Advance Access publication December 8, 2007

Political Analysis (2008) 16:70–92
doi:10.1093/pan/mpm030

Ignoramus, Ignorabimus? On Uncertainty in Ecological Inference

Martin Elff
Faculty of Social Sciences, University of Mannheim, A5, 6, 68131 Mannheim, Germany
e-mail: elff@sowi.uni-mannheim.de (corresponding author)

Thomas Gschwend
Center for Doctoral Studies in Social and Behavioral Sciences, University of Mannheim, D7, 27, 68131 Mannheim, Germany
e-mail: gschwend@uni-mannheim.de

Ron J. Johnston
School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK
e-mail: r.johnston@bristol.ac.uk

Models of ecological inference (EI) have to rely on crucial assumptions about the individual-level data-generating process, which cannot be tested because of the unavailability of these data. However, these assumptions may be violated by the unknown data and this may lead to serious bias of estimates and predictions. The amount of bias, however, cannot be assessed without information that is unavailable in typical applications of EI. We therefore construct a model that at least approximately accounts for the additional, nonsampling error that may result from possible bias incurred by an EI procedure, a model that builds on the Principle of Maximum Entropy. By means of a systematic simulation experiment, we examine the performance of prediction intervals based on this second-stage Maximum Entropy model. The results of this simulation study suggest that these prediction intervals are at least approximately correct if all possible configurations of the unknown data are taken into account. Finally, we apply our method to a real-world example, where we actually know the true values and are able to assess the performance of our method: the prediction of district-level percentages of split-ticket voting in the 1996 General Election of New Zealand. It turns out that in 95.5% of the New Zealand voting districts, the actual percentage of split-ticket votes lies inside the 95% prediction intervals constructed by our method.

1 Introduction

Many students and practitioners of social science have remained skeptical about the feasibility of sound ecological inference (EI). To some, the terms "ecological inference"

Authors' note: We thank three anonymous reviewers for helpful comments and suggestions on earlier versions of this paper. An appendix giving some technical background information concerning our proposed method, as well as data, *R* code, and *C* code to replicate analyses presented in this paper are available from the *Political Analysis* Web site. Later versions of the code will be packaged into an *R* library and made publicly available on *CRAN* (http://cran.r-project.org) and on the corresponding author's Web site.

and "ecological fallacy" appear almost synonymous: They doubt whether it will be possible at all to draw any conclusions about the behavior of individuals from aggregate data. There seem to be good reasons for such skepticism: EI has to rely on certain assumptions about the data-generating process at the level of individuals that cannot directly be tested, simply because the data on which such tests could be based are unavailable to begin with. The burden of assumptions becomes especially visible in a recent survey of Bayesian approaches to EI for 2×2 tables given by Wakefield (2004), which shows that even Markov chain Monte Carlo (MCMC) approaches are faced with this indeterminacy. Wakefield considers a total of 13 different variants of prior distributions for use in MCMC analysis of 2×2 tables including King's original truncated normal prior, beta exponential, and Student logistic gamma compound prior distributions and compares them with respect to prediction bias for King's (1997) Louisiana party registration data. Although there is one prior distribution that performs best in this application, one may still ask whether this result can be generalized to all possible EI applications to 2×2 tables. The question remains, as Fienberg and Robert aptly remark in their comments to Wakefield's review, "to what extent can we really distinguish between the fit of different models, hierarchical or otherwise, when only aggregate data are available?" (Fienberg and Robert 2004, 432).

Nonetheless, without certain restrictive assumptions about the process generating the unknown data, it is impossible to obtain any estimates for an EI problem, however preliminary these estimates may be. It should be noted that the challenge of EI is only a special case of the wider class of ill-posed inverse problems (King 1997). The task of EI is solving a problem that is *inverse* insofar as only a set of summaries of the data of interest are given and *ill posed* insofar as the information given by these summaries is not sufficient to identify a solution. Problems of this kind abound in the technical and scientific literature and numerous approaches to their solution have been proposed (see e.g., Groetsch 1993). Therefore, we think that a general rejection of EI procedures would be premature. Although the assumptions inherent in an EI procedure cannot be tested by means of statistical techniques, it is still possible to delimit the potential error that is associated with predictions from such a procedure. Constructing bounds to this potential error is the aim of the present paper. We derive a method to construct "robust" prediction intervals, that is, intervals that contain the true values of the unknown data with (at least approximately) known probability. Further, we assess the performance of these prediction intervals by means of simulation and a real-world example, the reconstruction of split-ticket votes in the 1996 General Election of New Zealand. As point of departure, we take an EI procedure recently presented in this journal, the entropy-maximizing approach of Johnston and Pattie (2000). When recast into a probability model, the Johnston-Pattie model imposes relatively mild and clearly structured restrictions on the unknown data-generating process and thus suits well the exploration of the consequences of model departures. As it will turn out, both in the simulation study and in the application to ticket splitting in New Zealand, the prediction intervals that we construct have a coverage that is almost identical to their nominal level.

The paper is organized as follows: In Section 2, we explain the fundamental dilemma of EI, which results from the necessity of EI procedures that employ certain restrictive assumptions that cannot be tested without the aid of the very data which are unavailable in a typical EI application. In Section 3, we propose a second-stage correction of the error distribution of EI estimates based on the Principle of Maximum Entropy and report the results of a simulation study to assess the performance of the proposed method. In Section 4, we discuss how this method can be adapted to cases where some of the data on which EI is based do not come from population-level aggregates but from a survey sample. In

Section 5, we illustrate our method with its application to split-ticket voting in the 1996 General Election of New Zealand. Section 6 discusses the limits of our proposed method, whereas Section 7 summarizes our results.[1]

2 A Basic Dilemma of EI

The situation of EI can be compared to reconstructing the "inner workings" of a "black box." These inner workings may be, for example, the numbers x_{ijk} of members from various ethnic groups ($i = 1, \ldots, I$) who do or do not turn out to vote ($j = 1, \ldots, J$) in voting districts ($k = 1, \ldots, K$) or the probabilities $\Pr(X_{ijk} = x_{ijk})$ in which these counts may occur. If the total sum of the counts is n, for example, if the total population of eligible voters is n, then the total number of possible configurations of counts x_{ijk} that sum to n can be expressed as the binomial coefficient $\binom{n + IJK - 1}{IJK - 1}$. In typical instances of EI, this is a vast number: Even in a country with only 1 million ($=n$) voters and three ($=I$) ethnic groups, whose members may or may not turn out to vote ($J = 2$) in one of 100 ($=K$) voting districts, there are

$$\binom{n + IJK - 1}{IJK - 1} = \binom{10^6 + 599}{599} \approx 10^{2189} \tag{1}$$

possible ways to arrange the voters into the black box.

In the absence of any information about the marginal sums $n_{.jk} = \sum_i x_{ijk}$, $n_{i.k} = \sum_j x_{ijk}$, or $n_{ij.} = \sum_k x_{ijk}$, that is, if only the total sum n of the counts x_{ijk} is known and if one has to make a point prediction about what configuration of counts is present inside the black box, it seems that one cannot do better than pick any of the $\binom{n + IJK - 1}{IJK - 1}$ possible configurations at random.[2] Such a random pick can be represented by a Uniform distribution on all possible configurations of counts $\mathfrak{x}=(x_{ijk})$ in the $(I \times J \times K)$-array that have a total sum of n. The probability of any specific configuration of being chosen then is $1 \Big/ \binom{n + IJK - 1}{IJK - 1}$. Also, the probability of hitting the true configuration of counts $\mathfrak{x}^* = (x^*_{ijk})$ by accident is also $1 \Big/ \binom{n + IJK - 1}{IJK - 1}$. Thus, one could also think of this true configuration as the outcome of a random variable $\mathfrak{X}=(X_{ijk})$ that has arrays $\mathfrak{x}=(x_{ijk})$ as values and has a Uniform distribution on its values.

Now, information about marginal sums of the cell counts, for example, the number of members of ethnic groups within each voting district, can vastly reduce the number of possible cell counts one has to consider. If we consider the case (for reasons of simplicity if not of plausibility) that the observed turnout rate in all voting districts is 50%, then one has

[1] An appendix, available online on the *Political Analysis* Web site, contains supplemental information: some background on maximum entropy distributions, properties of the Dirichlet and Dirichlet-multinomial distributions relevant for the argument of our paper, details on the computation of Beta-binomial prediction intervals, and some details about two nonparametric alternatives to the method proposed in our paper. The *Political Analysis* Web site also contains source code in *R* and *C* as well as data suitable for the replication of the analyses presented here.

[2] A note on notation: We abbreviate $\sum_{i=1}^{I}$, etc., as $\sum_i$ and $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K}$ as $\sum_{i,j,k}$ if the summation limits are understood.

only to consider, given a combination of voting district k and turnout status j, the number of possible triples of numbers that sum to $n/JK = 10^6/200$, that is,

$$\binom{n/(JK) + J - 1}{J - 1} = \binom{10^6/200 + 3 - 1}{3 - 1} \approx 10^{7.1}, \tag{2}$$

a number that is several orders of magnitude smaller than the number of possible configurations if no information about the marginal counts was available. This suggests that predictions about the unknown cell counts that take into account the marginal tables can have a greatly improved performance as compared to predictions that do not. EI methods can be considered as attempts to make such improved predictions. Yet, these methods are plagued by a serious problem, which we will expose in the following.

Statistical inference usually is made with the intention to describe a population based on a sample. For example, one may try to identify a model of voting behavior that connects voters' decisions to their own and the candidates' policy positions. If a correctly specified model is found, it can be used to make out-of-sample predictions, which may be predictions about future states of affairs or contrafactual states of affairs. One can also find both types of usage of results of EI. As an example of the first type of usage, Burden and Kimball (1998) try to find how much and why American voters engage in split-ticket voting between Presidential and Congress elections based on aggregated Presidential and Congressional votes for American voting districts. An example of the other type of usage is predicting the effects of changing the boundaries of voting districts on the representation of racial groups in these districts and on the chances of Democratic and Republican candidates in these rearranged districts (Cirincione, Darling, and O'Rourke 2000). As with statistical inference, both usages of EI need correctly specified models that describe the population. In contrast to statistical inference, EI is confronted with two problems that together pose a dilemma. The first problem is that of modeling indeterminacy: If only aggregates of the variables of interest are observed, there will always be more than one model of an interrelation between these variables that even fits perfectly to the observed aggregates. The second problem is that of inferential indeterminacy: If one arrives at identifying a model describing the interrelation of the variables of interest, this model will entail certain restrictive assumptions about the population, which cannot be tested based on aggregate data alone, which usually are the only data available. Hence the dilemma: If the first problem is solved, the second one is inevitably encountered. If one tries to avoid the second problem, one cannot solve the first one.

The challenge of the first problem lies in finding assumptions suitably restrictive for model identification. The criteria for suitability may vary with the application, but in most cases one will use assumptions that are plausible on the one hand and convenient on the other insofar as they lead to simple models for which estimation is feasible. However, plausibility and simplicity may conflict.

Consider the case in which someone has aggregate data on turnout and the proportion of African-Americans at the level of voting districts and wants to find out whether African-Americans differ from other citizens with respect to turnout. In this case, a model that presupposes that in each voting district turnout and race are conditionally independent will perfectly fit to the aggregate data and thus cannot be improved based on these data alone: Let $n_{1\cdot k}$ denote the number of African-Americans eligible to vote in district k, $n_{2\cdot k}$ the number of other citizens eligible to vote, $n_{\cdot 1k}$ the number of citizens eligible who actually turn out to vote, and $n_{\cdot 2k}$ the number of citizens eligible who do not turn out to vote in district k. Further, let $p_{ij|k}$ be the probability that an eligible citizen in district k is an

African-American ($i = 1$) and turns out to vote ($j = 1$), is an African-American ($i = 1$) and does not turn out to vote ($j = 2$), is not an African-American ($i = 2$) and turns out to vote ($j = 1$), or is not an African-American ($i = 2$) and does not out to vote ($j = 1$) and let x_{ijk} be the actual number of African-Americans/others who turn out/do not turn out to vote in district k. We thus have $n_{i\cdot k} = \sum_j x_{ijk}$ and $n_{\cdot jk} = \sum_i x_{ijk}$.

Then, a model that states

$$p_{ij|k} = p_{i\cdot|k} p_{\cdot j|k} \tag{3}$$

can be fitted to the aggregate data using the maximum likelihood estimates

$$\hat{p}_{ij|k} = \hat{p}_{i\cdot|k}\hat{p}_{\cdot j|k} = \frac{n_{i\cdot k}}{\sum_i n_{i\cdot k}} \frac{n_{\cdot jk}}{\sum_j n_{\cdot jk}}. \tag{4}$$

This model implies that within each race group i in district k, the probability to turn out to vote is equal:

$$p_{j|ik} = \frac{p_{ij|k}}{\sum_j p_{ij|k}} = \frac{p_{i\cdot|k} p_{\cdot j|k}}{\sum_j p_{i\cdot|k} p_{\cdot j|k}} = \frac{p_{\cdot j|k}}{\sum_j p_{\cdot j|k}}, \tag{5}$$

that is, turnout is unrelated to race. Of course, such a model seems implausible given the fact that survey data indicate that turnout and race are statistically related (e.g., Abramson and Claggett 1984). But this independence assumption cannot be tested based on the available aggregate data alone. On the other hand, any model that poses that race and turnout to be related, that is, $\hat{p}^*_{ij|k} = c_{ij}\hat{p}_{i\cdot|k}\hat{p}_{\cdot j|k}$ with $c_{ij} \neq 1$, $\sum_i c_{ij}\hat{p}_{ij|k} = \hat{p}_{i\cdot|k}$, and $\sum_i c_{ij}\hat{p}_{ij|k} = \hat{p}_{\cdot j|k}$, is empirically indistinguishable from the independence model. Thus, it seems that based on aggregate data alone, one cannot decide whether race and turnout are related or not.

Ecological regression (Goodman 1953, 1959) aims to overcome this problem by using a model that is in some ways more but in other ways less restrictive. In the case of race and turnout, it requires that the conditional probabilities of turnout are different across race groups but are the same in all voting districts, that is, $p_{ij|k} = p_{j|i}p_{\cdot j|k}$. Some refined ecological regression models even relax this assumption; they allow the conditional probability to vary across districts according to some known properties of the districts, that is, $p_{ij|k} = p_{j|ik}p_{i\cdot|k}$ and $p_{j|ik} = f(\beta_{ij0} + \beta_{ij1}z_{1k} + \cdots + \beta_{ijD}z_{Dk})$, where $z_{1k}, \ldots, z_{Dk}$ are the properties of the districts and $\beta_1, \ldots, \beta_D$ are the parameters to be estimated. In contrast to the conditional independence model discussed above, ecological regression models can, to some degree, empirically be tested. With maximum likelihood estimates $\hat{p}_{j|ik} = f(\hat{\beta}_{ij0} + \hat{\beta}_{ij1}z_{1k} + \cdots + \hat{\beta}_{ijD}z_{Dk})$ and $\hat{p}_{i\cdot|k} = n_{i\cdot k}/\sum_i n_{i\cdot k}$, it is still logically possible that predicted turnout numbers per district $\hat{n}_{\cdot jk} = \sum_j \hat{p}_{j|ik} n_{i\cdot k}$ differ from actually observed turnout numbers $n_{\cdot jk}$. Thus, in case of a poor fit between observed and predicted rates of eligible citizens who turn out in each of the districts (indicated by k), one may conclude that the ecological regression model is wrong and some of its assumptions have to be lifted. To account for departures of observed counts in a marginal table, such as the observed turnout per voting district, from predicted turnout, various authors extend this ecological regression by a random component: In this extended ecological regression model, conditional probabilities $p_{j|ik}$ are not fixed parameters but outcomes of a random variable. King proposes that these conditional probabilities have a truncated normal

distribution. Brown and Payne (1986) and Rosen et al. (2001) use the more natural assumption that the conditional probabilities have a Dirichlet distribution with mean

$$E(p_{j|ik}) : = \pi_{j|ik} = \frac{\exp(\beta_{ij0} + \beta_{ij1} z_{1k} + \cdots + \beta_{ijD} z_{Dk})}{\sum_s \exp(\beta_{is0} + \beta_{is1} z_{1k} + \cdots + \beta_{isD} z_{Dk})} \tag{6}$$

and precision parameter θ.

As Goodman (1953) notes, ecological regression models are suitable only if it is reasonable to assume that there is a causal relation between the properties corresponding to the marginal tables. But based on the marginal tables alone, that is, the district-level aggregates, it is possible neither to establish such a causal relation nor to disprove it: As just noted, in ecological regression models there may be a lack of fit between observed and predicted counts in the marginal table of turnout per district. The independence model (3) will thus almost always seem superior to any ecological regression model. But again, the fact that the marginal tables can perfectly be fitted by the independence model does not imply that, for example, turnout and race are unrelated since there are infinitely many models that pose a relation between race and turnout which may also perfectly fit to the marginal tables.

Statistical relations between variables represented by the observed marginal tables are often suggested by survey samples. For example, there is evidence from election studies (e.g., Abramson and Claggett 1984) that there is a clear statistical association between race and turnout. Therefore, it is worthwhile to combine evidence from aggregate data with evidence from survey data. An EI model that allows for this is the Entropy-Maximizing model proposed by Johnston and Pattie (2000) in an earlier issue of this journal. In contrast to the models just discussed, Johnston and Pattie aim to directly make predictions about the unknown counts x_{ijk} without the need of any statistical model. They show that a model of cell counts that are the most likely subject to the constraints

$$\sum_j x_{ijk} = n_{i\cdot k}, \quad \sum_i x_{ijk} = n_{\cdot jk}, \quad \text{and} \quad \sum_k x_{ijk} = n \frac{m_{ij}}{m} \tag{7}$$

has the log-linear form

$$\log \tilde{x}_{ijk} = \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \tau. \tag{8}$$

They also show that its parameters can be estimated by an iterative proportional scaling procedure.

Although Johnston and Pattie (2000, 337) state that their proposed procedure is "mathematical rather than statistical" and that therefore "no error terms" are attached to the predicted cell counts, this is not the case: Readers familiar with log-linear contingency table analysis (Fienberg, Holland, and Bishop 1977; King 1998) will realize that the Johnston-Pattie model is actually a log-linear model for counts without three-way interactions. However, the correct form of this model is

$$\log \mu_{ijk} = \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \tau, \tag{9}$$

where μ_{ijk} is the *mean* of a Poisson distribution, that is, the counts may indeed vary around this mean. The iterative scaling algorithm, which Johnston and Pattie propose, is the one usually employed to find maximum likelihood estimates for such log-linear models. That is, the "maximum likelihood solution" of Johnston and Pattie (2000, 335) is in fact the

maximum likelihood estimates of the means of Poisson distributions that have the structure of equation (9). If the sum $\sum_{i,j,k} x_{ijk} = \sum_{i,j,k} \mu_{ijk}$ is known in advance as n, which is usually the case in contingency table analysis and in EI, the distribution of the cell counts can also be modeled as a multinomial distribution with cell probabilities

$$p_{ijk} = \frac{\exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})}. \tag{10}$$

Good (1963) has shown that finding the multinomial distribution that maximizes the entropy $-\sum_{i,j,k} p_{ijk} \ln p_{ijk}$ subject to the constraints

$$\sum_j p_{ijk} = \frac{n_{i\cdot k}}{n}, \quad \sum_i p_{ijk} = \frac{n_{\cdot jk}}{n}, \quad \text{and} \quad \sum_k p_{ijk} = \frac{m_{ij}}{m} \tag{11}$$

leads to this model,[3] which requires the absence of three-way interactions among the cell probabilities p_{ijk}. The absence of three-way interactions among the cell probabilities p_{ijk} implies that the odds ratios $\left(p_{i_1 j_1 k}/p_{i_1 j_2 k}\right)/\left(p_{i_2 j_1 k}/p_{i_2 j_1 k}\right)$ are equal for all k. This imposes milder restrictions on the cell probabilities than the independence model or a ecological regression model without covariates since constant odds ratios still allow for variation of the conditional probabilities $p_{j|ik}$.

More recently, Judge, Miller, and Cho (2004) have developed an EI model based on information theoretic considerations, along similar lines as Good (1963) and Johnston and Pattie (2000). In contrast to these, they construct a model of the *conditional* cell probabilities $p_{j|ik}$, which requires information only about two of the three marginal tables, similarly to ecological regression and the conditional independence model.

The restrictions imposed by the various models so far discussed in this section can be compared easily by bringing the model specifications in logit form: For the independence model, we have

$$p_{ijk} = \frac{\exp(\beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} \exp(\beta_{rt} + \gamma_{st})} \text{ with } \exp(\beta_{ik}) \propto n_{i\cdot k} \text{ and } \exp(\gamma_{jk}) \propto n_{\cdot jk}; \tag{12}$$

for the ecological regression model (without district-level covariates), we have

$$p_{ijk} = \frac{\exp(\alpha_{ij} + \gamma_{jk})}{\sum_{r,s,t} \exp(\alpha_{rs} + \gamma_{st})} \text{ with } p_{j|i} = \frac{\exp(\alpha_{ij})}{\sum_s \exp(\alpha_{is})} \text{ and } \exp(\gamma_{jk}) \propto n_{\cdot jk}; \tag{13}$$

whereas for the Johnston-Pattie multinomial model, we have

$$p_{ijk} = \frac{\exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})}. \tag{14}$$

All these are special submodels of the *saturated* model

$$p_{ijk} = \frac{\exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk} + \delta_{ijk})}{\sum_{r,s,t} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st} + \delta_{rst})}, \tag{15}$$

[3]More on the application of information theoretic concepts to contingency table analysis and to statistics in general can be found in Kullback (1959). The argumentation of Johnston and Pattie (2000), however, follows a nonprobabilistic interpretation of entropy mentioned by Jaynes (1968).

which does not pose any restrictions on the structure of the cell probabilities. The complete-data log-likelihood of this model is

$$\ell = \sum_{i,j,k} x_{ijk} \log p_{ijk} = \sum_{i,j} n_{ij\cdot}\alpha_{ij} + \sum_{i,k} n_{i\cdot k}\beta_{ik} + \sum_{j,k} n_{\cdot jk}\gamma_{jk} + \sum_{i,j,k} x_{ijk}\delta_{ijk} - n\log\left(\sum_{i,j,k} \exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk} + \delta_{ijk})\right). \tag{16}$$

The implications of this expansion of the complete-data log-likelihood are quite ambivalent: If, for example, the assumptions inherent in the independence model apply, that is, $\alpha_{ij} = 0$ and $\delta_{ijk} = 0$ for all i, j, and k, the complete-data log-likelihood depends only on the observed aggregates $n_{i\cdot k}$ and $n_{j\cdot k}$ and the parameters of the independence model can be estimated by direct maximum likelihood. Also, if the assumptions of the Johnston-Pattie model apply and odds ratios are equal in all districts, that is, the three-way interaction parameters δ_{ijk} are zero for all i, j, and k, the complete-data log-likelihood depends only on the aggregates $n_{ij\cdot}$, $n_{i\cdot k}$, and $n_{j\cdot k}$. No such conclusion can be drawn with respect to ecological regression: Ecological regression models are used if the aggregate table $n_{ij\cdot}$ is not available. However, ecological regression models set parameters $\beta_{i\cdot k}$ and δ_{ijk} to zero but not α_{ij}. Therefore, the complete-data log-likelihood for ecological regression is *not* a special case of equation (16), but still it requires assumptions that are untestable without access to the complete array of counts x_{ijk}. Thus, any of the models so far discussed requires assumptions that can be tested only if the complete data x_{ijk} are available. In EI problems, however, they are not.

This problem, which may be called *inferential indeterminacy*, is much more serious than the fact that distributional assumptions of, for example, ecological regression models of King (King 1997; King, Rosen, and Tanner 1999; Rosen *et al.* 2001) or Brown and Payne (1986) cannot be checked or that under certain circumstances an ecological regression model like King's may be susceptible to "aggregation bias" (Openshaw and Taylor 1979, 1981; Cho 1998; Steel, Beh, and Chambers 2004). For example, one may want to make predictions about the number of African-Americans who will turn out to vote in a specific voting district. If one uses one of the models for EI discussed above and the assumptions inherent in this model do not hold, the predictions will be *systematically* biased, irrespective of the specific estimation procedure one uses: Suppose, for example, that in a specific application of EI, the complete data are generated from a multinomial distribution with cell probabilities p^*_{ijk} as in equation (15) and all structural parameters α^*_{ij}, β^*_{ik}, γ^*_{jk}, and δ^*_{ijk} are nonzero. Suppose further that one uses, for example, a procedure based on the Johnston-Pattie model (8) for EI. If n approaches infinity, the cell proportions x_{ijk}/n will converge to p^*_{ijk} and the proportions $n_{ij\cdot}/n$, $n_{i\cdot k}/n$, and $n_{\cdot jk}/n$ in the marginal tables will converge to $p^*_{ij\cdot} = \sum_k p^*_{ijk}$, $p^*_{i\cdot k} = \sum_j p^*_{ijk}$, and $p^*_{\cdot jk} = \sum_i p^*_{ijk}$, respectively. Further, the parameters of equation (8) will converge to values $\tilde{\alpha}_{ij}$, $\tilde{\beta}_{ik}$, and $\tilde{\gamma}_{jk}$ that maximize the scaled expected log-likelihood

$$\begin{aligned} n^{-1}\tilde{\ell} &= \sum_{i,j,k} p^*_{ijk}\log\tilde{p}_{ijk} = \sum_{i,j,k} p^*_{ijk}\log\left(\frac{\exp(\tilde{\alpha}_{ij} + \tilde{\beta}_{ik} + \tilde{\gamma}_{jk})}{\sum_{r,s,t}\exp(\tilde{\alpha}_{rs} + \tilde{\beta}_{rt} + \tilde{\gamma}_{st})}\right) \\ &= \sum_{i,j} p^*_{ij\cdot}\tilde{\alpha}_{ij} + \sum_{i,k} p^*_{i\cdot k}\tilde{\beta}_{ik} + \sum_{j,k} p^*_{\cdot jk}\tilde{\gamma}_{jk} - \log\left(\sum_{i,j,k}\exp(\tilde{\alpha}_{ij} + \tilde{\beta}_{ik} + \tilde{\gamma}_{jk})\right). \end{aligned} \tag{17}$$

Now even if $\tilde{\alpha}_{ij}$, $\tilde{\beta}_{ik}$, and $\tilde{\gamma}_{jk}$ are equal or very close to the corresponding parameters α^*_{ij}, β^*_{ik}, and γ^*_{jk}, $\tilde{p}_{ijk}$ and p^*_{ijk} will in general be different because of the nonzero δ^*_{ijk}. Consequently, estimates of cell probabilities based on the Johnston-Pattie model will have asymptotic bias $\tilde{p}_{ijk} - p^*_{ijk}$ and thus are not consistent. However, in order to obtain an estimate of this bias, one will need an estimate of δ^*_{ijk}, which is unavailable because only the marginal tables are known.

This is what makes the dilemma posed at the beginning of this section so serious: Without making certain identifying assumptions, one will not arrive, for example, at a prediction about the number of African-Americans who turn out to vote at all, apart from a random guess. The identifying assumptions, however, have crucial behavioral implications and if these assumptions are wrong, one will incur biased predictions. But these assumptions cannot be tested without access to the complete data, which are unavailable to begin with. Therefore, in any specific instance, one will not know how large this bias actually is.

3 Accounting for Inferential Uncertainty: A Maximum Entropy Approach

Estimating cell probabilities with the help of one of the models discussed in the previous section exhausts the resources for statistical inference. The models used for estimating the cell probabilities all employ certain restrictive assumptions that are necessary for the identification of cell probability estimates. Yet, as long as one is faced with the task of EI, these assumptions cannot be tested. They could be tested as a statistical hypothesis only if the complete data were available. But then, the problem no longer would be one of EI. This dilemma seems to sustain a skepticism with regard to the validity of EI. However, we would argue that a healthy amount of skepticism does not force us to give up the attempt at EI altogether. Doing so would mean ignoring the information contained in the marginal tables; it would mean throwing out the baby with the bathwater. On the other hand, it is clear that one should not put the same confidence in predictions from an EI model as one would in predictions from a well-tested statistical model.

In the present section, we propose a model that allows taking into account the uncertainty associated with estimates obtained from an EI. Since the resources for statistical inference are already exhausted, the model we propose cannot be justified in terms of theoretical statistics. All we can do is to appeal to some general principles that have some plausibility. The principle on which our proposition is based is the Principle of Maximum Entropy, which is a generalization of the (in)famous Laplace's Principle of Indifference. Before we present our proposed model, we need to explain what the Principle of Maximum Entropy entails and in what perspective we hold it to be plausible.

Laplace's Principle of Indifference, of which the Principle of Maximum Entropy is a generalization (Uffink 1995), postulates that one should assign to each elementary outcome $\{x_1\}, \ldots, \{x_n\}$ of a probability experiment, in absence of any prior information, the same probability $1/n$. For example, if the experiment is throwing a dice, one should assign to the outcome of each number one, two, three, four, five, or six the same probability 1/6.

The Principle of Maximum Entropy generalizes this principle to cases where some prior information of the probability distribution in question is available and where the probability distribution may be continuous with infinite support. It postulates that, if only some moments of the probability distribution (i.e., some nonrandom functions of the probability distribution, like, e.g., mean and variance) are given in advance, one should select the probability distribution that has maximal entropy among a set of probability distributions

with the same support with the given moments.[4] This principle leads to some common families of probability distributions such as the family of normal distributions or the family of exponential distributions. For example, the normal distribution with zero mean and variance σ^2 has maximal entropy of all continuous distributions over the real line with zero mean and variance σ^2. The exponential distribution with parameter λ has maximal entropy of all continuous distributions over the positive half real line with mean λ^{-1} (Shannon 1948).

The Maximum Entropy Principle has been used to specify "reasonable" null hypotheses for contingency table analysis (Good 1963; Golan, Judge, and Perloff 1996), to specify noninformative priors for Bayesian inference (Jaynes 1968), and for proposing solutions to ill-posed inverse problems (Vardi and Lee 1993). But it has also been used for EI (Johnston and Hay 1983; Johnston and Pattie 2000; Judge, Miller, and Cho 2004). Here, we use the Principle of Maximum Entropy to motivate and construct a second-stage probability model to account for inferential uncertainty.

Suppose (x_{ijk}) is an $(I \times J \times K)$-array of counts generated by a multinomial distribution with size index n (the "population size") and cell probabilities p^*_{ijk}. Suppose, further, that one has knowledge only about the marginal tables of this array and tries to make predictions about the cell counts. Using an EI method as discussed in the previous section, one may arrive at an estimate $\hat{p}_{ijk}$ for the cell probabilities. The model by which the cell probabilities are estimated may or may not be correctly specified, that is, the "true" cell probabilities p^*_{ijk} may or may not satisfy the constraints inherent in the model. Based on the EI procedure of choice, one makes the prediction $n\hat{p}_{ijk}$ about the cell count x_{ijk}. Then, the error of prediction of x_{ijk} by $\hat{p}_{ijk}$ can be decomposed as follows:

$$x_{ijk} - n\hat{p}_{ijk} = n\left(\frac{x_{ijk}}{n} - p^*_{ijk}\right) + n(p^*_{ijk} - \hat{p}_{ijk}). \tag{18}$$

The difference $(x_{ijk}/n) - p^*_{ijk}$ will have mean zero and variance $p^*_{ijk}(1 - p^*_{ijk})/n$ and thus will be the smaller the larger n is, while the difference $p^*_{ijk} - \hat{p}_{ijk}$ will not become smaller unless the model leading to $\hat{p}_{ijk}$ is correctly specified, that is, if the assumptions of the EI model employed are satisfied by p^*_{ijk}. In the analysis of bias at the end of the last section, we held p^*_{ijk} fixed and considered the estimator $\hat{p}_{ijk}$ as random. In contrast, we now propose to change the roles of $\hat{p}_{ijk}$ and p^*_{ijk}, that is, to treat $\hat{p}_{ijk}$ as fixed, in virtue of being a function of the known marginal tables, and p^*_{ijk} as the realization of a random variable P_{ijk}.

If we are completely ignorant about the data array (x_{ijk}) except for the total sum n, then any possible array of numbers (p_{ijk}) with $0 < p_{ijk} < 1$ and $\sum_{i,j,k} p_{ijk} = 1$ would seem equally plausible as having generated the unknown array of counts. We can represent this, according to Laplace's Principle of Indifference, by a probability distribution with a density function that is Uniform for all admissible arrays (p_{ijk}). Now, if each of the n individuals in the $(I \times J \times K)$-array falls into the cell (i, j, k) with probability P_{ijk}, which is an element of a random array (P_{ijk}) with a Uniform distribution, then each possible array (x_{ijk}) that may result from such a process has the same chance of occurrence. That is, there is a direct connection between the perspective that focuses on our ignorance about the counts x_{ijk} and the perspective that focuses on our ignorance about the data-generating process.

Taking this as a baseline, we can construct a distribution of plausible arrays (p_{ijk}) that reflects our ignorance about the true cell probabilities (p^*_{ijk}) that generated the unknown

[4] Although there is some relation between the two, "entropy" refers here to a functional of density or probability mass functions and not to entropy in the sense of statistical mechanics and thermodynamics. For the relation, see Jaynes (1957). There are several attempts at giving this principle a general axiomatic foundation, for example, Jaynes (1957), Vasicek (1980), and Cziszar (1991).

data (x_{ijk})—an ignorance that is only reduced by the information contained in the marginal tables $(n_{ij\cdot}) = \left(\sum_k x_{ijk}\right)$, $(n_{i\cdot k}) = \left(\sum_j x_{ijk}\right)$, and $(n_{\cdot jk}) = \left(\sum_i x_{ijk}\right)$ and recovered by the EI. Such a distribution should be as similar to the Uniform distribution described in the previous paragraph as possible under the restriction that its mean is given by the estimates produced by the EI method used. The Uniform distribution under consideration is a special case of a Dirichlet distribution, that is, a Dirichlet distribution with all shape parameters equal to one. Now, if we select a Dirichlet distribution that maximizes entropy under the constraint that its mean is equal to the estimates obtained from an EI procedure, then we have the distribution that is, under these constraints, the most similar to the Uniform distribution in terms of the Kullback-Leibler criterion for the similarity of distributions.[5] Informally speaking, the maximum entropy criterion leads here to the "flattest," that is least informative distribution with the given mean.

The selection of such an entropy-maximizing Dirichlet distribution is possible since by its mean a Dirichlet distribution is only partially specified: If a multidimensional random variable (P_{ijk}) has a Dirichlet distribution with parameters θ_{ijk}, its components have expectations $\pi_{ijk} := E(P_{ijk}) = \frac{\theta_{ijk}}{\theta_0}$, where $\theta_0 := \sum_{r,s,t} \theta_{rst}$. Therefore, each parameter θ_{ijk} of a Dirichlet distribution can be decomposed into a mean parameter π_{ijk} and a common scale parameter θ_0 by $\theta_{ijk} = \pi_{ijk}\theta_0$, so that an entropy-maximizing Dirichlet distribution with mean π_{ijk} fixed at $\hat{p}_{ijk}$ can be identified by maximizing

$$H_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{i,j,k} \ln\Gamma(\theta_0\hat{p}_{ijk}) - \ln\Gamma(\theta_0) + (\theta_0 - IJK)\psi(\theta_0) - \sum_{i,j,k} (\theta_0\hat{p}_{ijk} - 1)\psi(\theta_0\hat{p}_{ijk}) \tag{19}$$

for θ_0, where $\psi(\cdot)$ is the digamma function (Abramovitz and Stegun 1964, 258).[6]

Since the family of Dirichlet distributions is a multivariate generalization of the family of Beta distributions, we can use this family of distributions to illustrate what finding a maximum entropy distribution entails. A Beta distribution is usually characterized by its two shape parameters, which we call here ϕ_1 and ϕ_2. The mean of this distribution then is $\pi := \phi_1/(\phi_1 + \phi_2)$, so if we define $\theta_0 := \phi_1 + \phi_2$, the shape parameters can be reexpressed as $\phi_1 = \theta_0\pi$ and $\phi_2 = \theta_0(1 - \pi)$ and the variance as $\pi(1 - \pi)/(1 + \theta_0)$. Figures 1 and 2 depict how a Beta distribution will look like for π fixed at 0.5 and 0.2, respectively, for various values of θ_0 below, above, and equal to θ_{MaxEnt}, where θ_{MaxEnt} denotes the value of θ_0 for which the entropy of the Beta distribution is maximal. As Fig. 1 shows, the Uniform distribution over the interval [0, 1] is a special case of a Beta distribution: the Beta distribution with maximal entropy subject to the constraint that the expectation is equal to 0.5. Both figures indicate that, irrespective of the value of the expectation of the distribution, values of θ_0 above θ_{MaxEnt} lead to single-peaked densities, where the peak is the higher the larger θ_0 is. However, as θ_0 gets smaller, the density function puts more and more weight on values around zero and one. Since the variance approaches a supremum of $\pi(1 - \pi)$ as θ_0 approaches zero, the variance of a Beta density is not a good measure of uncertainty. Conversely, since the weight of the density is most evenly distributed if the θ_0 attains the entropy-maximizing value, the entropy seems to be a much better measure of uncertainty if the expectation π is fixed.

If each of the n individuals in the $(I \times J \times K)$-array falls into the cell (i, j, k) with probability P_{ijk}, where P_{ijk} itself is part of a random array with a Dirichlet distribution with

[5]For details see Appendix (Section A.1) to this paper on the *Political Analysis* Web site.
[6]For a formal proof of the validity of this formula, see Appendix (Section B.2) at the *Political Analysis* Web site.

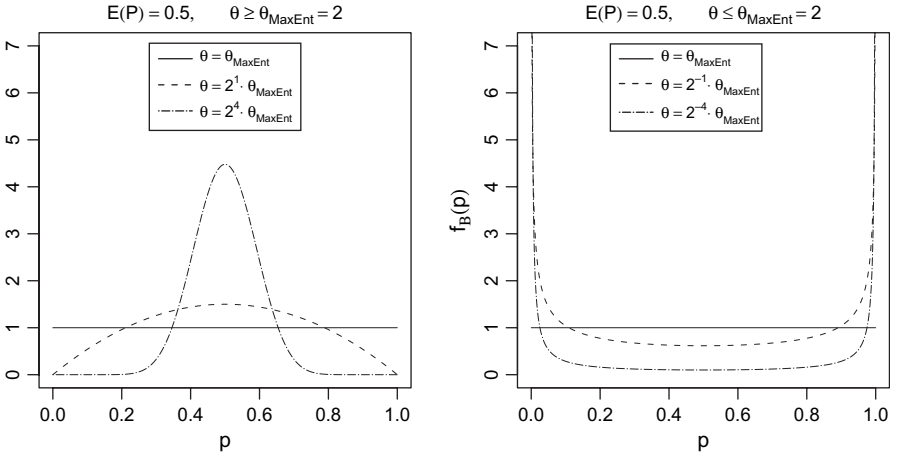

Fig. 1 Density of Beta distributions with mean held fixed at 0.5 and various values for the parameter θ_0.

parameters $\theta_{ijk} = \theta_0\pi_{ijk}$, then the distribution of the possible resulting arrays (x_{ijk}) is a mixture of a multinomial distribution with a Dirichlet distribution, a compound-multinomial or Dirichlet-multinomial distribution (Mosimann 1962; Hoadley 1969). Just as the family of Dirichlet distributions is a multivariate generalization of the family of Beta distributions, the family of Dirichlet-multinomial distributions is a multivariate generalization of the family of Beta-binomial distributions, which has been used to model overdispersed proportions and success counts (Skellam 1948; Crowder 1978; Prentice 1986). In fact, if an array of random variables (P_{ijk}) has a joint Dirichlet distribution with parameter array (θ_{ijk}), each component of the array has a Beta distribution with parameters $\theta_0\phi_{ijk}$ and $\theta_0(1 - \pi_{ijk})$ and each component of the corresponding array of counts (x_{ijk}) with a Dirichlet-multinomial distribution has a Beta-binomial distribution with parameters $\theta_0\phi_{ijk}$ and $\theta_0(1 - \pi_{ijk})$ (Mosimann 1962), where, as before, $\theta_0 = \sum_{i,j,k} \theta_{ijk}$. Now, the counts in each cell have expectation $n\pi_{ijk}$ and variance $n\pi_{ijk}(1 - \pi_{ijk})(n + \theta_0)/(1 + \theta_0)$. That is, the

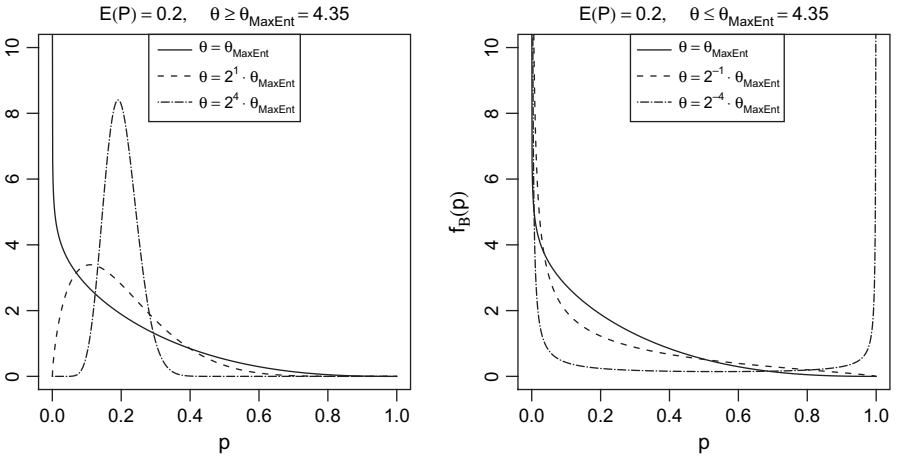

Fig. 2 Density of Beta distributions with mean held fixed at 0.2 and various values for the parameter θ_0.

expectation is the same as if the counts had a multinomial distribution with cell probabilities π_{ijk}, whereas the variance differs from the variance of such a multinomial distribution by the factor $(n + \theta_0)/(1 + \theta_0)$. Although proportions X_{ijk}/n of counts X_{ijk} with binomial distribution are asymptotically normal and although arrays of proportions X_{ijk}/n of arrays of counts X_{ijk} with multinomial distribution are asymptotically multivariate normal as n approaches infinity, it will be a mistake to assume asymptotic normality in the case of counts that have a Beta-binomial or Dirichlet-multinomial distribution, respectively: The distribution of the proportions will converge to a Beta distribution or Dirichlet distribution, respectively, which can have a shape quite dissimilar to the normal distribution if it has maximal entropy.

The above considerations make clear that one cannot assume that the asymptotic distribution obtained from an EI procedure is normal. Therefore, one should not use the normality assumption to construct confidence or prediction intervals based on the standard errors of the estimates. Rather, we propose to use the quantile function of the Beta distribution to construct approximate credibility intervals for the cell probabilities p_{ijk} and the quantile function of the Beta-binomial distribution to construct approximate prediction intervals for the cell counts x_{ijk}. If the counts in cell (i, j, k) have a Beta-binomial distribution with parameters $\theta_0\pi_{ijk}$ and $\theta_0(1 - \pi_{ijk})$, 95% prediction intervals would be, for example, delimited by $F^{-1}_{\mathrm{Bb}}(0.025; \theta_0, \pi_{ijk})$ and $F^{-1}_{\mathrm{Bb}}(0.975; \theta_0, \pi_{ijk}) + 1$, where $F^{-1}_{\mathrm{Bb}}(\alpha; \theta_0, \pi_{ijk}) := \sup\{x : \quad F_{\mathrm{Bb};\theta_0,\pi_{ijk}}(x) < \alpha\}$ is the quantile function and $F_{\mathrm{Bb}}(x; \theta_0, \pi_{ijk}) := \sum_{t=0}^{x} f_{\mathrm{Bb}}(t; \theta_0, \pi_{ijk})$ is the cumulative distribution function of the Beta-binomial distribution with parameters $\theta_0\pi_{ijk}$ and $\theta_0(1 - \pi_{ijk})$. That is, under the assumption that the counts in cell (i, j, k) have this Beta-binomial distribution, the probability that the counts are in these intervals will be 95%.[7]

Since the reasoning behind this procedure is rather heuristic than justifiable in terms of theoretical statistics or probability theory, it seems necessary to assess its performance by way of a simulation study. Therefore, we conducted a systematic simulation experiment in which we vary the size of the array of counts (x_{ijk}) and its total sum n. For each considered array size $I \times J \times K$ and each "population size" n, we (1) generated 2000 arrays of random numbers $(x^{(r)}_{ijk})$ with $r = 1, \ldots, 2000$, (2) computed the marginal tables $(n^{(r)}_{ij\cdot})$, $(n^{(r)}_{i\cdot k})$, and $(n^{(r)}_{\cdot jk})$, and (3) used the Johnston-Pattie procedure to generate estimates of cell probabilities $(\hat{p}^{(r)}_{ijk})$. We then (4) constructed prediction intervals for the cell counts based on the assumption of the Johnston-Pattie model that the cell counts have, jointly, a multinomial distribution and, individually, a binomial distribution with success probability $\hat{p}^{(r)}_{ijk}$ and (5) prediction intervals based on the procedure proposed in the present section. We then (6) recorded whether the counts $x^{(r)}_{ijk}$ are covered by the respective prediction intervals, that is, whether they are inside the intervals. The random counts are generated such that each array of counts (x_{ijk}) that sums to n has the same chance of occurrence. Thus, the generated arrays of counts are a *simple random sample* from all such arrays and the distribution of the generated counts represents the initial ignorance about the interior of the black box before a procedure of EI is applied. The simulation results may also be generalized such that they are representative for the average performance with respect to all possible interiors of the black box. By recording the performance of prediction intervals *both* based on assumptions of an EI procedure *and* based on the second-stage maximum entropy procedure, we are able, first, to demonstrate the consequences of the indeterminacy that besets EI and, second, to show the degree to which our proposed second-stage maximum entropy

[7]For details about the computation of Beta-binomial cumulative probability and quantile functions see Appendix (Section B.4) on the *Political Analysis* Web site.

Table 1 Simulation study of coverage of true cell counts and true cell probabilities after 2000 replications

		Population size	
	Array size	*100,000*	*10,000,000*
(a) Effective coverage by prediction intervals based on the assumption of a multinomial distribution	$3 \times 3 \times 50$	18.4	2.1
	$7 \times 7 \times 50$	28.0	2.6
	$3 \times 3 \times 200$	33.0	3.3
	$7 \times 7 \times 50$	54.1	6.0
(b) Effective coverage by prediction intervals based on the second-stage maximum entropy method	$3 \times 3 \times 50$	95.5	96.3
	$7 \times 7 \times 50$	93.1	94.1
	$3 \times 3 \times 200$	95.0	96.4
	$7 \times 7 \times 200$	88.5	93.8

procedure improves over a "naive" application of EI and represents the actual amount of uncertainty associated with EI procedures.

Table 1 presents the results of our simulation study regarding the coverage performance of 95% prediction intervals of these two types with respect to an arbitrary chosen cell in the respective arrays. Since the distribution from which cell counts are generated is symmetric with respect to the cells in the array, any cell is as representative for the whole set of counts in the array as any other cell. For convenience, we chose the cell with indices (1, 1, 1). Panel (a) of Table 1 reports the coverage performance of naive prediction intervals based on the assumption that the counts have a multinomial distribution with correctly specified cell probabilities, whereas panel (b) reports the coverage performance of prediction intervals constructed based on the second-stage maximum entropy method.

A comparison of the two panels (a) and (b) in Table 1 makes clear that prediction intervals are a large improvement in comparison to naive model–based prediction intervals. Panel (a) shows that the effective coverage of multinomial distribution–based prediction intervals that cover the true counts *decreases* with increasing population size. If the population size n is 10,000, the effective coverage may reach more than 50%, which still falls short from the nominal coverage of 95%. But if the population size is 10 million, the undercoverage of the naive prediction intervals is disastrous—in at most 6% of the cases are the true counts inside a 95% prediction interval. This is a direct consequence of the inconsistency of the estimates of the cell probabilities: The naive prediction intervals presuppose that the standard deviation of x_{ijk}, conditional on a given cell probability p_{ijk}, increases with the population size n only proportional to $\sqrt{n}$. Consequently, the length of the prediction intervals will decrease relative to the range of possible values of x_{ijk} (which is delimited by zero and n) proportional to $1/\sqrt{n}$. Due to the fact that by construction of the simulation experiment the cell probabilities will almost surely be misspecified, the actual root mean square error of predictions based on the cell estimates will also reflect the error incurred by misspecification. As n grows large, the effect of misspecification error obviously will dominate the random variation of the cell counts x_{ijk} around their expectation np_{ijk}.

In contrast, the maximum entropy Dirichlet-multinomial based prediction intervals show an effective coverage quite close to their nominal level, as appears in panel (b) of Table 1. That the prediction intervals based on the second-stage maximum entropy method still differ from their nominal level may have several reasons. These differences may just be a consequence of simulation error, in which case these differences disappear as the number of replications approaches infinity. However, these differences may also show that

our method is just an approximation. Our proposed method so far only takes into account the consequences of inferential indeterminacy but not the consequences of sampling variability. It takes into account that the true cell probabilities and true expected cell counts cannot be known completely even if n approaches infinity. It does not, however, take into account the fact that the cell probabilities are estimated on the basis of a finite n. It may thus be possible to improve on the coverage performance of the prediction intervals if a finite-n correction could be applied. However, this is beyond the scope of this paper since it is mainly concerned with the consequences of inferential indeterminacy. Another reason is that exact identity between effective coverage and nominal coverage would only be achievable if the counts were continuous and not discrete. With finite n, discrete values may just be too coarse. For example, for a population of size 100,000 and for $7 \times 7 \times 200 = 9800$ cells, the actual coverage of the cell counts falls clearly short. That may be caused by the fact that in this setting, the average counts per cell are less than four which is clearly below 100. Therefore, it will be almost impossible to obtain exact percentiles for such counts. Nevertheless, even if we admit that the proposed method is only an approximation, it works better than prediction intervals that rely on the identifiability assumptions of the EI procedure of Johnston and Pattie to hold.

4 Accounting for Sampling Variability

In the previous section, we considered the uncertainty about estimates that comes from the unavailability of the complete data. The observed data, however, were assumed to be *population-level* summaries. Probabilistic methods are employed in the previous sections, first, to take into account that even population-level counts are the outcome of some stochastic data-generating process and, second, to take into account the uncertainty about estimates based on these counts. If, however, at least one of the available marginal tables is not a population-level summary but comes from a sample, for example, from a survey sample with data on ethnic identity and voting behavior, another level of uncertainty is added.

For example, if instead of a population-level cross-tabulation of voting behavior and ethnic group membership only a cross-tabulation $\mathfrak{m}=(m_{ij})$ from a sample is available and the true counts in the array of combinations of ethnic group membership (with categories $i = 1, \ldots, I$), voting behavior (with categories $j = 1, \ldots, J$), and voting district (with running numbers $k = 1, \ldots, K$) are x_{ijk}, then (provided that we have a simple random sample) the array of counts m_{ij} has a multinomial distribution with cell probabilities $q_{ij} = n_{ij\cdot}/n$, where $n_{ij\cdot} = \sum_k x_{ijk}$. For given marginal tables $\mathfrak{n}_1=(n_{\cdot jk})$, $\mathfrak{n}_2=(n_{i\cdot k})$, and $\mathfrak{n}_3=(n_{ij\cdot})$, the method proposed in the previous section leads to one set of parameters $(\theta_{ijk})=\boldsymbol{\theta}(\mathfrak{n}_1,\mathfrak{n}_2,\mathfrak{n}_3)$ of a Dirichlet-multinomial model of the counts x_{ijk} in the array. But for a given sample cross-tabulation m_{ij}, there is more than one possible multinomial distribution with cell probabilities q_{ij} and thus more than one possible marginal population-level table $n_{ij\cdot}$ from which this sample may have been drawn. That is, there is more than one set of parameters to be considered for modeling the distribution of the unknown counts in the array $(\theta_{ijk})=\boldsymbol{\theta}(\mathfrak{n}_1,\mathfrak{n}_2,\mathfrak{n}_3)$.

Of course, given a sample cross-tabulation, not all possible $(I \times J)$ arrays $n_{ij\cdot}$ with $\sum_{i,j} n_{ij\cdot} = n$ are equally plausible candidates. Rather, the plausibility of these candidates is adequately expressed by the values $f(\mathfrak{n}_3|\mathfrak{m})=\Pr(\mathfrak{N}_3=\mathfrak{n}_3|\mathfrak{M}=\mathfrak{m})$ of the probability mass function of the conditional distribution of $\mathfrak{N}_3=(N_{ij\cdot})$ given the sample table $(M_{ij})=\mathfrak{M}$. Now, since $(N_{ij\cdot})$ is an unobserved random variable in this perspective, it is no longer possible to simply construct prediction intervals based on a quantile function

$F_{\mathrm{Bb}}^{-1}(\alpha; \theta_0, \pi_{ijk})$ whose parameters are computable from fixed, observed marginal tables $\mathfrak{n}_1$, $\mathfrak{n}_2$, and $\mathfrak{n}_3$. Rather, an appropriate quantile function is given by

$$F^{-1}(\alpha; \theta_0, \pi_{ijk} \mid \mathfrak{M}) = \sum_t F_{\mathrm{Bb}}^{-1}\left(\alpha; \theta_0, \pi_{ijk} \mid \mathfrak{n}_3^{(t)}\right) f\left(\mathfrak{n}^{(t)} \mid \mathfrak{m}\right), \tag{20}$$

where the sum is over all possible tables $\mathfrak{n}_3^{(t)} = (n_{ij\cdot}^{(t)})$ $(t = 1, \ldots)$ that satisfy $\sum_{i,j} n_{ij\cdot}^{(t)} = n$. This sum has no closed form, not the least because there is no function in closed form that leads from the marginal tables $\mathfrak{n}_1 = (n_{\cdot jk})$, $\mathfrak{n}_2 = (n_{i\cdot k})$, and $\mathfrak{n}_3 = (n_{ij\cdot})$ to θ_0 and π_{ijk} (these parameters have to be computed by numerical methods). Therefore, we propose a bootstrap method to construct the limits of prediction intervals, which involves the following steps:

1. Each replication $r = 1, \ldots, R$ starts by generating random counts from an approximation of the conditional distribution of (n_{ij}) given (m_{ij}). This is done by a double-bootstrap procedure. First, random counts $(m_{ij}^{(r)})$ from a multinomial distribution with size index m and cell probabilities $p_{ij}^{*} = m_{ij}/m$ are generated. From these random counts, a second, random set of cell probabilities $p_{ij}^{(r)} = m_{ij}^{(r)}/m$ is computed. These are the cell probabilities of a multinomial distribution with size index n from which a second set $(n_{ij\cdot}^{(r)})$ of counts is generated. This assures that the (random) marginal tables $(n_{ij\cdot}^{(r)})$ are integers from a multinomial distribution with size index n and reflect the variability of the observed sample (m_{ij}).
2. Based on the marginal tables $(n_{ij\cdot}^{(r)})$, $(n_{i\cdot k})$, and $(n_{\cdot jk})$, first-stage cell probability estimates $\hat{p}_{ijk}^{(r)}$ are obtained based on the Johnston-Pattie model for each replication $r = 1, \ldots, R$.
3. After setting $\pi_{ijk}^{(r)} = \hat{p}_{ijk}^{(r)}$ values, $\theta_0^{(r)}$ are determined for each r such that the Dirichlet distribution with parameters $\theta_{ijk}^{(r)} = \theta_0^{(r)} \pi_{ijk}^{(r)}$ has maximal entropy for given $\pi_{ijk}^{(r)}$.
4. Random numbers $p_{ijk}^{(r)}$ from the Dirichlet distribution with parameters $\theta_{ijk}^{(r)}$ are generated for each r.
5. For each r, random counts $x_{ijk}^{(r)}$ from a multinomial distribution with probability parameters $p_{ijk}^{(r)}$ are generated. For each r, the random array $(x_{ijk}^{(r)})$ has thus a Dirichlet-multinomial distribution, however, with different parameters $\theta_{ijk}^{(r)}$ for each r.

After R replications, the predictions of the unknown cell counts x_{ijk} are given by the averages $R^{-1} \sum_r x_{ijk}^{(r)}$ of the random counts for all i, j, and k, and the limits of the prediction interval for each cell count x_{ijk} are given by the respective quantiles of the random counts $x_{ijk}^{(r)}$. In the next section, we demonstrate the application of this procedure to the reconstruction of percentages of split-ticket votes in the 1996 General Election of New Zealand.

5 A Real-World Example: Split-Ticket Voting in New Zealand

The term "split-ticket voting" is a pattern of voting behavior that can emerge when voters have the opportunity to cast several votes on the same occasion. For example, American citizens have the opportunity to cast a vote both for a candidate who runs for the Presidency and for a candidate for the House of Representatives. In parallel and mixed-member electoral systems, voters may cast two votes in general elections, one with which they can choose the candidate to represent the voting district they live in and one with which they can choose which party they want to support for the proportional tier of the electoral system. Variants of mixed-member electoral systems can be found, for example, in legislative elections of Bolivia, Germany, New Zealand, and Venezuela; parallel voting systems can be found in general elections, for example, in Japan, Mexico, and South Korea.

Strategic voting accounts emphasize the effect of "Duverger's Law" on the amount of split-ticket voting: Voters will give their most preferred party their list vote but, if they expect that the district candidate of this party to have only little chances of electoral success, restrict their choice on those candidates they deem to be potentially successful, thus not to "waste their vote." Often, available empirical data do not suffice to test such accounts. Either there are only aggregate data on list and candidate votes in individual voting districts available or the survey data, if they are available, are too sparse with regard to candidate vote-list vote combinations in individual voting districts. Therefore, examining split-ticket voting is a typical field of application for EI (e.g., Burden and Kimball 1998; Gschwend, Johnston, and Pattie 2003; Benoit, Laver, and Gianetti 2004).

There are, however, some rare occasions where official district-level data are available not only on list vote and candidate vote results but also on the numbers of straight- and split-ticket votes. One of these occasions is the 1996 General Election of New Zealand. This gives us the opportunity to examine the performance of the methods developed in the two preceding sections in a "real-life" setting.

Three sorts of data are available on the 1996 General Election of New Zealand: (1) official data on electoral results for *party lists* and for *party candidates* and *independent candidates* for each of the 67 voting districts, (2) official data on total numbers of *straight-ticket* and *split-ticket* votes for each of the voting districts, and (3) an 8×8 table of combinations of list and candidate votes from a nation-level survey sample (Levine and Roberts 1997; Johnston and Pattie 2000). That is, whereas the district-level aggregates $\mathfrak{n}_1=(n_{\cdot jk})$ and $\mathfrak{n}_2=(n_{i\cdot k})$ are observed, the nation-level aggregate $\mathfrak{n}_3=(n_{ij\cdot})$ is not but only a sample cross-tabulation $\mathfrak{m}=(m_{ij})$. To make an approximately valid EI about the level of split-ticket voting in the individual voting districts, we thus need the bootstrap method developed in the preceding section.

In the context of research on split-ticket voting, one is usually not interested in all counts x_{ijk} in the array made of candidate votes, list votes, and voting districts. Rather, one is interested in the proportion of split-ticket and straight-ticket votes in the individual voting districts. This additional complication, however, can be tackled in a straightforward manner. The bootstrap method of the preceding section produces random counts $x_{ijk}^{(r)}$ ($r = 1, \ldots, R$) from which point and interval predictions of the true counts x_{ijk} are computed. Now, based on

$$f_{\text{straight},k}^{(r)} = \frac{\sum_{i=j} x_{ijk}^{(r)}}{\sum_{i,j} x_{ijk}^{(r)}} \quad \text{and} \quad f_{\text{split},k}^{(r)} = \frac{\sum_{i\neq j} x_{ijk}^{(r)}}{\sum_{i,j} x_{ijk}^{(r)}}, \tag{21}$$

we obtain random proportions of straight-ticket and split-ticket votes.[8] The averages $\frac{1}{R}\sum_r f_{\text{straight},k}^{(r)}$ and $\frac{1}{R}\sum_r f_{\text{split},k}^{(r)}$ can then serve as a point prediction of the proportions of straight-ticket and split-ticket votes, whereas the simulated quantiles can serve as limits of prediction intervals.

Figure 3 shows a comparison of actual against predicted percentages of ticket splitters in the voting districts of New Zealand along with 95% prediction intervals.[9] As can be seen

[8] A further note on notation: The expression $\sum_{i=j} x_{ijk}$ refers to the sum of all elements x_{ijk} in unit (voting district) k for which the first index (e.g., the vote for the candidate of party i) equals the second index (e.g., the vote for the list of party j). The expression $\sum_{i\neq j} x_{ijk}$ refers to the sum of all elements in which the first and the second index differ.

[9] Rather than showing an ordinary scatter plot of predicted against actual percentages, this plot contains a dot plot of predicted and actual percentages against the individual voting districts (sorted by the predicted percentages) to make it easier to discern the prediction intervals of cases with similar predicted or actual percentages of ticket splitters.

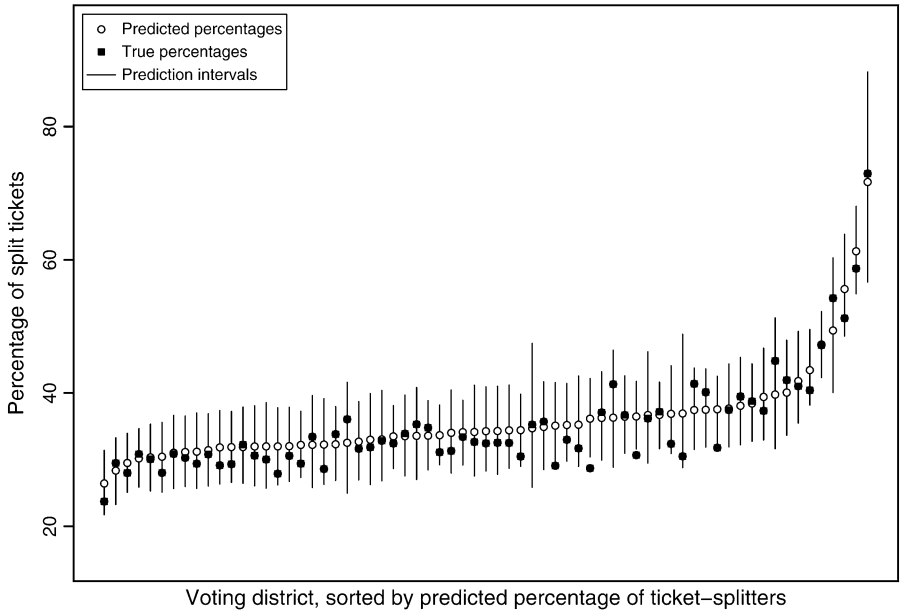

Fig. 3 Application of the second-stage maximum entropy approach to split-ticket voting in the 1996 General Election of New Zealand: true and predicted split-ticket percentages with prediction intervals based on bootstrap percentiles, after $R = 2000$ replications. The coverage of the true percentages is 95.5%.

in this plot, there are three instances in which the actual percentage of ticket splitters lies outside prediction intervals. The total coverage of the actual ticket-splitting percentages by these prediction intervals thus is 95.5%, which is close to their nominal level. This constitutes a slight overcoverage, but with only 67 voting districts, one can hardly expect to achieve a coverage exactly at nominal level.

The application to the case of split-ticket voting in the 1996 General Election of New Zealand is another corroboration of the method developed in the preceding sections. Although, again, the match between the nominal coverage of the prediction intervals and their effective coverage is not perfect, it is close enough to suggest that the methods are, if not a final solution to the problem, a substantial step in the right direction. However, the comparison of predicted and actual percentages of ticket splitting as well as the length of the prediction intervals shows that EI cannot produce the miracle of delivering predictions from mere aggregates that are comparable in quality to estimates and predictions obtained from fully observed data. In the New Zealand application, the predictions of percentages of ticket splitters are in some districts roughly 10 percentage points away from the actual percentages. Also, the length of the prediction intervals amounts to, on average, 15 percentage points. This makes again clear that the basic dilemma of EI should lead scholars to take greatest care when they interpret results obtained from an EI procedure.

6 Discussion

In Sections 3 and 4, we develop a method of accounting for the extra amount of uncertainty associated with estimates and predictions obtained from an EI procedure that stems from a problem that we exposed in Section 2, the problem of inferential indeterminacy. Without

the problem of inferential indeterminacy, it would be possible to model the data-generating process of the counts x_{ijk} as a multinomial distribution. To model the consequences of inferential indeterminacy for the uncertainty about cell probabilities, we use the conjugate family of the multinomial, the family of Dirichlet distributions. To model the consequences for uncertainty about cell counts, we use the family of mixtures of multinomial Dirichlet distributions.

The first to consider Dirichlet-multinomial mixture distributions for EI are Brown and Payne (1986). Their approach consists of an extension of the ecological regression model (12), which allows the conditional probabilities $p_{j|ik}$ to vary in different units k according to a Dirichlet distribution with mean parameters $E(p_{j|ik}) = \pi_{j|i}$ and precision parameter θ_0. According to Brown and Payne (1986), both the mean and the dispersion parameters can be estimated from the marginal tables $n_{\cdot jk}$ and $n_{i\cdot k}$, the mean parameters based on generalized least squares and the precision parameter based on the variance of the residuals of the regression of $n_{\cdot jk}$ on $n_{i\cdot k}$ (Rosen *et al.* [2001] consider estimation of this model via MCMC and a simpler model without a Dirichlet mixing distribution estimated via nonlinear least squares). Since the precision parameter θ_0 is estimated from the residuals of this regression, this model may account for model departures with respect to those aspects in which model (12) is more restrictive than the Johnston-Pattie model (8). But since the precision parameter of the model in Brown and Payne (1986) is estimated based on the observed marginal tables $(n_{i\cdot k})$ and $(n_{\cdot jk})$, the resulting Dirichlet model of the conditional probabilities can hardly account for those model departures that can be detected only if the full array (x_{ijk}) of counts is available. In contrast to the approach of Brown and Payne (1986), we do not try to estimate the precision parameter θ_0 from the observable marginal tables but rather determine its value according to an a priori criterion, the Principle of Maximum Entropy. As our simulation experiments show, the values of the precision parameter thus determined can capture much of the uncertainty caused by the possibility of undetectable model departures. Therefore, we suggest that any method of constructing prediction intervals for EI should lead to intervals at least as large as those based on our proposed method. Otherwise, they cannot account for those undetectable model departures that haunt the confidence in results of EI.

It seems that our method of accounting for the consequences of modeling indeterminacy itself rests on a crucial assumption that a priori, without taking into account the information contained in the marginal tables, any possible array of counts may occur with the same probability. This, however, would be a misunderstanding. The Uniform distribution reflects the ignorance about the true cell counts that characterizes the point of departure of EI problems. It plays a role similar to a "flat" or noninformative prior in Bayesian inference. Using an informative, non-Uniform distribution as a reference distribution may increase the risk of biased predictions unless this reference distribution is sufficiently close to the true distribution of the cell counts. We use a Uniform distribution as reference specifically to avoid such possible bias. If, however, prior information is available that can be summarized in a distribution of the cell probabilities, a generalization of the maximum entropy method can be used: One can then, instead of maximizing entropy, select a model that minimizes the directed Kullback-Leibler information divergence, also known as Kullback-Leibler information criterion, relative to this prior distribution.[10]

[10] As Kullback (1959) and Good (1963) have pointed out, the Principle of Maximum Entropy is just a special case of the Principle of Minimum Discriminating Information: Choosing the distribution with maximal entropy is equivalent to minimizing the directed Kullback-Leibler information divergence relative to a Uniform distribution. For details see Appendix (Section A) on the *Political Analysis* Web site.

As a method to construct a distribution of the cell counts, one may argue that the method developed in Section 3 is deficient insofar as it only corrects consequences of parametric assumptions of EI methods that could be bypassed by directly constructing a distribution of the unknown cell counts x_{ijk} that, without any prior parametric assumptions, maximizes entropy subject to those constraints that reflect the information contained in the known marginal tables. In principle, such a construction will be straightforward; in practice, however, finding a solution will be infeasible because of its prohibitive computational complexity.[11] The method developed in Section 3 contains some concepts of Bayesian statistical inference, that is, a probability distribution on the parameters (the cell probabilities) of a multinomial distribution, which is a member of the conjugate to the family of multinomial distributions. But it does not make use of Bayes's theorem. The combination of the Johnston-Pattie model with the second-stage maximum entropy construction may be viewed, at best, as an approximation to a posterior distribution of the cell counts. However, we are not able to show how good an approximation this is except by means of the simulation study, the results of which we report at the end of Section 3. Therefore, a direct Bayesian approach that makes use of the complete set of tools of this technique of statistical inference seems preferable to the approach proposed in this paper. As can be shown,[12] a posterior constructed by the straightforward application of Bayes's theorem to a noninformative prior distribution of the cell counts would be of surprisingly simple structure. However, the computation of the posterior probabilities of the counts will be as computationally demanding as a nonparametric maximum entropy model, making a direct Bayesian approach as unfeasible as a nonparametric maximum entropy approach to modeling the distribution of the cell counts.

Finally, we want to emphasize that one should not yield to the temptation of using cell counts or cell probabilities as "data" in second-stage regression models. Apart from the conceptual issues involved in fitting models to data predicted from another model and from possible bias incurred by using such second-stage regressions (Herron and Shotts 2003, 2004), conventionally estimated standard errors of a naive second-stage regression model will be highly inaccurate. Also, this problem cannot be cured by an increase in the number of spatial units considered or by the number of cases within the respective spatial units. Our simulation results reported at the end of Section 3 show how much one will be misled if one relies on asymptotic theory in these cases. Therefore, instead of treating estimates obtained from an EI as data, one should consider them as parameters of the predictive distribution for *unknown* data and use this distribution for generating values for multiple imputation (e.g., King et al. 2001). Of course, since the amount of unknown data is quite large relative to the amount of known data, one would need a large number of imputed data sets in order to get reliable estimates of quantities of interest and their variances. The issues connected with such a use of EI results clearly deserve some further research, which is, however, beyond the scope of this paper.

7 Conclusion

The point of departure of our paper is a fundamental dilemma of EI and of inference in ill-posed inverse problems in general: Estimates can be identified only if certain restrictive assumptions are made with respect to the structure of the data-generating process leading

[11]For details see Appendix (Section A.4) on the *Political Analysis* Web site.
[12]See Appendix (Section C) on the *Political Analysis* Web site.

to the unknown data one tries to reconstruct. However, these assumptions may not be satisfied by the unknown data, leading to serious bias in the estimates relative to the true values of the data. This may be called the dilemma of *fundamental indeterminacy* (Cho and Mansky forthcoming). We tackle this dilemma by proposing a method of delimiting the error one has to expect when the assumptions needed for the identification of a solution are violated by the unknown data.

We focus on a special model for EI, a version of the model proposed in Johnston and Pattie (2000), because this model requires only relatively mild assumptions as compared to those required by ecological regression models. We examine the consequences of possible departures of the model assumptions. We find that prediction intervals based on the model assumptions are far too narrow and show a very serious undercoverage of the unknown data. The main contribution of our paper, however, is the development of a method for the construction of prediction intervals that are at least approximately correct.

The method we propose consists in combining two stages. In the first stage, point estimates for the unknown data are constructed based on a model that contains certain assumptions inevitable for the identification of the solution. In the second stage, we consider the set of all possible solutions that satisfy or do not satisfy these identification assumptions. We construct a probability distribution that meets the requirement that its expectation is the solution from the first stage but is otherwise as neutral as possible with respect to other possible first-stage solutions. This distribution assigns the weights of possible solutions in terms of values of its density functions as equal as possible, that is, has maximal entropy, subject to the constraints on the expectation of this distribution. Results from a simulation experiment show that if the population for which the EI is to be made is large enough, prediction intervals based on our proposed two-stage method are approximately correct.

We supplement this simulation experiment with a real-world application: the prediction of district-level percentages of ticket splitting in the 1996 General Election of New Zealand based on aggregate data about candidate votes and list votes at the voting-district level and an 8×8 table obtained from a survey sample. This election is a rare opportunity to check the performance of an EI procedure against real data: Not only are discrete-level data available on candidate votes and list votes and a sample for candidate vote-list vote combinations at the national level but also district-level data on the percentages of straight-ticket and split-ticket votes. Therefore, we are able to compare predicted with actual percentages of split-ticket votes for each voting district. Within the context of this application, we address another issue that scholars dealing with aggregate data may face: Not always are these aggregate data exact summaries of the population but rather a sample that summarizes the population. As a solution for this problem, we propose a combination of our two-stage maximum entropy method with bootstrapping from the empirical distribution of the sample. Our application of this combination to the New Zealand data shows that it results in prediction intervals with a coverage performance roughly equal to their nominal level: In 95.5% of the New Zealand voting districts lies the actual percentage of split-ticket votes inside 95% prediction intervals constructed based on our proposed method.

The main conclusion of our paper thus is that standard errors and confidence intervals for EI problems that do not take into account the fundamental uncertainty associated with any solution to ill-posed inverse problems may be grossly misleading. However, the consequences of this fundamental uncertainty can be delimited, if not exactly, though approximately. Therefore, despite the problems discussed in this paper, to reject altogether the idea of EI on these grounds means throwing out the baby with the bathwater.

References

Abramovitz, Milton, and Irene A. Stegun, eds. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Washington, DC: National Bureau of Standards.

Abramson, Paul R., and William Claggett. 1984. Race-related differences in self-reported and validated turnout. *Journal of Politics* 46:719–38.

Benoit, Kenneth, Michael Laver, and Daniela Gianetti. 2004. Multiparty split-ticket voting estimation as an ecological inference problem. In *Ecological inference: New methodological strategies*, ed. Gary King, Ori Rosen, and Martin Tanner, 333–50. Cambridge, UK: Cambridge University Press.

Brown, Philip J., and Clive D. Payne. 1986. Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* 81:452–60.

Burden, Barry C., and David C. Kimball. 1998. A new approach to the study of ticket splitting. *American Political Science Review* 92:533–44.

Cho, Wendy K. Tam. 1998. Iff the assumption fits . . .: A comment on the King ecological inference solution. *Political Analysis* 7:143–63.

Cho, Wendy K. Tam, and Charles F. Manski. Forthcoming. Cross-level/ecological inference. In *Oxford handbook of political methodology*, ed. Janet Box-Steffensmeier, Henry Brady, and David Collier. Oxford, UK: Oxford University Press.

Cirincione, C., T. A. Darling, and T. G. O'Rourke. 2000. Assessing South Carolina's congressional districting. *Political Geography* 19:189–211.

Crowder, Martin J. 1978. Beta-binomial ANOVA for proportions. *Applied Statistics* 27:34–7.

Cziszar, Imre. 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics* 19:2032–66.

Fienberg, Stephen E., Paul W. Holland, and Yvonne Bishop. 1977. *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Fienberg, Stephen E., and Christian P. Robert. 2004. Comment to 'Ecological inference for 2×2 tables' by Jon Wakefield. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167:432–4.

Golan, Amos, George Judge, and Jeffrey M. Perloff. 1996. A maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association* 91:841–53.

Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics* 34:911–34.

Goodman, Leo A. 1953. Ecological regressions and the behavior of individuals. *American Sociological Review* 18:663–4.

———. 1959. Some alternatives to ecological correlation. *American Journal of Sociology* 64:610–25.

Groetsch, Charles W. 1993. *Inverse problems in the mathematical sciences*. Braunschweig and Wiesbaden: Vieweg.

Gschwend, Thomas, Ron Johnston, and Charles Pattie. 2003. Split-ticket patterns in mixed-member proportional election systems: Estimates and analyses of their spatial variation at the German federal election, 1998. *British Journal of Political Science* 33:109–27.

Herron, Michael C., and Kenneth W. Shotts. 2003. Using ecological inference point estimates as dependent variables in second-stage linear regressions. *Political Analysis* 11:44–64.

———. 2004. Logical inconsistency in EI-based second-stage regressions. *American Journal of Political Science* 48:172–83.

Hoadley, Bruce. 1969. The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association* 64:216–29.

Jaynes, Edwin T. 1957. Information theory and statistical mechanics. *Physical Review* 106:620–30.

———. 1968. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* 4:227–41.

Johnston, Ron J., and A. M. Hay. 1983. Voter transition probability estimates: An entropy-maximizing approach. *European Journal of Political Research* 11:405–22.

Johnston, Ron J., and Charles Pattie. 2000. Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting. *Political Analysis* 8:333–45.

Judge, George G., Douglas J. Miller, and Wendy K. Tam Cho. 2004. An information theoretic approach to ecological estimation and inference. In *Ecological inference: New methodological strategies*, ed. Gary King, Ori Rosen, and Martin Tanner, 162–87. Cambridge, UK: Cambridge University Press.

King, Gary. 1997. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton: Princeton University Press.

———. 1998. *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor, MI: Michigan University Press.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95:49–69.

King, Gary, Ori Rosen, and Martin A. Tanner. 1999. Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research* 28:61–90.

Kullback, Solomon. 1959. *Information theory and statistics*. New York: Wiley.

Levine, Stephen, and Nigel S. Roberts. 1997. Surveying the snark: Voting behaviour in the 1996 New Zealand general election. In *From campaign to coalition: New Zealand's first general election under proportional representation*, ed. Jonathan Boston, Stephen Levine, Elizabeth McLeay, and Nigel Roberts, 183–97. Palmerston North, NZ: Dunmore Press.

Mosimann, James E. 1962. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* 49:65–82.

Openshaw, S., and P. J. Taylor. 1979. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical methods in the spatial sciences*, ed. N. Wrigley, 127–44. London: Pion.

———. 1981. The modifiable areal unit problem. In *Quantitative geography: A British view*, ed. N. Wrigley and R. J. Bennett, 60–70. London: Routledge and Kegan Paul.

Prentice, R. L. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 81:321–7.

Rosen, Ori, Wenxing Jiang, Gary King, and Martin A. Tanner. 2001. Bayesian and frequentist inference for ecological inference: The $r \times c$ case. *Statistica Neerlandica* 55:134–56.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423, 623–56.

Skellam, J. G. 1948. A probability distribution derived from the binomial distribution by regarding the probability as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)* 10:257–61.

Steel, David G., Eric J. Beh, and Ray L. Chambers. 2004. The information in aggregate data. In *Ecological inference: New methodological strategies*, ed. Gary King, Ori Rosen, and Martin Tanner, 51–68. Cambridge, UK: Cambridge University Press.

Uffink, Jos. 1995. Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Modern Physics* 26B:223–61.

Vardi, Y., and D. Lee. 1993. From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society. Series B (Methodological)* 55: 569–612.

Vasicek, Oldrich Alfonso. 1980. A conditional law of large numbers. *Annals of Probability* 8:142–7.

Wakefield, Jon. 2004. Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167:385–426.