribution of Bush's margin of victory without the 680 invalid overseas absentee ballots

as well as the partial recount and did not include them in the certified official tally, thereby denying Gore a total of 349 votes (Purdum, 2000).

The panel of Table 5 marked "actual recounts" presents our prediction for Bush's margin and Gore's probability of victory in situations where the invalid overseas absentee ballots had been rejected *and* the recounts in one or both of these counties had been included in the final tally. For example, if the recounted votes in Miami Dade and Palm Beach had all been counted, Gore would have won with a 0.82 probability, with the uncertainty in this number coming only from our analysis of the bad overseas absentee ballots. If only the Palm Beach votes had been counted, Gore would have won with 0.29 probability. To put it another way, the massive differences in the probabilities from 0.002 to 0.82 for a Gore victory were all due to the decisions of Katherine Harris.

In the last panel of Table 5, we consider counterfactuals where the invalid overseas absentee ballots had not been counted and election day voting recounts had occurred in various ways, as suggested by a study conducted by a consortium of media organizations (Fessenden and Broder, 2001). For example, this analysis shows that if the U.S. Supreme Court had not stopped the recount in *Bush v. Gore*, the victor would have changed with only a 1% probability. However, if Gore's formal request that Broward, Miami Dade, Palm Beach, and Voulusia counties be recounted had been granted, then he would have been elected with a 73% probability. If the entire state had been recounted, according to almost any standard for judging the punch cards, Gore would have won election with a very high probability.

4.3 Indirect Evidence of Local Election Officials Responding to Republican Pressure

Six months of interviews and archival research on the ground in Florida and elsewhere led reporters from the *New York Times* to conclude that, "the Republicans mounted a legal and public relations campaign to persuade canvassing boards in Bush strongholds to waive the state's election laws when counting overseas absentee ballots.... Their goal was

	Bush's margin	Prob(Gore Wins)
Invalid overseas ballots alone	251	0.002
Actual recounts		
Miami Dade partial recount	94	0.19
Palm Beach recount	59	0.29
Miami Dade and Palm Beach	−98	0.82
Media recounts		
No U.S. Supreme Court decision[a]	242	0.01
Gore's request granted[b]	−26	0.73
only fully punched ballots[c]	−366	> 0.99
hanging chads and dimples[d]	−358	> 0.99
each county's standard[e]	−422	> 0.99

Table 5: Estimated margin and probability of victory if the invalid overseas absentee ballots had not been counted along with selected other counterfactuals (each of which also excludes the invalid absentee ballots). [a]Corresponds to what would have happened if the U.S. Supreme Court had not stopped the manual recount. [b]Gore requested that Broward, Miami-Dade, Palm Beach, and Voulusia counties be recounted. [c]Only fully punched ballots were recounted in all counties. [d]Recounting all counties using the standard that any hanging chad or dimple was counted. [e]A recount of the entire state, using the standards adopted by each county.

simple: to count the maximum number of overseas ballots in counties won by Mr. Bush, particularly those with a high concentration of military voters, while seeking to disqualify overseas ballots in counties won by Vice President Al Gore." The *Times* claimed that as a direct result of this pressure, "canvassing boards in about a dozen Republican-leaning counties had reconvened for a second round of counting. In each place, longstanding election rules were bent and even ignored. Boards counted ballots postmarked as many as seven days after the election, including some from within the United States. They counted two ballots sent by fax. Officials in Santa Rosa County even counted five ballots that arrived after the Nov. 17 deadline. Again and again, election officials crossed out the words 'REJECTED AS ILLEGAL' that had been stamped on ballot envelopes."

If these claims are correct, we ought to be able to find evidence of them in our data. We conduct two tests. In the first, we divide Florida's counties into three categories — the six counties mentioned explicitly in the *Times* story where the Republicans pressured officials to count illegal ballots, the four counties mentioned where Republicans pressured local election officials not to count the ballots, and the remaining counties which were not mentioned. We then compute various statistics for these three categories and present them for comparison in Table 6. (The results in this table were not available to the reporters before their article appeared and so Table 6 does represent an independent test.)

The evidence strikingly supports the *Times'* account of events. The first two columns of Table 6 report on the characteristics of the county, information available to Republican strategists before they started lobbying. With the exception of two counties with very few absentee ballots, the counties identified as areas where the Republicans focused their efforts to count ballots were those with large populations of military personnel and Republican voters. Similarly, the counties the *Times* identified as places where Republicans discouraged the ballots from being counted had consistently fewer military personnel and Republican voters.

	military ballots	Republican vote	Bad ballot acceptance[a]	Bad Ballots counted for Bush[b]	all ballots
Republican pressure to count					
Collier	46.7%	65.6%	53.7%	64.5%	60
Duval	83.8	57.5	62.3	67.8	637
Escambia	88.6	62.6	64.2	80.3	272
Okaloosa	88.9	73.7	42.0	69.4	189
Pasco	62.3	48.0	60.5	76.4	53
Santa Rosa	90.3	72.1	84.6	84.4	93
Average	*83.4*	*60.0*	*61.5*	*74.3*	1304
Counties not mentioned by the *Times*					
Average	*67.6*	*51.8*	*30.0*	*71.5*	1751
Republican pressure not to count					
Alachua	46.8	39.8	12.5	54.5	77
Broward	46.9	30.9	21.8	54.3	213
Miami Dade	44.4	46.3	11.7	57.1	306
Palm Beach	45.3	35.3	40.7	56.2	53
Average	*45.6*	*38.1*	*17.2*	*55.4*	649

Table 6: Counties classified by whether the *New York Times* reported evidence of Republican pressure to count or not count the overseas absentee ballots, compared to an average for the remaining counties not mentioned. Averages are weighted by the number of ballots. [a]The percent of bad ballots that arrived with local election officials and were included in the official count. [b]This column is estimated by our Bayesian model averaging ecological inference procedure.

The result of the Republican efforts also appears to have been successful. A larger fraction of bad ballots were counted in all counties where Republicans tried to get them counted than the average, and a smaller fraction than the average were counted in every county where the Republicans tried to have them not counted. The fraction of bad ballots accepted that had been cast for Bush also supports the same theory: Fewer of the counted bad ballots had been Bush voters when the Republicans tried not to have ballots counted than in every county where the Republicans tried to have them counted.

The *Times'* report also helps explain some interesting variations in this table. First, we would have expected more Bush votes among the bad ballots than we found in Duval County because it had so many military personnel. However, the *Times* reported that an election official on the Duval County canvassing board "held the line on counting ballots with missing postmarks." Similarly, Pasco County has relatively low numbers of military ballots and a small Republican vote share. So we might expect that this county to have had relatively few of the bad ballots being cast for Bush. However, the story also described the unusually strong Republican pressure applied in this county: "'It looks to me like we've got a lot of pressure here,' Judge Robert P. Cole, chairman of the Pasco board, said as he faced a throng of cheering Republicans and more than a dozen Bush representatives [and no officials from the Gore campaign]." Our quantitative results are certainly consistent with this qualitative evidence.

We also look for indirect evidence of local election officials succumbing to pressure from Republican Party officials by examining which component models in the Bayesian model averaging procedure were found to have the highest posterior probability. Table 7

	Bush's margin (95 % C.I.)		first difference	Posterior model probability
Bayesian Model Averaging	251	(69, 468)		
Individual models				
Registered Repub. absentee voters, 6[a]	269	(97, 475)	-52	0.565
Dem. vote share, 19	232	(69, 448)	3	0.239
Black absentee voters, 10	231	(69, 440)	-2	0.102
White absentee voters, 9	123	(-18, 315)	-23	0.033
Registered black Repubs., 25	229	(62, 441)	-6	0.021
Accepted absentee ballots, 27	218	(62, 409)	4	0.004

Table 7: Estimates of Bush's margin of victory after dropping the invalid overseas absentee ballots — overall and for the six component models with the highest posterior model probabilities among the 31 models estimated. The first differences represent the increase or decrease in Bush's estimated margin when the value of the covariate increases by 10 percentage points. [a]Each model is identified in the table by the covariate included, followed by the model number we assign to each in Footnote 6.

gives the top six such models listed in order. Generally, if the *Times'* hypothesis is right, we would expect that the covariates that have the biggest effects would be related to where Republicans tried hardest to influence local officials. If they were as rational and deliberate as the *Times* indicated, these would be counties where they expected the largest numbers of bad ballots that, if counted, would help Bush's cause. Obviously, we have no such variable, but there are a variety of variables related to this. In the top six, two have the largest effects and both are consistent with the theory: The more absentee voters registered as Republicans, and the more white absentee voters in a district, the more bad ballots were cast for Bush (the negative sign indicating that Bush's lead is reduced when these ballots are not counted). The other covariates have comparatively small effects.

The large variation in our prediction for Bush's margin across the six models in Table 7 emphasizes a clear advantage of our Bayesian model averaging procedure. The variation results from the large degree of model dependence in these data (because the data have fairly wide bounds). For example, the specification with white absentee voters gives a confidence interval which, when considered in isolation from the other models, would not enable us to reject the hypothesis that Gore won if only the overseas absentee ballots had been rejected. This is obviously quite different from our overall result of only a 0.2 percent probability that Gore won. Since different specifications yield very different inferences, an analyst having to choose one model would be in the untenable position of having to defend choices without a lot of prior evidence. Bayesian model averaging offers a way around this common problem. Instead of results jumping dramatically from one specification to the next, results in Bayesian model averaging do not change as much when new models are added to the specification, unless they have especially large probabilities of being true.

Finally, we give a summary of some of the diagnostics we used to check the fit of the model. First, we reran each model with different starting values to verify that we found the global maximum in each case. We also examined each the tomography plots with contours or confidence intervals to search for outliers or bad model fits. In addition, we plotted $E(T|X)$ or $E(T|X,Z)$ by X or Z, and checked whether the observed T fell within the (say) 90% confidence interval 90% of the time.

We also look at whether the first stage of the model (i.e., before conditioning on T_i) fits by comparing the county-level posterior predictive confidence interval for T_i to its observed value. Figure 3 gives several examples of this strategy for two of the models we ran (in the first two rows), as well as for the ultimate model implied by the entire Bayesian model averaging procedure (in the last row). The first column plots these predictive confidence intervals by X_i, the fraction of bad ballots. The second column plots the same intervals by the covariate included. The proportion of observed T_i points falling inside the 80% confidence intervals is always approximately 80%, which is one check on model fit. Another is the lack of any noticeable relationship between the degree to which the observed T_i deviates from the center of the conference interval and the variable plotted on the horizontal axis. All these diagnostics, and other similar graphs we examined, seem to confirm the fit of the model to these data.

4.4 The Minor Effects of Minor Candidates and Prior Densities

Finally, we analyze the effects on our results of ignoring the minor party candidates and also study the sensitivity of our results to our choices for model priors. Figure 4 plots the posterior distribution of Bush's new margin without invalid overseas absentee ballots for different model specifications for comparison with our posterior distribution. The solid line in this figure is an alternative specification of the ecological inference model where we combine the votes for the minor candidates together with those for Gore rather than with the votes for Bush, resulting in 31 new individual models corresponding to the 31 models already run. The posterior distribution for this model is very similar to the results of our previous analysis indicating that our results do not depend on which model specification is used.

The other two lines represent the use of different prior distributions for the coefficients of the covariate of each model. We use different prior standard deviations for the coefficients of the covariates to see if the results are sensitive to the choice of the prior distribution. While some feature of the posterior do vary across specifications, the graph shows that model results are not particularly sensitive to the prior specification. The median posterior point in particular is hardly changed across the models.

5 Concluding Remarks

Counterfactual analysis is normally very difficult, and especially so when the subject of the inference is far from the factual evidence. When the counterfactual is very close to the data, however, we stand an especially good chance of making valid inferences (King and Zeng, 2001; Lebow, 2000). The counterfactuals in the case of Florida are especially clear and easily could have happened, which makes the results of this case study somewhat more certain than usual. If the problem of the overseas absentee ballots had been litigated and the law applied equally in every county (as *Bush v. Gore* required of the votes cast on election day), the bad ballots might very well have been disqualified. In this situation, although Gore probably would have lost, we conclude that no one will ever be able to say with certainty who would have won the American presidential election if all American laws had been followed. If also the Florida Secretary of State had somewhat different views on issues that were at least somewhat open to discretion, the outcome of the election might very well have changed. Of course, a few different decisions by the candidates on visits to Florida, campaign spending, Elian Gonzalos, or any of a variety of other issues might also have produced a different outcome.

Our results also provide strong, independent, but indirect support for the thesis that

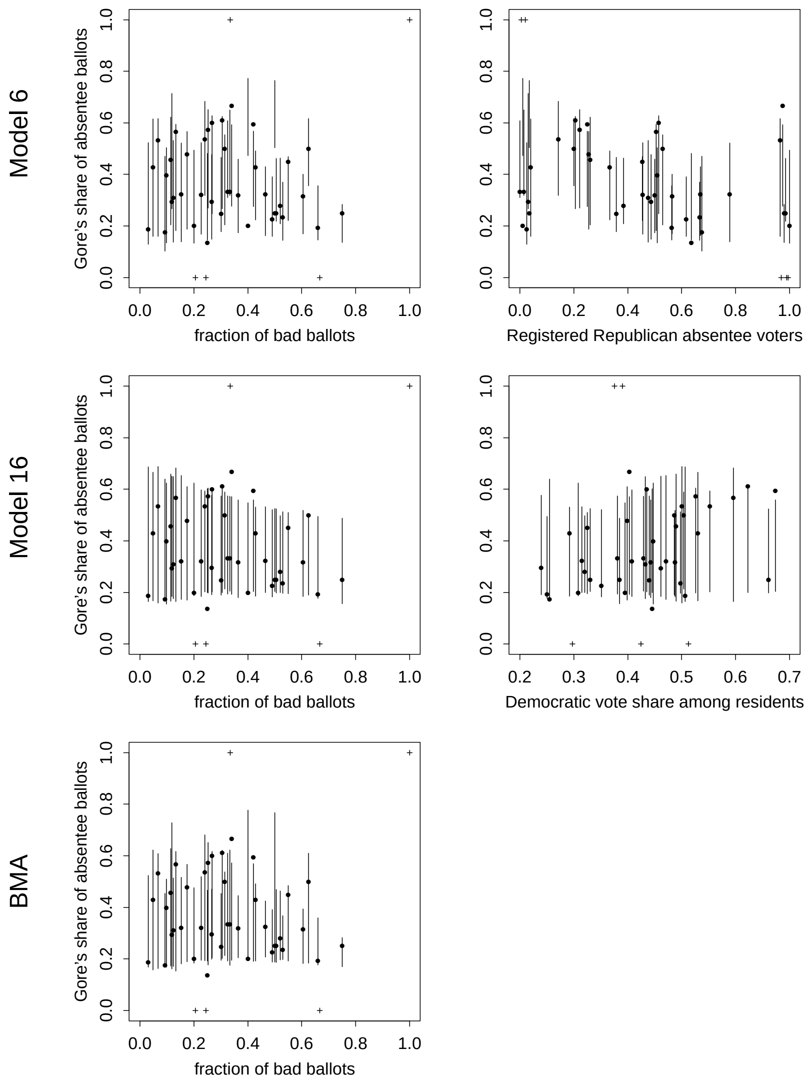

Figure 3: Evaluation of model fit: Predicted distribution of Gore's share of absentee ballots in each county, T_i. 80% confidence intervals of the predictive distribution of T_i given the observed values of X_i, in the first column, and a given Z_i, in the second column. The first two rows give selected individual models and the last is the full model computed via Bayesian model averaging. Dots represent the true values of T_i and "+" indicates that the model predicts exactly without uncertainty.

local election officials bent to the persuasive efforts of Republican strategists to follow the law in Gore counties and break it in Bush areas.

Finally, we think this paper also provides an especially good example of the use of Bayesian model averaging. We have developed the application of it to the ecological inference model and offer computer code for others to use it. In applied work, Bayesian model averaging has been approximated, such as with the BIC, but its full version has not seen as much use as it could.

Bayesian model averaging is a clear improvement on the usual situation of having to

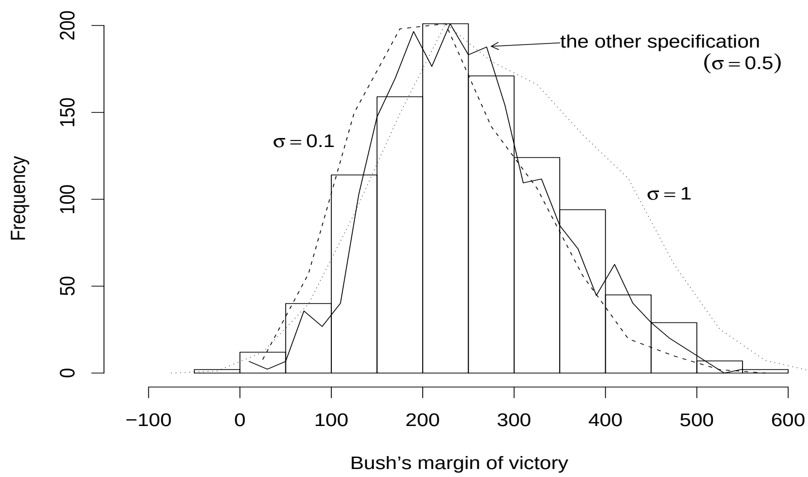

Figure 4: Sensitivity analysis of Bayesian model averaging. The histogram is the posterior distribution of Bush's new margin for our model, while the other lines give posterior distributions from different models and prior specifications.

select and defend a single model, but it is of course not a panecia. A researcher never knows whether all relevant models have been included and, although its results are more robust than single-model approaches, it is always possible to come up with a different list of models and produce a different substantive result. And so in the end, and as always, the investigator's judgment always plays an important role in making inferences. Model averaging cannot substitute for judgment, but it can help account for model uncertainties where prior knowledge is not available. Furthermore, in the present case, where 100% confidence intervals are available (in the form of bounds on the parameters), we also have additional constraints on possible results.

A Technical Issues in Modeling and Estimation

King's (1997) ecological inference model assumes that the joint distribution of $\beta^{\rm good}$ and $\beta^{\rm bad}$ is truncated bivariate normal with mean $\breve{\mathfrak{B}} = \{\breve{\mathfrak{B}}^{\rm bad}, \breve{\mathfrak{B}}^{\rm good}\}$ and variance matrix $\breve{\Sigma}$:

$$\mathrm{P}(\beta_i^{\rm bad}, \beta_i^{\rm good} | \breve{\mathfrak{B}}, \breve{\Sigma}) = \mathrm{N}(\beta_i^{\rm bad}, \beta_i^{\rm good} | \breve{\mathfrak{B}}, \breve{\Sigma}) \frac{\mathbf{1}(\beta_i^{\rm bad}, \beta_i^{\rm good} \in [0,1])}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}, \tag{9}$$

where $\mathbf{1}$ is the indicator function, $\mathrm{N}(\beta_i^{\rm bad}, \beta_i^{\rm good} | \breve{\mathfrak{B}}, \breve{\Sigma}) = (2\pi)^{-1/2} |\breve{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\beta_i - \breve{\mathfrak{B}})^T \breve{\Sigma}^{-1} (\beta_i - \breve{\mathfrak{B}})\}$, and $R(\breve{\mathfrak{B}}, \breve{\Sigma}) = \int_0^1 \int_0^1 \mathrm{N}(\beta^{\rm bad}, \beta^{\rm good} | \breve{\mathfrak{B}}, \breve{\Sigma}) d\beta^{\rm bad} d\beta^{\rm good}$.

This distributional assumption together with the accounting identity in Equation 3 determines the likelihood for the observed data, T_i, which is

$$\mathrm{P}(\Theta|T) \propto \prod_{i=1}^{67} N(T_i | \mu_i, \sigma_i^2) \frac{S(\breve{\mathfrak{B}}, \breve{\Sigma})}{R(\breve{\mathfrak{B}}, \breve{\Sigma})}, \tag{10}$$

where $\Theta = \{\breve{\mathfrak{B}}, \breve{\Sigma}\}$ is a vector of all parameters to be estimated,

$$S(\breve{\mathfrak{B}}, \breve{\Sigma}) = \int_{\max(0, T_i/X_i-(1-X_i)/X_i)}^{\min(1,T_i/X_i)} \mathrm{N}(\beta^{\text{bad}}|\breve{\mathfrak{B}}^b + \omega_i\epsilon_i/\sigma_i, \, \breve{\sigma}_b^2 - \omega^2/\sigma_i^2)d\beta^{\text{bad}}, \tag{11}$$

$\breve{\mathrm{E}}(T_i|X_i) \equiv \mu_i = \breve{\mathfrak{B}}^{\text{bad}}X_i + \breve{\mathfrak{B}}^{\text{good}}(1 - X_i)$, and $\breve{\mathrm{V}}(T_i|X_i) \equiv \sigma_i^2 = \breve{\sigma}_g + 2(\breve{\sigma}_{bg} - 2\breve{\sigma}_g^2)X_i + (\breve{\sigma}_b^2 + \breve{\sigma}_g^2 - 2\breve{\sigma}_{bg})X_i^2$.

Before adding covariates, we reparameterize for estimation as $\phi_1 = (\breve{\mathfrak{B}}^b - 0.5)/(\breve{\sigma}_b^2 + 0.25)$, $\phi_2 = (\breve{\mathfrak{B}}^g - 0.5)/(\breve{\sigma}_g^2 + 0.25)$, $\phi_3 = \ln\breve{\sigma}_b$, $\phi_4 = \ln\breve{\sigma}_g$, and $\phi_5 = 0.5\ln\{(1+\rho)/(1-\rho)\}$. For all but one model we also control for covariates $Z_i = \{Z_i^{\text{bad}}, Z_i^{\text{good}}\}$ via King's (1997: Ch. 9) extended model. This allows the mean (on the untruncated scale) of the two unknown parameters $(\beta_i^{\text{bad}}, \beta_i^{\text{good}})$ to be a function of these covariates:

$$\begin{aligned}
\breve{\mathfrak{B}}_i^b &= [\phi_1(\breve{\sigma}_b^2 + 0.25) + 0.5] + (Z_i^{\text{bad}} - \bar{Z}_i^{\text{bad}})\alpha^{\text{bad}} &(12)\\
\breve{\mathfrak{B}}_i^g &= [\phi_1(\breve{\sigma}_g^2 + 0.25) + 0.5] + (Z_i^{\text{good}} - \bar{Z}_i^{\text{good}})\alpha^{\text{good}}, &(13)
\end{aligned}$$

where α^{good} and α^{bad} are now additional parameters to be estimated and $\bar{Z}_i^{\text{good}}$ and $\bar{Z}_i^{\text{bad}}$ are means of the covariate vectors. This extension of the model relaxes the assumptions of mean independence, no spatial autocorrelation, and truncated bivariate normality, and replaces them with assumptions that are now conditional on the two sets of covariates. This also allows for the possibility of including X_i as a covariate, which is not possible in prior approaches. Our prior specification follows the recommendations of King (1997: Ch. 7).

From one application of one specification of this model, we compute the posterior density of a quantity of interest Δ by drawing it from its posterior, $\mathrm{P}(\Delta|M_k, T)$. To do this, we draw simulations of β^{bad} and β^{good} from their posterior and calculate simulations of Δ. By applying this procedure to each of the 31 models, we get the first of the two components of Equation 6 needed for Bayesian model averaging. The difficult part of computing the second piece is the computation of the marginal likelihood. In particular, the integration in Equation 8 is difficult, especially when the prior on the parameters is relatively flat. A number of estimators have been proposed to improve the precision of the estimation of the marginal likelihood (see Kass and Raftery (1995) for a comprehensive survey).

To compute the marginal likelihood, we use the Laplace approximation,

$$\mathrm{P}(T|M_k) \approx (2\pi)^{d_k/2}|\widehat{V}(\hat{\Theta}_k)|^{1/2}\mathrm{P}(T|\hat{\Theta}_k)\mathrm{P}(\hat{\Theta}_k), \tag{14}$$

where, for Model k, $\hat{\Theta}_k$ is the posterior mode, $\hat{V}(\hat{\Theta}_k)$ is minus the inverse Hessian of the posterior evaluated at the mode, and d_k is the number of parameters. The Laplace approximation is known to perform well compared with other methods (Raftery, 1996; Lewis and Raftery, 1997; DiCiccio et al., 1997).[8] Its rate of approximation is $O(n^{-1})$, which is considerably better than easier-to-apply methods such as Bayesian information criteria (BIC), which has a rate of only $O(n^{-1/2})$ (Kass, Tierney and Kadane, 1989). For example, Kass and Raftery (1995: 778) say that "even for very large samples, it [BIC] does not produce the correct value." Hence, in our application the Laplace approximation is more appropriate than BIC.

[8] If a model is estimated through a Markov chain Monte Carlo (MCMC) simulation, one can construct an algorithm where the chain visits each model stochastically so that you can compute the posterior model probability as a direct output of the MCMC posterior draws. For example, see George and McCulloch (1993), Carlin and Chib (1995) and Chib (1995). However, this requires constructing a new MCMC algorithm for each application.

References

Adams, Greg D. and Chris Fastnow. 2000. "A Note on the Voting Irregularities in Palm Beach, FL." `http://madison.hss.cmu.edu/`.

Barbanel, Josh. July 15, 2001. "How the Ballots were Examined." *New York Times*. `http://www.nytimes.com/2001/07/15/national/15METH.html?pagewanted=all`.

Barstow, David and Jr. Don Van Natta. July 15, 2001. "How Bush Took Florida: Mining the Overseas Absentee Vote." *New York Times* `http://www.nytimes.com/2001/07/15/national/15BALL.html?pagewanted=all`.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Carlin, Bradley P. and Siddhartha Chib. 1995. "Bayesian Model Choice via Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B* 57:473–484.

Chib, Siddhartha. 1995. "Marginal Likelihood From the Gibbs Output." *Journal of the American Statistical Association* 90:1313–1321.

DiCiccio, Thomas J., Robert E. Kass, Adrian Raftery and Larry Wasserman. 1997. "Computing Bayes factors by combining simulation and asymptotic approximations." *Journal of the American Statistical Association* 92:903–915.

Draper, David. 1995. "Assessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society, Series B* 57:45–97.

George, Edward I. and Robert E. McCulloch. 1993. "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association* 88:881–889.

Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663–666.

Herron, Michael C. and Jasjeet S. Sekhon. 2002, forthcoming. "Overvoting and Representation: An Examination of Overvoted Presidential Ballots in Broward and Miami-Dade Counties." *Electoral Studies*.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14:382–417.

Kass, Robert E. and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–795.

Kass, Robert E., Luke Tierney and Joseph B. Kadane. 1989. "Approximate methods for assessing influence and sensitivity in Bayesian analysis." *Biometrika* 76:663–674.

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

King, Gary and Langche Zeng. 2001. "How Factual is Your Counterfactual?" `http://GKing.harvard.edu`.

Lebow, Richard Ned. 2000. "What's so Different About a Counterfactual?" *World Politics* 52:550–85.

Lewis, Steven M. and Adrian E. Raftery. 1997. "Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator." *Journal of the American Statistical Association* 92:648–655.

Madigan, David and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89:1535–1546.

Purdum, Todd S. November 27, 2000. "Counting the Vote: The Overview." *New York Times*.

Raftery, Adrian E. 1996. *Markov Chain Monte Carlo in Practice (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter)*. London: Chapman & Hall Chapter 10 Hypothesis testing and model selection, pp. 163–187.

Robert, C. P. 1996. *Markov Chain Monte Carlo in Practice (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter)*. London: Chapman & Hall Chapter 24 Mixtures of distributions: inference and estimation, pp. 163–187.

Rosen, Ori, Wenxin Jiang, Gary King and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case." *Statistica Neerlandica* 55:134–156.

Rosen, Ori, Wenxin Jiang and Martin A. Tanner. 2000. "Mixtures of marginal models." *Biometrika* 87:391–404.

Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr., Michael C. Herron and Henry Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* forthcoming.