STATISTICS IN MEDICINE
Statist. Med. **17**, 2025–2043 (1998)

SPATIO-TEMPORAL MODELS WITH ERRORS IN COVARIATES: MAPPING OHIO LUNG CANCER MORTALITY

HONG XIA AND BRADLEY P. CARLIN*

Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A.

SUMMARY

In estimating spatial disease patterns, as well as in related assessments of environmental equity, regional morbidity and mortality rate maps are widely used. Hierarchical Bayes methods are increasingly popular tools for creating such maps, since they permit smoothing of the fitted rates toward spatially local mean values, with more unreliable estimates (those arising in low-population regions) receiving more smoothing. In this paper we blend methods for spatial-temporal mapping with those for handling errors in covariates in a single hierarchical model framework. Estimated posterior distributions for the resulting highly-parameterized models are obtained via Markov chain Monte Carlo (MCMC) methods, which also play a key role in our approach to model evaluation and selection. We apply our approach to a data set of county-specific lung cancer rates in the state of Ohio during the period 1968–1988. Our model uses age-adjusted death rates, and incorporates recent information regarding smoking prevalence, population density, and the socio-economic status of the counties. This information is critical to understanding the role played by a certain depleted uranium fuel processing facility on the elevated lung cancer rates in the counties that neighbour it. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

The issue of *environmental justice* (that is, the equitable distribution of exposures to and adverse outcomes from environmental hazards among various socio-demographic subgroups[1]) is attracting increasing public attention. Through accurate assessment of environmental justice, the biostatistician can contribute directly to fair and equitable protection for all persons, regardless of age, ethnicity, gender, race, or socio-economic status.

Issues of environmental justice inherently involve locations of environmental hazards and groups of people. Thus, maps of geographical variation in disease occurrence are important tools for identifying areas with potentially elevated risk, determining spatial patterns, and formulating

* Correspondence to: Bradley P. Carlin, Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A. E-mail: brad@muskie.biostat.umn.edu

Contract/grant sponsor: National Institute of Allergy and Infectious Diseases
Contract/grant number: 1-R29-AI33466
Contract/grant sponsor: National Institute of Environmental Health Sciences
Contract/grant number: 1-R01-ES07750

CCC 0277–6715/98/182025–19$17.50
© 1998 John Wiley & Sons, Ltd.

Received November 1996
Revised September 1997

or validating aetiological hypotheses about disease. As a result, disease maps play a central role in establishing a credible scientific foundation for evaluating environmental justice.

In recent years, there have been many efforts to map incidence and mortality from diseases such as lung cancer. One of the main problems has been the choice of the appropriate measure of cancer incidence or mortality to map. Atlases usually have displayed either measures of relative risk in each district, as measured by the district-specific standardized mortality ratio (SMR), or the statistical significance level for a test of the difference between the rates in the district and some reference standard. Neither of these approaches is fully satisfactory. When the disease is rare or the population size is small, Poisson variation in the area-specific counts causes maps based directly on the raw SMRs to show a misleading picture of the true underlying relative risks. The most extreme estimates will dominate the pattern of a map, even though these are typically the least stable. Moreover, this approach also fails to account for the anticipated similarity of relative risks in nearby or adjacent regions. On the other hand, mapping statistical significance alone completely ignores the size of the corresponding effect, so that on the map, the most extreme areas may simply be those with the largest populations.[2]

Waller *et al.*[3] review the literature on Bayes and empirical Bayes approaches for modelling and mapping disease rates, as now summarized. Clayton and Kaldor[2] and Manton *et al.*[4] introduced empirical Bayes approaches that account for spatial similarities among neighbouring and nearby rates. Ghosh[5] and Devine *et al.*[6–8] consider constrained empirical Bayes approaches which produce improved estimates of the true distribution of the unknown rates. Cressie and Chan[9] fit Markov random field models to a sudden infant death syndrome (SIDS) data set. Clayton and Bernardinelli[10] review Bayesian methods for modelling regional disease rates. Besag *et al.*[11] describe a fully Bayesian approach which separates spatial effect from overall heterogeneity in the rates. Bernardinelli and Montomoli[12] compare empirical and fully hierarchical Bayes methods, and conclude that the latter offer greater flexibility and convenience in the statistical analysis of geographical variation in disease rates. Breslow and Clayton[13] place the disease mapping problem within the larger framework of generalized linear mixed models and provide approximation schemes for inference.

More recently, Bernardinelli *et al.*[14] employ a spatial ecological regression model to account for imprecisely observed covariates. Bernardinelli *et al.*[15] address the issue of choosing the hyperprior distribution of the dispersion parameter in the fully Bayesian approach to disease mapping. Bernardinelli *et al.*[16] also perform an analysis of variation of risk for a given disease in space and time. These authors propose a fully Bayesian model in which both area-specific intercept and trend are modelled as random effects with allowance for correlation among them. Ghosh *et al.*[17] consider hierarchical Bayes generalized linear models for a unified analysis of both discrete and continuous data for small area estimation.

Most recently, Waller *et al.*[3] extended hierarchical spatio-temporal modelling to accommodate general temporal effects and space–time interactions. This method requires careful implementation via Markov chain Monte Carlo (MCMC) methods, which also permit related new approaches to model validation and comparison. The method is illustrated using a data set of annual county-specific lung cancer mortality rates during the period 1968–1988 in the state of Ohio. This work was motivated by a previous spatial-only analysis of the same data set,[18] which fit Poisson spatial models incorporating the sex and race covariates to each year's data separately.

The Ohio data set was originally studied by Devine,[19] who notes that its collection was motivated both by public concern and empirical evidence that lung cancer mortality rates tended to be elevated in the vicinity of the U.S. Department of Energy Fernald Materials Processing Center, located in the southwest corner of Ohio about 25 miles northwest of Cincinnati. The Fernald plant recycles depleted uranium fuel from U.S. Department of Energy and Department of Defense nuclear facilities, a process which creates a large amount of uranium dust. During peak production years (roughly 1951 to the early 1960s), some radioactive dust particles may have been released into the air due to inadequate filtration and ventilation systems. Lung cancer is therefore of interest because inhalation is believed to be the major exposure pathway for off-site populations, and because it is the most prevalent form of cancer potentially associated with exposure to uranium.

Unfortunately, the data-analytic conclusions of Waller *et al.*[3] are unsatisfying for two reasons. First, their analysis fails to include important covariate information, most notably on the prevalence of cigarette smoking (though they do appreciate the importance of this information, and state that they will include it in future modelling efforts). Second, their failure to include age in the model, either explicitly as a predictor or implicitly by age-standardizing the crude rates, appears to lead to the anomalous result that non-white females emerge as the healthiest socio-demographic subgroup (in fact, we show below that this difference is statistically significant under their model). Because death rates for most diseases (including lung cancer) generally increase with age, a population group having more older individuals will tend to have greater crude death rates. A comparison of the crude mortality rates could thus be distorted by the different age distributions. Personal communication with Ohio Department of Health staff encourages an analysis incorporating age standardization, to see if non-whites are in fact dying at younger ages.

In this paper we correct these two deficiencies by absorbing the errors in covariates idea proposed by Bernardinelli *et al.*[14] into the spatio-temporal framework of Waller *et al.*[3] Section 2 constructs the basic spatial models and investigates several potentially important covariates using the 1988 data only. Section 3 introduces the errors in covariates aspect of the model, which allows for both sampling error and spatial correlation in the covariate. After selecting an appropriate set of covariates, Section 4 fits the full spatio-temporal model with errors in covariates and discusses the results, highlighting significant differences from those obtained by Waller *et al.*[3] Section 5 addresses the issues of model selection and adequacy, which are complicated in our case by the presence of random effects and improper prior distributions for certain model parameters. Finally, Section 6 discusses our findings and suggests avenues for future research. (See also the next paper in this issue by Knorr-Held and Besag, pp. 2045–2060.)

2. SPATIAL MODELLING WITH COVARIATES

2.1. Description of data set

The Ohio lung cancer data set consists of C_{ijkt}, the numbers of lung cancer deaths in county i for gender j and race k (white and non-white) during year t, and n_{ijkt}, the corresponding population counts, where $i = 1, \dots, I$, $j = 1, 2$, $k = 1, 2$ and $t = 1, \dots, T$. These data were originally taken from the National Center for Health Statistics (NCHS) Compressed Mortality File,[20] which provides age-specific death counts by underlying cause and population estimates for every county in the U.S. Our subset of lung cancer data are recorded for each of the $I = 88$ Ohio counties over the period 1968–1988 (that is, $T = 21$). In addition, lung cancer deaths in each category are partitioned into 11 age classes: < 5, 5–9, 10–14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and $\geqslant 85$ years.

2.2. Description of covariates

It is well known that smoking is a very important risk factor for lung cancer. Other factors, such as gender, race, age, urban living and socio-economic status (SES), may also be involved. Gender, race and age information is available directly from our data set, but for the remaining covariates, we must rely on other sources.

In environmental epidemiology, covariates are often difficult to measure directly and can at best be assessed in only a small fraction of the study population. Therefore, investigators must rely on imperfect surrogate measures of the true covariates. Our cigarette smoking data is a case in point, coming from the Ohio Behavioral Risk Factor Surveillance System. These data, a total of 9525 observations, were collected by telephone for persons 18 years of age and older from 1988 to 1994. The summary data include county-level weighted numbers of current smokers, former smokers and non-smokers. To adjust the prevalence estimate to reflect the demographic composition of the Ohio population for persons over age 18, a weight is assigned to each respondent based on age, race, gender, and probability of selection in the survey. Aggregating the survey data over the 7-year survey period, we use the proportion of current smokers in county i as a surrogate for the true smoking proportion in this county. With over half of the counties having fewer than 50 respondents, the accuracy of these data is very much in doubt.

The two other important potential covariates for lung cancer risk, urban living and SES, can be obtained from the Southwest Ohio Regional Data Center home page on the world wide web (http://www.uc.edu/ ~ sordc/index.html). Here we find total resident population per square mile in 1992 as a proxy for urban living, and income per capita in 1989 as a proxy for SES, both recorded at the county level.

2.3. Model development

We start with the basic spatial model for the 1988 data only, and as such we suppress the subscript t for now. Since the data are lung cancer death counts by gender and race, an additive log-linear model with a Poisson likelihood is appropriate. We add the additional covariates age, smoking status, urban living, and SES to construct a Poisson spatial model with covariates.

We begin by adjusting for the age covariate. One approach is to use external standardization to obtain the expected deaths E_{ijk}, applying age group-specific reference death rates to the area-specific population subdivided by age.[12] However, since we lack an appropriate reference mortality table, we adopt the simpler alternative of age-standardizing the observed death counts C_{ijk}, and subsequently obtaining the expected deaths E_{ijk} through internal standardization. More explicitly, Table I shows the United States standard million age distribution, taken from the 1970 U.S. Census. We calculated adjusted deaths $C^*_{ijk} = \sum_l a_l C_{ijkl} / \sum_l a_l$, where a_l comes from the second column of Table I with l indexing the age groups, $l = 1, \ldots, 11$. We then set $E_{ijk} = Rn_{ijk}$, where $R = (\sum_{ijk} C^*_{ijk} / \sum_{ijk} n_{ijk})$, the statewide (age-adjusted) death rate. We emphasize that, while our rates are still internally standardized, unlike previous analyses of this data set,[3,18] they are now age-adjusted.

Letting C^*_{ijk} be the number of observed age-adjusted deaths in county i for sex j and race k in 1988, we assume

$$C^*_{ijk} \sim \text{Poisson}(E_{ijk} \exp(\mu_{ijk})) \quad (1)$$

where E_{ijk} are the expected death counts and

$$\mu_{ijk} = \mu + s_j\alpha + r_k\beta + s_j r_k \xi + q_i\rho + u_i\delta + v_i\kappa + \theta_i + \phi_i \quad (2)$$

Table I. U.S. standard million age distribution (source: U.S. Bureau of the Census, Census of Population 1970)

Age	Standard million (a_l)
All ages	1,000,000
< 5	84,416
5–9	98,204
10–14	102,304
15–24	174,406
25–34	122,569
35–44	113,614
45–54	114,265
55–64	91,480
65–74	61,195
75–84	30,112
85+	7,435

where s_j indicates sex (0 if male, 1 if female), r_k indicates race (0 if white, 1 if non-white), q_i denotes the current smoking proportion observed in our sample survey in county i, and u_i and v_i denote the urban living and SES covariates in county i, respectively. To avoid a potentially strong *a posteriori* correlation between μ and (ρ, δ, κ), we centred the covariates q_i, u_i and v_i around their means. The fixed effects α, β and ξ capture the effect of sex, race, and sex–race interaction, respectively, while μ is the overall log-relative risk for all the counties. Finally, the two random effect terms θ_i and ϕ_i capture the effect of unstructured heterogeneity and geographical clustering, respectively, as we now describe.

Since μ, α, β, ξ, ρ, δ and κ can be identified by the likelihood, we use a flat prior on these parameters, but since only the sum of the heterogeneity and clustering parameters, $\theta_i + \phi_i$, is identified by the likelihood for a given i, we require proper priors here. We thus assume

$$\theta_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1/\tau) \quad \text{and} \quad \phi_i | \phi_{j \neq i} \sim \mathrm{N}(\mu_{\phi_i}, \sigma^2_{\phi_i}),\ i = 1, \ldots, I \tag{3}$$

where

$$\mu_{\phi_i} = \frac{\sum_{j \neq i} w_{ij} \phi_j}{\sum_{j \neq i} w_{ij}} \quad \text{and} \quad \sigma^2_{\phi_i} = \frac{1}{\lambda_\phi \sum_{j \neq i} w_{ij}}.$$

The weights w_{ij} are fixed constants that measure the proximity of counties i and j. Specifically, we take $w_{ij} = 1$ if areas i and j are adjacent (that is, share a common boundary), and $w_{ij} = 0$ otherwise. Other forms of w_{ij} are potentially relevant.[6–9] For brevity, we write this *conditionally autoregressive* prior in vector notation as $\boldsymbol{\phi} \sim \mathrm{CAR}(\lambda_\phi)$. Since this CAR prior is translation invariant, to identify the grand mean μ we also add the constraint $\sum_{i=1}^{I} \phi_i = 0$.

The relative magnitude of the heterogeneity and clustering effects is controlled by the two scale parameters τ and λ_ϕ, for which we also require proper prior specifications. Here we choose gamma(a, b) and gamma(c, d) hyperpriors, respectively – flexible forms which have the benefit of being conjugate with the prior structure (3). Advice in Bernardinelli *et al.*[15] on how to make these choices 'equally informative' suggested a gamma(1, 100) and gamma(1, 7) for τ and λ_ϕ, respectively.

2.4. Computing

Estimated posterior distributions for the parameters in our model may now be obtained via MCMC methods. The aforementioned conjugacy of our hyperpriors results in *full conditional distributions* (that is, the distribution of the parameter in question given all remaining parameters) for τ and λ_ϕ that are also gamma-distributed, so that Gibbs sampling[21] may be adopted for these two parameters. However, no closed forms are available for the full conditionals of the remaining parameters, so they must be updated using Metropolis sampling (see, for example, Carlin and Louis,[22] Section 5.4.3, or Chib and Greenberg[24]). The intercept-identifying constraint $\sum_{i=1}^{I} \phi_i = 0$ is implemented simply by recentring the current $\boldsymbol{\phi}^{(g)}$ vector around 0 at the end of each MCMC iteration g, as recommended for example by Besag *et al.*[23]

We implemented all of the models in this paper using the C programming language and S-plus graphics routines. In each case, we ran five independent, initially overdispersed sampling chains of the appropriate Gibbs–Metropolis algorithm until convergence. Figure 1 shows a typical convergence plot, displaying sample traces for all of the model's fixed effects and variance components, its log-likelihood score, and a judiciously chosen subset of its random effects (here, representing both an urban and a rural county). The traces for each parameter are annotated with the median and 97·5 per cent upper bound of the Gelman and Rubin[25] convergence diagnostic (abbreviated 'G&R'), and the lag 1 sample autocorrelation in the third (middle) chain (abbreviated 'lag 1 acf').

2.5. Results

Using the above model and prior structure, we ran our MCMC sampler for 2200 iterations, discarding the first 200 samples from each chain as preconvergence burn-in. Thus, we used a total of 10,000 samples to compute the posterior summaries. We discovered that our data could not identify an effect of either urban living or SES. Figure 2 shows spatial distributions for the three area-specific covariates. Note that urban living and SES show a high degree of collinearity (that is, they exhibit very similar spatial patterns). Lung cancer mortality rates are occasionally high in the most populous counties, which are also very wealthy. On the other hand, some rural counties (which often have lower SES) also have higher rates. Overall, both covariates behave something like an intercept, leading to collinearity with each other (and with the grand mean μ), hence MCMC convergence failure.

On the contrary, smoking distribution in the map appears to be spatially different from the other two covariates, and in fact does emerge as the most statistically significant predictor of the three. By removing the urban living and SES terms from model (2), we obtained a readily convergent algorithm (Figure 1) and precise estimates of the remaining parameters in our model. Finally, we drew crude age-adjusted and fitted disease rate maps, with the latter made by plugging the estimated posterior medians of all the parameters into $R\exp(\mu_{ijk})$, the disease rate predicted by the model. Our findings were consistent with those obtained by Xia *et al.*[18] In the smoothed disease map, we found that the higher rates occur in the counties in some populous areas containing big cities, and also along the rural eastern border with West Virginia. Perhaps the common link between those areas is a larger smoking population, a higher percentage of non-agricultural employment, or poorer access to quality health care.

Finally, like Waller *et al.*[3] we found that the posterior medians of the heterogeneity parameters θ_i were tightly centred near zero, suggesting no significant additional heterogeneity in the data beyond that explained by the CAR prior. In the remainder of this paper, we thus remove these terms from our log relative risk models.

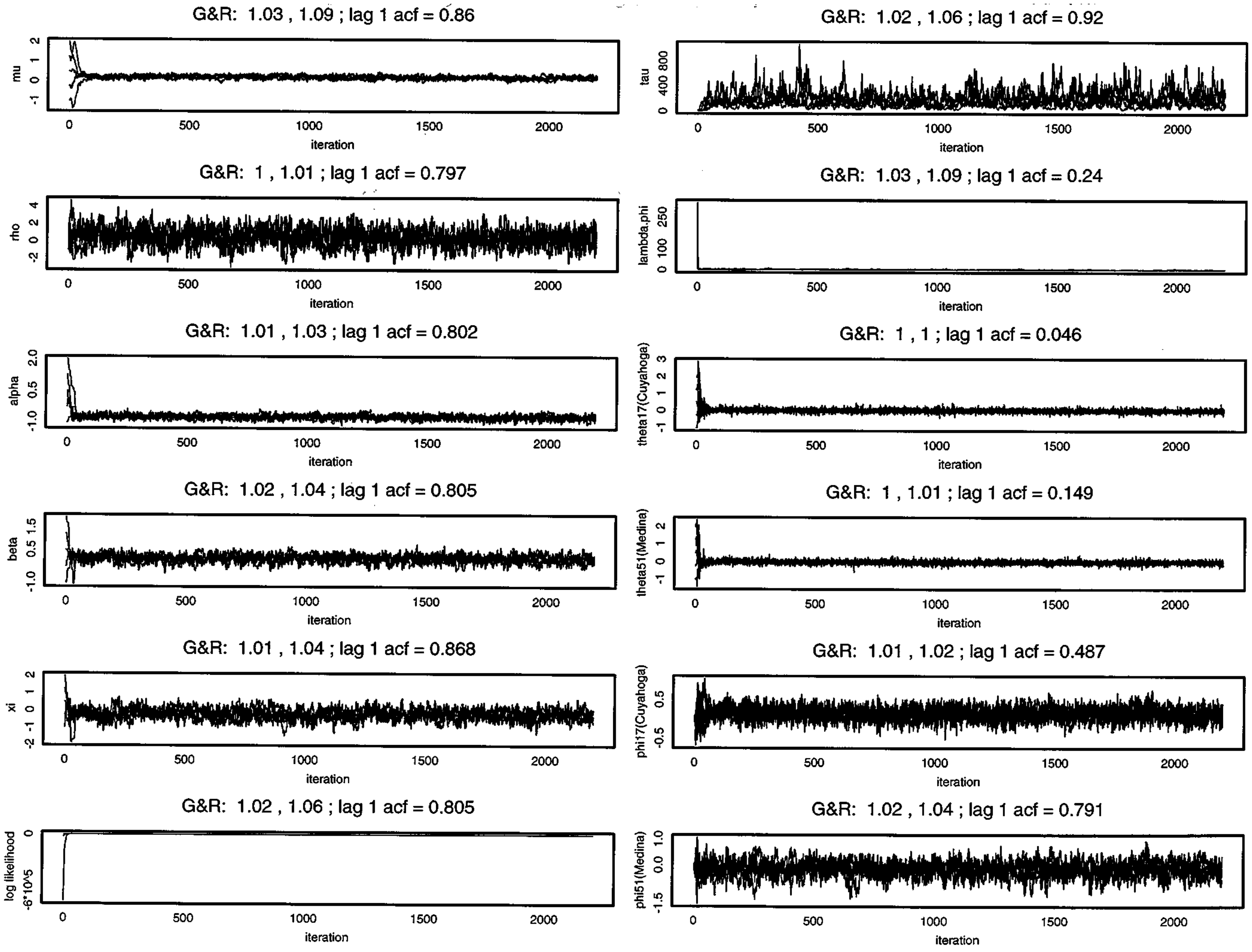


Figure 1. Sample convergence monitoring plot; 5 independent chains × 2200 iterations

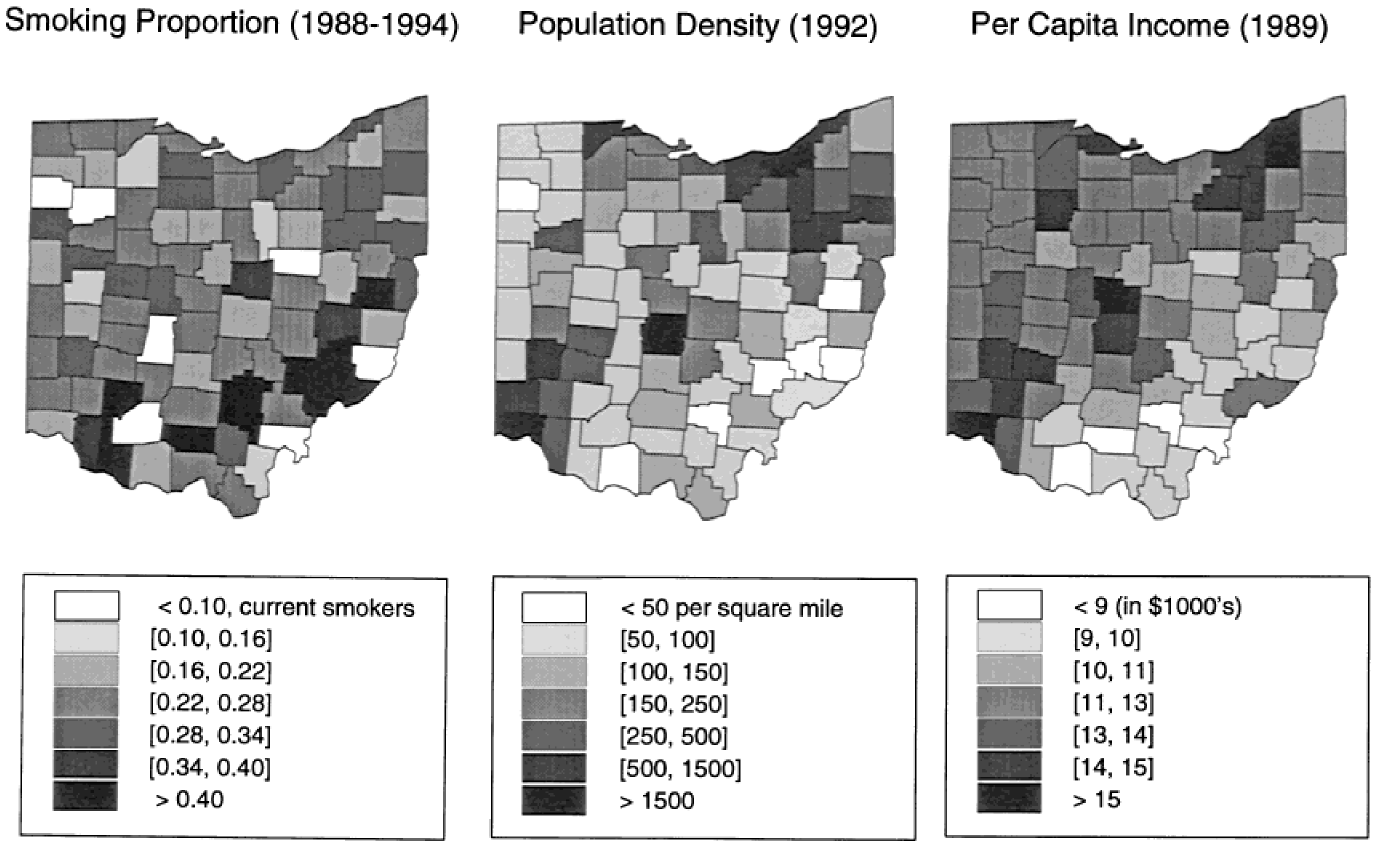


Figure 2. Spatial distributions for three ecological covariates

3. ERRORS IN COVARIATES MODEL

3.1. Introduction

The simple Poisson spatial model constructed in the last section incorporates spatial correlation through the prior distributions for estimating the underlying true log relative risk parameter μ_{ijk}. In this way each estimate of μ_{ijk} can 'borrow strength' from neighbouring areas so that the empirical map is smoothed, and geographical trends and epidemiologic inferences become more reliable.

The above model implicitly assumes that all of the covariates are measured without error, but in our case, the covariate (smoking proportion) cannot be observed directly. The observed smoking proportion q_i is an imperfect measurement or proxy for the underlying true smoking proportion, say, p_i. Treating a surrogate as the true covariate may produce biased results and unrealistically narrow interval estimates unless this surrogate is a very accurate measure of the true covariate.[26] This motivates an *errors in covariates* version of our model, wherein we replace the observed q_i in equation (2) with the unobserved p_i, and subsequently model the connection between the two.

In this new model, spatial smoothing priors are posited for both the clustering parameters (ϕ_i) and the true smoking covariates (p_i). In conjunction with the Poisson likelihood (1) and the simplifications adopted in the previous section, the log-relative risk becomes

$$\mu_{ijk} = \mu + s_j\alpha + r_k\beta + s_jr_k\xi + p_i\rho + \phi_i. \tag{4}$$

(Note that q_i in (2) has been replaced by p_i.) Following the approach of Bernardinelli *et al.*,[14] we introduce both sampling error and spatial correlation into the smoking covariate. Let

$$q_i|p_i \sim \mathrm{N}(p_i, \sigma^2_{q0}), \quad i = 1, \dots, I \tag{5}$$

and

$$\mathbf{p} \sim \mathrm{CAR}(\lambda_p) \Leftrightarrow p_i|p_{j\neq i} \sim \mathrm{N}(\mu_{p_i}, \sigma^2_{p_i}), \quad i = 1, \dots, I \tag{6}$$

where $\mu_{p_i} = \sum_{j\neq i} w_{ij}p_j / \sum_{j\neq i} w_{ij}$ and $\sigma^2_{p_i} = (\lambda_p \sum_{j\neq i} w_{ij})^{-1}$. Note that the amount of smoothing in the two CAR priors (3) and (6) may differ, since the smoothing is controlled by different parameters λ_ϕ and λ_p. Like λ_ϕ, λ_p is also assigned a gamma hyperprior – namely, a gamma(e,f). The additional full conditionals for p_i, like those for ϕ_i, are non-conjugate, but there is no additional conceptual difficulty in adding the errors in covariates aspect to the model.

3.2. Results

Posterior estimates were calculated for the above model (1) and (3)–(6). Recalling that the p_i's and q_i's are bounded between 0 and 1, we chose a fairly vague gamma(1, 100) prior for λ_p and set $\sigma^2_q = 0{\cdot}01$, allowing modest spatial correlation among the p_i's but substantial freedom for the q_i's. This resulted in a posterior median for ρ of 0·28, with 95 per cent Bayesian credible set $[-1{\cdot}15, 1{\cdot}73]$. While this point estimate is still positive (suggesting increased risk of lung cancer associated with higher smoking levels), the very wide credible interval confirms how weakly informative our small sample survey is. By comparison, the model in Section 2 (which ignored the sampling error and possible spatial correlation in smoking proportion) produced a posterior median for ρ of 0·52, with a 95 per cent credible set of $[-1{\cdot}42, 2{\cdot}44]$. The attenuation in the estimated magnitude for ρ (as well as the substantial narrowing of the credible set) supports use of the more sophisticated errors in covariates model. We remark that while for linear models with additive error structure, acknowledging covariate error usually leads not to attenuation but to an *increase* in the magnitude of the covariate effect, no such general rule applies to our non-linear, hierarchical model (see Carroll *et al.*,[27] p. 23).

4. SPATIO-TEMPORAL MODEL WITH ERRORS IN COVARIATES

4.1. Model statement

The study of the trend of risk for a given disease in space and time may provide important clues in exploring underlying causes of the disease and helping to develop environmental health policy. This can be done by constructing a Poisson log-linear spatio-temporal model whose linear predictor contains a variety of *main effect* and *interaction* terms. Waller *et al.*[3] summarize various choices for spatial, temporal, and spatio–temporal interaction effects. In this context, we assume

$$C^{*}_{ijkt} \sim \text{Poisson}(E_{ijkt} \exp(\mu_{ijkt})) \tag{7}$$

where again C^{*}_{ijkt} denotes the observed age-adjusted deaths in county i for sex j, race k, and year t, and E_{ijkt} are the expected death counts. The log relative risk is now modelled as

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_j r_k\xi + p_i\rho + \gamma t + \phi_{it} \tag{8}$$

where γ represents the fixed time effect, and the ϕ_{it} capture the random spatial effects over time, wherein clustering effects are nested within time. That is, writing $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{It})'$, we let $\boldsymbol{\phi}_t \sim \text{CAR}(\lambda_t)$ where $\lambda_t \overset{\text{iid}}{\sim} \text{gamma}(c, d)$. We assume that the socio-demographic covariates (sex and race) do not interact with time or space. We follow (5) and (6) in Section 3 to take sampling error and spatial correlation in the smoking covariate q_i into account, so that again $q_i | p_i \sim \text{N}(p_i, \sigma_q^2)$ and $\mathbf{p} \sim \text{CAR}(\lambda_p)$ where $\lambda_p \sim \text{gamma}(e, f)$. Certainly, other forms of model specification are available and we address several in the next section.

4.2. Results

We once again ran five independent chains using our Gibbs–Metropolis algorithm for 2200 iterations each; plots similar in appearance to Figure 1 suggested discarding the first 200 samples as an adequate burn-in period. Total computation time was about 50 minutes on a Sparc 10 workstation.

We obtained the 95 per cent posterior credible sets $[-1{\cdot}14, -0{\cdot}98]$, $[0{\cdot}07, 0{\cdot}28]$ and $[-0{\cdot}37, -0{\cdot}01]$ for α, β and ξ, respectively. Note that all three fixed effects are significantly different from 0, in contrast to previous results[3] which failed to uncover a main effect for race. The corresponding point estimates are translated into the fitted relative risks for the four socio-demographic subgroups in Table II. Non-white males experience the highest risk, followed by white males, with females of both races having much lower risks.

Regarding the aforementioned anomalous result of Waller *et al.*[3] that non-white females are the healthiest subgroup, consider the model which reverses the gender scores ($s_j = 1$ if male, 0 if female) and reparameterizes the log relative risk (8) as

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_j r_k(\xi - \alpha - \beta) + p_i\rho + \gamma t + \phi_{it}.$$

Under this model, β now unequivocally captures the difference in log relative risk between white and non-white females. Running our MCMC algorithm once again, we obtain point and 95 per

Table II. Fitted relative risks, four socio-demographic subgroups in the Ohio lung cancer data

Demographic subgroup	Contribution to ε_{jk}	Fitted log-relative risk	Fitted relative risk
White males	0	0	1
White females	α	− 1·06	0·35
Non-white males	β	0·18	1·20
Non-white females	$\alpha + \beta + \xi$	− 1·07	0·34

cent interval estimates of $-0{\cdot}01$ and $[-0{\cdot}20, 0{\cdot}18]$ for β; using the same reparameterization under the chosen model (10) in Waller *et al.*,[3] the point and interval estimates instead are $-0{\cdot}20$ and $[-0{\cdot}26, -0{\cdot}15]$. Thus, using this reparameterization we see that age-adjusting has eliminated the statistical significance of the difference between the two female groups.

The posterior median of the smoking effect ρ drops to 0·09 with 95 percent credible set $[-0{\cdot}47, 0{\cdot}66]$. This further attenuation is perhaps not surprising, since our survey data set, small to begin with, is most relevant only for the later years in our study, and cannot reflect the change in smoking patterns since 1968.

Figure 3 shows the fitted age-adjusted lung cancer death rates per 1000 population for one of our demographic subgroups (non-white females), as well as the corresponding variability indexed by interquantile range (IQR), for the years 1968, 1978 and 1988. First, obviously lung cancer death rates are increasing over time, as indicated by the gradual darkening of the maps from 1968 to 1988. Second, their variabilities are also going up somewhat, especially in the last decade. This variability is smallest for high-population counties, such as the one containing the Cleveland metro area (northern border, third from the right). These findings are consistent with those of Waller *et al.*[3]

However, as a result of the common scale in Figure 3, the time effect almost completely obscures the spatial clustering effect. To explore the spatial trend we plot Figure 4, drawing the fitted disease maps for 1968, 1978 and 1988 with three distinct range scales. For 1968, we see a strong spatial pattern of increasing rates as we move from northwest to southeast, perhaps the result of an unmeasured occupational covariate (farming versus mining). Except for persistent low rates in the northwest corner, however, this trend largely disappears over time – perhaps due to increased mixing of the population or improved access to quality health care and health education.

We close this section by returning to the impact of the Fernald facility, and the environmental equity issues it generates. Recall that the plant (and the city of Cincinnati) are located in Hamilton county, in the southwest corner of the state. Figure 5 shows the fitted lung cancer death rates per 1000 population by year for white males (the demographic subgroup most likely to have been working at the facility during the high-risk period) in Hamilton county, as well as Butler, Warren and Clermont counties, which are adjacent to Hamilton. The statewide age-adjusted death rate is also plotted by year for comparison. We observe substantially increased lung cancer death rates in Hamilton county over the whole time period, and rates in all four affected counties that appear to be rising more rapidly than the statewide average. Also, while the rates in the three 'collar counties' are similar to the statewide rate for most of the observation period, there is a dramatic

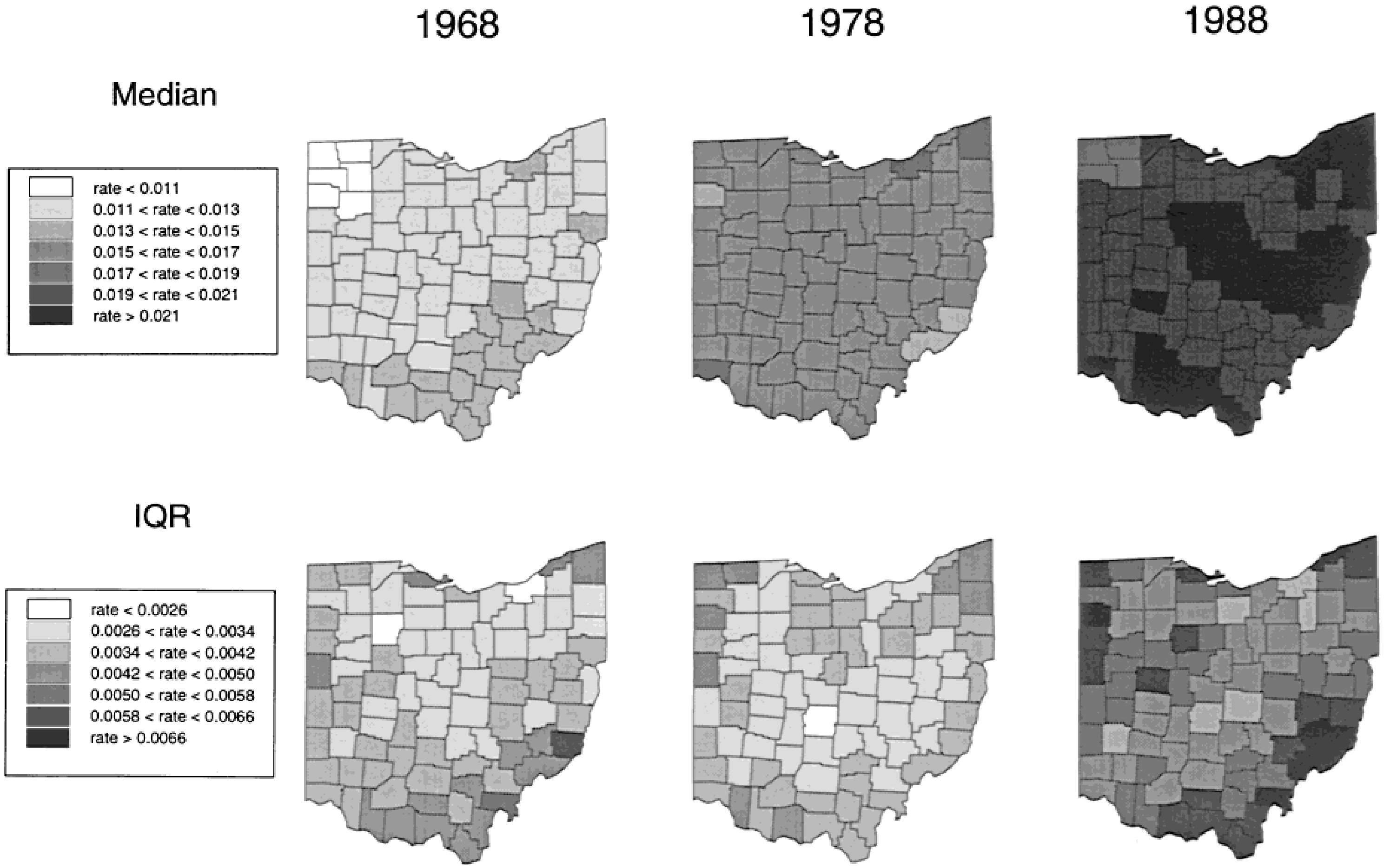


Figure 3. Fitted lung cancer death rates per 1000 population, non-white females

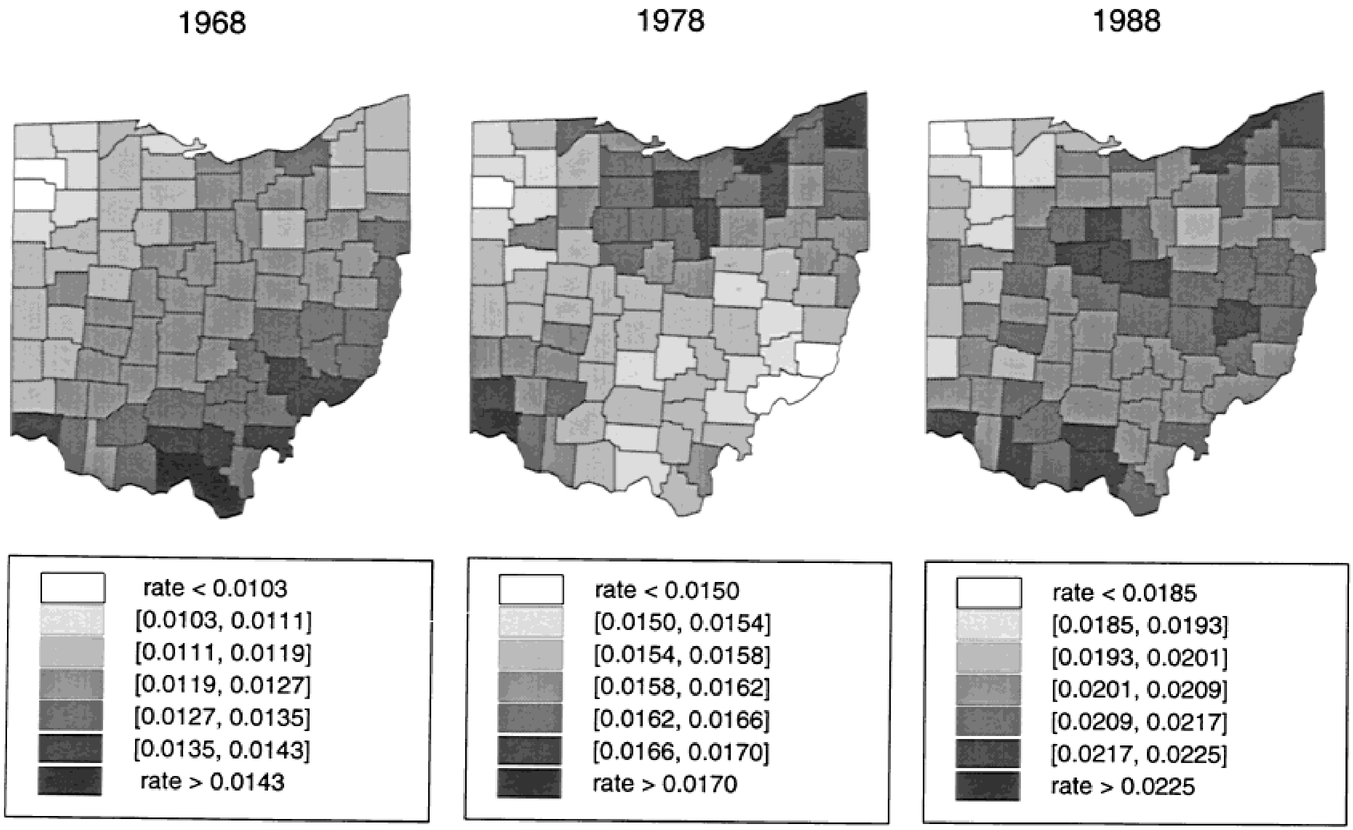


Figure 4. Fitted median lung cancer death rates per 1000 population, non-white females

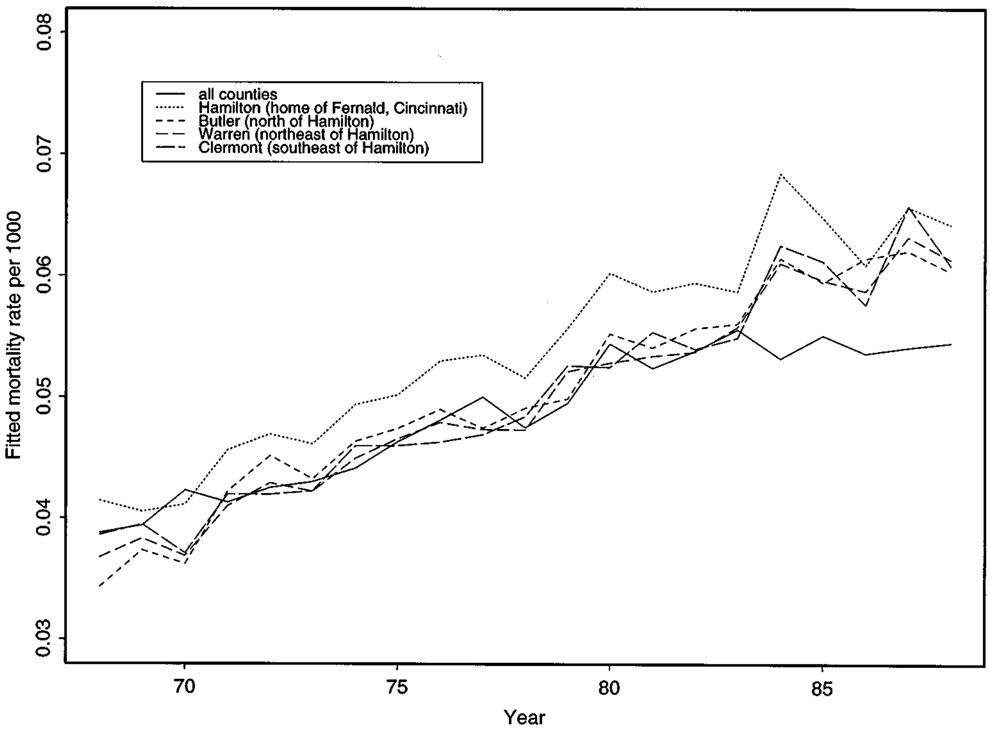


Figure 5. Estimated age-adjusted white male lung cancer death rates per 1000 by year, statewide average and counties near the Fernald Materials Processing Center

departure during the last five years of the study (1984–1988). This departure was not evident in the previous analysis by Waller *et al.*[3] Perhaps this is the result of urban flight to the suburbs or some new environmental factor, but it could also be the effect of plant exposure, since a 20–30 year lag between uranium dust inhalation and death from lung cancer is consistent with known disease aetiology. Our results suggest continued monitoring of cancer rates in the area, to determine whether or not this period of elevated rates is transient.

5. MODEL COMPARISON AND ASSESSMENT

5.1. Model comparison

Caution should be taken in interpreting any kind of model, and naturally our hierarchical spatio-temporal models are no exception. With models as complex as ours, it is particularly important to investigate the sensitivity of any conclusions to changes in model specification. As we have seen, important covariates exert substantial influence on our inferences. On the other hand, although armed with modern computer power and the consequent ability to fit remarkably complex models, we still seek suitably parsimonious choices.

Since our occasional use of improper priors precludes traditional Bayesian model comparison via Bayes factors, we instead follow Gelfand and Ghosh,[28] who propose penalized likelihood

Table III. Model choice statistics for the Ohio lung cancer data, where $\varepsilon_{jk} \equiv s_j\alpha + r_k\beta + s_jr_k\xi$

Model	μ_{ijkt}	EPD	PEN	LRS
1	$\mu + \varepsilon_{jk} + \gamma t + \phi_{it} + p_i\rho$	9988·9	8344·4	1644·5
2	$\mu + \varepsilon_{jk} + \delta_t + \phi_i + p_i\rho$	10383·9	8560·8	1823·2
3	$\mu + \varepsilon_{jk} + \gamma t + \phi_i + p_i\rho$	10353·0	8198·3	2154·7
4	$\mu + \varepsilon_{jk} + \gamma t + p_i\rho$	9782·6	7550·6	2231·9
5	$\mu + \varepsilon_{jk} + \phi_{it} + p_i\rho$	10198·0	8383·9	1814·0
6	$\mu + \varepsilon_{jk} + \phi_i + p_i\rho$	11533·3	8577·8	2955·5
7	$\mu + \varepsilon_{jk} + \gamma t$	9803·9	7533·5	2270·4
8	$\mu + \varepsilon_{jk}$	10522·8	7455·0	3067·7

criteria arising under the posterior predictive distribution. Consider $\mathbf{C}_{\text{new}}$, a posterior replicate of the observed data vector $\mathbf{C}_{\text{obs}} \equiv \{C^*_{ijkt}\}$. Given a posterior sample $\mu^{(g)}_{ijkt}$ from model M_i, we can draw $\mathbf{C}^{(g)}_{\text{new}}$ via (7). This permits evaluation of the expected posterior predictive loss under the given model. For example, under squared error loss, we define the *expected predictive deviance*,

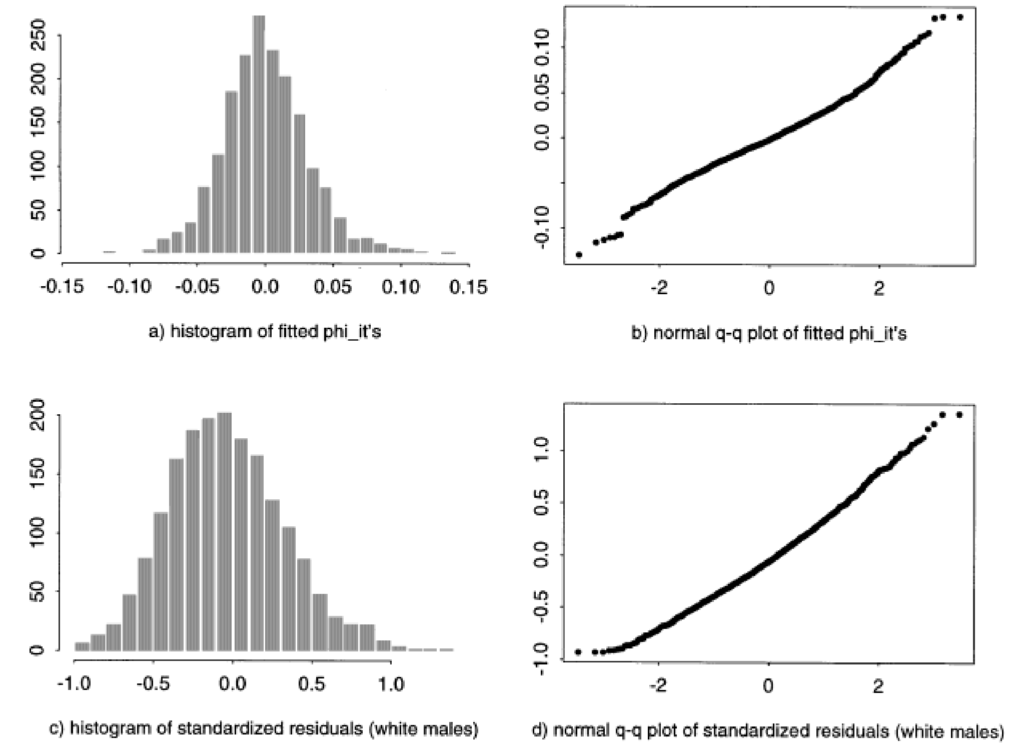
$$\begin{aligned}
\text{EPD} &= E\left[\sum_l (C_{l,\text{new}} - C_{l,\text{obs}})^2 | \mathbf{C}_{\text{obs}}, M_i\right] \\
&= E\left\{\sum_l [C_{l,\text{new}} - E(C_{l,\text{new}}|\mathbf{C}_{\text{obs}}, M_i) + E(C_{l,\text{new}}|\mathbf{C}_{\text{obs}}, M_i) - C_{l,\text{obs}}]^2 \middle| \mathbf{C}_{\text{obs}}, M_i\right\} \\
&= \sum_l E\{[C_{l,\text{new}} - E(C_{l,\text{new}}|\mathbf{C}_{\text{obs}}, M_i)]^2|\mathbf{C}_{\text{obs}}, M_i\} + \sum_l [E(C_{l,\text{new}}|\mathbf{C}_{\text{obs}}\, M_i) - C_{l,\text{obs}}]^2
\end{aligned}$$

where l generically denotes the subscript $ijkt$. The first summation on the right hand side of the above expression is equal to $\text{Var}(\mathbf{C}_{\text{new}}|\mathbf{C}_{\text{obs}}, M_i)$, a measure of the predictive variability in the model, which can be thought of as a penalty (PEN) for model complexity. The second term indicates model goodness-of-fit, and is essentially a likelihood ratio statistic (LRS). Thus we have the decomposition EPD = PEN + LRS, with all three terms easily estimated from the MCMC samples. Smaller EPD values indicate a better model.

Table III shows model choice diagnostics for several possible models. For each model, we ran five parallel, initially overdispersed MCMC chains for 400 post-convergence iterations each. We investigated the MCMC variation of the resulting EPD values, and found a ± 2 standard deviation range of roughly 90 points for each model. Model 1 is the full model proposed in Section 4. Model 2 eliminates the spatial-temporal interaction, but allows a slightly more general temporal main effect (δ_t). Models 3–8 are various simplifications of Model 1. Models 1, 4 and 7 have the smallest EPD estimates. Given the near equality of this diagnostic over these three models, model 1 seems preferable because it provides significantly improved goodness-of-fit, as measured by LRS value.

5.2. Model assessment

Finally, we check our preferred model (Model 1) using two different types of residuals.[17] First, the fitted ϕ_{it} themselves can be viewed as residuals, since they are random effects initially assigned



a) histogram of fitted phi_it's

b) normal q-q plot of fitted phi_it's

c) histogram of standardized residuals (white males)

d) normal q-q plot of standardized residuals (white males)

Figure 6. Model checking, Model 1 (full model), Ohio lung cancer data

a CAR normal prior. Figure 6(*a*) shows a histogram of these fitted effects, while their normal q–q plot is given in Figure 6(*b*). The plots disclose an acceptable degree of normality. Second, we consider the collection of cross-validation Bayesian standardized residuals for the white male subpopulation, compressing the $i11t$ subscript to it for notational convenience. Let

$$r_{it} = \frac{C^*_{it} - E(C^*_{it} | \mathbf{C}_{(it)})}{\sqrt{\{\mathrm{var}(C^*_{it} | \mathbf{C}_{(it)})\}}}$$

where $\mathbf{C}_{(it)}$ denotes the (age-adjusted) data vector with the (it)th point removed. $E(C^*_{it} | \mathbf{C}_{(it)})$ can be approximated by the estimated posterior mean $\frac{1}{G}\sum_{g=1}^{G} E_{it} \exp(\mu_{it}^{(g)})$, where g indexes the post-convergence samples. Denoting the entire parameter collection generically by $\boldsymbol{\theta}$, we also have

$$\begin{aligned}
\mathrm{var}(C^*_{it} | \mathbf{C}_{(it)}) &= E(C^{*2}_{it} | \mathbf{C}_{(it)}) - [E(C^*_{it} | \mathbf{C}_{(it)})]^2 \\
&= E[E(C^{*2}_{it} | \boldsymbol{\theta}, \mathbf{C}_{(it)})] - \{E[E(C^*_{it} | \boldsymbol{\theta}, \mathbf{C}_{(it)})]\}^2 \\
&= E\{\mathrm{var}(C^*_{it} | \boldsymbol{\theta}, \mathbf{C}_{(it)}) + [E(C^*_{it} | \boldsymbol{\theta}, \mathbf{C}_{(it)})]^2\} - \{E[E(C^*_{it} | \boldsymbol{\theta}, \mathbf{C}_{(it)})]\}^2 \\
&\approx E\{\mathrm{var}(C^*_{it} | \boldsymbol{\theta}, \mathbf{C}) + [E(C^*_{it} | \boldsymbol{\theta}, \mathbf{C})]^2\} - \{E[E(C^*_{it} | \boldsymbol{\theta}, \mathbf{C})]\}^2 \\
&\approx \frac{1}{G} \sum_{g=1}^{G} \{E_{it} \exp(\mu_{it}^{(g)}) + [E_{it} \exp(\mu_{it}^{(g)})]^2\} - \left[\frac{1}{G} \sum_{g=1}^{G} E_{it} \exp(\mu_{it}^{(g)})\right]^2 .
\end{aligned}$$

The approximation in the penultimate line should be very good in our case, since $\mathbf{C}$ and $\mathbf{C}_{(it)}$ differ by only one datapoint out of $(88)(21) = 1848$. This in turn enables the final approximation, Monte Carlo estimation of the cross-validatory variance using samples from the full posterior (a great convenience, since it avoids working with 1848 different cross-validatory posterior samples).

Figures 6(*c*) and (*d*) show the histogram and normal q–q plot for these standardized residuals, again using $G = 10{,}000$ post-convergence samples. Because these plots also resemble Gaussian specifications with few outliers, we conclude that our full model fits the data fairly well.

6. DISCUSSION

In this paper we have presented a spatio-temporal modelling framework that allows for (possibly spatially correlated) errors in the observed covariates. We have applied our methods to the smoothing of observed county-level lung cancer mortality rates over a 21-year period in the state of Ohio, accounting for cigarette smoking using an errors in covariates approach. (Two other potentially important covariates, urban living and socio-economic status, could not be included due to collinearity, leading to MCMC convergence failure.) Various posterior predictive model checks and comparisons indicate our final model (which accounts for age, gender, race and smoking, and also includes a simple linear temporal effect) provides adequate fit without sacrificing model parsimony.

Several items suggest themselves for future investigation. For example, a plot of the fitted δ_t posterior means on the same axes as our preferred simple linear trend, $E(\gamma|\mathbf{C}_{\text{obs}})t$, suggests a simple autoregressive time series structure as a possible compromise; Waller *et al.*[29] provide preliminary results in the case of an AR(1) model. Also, the mismatch in type between $C_{l,\text{new}}$ (integer) and $C_{l,\text{obs}}$ (real, due to the age-standardization) in our Section 5 model comparison diagnostics motivates investigation of an expanded model of the age bracket-specific death counts C_{ijklt}, where l indexes the age bracket. While this would mean a data set of $(88)(2)(2)(21)(11) = 81{,}312$ observations, a suitably tuned MCMC algorithm might still be able to produce sufficiently accurate posterior estimates in acceptably short runtimes.

ACKNOWLEDGEMENTS

The research of both authors was supported in part by National Institute of Allergy and Infectious Diseases (NIAID) FIRST Award 1-R29-AI33466. In addition, the research of the second author was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07750. The authors are grateful to Dr. Owen Devine of the Centers for Disease Control and Prevention for furnishing the Ohio lung cancer data set, Mary Plummer, M.S., of the Ohio Department of Health for furnishing the Ohio cigarette smoking survey data, Professor Tom Louis for helpful comments, and Professor Lance Waller for longtime assistance with spatio-temporal modelling.

REFERENCES

1. Perlin, S. A., Setzer, R. W., Creason, J. and Sexton, K. 'Distribution of industrial air emissions by income and race in the United States: an approach using the toxic release inventory', *Environmental Science and Technology*, **29**, 69–80 (1995).
2. Clayton, D. G. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, **43**, 671–681 (1987).

3. Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association*, **92**, 607–617 (1997).
4. Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pellom, A. C. 'Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates', *Journal of the American Statistical Association*, **84**, 637–650 (1989).
5. Ghosh, M. 'Constrained Bayes estimation with applications', *Journal of American Statistical Association*, **87**, 533–540 (1992).
6. Devine, O. J., Halloran, M. E. and Louis, T. A. 'Empirical Bayes methods for stabilizing incidence rates prior to mapping', *Epidemiology*, **5**, 622–630 (1994).
7. Devine, O. J. and Louis, T. A. 'A constrained empirical Bayes estimator for incidence rates in areas with small populations', *Statistics in Medicine*, **13**, 1119–1133 (1994).
8. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Empirical Bayes estimators for spatially correlated incidence rates', *Environmetrics*, **5**, 381–398 (1994).
9. Cressie, N. and Chan, N. N. 'Spatial modeling of regional variables', *Journal of the American Statistical Association*, **84**, 393–401 (1989).
10. Clayton, D. G. and Bernardinelli, L. 'Bayesian methods for mapping disease risk', *in* Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology*: *Methods for Small-Area Studies*, Oxford University Press, Oxford, 1992.
11. Besag, J., York, J. C. and Mollié, A. 'Bayesian image restoration, with two applications in spatial statistics' (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1–59 (1991).
12. Bernardinelli, L. and Montomoli, C. 'Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk', *Statistics in Medicine*, **11**, 983–1007 (1992).
13. Breslow, N. E. and Clayton, D. G. 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25 (1993).
14. Bernardinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. 'Disease mapping with errors in covariates', *Statistics in Medicine*, **16**, 741–752 (1997).
15. Bernardinelli, L., Clayton, D. and Montomoli, C. 'Bayesian estimates of disease maps: how important are priors?' *Statistics in Medicine*, **14**, 2411–2431 (1995).
16. Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M. 'Bayesian analysis of space-time variation in disease risk', *Statistics in Medicine*, **14**, 2433–2443 (1995).
17. Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. 'Generalized linear models for small area estimation', *Journal of the American Statistical Association*, (1998), in press.
18. Xia, H., Carlin, B. P. and Waller, L. A. 'Hierarchical models for mapping Ohio lung cancer rates', *Environmetrics*, **8**, 107–120 (1997).
19. Devine, O. J. *Empirical Bayes and Constrained Empirical Bayes Methods for Estimating Incidence Rates in Spatially Aligned Areas*, unpublished Ph.D. dissertation, Division of Biostatistics, Emory University, 1992.
20. Centers for Disease Control and Prevention, National Center for Health Statistics. *Public Use Data Tape Documentation Compressed Mortality File*, *1968–1985*, U.S. Department of Health and Human Services, Hyattsvile, Maryland, 1988.
21. Gelfand, A. E. and Smith, A. F. M. 'Sampling based approaches to calculating marginal densities', *Journal of the American Statistical Association*, **85**, 398–409 (1990).
22. Carlin, B. P. and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, London, 1996.
23. Besag, J., Green, P., Higdon, D. and Mengersen, K. 'Bayesian computation and stochastic systems' (with discussion), *Statistical Science*, **10**, 3–66 (1995).
24. Chib, S. and Greenberg, E. 'Understanding the Metropolis–Hastings algorithm', *The American Statistician*, **49**, 327–335 (1995).
25. Gelman, A. and Rubin, D. B. 'Inference from iterative simulation using multiple sequences' (with discussion), *Statistical Science*, **7**, 457–511 (1992).
26. Richardson, S. and Gilks, W. R. 'Conditional independence models for epidemiological studies with covariate measurement error', *Statistics in Medicine*, **12**, 1703–1722 (1993).
27. Carroll, R. J., Ruppert, D. and Stefanski, L. A. *Measurement Error in Nonlinear Models*, Chapman and Hall, London, 1995.

28. Gelfand, A. E. and Ghosh, S. K. 'Model choice: a minimum posterior predictive loss approach', *Biometrika*, (1998), in press.
29. Waller, L. A., Carlin, B. P. and Xia, H. 'Structuring correlation within hierarchical spatio-temporal models for disease rates', *in* Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G. and Wolfinger, R. D. (eds), *Modelling Longitudinal and Spatially Correlated Data*, Springer-Verlag, New York, 1997, pp. 308–319.