

# Chapter 11

## Multilevel and hierarchical models for disease mapping<sup>1</sup>

疾病地图的多层次和按等级划分层次模型

Jarvis T. Chen

### 11.1 Introduction

Georeferenced health data makes it possible to locate cases in space and to analyze spatially varying patterns of health, disease, and well-being. By linkage to geocoded data on exposures and covariates, it is also possible to enrich health data and the individual level variables contained therein with a variety of information on environmental exposures, socioeconomic context, and other variables. As these rich and diverse sources of georeferenced data have become more available in recent decades, increasingly sophisticated statistical methods have been developed for modelling these data. With these methods at their disposal, epidemiologists, sociologists, and environmental scientists can explore such questions as: (1) “How do population rates of health, disease, and wellbeing vary over space and across different levels of geographic aggregation?” (2) To what extent are observed spatial variations in health driven by geographically varying exposures? (3) To what extent are the observed spatial variations driven by the different characteristics of populations living in those areas? (4) Which geographic areas exhibit unusually high or unusually low rates of disease, taking into account known exposures?

In this chapter, I discuss modelling frameworks that enable one to explore the

---

<sup>1</sup>To appear in Boscoe, F. (ed) *Geographic Health Data: Fundamental Techniques for Analysis*, Wallingford, UK: CABI. Please do not copy or circulate without permission.

与`observed`协变量相关的`health outcome`的空间变异

spatial variation in health outcomes in relation to observed covariates, and to derive estimates of meaningful quantities for data presentation. The modelling approaches that are useful for spatial data are part of a broad family of models known as **hierarchical or multilevel models**, which have arisen from parallel lines of development in the biostatistical literature on disease mapping, analysis of longitudinal data, and methods for handling clustered data (Gelman et al., 2003; Banerjee et al., 2004; Gelman and Hill, 2007). The broad class of models includes so-called mixed effects, random effects, variance components, and generalized linear mixed models (GLMM). Latent variable models can also be viewed as hierarchical models since random effects can also have a latent variable interpretation.

What these approaches have in common is that they are useful in complex datasets where information is available on several different levels of observational units, the distributions of which are controlled by multiple parameters that can be regarded as connected by the structure of the problem. Because they are “related,” the joint probability model for these parameters needs to reflect the dependence among them. An appealing `approach` is to model such data hierarchically, with observable outcomes modelled conditional on certain parameters, which themselves are given a probabilistic specification in terms of other parameters (known as hyperparameters). This kind of approach helps us to model complex multiparameter problems efficiently, and can also facilitate computation for model fitting.

To give a concrete example, suppose that a researcher is interested in studying the relationship between health, as measured by a continuous variable  $y$ , and a measure of individual socioeconomic deprivation,  $x$ . Let us assume that the researcher has collected data on  $y$  and  $x$  in a 社会经济贫困 simple random sample of  $n$  individuals, indexed by  $i$ . A `classical linear regression` would model:

$$y_i = \alpha + \beta x_i + \epsilon_i, \text{ for individuals } i = 1, \dots, n$$

where  $\epsilon$  is a normally distributed random error term with mean 0 and variance  $\sigma_\epsilon^2$ . Now assume that the researcher has collected data on a sample of 50 neighbourhoods (indexed by  $j$ ), and within each neighbourhood, a random sample of individuals (indexed by  $i$ ). It would be reasonable to assume that the health scores of individuals from the same neighbourhood might be more similar than individuals from different neighbourhoods, perhaps due to unmeasured environmental exposures that vary from neighbourhood to neighbourhood. If so, the outcomes observed in individuals from the same neighbourhood will be correlated, which violates a key assumption of the classical linear regression model. A simple multilevel model for these data might be to assume instead that, at the level of the individual, the observed  $y$  for individual  $i$  in neighbourhood  $j$  depends on the observed value of  $x_{ij}$ , a `neighbourhood-specific`

某些参数对可观察结果进行建模，这些参数本身根据其他参数（称为超参数）给出了一个概率规范。这种方法有助于我们对复杂的多参数问题进行有效的建模，也便于模型拟合的计算。

intercept,  $\alpha_j$ , and an individual-level random error term  $\epsilon_{ij}$ ,

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

where  $\epsilon_{ij} \sim \text{Normal}(0, \sigma_\epsilon^2)$ . At the second level, a neighbourhood level model is specified for  $\alpha_j$ , for example

$$\alpha_j = \mu_\alpha + v_j$$

with  $v_j \sim \text{Normal}(0, \sigma_v^2)$ . In this simple example, one assumes that  $\alpha_j$  comes from a distribution of neighbourhood intercepts that is normally distributed with mean  $\mu_\alpha$  and variance  $\sigma_v^2$ . Whereas in the classical linear regression model, the intercept ( $\alpha$ ) was treated as invariant over neighbourhoods, in the multilevel model, the intercept parameter is allowed to vary by neighbourhood and is given a probabilistic model. This relatively simple 2-level model is often referred to as a **random intercept** model. The second stage model could be further elaborated to incorporate additional neighbourhood level predictors.

This 2-level multilevel model could also be written in a single line as

$$y_{ij} = \mu_\alpha + \beta x_{ij} + v_j + \epsilon_{ij},$$

which highlights another way to think about this model: one can also conceptualize the residual variation in  $y$ , conditional on the observed values of  $x$ , as having been partitioned into a between-neighbourhood component ( $v_j$ ) and a within-neighbourhood component ( $\epsilon_{ij}$ ).

A key concept in hierarchical models is that of **exchangeability** (Gelman et al. 2003). In a typical statistical analysis, the (often tacit) assumption is made that the outcomes  $y_i$  observed for  $n$  units of observation are exchangeable, meaning that their joint probability is invariant to permutation of the indexes. Thus, one usually models outcomes from an exchangeable distribution as independently, identically distributed (i.i.d.) conditional on some unknown parameter vector  $\boldsymbol{\theta}$  (that may include the effects of known and observed explanatory variables). In a hierarchical model, one can speak of exchangeability at each level of units. So, in the simple example above, one treats individuals and their outcomes within each neighbourhood as exchangeable and also treats the neighbourhoods as exchangeable (as reflected in the second stage model for  $\alpha_j$ ). In most analytic applications, one will want to make these assumptions of exchangeability at each level conditional on explanatory variables that are included as fixed effects in the model (e.g. in the above example, the assumption of exchangeability is conditional on the socioeconomic deprivation variable  $x_{ij}$ ).

An important consequence of exchangeability is that each parameter borrows strength from the other parameters at its level of the hierarchy, and as a result, estimates are “shrunk” toward the population mean. This behavior can be beneficial, especially when small numbers of individuals are observed in some of the units in the hierarchy. In these cases, the reduction in uncertainty for that unit can be quite large, since information from other groups or units with smaller variability is incorporated in the posterior estimates.

The goal in this chapter is to provide an introduction to the hierarchical modelling of spatial variation in health outcomes. In many health research applications, the georeferenced data of interest take the form of area-level aggregated counts, which are matched to population counts that are treated as known. Thus, I focus the discussion on multilevel Poisson regression models. An important consideration is specifying a reasonable model for the second stage parameters that appropriately captures the spatial variation in risk. I motivate the presentation by considering the setting in which one is interested in mapping smoothed estimates of disease rates and/or risks in order to reduce the statistical instability of small-area estimates. I then consider two of the most popular approaches for accommodating local clustering of area-level spatial variation in a hierarchical modelling framework: (a) a multilevel (nested) hierarchy of normally distributed random effects (Goldstein, 2011), and (b) the popular intrinsic Gaussian conditionally autoregressive (CAR) approach by which local clustering is induced by a set of conditional autoregressions (Besag et al., 1991).

The literature on hierarchical modelling approaches to spatially referenced health data is quite large, and excellent reviews and textbooks already exist that present the technical aspects in greater detail than can be accomplished here (Marshall, 1991; Clayton and Bernardinelli, 1995; Lawson et al., 1999; Bithell 2000; Wakefield et al. 2000; Richardson et al. 2003; Banerjee et al., 2004). Similarly, there is an extensive literature on multilevel models applied to problems in the social sciences and epidemiology (see, for example, Goldstein, 2011; Snijders and Bosker, 2011; Subramanian et al., 2003; Gelman and Hill, 2007). A particularly useful textbook that addresses the underlying theory and practical issues of implementation from both the disease mapping and multilevel perspectives is Lawson et al. (2003).

Here, it is useful to note a potential confusion in the nomenclature used to refer to these methods. Multilevel modelling grew out of work in the field of education in the 1980’s (Aitkin et al., 1981; Aitkin and Longford, 1986) where the units of analysis (e.g. students nested within classrooms within schools) could be organized into a nested hierarchy of levels. It makes sense that similar approaches could be applied to nested spatial data organized by geographical units (e.g. census tracts nested

within counties or states) and even temporally organized data (repeated observations “nested” within individuals in the longitudinal setting). In this way, the term “multilevel” has become synonymous with hierarchically nested data structures (Blakely and Subramanian, 2006). However, while nested data clearly lend themselves to hierarchical modelling approaches, more complex, non-nested data structures do as well. The observational units may no longer be neatly nested, but the models are still hierarchical in that the parameters that control the distribution of outcomes at the first level of the model may themselves be given a second-stage model specified in terms of hyperparameters. In the multilevel modelling literature, these may be referred to as cross-classified structures. For clarity, I will use the word “multilevel” to refer to models for nested data structures and use the word “hierarchical” to refer to the broader class of modelling approaches.

Throughout, I make use of an illustrative dataset of lung cancer deaths among the white non-Hispanic population observed in the 156 census tracts of city of Boston (Massachusetts, United States of America (US)). The 156 census tracts are nested in 16 neighbourhoods, which potentially permits a three-level nested hierarchy of individuals nested in census tracts, nested in neighbourhoods. To obtain the dataset, the 910 lung cancer deaths that occurred between 2000 and 2005 were geocoded to the census tract level and matched to age, race, and gender-specific census tract population estimates from the US Census 2000. These population estimates are multiplied by 6 (the number of years of data) to obtain person-time denominators.

### 11.1.1 Model implementation

Hierarchical model formulations have a particular affinity with the Bayesian inferential framework, as random effects naturally have a prior distribution (Gelman et al., 2003). Thus, fully Bayesian and empirical Bayes (posterior approximation) methods are commonly used for fitting disease mapping models (Wakefield et al., 2000; Best et al., 2005). Suppose  $\theta = (\theta_1, \dots, \theta_p)$  is the vector of model parameters to be estimated and  $\mathbf{y} = (y_1, \dots, y_n)$  is the observed data. In contrast to the usual frequentist approach which treats  $\theta$  as a fixed parameter to be estimated, the Bayesian approach models the observed data and all unknown parameters as random. In addition to specifying the likelihood function, denoted by  $f(\mathbf{y}|\theta)$ , a Bayesian model assumes that the vector  $\theta$  is a random variable sampled from a prior distribution,  $\pi(\theta|\lambda)$ , where  $\lambda$  is a vector of hyperparameters that characterize this distribution. Sometimes, values of  $\lambda$  can be picked to reflect a strong prior belief about the values of  $\theta$ . More often, however, one picks values so that  $\theta$  can be free to vary across a very wide range — this is known as placing an “uninformative prior” distribution on  $\theta$ .

In the fully Bayesian approach, inference is based on the posterior distribution  $p(\boldsymbol{\theta}|\mathbf{y})$ , obtained using Markov Chain Monte Carlo (MCMC) methods. From the Bayesian viewpoint, all current knowledge about  $\boldsymbol{\theta}$  is summarized by the posterior distribution. Although the computation of the posterior distribution can often be difficult, subsequent Bayesian inferences are relatively straightforward once the posterior distribution has been obtained. For example, let us assume, for simplicity, that the parameter  $\theta$  is univariate. Possible choices for a point estimate of  $\theta$  are the mean, median or mode of the posterior distribution. The 95% credible set or Bayesian confidence interval for  $\theta$ ,  $(q_L, q_U)$ , can be determined empirically from the posterior distribution by choosing  $q_L$  and  $q_U$  as the 2.5th and 97.5th quantiles of the posterior distribution of  $\theta$ . In addition, other useful summary quantities, perhaps involving combinations or manipulations of elements of  $\theta$  can be monitored at each iteration of the MCMC algorithm, and 95% credible intervals can be easily calculated. I give some examples of these in this chapter. Throughout, I take a Bayesian approach to model fitting using the popular WinBUGS (**B**ayesian **I**nference **U**sing **G**ibbs **S**ampling) software package (Lunn et al., 2000). For each model, I ran two independent chains starting from different initial values. The first 10,000 iterations were discarded as a burn-in, and each chain was run for a further 40,000 iterations and thinned by 8 (i.e. every 8th iteration was retained). Model convergence was evaluated by visual inspection of the time series plots of samples for each chain and by computing the Gelman-Rubin diagnostic (Gelman and Rubin, 1992). I report estimates and 95% credible intervals based on the mean and quantiles of the posterior samples.

To compare models, one can make use of the deviance information criterion (DIC), defined as the sum of the posterior mean of the deviance,  $\bar{D}$  (-2 times the log likelihood) and an estimate of the ‘effective’ number of parameters,  $p_D$  (Spiegelhalter et al. 2002, Best et al. 2005). The latter can be thought of as a penalty term reflecting the model complexity or degrees of freedom, so that, broadly speaking, the DIC has a similar interpretation to Akaike’s Information Criterion (AIC) in classical generalized linear models. **Smaller values of the DIC reflect better model fit**, and Spiegelhalter et al. (2005) suggests that models with DIC values more than 7 higher than the ‘best’ model are substantially inferior.

## 11.2 Disease mapping

As a first step in addressing the question of “How does the risk of this particular disease vary geographically?” consider the situation where one has geocoded records of deaths due to lung cancer, and wants to map these data to obtain a visual rep-



Figure 11.1: A map of (hypothetical) residential addresses of lung cancer deaths among white non-Hispanics in Boston

resentation of where death due to lung cancer is more common and where it is less common. The map may be intended for simple descriptive purposes, to provide information on the health needs of the population, or to generate hypotheses about disease aetiology.<sup>病因学</sup> As a first approach, one might be tempted to begin by mapping the locations of each case's geocoded address of residence, as I have done in Figure ?? for (hypothetical) lung cancer deaths among white non-Hispanics in Boston in 2000-2005. While this gives a visual representation of where cases were recorded as living when they died of lung cancer, this does not provide a very satisfying summary of lung cancer mortality. Firstly, there are likely to be confidentiality issues with plotting exact locations of cancer deaths (which is why I have only plotted hypothetical locations in Figure ??). Secondly, and more importantly, the absolute number of deaths observed is sensitive to the population at risk, which varies across space. Thus, a more epidemiologically meaningful quantity to map is the **mortality rate**, defined as the number of deaths divided by the population (or person-time) at risk.

Even though the dataset includes exact locations of deaths, population counts are only available aggregated over small areas (e.g. census block groups, census tracts, counties). Consequently, one has to aggregate the death counts to match the

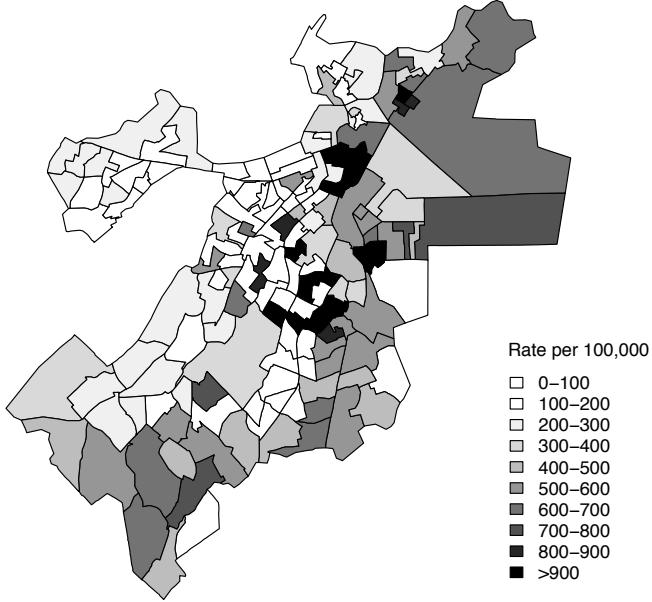


Figure 11.2: Crude census tract lung cancer mortality rates for white non-Hispanics in Boston, 2000-2005

available units in the population data. This results in aggregated *count* data (also known as *area* or *lattice* data, as opposed to *point* data). That is, one observes  $y_i$  (the number of deaths in area  $i$ ) and  $N_i$  (the population at risk in area  $i$ ), from which one can estimate the crude mortality rate in area  $i$  as  $\hat{r}_i = y_i/N_i$ . I have mapped these using (real) data on lung cancer deaths among white non-Hispanics in Boston for 2000-2005 in Figure ???. The resulting map allows one to visualize areas with low and high lung cancer mortality rates. However, it is still of limited use because it is well known that mortality rates are strongly related to age, and the age distribution may differ across areas. A well known way of dealing with this is to standardize the mortality rates by age using indirect standardization. If  $y_{ij}$  is the count of deaths in area  $i$  in age stratum  $j$ , define  $O_i$  as the observed number of deaths in area  $i$ , summed up over the age strata:  $O_i = \sum_j y_{ij}$ . The expected number of deaths,  $E_i$ , is computed by applying age-specific mortality rates from a reference population to

the age-specific population counts in each area:

$$E_i = \sum_j N_{ij} \times R_j$$

where  $N_{ij}$  is the person-time at risk in age group  $j$  in area  $i$  (computed by taking the population in age-group  $j$  in area  $i$   $\times$  number of years) and  $R_j$  is the mortality rate age group  $j$  of a suitable reference population. The ratio  $O_i/E_i$  is known as the Standardized Incidence Ratio (SIR), or, in the case of mortality outcomes, the Standardized Mortality Ratio (SMR). This ratio is an estimate of the *relative risk* within each area,  $\theta_i$ . For rare, non-communicable diseases, the standard statistical model for  $O_i$  is that of the Poisson distribution:

$$O_i \sim \text{Poisson}(\theta_i E_i).$$

The maximum likelihood estimator of  $\theta_i$  is

$$\hat{\theta}_i = \text{SMR}_i = \frac{O_i}{E_i}$$

with  $\text{Var}(\text{SMR}_i) = \theta_i/E_i$ , estimated by  $O_i/E_i^2$ . Note that  $\text{Var}(\text{SMR}_i)$  is inversely proportional to  $E_i$ . To translate the relative risk into rates, one simply multiplies the SMR by the overall mortality rate from the reference population.

SMRs have been commonly used in disease maps, and are particularly useful when the age-specific counts of deaths are not available for each area. The method also produces rate estimates with smaller asymptotic variance than the corresponding direct standardization method (Pickle and White, 1995). Because SMRs estimate the relative risk relative to the average risk over the entire study area, they are perhaps best visualized in colour with a divergent (or “double-ended”) colour map, which permits identification of unusually high or unusually low SMRs, SMRs close to the null value of 1.0 visualized with an unsaturated colour (Brewer, 2005). When maps are rendered in black and white, representation of the divergent scale is more difficult. In the print version of this chapter, maps of SMRs are presented using a sequential greyscale gradient; to see colour maps with a divergent colour map (with  $\text{SMR} < 1$  represented in blues and  $\text{SMR} > 1$  represented in reds) (see <http://www.albany.edu/~fboscoe/gisbook>).

As summary measures, SMRs also have certain drawbacks. They are based on ratio estimators, and are thus sensitive to small changes in  $E_i$ . In particular, when  $E_i$  is close to zero, the SMR will be very large for any positive count. As the estimate of  $\text{Var}(\text{SMR}_i)$  is proportional to  $1/E_i$ , SMRs of zero do not distinguish variation

in expected counts. Most importantly, interpretability and comparability of SMRs based on indirectly age standardized data across areas depends on the assumption of independent area and age effects with respect to the standard population. This is known as the proportionality assumption, and assumes that  $r_{ij} = \theta_i \times \alpha_j$  where  $\alpha_j$  is an age effect that does not vary over areas.

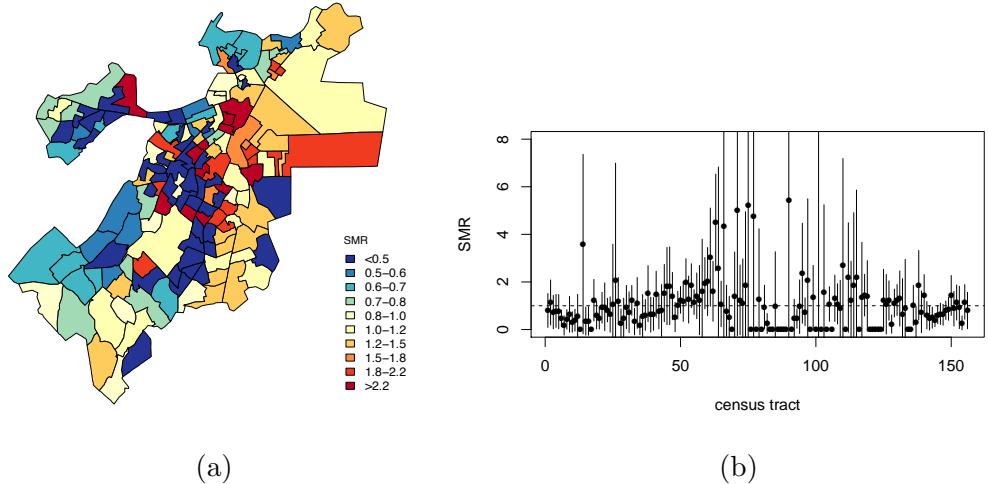


Figure 11.3: Observed SMRs for lung cancer mortality among white non-Hispanics in Boston, 2000-2005. (a) map and (b) estimates and 95% confidence intervals

Applying indirect age standardization to lung cancer mortality rates for white non-Hispanics in Boston (using age-specific lung cancer mortality reference rates from Massachusetts for the same period) yields the map in Figure 11.1a. Here, the quantity mapped for each area is the ratio of the mortality rate in that particular area compared to the overall mean rate in the reference population. Since  $\text{Var}(\text{SMR}_i) \approx O_i/E_i^2$ , one can use this to calculate 95% confidence limits for the SMRs in the Boston example (Figure 11.1b). Most of the confidence limits are wide and include the null value of 1. Thus, the resulting estimates are dominated by sampling variability, making comparison of mortality risks between areas difficult.

To overcome this variability, hierarchical models can be used to “smooth” the raw rates. When faced with the problem of making inferences on many parameters  $\{\theta_i\} = \theta_1, \dots, \theta_n$ , measured on  $n$  areas, one can imagine two possible extreme assumptions. At one extreme, one could assume that all of the  $\{\theta_i\}$  are identical, in which case all the data can be pooled, and the individual units ignored. This is what one typically does when presenting summary rates and rate ratios over the whole area. At the other

extreme, one could assume that all the  $\{\theta_i\}$  are independent and entirely unrelated, in which case the results from each area would need to be analyzed independently (no pooling). As we have seen in the SMR example above, this leads to statistical instability if the numbers are small.

A third possible assumption lies somewhere between these two extremes. One could assume that the  $\{\theta_i\}$  are ‘similar’ in the sense that the area labels convey no additional information. This is known as **exchangeability**, and is equivalent to assuming that  $\{\theta_i\}$  are drawn from a *common prior distribution with unknown parameters*.

### 11.3 Poisson gamma model

A classic example of this approach is presented by Clayton and Kaldor (1987), who developed a Bayesian analysis of a Poisson likelihood model. Their model is a useful introduction to the idea of hierarchical modelling of disease rates, as the second stage distribution for the area variability is analytically tractable. In the first stage of the hierarchy, one assumes that the observed death counts for each area are Poisson distributed:

$$O_i \sim \text{Poisson}(\theta_i E_i).$$

In the second stage, a hierarchical prior is placed on  $\theta_i$ :

$$\theta_i \sim \text{Gamma}(a, b).$$

Recalling that the gamma distribution with parameters  $a$  and  $b$  has mean  $a/b$ , this simply states that we expect the distribution of  $\{\theta_i\}$  to follow a gamma distribution with mean  $a/b$  and variance  $a/b^2$ . Since the gamma distribution is the conjugate prior of the Poisson, the posterior distribution of  $p(\theta_i|O_i, E_i)$  also follows a gamma distribution:

$$\text{Gamma}(a + O_i, b + E_i)$$

with mean given by

$$E(\theta_i|O_i, a, b) = \frac{a + O_i}{b + E_i} = w_i \text{SMR}_i + (1 - w_i) \frac{a}{b}$$

where

$$w_i = \frac{E_i}{b + E_i}.$$

Thus, the posterior mean of the relative risk for the  $i$ th area is a weighted average of the SMR for the  $i$ th area and the average relative risk ( $a/b$ ) in the overall map. The



weight is inversely proportional to the variance of the SMR. Accordingly, when  $E_i$  is small (for rare diseases or small population counts), the variance is large, so the weight  $w_i$  is small. In this situation, the posterior mean is dominated by the prior mean,  $a/b$ . In areas with abundant data, the posterior mean is close to  $O_i/E_i$  i.e. the observed SMR. This feature, whereby the amount of smoothing is proportional to the amount of information available for a particular area, is known as *precision weighting*. It has an intuitive appeal in that, when one does not observe a lot of information about an area (because the sample size is small and the risk estimate is unstable), one's "best guess" concerning that area's mortality risk should be weighted towards what little is known from prior knowledge, i.e. that the risk is, on average,  $a/b$ . In contrast, if one observes a lot of information for an area (e.g. because the sample size is large), one is more likely to believe what the data say about mortality risk in that particular area, and thus the "best guess" would reasonably be weighted towards the observed SMR for that specific area.

In the fully Bayesian framework, one treats the parameters of the gamma prior as unknown parameters to be estimated as well, and assign prior distributions to  $a$  and  $b$ . This allows the data to inform the prior for the gamma distribution. If one selects relatively uninformative priors, such as

$$\begin{aligned} a &\sim \text{Exponential}(0.01) \\ b &\sim \text{Exponential}(0.01) \end{aligned}$$

then the choice of hyperparameters for these distributions does not heavily influence the lower level variation. In general, the posterior distribution of these parameters is not of closed form, but MCMC methods can be used to obtain posterior samples.

The resulting estimate of each  $\theta_i$  "borrows" strength from the likelihood contributions for *all* of the areas, via their joint influence on the estimate of the unknown prior parameters  $a$  and  $b$ . This leads to *global smoothing* of the relative risks: as seen in Figure 11.2, the posterior estimates of the relative risks have been "shrunk" towards the null value of 1. One obtains posterior distributions for each  $\theta_i$ , and 95% credible intervals can be computed from quantiles of these distributions. They show that the precision of the relative risk estimates is also improved by borrowing strength across areas. These estimates also reflect the uncertainty about the true values of  $a$  and  $b$ .

A comparison of the raw SMR map to the smoothed map from the Poisson-gamma model also shows the substantial smoothing that has occurred. Most of the areas that appeared to have markedly higher or lower risk of lung cancer mortality in the raw SMR map now appear to have fairly average risks, and only a few areas stand out as unusual in the second map.

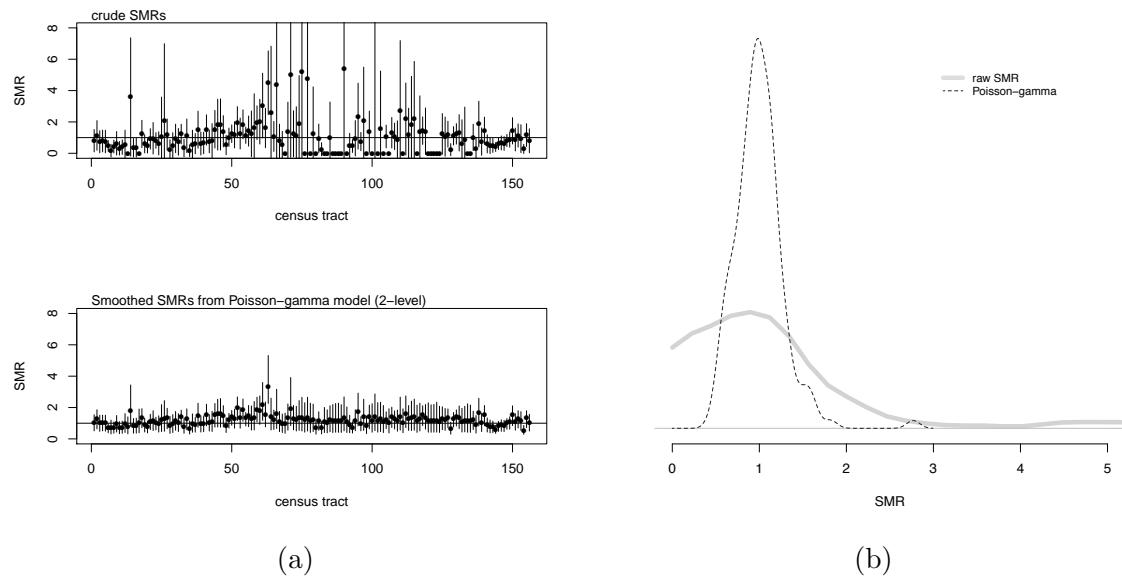


Figure 11.4: Comparison of raw SMRs and smoothed SMRs from the Poisson-gamma model. (a) Raw SMRs and 95% confidence limits; smoothed SMRs and 95% credible intervals. (b) Distributions of raw SMRs and smoothed SMR estimates

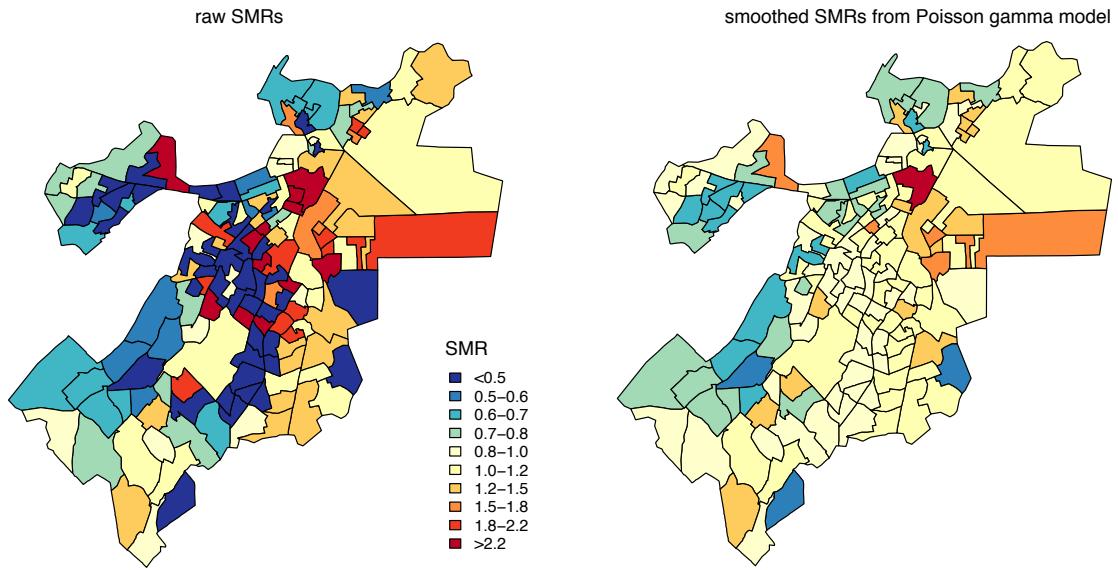


Figure 11.5: Map of the smoothed relative risk estimates for lung cancer mortality from the Poisson-gamma model compared to the raw SMRs, for white non-Hispanics in Boston, 2000-2005

## 11.4 Poisson log normal model

A gamma prior for  $\{\theta_i\}$  is mathematically convenient, but may be restrictive. If the goal in hierarchical modelling is to be able to treat the units at each stage of the model as exchangeable, one will likely want to adjust for observed covariates of disease risk. However, this is difficult in the Poisson-gamma model. A normal prior for  $\log(\theta_i)$  yields a more flexible model (Wakefield et al., 2000; Lawson et al., 2003):

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \alpha + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

where  $\alpha$  is an intercept term representing the overall log relative risk of disease in the whole study region compared to the reference rate and  $v_i$  is the residual log relative risk in area  $i$  compared with the average over the study region. Here, the  $\{v_i\}$  are assumed to arise from a normal distribution with mean 0 and variance  $\sigma_v^2$ . This model is recognizable as a generalized linear mixed model, where “mixed” refers to the fact that the model contains both “random effects” (the  $\{v_i\}$ ), as well as accommodating “fixed” covariate effects, e.g.

$$\log(\theta_i) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p + v_i.$$

In this case the  $\{v_i\}$  are interpretable as *residual* area-specific effects conditional on the fixed covariates. Similarly, the  $\beta$ s are interpretable as covariate effects *conditional* on the area random effects.

It should be noted here that, as pointed out by Wolpert and Ickstadt (1998), the Poisson log normal model does not aggregate consistently. That is, if one specifies a log normal distribution for each of the relative risks and then combines two areas and specifies a log normal distribution for the relative risk of the combined area, then these distributions are inconsistent (because the sum of log normal distributions is not log normal). This can be understood as a form of ecologic bias whereby risk relationships do not remain constant across levels of aggregation (Wakefield et al., 2000). Nevertheless, a normal second-stage distribution has been observed empirically to provide a good model for log relative risks over a range of aggregations, and does present advantages with respect to model flexibility and ease of computation.

To complete the Bayesian specification, one needs to place hyperpriors on  $\alpha$  and  $\sigma_v^2$ . For this example, I have specified a diffuse normal prior for  $\alpha \sim \text{Normal}(0, 2000)$ . Note that in WinBUGS, the normal distribution is specified in terms of the mean and the precision  $\tau^2 = 1/\sigma^2$ . Thus, the diffuse normal prior specified by `dnorm(0,`

0.0005) corresponds to a normal distribution with mean zero and variance of 2000. It is convenient in WinBUGS to specify a gamma prior for the precision of  $\sigma_v^2$ , e.g.  $1/\sigma_v^2 \sim \text{Gamma}(0.5, 0.0005)$  (Kelsall and Wakefield, 1999; however, see Gelman, 2006, for a discussion about selecting appropriate prior distributions for the precision).

The estimate of the SMR for area  $i$  relative to the reference mortality rate (from the indirect age standardization) is  $\exp(\alpha + v_i)$ , while  $\exp(v_i)$  is the residual relative risk for area  $i$  relative to the mean over the study area. Note that if an internal set of reference rates for indirect age standardization has been used, i.e. based on the observed rates over the whole study area itself, then  $\alpha$  will generally be close to zero. The  $\{v_i\}$  can also be thought of as latent variables that capture the effects of unknown or unmeasured area level covariates.

As shown in Figure 11.4, the Poisson log normal smooths the raw SMRs considerably, and yields a map similar to what we saw from the Poisson gamma model in Figure 11.3. Similarly, in the density plot in Figure 11.5, we see that the posterior distribution of the SMRs from the Poisson gamma and Poisson log normal models are virtually the same, with both showing substantial smoothing compared to the raw SMRs. Comparison of the DICs from the Poisson gamma model (DIC=682.1) to the Poisson log normal model (DIC=681.1) shows that both models are equally supported by the data and perform similarly in capturing the spatially varying risk surface.

The variance of the random effects ( $\sigma_v^2$ ) reflects the amount of extra-Poisson variation in the data, but may be difficult to interpret on its own as it represents this variation on the log(SMR) scale. A useful alternative summary of the random effects variability is to look at the ratio of quantiles of their empirical distribution. For example, one can define  $\text{QR}_{90} = \exp(q_{95\%} - q_{5\%})$ , where  $q_{5\%}$  is the log relative risk for the area ranked at the 5th percentile and  $q_{95\%}$  is the log relative risk for the area ranked at the 95th percentile.  $\text{QR}_{90}$  is thus interpretable as the **relative risk** between the top and bottom 5% of areas. Computing this based on the posterior samples of  $\{v_i\}$  from the Poisson lognormal model, one obtains an estimated  $\text{QR}_{90}$  of 3.80 (95% CI 2.90, 5.53). That is, there is an almost four-fold disparity in mortality comparing the 5% of census tracts with the highest mortality rates to the 5% with the lowest mortality rates.

In both the Poisson gamma and Poisson log normal models, the smoothing is global: the random effects  $\{v_i\}$  are treated as exchangeable and come from the same common distribution. As a result, all of the area specific relative risks are shrunk to the same overall mean. This kind of global smoothing does not allow for spatial correlation between risks in nearby areas, as might be expected if there is local

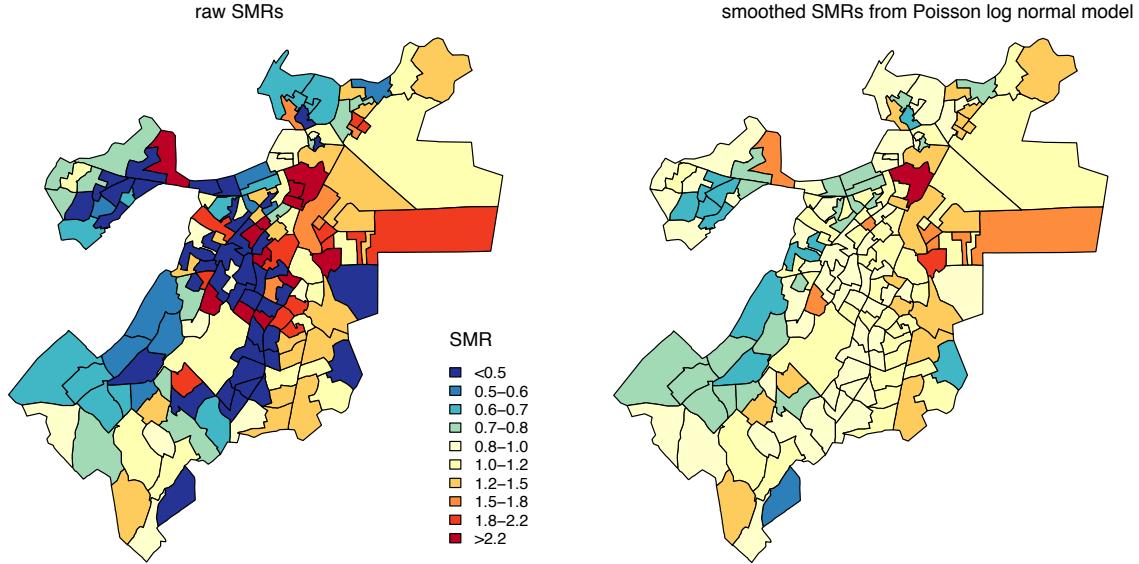


Figure 11.6: Map of the smoothed relative risk estimates for lung cancer mortality from the Poisson log normal model compared to the raw SMRs, for white non-Hispanics in Boston, 2000-2005.

clustering in the spatial pattern of risks. Local clustering of risks may be due to risk factors that are shared within and across census tracts: individuals who share these spatially varying risk factors would be expected to have similar outcomes. If such local clustering exists, this is a violation of the assumption of exchangeability. If these risk factors are known and observed, the most straightforward solution would be to include them as covariates in the model. However, in most observational settings, one can rarely measure, or even know about, all the relevant risk factors. If one suspects that these risk factors vary by area, it will be necessary to consider ways to allow for local clustering in models for  $\{\theta_i\}$ .

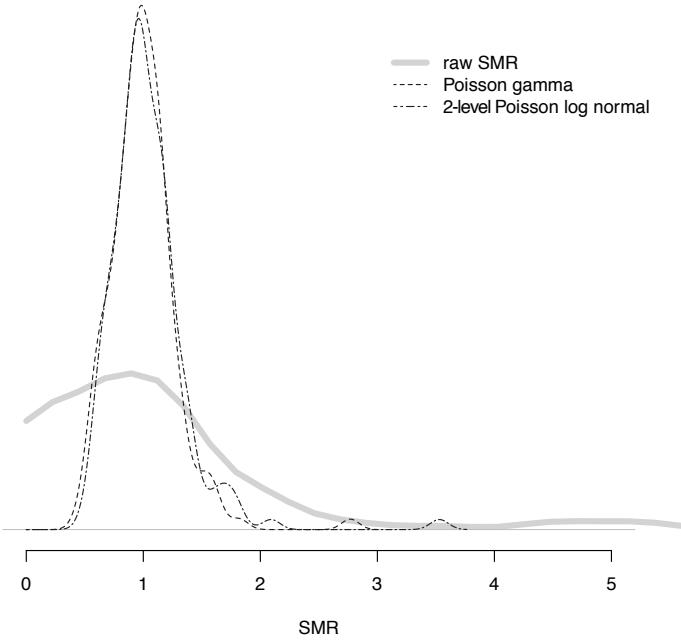


Figure 11.7: Density plots of the posterior estimates of  $\{\theta_i\}$  from the Poisson log normal model compared with those from the Poisson-gamma model and the raw SMRs.

## 11.5 Multilevel approaches to clustering

If one has some *a priori* information about how census tracts are organized, one might consider modelling clustering by specifying an additional level in the multilevel hierarchy. For example, Boston has well defined neighborhoods that are aggregations of census tracts. The result is that Boston residents have a strong sense of neighborhood identity. Additionally, neighborhoods are used as administrative units, e.g. by the Boston Public Health Commission and Boston Housing Authority, so that there is good reason to believe that neighbourhood level policies might lead to clustering among census tracts (Chen et al., 2006).

To adopt a multilevel (nested) approach to clustering, one can specify a three-level Poisson log normal model with  $i$  indexing census tracts and  $j$  indexing neighbour-



Figure 11.8: Boston neighbourhoods defined by the Boston Public Health Commission

hoods:

$$\begin{aligned} O_{ij} &\sim \text{Poisson}(\theta_{ij} E_{ij}) \\ \log(\theta_{ij}) &= \alpha + v_{ij} + z_j \\ v_{ij} &\sim \text{Normal}(0, \sigma_v^2) \\ z_j &\sim \text{Normal}(0, \sigma_z^2) \end{aligned}$$

Here, the  $\{v_{ij}\}$  are independently normally distributed *census tract* effects and the  $\{z_j\}$  are independently normally distributed *neighbourhood* effects. Note that the census tracts and neighbourhood effects are conditionally independent. As with the two-level Poisson log normal model, the model for  $\log(\theta_i)$  could also include any number of fixed covariate effects, although for the moment I include just an intercept ( $\alpha$ ). Parameters of interest from this model include the census tract specific SMRs ( $\text{SMR}_i = \exp(\alpha + v_{ij} + z_j)$ ), the variance of the census tract random effects ( $\sigma_v^2$ ), the variance of the neighbourhood random effects ( $\sigma_z^2$ ), and the empirical ratio of the 95th to 5th percentile of the SMRs ( $\text{QR}_{90}$ ).

Figure 11.7 compares the resulting maps of the SMRs from the 2- and 3-level Poisson log normal models. The heavier lines in the second map correspond to the neighbourhood boundaries. One can see that the 3-level model allows for local

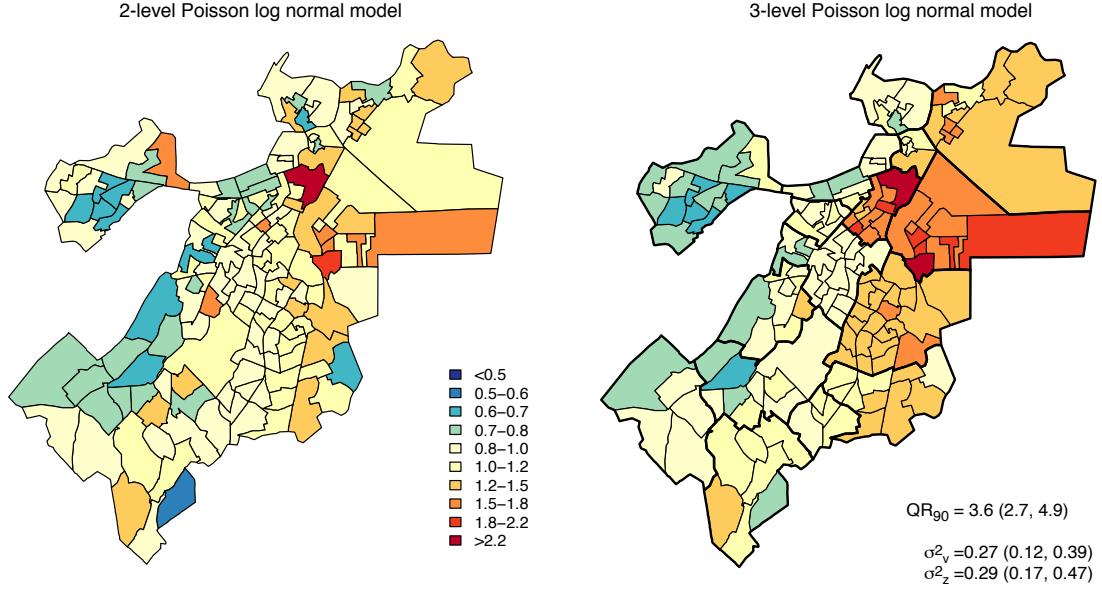


Figure 11.9: Comparison of the mapped smoothed SMRs from the 2-level and 3-level Poisson log normal models

clustering of risk in certain neighbourhoods. The census tract level variance ( $\sigma_v^2$ ) is 0.27 (95% CI 0.12, 0.39) and the neighbourhood level variance ( $\sigma_z^2$ ) is 0.29 (95% CI 0.17, 0.47), and the empirical ratio of the SMRs ( $QR_{90}$ ) is 3.6 (95% CI 2.7, 4.9). Comparison of the DIC for the 3-level Poisson log normal model (DIC=672.3) to the 2-level log normal model (DIC=681.1) shows that the 3-level model offers a substantially better fit to the data.

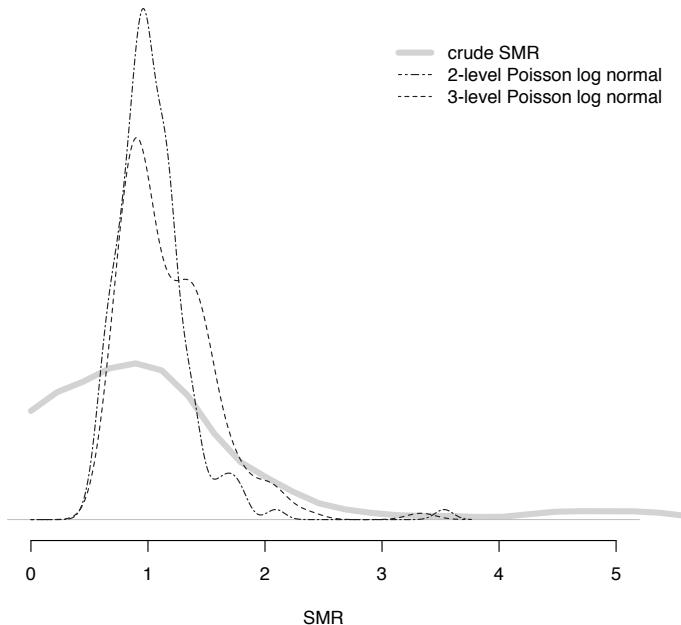


Figure 11.10: Density plots comparing the raw SMRs and the posterior estimates of  $\{\theta_i\}$  from the 2-level and 3-level Poisson log normal models.

## 11.6 Besag, York and Mollié model

In the multilevel Poisson log normal model with independent census tracts and neighbourhoods, the neighbourhood units were specified *a priori*. In a sense, the neighborhood boundaries are “hard” boundaries, and crossing a boundary means that one’s relative risk may jump by quite a lot. What if one wants to account for spatial correlation in a “smoother” manner? One way to do this is to specify a multivariate normal prior distribution for all the area parameters with a spatially-structured covariance matrix. Many different ways have been proposed to specify spatially structured multivariate normal distributions for  $\log(\theta_i)$  (Wakefield et al., 2000; Banerjee et al., 2004).

One particular form of the multivariate normal distribution that is commonly used is the *intrinsic Gaussian conditional autoregression* (CAR) prior suggested by Clayton and Kaldor (1987) and developed by Besag et al. (1991). This is one of

the most popular ways of dealing with spatial autocorrelation. The spatial structure is formulated through a set of conditional autoregressions, which uses the fact that if a vector of random variables has a multivariate normal distribution, then the distribution of each element of that vector conditional on all the other elements in the vector is also normal, with mean and variance that depend on the original multivariate mean and covariance matrix.

As with the other models, the first stage model assumes that the observed counts are Poisson distributed, and that an additive model for  $\log(\theta_i)$  can be specified for accommodating covariate effects:

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \alpha + u_i \end{aligned}$$

Instead of the independent normal prior for the distribution of the census tract effects, one models

$$u_i | u_{j, j \neq i} \sim \text{Normal}(\mu_i, \tau_u^2 / m_i)$$

where

$$\begin{aligned} \mu_i &= \frac{\sum_j w_{ij} u_j}{\sum_j w_{ij}} \\ w_{ij} &= \begin{cases} 1 & \text{if } i, j \text{ are adjacent,} \\ 0 & \text{if they are not} \end{cases} \end{aligned}$$

and  $m_i$  is the number of adjacent areas. To understand this, consider the highlighted census tract  $i$  in Figure 11.9. If  $\partial_i$  is the set of areas adjacent to  $i$ , and one sets  $w_{ij}$  to 1 for areas  $j \in \partial_i$  and zero otherwise, then the prior distribution for  $u_i$  has conditional mean equal to the average of the neighbouring  $u_j$ 's and variance inversely proportional to the number of adjacent neighbours. The effect is to smooth  $u_i$  toward the mean risk in the set of neighbouring areas. Note that  $\tau_u^2$  is the *conditional* variance (scaled by  $\sum_{j \in \partial_i} w_{ij} = m_i$ , i.e. the number of neighbours); to emphasize that it is only interpretable conditionally, I have changed the notation from  $\sigma_u^2$  to  $\tau_u^2$ .

Besag et al. (1991) recommended combining the CAR prior and the standard normal prior to allow for both spatially unstructured and spatially correlated random effects:

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \alpha + v_i + u_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \\ u_i | u_{j, j \neq i} &\sim \text{Normal}(\mu_i, \tau_u^2 / m_i) \end{aligned}$$

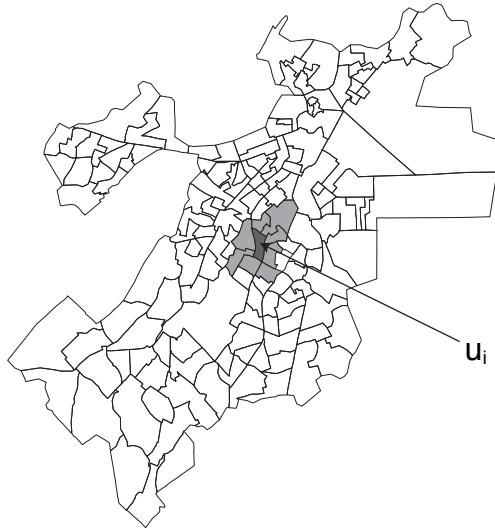


Figure 11.11: An illustration of the CAR idea:  $U_i$  is smoothed toward the mean risk of the neighbouring areas

This model is often referred to as the convolution model or the Besag, York and Mollié (BYM) model. The unstructured  $\{v_i\}$  can be thought of as capturing unstructured heterogeneity, or correlation *within* areas. In contrast, the spatially structured  $\{u_i\}$  capture spatial correlation *across* areas (Bernardinelli et al., 1995). Note, however, that  $\sigma_v^2$  (unstructured heterogeneity variance) and  $\tau_u^2$  (spatial variance) are not directly comparable:  $\sigma_v^2$  reflects the **marginal** variability of the unstructured random effects between areas, while  $\tau_u^2$  is the **conditional** variance of the spatial effect in area  $i$ , conditional on values of neighboring spatial effects. No closed-form expression is available for the **marginal** between-area variance of the spatial effects. However, in the Bayesian approach, the marginal spatial variance  $s_u^2$  can be estimated empirically from the posterior samples of  $\{u_i\}$ :

$$s_u^2 = \sum_i (u_i - \bar{u})^2 / (n - 1)$$

where  $\bar{u}$  is the average of the  $\{u_i\}$ .

Technically, although the univariate conditional prior distributions for each  $u_i$  are well defined, the corresponding joint prior distribution for  $\{u_i\}$  is now improper (undefined mean and infinite variance). It is necessary to impose a constraint to

ensure that the model is identifiable. This is accomplished by constraining the  $\{u_i\}$  to have a zero mean and specifying an improper uniform prior distribution for the intercept  $\alpha$ .

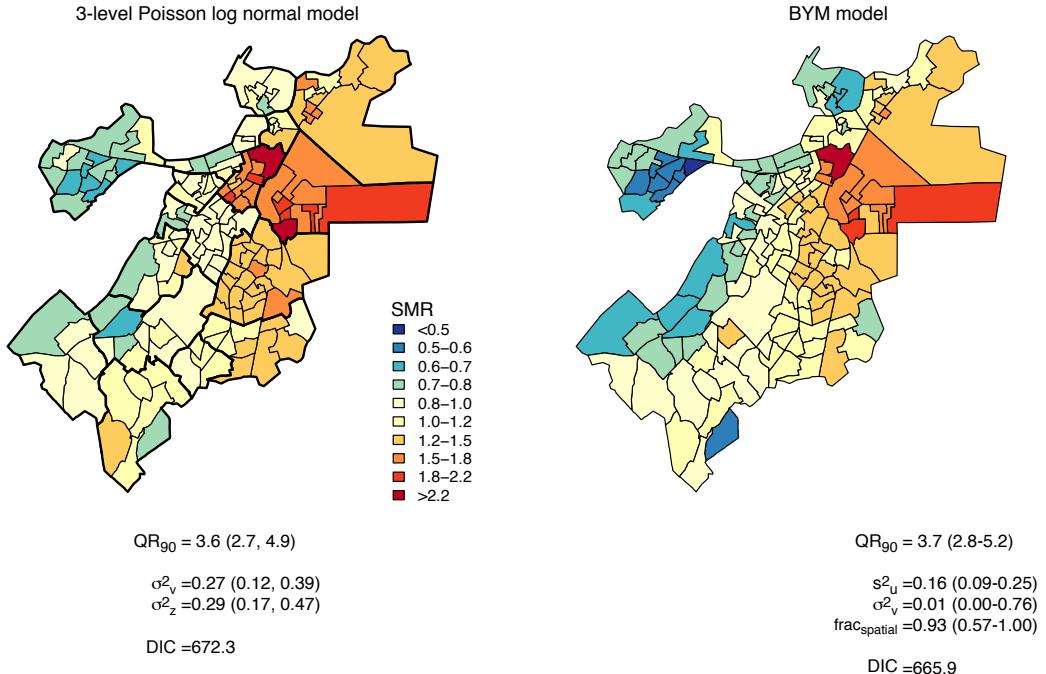


Figure 11.12: Mapped smoothed SMRs from Poisson CAR model (right) compared with Poisson log normal model (left)

With an estimate of the marginal between-area variance, one can also characterize the relative contribution of spatial vs. unstructured heterogeneity:

$$frac_{spatial} = s^2_u / (s^2_u + \sigma^2_v)$$

When  $frac_{spatial}$  is close to 1, spatial heterogeneity dominates. When  $frac_{spatial}$  is close to 0, unstructured heterogeneity dominates (Best et al. 2005).

Figure 11.10 compares the 3-level Poisson log normal model to the BYM model. Both models accommodate local clustering of census tract risks, compared with the

2-level Poisson log normal model in Figure 11.4, with some differences that are also reflected in the different distributions in Figure 11.11. The marginal variance of the spatially structured effects,  $s_u^2=0.16$  (95% CI 0.09, 0.25), is substantially larger than the variance of the unstructured effects,  $\sigma_v^2=0.01$  (95% CI 0.00, 0.76), showing that the map is dominated by spatially structured heterogeneity ( $\text{frac}_{\text{spatial}}=0.93$ , 95% CI 0.57, 1.00). Examination of the DICs suggests that the BYM model yields a better fit to the data (DIC=665.9), compared with the 3-level Poisson log normal model (DIC=672.3). One way to think about this is that, while the defined Boston neighbourhoods tend to capture how the census tract level risks are clustered, the *a priori* defined neighbourhood structure does not compare this perfectly, while in the BYM model the clustering is informed by the data.

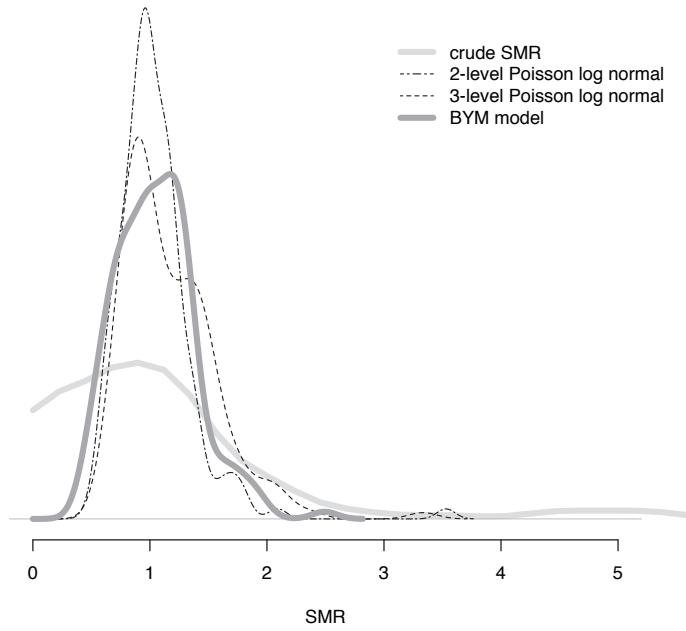


Figure 11.13: Density plots comparing the raw SMRs and the posterior estimates of  $\{\theta_i\}$  from the 2-level Poisson log normal, 3-level Poisson log normal, and BYM models.

In the spatial modelling literature, the intrinsic CAR model based on adjacency

weights has been a popular choice for modelling spatial variation in disease rates with local smoothing. However, the method entails making several assumptions about the weighting structure that should be acknowledged. Firstly, there is the question of how to define adjacency, particularly for irregularly shaped areas. One basic choice involves deciding whether adjacent neighbours are to be defined based only on shared edges (the Rook method) or shared edges and vertices (the Queen method; see Chapter 1). Secondly, while the intrinsic CAR model defines the conditional distributions using weights of 1 for adjacent areas and 0 for non-adjacent areas, other weighting schemes are possible for the general class of CAR models. For example, weights could be defined in a smoother manner as a decay function of distance, e.g. distance between population-weighted centroids of the areas. This function may be obtained empirically by fitting a correlogram to residuals from a model that does not include a spatially-structured random effect, in order to determine the likely range of spatial autocorrelation (Cressie and Chan, 1989).

Published studies using various CAR approaches have rarely provided justification for the choice of the neighbourhood weight matrix structure. Best et al. (2001) investigated the relationship between benzene emissions and the incidence of childhood leukaemia in Greater London, and considered three alternative levels of data aggregation in their analysis. They examined adjacency vs. distance-based neighbourhood spatial weights, and found, for both grid-level and ward-level analyses, that the adjacency based neighbourhood structure provided a better fit of the data, based on DIC comparisons. They also found a significant difference in the estimates of the spatially structured random effects between the distance-based and adjacency-based neighbourhood structures for the ward-level analysis without any covariates. Earnest et al. (2007) explored a variety of adjacency- and distance-based weight matrices in an study of birth defects in New South Wales, Australia. They found that, in terms of agreement between observed and predicted relative risks, the distance-based matrices tended to perform better than those based on adjacency. The attributed this finding in part to the highly irregular shapes of the areas, and suggested that adjacency models would be expected to perform better with regular shaped areas such as grids. Thus, while the intrinsic CAR approach provides a convenient implementation of local smoothing, when possible, it is advisable to conduct an exploratory analysis of the neighbourhood weight matrix to guide the choice of a suitable weight scheme.

While the multilevel modelling framework is usually specified in terms of nested random normal effects, multilevel models *can* also accommodate spatial correlations via their extension to multiple membership models (Hill and Goldstein, 1998). Langford et al. (1999) showed how to use such models to fit spatial data and Browne et al. (2001) present Bayesian extensions of this model as a member of the family of

models called multiple-membership multiple-classification (MMMC) models.

## 11.7 CAR and neighbourhood effects

Given an *a priori* belief on the one hand of the importance of neighbourhoods and an appreciation, on the other hand, of the idea of smooth, spatially autocorrelated effects, one might be tempted to combine the CAR structure at the census tract level with independent neighbourhood effects by modelling

$$\begin{aligned} O_{ij} &\sim \text{Poisson}(\theta_{ij} E_{ij}) \\ \log(\theta_{ij}) &= \alpha + v_{ij} + u_{ij} + z_j \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \\ u_i | u_j, j \neq i &\sim \text{Normal}(\mu_i, \tau_u^2/m_i) \\ z_j &\sim \text{Normal}(0, \sigma_z^2). \end{aligned}$$

This model includes spatially structured and unstructured effects at the census tract level along with neighbourhood level unstructured effects.

I compare this model to the BYM model and the 3-level Poisson log normal model in Figure ???. The map that this model generates is virtually indistinguishable from what is obtained from the BYM model, but there is evidence that the more complex model is overspecified. The variance of the neighbourhood effects ( $\sigma_z^2$ ) is close to zero and the DIC for this model is not appreciably different for DIC from the BYM model. Figure 11.12 shows that the distribution of census tract effects is essentially the same as the BYM model, leading one to conclude that the neighbourhood level effects do not add much beyond the clustering already captured by the conditionally autoregressive structure.

This exercise in model building suggests that, while one may have compelling conceptual reasons to specify a complex multilevel structure, care must be taken to ensure that the model is not overspecified and that the data support the proposed structure. In the Boston lung cancer mortality example, the conditionally autoregressive spatial structure captures most of the census tract level variability, including the local clustering, as evidenced by the high value for  $\text{frac}_{\text{spatial}}$  and the fact that the neighbourhood level structure adds little to the model. Notably, the BYM model does support the notion of neighbourhood level variability in the sense of clustered census tracts that exhibit similarly elevated or depressed risks. However, the improved fit of the BYM model over the 3-level Poisson log normal model with pre-defined Boston neighbourhoods suggests that the official neighbourhood bound-

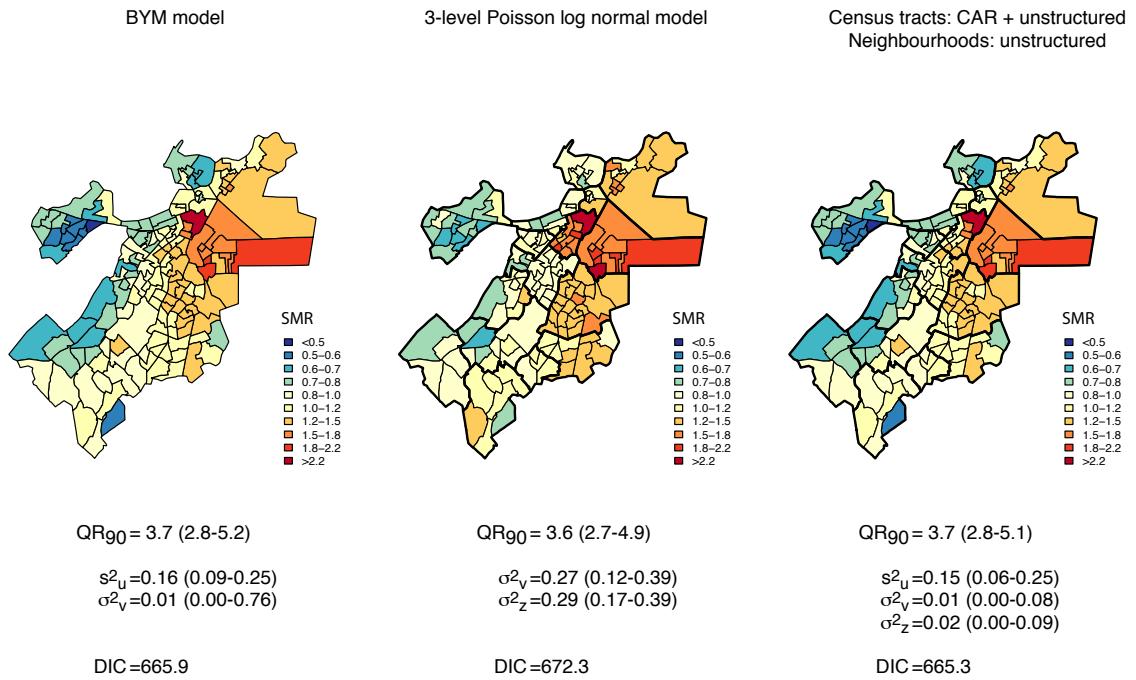


Figure 11.14: Comparing the BYM model, the 3-level Poisson log normal model, and a model with CAR + unstructured variability at the census tract level and unstructured neighbourhood variability.

aries do not perfectly coincide with the actual pattern of spatial clustering of lung cancer mortality risks.

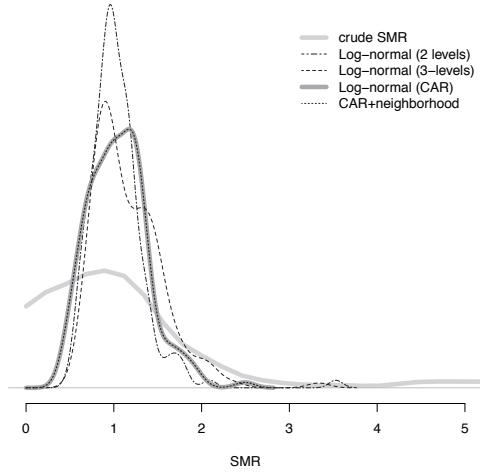


Figure 11.15: Density plots comparing the raw SMRs and the posterior estimates of  $\{\theta_i\}$  from the BYM model, 3-level Poisson log normal, and a model with CAR + unstructured variability at the census tract level and unstructured neighbourhood variability.

## 11.8 Covariate effects

Having established that the BYM model best describes the census tract level variation in lung cancer mortality risk among white non-Hispanics, I now turn to fitting the BYM model with covariate effects. Documenting and monitoring socioeconomic disparities in cancer outcomes is of great interest to social epidemiologists and public health policy makers, and a substantial literature has linked socioeconomic deprivation to poorer cancer outcomes. To estimate socioeconomic disparities in lung cancer mortality in white non-Hispanics in Boston, geocoded death records from 2000-2005 were linked to US Census 2000 estimates of the percent of individuals living below the US defined poverty line at the census tract level. Figure 11.13 shows the geographic variation in census tract poverty in Boston, and shows many areas where a substantial proportion of residents are living below the poverty line ( $\geq 20\%$  below poverty). Census tract level poverty ranged from 2.3% to 57.4%, with a median of 17.4%. Over the whole city, 19.5% of Boston residents in 2000 were living below the poverty line, compared with 9.3% of Massachusetts residents.

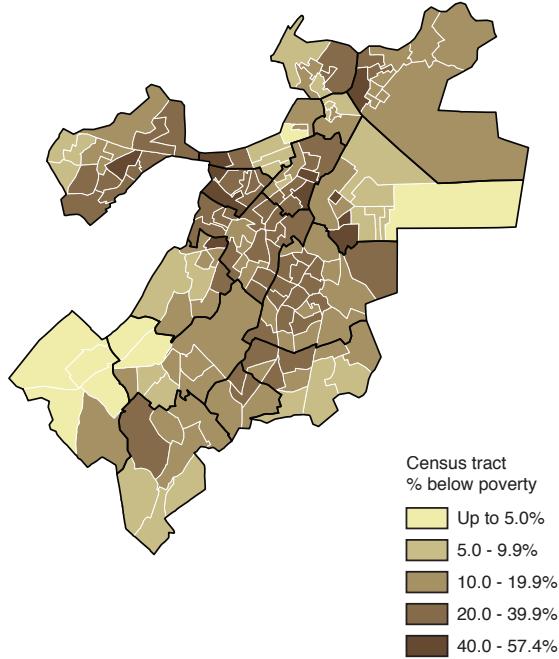


Figure 11.16: Census tract % below poverty (US Census 2000 Summary File 3) in Boston, Massachusetts, US.

If one compares this map to the smoothed SMRs from the BYM model in Figure 11.10, one sees that there are numerous areas where high poverty coincides with high lung cancer mortality. However, there are also areas that do not follow this pattern, where high poverty coincides with more modest lung cancer mortality risk or where lower poverty coincides with higher lung cancer mortality risk.

To evaluate the overall association between census tract poverty across Boston census tracts, I fit the following model, which includes the census tract percent below poverty measure as a continuous covariate:

$$\begin{aligned}
 O_{ij} &\sim \text{Poisson}(\theta_{ij} E_{ij}) \\
 \log(\theta_{ij}) &= \alpha + \beta * \text{poverty} + v_{ij} + u_{ij} \\
 v_i &\sim \text{Normal}(0, \sigma_v^2) \\
 u_i | u_j, j \neq i &\sim \text{Normal}(\mu_i, \tau_u^2 / m_i).
 \end{aligned}$$

To fit the model in the fully Bayesian framework, I specified a diffuse normal prior for  $\beta$  ( $\sim \text{Normal}(0, 10^4)$ ). Table 11.1 presents the model results, including the estimated poverty gradient and summaries of area variation from the BYM model. I report the intercept and  $\beta$  for census tract poverty on the exponentiated scale so that these estimates are interpretable on the relative risk scale.

Recalling that the expected deaths in each census tract were computed using age specific reference rates for Massachusetts as a whole, the intercept effect of  $\exp(\alpha) = 1.39$  shows that White Non-Hispanics in Boston experienced lung cancer mortality rates nearly 40% higher than the statewide lung cancer mortality rates. There was a strong socioeconomic gradient in lung cancer mortality, with those living in more impoverished census tracts experiencing higher lung cancer mortality rates. As with the BYM model, the residual census tract variation in mortality was dominated by spatially structured heterogeneity ( $\text{frac}_{\text{spatial}} = 3.61$ , 95% CI 2.73, 3.64).

Interestingly, while one observes a significant socioeconomic gradient in lung cancer mortality by census tract poverty, adjusting for census tract poverty does not substantially reduce the census tract level variation. The empirical marginal variance of the CAR effects ( $s_u^2$ ) was estimated as 0.16 in both this and the earlier BYM model. Similarly, in the model adjusting for census tract poverty, the DIC was 662.6, compared with 665.9 in the previous model. This leads one to the somewhat counterintuitive conclusion that census tract poverty is, on the one hand, a significant predictor of lung cancer mortality, but that it contributes only a small fraction to the overall census tract variation in lung cancer mortality risk.

Since  $QR_{90}$  expresses the residual census tract variation in lung cancer mortality on the relative risk scale, one can gain a sense of what this variation means in terms of risk: the risk ratio comparing the 5% of census tracts with the highest mortality rates to the 5% with the lowest mortality rates is 3.61 (95% CI 2.73, 3.64). In contrast, based on the model estimate of the poverty gradient, comparing the census tract with the maximum census tract poverty level in Boston (54.7% below poverty) to census tracts with 0% poverty yields a risk ratio of around 2.0. Thus, the variation in lung cancer mortality risk across census tracts exceeds the variation due to poverty, and even after adjusting for census tract poverty, there remains a great deal of unexplained census tract level variation in mortality.

Figure 11.14 maps the posterior estimates of the census tract SMRs from the BYM model before and after adjusting for census tract poverty. As suggested by the model results concerning the lack of change in the variances of the structured and unstructured spatial heterogeneity, the maps show that adjustment for census tract poverty does little to change the observed pattern of lung cancer mortality risk, other than minor changes in the census tract SMR for some of the census tracts.

Of course, census tract poverty is but one area-based socioeconomic variable, and there remain many potential risk factors, both individual and contextual, that may contribute to the area variation in lung cancer mortality rates. On a related note, since this estimated socioeconomic gradient is based on an area-based socioeconomic measure, one must be cautious about making inferences about the relationship between an individual's socioeconomic position and their risk of lung cancer death based on the observed ecologic association. For an in depth discussion of ecologic bias, see Greenland (1992) and Elliott and Wakefield (1999). Similarly, while an area-based socioeconomic variable can be conceptualized in its own right as reflecting "contextual" socioeconomic factors and not just as a proxy for individual-level socioeconomic position, the estimated socioeconomic effect from this model reflects both individual-level and contextual processes.

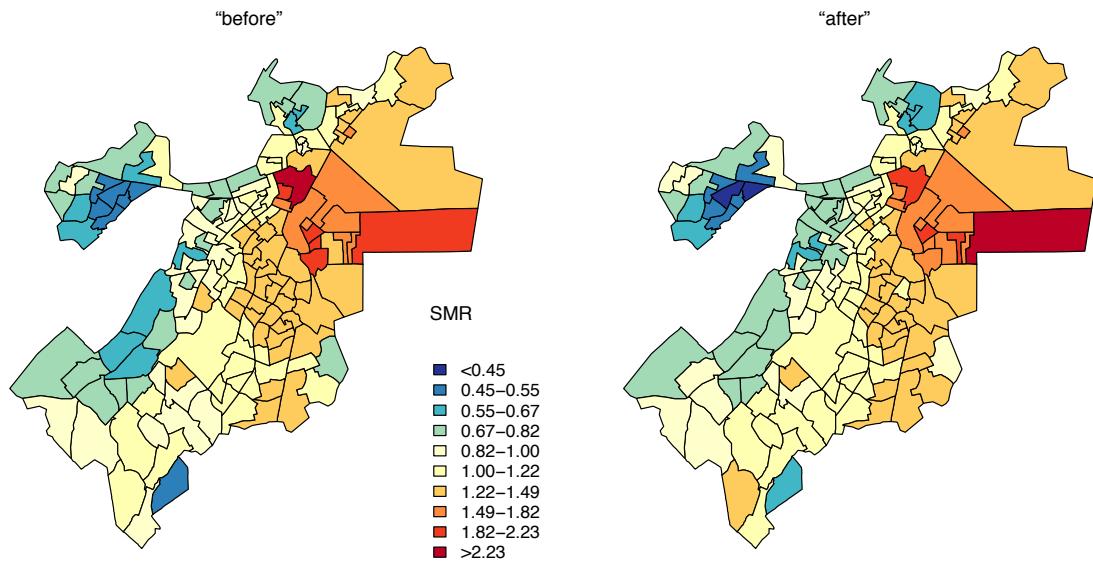


Figure 11.17: Census tract SMRs from the BYM model, before and after adjustment for census tract poverty.

Table 11.1: Covariate effects and summaries of area variation from BYM model before and after adjustment for census tract poverty.

	<b>Before</b>	<b>After</b>
<b>Covariate Effects</b>	<b>RR (95% CI)</b>	<b>RR (95% CI)</b>
Intercept ( $e^\alpha$ )	1.13 (1.03, 1.24)	1.39 (1.26, 1.38)
Census tract % below poverty	-	3.54 (1.45, 8.66)
<b>Other Summary Quantities</b>	<b>Estimate (95% CI)</b>	<b>Estimate (95% CI)</b>
$s_u^2$	0.16 (0.09, 0.25)	0.16 (0.09, 0.24)
$\sigma_v^2$	0.01 (0.00, 0.08)	0.01 (0.00, 0.05)
frac <sub>spatial</sub>	0.93 (0.57, 1.00)	0.96 (0.70, 1.00)
QR <sub>90</sub>	3.73 (2.83, 5.17)	3.61 (2.73, 3.64)
DIC	665.9	662.6

## 11.9 Summary

This chapter has presented an introduction to how health data in the form of georeferenced counts can be analyzed using hierarchical models. These models are useful when one is interested in estimating, visualizing, and comparing area-specific relative risks, particularly when the areas are small and the number of events in each area may be sparse. A key feature of this approach is that the unstable area-level parameters from the first stage of the model can be smoothed by specifying a second stage model that may, in turn, depend on a set of hyperparameters. This allows for more efficient estimation of regression parameters and also allows the standard errors of estimates to account appropriately for correlation between observations within areas and uncertainty in prediction and estimation.

I have compared several different modelling choices that allow for local clustering, including a multilevel approach based on nested independent normal effects at the census tract and neighbourhood levels and the popular intrinsic Gaussian CAR model (BYM). Both models also allow for covariate adjustment and can be used to estimate the effect of fixed covariates conditional on the random effects structure. In our specific example focusing on estimation and mapping of the census tract level SMRs for lung cancer mortality among white non-Hispanics in Boston, the BYM model gave the best fit to the observed data. In practice, however, the choice of modelling approach may depend on *a priori* beliefs about multilevel structure for a particular dataset (e.g. the salience of administrative and policy-relevant levels such as census tracts, counties, provinces or states) vs. taking a more empirical approach to local smoothing. The inferential targets (i.e. estimates of the covariate effects vs. posterior estimates of the area-specific effects) should also inform modelling choices.

In particular, while hierarchical models are widely used in the spatial epidemiologic literature, it is important to note that estimates of the covariate effects from these hierarchical models are conditional on the correct specification of the random effects structure. If the model is not correctly specified, including if the random effects structure does not incorporate the relevant levels, this can potentially lead to bias. When interest focuses on the estimating covariate effects and the correlation between observations from the same area is to be treated as a nuisance, an alternative strategy to modelling may be to use a generalized estimating equations (GEE) approach (Liang and Zeger, 1986). Parameter estimates from the GEE are consistent even when the covariance structure is misspecified and Hubbard et al. (2010) argue that, when population averaged covariate effects are the primary interest, robustness to model mis-specification may be desirable.

Hierarchical models lend themselves to specification in the Bayesian framework, and as shown here, the posterior samples obtained from MCMC simulation can be used to obtain empirical estimates and credible intervals for several useful quantities (including, in these examples,  $QR_{90}$ , the marginal variance of CAR effects  $s_u^2$ , and  $\text{frac}_{\text{spatial}}$ ). Caution must be used, however, as the Bayesian approach requires specifications of prior distributions of all model parameters, and model behaviour may be sensitive to these priors. While one typically tries to specify uninformative prior distributions so that the posterior likelihoods will be dominated by the data, in practice it is a good idea to try several different prior specifications in order to evaluate sensitivity. For example, in these models, I specified  $\text{Gamma}(\varepsilon, \varepsilon)$  priors for the inverse of the variances of the random effects (i.e.  $\sigma_v^2, \tau_u^2, \sigma_z^2$ ), with  $\varepsilon$  very small (say  $10^2$  or  $10^3$ ) (Lawson et al. 2003). However, Gelman (2006) and Kelsall and Wakefield (1999) point out that even small values of  $\varepsilon$  can be informative and inferences become sensitive to  $\varepsilon$ . As an alternative, Gelman (2006) suggests using a half Cauchy prior for the precision of variance parameters.

While extremely flexible and able to deal with complex models, MCMC methods do involve computationally intensive simulations to obtain the posterior distribution for the parameters, and models for large datasets with many parameters may take a long time to fit. While multilevel Poisson models can also be fit in a frequentist setting using a penalized quasi-likelihood algorithm (PQL), inference based on posterior samples is not available, and estimates of random effects may be biased in multilevel logistic models for binary response data. The Integrated Nested Laplace Approximation (INLA; Rue et al., 2009) approach has been recently developed as a computationally efficient alternative to MCMC. INLA is designed for the broad class of latent Gaussian models, including (generalized) linear mixed, spatial, and spatio-temporal models, and is gaining in popularity for spatial epidemiologic applications.

While the BYM based on the intrinsic CAR prior is widely used in spatial epidemiologic analyses, development of hierarchical models for spatial and spatio-temporal data continues to be an active and vibrant research area. For example, some authors have argued that the parametric BYM model may result in oversmoothing of the risk surface, and have developed semiparametric spatial models that allow for discontinuities in the risk surface and make fewer distributional assumptions (Best et al. 2005, Richardson et al. 2004). The BYM model also requires a well-defined set of areal boundaries in order to determine adjacency, and this can be complicated in the spatio-temporal setting if boundaries shift. Kelsall and Wakefield (2002) argue for the use of a smooth underlying surface based on a Gaussian random field prior to model areal data, with each areal observation related to the average of the surface over the defined area. Hund et al. (2012) build on this to develop a model for temporally misaligned data.

As a general framework for analyzing disease risk across geographic areas, hierarchical modelling provides powerful and flexible tools for incorporating available information about how disease risk varies over populations in space. A multilevel statistical perspective can also encourage one to think creatively about causal processes that take place at multiple levels, and lead one to consider relevant exposures and covariates on multiple spatial scales. As methods and computational techniques for hierarchical models continue to evolve, and as rich georeferenced databases continue to proliferate, it is anticipated that these kinds of models will be used more and more widely by epidemiologists and health researchers.

## References

- Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society* 144, 148–161.
- Aitkin, M. and Longford, N. (1986) Statistical modelling in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society* 149, 1–43.
- Banerjee, S., Carlin, B.P., Gelfand, A.E. (2004) *Hierarchical modeling and analysis for spatial data*. Chapman & Hall, Boca Raton, FL.
- Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1–59.
- Best, N., Cockings, S., Bennett, J., Wakefield, J., Elliott, P. (2001) Ecological re-

- gression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society Series A* 164:155–174.
- Best, N., Richardson, S. and Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14(1), 35–49.
- Bitthell, J.F. (2000) A classification of disease mapping methods. *Statistics in Medicine* 19, 2203–2215.
- Blakely T. and Subramanian, S.V. (2006) Multilevel studies. In: Oakes, J.M. and Kaufman, J.S. (eds). *Methods in social epidemiology*. Jossey-Bass, San Francisco, CA. pp 316–340.
- Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of American Statistical Association* 88, 9–25.
- Breslow, N.E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 82, 81–91.
- Browne, W., Goldstein, H., and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling* 1, 103–124.
- Chen, J.T., Rehkopf, D.H., Waterman, P.D., Subramanian, S.V., Coull, B.A., Cohen, B., Ostrem, M. and Krieger, N. Mapping and measuring social disparities in premature mortality: the impact of census tract poverty within and across Boston neighborhoods, 1999–2001. *Journal of Urban Health* 83(6), 1063–1084.
- Clayton, D.G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In: Elliott, P., Cuzick, J., English, D. and Stern, R. (ed.) *Geographical and environmental epidemiology: methods for small-area studies*. Oxford University Press, London, UK. pp. 205–20.
- Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43(3), 671–681.
- Cressie, N. and Chan, N.H. (1989). Spatial modelling of regional variables. *Journal of the American Statistical Association* 84, 393–401.
- Elliott, P. and Wakefield, J.C. (1999). Bias and confounding in spatial epidemiology. In: Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. *Spatial Epidemiology: Methods and Applications*, Oxford University Press, London, UK. pp. 68–84.
- Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple

- sequences (with discussion). *Statistical Science* 7, 457–511.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003) *Bayesian Data Analysis* Chapman & Hall, New York, NY.
- Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1(2), 1–19.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.
- Goldstein, H. (2011). *Multilevel Statistical Models*. 4th ed. Edward Arnold, London, UK.
- Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics In Medicine* 11, 1209–23.
- Hill, P. and Goldstein, H. (1998). Multilevel modelling of educational data with cross-classifications and missing identification of units. *Journal of Educational and Behavioral Statistics* 23, 117–128.
- Hubbard, A.E., Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N., Bruckner, T. and Satariano, W.A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 21(4), 467–474.
- Hund, L., Chen, J.T., Krieger, N., and Coull, B.A. (2012). A geostatistical approach to large-scale disease mapping with temporal misalignment. *Biometrics* 68(3), 849–58.
- Kelsall, J.E., and Wakefield, J.C. (1999) Discussion of Bayesian models for spatially correlated disease and exposure data by N. Best, L. Waller, A. Thomas, E. Conlon and R. Arnold. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., editors. Sixth Valencia International Meeting on Bayesian Statistics. London: Oxford University Press.
- Langford, I., Leyland, A., Rasbash, J., Goldstein, H. (1999) Multilevel modelling of the geographical distribution of rare diseases. *Journal of the Royal Statistical Society* 48, 253–268.
- Lawson, A., Böhning, D., Biggeri, A., Lesaffre, E., Viel, J-F. (1999) Disease mapping and its uses. In: Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J-F., Bertollini, R. (ed.) *Disease mapping and risk assessment for public health*. John Wiley & Sons, Chichester, UK. pp. 3–13.

- Lawson, A., Browne, W.J., Vidal Rodeiro, C.L. *Disease Mapping with WinBUGS and MLwiN*. John Wiley & Sons, Chichester, UK.
- Lin, X and Breslow, N.E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of American Statistical Association* 91, 1007–1016.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.
- Richardson, S. (2003). Spatial models in epidemiological applications. in: Green, P., Hjort, N. and Richardson, S. (ed.) *Highly Structured Stochastic Systems*. Oxford University Press, London, UK. pp. 237–259.
- Rue, H., Martino, S., and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B* 71(2), 1–35.
- Spiegelhalter, D.J, Best, N.G., Carlin, B.P., and Van der Linde, A. (2002) Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Journal of the Royal Statistical Society, Series B*, 64:583-640.
- Subramanian, S., Jones, K. and Duncan, C. (2003). Multilevel methods for public health research. In: Kawachi, I. and Berkman, L. (ed.) *Neighborhoods and health*. Oxford Press, NY. pp. 65–111.
- Snijders, T. A. B. and Bosker, R. J. (2011). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Sage, London, UK.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(2), 319–392.
- Wakefield, J.C., Best, N.G. and Waller, L. (2000). Bayesian approaches to disease mapping, In: Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. *Spatial Epidemiology: Methods and Applications*, Oxford University Press, London, UK. pp 104–127.
- Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics, *Biometrika* 85, 251–267.