A hierarchical aggregate data model with spatially correlated disease rates

Katherine A. Guthrie\textsuperscript{1,*}, Lianne Sheppard\textsuperscript{2,**}, and Jon Wakefield\textsuperscript{3,***}

\textsuperscript{1}Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, USA

\textsuperscript{2}Departments of Biostatistics and Environmental Health

\textsuperscript{3}Departments of Biostatistics and Statistics

Box 357232, University of Washington, Seattle, WA 98195-7232, USA

*\textit{email}: kguthrie@fhcrc.org

**\textit{email}: sheppard@u.washington.edu

***\textit{email}: jonno@u.washington.edu

Summary

The aggregate data study design (Prentice and Sheppard, 1995, \textit{Biometrika}) estimates individual-level exposure effects by regressing population-based disease rates on covariate data from survey samples in each population group. In this work, we further develop the aggregate data model to allow for residual spatial correlation among disease rates across populations. Geographical variation that is not explained by model predictors and has a spatial component often arises in studies of rare chronic diseases, such as breast cancer. We combine the aggregate and Bayesian disease-mapping models to provide an intuitive approach to the modeling of spatial effects while drawing correct inference regarding the exposure effect. Based on the results of simulation studies, we suggest guidelines for use of the proposed model.

\textbf{Key Words}: Ecological bias; Aggregate data analysis; Spatial dependence; Bayesian disease mapping.
1 Introduction

The aggregate data study design (Prentice & Sheppard, 1995) estimates exposure effects by regressing population-based disease rates on covariate data from survey samples in each population group, providing an alternative group-level analysis method to ecological studies in the estimation of individual-level health risks. The goal of this paper is to develop a hierarchical aggregate data model that allows for residual spatial dependence among disease rates across populations.

Ecological studies, which define a group or cohort as the unit of analysis, offer several advantages over individual-level studies. First, ecological studies provide a means of increasing statistical power for studying small disease risks, since large populations may be considered. Second, disease incidence and mortality rates for large groups such as those defined by cities, states, or countries, are often available through public-use datasets; therefore ecological studies can often be performed quickly and inexpensively. Third, group-level studies can estimate exposure effects that may be difficult to detect within any one group of individuals. Disease rates and exposures often show more variation between rather than within areas. In addition, measurement error may cause less bias in group-level than individual-level studies.

However, the aim of most epidemiological studies is to determine risks to the health of individuals. Thus, the main disadvantage of the ecological study design stems from its use of group-level data to draw conclusions about individuals. In a limited number of settings, these conclusions may be valid, but the lack of information about the within-group distribution of exposures and potential confounders in ecological studies can lead to bias (Greenland & Robins, 1994). For a review of ecological bias, see Morgenstern (1998).

The Prentice & Sheppard (1995) aggregate data study design addresses many of the potential pitfalls of ecological studies. The model is derived by aggregating a plausible individual-level
relative rate model within groups. Population-based disease rates are modeled as functions of individual-level covariate data. Thus, the aggregate model estimates the parameters of interest at the individual level. Given individual covariate values from at least a random subsample of each group, the aggregate model can control for potential confounders and include terms for possible effect modification. In addition, this estimation method is fairly robust to the effects of classical measurement error in the exposure of interest (Sheppard, Prentice, & Rossing, 1996) and nondifferential misclassification in confounders (Guthrie & Sheppard, 2001). While improving exposure effect assessment, the aggregate study design retains the ability of ecological studies to exploit group-level information in the data that is less realistic to obtain from single-area individual-level studies. In many practical settings, group-level analyses may provide unique insights into the origins of disease, complementing the knowledge gained from individual-level studies (Sheppard & Prentice, 1995).

We consider the context of environmental epidemiology, so that groups are defined by geographical region, and we are interested in the relationship between one or more environmental exposures and a disease outcome. An important consideration in the analysis of these data is the potential for spatially correlated disease rates. It is reasonable to assume that disease rates in neighboring areas may show similarities not attributable to measured covariates. These similarities may arise from patterns of cultural or ethnic affiliations, geographical characteristics or other unmeasured environmental risk factors across regions, and data anomalies such as double-counting of cases and under-enumeration at census. For example, rates of childhood leukemia show significant geographic variation, which has led to the development of several etiologic hypotheses (Ross et al., 1994). Breast cancer rates also vary widely across countries, with neighboring countries often showing similar rates (Anderson & Prentice, 1998). When modeling such rates in an ecological regression context, ignoring positive dependence in the outcomes may lead to inappropriate estimates of the
standard errors of the regression coefficients.

The original aggregate data model assumes independence of the disease rates across groups. Anderson & Prentice (1998) extended this work to allow for correlations between groups. In particular, the authors proposed an a priori partition of the population groups into known clusters, within which correlations are allowed between geographically contiguous regions. Simulation studies suggest that their generalized estimating equations approach gives rise to underestimation of the variance and correlation parameters. In addition, the need for a substantial number of independent clusters may be too restrictive for many group-level studies.

As an alternative, we apply a hierarchical model-based approach to estimation of the exposure effects in the presence of spatial correlation between groups. We begin in Section 2 by reviewing the two methods that are combined to create the proposed model: the aggregate data study design and the Bayesian disease-mapping model. In Section 3, we present the new model and describe the results of simulation studies under complete covariate data. Section 4 examines the effects of fitting the proposed model with subsamples of covariate data from each group. Finally, Section 5 provides a discussion of our findings.

## 2 Statistical framework

Suppose there are $k = 1, \ldots, K$ groups of interest, each consisting of $i = 1, \ldots, n_k$ individuals. We may stratify the outcome by confounders such as age and gender, so that $n_k = \sum_{s=1}^{S} n_{ks}$ for strata $s = 1, \ldots, S$. For simplicity, we consider a model with one exposure variable. Let $Y_{ksi}$ be a binary disease indicator, with $Y_{ksi} = 1$ for a case. Given a rare disease setting, we assume that

$$Y_{ksi} | \alpha, x_{ksi}, v_k \sim \text{Bernoulli}(e^{\gamma_s + x_{ksi} \beta + v_k})$$
where $\mathbf{v} = (v_1, \ldots, v_k)^T$ are independent and identically distributed (iid) with mean 0 and variance $\sigma_v^2$. The relative rate parameter for the exposure, $\beta$, is assumed to be constant across strata. We sum over individuals to model $Y_{ks}$, the observed number of cases in strata $s$, group $k$ as:

$$Y_{ks|\alpha, \mathbf{x}_k, v_k} \sim \text{binomial}\{n_{ks}, e^{\gamma_s p_k(\beta, \mathbf{x}_k) e^{v_k}}\},$$

where $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn_k})^T$ and $p_k(\beta, \mathbf{x}_k) = \int e^{x_k \beta} f_{ks}(x) \, dx$. If the distribution of the exposure is independent of stratum, we can write

$$p_k(\beta, \mathbf{x}_k) = \int e^{x_k \beta} f_k(x) \, dx. \hspace{1cm} (1)$$

We approximate the binomial by the Poisson model and sum over strata to obtain:

$$Y_k|\alpha, \mathbf{x}_k, v_k \sim \text{Poisson}\{E_k p_k(\beta, \mathbf{x}_k) e^{v_k}\}, \hspace{1cm} (2)$$

where $E_k = \sum_{s=1}^{S} n_{ks} e^{\gamma_s}$ is the expected number of events.

The aggregate data model assumes only the Bernoulli model for the individual outcomes, with

$$E(Y_k|\alpha, v_k) = E_k p_k(\beta, \mathbf{x}_k) e^{v_k}, \hspace{1cm} (3)$$

where $p_k(\beta, \mathbf{x}_k)$ is given by $n_k^{-1} \sum_{i=1}^{n_k} e^{x_{ki} \beta}$. The parameters are estimated by solving for $\alpha$ in estimating equations, with a moment estimator for $\sigma_v^2$ (Prentice & Sheppard, 1995). The method is extended to the more realistic scenario where only a subsample of covariate values is available by replacing the functions of the group covariates with the corresponding functions of the subsample data. For $m_k < n_k$, the sampled-covariate aggregate model is

$$E(Y_k|\alpha, v_k) = E_k p_k(\beta, \mathbf{x}_k^m) e^{v_k}. \hspace{1cm} (4)$$

where $\mathbf{x}_k^m = (x_{k1}, \ldots, x_{km_k})^T$ and $p_k(\beta, \mathbf{x}_k^m) = m_k^{-1} \sum_{i=1}^{m_k} e^{x_{ki} \beta}$. A correction for potential sampling error bias was derived by Prentice & Sheppard (1995).
correction term will be determined by the data characteristics. We discuss these conditions in Section 4.

Disease mapping is a process of describing geographical variation in disease incidence and mortality. A disease-mapping model may also include an ecological regression on potential risk factors for disease. Clayton & Kaldor (1987) were the first to apply a Bayesian hierarchical model to account for extra-Poisson variation in disease-mapping studies. A fully Bayesian version of this model is now a commonly used approach. The Besag, York, & Mollié (1991) formulation in an ecological regression context models the disease counts as Poisson random variates with

$$E(Y_k|\alpha, x_k, u_k, v_k) = E_k p_k(\beta, x_k) e^{u_k + v_k}.$$  \hspace{1cm} (5)

Here, $x_k$ denotes an area-level exposure value, so $p_k(\beta, x_k) = e^{x_k \beta}$. The sets of $u_k$ and $v_k$ represent spatially clustered and unstructured variation, respectively. See Wakefield, Best, & Waller (2000) for a review of the statistical aspects of this model. The random effects are intended to account for unknown or unmeasured disease risk factors. Each $u_k$ and $v_k$ are assumed to be independent of one another, as well as independent of the exposure distribution. Therefore, they should not be used to adjust for unmeasured group-level confounders (Wakefield & Salway, 2001).

Besag et al. (1991) assumed a Gaussian intrinsic autoregressive distribution for $u_k$:

$$u_k|u_l, l \in \delta k \sim N(\bar{u}_k, w_{k+}^{-1} \sigma_u^2)$$

where $\delta k$ indicates the set of neighboring areas, $w_{k+}$ is the number of neighbors, and $\bar{u}_k$ is the mean of these neighbors. To ensure identifiability of this random effect distribution one must assume that $\sum_{k=1}^{K} u_k = 0$, or alternatively fit the model without an intercept term.

The inclusion of the two sets of random effects can be thought of as a way of smoothing the disease rates at both a global and a local level. An estimate of the heterogeneity variance
parameter, $\sigma_v^2$, provides information about the overall non-spatial variability of the disease map. The clustering component, $\sigma_u^2$, describes the variability of disease risks relative to neighboring areas. The definition of the spatial random effects, $u_k$, in terms of the conditional distributions leads to computational efficiency. However, it also leads to some difficulty in the interpretation of the estimated magnitude of $\sigma_u^2$.

The regression coefficients are usually assumed to be distributed as uniform random variates with large domains; the limiting improper priors do not lead to an improper posterior distribution (Mollié, 1996). For the variance components, it is computationally convenient to assume conjugate gamma distributions on the inverse variance or precision quantities. Define $\lambda_u = \sigma_u^{-2}$ and $\lambda_v = \sigma_v^{-2}$, and assume $\lambda_u \sim \Gamma(a, b)$ and $\lambda_v \sim \Gamma(c, d)$ where $a$, $b$, $c$, and $d$ are known quantities. Since this is a fully Bayesian model, inference about the unknown parameters is based on the joint posterior distribution. The joint distribution, up to a normalizing constant, is analytically intractable, thus MCMC is used for estimation.

Richardson, Stücker, & Hémon (1987) proposed a two-term approximation to $p_k(\beta, x_k)$, given by equation (1). For example, if $\hat{\mu}_k$ and $\hat{\tau}_k^2$ are the sample mean and variance of $x_k$, then

$$p_k(\beta, x_k) \doteq \exp(\hat{\mu}_k \beta + \hat{\tau}_k^2 \beta^2 / 2).$$

(6)

This model was recently applied in a disease-mapping context with adjustment for residual spatial dependence by Best et al. (2001).
3 Complete covariate data model

3.1 Introduction

In this section, we present the hierarchical aggregate data model under the idealized setting of access to complete individual-level covariate data. This is not a realistic study design, but this assumption allows us to develop and assess our model before adding the complexity of the more practical sampled-covariate scenario.

3.2 The model

To specify the new model, we combine the original aggregate data model and the Bayesian disease-mapping model. That is, we assume a Poisson model for the observed number of cases with

\[ E(Y_k|\alpha, x_k, u_k, v_k) = E_k p_k(\beta, x_k) e^{u_k+v_k}, \]  

where \( p_k(\beta, x_k) = n_k^{-1} \sum_{i=1}^{n_k} e^{x_i \beta} \). Under the aggregate data model, the distribution of the observed number of events, \( Y_k \), is only approximately binomial (which is then approximated by the Poisson model). Since we consider the complete exposure data to be a fixed set of values for each group, the Bernoulli random variates are not independently distributed; therefore, the likelihood is not binomial (Wakefield & Salway, 2001). This assumption implies an approximation to the form of the conditional variance of \( Y_k \). For any given area, the difference between the binomial and the aggregate data model variance is proportional to the sample variance of the individual disease probabilities. We have examined the aggregate model approximation to the binomial variance via simulation and found it to be accurate. The approach of Richardson et al. (1987) assumes iid sampling from \( f_k(x) \), so this problem does not arise.
3.3 Parameter estimation

To estimate the parameters of the hierarchical aggregate data model we use a slight variation on Gibbs sampling. The full conditional distribution (fcd) of $\beta$ under the aggregate model introduces a level of complexity not present in the disease-mapping model structure. In the disease-mapping model with a uniform prior for $\beta$, the fcd for $\beta$ is normal (Mollié, 1996). Under the aggregate model, the fcd does not have a standard form. We simulate values of $\beta$ via a single-component Metropolis-Hastings algorithm. Under the conjugate priors, we can apply standard algorithms to sample from the fcd of $\lambda_u = \sigma_u^{-2}$ and $\lambda_v = \sigma_v^{-2}$. Since the univariate distributions for each $u_k$ and $v_k$ are log-concave, we sample from them using adaptive rejection sampling (Gilks & Wild, 1992).

3.4 Simulation studies

3.4.1 Independent disease rates

For the initial simulations, we assumed independence of disease risk across countries, so that the ‘clustering’ random effect term is excluded from this model. For each of 300 simulated datasets, we fit both the Prentice & Sheppard (1995) estimating equations and the MCMC estimation procedures. Our goal in these simulations is to understand the impact of different prior distributional assumptions on the relative consistency and efficiency of the two methods.

The data are motivated by studies relating breast cancer incidence and per capita dietary fat intake among women aged 55-69 in 21 countries. A complete description of these data can be found in Prentice & Sheppard (1995). The parameter values are $\gamma = -6.079$, $\beta = 0.002937$, and $\sigma_v^2 = 0.0476$. For each simulation run, we generated $n_k = 10,000$ independent individual fat calorie data points $x_{ki}$ from a log-normal distribution as $x_{ki} = 200 + \exp(z_{ki})$, where
\( z_{ki} \sim iid \, N \{ \log(\bar{x}_k - 200), 0.373^2 \} \). The values of \( \bar{x}_k \) ranged from 271 to 618 fat calories. We generated the random effects as \( v_k \sim iid \, N(0, \sigma_v^2) \). These choices gave a range of roughly 5 to 50 events per area, with an average of 19.

A Newton-Raphson algorithm was applied to estimate the parameters \( \alpha^T = (\gamma, \beta) \) in the estimating equations (EE) approach. We adapted a program in the C computing language called BEAM (Clayton, 1992) to fit the hierarchical aggregate data model. The MCMC estimates were generated from Markov chains of length 5000 with a burn-in period of 500 iterations. The chain length was chosen by graphing the estimates across iterations and judging when the chain appeared to have become stable. After checking several randomly chosen runs for convergence, we assumed that the chain length was sufficient for all subsequent simulations. The prior distribution for the precision of the random effects, \( \lambda_v = \sigma_v^{-2} \), was initially set to a \( \Gamma(2, 0.1) \), as chosen through an empirical process based on the expected residual variation in the crude relative rates of disease. Under this prior, \( \sigma_v \) takes on values between approximately 0.15 and 0.6. To assess the sensitivity of the exposure effect estimates to the choice of prior distribution for \( \beta \), we assumed first a uniform distribution and then a zero mean Gaussian distribution. We also examined the sensitivity of estimation to the choice of prior distribution of the random effects precision.

Table 1 shows the average exposure effect estimates from both the EE and MCMC methods, with the true value displayed in the column heading. Also shown are the sample standard deviations of the exposure effect estimates across simulations (simulation standard deviations), the estimation method-specific standard deviations of the estimates, estimates of the coverage probabilities for a 95% confidence or credible interval, and estimated random effects variances. The estimation standard deviation was calculated as the square root of the average variance estimate. For the EE procedure, the variance is estimated by the sandwich variance estimator. For the MCMC procedure, the variance is estimated by the sample vari-
ance of the posterior distribution of $\hat{\beta}$. Also for the MCMC results, the 2.5% and 97.5% quantiles of the posterior distribution of $\hat{\beta}$ are used to determine coverage of the 95% credible interval.

(Place Table 1 here)

Overall, the MCMC and EE estimates of $\beta$ are approximately unbiased, with the MCMC estimates slightly smaller on average than the EE estimates. The simulation estimates of variation are similar across methods. As in Prentice & Sheppard (1995), the average standard error estimates for the EE approach are smaller than the empirical standard deviations, and the coverage estimates are lower than the nominal 95% level. The MCMC estimates of coverage accurately reflect the 95% credible interval.

The second set of MCMC estimates in Table 1 are based on a Gaussian prior distribution for $\beta$ with standard deviation 0.75, which allows for rate ratios in the range of 0.2 to 5.0. The third set of MCMC estimates again assumed a Gaussian prior distribution for $\beta$, but increased the variance of the prior for $\lambda_v$ to nearly twice the previous value. The exposure effect estimates appear to be insensitive to these changes in the prior distributions of $\beta$ and $\lambda_v$.

### 3.4.2 Spatially dependent disease rates

For these simulation studies, we assumed the disease rates to be spatially dependent across areas, fitting the full model as shown in equation (7). Our goal in these simulations is to understand the impact on exposure effect estimation of modeling spatially correlated disease rates. We also seek to examine the performance of the aggregate data model in a more challenging setting, with smaller areas, smaller numbers of events per area, a larger amount of exchangeable variation, and spatially-correlated variation. We compare the exposure effect
estimates across various specifications of the prior precision component distributions, varying levels of spatial dependence, and two different distributions of the exposure across areas.

The data are based on the dietary fat and breast cancer study, using the spatial structure of the 88 counties of Ohio. For the first study, the mean exposure values, \( \bar{x}_k \), were sampled with replacement from the original set of 21 dietary fat calorie means. The distribution of these mean fat calorie values is skewed to the left, with no extreme values. For each simulation, we generated \( n_k = 5,000 \) independent exposures \( x_{ki} \) from a log-normal distribution as in the previous study.

We set \( \gamma = -5.8 \) and \( \beta = 0.002937 \). The heterogeneity random effects, \( (v_1, \ldots, v_{88}) \), were drawn independently from \( N(0, 0.32^2) \). For the clustering random effects, \( u = (u_1, \ldots, u_{88}) \), we assume \( u \sim N_{88}(0, V) \), with \( V \) as described below. This data formulation gives a range of 0 to about 80 events per area, with an average of 13.

We generated the elements of \( V \) using a model based on distances between county centroids, with

\[
V_{jl} = \omega_u^2 \exp(-d_{jl} \theta),
\]

where \( \omega_u^2 = 0.1 \), \( \theta = 0.016 \) initially, and \( d_{jl} \) = miles between county \( j \) and county \( l \) for all \( j, l = 1, \ldots, K; j \neq l \). The initial value for \( \theta \) was chosen such that the correlation \( \rho = \exp(-d_{jl} \theta) \) was 0.7 for \( d_{jl} = 20 \) miles, which approximates the average distance between county centroids in the state of Ohio. The state is approximately a 200 by 200 mile square. The resulting levels of spatial correlation by distance are shown in Figure 1. The parameter estimates were generated from Markov chains of length 10,000 with a burn-in period of 1000 iterations. We assumed \( \beta \sim N(0, 0.75^2) \) and \( \lambda_u \sim \Gamma(1.1, 0.11) \).

We first assessed the sensitivity of the exposure effect estimates to the choice of prior distribution of \( \lambda_u \). The second column of Table 2 denotes whether the variance component
prior distribution for the clustered random effects takes on lower, equal, or higher values on average than that assigned for the exchangeable random effects. The right side of the table shows estimates of the conditional variance of the spatial random effects, and the marginal variances of both exchangeable and spatially correlated random effects.

(Place Table 2 here)

The accuracy and precision of the exposure effect estimates are not affected by prior specification of the spatially correlated random effects variance component, as shown in Table 2. However, the variability in the random effects estimates is dependent on the specifications of the prior distributions. The overall variability of the random effects, as measured by the sum of the marginal random effects variances $\text{var}(\hat{u}) + \hat{\sigma}_v^2$, increases as $f(\lambda_u)$ shifts to the right. As expected, estimates of the spatial and non-spatial random effects variances are negatively correlated. Robustness of the exposure effect, and sensitivity of the random effects variance component estimates to the specification of $f(\lambda_u)$ have been seen previously in a disease-mapping setting (Richardson & Monfort, 2000).

Table 3 shows the results of simulations comparing the performance of the MCMC and EE approaches under various levels of spatial correlation. For all levels of correlation, the EE analysis assumes independent disease rates, and the MCMC analysis includes the spatially clustered random effects. For all MCMC analyses, we assume $\beta \sim N(0, 0.75^2)$, and both $\lambda_u$ and $\lambda_v \sim \Gamma(1.1, 0.11)$. The MCMC estimates of the exposure effect tend to be slightly smaller (less than 1%) on average than the EE estimates across all levels of spatial correlation. The MCMC estimates of the exposure effect tend to be 10% less variable, with respect to the simulation standard errors, than the EE estimates when $\rho > 0$. Overall, Table 3 shows that the accuracy and precision of the EE and the MCMC estimates do not differ substantially.

(Place Table 3 here)
In contrast, when we consider an alternative distribution of the exposure means (i.e. \( \bar{x}_k \)), the standard errors of \( \hat{\beta} \) given by the EE method are underestimated. Table 4 shows the results of a study in which the area-level exposure means are fixed at 375 for counties in the western half, and 575 for counties in the eastern half of the state. The individual-level exposures are simulated as in the previous study. When \( \rho = 0.9 \), the estimated variance of \( \hat{\beta} \) from the independence model is approximately 40% of the estimate under the spatial model. The estimated coverage of the 95% confidence interval is correspondingly low (65%). Under a lower level of spatial correlation (0.45), the underestimation is less severe.

(Place Table 4 here)

These findings are consistent with those of Heagerty & Lumley (2000) and Wakefield (2002), who referred to such covariate distributions as slowly (Table 4) or quickly (Table 3) varying with respect to space. Valid inference with regard to the exposure effect depends on the distribution of the exposure over space, as well as the level of spatial correlation. We note that accounting for apparent spatial dependence in the residuals will always be an important consideration when the goal of analysis is estimation of relative risks of disease in specific areas rather than exposure effect estimation.

4 Sampled-covariate data model

4.1 Introduction

We now consider the more realistic setting of having access to a subsample of covariate values from each group in the study. Our goal is to be able to apply the hierarchical aggregate data model to publicly available datasets, or in situations where reasonably sized sample surveys can be carried out.
There are many aspects that affect the ability of the sampled-covariate aggregate data model to correctly estimate the exposure effect. The size of the survey sample required for each group is inversely proportional to the expected number of disease cases, with the constant of proportionality based on the relative between-to-within-group variability in the covariates (Plummer & Clayton, 1996). A larger sample will be necessitated by a small number of disease cases, a large exposure effect, and greater within-group variation in the exposure distribution. A greater degree of asymmetry in the within-group exposure distributions will also increase the necessary size of the survey sample, since such a distribution requires accurate estimation of third- and higher-order moments (Wakefield & Salway, 2001).

### 4.2 Simulation studies

We characterize the performance of the sampled-covariate model under two different within-area exposure distributions in order to gain insights into aggregate study design considerations. In particular, we want to compare the accuracy of the exposure effect estimates from the complete-data aggregate model (7) to the sampled-covariate aggregate model:

\[
E(Y_k|\alpha, x_k^m, u_k, v_k) = E_k p_k(\beta, x_k^m) e^{u_k+v_k}
\]  

(8)

with \(p_k(\beta, x_k^m)\) defined in equation (4).

The first data scenario is equivalent to that used for the complete-data simulations in Table 3. In the second data scenario, the individual exposure measurements are sampled from area-specific normal distributions with the same means and variances as in the log-normal distributions. In both scenarios, we assume that the disease rates are spatially dependent with \(\rho = 0.45\). We compare the complete-data aggregate model with \(n_k = 5000\), to sampled-covariate analyses with \(m_k = 100\) and 50. We also compare the aggregate analysis results to those from ecological log-linear models (5) and parametric models (6), with the exposure
means and variances estimated from \( m_k \) covariate values.

Table 5 shows the average exposure effect estimates from the ecological and aggregate models under the log-normal exposure distribution. The results of the aggregate analyses show increasing attenuation in the exposure effect estimates as \( m_k \) decreases. When \( m_k = 100 \), the estimates are approximately 3.5% biased, and when \( m_k = 50 \), the estimates are approximately 6% biased, relative to the complete-data estimates. It is interesting to note that the standard deviation estimates vary only slightly with the size of \( m_k \), since the variance of the exposure effect estimates increased with smaller \( m_k \) under the original aggregate data model (Sheppard et al., 1996; Wakefield & Salway, 2001).

(Place Table 5 here)

The ecological analysis results show the effects of two competing sources of bias. With complete data, the ecological analysis yields positively biased exposure effect estimates. This finding is consistent with the results of Wakefield (2002) since the variance can be viewed as an unmeasured confounder with a positive effect. If the variance increases with the mean, as is the case here, overestimation of the exposure effect will result if the variance term is not included in the model. Sheppard (2002) saw this phenomena as well. This bias decreases as \( m_k \) gets small, showing an attenuation away from the ‘true’ ecological estimate. In general, the bias in an ecological analysis will depend on the within-group exposure distributions and the relationships between the exposures and confounders in the model.

Table 6 shows that the sampled-covariate aggregate exposure effect estimates are not as biased under the normal as under the log-normal exposure distribution. As expected, the symmetry of the normal within-area exposure distributions reduces the bias of the aggregate estimates compared to those under the log-normal distribution with equivalent means and variances. The estimates are less than 3% biased with a covariate sample of 50 values in
each group.

(Place Table 6 here)

The parametric results are nearly equivalent to the aggregate analysis results. Such attenuation toward the null under the parametric model was also seen in simulation studies by Wakefield & Salway (2001). The bias with $m_k = 50$ is due to the reduced accuracy in estimating the means and variances of the within-area distributions. The ecological estimates again show the effects of two competing biases, and are inefficient relative to the aggregate estimates.

Overall, these simulations show that there is potential for bias in the sampled-covariate aggregate exposure effect estimates, with the magnitude of attenuation related to the size of $m_k$ relative to $n_k$, as well as characteristics of the within-area exposure distributions. In these data scenarios, the bias is small. In both data scenarios, the aggregate model estimates are much more accurate than the ecological estimates.

## 5 Discussion

This paper describes a new model of group-level rare disease outcomes which incorporates elements of a Bayesian disease-mapping model structure into the aggregate data model. The hierarchical aggregate data model can overcome ecological bias by using subsamples of individual-level covariate data, and accounts for spatial dependence in the disease rates; parameters can be estimated using MCMC.

When estimating an exposure effect, an adjustment for spatial correlation in the outcome needs to be considered in the context of the exposure distribution over space. The results of simulation studies in Section 3.4.2 suggest that the importance of adjusting for spatial
correlation will depend on the strength of that correlation, as well as the similarity in the degree of spatial variation of the area-level exposure with that of the extra-Poisson variation in the outcome. If the area-level exposure means vary on a larger scale over space relative to the spatial structure of the residuals, failing to account for the dependent errors may lead to underestimation the exposure effect variance (see Table 4). If the spatial structures are similar, the exposure effect may be confounded by the inclusion of spatial random effects, as was seen in the data example of Clayton & Bernardinelli (1992). When the covariate distribution varies quickly with respect to space, assuming independence may be adequate with respect to exposure effect estimation, thus allowing for use of the less computer-intensive EE method (see Table 3).

In other work, we considered several approaches to correcting for sampling bias under incomplete covariate data, including kernel density estimation and a resampling algorithm to extrapolate information about the exposure distribution from the data samples (Guthrie, 2001). Especially under an exposure distribution with high within- relative to between-group variability, these approaches were not able to capture the shape and size of the density for the most extreme values. To explore the sensitivity of the model to the information in the tails, we modeled order statistics that were estimated from the complete covariate data. We found that the 10th percentiles as well as the 5th and 95th percentiles were not adequate to overcome the sampling bias, but including the 2.5th and 97th percentiles was sufficient. For our simulations, these percentiles suggest $m_k = 40$ as a minimum sample size for the covariate data in each area.

This work underscores the need for careful treatment of the model which relates exposure to outcome in a group-level setting. If there is a sufficient amount of variability in the exposure and other covariates between groups, relative to that within groups, then a fairly small sample of covariate data can yield accurate estimates in an aggregate data model.
The proposed hierarchical aggregate data model gives relatively unbiased estimates of the exposure effect using covariate data from a 1% sample (50 out of 5000) individuals per group, under a variety of data scenarios. The development of a sampling bias adjustment for less amenable exposure data distributions remains a topic for further investigation.

Acknowledgements

This research was supported by grant ES08062 from the National Institute of Environmental Health Sciences, and by the National Research Center for Statistics and the Environment.

References


Table 1: Exposure effect estimates from 300 simulations of breast cancer incidence and fat calorie intake among women aged 55-69 in 21 countries, assuming independence of disease rates across countries.

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>$\bar{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
<th>$\bar{\sigma}_v^2 \times 10^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMC: Priors for $\beta, \lambda_v$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform, $\Gamma(2, 0.1)$</td>
<td>2.902</td>
<td>0.570</td>
<td>0.583</td>
<td>96</td>
<td>4.39</td>
</tr>
<tr>
<td>$N(0, 0.56)$, $\Gamma(2, 0.1)$</td>
<td>2.902</td>
<td>0.570</td>
<td>0.583</td>
<td>96</td>
<td>4.39</td>
</tr>
<tr>
<td>$N(0, 0.56)$, $\Gamma(1.1, 0.06)$</td>
<td>2.903</td>
<td>0.569</td>
<td>0.579</td>
<td>95</td>
<td>4.39</td>
</tr>
<tr>
<td>EE</td>
<td>2.935</td>
<td>0.573</td>
<td>0.505</td>
<td>88</td>
<td>3.27</td>
</tr>
</tbody>
</table>
Table 2: Exposure effect estimates from 300 simulations of spatially correlated disease rates across the 88 counties of Ohio, with $\rho=0.7$ between adjacent counties. The prior distributions are $\beta \sim N(0, 0.75^2)$ and $\lambda_v \sim \Gamma(1.1, 0.11)$.

<table>
<thead>
<tr>
<th>$f(\lambda_u)$</th>
<th>$f(\lambda_v)$</th>
<th>Weight of $f(\lambda_u)$</th>
<th>$\hat{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
<th>$\bar{\sigma_u}$</th>
<th>var($\hat{u}$)</th>
<th>$\bar{\sigma_v}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform $&lt; f(\lambda_v)$</td>
<td>2.920 0.425 95 0.004 0.002 0.130 2.917 0.422 94 0.162 0.067 0.092 2.918 0.427 95 0.240 0.090 0.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Gamma(1.1, 0.11) = f(\lambda_v)$</td>
<td>2.917 0.421 94 0.162 0.067 0.092</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Gamma(1.1, 0.33) &gt; f(\lambda_v)$</td>
<td>2.918 0.427 95 0.240 0.090 0.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Exposure effect estimates from 300 simulations of spatially correlated disease rates across the 88 counties of Ohio, with area-level exposure means varying randomly. The EE method assumes independent disease rates, while the MCMC method assumes spatially correlated disease rates. For the MCMC estimation, the prior distributions are $\beta \sim N(0, 0.75^2)$ and $\lambda_u, \lambda_v \sim \Gamma(1.1, 0.11)$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>Method</th>
<th>$\bar{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>MCMC</td>
<td>2.932</td>
<td>0.429</td>
<td>0.457</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td><strong>2.956</strong></td>
<td><strong>0.425</strong></td>
<td><strong>0.447</strong></td>
<td><strong>95</strong></td>
</tr>
<tr>
<td>0.45</td>
<td>MCMC</td>
<td>2.905</td>
<td>0.460</td>
<td>0.441</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>2.942</td>
<td>0.483</td>
<td>0.443</td>
<td>91</td>
</tr>
<tr>
<td>0.7</td>
<td>MCMC</td>
<td>2.917</td>
<td>0.434</td>
<td>0.422</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>2.943</td>
<td>0.466</td>
<td>0.427</td>
<td>92</td>
</tr>
<tr>
<td>0.9</td>
<td>MCMC</td>
<td>2.921</td>
<td>0.376</td>
<td>0.402</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>2.937</td>
<td>0.394</td>
<td>0.403</td>
<td>95</td>
</tr>
</tbody>
</table>
Table 4: Exposure effect estimates from 300 simulations of spatially correlated disease rates across the 88 counties of Ohio, with one area-level exposure mean each for the eastern and western halves of the map. The EE method assumes independent disease rates, while the MCMC method assumes spatially correlated disease rates. For the MCMC estimation, the prior distributions are $\beta \sim N(0, 0.75^2)$ and $\lambda_u, \lambda_v \sim \Gamma(1.1, 0.11)$.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>Method</th>
<th>$\bar{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.45</td>
<td>MCMC</td>
<td>2.923</td>
<td>0.592</td>
<td>0.618</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>2.945</td>
<td>0.616</td>
<td>0.385</td>
<td>77</td>
</tr>
<tr>
<td>0.9</td>
<td>MCMC</td>
<td>2.819</td>
<td>0.570</td>
<td>0.563</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>2.861</td>
<td>0.658</td>
<td>0.338</td>
<td>65</td>
</tr>
</tbody>
</table>
Table 5: Exposure effect estimates from 300 simulations of spatially correlated disease rates across the 88 counties of Ohio, using complete-data aggregate (7), sampled-covariate aggregate (8), and ecological regression (equation (5) with $x_k = \bar{x}_k$) analyses. Simulated disease rates based on 5000 individuals per area. The within-area exposure distribution is log-normal.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$m_k$</th>
<th>$\bar{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate</td>
<td>5000</td>
<td>2.930</td>
<td>0.413</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>2.824</td>
<td>0.405</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>2.757</td>
<td>0.404</td>
<td>93</td>
</tr>
<tr>
<td>Ecological</td>
<td>5000</td>
<td>3.438</td>
<td>0.575</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3.377</td>
<td>0.573</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>3.324</td>
<td>0.566</td>
<td>89</td>
</tr>
</tbody>
</table>
Table 6: Exposure effect estimates from 300 simulations of spatially correlated disease rates across the 88 counties of Ohio, complete-data aggregate (7), sampled-covariate aggregate (8), parametric (6), and ecological regression (equation (5) with $x_k = \bar{x}_k$) analyses. Simulated disease rates based on 5000 individuals per area. The within-area exposure distribution is normal.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>$\hat{\beta} \times 10^3$</th>
<th>SD($\hat{\beta}$) $\times 10^3$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2.951</td>
<td>0.459</td>
<td>95</td>
</tr>
<tr>
<td>100</td>
<td>2.913</td>
<td>0.460</td>
<td>95</td>
</tr>
<tr>
<td>50</td>
<td>2.887</td>
<td>0.460</td>
<td>96</td>
</tr>
<tr>
<td>Parametric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2.952</td>
<td>0.459</td>
<td>95</td>
</tr>
<tr>
<td>100</td>
<td>2.914</td>
<td>0.459</td>
<td>95</td>
</tr>
<tr>
<td>50</td>
<td>2.879</td>
<td>0.456</td>
<td>96</td>
</tr>
<tr>
<td>Ecological</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>3.353</td>
<td>0.579</td>
<td>87</td>
</tr>
<tr>
<td>100</td>
<td>3.304</td>
<td>0.574</td>
<td>89</td>
</tr>
<tr>
<td>50</td>
<td>3.252</td>
<td>0.567</td>
<td>91</td>
</tr>
</tbody>
</table>