

# Hierarchical models for disease mapping

The analysis of disease rates or counts from small areas often involves a trade-off between statistical stability of estimates and geographic precision. Detection of locally elevated risk or rates requires geographically small units to distinguish local risk/rates from area-wide values. On the other hand, smaller regions result in rate estimates based on smaller population sizes. For a rare disease, small population sizes result in particularly unstable rate estimates. The statistical literature contains various methods of combining information or ‘borrowing strength’ (*see Shrinkage*) between regions to achieve local rate stabilization without losing geographic resolution.

The most common approaches involve *hierarchical models* with random effects (intercepts) for each region. The use of random effects presumes that regional rates are drawn from some common superpopulation (or superpopulations) of rates, and allows the analyst to combine information from several regions. The resulting estimates typically involve a weighted average of the specific region’s crude rate and the rates from other regions. Manton, et al. [23] offer a maximum likelihood approach. Tsutakawa [31] considers an approach based on normally distributed logits of disease risk in a Bayesian setting. Both approaches result in a compromise between individual regional estimates and the overall disease rate observed in all counties. Clayton and Kaldor [10] introduce hierarchical models and associated empirical Bayesian inference for region-specific standardized mortality/morbidity ratios (SMRs) which allow spatial correlation between neighboring regions. Besag et al. [5] extend these to a fully Bayesian setting using Markov chain Monte Carlo algorithms. Kaldor and Clayton [19] provide a conceptual overview of **empirical Bayes methods**, while Clayton and Bernardinelli [9] and Mollié [24] provide thorough introductions to the fully Bayesian approach (*see Bayesian methods and modeling*). The Clayton and Kaldor [10] and Besag et al. [5] models motivate much of the recent disease mapping literature and these are detailed below.

Typical **disease mapping** data contain observed ( $Y_i$ ) and expected (often age-standardized) disease counts ( $E_i$ ) for subregion  $i$ ,  $i = 1, \dots, I$ , where the

$I$  subregions partition the study area of interest. The **maximum likelihood estimate** (MLE) of the SMR for region  $i$  is  $Y_i/E_i$ . The first stage of the model presumes a Poisson distribution for the regional counts, conditional on log relative risks ( $\mu_i$ ) associated with each region  $i$ , i.e.

$$Y_i | \mu_i \stackrel{\text{ind}}{\sim} \text{Poisson} [E_i \exp(\mu_i)],$$

$$i = 1, \dots, I \quad (1)$$

(The reason for considering log relative risks becomes apparent below.)

For the second stage, Clayton and Kaldor [10] consider a variety of **prior distributions** for  $\exp(\mu_i)$ . In the simplest case, we assign a conjugate gamma distribution with scale parameter  $\alpha$  and shape parameter  $\nu$  (i.e. mean  $\nu/\alpha$  and variance  $\nu/\alpha^2$ ). This yields a **negative binomial distribution** for  $Y_i$ , unconditional on the  $\exp(\mu_i)$ s, providing posterior expectation

$$E[\exp(\mu_i) | Y_i; \alpha, \nu] = \frac{Y_i + \nu}{E_i + \alpha} \quad (2)$$

Clayton and Kaldor [10] provide an iterative algorithm to obtain MLEs  $\hat{\alpha}$  and  $\hat{\nu}$ , which, when substituted into (2), yield closed form empirical Bayes estimates of  $\exp(\mu_i)$ ,  $i = 1, \dots, I$ . The form of (2) indicates the compromise between region-specific data ( $Y_i$  and  $E_i$ ), and overall population parameters ( $\alpha$  and  $\nu$ ).

Sometimes the data also include covariate information for region  $i$ , denoted by the vector  $x_i$ . In such cases, Clayton and Kaldor [10] propose a log normal second stage (*see Lognormal distribution*) as an alternative to the computationally convenient gamma prior, and consider the addition of random effects resulting in posterior estimates compromising between a region’s SMR and that of its spatial neighbors using a prior based on conditional autoregressions [3]. Besag et al. [5] expand the approach to include influence of both the overall disease rate (in the entire study area) and that of the spatial neighbors within a fully Bayesian setting. In this case, one replaces  $\mu_i$  by a linear combination of covariate effects and **random effects**, i.e.

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + v_i \quad (3)$$

where  $u_i$  and  $v_i$  denote random effects (intercepts) measuring spatial similarity and excess heterogeneity, respectively.

## 2 Hierarchical models for disease mapping

One typically assumes independence between  $\mathbf{u}$  and  $\mathbf{v}$ , and models excess heterogeneity (i.e. overdispersion, extra-Poisson variation) through a set of exchangeable priors for the  $v_i$ s, e.g.

$$v_i \stackrel{\text{ind}}{\sim} N\left(0, \frac{1}{\tau}\right), \quad i = 1, \dots, I. \quad (4)$$

To model spatial similarity in residuals, Clayton and Kaldor [10] and Besag et al. [5] assign an *intrinsic autoregressive* structure for the set of  $u_i$ s. Specifically, we define the prior distribution of each  $u_i$  conditional on the other  $u_j$ ,  $j \neq i$ , as

$$u_i | u_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{1}{\lambda \sum_{j \neq i} w_{ij}}\right), \quad i = 1, \dots, I \quad (5)$$

where the  $w_{ij}$ s denote weights defining which regions  $j$  are neighbors to region  $i$  (by convention  $w_{ii} = 0$ , for all  $i$ ), and  $\lambda$  denotes a hyperparameter controlling how similar  $u_i$  is to its neighbors. Typical applications consider adjacency-based weights where  $w_{ij} = 1$  if region  $j$  is adjacent to region  $i$ ,  $w_{ij} = 0$  otherwise, although other options appear in the literature (e.g. [7]).

Several features of the set of spatially autoregressive priors merit attention. First, Besag [3] shows that the collection of conditional distributions uniquely defines a corresponding multivariate normal joint distribution. However, the choice of adjacency weights leads to a singular precision matrix in the joint distribution, so that the spatial similarity implied by the conditional distributions does not translate directly into a model of spatial correlation [4]. Secondly, such priors are improper, since they only define contrasts between pairs of  $u_i$ s, but the inclusion of any informative data (through the likelihood function) results in a proper posterior (see [6, p. 11]). Thirdly, in order to allow identifiability of an intercept in  $\mathbf{x}_i^T \boldsymbol{\beta}$ , one often adds a constraint  $\sum_{i=1}^I u_i = 0$ . Besag [3], Besag and Kooperberg [4], Besag et al. [5], and Cressie [12, pp. 407–408, 410–423], provide detailed discussion of conditional autoregressive structures.

One completes the hierarchical model by defining vague priors for the regression parameters  $\boldsymbol{\beta}$ , and proper hyperprior distributions for hyperparameters  $\tau$  and  $\lambda$ . In practice, conjugate inverse gamma distributions are popular, and Ghosh et al. [15] and

Sun et al. [29] discuss restrictions on parameters for these hyperpriors to ensure posterior propriety.

The model is clearly overparameterized by including both  $\mathbf{u}$  and  $\mathbf{v}$  (two random intercepts for each region) so the likelihood will only identify their sum ( $u_i + v_i$ ) for each region, although the nature of the prior distributions allows posterior identifiability ([8], p. 308). A related research question involves determining a ‘fair’ assignment of prior variability for  $\tau$  and  $\lambda$  to avoid prior overemphasis on the role of the global or local rates, a question complicated by the marginal nature of  $\tau$  and the conditional nature of  $\lambda$  [1, 7, 13].

Inference proceeds via **Markov chain Monte Carlo** algorithms providing the analyst with sample-based posterior SMRs, counts, and rates. In addition, Conlon and Louis [11] and Stern and Cressie [27, 28] consider inference for ranks and extremes of regional rates based on related models.

Recently, several authors have provided spatio-temporal extensions to the approaches outlined above, allowing temporally evolving spatial structures [2, 17, 20, 21, 30, 32, 33] (see **Hazardous waste site**). Knorr-Held and Besag [21] in particular provide an insightful discussion of the interpretation of random effects  $\mathbf{u}$  and  $\mathbf{v}$  as residual patterns in the data, after accounting for standardization and covariate effects. These random effects often act as surrogates for unmeasured (perhaps unmeasurable) covariates missing from  $\mu_i$  (the linear collection of covariates in region  $i$ ).

Though popular, the hierarchical models outlined above are not the only solutions to the inference problems associated with disease mapping; see [14, 16, 18, 22, 25], and [26]. See also the discussion by Best et al. [7] for inferential problems, related topics, and alternative approaches.

### References

- [1] Bernardinelli, L., Clayton, D. & Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors?, *Statistics in Medicine* **14**, 2411–2431.
- [2] Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995). Bayesian analysis of space–time variation in risk, *Statistics in Medicine* **14**, 2433–2443.
- [3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 192–236.
- [4] Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions, *Biometrika* **82**, 4, 733–746.

- [5] Besag, J., York, J.C. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**, 1–59.
- [6] Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**, 3–66.
- [7] Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. & Conlon, E.M. (1999). Bayesian models for spatially correlated disease and exposure data, in *Bayesian Statistics*, Vol. 6, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds, Oxford University Press, Oxford, pp. 131–156.
- [8] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London.
- [9] Clayton, D.G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds, Oxford University Press, Oxford, pp. 205–220.
- [10] Clayton, D.G. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–681.
- [11] Conlon, E.M. & Louis, T.A. (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods, in *Disease Mapping and Risk Assessment for Public Health*, A.B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel & R. Bertollini, eds, Wiley, Chichester.
- [12] Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition*, Wiley, New York.
- [13] Eberly, L.E. & Carlin, B.P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models, *Statistics in Medicine* **19**, 2279–2294.
- [14] Gelman, A. & Price, P.N. (1999). All maps of parameter estimates are misleading, *Statistics in Medicine* **18**, 3221–3234.
- [15] Ghosh, M., Natarajan, K., Waller, L.A. & Kim, D. (1999). Hierarchical GLMs for the analysis of spatial data: an application to disease mapping, *Journal of Statistical Planning and Inference* **75**, 305–318.
- [16] Heagerty, P.J. & Lele, S.R. (1998). A composite likelihood approach to binary spatial data, *Journal of the American Statistical Association* **93**, 1099–1111.
- [17] Heisterkamp, S.H., Doornbos, G. & Nagelkerke, N.J.D. (2000). Assessing the impact of environmental pollution sources using space–time models, *Statistics in Medicine* **19**, 2569–2578.
- [18] Kafadar, K. (1999). Simultaneous smoothing and adjusting mortality rates in US counties: melanoma in white females and males, *Statistics in Medicine* **18**, 3167–3188.
- [19] Kaldor, J. & Clayton, D. (1989). Role of advanced statistical techniques in cancer mapping, in *Cancer Mapping (Recent Results in Cancer Research)*, P. Boyle, C.S. Muir & E. Grundmann, eds, Springer-Verlag, Berlin.
- [20] Knorr-Held, L. (2000). Bayesian modelling of inseparable space–time variation in disease risk, *Statistics in Medicine* **19**, 2555–2567.
- [21] Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine* **17**, 2045–2060.
- [22] Lawson, A.B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F. & Bertollini, R., eds (1999). *Disease Mapping and Risk Assessment for Public Health*, Wiley, Chichester.
- [23] Manton, K.G., Woodbury, M.A. & Stallard, E. (1981). A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties, *Biometrics* **37**, 259–269.
- [24] Mollié, A. (1996). Bayesian mapping of disease, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds, Chapman & Hall, London, CRC Press/Boca Raton, pp. 360–379.
- [25] Mungiole, M., Pickle, L.W. & Hansen-Simonson, K. (1999). Application of a weighted head-banging algorithm to mortality data maps, *Statistics in Medicine* **18**, 3201–3209.
- [26] *Statistics in Medicine* (2000). (Issue devoted to Disease Mapping with a Focus on Evaluation), **19**, 2201–2594.
- [27] Stern, H. & Cressie, N. (1999). Inference for extremes in disease mapping, in *Disease Mapping and Risk Assessment for Public Health*, A.B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel & R. Bertollini, eds, Wiley, Chichester.
- [28] Stern, H. & Cressie, N. (2000). Posterior predictive model checks for disease mapping models, *Statistics in Medicine* **19**, 2377–2397.
- [29] Sun, D., Tsutakawa, R.K. & Speckman, P.L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions, *Biometrika* **86**, 341–350.
- [30] Sun, D., Tsutakawa, R.K., Kim, H. & He, Z. (2000). Spatio-temporal interaction with disease mapping, *Statistics in Medicine* **19**, 2015–2035.
- [31] Tsutakawa, R.K. (1985). Estimation of cancer mortality rates: a Bayesian analysis of small frequencies, *Biometrics* **41**, 69–79.
- [32] Waller, L.A., Carlin, B.P., Xia, H. & Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical Association* **92**, 607–617.
- [33] Xia, H. & Carlin, B.P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality, *Statistics in Medicine* **17**, 2025–2043.

(See also **Epidemic models; Hierarchical Bayes; Morbidity and mortality; Small area estimation; Spatial statistics in environmental epidemiology**)

LANCE A. WALLER