Research Article

Statistics in Medicine

Received 25 May 2010, Accepted 3 January 2011 Published online 22 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4214

Bayes computation for ecological inference

Jon Wakefield,[a,b*†] Sebastien Haneuse,[c] Adrian Dobra[a,d,e] and Elizabeth Teeple[f]

Ecological data are available at the level of the group, rather than at the level of the individual. The use of ecological data in spatial epidemiological investigations is particularly common. Although the computational methods described are more generally applicable, this paper concentrates on the situation in which the margins of 2×2 tables are observed in each of n geographical areas, with a Bayesian approach to inference. We consider auxiliary schemes that impute the missing data, and compare with a previously suggested normal approximation. The analysis of ecological data is subject to ecological bias, with the only reliable means of removing such bias being the addition of auxiliary individual-level information. Various schemes have been suggested for this supplementation, and we illustrate how the computational methods may be applied to the analysis of such enhanced data. The methods are illustrated using simulated data and two examples. In the first example, the ecological data are supplemented with a simple random sample of individual-level data, and in this example the normal approximation fails. In the second example case–control sampling provides the additional information. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: auxiliary data; case–control sampling; ecological bias; Markov chain Monte Carlo

1. Introduction

Ecological data, in which data are available for groups rather than individuals, are ubiquitous in many disciplines including epidemiology, the social sciences and education research [1, 2]. In this paper, we focus on applications in spatial epidemiology in which data are available across multiple areas. Within each of the areas, the observed data consist of the margins of a 2×2 table. As an example, in Section 5.1 we consider a situation in which we observe, across zip codes in Washington State, the number of women of white and non-white races and the number diagnosed with diabetes. Crucially, in the ecological data we do not observe the cross-classification of diabetes status by race and, hence, inference is fraught with difficulties, since there is an intrinsic non-identifiability problem, which can lead to ecological bias. The latter describes the difference between estimated parameters from individual- and ecological-level analyses. Understanding and characterizing ecological bias has a long history [3–7].

As illustrated by a number of authors, the only reliable method for removing ecological bias is to supplement the ecological data with individual-level information. The benefits of simple random sampling have been previously illustrated [8–12]. In the case of a rare event, random sampling will produce few cases. This prompted the development of the *aggregate data* method in which ecological data are combined with random subsamples of individuals who provide covariate information, but no individual response data [13]. Inference is carried out using estimating functions, with adjustment for the effects of subsampling. Again for rare diseases, an approach has been developed to combine ecological data with case–control samples gathered within areas, [14, 15], with computation via both maximum

[a] *Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
[b] *Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.*
[c] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, U.S.A.*
[d] *Department of Biobehavioral Nursing and Health Systems, Washington, Seattle, WA, U.S.A.*
[e] *Center for Statistics and the Social Sciences, University of Washington, Seattle, WA, U.S.A.*
[f] *Biostatistics Unit, Group Health Research Institute, Seattle, WA, U.S.A.*
**Correspondence to: Jon Wakefield, Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
† *E-mail: jonno@uw.edu*

likelihood (based on the EM algorithm) and an auxiliary variable scheme. Continuing this thread of research, two-phase sampling [16, 17] has been considered in the ecological context, with inference via maximum likelihood [18]. In this paper, we consider Bayesian computation for ecological data, with or without supplementary information. Computation is challenging for the models we consider. One popular approach depends on an analytic normal approximation, while another is computationally prohibitive in typical applications. Our contribution is to describe MCMC algorithms that are easy to implement in a range of different discrete data settings, and to demonstrate their use on both simulated data and on two illustrative data sets.

The structure of the paper is as follows. In Section 2 Bayesian models are described for three situations: ecological data only; ecological data plus individual data from a random sample; ecological data plus individual data from a case–control sample. Section 3 describes the methods that we propose for computation, while Section 4 compares the various algorithms on simulated data. Section 5 describes the use of the algorithm on two real data sets, and we conclude with a discussion in Section 6. Two appendices contain more technical material.

2. Model formulation

2.1. Ecological data

We let y_{i0} and y_{i1} represent the number of individuals with disease from populations of sizes N_{i0} and N_{i1}, respectively, in each of $i=1,\ldots,n$ areas. Table I summarizes notation. In different contexts the two populations, $j=0$ and $j=1$ might represent two races (Section 5.1), female and male (Section 5.2) or unexposed/exposed. In the ecological inference problem, y_{i0} and y_{i1} are unobserved, with the total, $y_{i+}=y_{i0}+y_{i1}$ only, being available, in addition to N_{i0} and N_{i1}.

If the individual-level data were observed then a starting model would be $y_{ij}|p_{ij} \sim \text{Binomial}(N_{ij}, p_{ij})$, for $j=0,1$. Letting $\boldsymbol{p}_i=(p_{i0}, p_{i1})$ and $\boldsymbol{N}_i=(N_{i0}, N_{i1})$, the distribution of the sum, y_{i+}, is the convolution

$$\Pr(y_{i+}|\boldsymbol{p}_i, \boldsymbol{N}_i) = \sum_{y_{i0}=l_i}^{u_i} \binom{N_{i0}}{y_{i0}} \binom{N_{i1}}{y_{i+}-y_{i0}} p_{i0}^{y_{i0}}(1-p_{i0})^{N_{i0}-y_{i0}} p_{i1}^{y_{i+}-y_{i0}}(1-p_{i1})^{N_{i1}-y_{i+}+y_{i0}}, \tag{1}$$

which we label as Binomial-Conv($\boldsymbol{N}_i, \boldsymbol{p}_i$), and where

$$l_i=\max(0, y_{i+}-N_{i1}), \quad u_i=\min(N_{i0}, y_{i+}) \tag{2}$$

gives the range of admissible values that y_{i0} can take, given the margins in Table I. This likelihood has been considered by a number of authors [9, 19, 20].

Many authors, [1, 9–12, 21], have suggested hierarchical models for the analysis of ecological data. Generically, (1) is combined with a second-stage model in which the distribution of $\boldsymbol{p}_i|\boldsymbol{\phi}$ is specified as a function of hyperparameters $\boldsymbol{\phi}$. At the third stage $\boldsymbol{\phi}$ are assigned a hyperprior.

For ecological analysis we consider a previously suggested three-stage hierarchical model [9]:

Stage 1. Likelihood: $y_{i+}|\boldsymbol{p}_i, \boldsymbol{N}_i \sim \text{Binomial-Conv}(\boldsymbol{N}_i, \boldsymbol{p}_i)$, $\quad i=1,\ldots,n$.
Stage 2. Random Effects Distribution: $\text{logit}(p_{ij})=\mu_j+\delta_{ij}$, $\quad \delta_{ij}\sim \text{N}(0,\sigma_j^2)$, $\quad i=1,\ldots,n,\ j=0,1$.
Stage 3. Hyperprior for $\boldsymbol{\phi}=(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2)$: $\mu_j\sim \text{N}(a,b)$, $\quad \sigma_j^{-2}\sim \text{Gamma}(c,d)$, $\quad j=0,1$, with a,b,c,d specified *a priori*.

In an alternative specification, we could replace the normal distributions at Stage 2 with a pair of beta distributions for p_{ij}, with the parameters of these distributions being assigned hyperpriors at Stage 3.

Table I. Table summarizing data in area i; in an ecological study the margins only are observed. There are N_{i+} individuals in area i, with y_{i+} responding $Y=1$, and N_{i0}, N_{i1} individuals with $j=0,1$, respectively.

	$Y=0$	$Y=1$	
$j=0$		y_{i0}	N_{i0}
$j=1$		y_{i1}	N_{i1}
	$N_{i+}-y_{i+}$	y_{i+}	N_{i+}

We let $\boldsymbol{p}=(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n)$, $\boldsymbol{y}_+=(y_{1+},\ldots,y_{n+})$ and $\boldsymbol{N}=(\boldsymbol{N}_1,\ldots,\boldsymbol{N}_n)$. The posterior of interest is available as the product of the three stages:

$$\begin{aligned}\pi(\boldsymbol{p},\boldsymbol{\phi}|\boldsymbol{y}_+,\boldsymbol{N}) &\propto \Pr(\boldsymbol{y}_+|\boldsymbol{p},\boldsymbol{N})\pi(\boldsymbol{p}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\\ &=\left\{\prod_{i=1}^{n}\Pr(y_{i+}|\boldsymbol{p}_i,\boldsymbol{N}_i)\pi(\boldsymbol{p}_i|\boldsymbol{\phi})\right\}\pi(\boldsymbol{\phi}). \qquad (3)\end{aligned}$$

2.2. *Ecological data with random sampling*

In this section we assume that a random sample of size $0\leqslant m_{ij}\leqslant N_{ij}$ is drawn, *within* population j and area i. The disease status of the sampled individuals is then determined, with z_{ij} denoting the number of diseased individuals. We let $y^{\star}_{i+}=y_{i+}-z_{i+}$ and $N^{\star}_{ij}=N_{ij}-m_{ij}$, $j=0,1$, represent the residual aggregate data that are constructed from the full totals, with the observed individual-level data subtracted, as detailed in Table II. Let $\boldsymbol{y}^{\star}_+=(y^{\star}_{1+},\ldots,y^{\star}_{n+})$, $\boldsymbol{N}^{\star}_i=(N^{\star}_{i0},N^{\star}_{i1})$, $\boldsymbol{N}^{\star}=(\boldsymbol{N}^{\star}_1,\ldots,\boldsymbol{N}^{\star}_n)$, $z_i=(z_{i0},z_{i1})$, $\boldsymbol{z}=(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)$, $\boldsymbol{m}_i=(m_{i0},m_{i1})$ and $\boldsymbol{m}=(\boldsymbol{m}_1,\ldots,\boldsymbol{m}_n)$.

We decompose the posterior as

$$\begin{aligned}\pi(\boldsymbol{p},\boldsymbol{\phi}|\boldsymbol{y}^{\star}_+,\boldsymbol{z},\boldsymbol{N}^{\star},\boldsymbol{m}) &\propto \Pr(\boldsymbol{y}^{\star}_+,\boldsymbol{z}|\boldsymbol{p},\boldsymbol{N}^{\star})\pi(\boldsymbol{p}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\\ &=\Pr(\boldsymbol{y}^{\star}_+|\boldsymbol{p},\boldsymbol{N},\boldsymbol{m})\Pr(\boldsymbol{z}|\boldsymbol{p},\boldsymbol{m})\pi(\boldsymbol{p}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\\ &=\left\{\prod_{i=1}^{n}\Pr(y^{\star}_{i+}|\boldsymbol{p}_i,\boldsymbol{N}^{\star}_i)\Pr(\boldsymbol{z}_i|\boldsymbol{p}_i,\boldsymbol{m}_i)\pi(\boldsymbol{p}_i|\boldsymbol{\phi})\right\}\pi(\boldsymbol{\phi})\end{aligned}$$

with the *residual* ecological data modeled as $y^{\star}_{i+}|\boldsymbol{p}_i\sim\text{Binomial-Conv}(\boldsymbol{N}^{\star}_i,\boldsymbol{p}_i)$, and the individual data as $z_{ij}|p_{ij}\sim\text{Binomial}(m_{ij},p_{ij})$, for $j=0,1$ and $i=1,\ldots,n$. Stages 2 and 3 are as in Section 2.1. Note that there is no need for random samples to be available in all areas, so that $m_{i0}=0$ and/or $m_{i1}=0$ in some areas.

2.3. *Ecological data with case–control sampling*

We now consider the hybrid ecological case–control design, [15], in which samples of $Y=0$ and $Y=1$ individuals are gathered within areas, rather than random samples from the $j=0$ and $j=1$ populations. Table III summarizes notation for this design. Within area i, M_{i0} controls are sampled from the available $N_{i+}-y_{i+}$ non-cases, and M_{i1} cases are sampled from the total available cases, y_{i+}. Hence, M_{i0} and M_{i1} are the control and case sample sizes in area i. Subsequently, we determine the population ($j=0/1$) to which each of these M_{i0} and M_{i1} individuals belong. We suppose that a number z^1_{i0} of the controls and z^1_{i1} of the cases fall in the $j=1$ population; these data constitute the case–control outcomes. Let $z^1_i=(z^1_{i0},z^1_{i1})$, $\boldsymbol{z}^1=(\boldsymbol{z}^1_1,\ldots,\boldsymbol{z}^1_n)$, $\boldsymbol{M}_i=(M_{i0},M_{i1})$ and $\boldsymbol{M}=(\boldsymbol{M}_1,\ldots,\boldsymbol{M}_n)$. The *case-only* design in which $M_{i0}=0$, for $i=1,\ldots,n$, is particularly appealing since it requires the population status of only the cases to be determined, and information about cases may be more straightforward to obtain than information about controls, for example from a cancer registry.

The posterior is given by

$$\begin{aligned}\pi(\boldsymbol{p},\boldsymbol{\phi}|\boldsymbol{y}^{\star}_+,\boldsymbol{z}^1,\boldsymbol{N},\boldsymbol{M}) &\propto \Pr(\boldsymbol{y}^{\star}_+,\boldsymbol{z}^1|\boldsymbol{p},\boldsymbol{N},\boldsymbol{M})\pi(\boldsymbol{p}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\\ &=\prod_{i=1}^{n}\left[\Pr(y_{i+}|\boldsymbol{p}_i,\boldsymbol{N}_i)\Pr(\boldsymbol{z}^1_i|y_{i+},\boldsymbol{p}_i,\boldsymbol{N}_i,\boldsymbol{M}_i)\pi(\boldsymbol{p}_i|\boldsymbol{\phi})\right]\pi(\boldsymbol{\phi})\end{aligned}$$

Table II. Summary of notation for the situation in which we have both individual survey data with sample sizes m_{i0} and m_{i1}, and ecological (aggregate) marginal data, in area i. The right panel of the table describes the residual ecological portion of the data.

	Survey data			Residual ecological data		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$j=0$		z_{i0}	m_{i0}			$N^{\star}_{i0}=N_{i0}-m_{i0}$
$j=1$		z_{i1}	m_{i1}			$N^{\star}_{i1}=N_{i1}-m_{i1}$
	$m_{i+}-z_{i+}$	z_{i+}	m_{i+}	$N^{\star}_{i+}-y^{\star}_{i+}$	$y^{\star}_{i+}=y_{i+}-z_{i+}$	$N^{\star}_{i+}=N_{i+}-m_{i+}$

Table III. Summary of notation for the situation in which we have both individual case–control data with sample sizes M_{i0} and M_{i1}, and ecological (aggregate) marginal data, in area i.

	Case–control data			Ecological data		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$j=0$						N_{i0}
$j=1$	z^1_{i0}	z^1_{i1}				N_{i1}
	M_{i0}	M_{i1}	M_{i+}	$N_{i+}-y_{i+}$	y_{i+}	N_{i+}

with $\Pr(\boldsymbol{y}^\star_{i+}|\boldsymbol{p}_i,\boldsymbol{N}_i)$ providing the likelihood contribution from the ecological data (i.e. the binomial convolution) in area i, and $\Pr(\boldsymbol{z}^1_i|\boldsymbol{y}_{i+},\boldsymbol{p}_i,\boldsymbol{N}_i,\boldsymbol{M}_i)$ the contribution from the case–control data, given the ecological data. Detailed arguments given elsewhere [15] show that the latter distribution is

$$
\begin{aligned}
&\Pr(\boldsymbol{z}^1_i|\boldsymbol{y}_{i+},\boldsymbol{p}_i,\boldsymbol{N}_i,\boldsymbol{M}_i)\\
&\quad=\sum_{y_{i0}\in R_i}\Pr(z^1_{i0}|y_{i0},y_{i+},N_{i0},M_{i0})\Pr(z^1_{i1}|y_{i0},y_{i+},N_{i1},M_{i1})\Pr(y_{i1}|y_{i+},\boldsymbol{p}_i,\boldsymbol{N}_i)\\
&\quad=\sum_{y_{i0}\in R_i}\frac{\binom{N_{i1}-y_{i+}+y_{i0}}{z^1_{i0}}\binom{N_{i0}-y_{i0}}{M_{i0}-z^1_{i0}}}{\binom{N_{i+}-y_{i+}}{M_{i0}}}\frac{\binom{y_{i+}-y_{i0}}{z^1_{i1}}\binom{y_{i0}}{M_{i1}-z^1_{i1}}}{\binom{y_{i+}}{M_{i1}}}\\
&\qquad\times\frac{\binom{N_{i0}}{y_{i0}}\binom{N_{i1}}{y_{i+}-y_{i0}}\exp(y_{i0}\theta_i)}{\sum_{u\in R_i}\binom{N_{i0}}{y_{i+}-u}\binom{N_{i1}}{u}\exp(u\theta_i)},
\end{aligned}
\tag{4}
$$

where the support of y_{i0} is

$$R_i=\{\max(M_{i1}-z^1_{i1},y_{i+}-N_{i1}+z^1_{i0}),\ldots,\min(y_{i+}-z^1_{i1},N_{i0}-M_{i0}-z^1_{i0})\}$$

and

$$\theta_i=\log\{p_{i0}/(1-p_{i0})\}-\log\{p_{i1}/(1-p_{i1})\} \tag{5}$$

is the log odds ratio comparing the log odds of the probability of $Y=1$ for $X=0$, versus $X=1$.

3. Computation

In this section we discuss computation for three designs. In the first, ecological data only are available; in the second and third situations we supplement the ecological data, first with random samples, and then with case–control samples.

3.1. Ecological data

The posterior distribution in (3) is analytically intractable, and various approaches to computation have been suggested. In a truncated normal random effects model [1] numerical integration was used, while in a binomial-beta model [21] MCMC was utilized. In the context of a Poisson model for rare events, an auxiliary variable MCMC scheme was proposed [22] and this scheme was subsequently used in the non-rare situation [9]. With respect to the posterior (3), a generic MCMC Metropolis–Hastings algorithm cycles through the following steps:

(1) $\pi(\boldsymbol{p}_i|\boldsymbol{y}_i,\boldsymbol{\phi},\boldsymbol{N}_i)$, $i=1,\ldots,n$.

(2) $\pi(\boldsymbol{y}_i|\boldsymbol{p}_i, \boldsymbol{\phi}, \boldsymbol{y}_{i+}, \boldsymbol{N}_i) = \pi(\boldsymbol{y}_i|\boldsymbol{p}_i, \boldsymbol{y}_{i+}, \boldsymbol{N}_i)$, $i = 1, \ldots, n$.
(3) $\pi(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{p}, \boldsymbol{N}) = \pi(\boldsymbol{\phi}|\boldsymbol{p})$.

We emphasize that in step 1, we systematically cycle through each table in turn, on each occasion sampling $\boldsymbol{p}_i$, and similarly for $\boldsymbol{y}_i$ in step 2, $i = 1, \ldots, n$. It is straightforward to construct Metropolis–Hastings/Gibbs proposals for steps 1 and 3 since, when the individual-level data are available, we simply have a generalized linear mixed model. Specifically, in step 1 we parameterize in terms of the pair of logits $\{\log[p_{i0}/(1-p_{i0})], \log[p_{i1}/(1-p_{i1})]\}$, and use a random walk chain with a multivariate normal proposal (so that we carry out block updating, with each block consisting of the pair of logits). In step 3 we assume conjugate distributions (normal and gamma distributions) so that Gibbs steps are possible.

3.1.1. Full enumeration. The obvious method for generation in step 2 is a Gibbs step in which we evaluate the full conditional distribution. Since we observe y_{i+} we need to sample from

$$\pi(\boldsymbol{y}_i|\boldsymbol{p}_i, \boldsymbol{\phi}, y_{i+}, \boldsymbol{N}_i) \equiv \Pr(y_{i0}|y_{i+}, \boldsymbol{p}_i, \boldsymbol{N}_i) = \begin{cases} \dfrac{\binom{N_{i0}}{y_{i0}}\binom{N_{i1}}{y_{i+}-y_{i0}}\exp(y_{i0}\theta_i)}{\sum_{z=l_i}^{u_i}\binom{N_{i0}}{z}\binom{N_{i1}}{y_{i+}-z}\exp(z\theta_i)}, & y_{i0} = l_i, \ldots, u_i, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where l_i and u_i are given in (2), and θ_i is the log odds ratio defined in (5).

3.1.2. A normal approximation. Sampling from (6) for each area requires the evaluation of the summation in the denominator, and at each iteration of the Markov chain, which in many applications is a computationally prohibitive task, because N_{0i} and N_{1i} are large. Instead, a normal approximation to the convolution was introduced for the situation in which the marginal totals are large [9]. Specifically, the normal approximation is

$$Y_{i+}|p_{i0}, p_{i1} \sim \mathrm{N}\{N_{i0}p_{i0} + N_{i1}p_{i1}, N_{i0}p_{i0}(1-p_{i0}) + N_{i1}p_{i1}(1-p_{i1})\}. \tag{7}$$

`WinBUGS` code for ecological inference using this approximation is available, [23]. The approximation is also implemented in the `MCMCpack R` function `MCMChierEI`, [24]. In Sections 4 and 5 we examine the accuracy of the normal approximation, (7), and in Appendix A we describe a new implementation method that acknowledges the lack of identifiability in the likelihood. In the `JAGS` (Just Another Gibbs Sampler) software, [25], there is a novel distribution `dsum` that may be used in the ecological inference context, as we now describe (Plummer, personal communication). The specification `y ~ dsum(y0, y1)`, where `y` is observed and `y0, y1` are unobserved discrete-valued stochastic nodes, creates an MCMC sampler that will simultaneously update `y0` and `y1`, while respecting the constraint `y0 + y1 == y`. In `JAGS 1.0.x` discrete slice sampling [26] was used. In `JAGS 2.0.0` a random walk Metropolis step is used, with the step size drawn from a geometric distribution. We now describe a related approach.

3.1.3. Markov basis method. We consider a generic scheme based on Markov bases [27], and previously employed, without example, for ecological data [28]. The domain of applications of Markov bases includes the calculation of exact p-values [29], testing goodness-of-fit for categorical data [30] and data augmentation in multi-way contingency tables [31]. Here, we present its application as a method for missing data imputation in two-way tables with fixed row and column totals, and possibly with supplementary individual-level data. Appendix B presents a more technical discussion of the use of Markov basis methods, including the consideration of more complex situations in which data on more than two variables are available.

The idea is to replace the expensive sampling in (6) with a Metropolis step in which, given a current point y_{i0}, a new point is proposed as $\tilde{y}_{i0} = y_{i0} + t$ or $\tilde{y}_{i0} = y_{i0} - t$, each with probability 0.5, where t is drawn uniformly from $\{1, 2, \ldots, T\}$, for a fixed $T > 0$, with y_{i1} correspondingly replaced with $\tilde{y}_{i1} = y_{i1} - t$ or $\tilde{y}_{i1} = y_{i1} + t$, respectively. The new point will clearly respect the margins (as given in Table I), but first we must check that it is valid with respect to the constraints (2). If the latter check fails, we remain at the current point. If a 'legal' full table is generated, we move to the new point via a Metropolis move with probability $\min\{p,1\}$ where

$$p=\frac{\Pr(\tilde{y}_{i0}|\boldsymbol{p}_i,y_{i+})}{\Pr(y_{i0}|\boldsymbol{p}_i,y_{i+})}=\frac{\binom{N_{i0}}{\tilde{y}_{i0}}\binom{N_{i1}}{y_{i+}-\tilde{y}_{i0}}\exp(\tilde{y}_{i0}\theta_i)}{\binom{N_{i0}}{y_{i0}}\binom{N_{i1}}{y_{i+}-y_{i0}}\exp(y_{i0}\theta_i)} \tag{8}$$

In the case $T=1$, we propose either $\{\tilde{y}_{i0}=y_{i0}+1, \tilde{y}_{i1}=y_{i1}-1\}$, or $\{\tilde{y}_{i0}=y_{i0}-1, \tilde{y}_{i1}=y_{i1}+1\}$, each with probability 0.5. The acceptance ratio (8) takes a particularly simple form:

$$p=\begin{cases}\dfrac{(N_{i0}-y_{i0})(y_{i+}-y_{i0})}{(y_{i0}+1)(N_{i1}-y_{i+}+y_{i0}+1)}\exp(\theta_i) & \text{if } \tilde{y}_{i0}=y_{i0}+1,\\[2ex] \dfrac{y_{i0}(N_{i1}-y_{i+}+y_{i0})}{(N_{i0}-y_{i0}+1)(y_{i+}-y_{i0}+1)}\exp(-\theta_i) & \text{if } \tilde{y}_{i0}=y_{i0}-1.\end{cases}$$

3.2. Ecological data with random sampling

The extension to the case in which random samples of individual-level data are drawn within areas is straightforward. In this case the auxiliary variable scheme is applied to $\pi(\boldsymbol{y}_i^\star|\boldsymbol{p}_i,y_{i+}^\star,\boldsymbol{N}_i^\star)$. Specifically, (8) holds but with (N_{i0},N_{i1}) replaced by $(N_{i0}^\star,N_{i1}^\star)$, and (y_{i0},y_{i1}) replaced by $(y_{i0}^\star,y_{i1}^\star)$, with (2) similarly amended.

3.3. Ecological data with case–control sampling

In previous analyses of ecological plus case–control data, exact sampling was used for the unobserved cases and non-cases [14]. As with the extension to the situation in which ecological data are supplemented with individual-level samples drawn at random, the extension to the situation in which case–control samples are drawn from the areas is relatively straightforward. Specifically, the acceptance ratio in the Metropolis step for updating the auxiliary variable, y_{i0}, is modified to accommodate the additional information to give:

$$\begin{aligned} p&=\frac{\Pr(\tilde{y}_{i0}|\boldsymbol{p}_i,y_{i+},\boldsymbol{N}_i,\boldsymbol{z}_i^1,\boldsymbol{M}_i)}{\Pr(y_{i0}|\boldsymbol{p}_i,y_{i+},\boldsymbol{N}_i,\boldsymbol{z}_i^1,\boldsymbol{M}_i)}\\ &=\frac{\binom{N_{i0}-\tilde{y}_{i0}}{M_{i0}-z_{i0}^1}\binom{y_{i+}-\tilde{y}_{i0}}{z_{i0}^1}\binom{\tilde{y}_{i0}}{M_{i1}-z_{i1}^1}\binom{N_{i1}-y_{i+}+\tilde{y}_{i0}}{z_{i1}^1}\binom{N_{i0}}{\tilde{y}_{i0}}\binom{N_{i1}}{y_{i+}-\tilde{y}_{i0}}\exp(\tilde{y}_{i0}\theta_i)}{\binom{N_{i0}-y_{i0}}{M_{i0}-z_{i0}^1}\binom{y_{i+}-y_{i0}}{z_{i0}^1}\binom{y_{i0}}{M_{i1}-z_{i1}^1}\binom{N_{i1}-y_{i+}+y_{i0}}{z_{i1}^1}\binom{N_{i0}}{y_{i0}}\binom{N_{i1}}{y_{i+}-y_{i0}}\exp(y_{i0}\theta_i)}. \end{aligned} \tag{9}$$

Similar to the ecological only case of Section 3.1, the acceptance ratio can be simplified considerably. For example, when $T=1$, we have

$$p=\begin{cases}\dfrac{(N_{i0}-y_{i0}-M_{i0}+z_{i0}^1)(y_{i+}-y_{i0}-z_{i1}^1)}{(y_{i0}+1-M_{i1}+z_{1i}^i)(N_{i1}-y_{i+}+y_{i0}+1-z_{i0}^1)}\exp(\theta_i) & \text{if } \tilde{y}_{i0}=y_{i0}+1,\\[2ex] \dfrac{(y_{i0}-M_{i1}+z_{1i}^i)(N_{i1}-y_{i+}+y_{i0}-z_{i0}^1)}{(N_{i0}-y_{i0}+1-M_{i0}+z_{i0}^1)(y_{i+}-y_{i0}+1-z_{i1}^1)}\exp(-\theta_i) & \text{if } \tilde{y}_{i0}=y_{i0}-1.\end{cases}$$

For the case-only analysis, (9) is modified by removing the terms that involve M_{i0} and/or z_{i0}^1.

4. Simulation study

We report a small simulation study that compares the Markov basis and normal approximation methods in terms of speed and accuracy. For a subset of simulations we also consider the full enumeration approach. We generate data with $n=100$ areas and $\mu_0=-2$, $\mu_1=0$, $\sigma_0=\sigma_1=0.5$. This choice gives

90 per cent intervals for p_{i0} and p_{i1} of (0.06, 0.24) and (0.31, 0.69), respectively. We consider three sample sizes of $N_{i+}=1000, 500, 100$ and two choices of the proportions, across areas, within each of the $j=0,1$ categories. In the 'balanced' design we take N_{i0}/N_{i+} to be approximately uniformly distributed over (0,1). Specifically, $N_{i0}=10+\text{floor}\{(i-1)(N_{i+}-10)/(n-1)\}$ and $N_{i1}=N_{i+}-N_{i0}$. Hence, with $N_{i+}=1000$ we obtain $N_{i0}=\{10, 19, 29, \ldots, 970, 980, 990\}$ and $N_{i1}=\{990, 981, 971, \ldots, 30, 20, 10\}$. In the 'unbalanced' design we take N_{i0}/N_{i+} to be approximately uniformly distributed over (0, 0.5). Specifically, $N_{i0}=10+\text{floor}\{(i-1)(N_{i+}/2-20)/(n-1)\}$ and $N_{i1}=N_{i+}-N_{i0}$. Hence, with $N_{i+}=1000$ we obtain $N_{i0}=\{10, 14, 19, \ldots, 490, 495, 500\}$ and $N_{i1}=\{990, 986, 981, \ldots, 510, 505, 500\}$. For each set of simulations, six models are fitted. As a baseline, an individual-level model is fitted to data in which all of the entries are observed in the 100 2×2 tables. For the analysis of the ecological data alone, the Markov basis computational approach is compared with the normal approximation. For this pair of analyses, there is potential for ecological bias, particularly in the unbalanced case. Hence, we supplement the ecological data with random subsamples of size 5, 10, 20 per cent in each area and in each of the $j=0,1$ categories. In these three situations we analyze with the Markov basis approach. In all simulations and for all methods the Markov chain was initially run for 250 000 iterations, following 10 000 burn-in, with the summaries based on thinning by 250 (to give a sample size of 1000). For the $N_{i+}=100$ results further runs were required due to poor convergence, as detailed below.

Figure 1 displays the results, with a panel for each of μ_0, μ_1, σ_0 and σ_1, and the horizontal dashed line indicating the true values from which the data were generated. Since we only carry out a single simulation for each scenario, the filled black circles, which correspond to the individual-level analysis, should be viewed as the gold standard.

For the balanced design with $N_{i+}=1000$ individuals in each area, there is ecological bias for each parameter though little difference between the Markov basis and normal approximation, indicating that the latter is accurate. The bias is quickly reduced when individual data supplement the ecological data, and accurate inference results when samples of both 10 and 20 per cent are added. For the unbalanced $N_{i+}=1000$ simulation the ecological bias is more pronounced for σ_0, though samples of 10 and 20 per cent correct the bias. For the $N_{i+}=500$ results similar broad patterns emerge, though in the unbalanced case the normal approximation starts to fail (in particular for μ_0 and σ_0). For the $N_{i+}=100$ results ecological bias is severe in the unbalanced case, and a 20 per cent sample of individuals

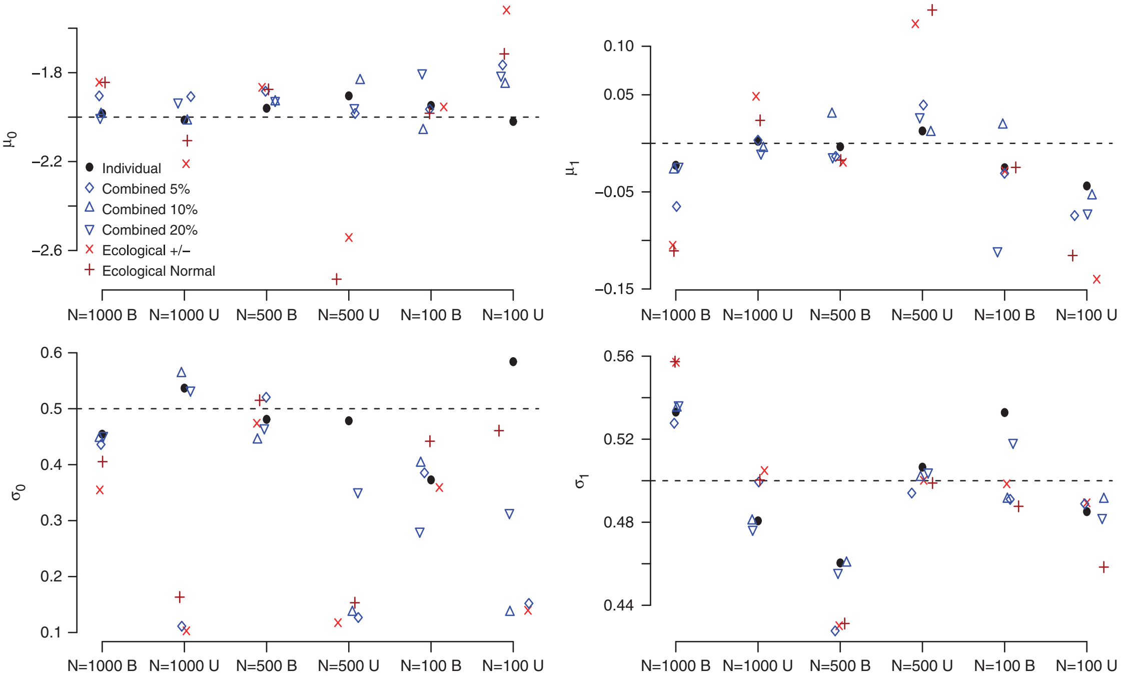

Figure 1. Results of the simulation study. 'B' and 'U' correspond to the balanced and unbalanced data sets. The posterior medians for each parameter are plotted for each of the six scenarios. The dashed lines correspond to the values that were used to generate the data. On each of the four plots, for each of the six experiments, the points have been jittered in a horizontal direction.

(20 in each area) does not redeem the situation. In this scenario the convergence of the Markov chains for each of the analyses based on the ecological data along was very poor, and so runs of 1 million iterations were performed.

We report timings based on the first figure (total user time) produced by the `proc.time()` function in `R`. The analyses were run on an Apple Macintosh laptop with a 2.4 GHz Intel Core 2 Duo processor and 4 GB of memory. The individual analyses all took a similar time, with an average of 6153 s. The Markov basis method took an average of 8685 s, which is a 40 per cent increase in time. With the addition of the individual-level data, this time increased to around 8996 s (averaged across the three percentage sample sizes, they were all approximately equal). The normal approximation averaged 9823 s, but this was for the code that was written by the authors of this paper. The normal approximation available in `MCMCpack` is orders of magnitude faster, since it utilizes compiled `C++`. All timings were for analyses that were run for the same number of iterations.

The full enumeration method took a very large amount of time to run, and so was only used on a limited subset. For example, the method took 28 379 s for the case of $N_{i+}=100$ individuals. This method does not scale well, and for the $N_{i+}=500$ case the run time was 102 863 s, which is 12 times the run time of the Markov basis method. Approximate versions of this technique are available, for example, by truncating the range of y_{i0} when the probabilities become 'negligible' [9], but care is required with such shortcuts. We would clearly expect the individual-level analysis to have the shortest convergence time, and the full enumeration method should display better convergence than the method based on Markov bases.

The simulated data sets and `R` code to implement the methods are available at: http://faculty.washington.edu/jonno/software.html.

5. Examples

We present two real-world examples. In the first example we illustrate the inaccuracy of the normal approximation when compared with the Markov basis method, and show the benefits of supplementing the ecological data with individual random samples. In the second example we illustrate how the Markov basis method can be implemented in a non-standard scenario, namely the ecological plus case–control sample design. In both examples we have access to the individual-level data, which allows examination of the extent of ecological bias.

5.1. Diabetes and race

To illustrate the approach described in Sections 3.1 and 3.2 we analyze data from the Behavioral Risk Factor Surveillance System (BRFSS), collected in Washington State in 2006. We consider females only, and take as the outcome of interest a binary response, with $Y=1$ corresponding to 'Ever told you had diabetes,' and $Y=0$ to the complementary event. The binary explanatory variable is white/non-white (corresponding to $j=0/1$), and zip codes provide the geographical level of aggregation, with 506 in total. We wish to compare the algorithm of Section 3.1 with the normal approximation, (7). Table IV gives numerical summaries of the data. Let y_{i0} and y_{i1} represent the number of white and non-white diabetes cases in area i, with N_{i0} and N_{i1} the respective denominators. Based on the complete data, the proportion of whites with diabetes across areas, y_{i0}/N_{i0}, ranges between 0 and 0.5 for whites, with mean 0.091, and the proportion of non-whites with diabetes across areas, y_{i1}/N_{i1}, ranges between 0 and 1, with mean 0.11. We carry out ecological inference based on y_{i+} only. Subsequently, we examine the case in which subsamples are taken within areas, to provide a subset of individual-level data.

The unobserved fractions of white and non-white diabetes cases may be bounded, based on the observed row and column margins, as given in (2). Given the margins, there is only a single unknown within each table, and the two unobserved fractions are linearly related. So-called tomography lines [1] plot these linear combinations, and graphically allow the bounds on the fractions with diabetes for whites and non-whites to be examined. Figure 2 illustrates these bounds for the diabetes data. The bounds are far wider for non-whites since, as noted, the latter are a far smaller population in the sample. Consequently, we would expect inference for the non-whites, based on the ecological data only, to be particularly unreliable.

We use the three-stage hierarchical model described in Section 2.1. As hyperpriors we assume $\mu_j \sim \text{N}(0, 2.9)$, $\sigma_j^{-2} \sim \text{Gamma}(0.5, 0.008)$ for $j=0, 1$. The priors on μ_j are such that the medians, across

Table IV. Summary information for the BRFSS data. Rows 1 and 2 of the table contain summaries for the sizes of the surveys for whites, N_{i0}, and non-whites, N_{i1}. Rows 3 and 4 contain summaries for the number of diabetes cases who are white, y_{i0}, and non-white, y_{i1}.

	Summaries across areas				
	Mean	Min	Median	Max	Total
White survey	25.5	0	26.0	200	12 919
Non-white survey	3.8	0	2.0	65	1943
White diabetes	1.0	0	2.2	23	1133
Non-white diabetes	0.41	0	0.0	10	209

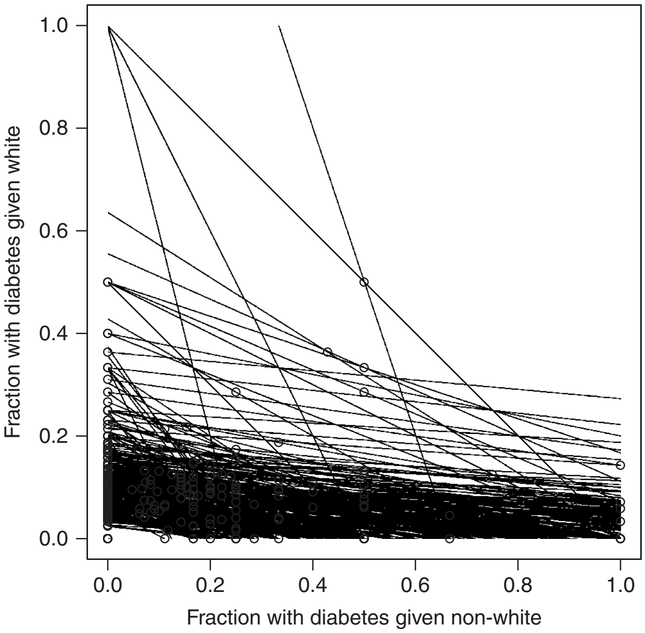

Figure 2. Tomography plot illustrating the bounds on the fractions with diabetes for whites and non-whites. The majority of the bounds for the non-white populations cover (0, 1) and so are uninformative, which explains why an aggregate data analysis suffers from ecological bias.

areas, of p_{ij} are approximately uniform on [0, 1]. The priors on σ_j^{-2} give residual odds with a 95 per cent range (across areas) of [0.2, 5]; further details on these derivations are provided elsewhere [32]. We initially report three analyses: (1) individual-level based on the full data, (2) aggregate-level using the normal approximation to the convolution likelihood, (3) aggregate-level using the Metropolis Markov basis algorithm described in Section 3 using $T = 1$ (since the denominators are relatively small here).

Table V presents results for the different analyses. For each of the analyses based on the ecological data, mixing of the Markov chain was slow, and 250 000 iterations were performed, following burn-in. When compared with the results from the individual-level data, the ecological analyses clearly produce biased results, in particular for μ_1 and σ_1 (because there is far less information for the non-white sample). Results from the normal approximation, (7), showed inaccuracy for μ_1 and σ_1 in particular, due to the small denominators for the non-white sample.

To supplement the ecological data, we adopt a design in which random samples of size 10 were taken from the white sample in all areas containing 15 or more whites (there were 257 such areas), with the diabetes status of these individuals being recorded. For non-whites, samples of size 3 were drawn from areas containing at least 4 or more non-whites (there were 189 such areas). Of the 506 areas, 161 contributed both white and non-white individual data. This design led to individual-level data for a total of 2570 white and 567 non-white individuals, corresponding to 20 and 29 per cent of the total. Table V gives the results from the combined aggregate/individual data, based on the algorithm described in Section 3.2, and illustrates that ecological bias has been greatly reduced. For the white group little is lost when moving between the individual and combined data situations,

Table V. Posterior summaries for the means and standard deviations of the logits for each population, for individual, aggregate and combined analyses of the diabetes and race data. For each method and parameter we present the posterior median and (5 per cent, 95 per cent) interval. The 'Individual' analysis is based on the totality of data, as described in Table IV, so that there are 12 919 and 1943 individuals of white and non-white races, respectively. The combined analysis is based on the ecological data combined with individual-level data on 2570 white and 567 non-white individuals.

	μ_0		μ_1		σ_0		σ_1	
	Median	90 per cent CI	Median	90 per cent CI	Median	90 per cent CI	Median	90 per cent CI
Individual	−2.35	(−2.40, −2.30)	−2.14	(−2.29, −2.03)	0.13	(0.053, 0.23)	0.12	(0.057, 0.30)
Ecological normal	−2.54	(−2.66, −2.42)	−1.61	(−2.16, −1.27)	0.23	(0.080, 0.40)	0.67	(0.086, 1.24)
Ecological Markov basis	−2.50	(−2.60, −2.41)	−1.55	(−1.83, −1.31)	0.15	(0.070, 0.28)	0.13	(0.061, 0.34)
Combined Markov basis	−2.38	(−2.45, −2.32)	−1.99	(−2.17, −1.80)	0.13	(0.059, 0.24)	0.13	(0.059, 0.34)

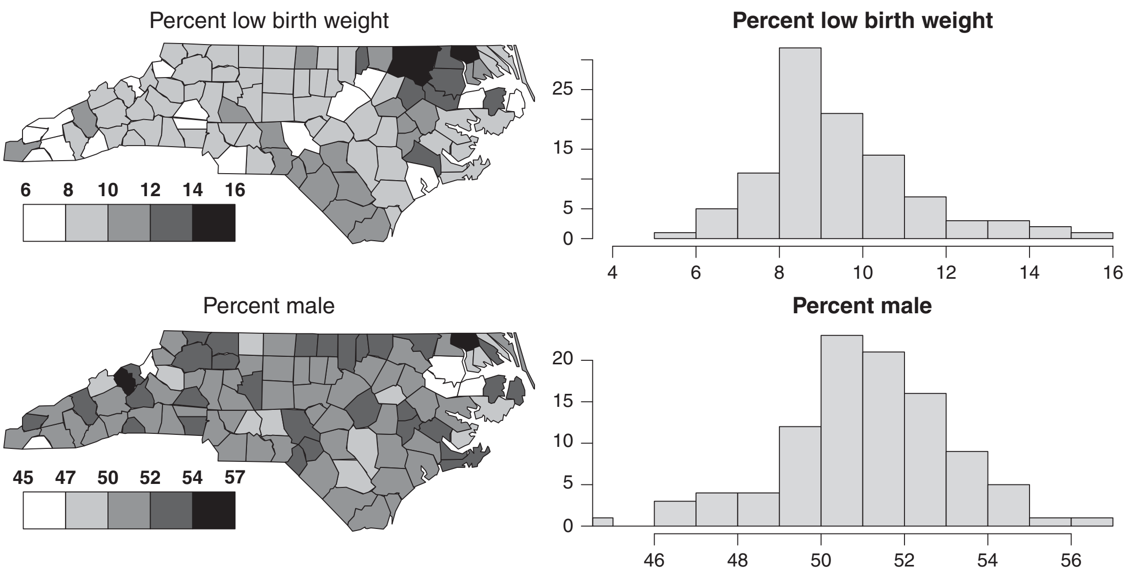

Figure 3. County-specific outcome (LBW) and exposure (gender) information for the North Carolina data.

Table VI. Summary information for the low birth weight data. Rows 1 and 2 of the table contain summaries for the county-specific number of female births, N_{i0}, and male births, N_{i1}. Rows 3 and 4 contain summaries for the number of LBW births for females, y_{i0}, and males, y_{i1}.

	Summaries across areas				
	Mean	Min	Median	Max	Total
Female births	1161	41	610	12 523	116 111
Male births	1219	48	686	13 195	121 871
Female					
LBW births	114	4	62	1172	11 394
Male					10 099
LBW births	101	5	60	1000	10 099

while for the non-white group the estimate of μ_1 from the combined analysis is subject to a little ecological bias. In addition, interval estimates from the combined analysis have increased in width when compared with the individual data analysis.

For the individual-level analysis the time for 55 000 iterations was 5 802 s. For the normal approximation and Markov basis analyses a greater number of iterations were required. For 255 000 iterations the timing for the normal approximation (using our own code) was 39 572 s. For the Markov basis method, the timings were 33 390 s for the ecological data alone, and 31 779 s for the combined data.

5.2. *Low birth weight and gender*

In this example we illustrate the method of Section 3.3. In particular, we consider a study of the impact of gender on the probability of a low birth weight (LBW: weight less than 2500 g), using data compiled by the North Carolina State Center for Health Statistics (`http://www.irss.unc.edu/`).

Restricting our analysis to the years 2003 and 2004, there were a total of 237 982 births in the state of North Carolina, of which 21 493 were LBW. Across the $n=100$ counties in North Carolina, the number of births, N_{i+}, varied from 89 to 25 718, and the number of low birth weight babies, (y_{i+}), varied from 9 to 2172. The corresponding LBW rate varied from 6.0 to 15.9 per cent, while the percent male varied from 45.0 to 56.8 per cent. Figure 3 presents maps and histograms of the percentages of LBW and male individuals. Further summaries are provided in Table VI.

The information presented in Figure 3, along with county-specific population totals, corresponds to the information that one would have access to in an ecological study: marginal outcome and gender information. Before presenting the results based on the methods of Section 3, we first briefly consider naive analyses that are commonly performed given ecological data alone in this rare outcome setting. Specifically, letting $Y=0/1$ denote normal/low birth weight, we consider a simple county/gender-specific log-linear model (for a rare outcome) for $p_{ij}=\Pr(Y=1|\text{county } i, j)$:

$$p_{ij}=\exp(\alpha_0+\alpha_1 j) \tag{10}$$

with $j=0/1$ corresponding to female/male. In an ecological setting it may be tempting to fit the *ecological regression model*:

$$E[y_{i+}|\alpha_0^\star,\alpha_1^\star]=N_{i+}\exp(\alpha_0^\star+\alpha_1^\star q_i), \tag{11}$$

where $q_i=N_{i1}/N_{i+}$ denotes the proportion of males in county i. However, it is simple to show that this model bears little resemblance to the *induced aggregate regression model* which averages the individual-level model (10) over the gender distribution within area i. Specifically, from (10), the average risk in county i is

$$\begin{aligned} p_{i+} &= \Pr(Y=1|\text{county } i)=p_{i0}\times\Pr(\text{female}|\text{county } i)+p_{i1}\times\Pr(\text{male}|\text{county } i)\\ &= p_{i0}\times(1-q_i)+p_{i1}\times q_i \end{aligned}$$

to give

$$E[Y_{i+}|\alpha_0,\alpha_1]=N_{i+}\{(1-q_i)\mathrm{e}^{\alpha_0}+q_i\mathrm{e}^{\alpha_0+\alpha_1}\}, \tag{12}$$

which is of completely different form to (11). In the latter the *contextual* effect of gender is being estimated, that is, the effect on birth weights of having differing proportions of female births in the county. Hence, the parameters in the ecological and induced aggregate regression models, have completely difference meanings, which is why we label these parameters as $\alpha_j^\star$ and α_j, $j=0,1$, respectively. An alternative derivation of (12), follows by recognizing that when $Y_{ij}|p_{ij}\sim\text{Poisson}(N_{ij}p_{ij})$ for $j=0,1$, the convolution (1) is simply the sum of two Poisson distributions, leading immediately to the mean model (12).

Assuming a Poisson likelihood for the counts Y_{i+}, the maximum likelihood estimates for the ecological and induced aggregate regressions are $\exp(\widehat{\alpha_1^\star})$ = 0.45 (95 per cent confidence interval: 0.17, 1.2) and $\exp(\widehat{\alpha}_1)$ = 0.33 (95 per cent confidence interval: 0.03, 3.8), respectively. In contrast, an analysis based on the complete individual-level data, with model (10), yields a maximum likelihood estimate of $\exp(\widehat{\alpha}_1)=0.84$ (95 per cent confidence interval: 0.82, 0.86). As such, despite wide confidence intervals, point estimates from both the ecological and induced aggregate regression analyses vastly overstate the relative differences between male and female birth weights. A major difficulty with the aggregate analyses is that there is little information in the data, due to the narrow spread in the gender ratio across counties. As we see below, analyses that combine ecological and case–control data, can exploit within-, as well as between-county, information.

Returning to the methods of Section 3, suppose the county/gender-specific LBW probabilities are specified via the logistic model

$$\text{logit}(p_{ij})=\beta_0+\beta_1 j+\delta_i, \tag{13}$$

where $\boldsymbol{\delta}=(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_n)$ with $\boldsymbol{\delta}_i\sim_{\text{iid}}\text{N}(0,\sigma^2)$. It is usual to assume a common random effect for both populations, when the outcome is rare. Hence in the notation of Section 2, $\boldsymbol{\phi}=(\beta_0,\beta_1,\sigma^2)$. Independent normal priors with large variances were assumed for β_0 and β_1, with $\sigma^{-2}\sim\text{Gamma}(0.5,0.008)$, as in the analysis of the BRFSS data.

Table VII presents posterior medians and 90 per cent credible intervals from three analyses of the LBW data. The first row reports on an individual-level analysis of the full data. The second and third analyses consider the incorporation of supplemental case–control data consisting of 10 cases and 10 controls, sampled from each of the 100 counties. The *case–control* analyses uses all of the resulting 1999 samples (one county only had 9 cases); the *cases only* analyses solely uses the 999 cases.

For each analysis, three chains were run, each for 50 000 iterations, again with $T=1$ for the sampling of the population counts, using the Metropolis Markov basis methodof Section 3.3. Posterior summaries

Table VII. Posterior summaries for the baseline odds, odds ratio and random effects standard deviation, for individual and combined analyses of the low birth weight and gender data. For each method and parameter we present the posterior median and (5 per cent, 95 per cent) interval. The 'Individual' analysis is based on the totality of data, as described in Table VI so there are 116 111 female and 121 871 male births. The combined analysis is based on the ecological data, and either on 999 cases and 1000 controls (the 'Combined case–control' analysis) or on 999 cases only (the 'Combined cases only' analysis).

	$\exp(\beta_0)$		$\exp(\beta_1)$		σ	
	Median	90 per cent CI	Median	90 per cent CI	Median	90 per cent CI
Individual	0.109	(0.107, 0.111)	0.831	(0.812, 0.850)	0.011	(0.010, 0.013)
Combined case–control	0.108	(0.103, 0.114)	0.848	(0.760, 0.935)	0.011	(0.010, 0.013)
Combined cases only	0.108	(0.103, 0.114)	0.851	(0.765, 0.936)	0.011	(0.010, 0.013)

were then evaluated after removing a burn-in of 20 per cent of the resulting sample. Focusing on the odds ratio parameter, $\exp(\beta_1)$, the complete data individual-level analysis indicates that the odds of LBW among male births are estimated to be 0.83 (95 per cent credible interval: 0.81, 0.85) times those among female births. These results are therefore very similar to the maximum likelihood analysis of the individual-level data, which is not surprising given the very low estimate of σ. Both combined analyses yield similar results, highlighting the benefit of individual-level data. Not surprisingly, the credible intervals for the analyses based on the subsample of individual-level data are wider than those based on the complete data, although, in this example at least, there appears to be little loss associated with solely using information about the cases. Finally, an analysis was carried out based solely on the case–control data and yielded similar results for the gender odds ratio (0.89, 95 per cent credible interval: 0.76, 1.0), as compared with the gold standard, complete data, results. Interestingly, compared to the results based on the combined aggregate/individual-level data, the credible interval for the odds ratio parameter is clearly wider. This reflects that, once identifiability of the underlying individual-level model is established (via the individual-level data), the incorporation of the ecological data can provide substantial efficiency gains. The timings for each of the three analyses in Table VII were approximately equal.

6. Discussion

In this paper we have described algorithms for carrying out inference when we have ecological data, or ecological plus individual-level data, in the case in which the basic sampling unit is a 2×2 table. The extension to $R \times C$ tables, as previously considered, [12, 33], is simple, with the Markov basis being applied to a pair of randomly selected rows and columns (see Appendix B). When individual-level data are available, it is natural to include random effects with a spatial structure in the linear predictor, and computation for such models is also straightforward using the methods described here. The computational approach described in the paper is designed for discrete outcome and covariates, however, and so cannot be extended to continuous variables.

In Section 4 we saw that the normal approximation breaks down as the number of individuals in each area decreases, as we would expect. Unfortunately we cannot offer universal guidelines as to when the approximation will be accurate, but the Markov basis method offers an alternative with improved accuracy, which is straightforward to implement.

Appendix A

A computational advantage of the normal approximation is that it removes the need to introduce auxiliary variables. Instead, in an MCMC scheme one simply needs to generate for (p_{i0}, p_{i1}), and each of the hyperparameters, μ_0, μ_1, σ_0^2, σ_1^2. Generation for the latter are straightforward (at least for the normal and inverse gamma priors that we assume). An obvious approach to sampling for (p_{i0}, p_{i1}) is via a pair of random walk moves, but the identifiability in the likelihood leads to very low acceptance probabilities. Hence, we describe a new method for sampling the pair (p_{i0}, p_{i1}) under the normal approximation. The (conditional) posterior is

$$\begin{aligned}
&p(p_{i0}, p_{i1}|\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, y_{i+})\\
&\quad\propto (2\pi[N_{i0}p_{i0}(1-p_{i0})+N_{i1}p_{i1}(1-p_{i1})])^{-1/2}\exp\left(-\frac{(y_{i+}-[N_{i0}p_{i0}+N_{i1}p_{i1}])^2}{2[N_{i0}p_{i0}(1-p_{i0})+N_{i1}p_{i1}(1-p_{i1})]}\right)\\
&\qquad\times(2\pi\sigma_0^2)^{-1/2}\exp\left(-\frac{\{\log(p_{i0}/(1-p_{i0})-\mu_0\}^2}{2\sigma_0^2}\right)(2\pi\sigma_1^2)^{-1/2}\exp\left(-\frac{\{\log(p_{i1}/(1-p_{i1})-\mu_1\}^2}{2\sigma_1^2}\right)\\
&\qquad\times[p_{i0}(1-p_{i0})p_{i1}(1-p_{i1}]^{-1}, \qquad\qquad \text{(A1)}
\end{aligned}$$

where the third line corresponds to the Jacobean. The approach is based on a random walk chain in which we propose a point, $(p_{i0}^{\star}, p_{i1}^{\star})$, uniformly within a rectangle whose center is at the current point, and whose sides are parallel to, and perpendicular to, the tomography line, which is given by

$$p_{i1} = \frac{y_{i+}}{N_{i1}} + \frac{N_{i0}}{N_{i1}} p_{i0}.$$

Parallel to the tomography line we sample uniformly in the range $(-\varepsilon_0, \varepsilon_0)$, and in the perpendicular direction in the range $(-\varepsilon_1, \varepsilon_1)$. In the (p_{i0}, p_{i1}) space we have a symmetric proposal, and so the acceptance probability is simply the ratio of conditional posteriors, each of which takes the form (A1), evaluated at the new and old points. Once a new point is proposed we first check that it lies within $[0, 1]\times[0, 1]$, and if this test fails we remain at the current point. In the simulations of Section 4 we used the values $\varepsilon_0 = 0.05$ and $\varepsilon_1 = 0.02$, and achieved acceptance probabilities greater than 50 per cent.

Appendix B

Mehta and Patel, [34], described one of the first algorithms for sampling two-way contingency tables with known row and column totals. More recent key developments include importance sampling algorithms [35–37] and MCMC algorithms [29, 38–40]. One of the central contributions to the MCMC literature was the seminal paper of Diaconis and Sturmfels [27]. They generate tables in a reference set $\mathscr{S}$ through a Markov basis. The fundamental concept behind a Markov basis is easily understood by considering all the possible pairwise differences of tables in $\mathscr{S}$, i.e. $\mathscr{M} = \{n' - n'' : n', n'' \in \mathscr{S}\}$. The elements of $\mathscr{M}$ are called moves. A key issue is the ability to define moves so that all tables within $\mathscr{S}$ can be visited. Any table $n' \in \mathscr{S}$ can be transformed to another table $n'' \in \mathscr{S}$ by applying the move $n'' - n' \in \mathscr{M}$. Clearly, not all the moves in $\mathscr{M}$ are needed to connect any two tables in $\mathscr{S}$ through a series of moves. A Markov basis for $\mathscr{S}$ is obtained by eliminating some of the moves in $\mathscr{M}$ such that the remaining moves still connect $\mathscr{S}$.

The simplest Markov basis contains only moves with two entries equal to 1, two entries equal to -1, and the remaining entries equal to zero. It connects all the two-way tables with the same row and column totals [27]. This result is actually straightforward to prove. In this appendix we change notation, in order to be able to simply generalize to three-way and greater tables, and denote a 2×2 table by $n = (n_{ij})_{1\leqslant i,j\leqslant 2}$. The known row totals are $(n_{i+})_{1\leqslant i\leqslant 2}$ and the known column totals are $(n_{+j})_{1\leqslant j\leqslant 2}$. The set of tables $\mathscr{S}$ with the same row and column totals, i.e.

$$\mathscr{S} = \{n' = (n'_{ij})_{1\leqslant i,j\leqslant 2} : n'_{1+} = n_{1+}, n'_{2+} = n_{2+}, n'_{+1} = n_{+1}, n'_{+2} = n_{+2}\}$$

can be described as:

$$\mathscr{S} = \{(x, n_{1+} - x, n_{+1} - x, n_{2+} - n_{+1} + x) : \max\{0, n_{1+} + n_{+1} - n_{++}\} \leqslant x \leqslant \min\{n_{1+}, n_{+1}\}\}. \qquad \text{(B1)}$$

The lower and upper bounds for the (1,1) cell are identical to (2), on recognizing that $n_{+1} = y_{i0}$ and $n_{1+} = N_{i0}$. In equation (B1) the cells of a 2×2 table are arranged in the order (1,1), (1,2), (2,1), (2,2).

For two consecutive values of the (1,1) cell, say x and $x+1$, the difference between the corresponding 2×2 tables is

$$g=(1,-1,-1,1)$$

Therefore, for any two tables in the set $\mathscr{S}$, the difference between them is $\pm tg=\pm(t,-t,-t,t)$, where $1\leqslant t\leqslant \min\{n_{1+},n_{+1}\}-\max\{0,n_{1+}+n_{+1}-n_{++}\}$. It follows that $\{g,-g\}$ is a Markov basis for $\mathscr{S}$. Any set of moves that includes $\{g,-g\}$ is also a Markov basis for $\mathscr{S}$. Diaconis and Sturmfels [27] prove that a Markov basis associated with an arbitrary $I\times J$ table contains moves that have two entries equal to 1, two entries equal to -1, and the remaining entries equal to zero—see also Proposition 4.1 of [41]. In fact, Dobra [41] proves that moves of this type connect sets of multi-way tables defined by known marginals that are the cliques of decomposable graphs. These are a smaller class of graphs with special properties (e.g. they can be broken into components without losing information) that considerably simplify sampling and inference—see Lauritzen [42]. Markov bases based on decomposable graphs include, for example, sets of three-way tables with (i) all the three one-way marginals known; (ii) one one-way marginal known and the two-way marginal associated with the remaining variables known; (iii) two two-way marginals known. We may translate these into the ecological context. Suppose we have three binary variables which we label as disease outcome, exposure and gender. Then knowing the marginal disease, exposure and female/male rates across areas corresponds to (i). In addition if we know the exposure–gender margin (say) we obtain (ii), and knowing both the exposure–gender and the disease–gender margins (say) corresponds to (iii).

A case not covered by Dobra's results are sets of three-way tables defined by all three two-way marginals, since this is no longer a decomposable model. In the ecological setting this corresponds to knowing the exposure–gender, exposure–disease and gender–disease margins across areas. The corresponding graph has three vertices and it is complete (that is, all possible edges are present), while the fixed marginals are associated with its three edges. Hence, the simple scheme used in this paper would not work for data of this type. A Markov basis connecting such a set of tables would have to be identified using other methods, such as computational algebra [43]. Examples of more complex Markov bases can be found in the repository maintained by Kahle and Rauh (http://mbdb.mis.mpg.de). Once a Markov basis $\mathscr{M}$ is available for the set of tables $\mathscr{S}$, a Markov chain that involves moving from one table in $\mathscr{S}$ to another table in $\mathscr{S}$ proceeds as follows. We denote by n' the current state of the chain. We draw a move g' from the uniform distribution on $\mathscr{M}$. Let $n^*=n'+g'$ be a candidate table obtained by applying the selected move to n'. If n^* contains strictly negative entries, the chain stays at n'. Otherwise, since $\mathscr{M}$ is a Markov basis for $\mathscr{S}$, we must have $n^*\in\mathscr{S}$. The next state of the Markov chain is n^* with probability $\min\{1,\Pr(n^*)/\Pr(n')\}$. Otherwise the chain stays at n'.

Acknowledgements

The authors thank Dick Hoskins and Katrina Simmons for advice on the analysis of the BRFSS data, and Martyn Plummer and Kevin Quinn for providing details on their own computational methods for ecological inference. This research was supported by grants R01 CA095994 and R01 CA125081 from National Institutes of Health.

References

1. King G. *A Solution to the Ecological Inference Problem*. Princeton University Press: Princeton, 1997.
2. Salway R, Wakefield J. A comparison of approaches to ecological inference in epidemiology, political science and sociology. In *Ecological Inference*: *New Methodological Strategies*, King G, Rosen O, Tanner M (eds). Cambridge University Press: Cambridge, 2004.
3. Robinson WS. Ecological correlations and the behavior of individuals. *American Sociological Review* 1950; **15**:351–357.
4. Richardson S, Stucker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* 1987; **16**:111–120.
5. Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. *International Journal of Epidemiology* 1989; **18**:269–274.
6. Greenland S, Robins J. Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology* 1994; **139**:747–760.
7. Wakefield J. Ecologic studies revisited. *Annual Review of Public Health* 2008; **29**:75–90.
8. Raghunathan T, Diehr P, Cheadle A. Combining aggregate and individual level data to estimate an individual level correlation coefficient. *Journal of Educational and Behavioral Statistics* 2003; **28**:1–19.

9. Wakefield JC. Ecological inference for 2×2 tables (with Discussion). *Journal of the Royal Statistical Society*, *Series A* 2004; **167**:385–445.
10. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006; **25**:2136–2159.
11. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society*, *Series A* 2008; **171**:159–178.
12. Greiner D, Quinn K. $r\times c$ ecological inference: bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society*, *Series A* 2009; **172**:67–81.
13. Sheppard L, Prentice R. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics* 1995; **51**:853–863.
14. Haneuse S, Wakefied J. Hierarchical models for combining ecological and case–control data. *Biometrics* 2007; **63**: 128–136.
15. Haneuse S, Wakefield J. The combination of ecological and case–control data. *Journal of the Royal Statistical Society*, *Series B* 2008; **70**:73–93.
16. Breslow N, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society*, *Series B* 1997; **59**:447–461.
17. Scott A, Wild C. Fitting regression models to case–control data by maximum likelihood. *Biometrika* 1997; **51**:54–71.
18. Wakefield J, Haneuse S. Overcoming ecological bias using the two-phase study design. *American Journal of Epidemiology* 2008; **167**:908–916.
19. McCullagh P, Nelder J. *Generalized Linear Models*, *Second Edition*. Chapman & Hall: London, 1989.
20. Achen CH, Shively WP. *Cross-level Inference*. University of Chicago Press: Chicago, 1995.
21. King G, Rosen O, Tanner MA. Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research* 1999; **28**:61–90.
22. Byers S, Besag J. Inference on a collapsed margin in disease mapping. *Statistics in Medicine* 2000; **19**:2243–2249.
23. Wakefield J. Prior and likelihood choices in the analysis of ecological data. In *Ecological Inference*: *New Methodological Strategies*, King G, Rosen O, Tanner M (eds). Cambridge University Press: Cambridge, 2004; 13–50.
24. Martin A, Quinn K, Park J. *Package* ‘*MCMCpack*’ 2009.
25. Plummer M. Jags version 1.0.3 manual. *Technical Report*, 2009.
26. Neal R. Slice sampling (with Discussion). *Annals of Statistics* 2003; **31**:705–767.
27. Diaconis P, Sturmfels B. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* 1998; **26**:363–397.
28. Besag J. Bayesian inference for incomplete $2\times R\times C$ tables. Unpublished Manuscript, 2006.
29. Besag J, Clifford P. Generalized Monte Carlo significance tests. *Biometrika* 1989; **76**:633–642.
30. Dobra A. Markov bases for decomposable graphical models. *Bernoulli* 2003; **9**:1093–1108.
31. Dobra A, Tebaldi C, West M. Data augmentation in multi-way contingency tables with fixed marginal tables. *Journal of Statistical Planning and Inference* 2006; **136**:355–372.
32. Wakefield J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology* 2009; **38**:330–336.
33. Rosen O, Jiang W, King G, Tanner MA. Bayesian and frequentist inference for ecological inference: the $R\times C$ case. *Statistica Neerlandica* 2001; **55**(2):134–156.
34. Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in $r\times c$ contingency tables. *Journal of the American Statistical Association* 1983; **382**:427–434.
35. Booth JG, Butler JW. An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* 1999; **86**:321–332.
36. Chen Y, Diaconis P, Holmes SP, Liu JS. Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* 2005; **100**:109–120.
37. Chen Y, Dinwoodie IH, Sullivant S. Sequential importance sampling for multiway tables. *Annals of Statistics* 2006; **34**:523–545.
38. Guo SW, Thompson EA. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* 1992; **48**:361–372.
39. Forster JJ, McDonald JW, Smith PWF. Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society*, *Series B* 1996; **58**:445–453.
40. Caffo BS, Booth JG. A Markov chain Monte Carlo algorithm for approximating exact conditional probabilities. *Journal of Computational and Graphical Statistics* 2001; **10**:730–745.
41. Dobra A. Markov bases for decomposable graphical models. *Bernoulli* 2003; **9**:1–16.
42. Lauritzen SL. *Graphical Models*. Oxford University Press: Oxford, 1996.
43. Drton M, Sturmfels B, Sullivant S. *Lectures on Algebraic Statistics*, *Series*: *Oberwolfach Seminars*, vol. 39. Birkhäuser Verlag: Basel, Boston, Berlin, 2009.