Improving the Accuracy of Ecological Inference[1]

Micah Altman

Harvard-MIT Data Center

Department of Government

Littauer Center G-4

Harvard University

Cambridge, MA 02138

(617) 496-3847 (w)

(509) 562-0658 (fax)

Michael P. McDonald

University of Illinois, Springfield

mcdonald.michael@uis.edu

Illinois Legislative Studies, PAC 480

P.O. Box 19243

---

[1] Pre-print versions of this paper, replication data and code, and links to all on-line resources mentioned (and a number of others, as well) are all available through our numerical stability web page: <http://data.fas.harvard.edu/maltman/numal>

We would like to thank Barry Burden, Wendy Tam-Cho, Brian Gaines, David Kimball, and Gary King for providing assistance (data, programs, and advice) with replication. Their help and the conscientious effort to assist our replication effort are greatly appreciated. We thank Gary King for additional comments, and, we thank Jamie Holmes and Noah Dauber for their excellent research assistance.

Springfield, Illinois 62794-9243

(217) 206-6574 (w)

(217) 206-6542 (fax)

Abstract. Failure to understand the limitations of computational techniques can result in errors in inference. We show how data perturbation and other techniques can be used to probe the numerical stability of Gary King's (1997) solution to the ecological inference problem. We find substantial evidence that computational details are important in evaluating complex political science models.

1. Introduction

The ecological inference problem occurs when inferences pertaining to aggregate data are applied to sub-aggregates of the same data. Ecological inference problems abound in the social sciences. The problem is frequently encountered in voting rights litigation, where the voting patterns of individuals are estimated from aggregate voting precinct election returns. Recently new political science applications are found in determining individual political behavior from congressional district election returns (Burden and Kimball 1998).

Ecological inference has been claimed to be the "longest standing, *hitherto* unsolved problems in quantitative social science" (King 1997, Emphasis Added). The ecological inference problem is easy to define, but the solution is not as simple to resolve. Goodman (1954) was the first to propose a simple linear regression solution.[2] "Goodman's regression" is still widely used in voting rights litigation since the United States Supreme Court adopted it as the standard for evaluating racially polarized voting in *Thornburg* v. *Gingles* (1986). However, linear regression can provide nonsensical estimates when near the logical bounds, such as estimates of turnout rates above one hundred percent.

Gary King (1997) proposes a solution that addresses the limitations of linear regression by combining the constraints of the logical bounds with constrained maximum likelihood estimation. King's proposed solution has drawn particular attention, being the subject of "unprecedented" press releases from NSF and the Council of Scientific Society Presidents', as well as appearing in the popular press. (McCue 2001, Freeman 1997; Yemma 1997). It has also garnered significant academic attention, including a symposium in *Historical Methods* and a book review symposium in *Annuals of the Association of American Geographers (Johnston 2000; O'Loughlin 2000; Sui 2000; Kousser 2002; Lewis 2002; Palmquist 2002]*. Although much of the attention has been positive, the method has also been harshly criticized (Cho 1999; Ferree 1999; Freedman, et al. 1999; McCue 2001).

[2] Another simple solution is to analyze voting patterns in homogeneous precincts. This type of precinct is not always available, and voting patterns may vary differently in non-homogeneous precincts.

In this manuscript we examine the numerical accuracy of Gary King's approach, ways to detect numerical accuracy, and ways improve King's approach. Recently, a variety of econometric software packages have been evaluated for numerical accuracy (McCullough & Vinod, 1999; McCullough 1998; McCullough 1999a; McCullough 1999b; Altman and McDonald 2001). Simply put, numerical methods matter – different ways of computing results can lead to vastly different "solutions" for the same data and model. Often social scientists do not have the luxury of knowing the true solution, so we may find ourselves in the uncomfortable situation of our conclusions being dependent on our choice of software. The good news from the studies of numerical accuracy is that some commonly used statistical tools, such as linear regression and maximum likelihood estimation of single peaked likelihood functions, perform with little or no discernable error on problems of moderate difficulty. Poorly written software (including several commercial statistical packages), more complex models, or difficult data, however, are susceptible to numeric inaccuracy, and we have documented such cases even in published research (Altman and McDonald 2002).

Gary King's solution to the ecological inference problem makes use of what is perhaps one of the most complicated statistical computations in political science[3], and thus its numerical accuracy deserves special attention. The software King has devised to implement his solution is lengthy and, to our knowledge, there exists no other independently-written and publicly available software to estimate the same EI model. King publicly disseminates the most recent version of his solution on his website in two forms, as a Gauss program, which he refers to as EI, and as a standalone version for Windows called EzI.[4]

This paper will proceed as follows: we describe the potential sources of numeric inaccuracies within King's solution, provide ways to test for the presence of numeric inaccuracies (not only for this context but in a broader context as well), perform the tests, and seek ways to ameliorate the inaccuracies we identify. Our analysis of King's solution finds indications that numeric inaccuracies affect not only in

[3] Markov-Chain Monte-Carlo (MCMC) estimations are arguably more complicated, but we are aware of no 'packaged' analysis software that makes use of these.

[4] The EI software is available from <http://gking.harvard.edu>.

examples from King's (1997) book, but also an article that uses King's approach published in *American Political Science Review* (Burden and Kimball 1998). Our analysis shows that numerical accuracy may mislead social scientists that are caught by it unawares – so we must pay attention, if only to place our work and possible problems in the appropriate context. Surprisingly, this does not necessarily complicate our analyses. Since numerical inaccuracy potentially amplifies the effects of other statistical problems such as colinearity, measurement error, and misspecification; increasing accuracy can reduce these other problems, which might otherwise be addressable only through further data collection.

2. Methodology

There are three sources of numerical inaccuracy that King's solution, and many other statistical software, may run afoul. The first is the numerical inaccuracy inherent in all statistical software, which is a consequence of the errors introduced in translating pencil and paper numbers into the binary arithmetic of computers. The second results from the limits of (pseudo) random number generators (PRNGS) that are used in simulation. The third results from the limits of, maximum likelihood optimization algorithms (and other optimization algorithms), which, in addition to being affected by the limited precision of computer arithmetic, are not designed to find the global optima of the likelihood function (except in the rare case where the function is guaranteed to be single-peaked).

These sources of inaccuracy can be more formally described. In the remainder of this section, we discuss them. We conclude the section with a heuristic to measure the sensitivity of a solution to the presence of numerical instability underlying the computations by perturbing the data (adding small artificial measurement error) and noting the distribution of the resulting estimates.

2.1. Background: Computational Sources of Inaccuracy

Floating Point Arithmetic. Most computational algorithms use floating point arithmetic. Numeric inaccuracies are introduced because statistical programs use a fixed number of bits, called *precision*, to

store and manipulate binary numbers and calculations. When the binary representation of a number of exceeds the available number of bits, *overflow* occurs.[5] Surprisingly, often decimals numbers, particularly those with fractions, do not have an exact binary representation, forcing software to *round* or *truncate* numbers to the level of precision. It is here that small errors occur in the translation of numbers from the world of pencil and paper to computers. And manipulating these numbers, such as adding the squares of a large and small number, may propagate these errors or introduce new ones that may produce wildly different answers from the truth. Higham (1994) observes that researchers are often overconfident of the accuracy of their calculations because they are unaware of the consequence of floating point inaccuracy. Even a simple computation may be subject to numerical inaccuracy. Even one small error propagated through a calculation may cause significant inaccurate results. And increasing the precision of a calculation does not monotonically increase the accuracy of the final result.

Random Numbers. Random numbers are fundamental to many types of mathematical simulation, including Monte-Carlo simulation, which is used in King's programs, and in other areas of political science (Mooney 1998; Jackman 2000; King, Tomz & Wittenberg, 2000). Random numbers are also used in sub-sampling techniques, re-sampling techniques such as jack-knife and the bootstrap (Efron 1984), and to pick starting parameters and search directions for some non-linear optimization algorithms (including some of the optional computation methods for the Gauss maximum likelihood algorithm used by King). In simulation, increasing the number of simulations used in the estimation generally reduces this variance, but is subject to bounds determined by tquality of the random number generator used (Fishman 1996), as well as by floating point accuracy. A number of problems have been traced to

[5] For example, usually the number of bits, b=32, so the number "2147483648" (one plus $2^{32-1}-1$) would overflow, and be treated as "-1."

inadequate random number generators, some occurring as early as 1968 (Knuth 1997), and poor generators continue to be used, recommended, and invented anew (L'Ecuyer 1994).[6]

The numbers provided by computer algorithms are not genuinely random. Instead, they are *pseudo-random number generators* (PRNGs), deterministic processes that create a sequence that is statistically similar, in limited respects, to random draws from a uniform distribution. Pseudo-random number generators start with a single "seed" value and generate a repeating sequence with a certain fixed length, or period. In order for simulation or sampling results to be accurate, a PRNG should satisfy three criteria: long period, independence, and uniform distribution. In addition, all require a truly random 'seed' to produce independent sequences (Ripley 1990; Knuth 1997; Gentle 1997).

Nonlinear Optimization. King's programs use a constrained maximum likelihood solver to find the parameters of its solution. Standard techniques for programming an algorithm to find a local optima, which may or may not be the global optima, typically involve examining the numerically calculated or analytic gradients of the likelihood function at the current guess for the solution, and then use these to determine a direction to head "uphill." However, poorly conditioned data or implementations with numerical inaccuracies may prevent the location of local optima, even when the search algorithm itself is mathematically correct.

Non-linear optimization problems, including maximum likelihood estimation, suffer from a further and larger complication, that of finding the global optimum. Although we can determine conditions for global optima for some classes of problems (for example, those functions known to be single-peaked and twice continuously-differentiable), as one set of practitioners in the field a decade ago wrote, "Virtually nothing is known about finding global extrema in general" (Press, et. al, 1989: 290). Most techniques for finding global optima involve some degree of guesswork, or *heuristics*: either the algorithm guesses at

[6] Inadequate random number generators have caused errors in statistical and physical simulation, as well as problems in disparate areas such as gambling devices and encryption. (Vattalainen, Ala-Nissala, and Kankaala 1994; Selke, Talapov, Schchur 1993; Neuman 1995)

initial values for parameters and proceeds to find a local optima from there, or it perturbs a local optima in an attempt to dislodge the search from it.

In fact, for many non-linear optimization problems, solving for the global optima is provably computationally intractable. (Garey & Johnson 1983) Furthermore, it has been proven that that there is "no free lunch" for optimization -- all black-box optimizers (heuristic optimizers) must perform *no better than random search* (or better than any other heuristic), when averaged over all possible optimization problems.[7] (Wolpert & Macready 1997) In other words, all practical optimization algorithms are limited, and in order to choose or build an algorithm wisely, one needs to use specific knowledge about the structure of the particular problem to be solved by that algorithm.

2.2. Analyzing the Sensitivity of EI to Computational Issues

Formally, an algorithm for computing a function is said to be *numerically stable* if small errors in input cause only small errors in output, i.e.: $\hat{y} + \Delta y = f(x + \Delta x)$, where x is the true input, $\hat{y}$ is the true output, and $\hat{y} + \Delta y$ is the computed value (Higham 1994). Δx is error that enters into computations through, for example, converting a decimal number into a binary number with a finite degree of precision. Less formally, a stable algorithm gives "almost the right answer to almost the right problem."

The accuracy and stability of an algorithm are measured both through analysis of the algorithm itself (e.g., examining every operation in the algorithm to compute bounds on the numeric error) and through an analysis of the algorithm's behavior, where the results are better known (e.g., computed using analytic techniques, or using supercomputers to compute the answers with hundreds of digits of precision). [8] Unfortunately, the complexity of King's

---

[7] These theorems apply to such popular black-box optimization techniques as neural networks, genetic algorithms, and simulated annealing. In addition, some of these methods raise other practical problems that render their theoretical properties invalid in all practical circumstances: E.g., 'optimal' training of networks is computationally intractable, in general (Judd 1990), and practical training methods run the risk of getting trapped in local optima. (Michalewicz and Fogel 2000)

[8] This problem is not tractable even when the true precinct level parameters are known. Since correct application of the EI model does not imply that the estimates equal the true parameters we cannot know the true *estimates* that an infinitely accurate EI model would produce.

implementation makes either treatment well beyond the scope of a single paper. However, there are other tests of robustness that we can use, even where the "right" answers are not known.

Particularly, the definition of stability above suggests an exploratory test for a given problem: introduce small random perturbations to the data, on the order of the measurement error of the instruments used to collect it, and recalculate the estimate. Along these lines, and roughly following the recommendations of Gill, et. al, (1981, particularly section 8.3.3) we examine the sensitivity of King's programs to variations in the random draws produced in its simulation phase, to small perturbations in the data, and to variations in the optimization methods and other functions used. Using these techniques, we can assess whether the results are approximately deterministic, whether they are consistent with respect to small perturbations in the data, and whether other implementation factors affect the estimates. This methodology complements standard diagnostic plots in three ways. First, one can use the strictly numeric results as an unambiguous check: Simply check that the range of results across input perturbations and alternative algorithms supports the same substantive conclusions about the results. Second, this methodology may sometimes reveal numerical problems that may be missed in diagnostic plots. Third, this methodology can help to distinguish numeric inaccuracies from, problems that may have appeared in the diagnostic plots to be statistical in nature.

With regard to variations in algorithm, it is important to note that the default algorithm used by a particular statistical package usually has no special claim to accuracy. In Gauss, and by inheritance, EI and EzI, the default algorithms used to estimate maximum likelihood and constrained maximum likelihood are chosen to balance accuracy and performance, not to maximize accuracy (Aptech, 1995a,b). In fact, the default in one package can be an optional method in another. For example, SST uses the BHHH maximum likelihood search algorithm as its default while Gauss implements it as an option. Moreover, no one algorithm is guaranteed to give the most accurate results in all cases.

With regard to input perturbation, what is considered "small" for any particular case, is a matter of judgment. There is an obvious lower limit, perturbations of the data that are at the level below the precision of the machine should not be expected to cause meaningful changes in output. However, the upper limit on perturbations is less clear, and should be roughly equal to the accuracy of data measurement. In political science measurement error certainly dominates machine precision as a source of input inaccuracy.

It should be noted that "instability" discovered by this approach may have several sources. Sensitivity to input perturbations is often the result of floating point error (classic "numerical instability" discussed above), but can also result from sensitivity to measurement error, sampling error that the model has not taken into account, the misspecification of the model implied by ignoring these sources of error, and from the sensitivity of an optimization algorithm to the presence of multiple local optima. These causes may be teased apart if the theoretical behavior of the likelihood function is well understood (e.g., if it is known that the likelihood function is globally concave), and if multiple precision implementations are available, since these will help to reduce floating-point errors.

Additionally, although this methodology does not require that the likelihood function being examined be statistically well-behaved, it does have a natural interpretation for well-behaved maximum-likelihood estimations. If the likelihood function a MLE is well behaved (as King surmises EI's is [pg. 310-311]) then there is a simple mapping between perturbations of data and perturbations of the model. For example, a small normally-distributed noise added to the data should induce a small mean-shift in the likelihood curve. (Cook 1986; St. Laurent & Cook 1993)

. Of course most estimations in use by political scientists, are conditional on zero measurement error in X, and hence the standard errors of such estimations would not be expected to include the variation induced by perturbation. Nevertheless, if the model and software are well behaved, the variations induced should be small. If the amount of variation is large, regardless of whether the cause is multiple optima, sensitivity to measurement error, or pure floating-point error, however, the end-result is that coefficients estimated under that model are less trustworthy. None of these techniques proves accuracy, but they are evidence for it..

Care must be used to distinguish between numerical instability within an implementation of a particular algorithm, and differences between algorithms used to compute the same quantity of interest. We expect that new versions of software will be made more accurate as implementations are improved, and better algorithms found. Software writers have a responsibility not only to make improvements, but also to document the range of acceptable conditions for running their software and the accuracy that can be expected from it. Furthermore, as improvements are made, facilities should be provided to replicate the results

from previous versions.[9]

3. Estimating EI: Two Case Studies

The proceeding section suggests two strategies to identify the presence of numerical inaccuracies in a statistical program. One is to perturb the data and note how sensitive the estimated results are to these perturbations. Another is to run the program under different circumstances, by varying the default options of the algorithm, if such options are available. Similarly, the analysis can be replicated on multiple software packages and computer platforms (e.g., Altman and McDonald 2002). In addition, sophisticated users might provide alternative algorithms for some components of the analysis, if the software platform is sufficiently open to allow it. If the range of the solutions is robust to these variations, then numerical inaccuracies probably do not greatly affect the analysis. We offer no explicit statistical test since these are a heuristics, not classical statistical tests.

King devotes an appendix in his book discussion to computational issues, which is unusual in political science, and is to be lauded. This discussion does not, however, include either any proofs of EI's statistical well-behavedness or provides any numerical accuracy benchmarks or robustness tests. Generally, King is sanguine with respect to this area, stating that the precision of simulated results can be increased indefinitely by increasing the number of simulations used (pgs. 143, 309), that convergence occurs "almost every time" with exceptions being mainly "artificially generated data sets that the model does not remotely fit", and that, EI's results are theoretically robust to local minima, although King believes that it (EI's log posterior) probably has none. (pg. 310-311)

In this section be apply these heuristic tests to applications of King's solution to the ecological inference problem. We examine examples drawn from King's book, *A Solution to the Ecological*

[9] EI itself provides considerable built-in support for replication, diagnosis of statistical and computational problems, and support for different computational options. Such support is quite rare, and to be lauded. EI's reliance on Gauss as a statistical environment sometimes interferes with replication, however, as changes across Gauss versions cannot be controlled for in EI. Furthermore, although many computational options are provided in EI, the accuracy of these options is not always well documented.

Inference Problem (King 1997), and from a study of split-ticket voting appearing in *American Political Science Review* (Burden and Kimball 1998): We automate perturbation of the data and variation of default maximum likelihood search options. We then run the program on different platforms, such as Windows, HP-Unix, and Linux. Since source code is available for many Gauss functions, we are able to apply alternative algorithms, such as an improved PRNG and more accurate distribution functions.[10] Throughout, we note the sensitivity of the estimates to the different circumstances as indications of numeric inaccuracies severe enough to caution inference from these models.

3.1. Replicating Ecological Inference: Random Numbers, Implementation, Optimization Method and Perturbations

We begin by testing the sensitivity of the analyses used in King's (1997) book. The data and Gauss programs used to replicate the tables in King's book are available at the ICPSR replication archive as study #1132. For each case presented in King's book, named according to conventions found in the replication archive, we compare ecological inference results using the same data and model, but vary the PRNG, platform, optimization algorithm and options, and input perturbations. We present these results in Table 1.[11] The code to create these variations is available from our web page.

PRNG. King performs maximum likelihood estimation on a transformation of the variables and then uses simulation, rather than analytical methods, to transform estimated coefficients into the coefficients of

[10] As noted above, to our knowledge, no independent implementation of King's program exists, so we are unable to apply the heuristic test of running the model on different software packages.

[11] Runs under "Unix" were created under HP-Unix 10.10 K900, running Gauss 3.2.43, with CML v. 1.0.41, and EI v. 1.63 (the latest at time of writing) installed. The "Windows" system is a Pentium II, running Windows 95 Sr 2, and EZI v. 2.23. Note however, that the "latest" version number described on the official EI distribution web site did not match that embedded in the EI code itself (we used the latter). Also note that the version numbers do not seem to precisely track the history of the EI code: For example, when new EI code was released incorporating the changes to CDFBVN described in this chapter, the EI version number was not incremented. Nor are previously versions of the EI code (with the exception of EI 1.0) systematically available .

interest to ecological inference.[12] These simulations require a moderate number of pseudo-random numbers be generated, on the order of the number of simulations requested times the number of observations in the analysis. In addition, Gauss's constrained maximum likelihood solver, may in some circumstances, make use of pseudo-random numbers generation.

Our analysis of Gauss's built-in PRNG shows it fails a number of tests of randomness (Altman and McDonald 2001). In addition, most of the replication data sets that we examine require hundreds of thousands of random numbers, which far exceeds the suggested maximum for Gauss's PRNG. To test if these theoretical failings had an observable effect, we reconfigured EI to use a better generator, Marsaglia's KISS generator (1994),[13] and then replicated King's results using this new generator. (Column 4 of ***Table 1***).[14] The good news is that this change of generators had no discernable effect on the estimates, despite Gauss's poor native PRNG.

Platform and version sensitivity. Professor King distributes two versions of his ecological inference solution – source code that can be run under any version of the Gauss platform, called EI, and a standalone program for Windows, called EzI. We test the sensitivity of the implementation to platform by running simple replications using the built-in *eirepl* function in EI running in HP-Unix (the original development platform for EI, see King, 1997, Appendix F), and EzI running in Windows. In addition, we

[12] King uses simulation, in part, because this problem "appears to have no closed form solution" (King 1997, pg. 145). See, McCue 2001, however, for a contrasting view of the analytical tractability of the problem.

[13] We verified output from our random number generator code with the DIEHARD test suite and found it indeed passed all of the DIEHARD tests, as expected.

[14] In order to eliminate simulation variance as the cause of differences between the two sets of runs, we increased the number of simulations to 1000. There was no distinguishable difference between PRNGS, using this number of simulations, but using either PRNG, the third significant digits of the parameters would change between runs.

replicate King's estimations using the original version of EI (v.1.0),[15] and we report the values of these parameters from King's own analysis.[16]

The results of our analysis are presented in the second, third and fifth columns of Table 1. In four of the seven replication sets EzI reported log-likelihoods that were up to *four times* as large as those reported for the same analysis on the Unix platform. Clearly, this could have a tremendous effect on likelihood-ratio tests, and other tests that use likelihood as a for model comparison. In addition, in four cases EzI reported standard errors were substantially different, and in two the betas were substantially different as well.[17]

In one sense, these discrepancies are not problems with *EI* per-se. The differing results across platforms are more likely to be a result of differences between Gauss's implementation on Unix and Windows than from the small differences in EI's code between these two platforms. Furthermore, sophisticated users should expect computational methods to be improved across versions of statistical software[18].

Nevertheless, these results lead us to be somewhat wary: While the documentation of EI leads us to believe that accuracy is improving across versions, there is no formal documentation of EI's accuracy, no

---

[15] We note that although we use the original version of EI, and the original OS platform, we do not use the original Gauss environment or HPUX operating system version that were once used along with the original EI version (v1.0), since these versions are no longer available from the distributors. In addition, we attempted to run the EI replication files using the current version of Gauss on Linux, but EI's replication procedure failed to read the replication files correctly on that platform.

[16] King's replication files contain model parameters, EI settings, and the results from the EI analysis , as well as the data used to produce those results. We used these embedded results to create column one in ***Table 1***. Note, however, that there appear to be a few minor omissions in the text of King 1997. In particular, the starting values in CENS1910 and NJ, cases selection vectors in MATPROII, NJ, and SCSP were supplied by King (i.e., some cases were selected out, using _Eselect), but were not mentioned in the text. These settings do affect the estimations, so we preserved them in our replications.

[17] In this last case, however, EZI reported numerical problems from CML.

benchmarks tests for it, and no 'regression' tests to ensure that options provided for backwards-compatibility produce the results one would have obtained from previous versions. Furthermore, since no independent implementation of EI exists separate from the Gauss platform, and previous versions are not archived, it is not possible, in practice, to escape platform and version sensitivity when using EI.

Optimization Method. In theory, a solution to a mathematical problem should be invariant to the optimization method. In practice, this is often not the case. To test EI's robustness to changes in optimization method, we vary the multiple algorithms available in Gauss. We were able to use the options to the "Constrained Maximum Likelihood" (CML) library to vary across 40 combinations of optimization (secant) method, step (line search) method, and numerical derivative methods.[19]

The results of our analysis are presented in the sixth column of Table 1. The maximum, minimum, and mean values across the 40 combinations are presented for the log-likelihood, betas coefficeints, and standard errors. In two cases EI showed great sensitivity to the optimization algorithm used to perform the constrained maximum likelihood estimation. In four cases, varying the optimization algorithm revealed significant sensitivity of the parameters to method. In one case, SCSP, EI's estimate of the betas varied by 25% of the possible range. Even more dramatically, the estimate of log-likelihood varied by a factor of over one thousand. It is interesting to note that in all of these cases, the EI and EzI results also showed some discrepancies as well, and in only one case was there platform sensitivity (and in that case, minor) that was not also accompanied by algorithmic sensitivity.

Perturbations. We test the robustness of EI against input perturbations by performing 100 additional EI replications running under HP-Unix while perturbing the data. For each precinct, X and T (using King's terminology) are perturbed a small amount, uniformly distributed in [-0.025,0.025]. Measurement

---

[18] Although those critics who claim that issues of statistical computation are irrelevant to political science presumably assume that these changes do not substantively affect the results from models that political scientists use.

[19] In a significant minority of runs for each of the datasets, CML failed to converge properly, almost invariably because the maximum number of iterations had been exceeded. Although EI would attempt to produce parameter estimates in these cases, we did not include the results for these cases in our analysis.

error of +/– 2.5% is not unreasonable in census counts of minority population, and voting rates thereof, given the census undercount (U.S. GAO 1997).

The results of our analysis are presented in the last column of Table 1. The maximum, minimum and mean values across the 100 perturbations of the data are presented for the log-likelihood, betas, and standard errors. In five replications, the perturbations had very little effect – causing variations of only approximately 1%. In contrast, in two of the replications B^b was sensitive to input perturbations, varying by much as 50% of its possible range. This level of variation would be substantively important in almost any application of EI. The variation is particularly troubling because it considerably exceeds the variation implied by the standard error of the estimate.[20]

Errors, warnings, and diagnostics. In a number of cases, the constrained-maximum-likelihood program used by EI reported errors, such as a failure to converge. EI would still attempt to produce an estimate despite these errors. However, EI produces an indicator if a CML error occurs, and we used this return code to exclude all such results from Table 1.

A cautious researcher is admonished by the EI and EzI documentation (and may be admonished by warnings, during runtime) to make use of the many diagnostic tools available in EI to check results.[21] Such checks should include at least: examination of the tomography plots and "fit" plots for fit, examination of the bounds plots for aggregation bias, and examination of the nonparametric plots for

[20] Of course, EI, like OLS and most other estimations in use by political scientists, is conditional on zero measurement error in X, and hence the standard errors would not be expected to include variation. As we argue above, this is beside the point: We know that almost all the data used in political science is measured with error, so if our estimation procedures are strongly affected by it, we have real reasons to doubt our results. Furthermore, we believe that much of the sensitivity to input perturbations displayed by EI is a result of numerical instability, not statistical problems. For example, sensitivity to perturbations is much reduced when we introduce a more accurate cdfbvn (as described below), which would not occur if the sensitivity were purely statistical in nature.

[21] The vast majority of warnings in the course of replications either reported that simulated betas were close to their bounds, or that special methods were needed to invert the Hessian. The original version of EI, although it, too, used special method to treat some Hessians, did not report warnings in such cases.

multimodality (See King, 1997, ch. 16, and the EI documentation). (In addition, we recommend examination of the profile plots of the likelihood. These were not, however, available in EI version 1.0.)

We produced thousands of replication runs, so it was not feasible to examine diagnostics for each run. Furthermore, in contrast to its treatment of errors, EI does not return a code when a warning occurs, so the detection of warnings is difficult to automate. Instead, we visually scanned output for warnings and additionally "spot checked" a number of runs, using all of the diagnostics above. We discovered that warnings are not enough. For example, a reported log-likelihood of over one million in SCSP, one of the most extreme cases, was reported with no accompanying warnings.

Diagnostic plots can point out statistical problems that our diagnostics do not, and can give a richer view of the results. The standard diagnostic plots can also sometimes reveal numerical problems. For example, tomography plots of KYCK88 revealed some evidence of poor fit in EzI and nonparametric plots of MATPROII revealed a slight bi-modality that probably exacerbated the perturbation instability. Even a full set of diagnostic plots, however, is often insufficient. After all, these plots will rarely, if ever, demonstrate a perfect fit – so subjective judgment must be used to determine if the fit is adequate. For example, some plots of the numerically inaccurate FULTON runs under HP-Unix showed no definitive signs of trouble.[22]

Our routines for testing perturbation and alternate algorithms, however, complement standard diagnostic plots in three ways. First, since they are strictly numeric, they are often a clearer test than diagnostic plots: Simply check that the range of results across input perturbations and alternative algorithms supports the same substantive conclusions about the results.[23] Second, they sometimes reveal numerical problems that may be missed in diagnostic plots. Third, they may reveal, as numeric, problems that may have appeared in the diagnostic plots to be statistical in nature.

---

[22] Note that, when viewed in isolation, each of the results was plausible, but when fits were compared across columns, King's original results clearly fit better than the errant replications.

[23] Multiple imputation could be used to provide more precise estimates of the uncertainties resulting from numerical factors, but is not necessary in most cases.

Discussion. These results should be of serious concern to anyone who applies King's technique to substantive questions: They are based upon the same data used to illustrate the King's original analysis, and they are large enough to cause the substantive conclusions drawn from the analysis to change completely.[24] Furthermore, it should be noted that these implementation instabilities are not limited to cases with small numbers of observations, nor would users necessarily have been given any explicit warning or error messages that would have alerted them to these instabilities.

Given these results, we would normally recommend that the practitioners verify their results using multiple independent implementations of the statistical technique before publication. Since no separate *independent* implementations of King's method currently exist, however, we recommend that practitioners continue to use the current program, but that they take a number of steps to guard against instability of results before publishing final results. These steps should include, testing the sensitivity of these results to optimization techniques and to input perturbations, as well as using the rich array of built-in diagnostic graphs. (Code for testing the sensitivity of results to optimization techniques and input perturbations is simple to use, and is available from our website.)

Note that we are not questioning the validity or utility of the King's technique. Any sufficiently complex statistical technique is likely to be affected by implementation to some extent. Our point is that practitioners, in general, should take steps to ensure that their conclusions are not sensitive to implementation, not that they should fear using King's method, in particular.

This story has a reasonably happy ending. After discussing these results with King, he convinced us that much of this instability could be corrected by increasing the accuracy of the cumulative distribution function for the bivariate normal distribution. We located and consulted an expert in this area, Professor Allen Genz, who supplied us with a quadruple-precision function based on an extension of Drezner and

[24] King suggests that many of the variations in Table 10 are a result of numerical instabilities in the tails of the *lncdfbvn* function, that is provided by Gauss, and upon which EI relies. (Personal Communication, 2000) This underscores both the need for software developers to pay attention to numerical stability, and for political scientists to understand the limits of the packages that they use.

Wesolowsky (1989). After porting this function to Gauss, and integrating it into King's program, we tested the areas of previous instability. These were greatly improved, although not eliminated.[25] The more accurate function has now been incorporated into a new version of King's programs as an option. This approach to remove numerical inaccuracies may prove fruitful for sophisticated consumers of statistical software.

[25] Unfortunately, Genz's improved function runs approximately 300 times slower than the original, which rendered reproduction of the entire table infeasible.

Replication Case		King's Replic-ation Files	EI *(v. 1.0)*	EI: *(v. 1.63)*	EI w/ *Better PRNG*	EzI *2.23*	EI w/ Alternative Optimization *[min-max] (mean)*	EI w/ Measurement Error *[min-max] (mean)*
Cen1910	*LL*	*2449*	*2448*	*2448*	*2448*	*2448*	*[2448-2448] (2448)*	*[2316-2402] (2359)*
(n=1040)	Bb	*0.6383*	*0.6363*	*0.6360*	*0.6355*	*0.6357*	*[0.6348-0.6364]*	*[0.6365-0.6547]*
	Stderr βb	*0.0082*	*0.0043*	*0.0047*	*0.0050*	*0.0042*	*[0.0038-0.0050]*	*[0.0041-0.0550]*
	Bw	*0.9497*	*0.9506*	*0.9570*	*0.9509*	*0.9508*	*[0.9506-0.9512]*	*[0.9405-0.9508]*
	Stderr βw	*0.0034*	*0.0018*	*0.0019*	*0.0021*	*0.0017*	*[0.0016-0.0021]*	*[0.0017-0.0229]*
Fulton	*LL*	*591.5*	*369.37*	*370.3*	*370.3*	*589.2*	*[365.0-369.4] (369.0)*	*[365.2-373.4] (368.9)*
(n=289)	Bb	*0.5705*	*0.4149*	*0.4266*	*0.4262*	*0.5723*	*[0.3524-0.4141]*	*[0.4070-0.4174]*
	Stderr βb	*0.0052*	*0.0128*	*0.0194*	*0.0201*	*0.0046*	*[0.0118-0.0332]*	*[0.0109-0.0153]*
	Bw	*0.0428*	*0.4096*	*0.3821*	*0.3829*	*0.0385*	*[0.4115-0.5570]*	*[0.4079-0.4228]*
	Stderr βw	*0.0122*	*0.0302*	*0.0458*	*0.0474*	*0.0108*	*[0.0278-0.0782]*	*[0.0257-0.0361]*
Kyck88	LL	*198.6*	*44.90*	*73.70*	*73.70*	*214.8*	*Fatal Errors*	*[22.54-64.12] (47.87)*
(n=118)	Bb	*0.4216*	*0.5258*	*0.5230*	*0.5158*	*0.5212*	---	*[0.4967-0.5433]*
	Stderr βb	*0.1000*	*0.1103*	*0.1198*	*0.1144*	*0.1581*	---	*[0.0863-0.1269]*
	Bw	*0.7703*	*0.7628*	*0.7630*	*0.7634*	*0.7632*	---	*[0.7556-0.7718]*
	Stderr βw	*0.0071*	*0.0079*	*0.0085*	*0.0082*	*0.0113*	---	*[0.0060-0.0095]*
Lavote	LL	*6487*	*3295*	*3621*	*3621*	*6812*	*[3294-6486] (5361)*	*[3203-3283] (3237)*
(n=3262)	Bb	*0.6259*	*0.6261*	*0.6268*	*0.6266*	*0.6251*	*[0.6247-0.6307]*	*[0.6230-0.6282]*
	Stderr βb	*0.0016*	*0.0080*	*0.0091*	*0.0094*	*0.0016*	*[0.0016-0.0137]*	*[0.0062-0.0093]*
	Bw	*0.7068*	*0.7063*	*0.7061*	*0.7062*	*0.7067*	*[0.7047-0.7069]*	*[0.7055-0.7074]*
	Stderr βw	*0.0006*	*0.0029*	*0.0033*	*0.0034*	*0.0006*	*[0.0006-0.0050]*	*[0.0023-0.0034]*
Matproii	LL	*412.6*	*411.2*	*411.2*	*411.2*	*412.6*	*[411.2-412.6] (412.0)*	*[393.9-412.0] (403.3)*
(n=268)	Bb	*0.5922*	*0.6071*	*0.6123*	*0.6115*	*0.5924*	*[0.5901-0.6085]*	*[0.3309-0.8344]*
	Stderr βb	*0.0440*	*0.0468*	*0.0540*	*0.0541*	*0.0370*	*[0.0384-0.0553]*	*[0.0443-0.2334]*
	Bw	*0.8143*	*0.8101*	*0.8086*	*0.8088*	*0.8142*	*[0.8097-0.8149]*	*[0.7473-0.8873]*
	Stderr βw	*0.0125*	*0.0133*	*0.0153*	*0.0154*	*0.0240*	*[0.0109-0.0157]*	*[0.0126-0.0673]*
NJ	LL	*1043*	*1043*	*1043*	*1043*	*1043*	*[1043-1043] (1043)*	*[1007-1034] (1020)*
(n=493)	Bb	*0.0627*	*0.0642*	*0.0616*	*0.0611*	*0.0629*	*[0.0643-0.0695]*	*[0.0559-0.0761]*
	Stderr βb	*0.0097*	*0.0142*	*0.0095*	*0.0095*	*0.0105*	*[0.0122-0.0175]*	*[0.0081-0.0111]*
	Bw	*0.3791*	*0.3787*	*0.3794*	*0.3795*	*0.3790*	*[0.3775-0.3787]*	*[0.3742-0.3808]*
	Stderr βw	*0.0023*	*0.0033*	*0.0022*	*0.0022*	*0.0024*	*[0.0028-0.0041]*	*[0.0019-0.0026]*
SCSP	LL	*5339*	*5337*	*5657*	*5657*	*5688*	*[5094-1748000]*	*[5222-5305] (5270)*
(n=3185)	Bb	*0.1267*	*0.1197*	*0.1156*	*0.1162*	*0.1139*	*[0.0198-0.2572]*	*[0.0358-0.1281]*
	Stderr βb	*0.0041*	*0.0109*	*0.0278*	*0.0288*	*0.0424*	*[0.0015-0.1340]*	*[0.0038-0.0529]*
	Bw	*0.1786*	*0.1841*	*0.1888*	*0.1883*	*0.1904*	*[0.0548-0.2770]*	*[0.1771-0.2622]*
	stderr βw	*0.0038*	*0.0100*	*0.0256*	*0.0265*	*0.0391*	*[0.0014-0.1234]*	*[0.0035-0.0487]*

Table 1. Comparison EI results for identical data and models, across variations of platform, algorithm, and input perturbations.

(Notes: Shaded cells have a log relative error with respect to the base EI replication approximately <=1 – this indicates, roughly, disagreement at the first significant digit.)

3.2. Replicating an Analysis of Split-Ticket Voting

Our next case study is of Burden and Kimball's (1998) exploration of split ticket voting, and was the first application of King's method to independently appear in publication. Burden's and Kimball's (1998) aggregate data is the percentage of persons who voted for the two major party presidential and congressional candidates in 1988 congressional districts. From this, they wish to estimate the percentage that split their tickets among candidates from different parties, which they refer to as Bush and Dukakis splitters. Since fewer persons vote in congressional than in presidential races (i.e., roll-off), Burden and Kimball must estimate a two-stage model.[26] In the first stage Burden and Kimball estimate the percentage of roll-off between two types of ticket splitters, and in the second stage, they use the first stage estimates to estimate the rates of ticket splitting.

Determinism, Numerical Instability, and Implementation Dependence

Table 2 presents the results of 195 replications[27] of Burden and Kimball's analysis on HP-Unix (EI) and Windows (EzI) platforms.[28] For our HP-Unix analysis we produce 25 replications for each setting.[29] Through programming, we automate the replication process using the original (v1.0) and current (v1.63) version of EI on a HP-Unix system running the current version of HP-

[26] King's solution specifically applies to 2 x 2 tables of aggregate data. King further leverages information about the logical bounds by using a truncated [0,1] bivariate normal distribution to characterize the distribution of district propensities. King recommends a multi-stage approach to solving the more general R x C case (King 1997: see section 8.4 and Chapter 15), which essentially involves iteratively applying the estimation to data analyzed through a previous 2 x 2 estimation.

[27] Burden and Kimball, following the "Replication Standard" (King 1995), deposited a replication dataset with *ICPSR* (ICPSR #1140), prior to the publication of their research. The code necessary to reproduce their analysis was not provided, so we used the text of their article to guide our replications.

[28] EI reported estimates despite the non-convergence of the CML routine, but we felt that a cautious researcher would reject these estimates out of hand. This behavior is however, yet more evidence of platform-related variance.

[29] In our analysis, we control for an undocumented and apparent bug in EI's second stage estimation that could mimic large simulation variance in some circumstances. EI's second stage estimation procedure alters the EI global variables that store starting values for the likelihood estimation and number of simulations. These values, although suitable for the second stage, can interfere with the functioning of the CML estimation if not reset, the next time that a first stage estimate is run.

Unix Gauss (v3.2.43) and the constrained maximum likelihood (CML) algorithm (v1.0.41).[30] Our Windows analysis uses the stand-alone Gauss module EzI (v2.23), which is based on EI (v1.63), Gauss for Windows (v.3.2.2) and an unknown version of the CML algorithm. We produce only five replications for each setting since the EzI interface is menu-driven and we are forced to run those replications by hand.

Determinism and Numeric Stability. Our analysis indicates that for this case, EI is both approximately numerically deterministic and stable, at least in a restricted sense. By producing many replications we assess "simulation variance" induced by EI's use of random draws to calculate some distributions in lieu of analytic solutions. Simulation variance is not unique to EI and should not be mistaken for numeric instability.[31] Our results are presented as intervals, rather than point estimates, to capture the variability that arises in each replication.

[30] The Gauss environment used with the original EI version (v1.0) is no longer distributed.

[31] Even if the model is exactly the same, the random component of simulation that induces some variance from run to run.

	First Stage			Second Stage		
	β^b	β^w	Log-Likelihood	Dukakis Splitters	Bush Splitters	Log-Likelihood
Original EI (v1.0)						
HP-Unix, 25 Replications	*[.9711-.9713]*			*[.2185-.2214]*		
Original cdfbvn setting	*([.0008-*	*[.8664-.8666]*		*([.0060-*	*[.3094-.3121]*	
Original Hessian setting	*.0010])*	*([.0006-.0008])*	*-671.9*	*.0082])*	*([.0057-.0078])*	*[494.4-496.8]*
Current EI (v1.63)	*[.9304-.9314]*			*[.2125-.2181]*		
Default cdfbvn setting	*([.0024-*	*[.9006-.9013]*		*([.0068-*	*[.3220-.3269]*	
Default Hessian setting	*.0032])*	*([.0021-.0027])*	*1057*	*.0088])*	*([.0059-.0078])*	*[492.3-499.7]*
	[.9271-.9452]			*[.2074-.2134]*		
Original cdfbvn setting	*([.0344-*	*[.8886-.9041]*		*([.0065-*	*[.3145-.3222]*	
Default Hessian setting	*.0488])*	*([.0294-.0416])*	*1034*	*.0092])*	*([.0063-.0085])*	*[495.8-507.9]*
	[.8984-.9158]			*[.2091-.2178]*		
Default cdfbvn setting	*([.0186-*	*[.9138-.9286]*		*([.0072-*	*[.3290-.3374]*	
Original Hessian setting	*.0274])*	*([.0159-.0234])*	*1057*	*.0099])*	*([.0059-.0089])*	*[490.4-498.9]*
	[.9404-.9411]			*[.2129-.2170]*		
Original cdfbvn setting	*([.0020-*	*[.8922-.8928]*		*([.0061-*	*[.3178-.3230]*	
Original Hessian setting	*.0026])*	*([.0017-.0023])*	*1034*	*.0088])*	*([.0056-.0080])*	*[494.3-499.9]*
Perturbation Analysis						
25 Replications	*[.8271-.9943]*			*[.1791-.2211]*		
	([.0012-	*[.8469-.9892]*		*([.0064-*	*[.2971-.3467]*	
Original EI (v1.0)	*.0451])*	*([.0011-.0385])*	*[-645.3-1045]*	*.0094])*	*([.0057-.0084])*	*[490.7-505.1]*
	[.9016-.9555]			*[.2075-.2170]*		
	([.0022-	*[.8797-.9255]*		*([.0072-*	*[.3118-.3298]*	
EI (v1.63)	*.0635])*	*([.0019-.0542])*	*[939.4-996.5]*	*.0203])*	*([.0067-.0176])*	*[492.8-503.6]*
Windows EzI						
5 Replications						
Burden and Kimball				*.198*	*.331*	
EzI (v.1.21)		*Not Reported*		*(.007)*	*(.006)*	*Not Reported*
Current EzI (v2.3)	*[.9026-.9122]*					
Current cdfbvn setting	*([.0366-*	*[.9169-.9250]*		*[.2015-.2094]*	*[.3240-.3294]*	
Current Hessian setting	*.0391])*	*([.0312-.0334])*	*1060*	*([.0080-.0092)*	*([.0071-.0080])*	*[494.6-502.0]*
	[.9060-.9169]			*[.2048-.2388]*		
Original cdfbvn setting	*([.0338-*	*[.9128-.9221]*		*([.0074-*	*[.3246-.3534]*	
Current Hessian setting	*.0401])*	*([.0289-.0343])*	*1057*	*.0787])*	*([.0058-.0666])*	*[496.5-501.5]*
	[.8837-.8842]			*[.1917-.2000]*		
Current cdfbvn setting	*([.0020-*	*[.9407-9412]*		*([.0069-*	*[.3298-.3319]*	
Original Hessian setting	*.0026])*	*([.0017-.0023])*	*1060*	*.0086])*	*([.0055-.0069])*	*[500.7-501.7]*
	[.8838-.8841]			*[.1975-.1991]*		
Original cdfbvn setting	*([.0015-*	*[.9408-.9411]*		*([.0074-*	*[.3298-.3315]*	
Original Hessian setting	*.0017])*	*([.0013-.0014])*	*1057*	*.0087])*	*([.0059-.0070])*	*[501.2-501.7]*

Table 2. 195 replications of Burden's and kimball's EI analysis, and varying version and default settings. Note that estimates are presented as intervals to represent simulation variance.

Table 2 shows that EI produces approximately deterministic first stage estimates with regards

to Burden and Kimball's data.[32] For all first stage replications, where we vary only the version and some important default settings, the constrained maximum likelihood estimation consistently found the same optimum for each configuration. The point estimates of the coefficients fluctuate around the point estimate of the standard error, as is expected with simulation variance.

We additionally test the numeric stability of the original and current versions of EI by adding 1% simulated measurement error to T (see King 1997: 30, for notation), the House election turnout divided by the Presidential turnout. The perturbation analysis in Table 2 shows the CML algorithm finds different optima and greater variation in the point estimates, as is expected when noise is added to the data. The analysis indicates that changes to EI have improved the sensitivity of the first stage estimates to perturbations of data. The original version of EI showed considerable sensitivity to perturbations, with large swings in the value of the log-likelihood and point estimates of the coefficients. The current version of EI shows roughly a three-fold decrease in the sensitivity of EI's point estimates to perturbation. Further, for no replication did the constrained maximum likelihood procedure fail to converge or find an optimum far from the optimum found in the unperturbed replications of the current version

Second stage results are based on first stage estimates generated by simulation. Simulation is similar to perturbation analysis in that noise is added to the second stage analysis, as the first stage precinct level point estimates used by the second stage model randomly vary with each replication. As such, we can see that, even when data is unperturbed, the second stage optimum found by the CML algorithm varies across each random draw from the first stage simulation. However, the range of the log-likelihood is small and in no case did the algorithm fail to converge.

Our perturbation analysis of the second stage captures two forms of simulation variance, one built into King's program through simulation, and one artificially supplied. We see that the value of the second stage log-likelihood varies across the doubly perturbed data sets, but the optima are close to one another, indicating stability. The second stage point estimates show stability within the range of their simulation variance, even where the first stage estimates show greater sensitivity to perturbations. Again, the perturbation analysis of the second stage estimates suggests improvements to King's program have led to a three-fold improvement of the

[32] It may be possible that particularly ill-behaved data or models could increase the instability of the results.

sensitivity of the second stage estimates to perturbations.

Option Dependence. We begin with our analysis of option dependence of the estimates of EI by investigating the original and current version of EI run on the HP-Unix operating platform since we can automate the analysis to produce a large number of replications, and thus bolster the confidence of our results. In Table 2, the first stage point estimate for β^b in the original version is in the range [.9711-.9713] and for the current version the estimate is [.9304-.9314], while for β^w in the original version estimate is [.8664-.8666], while for the current version the estimate is [.9006-.9013]. The second stage estimates show more consistency; the estimate for Dukakis splitters is in the original version is [.2185-.2214] whereas the current version estimate is roughly [.2125-.2181], while for Bush splitters the original version estimate is [.3094-.3121] and the current version is [.3220-.3269]. For all estimates, the range of the estimates does not overlap, suggesting the differences are real and not the result of simulation variance. The value of the log-likelihood shows a difference of an order of magnitude, it is unclear if this is a result of option dependence – choices in the generation of internal functions – or instability induced as a result of the use of less accurate functions in earlier versions.

We identify two changes in EI that are sources of option dependence, changes in the cumulative bivariate normal distribution algorithm (cdfbvn) and method of inverting the Hessian. By forcing EI to use the methods from previous versions (by changing EI 'global' variables), we can assess how these changes affect the performance of EI.[33]

The cumulative bivariate normal distribution algorithm is an important factor in determining the shape of the likelihood function. The shape of the likelihood function determines not only the optimum of the function but also the Hessian, the second derivative of the likelihood function used to determine the standard errors of the estimation. The original default cdfbvn is a fast algorithm, but subject to inaccuracies for small values, while the current default represents a tradeoff between accuracy and speed. King recommends the use of the current default, though provides a slower and more accurate version as well (King 2000: 8). The maximum likelihood

[33] We also discovered that EI's built-in replication facility, *Eirepl*, only partially controls for option and global variable settings that control the details of statistical computation. First, *Eirepl* fails to restore settings for behavior that was fixed at the time of the replication, but made optional in later versions of EI. Second, *Eirepl* makes no attempt to capture the settings that affect statistical computations performed for EI by the underlying CML libraries and Gauss system.

solution is slightly different between the two versions, as evident in the value of the log-likelihood.[34] The new default appears to be more consistent than the older version, supported by the spread of the first stage point estimates, which are an order of magnitude greater when the cdfbvn is set to the old version. The spread of the estimates (induced by simulation variance) using the old cdfbvn cover the current version estimates, though there is a slight skew. Generally, we would recommend that before publication, researchers should confirm their results using the most accurate version of cdfbvn available in King's program.

A second source of option dependence is the method used to invert a non-positive definite Hessian. In all of our replications the first stage estimation results in a Hessian is not positive definite. In these circumstances the program uses specialized methods to find a "close" Hessian that is invertible. The program attempts a number of methods in sequence, and exits on the execution of the first successful method. As new versions of the program have been developed, new techniques have been devised to handle situations when the Hessian is not positive definite. The sequence of methods applied, when the normal method fails, has also changed. In early versions of the program, such as the one used by Burden and Kimball, the first specialized method that the program will attempt is documented as a "wide step procedure" or "quadratic approximation with falloff"(King 2000: 9). In later versions of EI, the program attempts a generalized inverse cholesky alternative proposed in a working paper by Gill and King (2000) and based on Schnabel and Eskow (1990). These methods are not guaranteed to produce meaningful results in all cases; the researcher must exercise caution, since there is a paucity of theoretical and empirical work in favor of any particular method.

Replicating Burden and Kimball while changing the current version of the method of inverting the Hessian to the wide step method shows option dependence. The spread of the first stage estimates are not within the same range for the two versions. For example, for the current version default settings β^b is [.9304-.9314], whereas for the old method of inverting the Hessian β^b is [.8984-.9158]. The spread of the second stage estimates are only slightly covered for Dukakis splitters only. We also note that there is a greater spread to the range using the old version, suggesting that the new cdfbvn method is indeed more stable. Furthermore the cholesky

[34] An examination of the constrained maximum likelihood solution shows differences in estimated coefficients as well.

method produces systematically larger (and in our opinion, more believable) estimates of standard errors than the wide step method.

Interestingly, the interaction of the implementation changes has a greater effect on the program's estimates than one change in isolation. The cdfbvn algorithm alone mostly affects the spread of the estimates, and changing the method of inverting the Hessian in the current version to the old method actually moves the range of the point estimates *away* from those of the original version. Only varying both cdfbvn and the method of inverting the Hessian in the current version produces point estimates that fall between the original and current versions. Even so, we note that we still have not been able to reproduce the EI (v1.0) results with the current version of EI (1.62). There are still a number of changes, beyond option settings, that affect the performance of the program.

Platform Dependence. We assess implementation dependence in regards to operating system by replicating our HP-Unix analysis using the Windows version of EI, EzI (v2.3). We only run five replications for each permutation of settings since EzI's graphical interface forces us to run the analysis by hand. Unfortunately, we do not have access to any definitive earlier versions of EzI, as King does not provide a software archive. In our analysis presented in Table 2, EzI shows signs of internal implementation dependence, just like the HP-Unix version. The critical factor appears to be the choice of the method of inverting the Hessian, while the choice of cdfbvn appears to affect the estimates only slightly. Only when the old method of inverting the Hessian is used does the range of coefficients cover the second stage EI results as originally published by Burden and Kimball.

Comparing Windows to HP-Unix, we see that operating system makes a difference as well, particularly for the first stage estimates, where there are greater discrepancies between the two platforms. Again, we find only in our Windows analysis ranges of coefficients that cover the original published results, though generally the second stage estimates between platforms are close, though do not cover, each other.

To summarize, even where the model and the data remain exactly the same as that used by Burden and Kimball, changes in the details of statistical computation in EI can significantly affect the estimated model parameters. We hope and expect that software will generally be improved across versions, and we believe that new versions of EI are more accurate. Neither the improvements (especially the changes in Hessian methods), however, nor the level of accuracy

that can be expected from the estimation are well-documented. Given the differences in behavior among versions exhibited here, users may wish to proceed with caution.

Incorporating Implementation Uncertainty into Burden and Kimball's Results

Burden and Kimball are not merely interested in the estimates produced by King's method. They use these estimates as a dependent variable in a regression equation. As we have shown above, these estimates should not be considered simply as point estimates. Variation enters into the estimates through the use of simulation and the numerical inaccuracies of the program. Here we attempt to summarize these uncertainties on the Burden and Kimball's regression estimates.

How does implementation dependence affect Burden and Kimball's analysis? The first stage estimates are more sensitive to implementation dependence than the second stage estimates because of the distribution of the first stage parameters to be estimated. The first stage estimates roll-off, the number of ticket splitters that voted in the Presidential election, but not the House election. Given the small number of voters that engage in roll-off, the point estimates are close to the [0,1] bound. Further, the point estimates presented in Table 2 are averages across districts, and in some cases are pressed rather tightly against their bound, too. A method that changes the mean or the standard error of the estimate will produce different simulation variation, affect how many estimates are censored at the bound, and thus greatly affects the overall estimate average. Although an analysis of roll-off is not central to Burden and Kimball's analysis, the estimates vary between 2.9 to 13.4 percentage points, suggesting implementation dependence could be of significant importance of research into roll-off.

The second stage estimates appear relatively robust to implementation dependence. Burden and Kimball estimate the percentage of Bush splitters – those that voted for President Bush in 1988 and a Democratic House member – to be .331, while a representative estimate using the current version of EzI is .324, for Dukakis splitters the two estimates are .198 and .215, respectively. These estimates are feasibly within simulation variation of each other. However, it may be possible that these district averages mask some changes in the underlying distributional patterns.

Do these differences have any importance to Burden and Kimball's ultimate analysis of split ticket voting? This may be the wrong question to ask, since, as we show below, there are

reasons to believe the Burden and Kimball model is misspecified. But, for the moment we will consider the question in order to assess the effect of EI's implementation dependence. We focus on one result in particular, Burden and Kimball's President/House OLS analysis presented in their Table 6, a model that estimates the importance of some contextual effects on the second stage EI estimate of those voters who split their tickets between President Bush and Democratic House candidates in 1988 (Burden and Kimball 1998: 539). We choose this one result for replication since split-ticket voting in 1988 was predominantly in the direction of a Republican presidential candidate and a Democratic House candidate.

In analyzing Burden and Kimball's OLS analysis, we would like to capture as much of the uncertainty in King's estimation process as we can. To minimize numeric inaccuracies, we run the analysis on the most current version available with the most accurate bivariate normal distribution option, a function that is not the default setting. To capture implementation dependence we run EI on three operating system platforms – Windows, HP-Unix, and Linux – and to capture robustness to instability, we introduce data perturbations (as above).[35] We then combine the estimation uncertainty of the EI estimates by following Rubin (1987) to create six imputations of the EI estimates, two for each platform; run Burden and Kimball's OLS analysis on each imputed data set; and then generate the "average" coefficients and standard errors across the data sets.[36] Multiple imputation is recommended by King (1997: 151) to capture simulation variation of estimates used in subsequent analysis, such as Burden and Kimball's OLS analysis.

---

[35] We introduce the input perturbation, as well as the cross-platform results, into the imputations primarily as a way to concisely incorporate the results of all the implementation specific uncertainties into a single analysis. In fact, as Table 2 shows, the effect of the input perturbation on the estimates is small compared to that of the cross-platform difference, and our conclusions in this section can stand on the latter alone.

[36] Following Rubin (1987), the point estimate of a coefficient (q) is simply the average of the point estimates across the multiple imputations (m) of the data, $\overline{q} = \frac{1}{m}\sum_{j=1}^{m} q_j$. The standard error is a combination of an average of the standard error of the point estimates, $SE(q_j)$, and the sample variance of the point estimates, $S_q^2 = \sum_{j=1}^{m} \frac{(q_j - \overline{q})^2}{m-1}$, and is given by $SE(q)^2 = \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + S_q^2\left(1 + \frac{1}{m}\right)$.

	Imputation							
	Windows		*HP-Unix*		*Linux*			
	Run 1	*Run 2*	*Run 1*	*Run 2*	*Run 1*	*Run 2*	*Overall Results*	*Published Results*
Democratic Incumbent	.111 (.015)	.106 (.015)	.106 (.014)	.106 (.014)	.089 (.026)	.084 (.026)	.100 (.024)	.107 (.015)
Spending Ratio	.331 (.020)	.337 (.021)	.324 (.020)	.322 (.020)	.374 (.039)	.377 (.038)	.344 (.040)	.350 (.021)
Ballot Format	-.027 (.008)	-.031 (.008)	-.025 (.008)	-.024 (.007)	-.021 (.013)	-.019 (.012)	-.024 (.011)	-.032 (.008)
South	.056 (.009)	.062 (.009)	.056 (.008)	.053 (.009)	.055 (.015)	.051 (.015)	.055 (.012)	.065 (.008)
Constant	.067 (.008)	.065 (.008)	.072 (.008)	.074 (.008)	.052 (.014)	.052 (.014)	.064 (.016)	.065 (.008)

Table 3. Replication of Burden and Kimball Table 6, President/House OLS analysis.

The results of our analysis are presented in Table 3. Generally, we see that the coefficients and standard errors for a run on any single imputation are close to the original Burden and Kimball results, though there are slight differences that extend beyond simulation variation. Note that since we are using the newest version of King's program available, and using the slowest and most accurate versions of cdfbvn, the variation across imputations is smaller than would be expected from normal use of the program or previous versions.[37]

In no case does our new analysis reverse the substantive conclusions of Burden and Kimball. However, we see that some refinement of Burden and Kimball's original point estimates can be made, and these refined estimates deviate somewhat from the published results. More importantly, we show that the original published results failed to capture the distribution of coefficients and standard errors that result from simulation variation and from other implementation uncertainties. For example, once these uncertainties are captured, the 'Ballot Format' variable is seen to be barely significant, whereas before it was strongly significant.[38]

Numerical Inaccuracy and Misspecification

If the model does not fit the data, the accuracy of its estimation is beside the point. In

[37] Our implementation analysis shows that users of EI in the default mode, or older versions, would expect to see a much greater degree of implementation variance. And we recommend that current users of EI upgrade, and that they use the accurate cdfbvn to check their results before publication.

[38] This is merely by way of example, since significance was reported in the original analysis. In general we believe that confidence intervals are more useful for inference than significance tests, and our imputation techniques is oriented towards this goal. (See Gill 1999, for a treatment of significance testing, and its shortcomings.)

addition, misspecification in the model may exacerbate numerical inaccuracies in estimation. A cautious researcher is admonished by the EI documentation to make use of the many diagnostic tools available in EI to check their results (we suspect that one would also have received at least some warnings when running any version of EI with Burden and Kimball's data).[39] Such checks should include at least: examination of the tomography plots and "fit" plots for fit, examination of the bounds plots for aggregation bias, and examination of the nonparametric plots for multimodality (See King, 1997, Chpt. 16, and the EI documentation).[40] Our examination of these diagnostics demonstrates that potential problems exist at both the first and second stage of the EI estimation.

[39]In all of our replications *using the current version of EI,* we were routinely greeted by a host of warnings: a warning against using a covariate without a prior, a warning that the Hessian was not positive definite and that a specialized method had been used to calculate the covariance matrix and the fit of the model should be examined carefully, and (often) warnings that the simulations were close to the distribution bound. None of these warnings was severe enough to signal an invalid result, but any one should have put a researcher on their guard. The original version of EI, although it performed the same actions that triggered warnings in later versions, produced only one warning consistently in our runs: a warning that simulations were close to the distribution bound It is impossible to reconstruct precisely, however, which warnings one would have received with other versions of EI, since only the original and current versions are available to us, and no others are archived, even by the author.

[40] King also recommends that one use the fit plot. We did not include this in our comparisons since it does not function when covariates are present unless _Eeta is set, which then alters the model estimated. (When we did set _Eeta to 3, as recommended, and double-checked a number of replications that apparently showed good fit, the new "fit" plots also fit well – so use of 'fit' would not affect our analysis.) In addition, we recommend examination of the profile plots of the likelihood. These were not, however, available in early versions of EI.

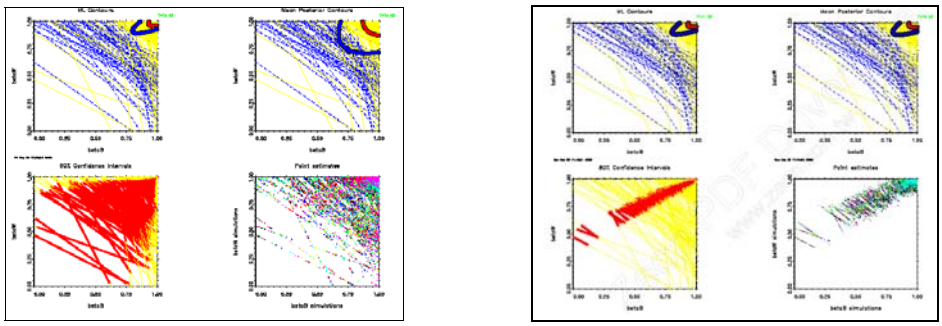

Figure 1: Two sets of four tomography plots for different replications of the Burden and Kimball model. In each set, the top left plot is a tomography plot plus likelihood contours, the top right plot is a tomography plot plus mean posterior contours, the lower left is a tomography plot with 80% confidence intervals, and the lower right plot is a density plot of simulated values. Representative plots using the current version of EzI with the default settings is on the left, on the right is EzI (v1.12).

King states, "It is *essential* to verify the model fits the data" (2000: 39, original emphasis). Two sets of diagnostics that King recommends to check, taken from a representative run of EzI (v.2.23) and from EzI (v.1.21), are presented in Figure 1. The set of diagnostics on the left show that something has gone wrong: Confidence intervals, likelihood contours, and mean posterior contours are all of supposed to roughly match each other clearly, but clearly do not. Interestingly, the problems with fit only appear in newer versions of EI, which use the generalized inverse cholesky to invert the Hessian. Older versions, such the representative plots presented on the right of Figure 1 and those available to Burden and Kimball, did not reveal the problems with fit in the model.

The first stage estimates pose a difficulty for EI because of the nature of the problem EI is attempting to solve. Recall that the first stage estimates are "roll-off," persons that voting in the presidential but did not vote in the congressional races. Since this is a small, but non-negligible, percentage of presidential voters, the mass of the tomography plot is centered in the upper right-hand corner. King and Benoit warn in the EzI FAQ that when the contours must fit a very small area of the tomography plot, calculations by cdfbvn can be imprecise (King and Benoit 1999). More generally, any change to the program that alters the distribution of coefficients when the distribution is centered close to the bounds may produce unexpected results. This is seen in our analysis of the implementation dependence of EI, where first stage estimates were more sensitive to changes in the cdfbvn algorithm and method of inverting the Hessian than second stage estimates, which are not centered near the bounds. Given the poor fit

in the first stage, the second stage estimates cannot be trusted.

Discussion

Our analysis shows that by properly replicating the Burden and Kimball analysis we confirm that implementation dependencies exist in across versions of the King's program. Through our improved analysis we are able to pinpoint that the primary causes, although perhaps not all, of the implementation dependence involves the method of inverting a non-positive definite Hessian and to changes in the cumulative bivariate normal distribution. When we use multiple imputation to incorporate the additional uncertainties in these estimates that arise from simulation variation, instability, and platform variation, we find that Burden and Kimball's primary substantive conclusions remain uncontradicted, but that some of the published point estimates can be refined, and that many of the reported standard errors are too small. Nevertheless, these implementation uncertainties are large enough to be of general concern since they lead to differences of up to 9.5 percentage points in the estimate of roll-off.

Moreover, by examining the diagnostic plots available in EzI, and running additional diagnostics ourselves, we discover evidence of uncontrolled violations of EI's distributional assumption in Burden's and Kimball's data. These distributional problems may exacerbate any instabilities in the statistical computation, and raises additional questions about the statistical fit of their model.

4. Avoiding Problems with Numerical Accuracy in EI

Check for Substantive and Statistical Plausibility. Look at the results and make sure that they make substantive sense and fit the data. Examine the data, results, intermediate output[41], return codes, and diagnostic plots for evidence of misspecification, outliers, or other signs of measurement or coding errors. Use other methods of ecological regression as a comparison.

These steps have two benefits. First, by checking for substantive and statistical plausibility, one is less likely to be misled if numerical accuracy does affect the model. Second, if, in fact, the model is correctly

[41] It is particularly important to check intermediate output for EI2 analyses using EZI, since improper convergence and other problems encountered during isolated earlier stages may not be returned in the result codes returned by the EZI program.

specified, and the data itself based on accurate measurements, any inaccuracy is likely to have a smaller impact on the measurements. These steps are not, however, a panacea – we were able to generate individually plausible results that showed no sign of problems in the diagnostics, yet were completely inaccurate.

Use High-Accuracy Options to Confirm Exploratory Results

As a direct result of this research, newer version of EI allow the user to select a high-quality cumulative bivariate normal distribution function. Although this is not the default, we recommend that this option be used to confirm EI results prior to publication. We also recommend that the CML options for more precise (central) derivatives be used, if this is not the default on your platform.

If EI result codes *or* intermediate output suggest any problems with the computed variance-covariance matrix, we suggest using several of the alternative algorithms to compute the Hessian matrix. The current default, although arguably the most sophisticated method, does not have strong basis in theoretical or practice that would justify its exclusive use.

In addition, those wishing to produce replicable work should record as much as possible about the details of statistical computation. Although EI's built-in replication facility, *eirepl*, replicates most of these EI-specific settings (with the exception noted in footnote XX, above), it omits platform related details (e.g. OS version, Gauss version, CML version, CML settings such as optimization method and tolerances, and Gauss settings, such as precision and tolerances) – which should be noted by the user.

Test Complex or Problematic Models for Accuracy. In problematic cases, where a model does give implausible results or where complex models or simulations are being used (especially for the first time on a particular problem), researchers should test for numerical inaccuracy explicitly**.** Ideally, when a new model is implemented, the author would create a comprehensive set of tests to which answers are known. Sometimes we can create such tests through Monte-Carlo simulation, by solving the problem closed-form, or by using high precision arithmetic in programs such as Maple or Mathematica. None of these options is available when the problem is sufficiently complex (although see McCue 2001, for an

alternative method of computing EI that makes use of more closed-form, and less numerically-intensive, solutions).

Where it is not possible to create such tests, the next best thing is to test the robustness of the simulation or estimation to differences in implementation. In section 4.2, we described a set of procedures that test the sensitivity of results to changes in PRNG's, small perturbations of input data, and changes in optimization methods. These are easy to implement, and can reveal a variety of numerical issues. (Example code is available from our web site.)

5. Conclusion

We have bad news, good news, and some cautious optimism. First, the bad news: we show that in the accuracy of statistical packages and other computational details can severely affect complex statistical models such as ecological inference. Failure to understand the limitations in statistical packages, and the details of their implementation, can lead to incorrect inferences.[42]

Now the good news: When we, as a discipline, pay attention to numerical accuracy, statistical software companies (and individual authors) respond: In response to an earlier version of this article, King has updated EI as well, to correct some of the inaccuracies that we uncovered.

Social scientists know that the details of collecting and analyzing data are important, but rarely pay attention to one aspect of their analysis, the numerical accuracy of their computational methods. As social scientists, we routinely deal with a host of other technical details that are irrelevant to our theories of social processes, but which we know are critical to a rigorous scientific analysis. For data gathering in survey research alone, these details might include sampling design, selection bias, question design,

[42] Understanding the limits of numeric algorithms is also important to how we use and interpret our analyses. For example, our exploration into numerical stability demonstrates that algorithms will often disagree on the third or fourth significant digit of a solution, including tests for statistical significance. The emphasis placed upon the 0.05 significance level has been repeatedly criticized for its arbitrariness since Fischer first proposed it in 1934 (e.g, Rosnow and Rosenthal 1989; Gill 1999). Numeric instability is

interviewer effects, inter-coder consistency, and respondent incentives. To analyze the resulting data requires further attention to technical details in statistics such as colinearity, measurement error, heteroskedasticity, misspecification, and missing data. Are we justified in ignoring numerical accuracy? And if not, does this add yet another layer of complexity to the process of social inquiry?

Our analysis of EI in this article leads us to argue that the answer to both questions is "no." Numerical accuracy can mislead social scientists that are caught by it unawares – so we must pay attention, if only to place our work and possible problems in the appropriate context. Surprisingly, this does not necessarily complicate our analyses. When researchers are cognizant of engineering principles for statistical software, and use appropriate tests, they can often detect, avoid, or fix problems of numerical inaccuracy. Since numerical inaccuracy potentially amplifies the effects of other statistical problems such as colinearity, measurement error, and misspecification; increasing accuracy can reduce these other problems, which might otherwise be addressable only through further data collection. We must remain cautious, however, and we must begin to test complex new methods more thoroughly.

---

another argument against bright-line tests of significance, since a t-statistic of 1.96 may contain significant numerical error.

References

Altman, Micah and Michael P. McDonald, 2002. "Replication with Attention to Numerical Accuracy", Political Analysis (with Michael P. McDonald)

Altman, Micah and Michael P. McDonald, 2001. "Choosing Reliable Statistical Software." *, PS: Political Science & Politics* 24(3): 681-8

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman, ``Estimation and Inference in Non-linear Structural Models", `Annals of Economic and Social Measurement', vol. 3, no. 4 (1974), pp. 653-665.

Burden, Barry C. and David C. Kimball. 1998. "A New Approach to the Study of Ticket Splitting." *American Political Science Review* 92 (September): 533-544.

Cho, W.K.T. 1998. Iff the assumption fits…: A comment on the King ecological inference solution. *Political Analysis* 7, 143-163.

Cook, R. Dennis, 1986 "Assessment of Local Influence", *J.R. Statist. Soc. B.* 48(2): 133-69.

Drezner, Z. and G.O. Wesolowsky 1989, "On the computation of the bivariate normal integral", *Journal of Statistical Computation & Simulation* 35: 101-107

Efron, Bradley, and Robert J. Tibshirani, 1993. *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Ferree, Karen E. 1999. ""Iterative Approaches to R x C Ecological Inference Problems: Where They Can Go Wrong." Presented at 1999 Summer Methodology Conference.

Fishman, George S., 1996. *Monte Carlo*: *Concepts, Algorithms and Applications*, Springer Verlag: New York.

Freedman, D.A., Klein, S., Ostland, M. and M. Roberts. 1998. Book review of "A Solution to the Ecological Inference Problem." *Journal of the American Statistical Association* 93, 1518-1522.

Freedman, D.A., Ostland, M., Roberts, M. and S. Klein. 1999. Response to King's comments. *Journal of the* American Statistical Association 94, 355-357

Freeman, Karen 1997, " Statistician Builds What May Be a Better Data Mousetrap", *New York Times*, July 8, 1997, *Science Times* Section, Page C8

Garey, Micheal R. and David S. Johnson. 1983. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman and Company.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. New York: Springer-Verlag.

Gill, Jeff, 1999. The Insignificance of Null Hypothesis Significance Testing. Political Research Quarterly, September 1999, p.647-674

Gill, Jeff, and Gary King. 2000. "'Hessian Not Invertible' Help! A Generalized Inverse/Generalized Cholesky Alternative to Model Respecification in Nonlinear Estimation." Working Paper.

Gill, Phillip E., Walter Murray and Margaret H. Wright, 1981. *Practical Optimization*, San Diego: Academic Press, Inc.

Higham, Nicholas J., 1996. *Accuracy and Stability of Numerical Algorithms*. Philadelphia: SIAM.

Jackman, Simon, "Estimation and Inference via Bayesian Simulation," *American Journal of Political Science* 44(2): 375-402

Johnston, Ron, 2001. "On geographers and ecological inference" *Annals of the Association of American Geographers* 91(2): 281, 283

King 1989, Unifying Statistical Methodology, Ann Arbor: University of Michigan Press.

King, Gary and Kenneth Benoit. 1999. "EzI: A Program for Ecological Inference, Frequently Asked Questions." Cambridge: Harvard University.

King, Gary, 1995. "Replication, Replication" [with comments from nineteen authors and a response], *PS: Political Science and Politics* 28(3): 443-499.

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

King, Gary, Micheal Tomz, and Jason Wittenberg, "Making the Most of Statistical Analyses," *American Journal of Political Science*, Vol. 44, No. 2, April 2000, Pp. 341–355

King, Gary. 2000. "EI: A Program for Ecological Inference, Version 1.63." Cambridge: Harvard University.

Knuth, Donald E., 1997. *The Art of Computer Programming, (3rd Edition)*. Reading, MA: Addison-Wesley.

Kousser, Morgan, 2002. "Ecological Inference from Goodman to King", *Historical Methods* (Forthcoming).

L'Ecuyer, Pierre and Peter Hellekalek, 1998. "Random Number Generators: Selection Criteria and Testing" in Peter Hellekalek (Ed.), *Random and Quasi-Random Point Sets*. Springer: New York.

St. Laurent, Roy T. and R Dennis Cook 1993, "Leverage, Local Influence and Curvature in Nonlinear Regression", *Biometrika* 80(1): 99-106.

Lewis, Jeffrey B., 2002, "Understanding King's ecological inference model: A method-of-moments approach", *Historical Methods* (Forthcoming).

McCue, Kenneth, 2001. "The Statistical Foundations of the 'EI' Method," *The American Statistician.* 55(2):106-11

McCullough, Bruce D. 1998. "Assessing the Reliability of Statistical Software: Part I,", *The American Statistician* 52(4): 358-366.

McCullough, Bruce D. 1999b. "Assessing the Reliability of Statistical Software: Part II,", *The American Statistician* 53(2): 149-159.

McCullough, Bruce D. and H. D. Vinod, 1999. "The Numerical Reliability of Econometric Software" *Journal of Economic Literature* 37: 633-665.

Michalewicz, Zbigniew, and David B. Fogel, 2000. *How to Solve It: Modern Heuristics,* Berlin, Springer-Verlag.

Mooney, Christopher Z. and George A. Krause. 1998. "Review Article: Of Silicon and Political Science – Computationally Intensive Techniques of Statistical Estimation and Inference" British Journal of Political Science 27, 83-110.

Neumann, Peter G., 1995. *Computer Related Risks*. Reading: Addison-Wesley.

O'Loughlin, John (2001). "Can King's Ecological Inference Method Answer a Social Scientific Puzzle: Who Voted for the Nazi Party in Weimar Germany?", *Association of American* (90) 3:592-602.

Palmquist, Bradley M, 2002, " Unlocking the Aggregate Data Past", *Historical Methods*, Forthcoming.

Ripley, B.D. 1990. “Thoughts on Pseudorandom Number Generators,” *Journal of Computational Applied Math* 31(1), 153-163.

Robert B. Schnabel and Elizabeth Eskow. 1990. "A New Modified Cholesky Factorization," *SIAM Journal of Scientific Statistical Computation*, 11(6): 1136-58.

Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Selke, W.; A.L. Talapov; and L.N. Schur (1994), “Cluster-Flipping Monte Carlo Algorithm and Correlations in ‘good’ random number generators”, JETP Letters 58. 665-8.

Sui, Daniel, 2001. "New Directions in Ecological Inference: An Introduction" *Annals of the Association of American Geographers* 90(3): 579-83.

Thornburg v. *Gingles*, 478 U.S. 30 (1986)

Vattulainen, L.; T. Ala Nissila, and K. Kankalla 1994. “Physical tests for random numbers in Simulations,” *Physical Review of Letters 73, 2513-1516.*

Wolpert, David H., and William G. MacReady, 1997. “No Free Lunch Theorems for Optimization”, *IEEE Transactions on Evolutionary Computation* 1(1): 67-82.

Yemma, John 1997 "Ideas", *Boston Globe Magazine* , Sunday, July 27, 1997.