Local Likelihood Bayesian Cluster Modeling for small area health data

Andrew Lawson

Arnold School of Public Health

University of South Carolina

Local Likelihood Bayesian Cluster Modelling for Small Area Health Data

Andrew Lawson
Arnold School of Public Health
University of South Carolina

DIMACS 2006

Talk Outline

- Background
- Model Development
- Parameters and Priors
- Posterior Distribution
- Profile likelihood variant
- Posterior Sampling
- Cluster Assessment
- Examples and Simulations
- Sensitivity to prior specification
- Extensions and Conclusions

Background

- There are many advantages to the adoption of a model based approach to clustering in small area health studies.
- The flexibility of model formulation, with the ability to add or remove covariates within an analysis and the ability to specify the form of clusters to be examined adds significantly to the desirability of this approach.
- The ability of modelling approaches to formulate clustering within a likelihood and to extend this to Bayesian methodology via the use of suitable prior distributions supports its wider adoption as a clustering paradigm.

- Such parametric modelling has many advantages but also has the disadvantage that a parametric model for clusters may be too rigid for some applications. For example, in spatial epidemiology, clustering of small area health data may take a variety of forms and usually the main concern in public heath is the assessment of where clusters occur and their significance.

Local Likelihood

- Local Likelihood: introduced by Tibshirani and Hastie (1987)
- Recent reference: Kauermann and Opsomer (2003) Local Likelihood estimation in generalised additive models *Scandinavian Journal of Statistics*, 30,317-337
- The advantage of a local likelihood approach to clustering is that it is relatively straightforward to incorporate covariate information in the analysis, to utilise a variety of prior distributions for the covariance effects which are excluded from the likelihood, and to be able to isolate regions of ‘significantly’ elevated or reduced risk.

- What is local likelihood for clustering?
  - Define ‘local’ (spatial) neighbourhoods within which a likelihood can be defined
  - Assign a parameter to the neighbourhood to define size and overall form of the area of excess risk (cluster)
  - Local estimates of excess risk are available if required
  - Because local neighbourhoods can overlap, the risk associated with these areas may be correlated

Model Development

- The general development is given for case event data.

Case Event Data

- A study region A is defined within which n cases of the disease of interest are observed. The specific type of disease is not of particular concern in this development. The disease will have been observed within a temporal window, often a period of years.
- The case event locations are $\{x_i\}, i = 1, \ldots, n$.

- We assume that the cases are governed by a first order intensity of the general form:

$$\lambda(x_i) = \rho\lambda_0(x_i)\theta(x_i|\delta_{x_i}),$$

  where $\theta(x|\delta_x)$ is the relative risk at the location x and $\lambda_0(x)$ is a baseline intensity which is usually a function of the spatial distribution of the population of interest.
- Relative risk modelled in relation to an unobserved process (δ_x).

- Relative risk is defined to be a function of the 'local' concentration of cases.
- Support for the local density of cases comes from the case locations themselves and their individual contribution to the likelihood.
- A *conditional logistic likelihood* is appropriate. In that formulation a control disease distribution is also available and the joint distribution of the cases and controls is modelled.
- The probability of a case at x_i is just $\rho\theta(x_i)/(1 + \rho\theta(x_i))$. In this situation the background intensity $\lambda_0(x_i)$ is conditioned out of the likelihood.
- $\theta(x|\delta_x)$ depends on a neighbourhood of location x. The neighbourhood is controlled by δ_x, which in turn is controlled by a prior distribution.

- δ_x defines the local scale of clustering.
- In more general settings, δ_x could be regarded as a random field and modelled accordingly.
- Here we simply assume that any realization of the field is derived from a spatially correlated prior distribution.
- As δ_{x_i} controls the size of the lasso, then it is possible to estimate δ_{x_i}, provided some suitable prior distributional assumptions are made.

- Note δ_{x_i} could include events which are also included in other $\delta_x s$. Hence correlation between the $\delta_{x_i} s$ must be considered.
- Advantages of this specification are that the cluster form is not restricted and low and high intensity clusters are admitted. Here, high intensity areas are the only focus.

- Specify the probability of a case at x :

$$\lambda(x)/(\lambda_0(x) + \lambda(x)).$$

- The local likelihood within an area δ_x is defined as:

$$\left(\frac{\rho\theta}{1 + \rho\theta} \right)^{n_{\delta x}} \left(\frac{1}{1 + \rho\theta} \right)^{e_{\delta x}}.$$

- Here we define $\theta \equiv \theta(x|\delta_x)$, the number of cases within δ_x as n_{δ_x}, and the number of controls as e_{δ_x}.
- Hence the full local likelihood is defined as :

$$L_l = \prod_{l \in \textit{ cases and controls}} \left\{ \frac{\rho\theta_l}{1 + \rho\theta_l} \right\}^{n_{\delta_l}} \left\{ \frac{1}{1 + \rho\theta_l} \right\}^{e_{\delta_l}}.$$

- One interesting result of this formulation is that the saturated maximum local likelihood estimate of θ is just

$$\widehat{\theta} = \frac{1}{\rho} \frac{n_{\delta_x}}{e_{\delta_x}},$$

which is just the local ratio of cases to controls scaled by the overall rate.

Count Data

- An alternative scenario, that arises more frequently in small area health studies, is that counts of the disease of interest arise within arbitrary small administrative areas. In addition, expected rates for the disease are available within the same areas.
- Assume that counts of cases and expected rates are available within p arbitrary regions.
- The observed counts in the regions are $\{n_i\}$, $i = 1, .., p.$ The expected rates are $\{e_i\}$, $i = 1, .., p.$
- The usual assumption is made: observed counts are Poisson distributed with expectation: $E(n_i) = e_i.\theta_i.$

- It is assumed that the expected rates have absorbed the overall rate and ρ is not required.
- In the local likelihood definition, the contribution to the likelihood for the i th region is locally Poisson, and, bar a constant, given by:

$$(e_{\delta_i}\theta_{\delta_i})^{n_{\delta_i}} \exp(-e_{\delta_i}\theta_{\delta_i})/n_{\delta_i}!.$$

- The resulting likelihood is just:

$$L_l = \prod_{i=1}^{p} (e_{\delta_i}\theta_{\delta_i})^{n_{\delta_i}} \exp(-e_{\delta_i}\theta_{\delta_i})/n_{\delta_i}!.$$

- Here n_{δ_i} is a count of cases within the neighbourhood δ_i, likewise e_{δ_i} is the total expected rate within δ_i.
- Note that in this case the saturated maximum likelihood estimate of θ_{δ_i} is also given by

$$\widehat{\theta}_{\delta_i} = \frac{n_{\delta_i}}{e_{\delta_i}}.$$

This is just the standardised mortality ratio for the area defined by δ_i.

- Note that these local likelihood estimates based on δ_i have two important limiting values. When $\delta_i \to \infty$ then $\widehat{\theta}_{\delta_i} \to \rho^*$ an overall constant rate. When $\delta_i \to 0$ then $\widehat{\theta}_{\delta_i} \to \widehat{\theta}_i = n_i/e_i$ the point-wise standardised mortality ratio (*smr*).

Specification of the δ_x surface and prior distributions

- The δ_x surface is assumed to be continuous and that values on the surface will be correlated if locations are close. It is natural therefore to consider a probability model which allows for positive spatial correlation.

The δ_x surface prior distribution

- The δ_x surface prior distribution could be specified in various ways.
- One approach is to consider the separation distance (d_{ij} say) between any two arbitrary locations (i, j) and the intersection of the two balls with radius $\delta_{x_i,}$and δ_{x_j}.

- Then it is possible to consider a parametric covariance function with i,j th element $c(i,j) \propto f(d_{ij}, \delta_{x_i,}, \delta_{x_j}; \theta)$. This could be structured to allow for changes in covariance when $d_{ij} > \delta_{x_i,} + \delta_{x_j}$ or $d_{ij} < \delta_{x_i,} + \delta_{x_j}$.
- A simpler approach has been adopted where the smoothness of the δ_x surface is controlled by a intrinsic Gaussian prior distribution and the strength of spatial correlation in the surface is controlled by a parameter β_δ (Kunsch(1987)).
- This is a singular distribution which is simple to implement within a posterior sampling scheme.

- Specifically we assume that:

$$[\delta_{x_1}, \ldots\ldots \delta_{x_n}]$$

$$\propto (\beta_\delta)^{-n/2} \exp\left\{ -\frac{1}{2\beta_\delta} \sum_{i<j} \sum_{x_j \in B(\delta_{x_i})} \{\delta_{x_i} - \delta_{x_j}\}^2 \right\},$$

where $B(\delta_x)$ is the ball of radius δ_x centred at x.

- The prior distribution $[\theta|\delta_x]$ describes the dependence of the relative risks on δ_x. In general this dependence is likely to be weak apriori across ranges of δ_x. Hence it would be reasonable to assume a uniform prior distribution for $[\theta|\delta_x]$.
- In addition, at the next level of the hierarchy, a hyperprior for β_δ must be assumed ($[\beta_\delta]$). This parameter is strictly positive and a conventional inverse gamma prior is assumed: $IG(\alpha, \beta)$. The choice of α and β will be discussed in the examples.
- The remaining parameter which must be given a prior distribution is the overall rate ρ. As this parameter is strictly positive it is also given an inverse Gamma distribution. Denote this prior distribution as $[\rho]$.

The Posterior Distribution

- Define the local likelihood of the data as $L(data|\boldsymbol{\psi})$ or $L(data|\boldsymbol{\psi}')$ for case events or count data respectively, and

$$\boldsymbol{\psi} = \left\{ \begin{array}{c} \rho \\ \beta_\delta \\ \{\delta_x\} \\ \{\theta_\delta\} \end{array} \right\}$$

and $\boldsymbol{\psi}'$ is the count data parameter vector minus ρ.

- The appropriate posterior distribution for case event data is then given by

$$P(\boldsymbol{\psi}) \propto L(data|\boldsymbol{\psi}).[\theta|\delta_x].[\delta_x|\beta_\delta].[\beta_\delta].[\rho],$$

where $[\delta_x|\beta_\delta]$ is the joint prior distribution for $\{\delta_x\}$, and $[\beta_\delta]$, and $[\rho]$ are prior distributions for β_δ and ρ respectively.

- For count data $\boldsymbol{\psi}$ is replaced by $\boldsymbol{\psi}'$ and $[\rho]$ disappears from the right hand side.
- The full model prior distributions are defined as:

$$[\theta|\delta_x] \sim U(a,b),$$
$$[\delta_x|\beta_\delta] \sim IG_a(\overline{\delta}_x,\beta_\delta),$$
$$[\beta_\delta] \sim IG(\tau_\beta,\epsilon_\beta),$$
$$[\rho] \sim IG(\tau_\rho,\epsilon_\rho),$$

where IG_a and IG denote the intrinsic Gaussian and inverse Gamma distributions.

Profile likelihood Variants

- It is interesting to note that the saturated ML estimators could play a role in a further likelihood variant which has appeal due to the intuitive nature of the estimators defined and simple interpretation.
- Consider a profile likelihood representation of the models whereby θ is replaced by an estimator. In this case a natural estimator would be the saturated ML estimator based on δ_x:

case event data:

$$\widehat{\theta}_s = \frac{1}{\rho}\frac{n_{\delta s}}{e_{\delta_s}},$$

where s here denotes a case event or a control event,

count data :

$$\widehat{\theta}_i = \frac{n_{\delta_i}}{e_{\delta_i}}.$$

- In both cases, a natural profile likelihood representation leads to, for

case event data:

$$L_c = \prod_{l \in cases} \frac{\rho\widehat{\theta_l}}{1+\rho\widehat{\theta_l}} . \prod_{q \in controls} \frac{1}{1+\rho\widehat{\theta_q}}$$
$$= \prod_{l \in cases} \frac{n_{\delta l}}{e_{\delta l}+n_{\delta l}} . \prod_{q \in controls} \frac{e_{\delta q}}{e_{\delta q}+n_{\delta q}} .$$

and for count data:

$$L_c = \prod_{i=1}^{p} (e_i.\widehat{\theta_i})^{n_{\delta_i}} \exp(-e_i.\widehat{\theta_i})/n_{\delta_i}!.$$

Clearly, for the count data case, if $\delta_i = 0$ then the resulting estimator is just $\widehat{\theta}_i = \frac{n_i}{e_i}$, the standardised incidence ratio for the i th region.

Posterior Sampling

- The hierarchical Bayesian model could be sampled in a conventional way using a Gibbs update where conditional distributions are available or via the more general Metropolis Hastings (*mh*) update.
- Because not all full conditionals are available, an *mh* update has been used for all the parameters in the examples, except for the $\{\theta\}$. These parameters can be sampled efficiently within a conditional Gibbs step, and so we have used a hybrid Gibbs-Metropolis algorithm for all updating.
- The prior distribution $[\theta|\delta_x]$ takes different forms depending on whether the profile likelihood or full likelihood is used. In the profile likelihood, θ is explicitly defined by δ_x and hence it does not require a prior distribution.

- In the full likelihood a parametric definition must be adopted. However it is unlikely that θ would be assumed to have strong dependence on $\{\delta_x\}$ and so an uninformative prior distribution would usually be assumed.
- In fact, given the local likelihood definition, it is straightforward to sample, in the count data case, θ from:

$$[e_{\delta_i}\theta_i|\delta_i] \sim G(n_{\delta_i}+1,1).$$

- In the case event example, the probability of a case is just a binomial probability and with a uniform prior distribution for θ, this probability can be sampled from a Beta $(n_{\delta_i}+1, e_{\delta_i}+1)$ and θ_i obtained by back transformation.
- Proposal distributions for the *mh* updates were assumed to be normal with the mean being the current parameter value and variance tuned to yield a reasonable rejection rate (see e.g. Robert and Casella(1999)).

- Multiple start points were examined as well as different iteration length trials.
- It was apparent that convergence was achieved in most cases by 2000 iterations. Convergence was assessed via both graphical diagnostic methods and monitoring of the logarithm of the posterior distribution itself (Yu and Mykland(1998)).

Cluster Assessment based on δ_x or $\theta(\delta_x)$

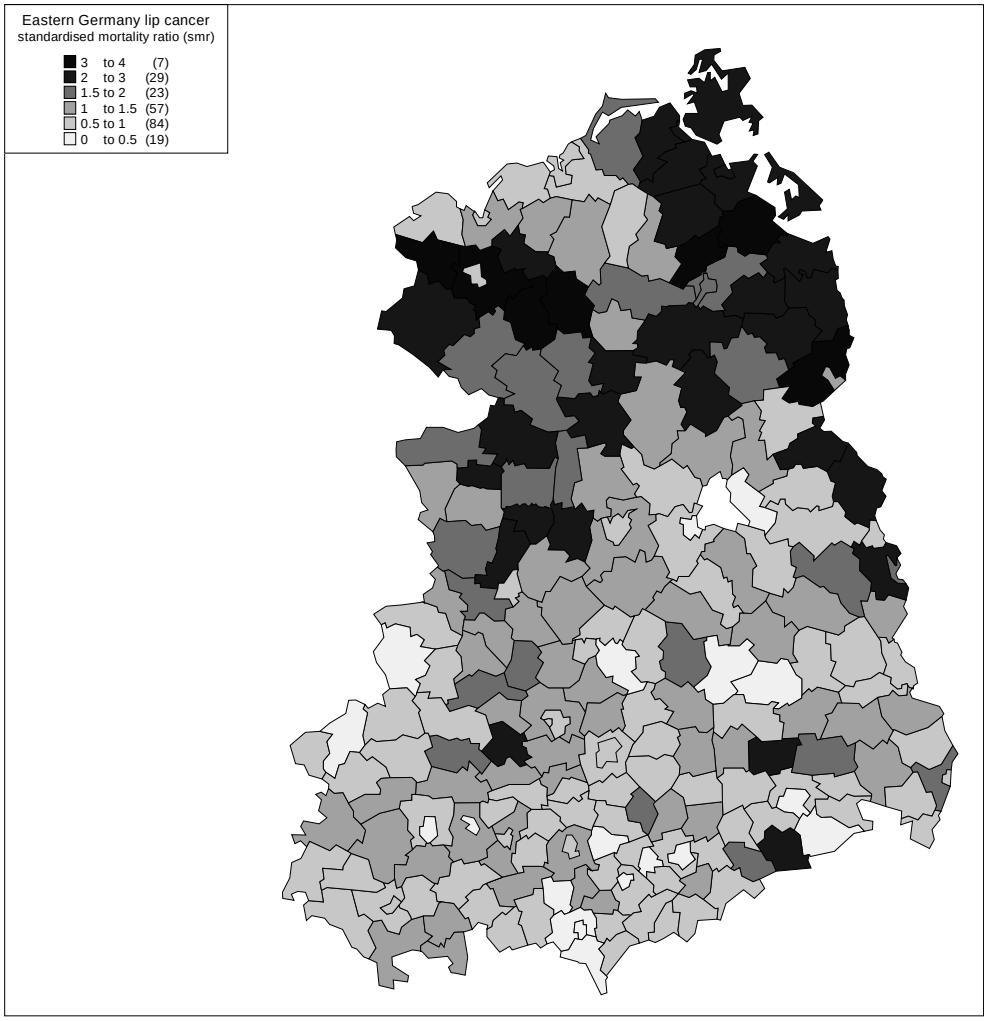
- δ_x represents a form of cluster variance or scale
- One derived measure could be $(\max(\delta_x) - \delta_x)/(\max(\delta_x) - \min(\delta_x))$: a relative clustering scale index.
- $E\{\theta_i\}$, the estimated posterior expected relative risk, can be used directly to assess excess risk.
- Cluster detection capability: examine the estimate of θ in each region (θ_i) and estimate $P(\theta > a)$ from the number of estimates in the final sample which exceed this limit. In the case of count or case event data $a = 1$.

- Contour maps of the resulting probability surface can be examined for areas of excess risk.
- A nominal level of $P(\theta > a) = 0.95$ or 0.975 could be used.

Examples

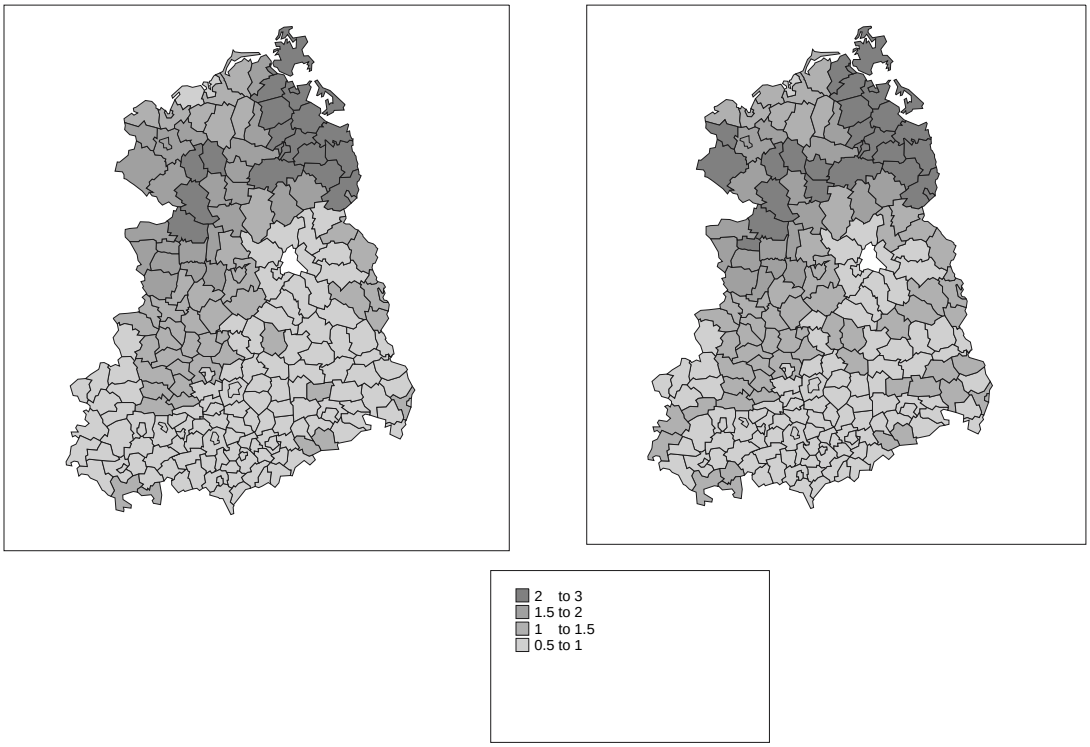
East German Lip Cancer

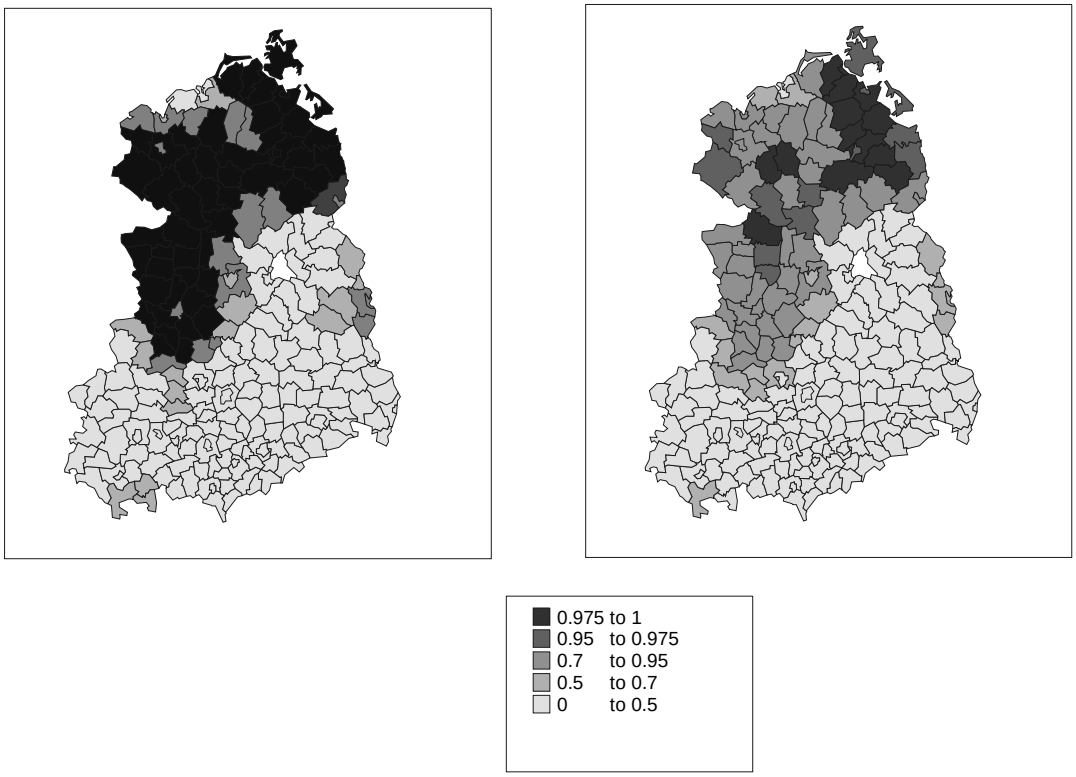
- The data has been analysed previously by Lawson et al (1999), and consists of mortality counts within 219 landkriese in the former East Germany for the period 1980-1989. The expected rates for lip cancer were also available, computed from nine age and two gender groups based on the national East Germany rates for the period. Lip cancer mortality is relatively rare and is strongly related to exposure to UV radiation.



- Evidence of a north-south gradient and possibly localized clustering.

- the δ_x sampler was run using the profile likelihood and the full local likelihood with the CAR prior distribution for $\delta_x$with an $IG(3, 0.01)$ for β_δ. The sensitivity to this prior specification is discussed in later.
- Convergence for this sampler is relatively fast. Based on checks of multiple start point chains and posterior checks, convergence was achieved within 2000 iterations.


2 to 3
1.5 to 2
1 to 1.5
0.5 to 1


0.975 to 1
0.95 to 0.975
0.7 to 0.95
0.5 to 0.7
0 to 0.5

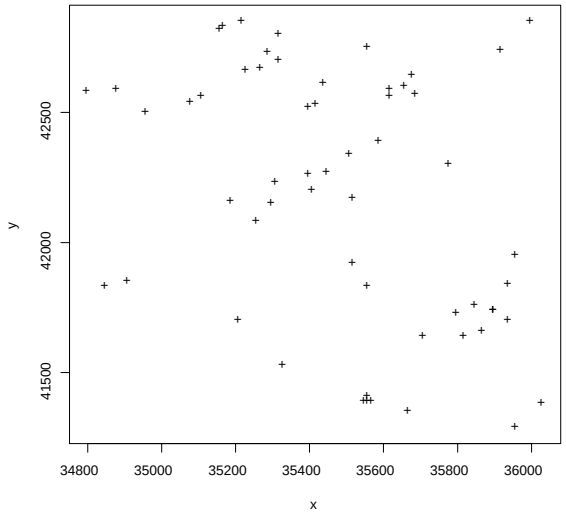
DIMACS 2006

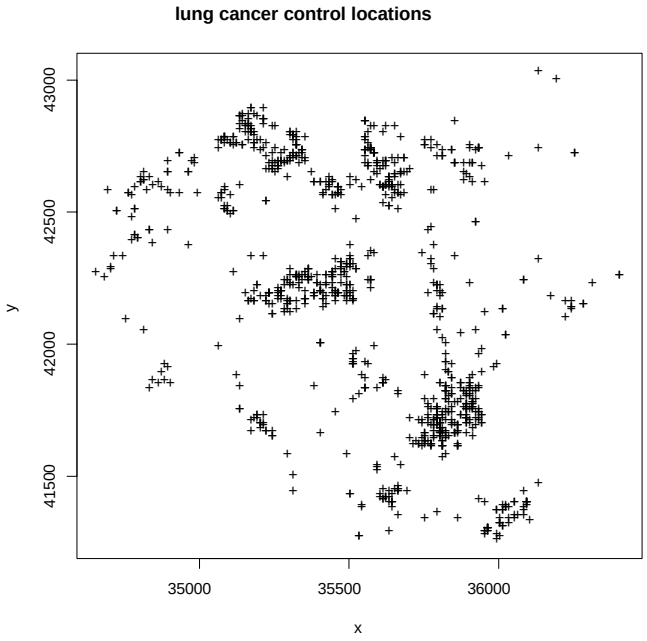
- Features:
  - The range of relative risk estimates is reduced from that of the smr. The profile likelihood range is 0.59 to 2.81. The full likelihood range is 0.61- 2.95.
  - The $Pr(\theta > 1)$ surfaces show evidence for areas with large probabilities corresponding to the northern elevated risk areas. The profile likelihood seems to yield false positives
  - The posterior expected values for β_δ was found to be 0.3336 (sample sd: 0.0187).

Lancashire Larynx Cancer

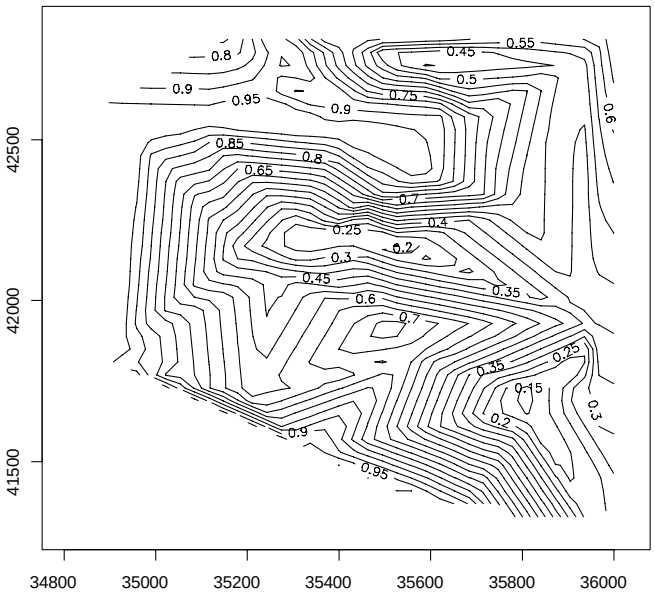
- Case event realization of larynx cancer death certificate addresses for the period of 1974-1983 in Lancashire UK.
- First presented by Diggle(1990) in relation to the analysis of risk around a putative source of health hazard (an incinerator).
- The dataset consists of 58 deaths due to larynx cancer. As a control disease for this example the 978 death certificate addresses for lung cancer in the same period and study area were used.
- Lung cancer is not a particularly good choice of control for larynx cancer in a putative air pollution hazard study, as it too can be influenced by confounding inhalatory insults.
- Conditional logistic likelihood form used: does not require the estimation of $\lambda_0(x)$.

- Here we report only the simpler *profile* local likelihood:
  - A CAR prior distribution for δ_x and $IG(3, 0.01)$ for β_δ and and $IG(3, 0.01)$ distribution for ρ. Both prior distributions for β_δ and ρ are reasonably well dispersed.
  - Convergence for this sampler was relatively fast and 2000 iterations was found to be adequate.




lung cancer control locations
y
x
43000
42500
42000
41500
35000
35500
36000


42500
42000
41500
34800
35000
35200
35400
35600
35800
36000



- Overall this distribution shows some evidence of elevated risk in the northwest and south east and a large excess in the north east. This is consistent with nonparametric extraction maps for the disease.
- The Pr($\theta > 1$) surface demonstrates that the area of high probability lie in these three regions. It is noteworthy that the north west and south east have areas of Pr($\theta > 1$) >0.95.
- The posterior expected values for ρ and β_δ were found to be 0.0587 (sample sd: 0.0090) and 4.6613 (sample sd: 0.3395) respectively.

DIMACS 2006

Sensitivity to prior distributional specification

1. Sensitivity of the CAR prior was examined by using an exchangeable Gamma prior for the δ_x. This was examined for the count data example only.
2. The range of relative risks estimated was slightly increased but the relative ranking of the areas and their probabilities was unaffected.
3. Essentially the CAR prior provides increased smoothing of the risks, as might be expected. However, it is also reassuring that using an uncorrelated prior appears to provide reasonable results for risk estimates also.

4. We have also examined the use of a range of *IG* parameters for the β_δ and ρ parameters and these were found to be relatively insensitive to changes. Some increased variability of ρ values can lead to increases in variability of θ estimates but this was found to be relatively small, by comparison to the differences found between CAR and exchangeable Gamma priors.

Extensions

- An important extension to this approach is the inclusion of covariates. The posterior distribution can be written with a likelihood dependent on a linear predictor. In the count data example, this would be:

$$L_c = \prod_{i=1}^{p} (e_{\delta_i} m_i \theta_{\delta_i})^{n_{\delta_i}} \exp(-e_{\delta_i} m_i \theta_{\delta_i}) / n_{\delta_i}!.$$

- where, for example, $m_i = \exp(d_i \beta)$, and d_i is the i th row of the $p \times q$ covariate design matrix and β is a column vector of q parameters.
- Whereas for the conditional logistic model $\rho \theta_{\delta_x}$ would be replaced by $\rho m \theta_{\delta_x}$, where we assume spatially continuous covariates.

- **Edge effects**: the sampled δ_x values have been allowed to overlap the boundary of the study region. In areas close to the boundary it is possible that there could be bias in the local θ estimator.
  - One approach to this problem is to conditionally simulate events or counts into exterior regions (if available) and treat these as parameters which can be updated within the algorithm. This is essentially the use of an external guard area within which data is imputed.
  - This form of data augmentation is well suited to iterative sampling algorithms (Tanner(1996)).

Discussion and Conclusions

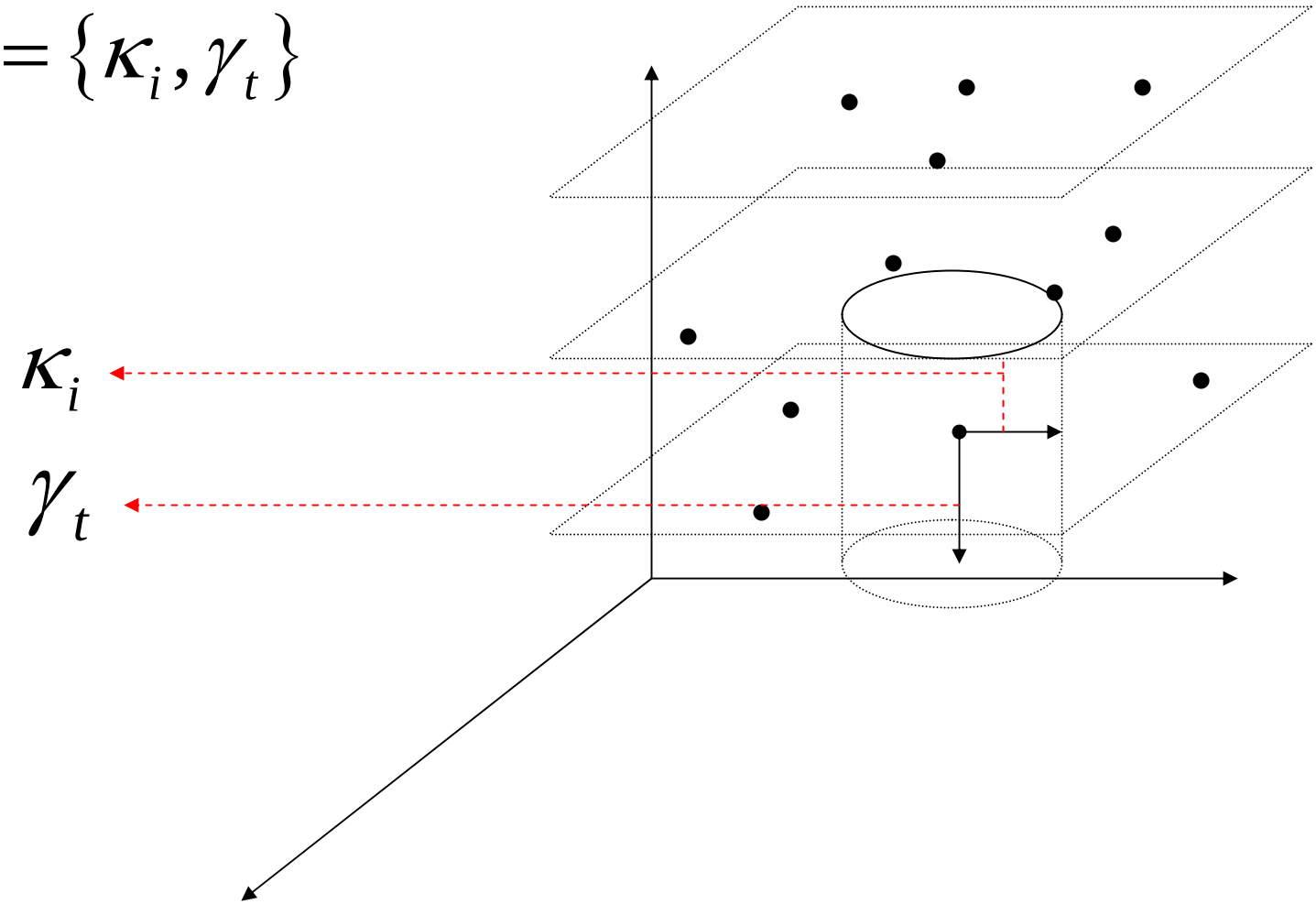
- General *specific* clustering is difficult to model without recourse to complex models and usually requires the modeling of specific localised behaviour.
- Local likelihood Bayesian modelling, is one fruitful avenue to pursue. The weak correlation within clusters leads to a relatively simple model formulation that can be sampled easily and leads to reasonable reconstructions of relative risk.

- Many fruitful applications of this methodology are possible not only in static clustering applications but also in the application of clustering methodology to *surveillance* problems. The need for faster methods for detecting clustering in that area of application could provide an important future application area.

Space-time Extension

- Space-time cylinder:

$$\delta_{it} = \{\kappa_i, \gamma_t\}$$



Space-Time Extension

- Count data: y_{it} and e_{it}; $i = 1, \ldots, n, \quad t = 1, \ldots, T$

$$y_{\delta_{it}} = \sum_{\substack{d_{ii'}^{(c)} \le \kappa_i \\ d_{tt'}^{(y)} \le \gamma_t}} y_{i't'}, \qquad e_{\delta_{it}} = \sum_{\substack{d_{ii'}^{(c)} \le \kappa_i \\ d_{tt'}^{(t)} \le \gamma_t}} e_{i't'}$$

$$i' = 1, \ldots, n, \quad t' = 1, \ldots, T$$

Space-Time Extension

- Model:

$$y_{\delta_{it}} \mid e_{\delta_{it}}, \theta_{it} \sim \text{Poisson}\left(e_{\delta_{it}} \theta_{it}\right)$$

$$\nu_{it} = \ln\left(\theta_{it}\right) = \kappa_i + \gamma_t + \varepsilon_i + \xi_t$$

Space-Time Extension

- Priors:

$$\kappa_i \mid \{\kappa_{-i}\}, \sigma_\kappa \sim \mathrm{N}\left(\bar{\kappa}_i, \frac{\sigma_\kappa}{d_i}\right)$$

$$\lambda_t \mid \gamma_{t-1}, \sigma_\gamma^2 \sim \mathrm{N}\left(\gamma_{t-1}, \sigma_\gamma^2\right)$$

$$\varepsilon_i \mid \sigma_\varepsilon^2 \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right)$$

$$\xi_t \mid \sigma_\xi^2 \sim \mathrm{N}\left(0, \sigma_\xi^2\right)$$

References

Hossain, M. and A. B. Lawson (2005). Local likelihood disease clustering : Development and evaluation. *Environmental and Ecological Statistics 12*, 259–273.

Hossain, M. and A. B. Lawson (2006). Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. *Statistics in Medicine 25*, 771–786.

Lawson, A. B. (2006a). Disease cluster detection : a critique and a bayesian proposal. *Statistics in Medicine 25*, 897–916.

Lawson, A. B. (2006b). *Statistical Methods in Spatial Epidemiology* (2 ed.). New York: Wiley.

Lawson, A. B. and D. Denison (Eds.) (2002). *Spatial Cluster Modelling*. New York: CRC press.