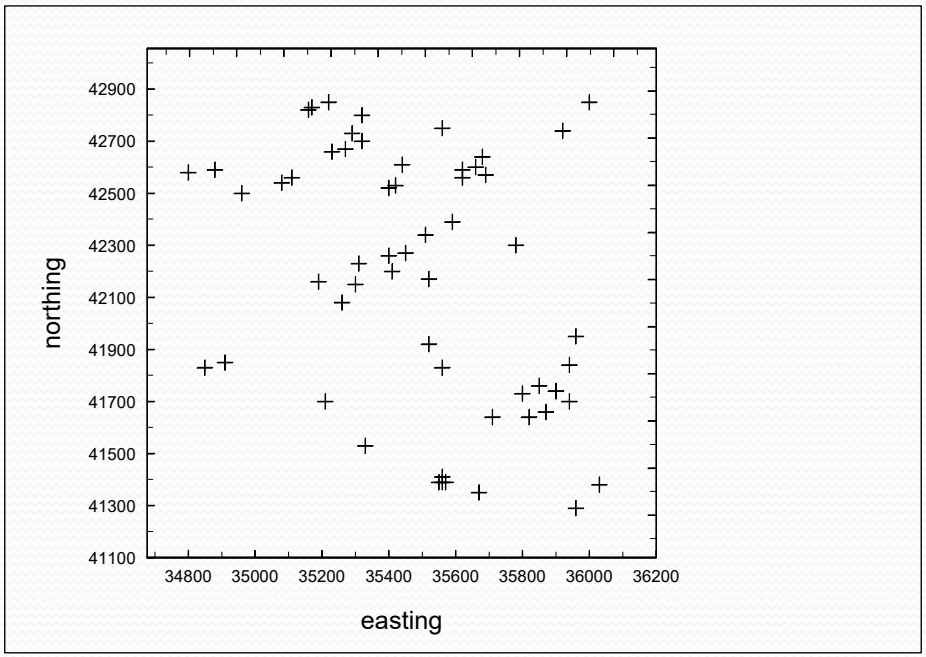
Case event data

- Count data is the commonest format found in spatial epidemiology
- However this is just an aggregation of case event data where the (residential) location of a case of disease is the primary data focus
- Often case event data is important when small spatial scales are of interest (1-10kms for example)

Example: larynx cancer in NW England



Case event notation

- Define the study area as T

  s_i : x,y coordinate pair of the i th location

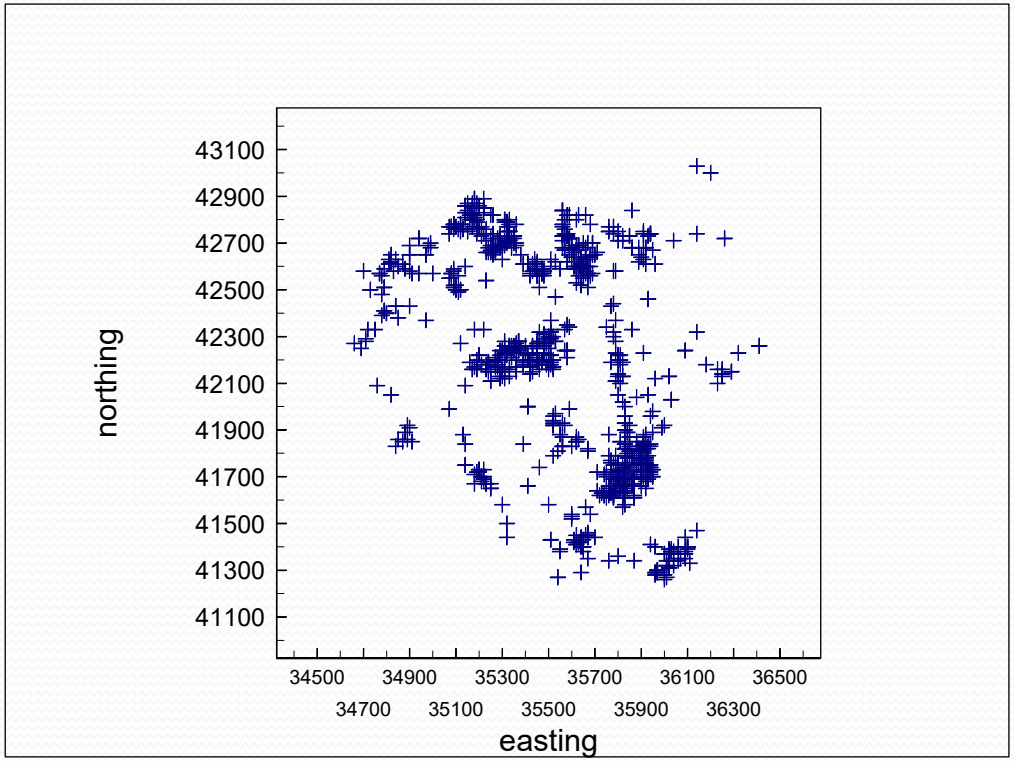
  m events in T

  $\{s_i\}, \ i = 1,....,m$

Control disease

- Usually the cases have associated with them a control disease realization
- This is used as a geographical control for the case distribution (acting like a expected count in the count data examples)

Control: lung cancer



Control notation

n control locations in T

$\{s_j\},\ j = m+1,\ m+n$

- Hence we treat the controls and cases as one vector of length $m+n$

Case event Models

- Natural models are Point processes
- Both cases and controls can be assumed to have Poisson point process (PPP) models governed by an intensity function

Case Event Direct Modeling

- Bayesian modeling of PPPs can be achieved directly (but not conveniently as an integral must be approximated)
  - Berman-Turner Approximation is useful
  - See Lawson (2006) ch 8, Section 8.4.3 and WinBUGS code in Appendix c.4.5
  - (in participant files: zeroes_PP...odc)
- We don't pursue this here

Conditional Logistic models

- Instead we use CONDITIONING to give us a simpler labeling approach
- Intensity of the case and control events is defined to be

$$control : \lambda_0(s)$$

$$case : \lambda_0(s)\lambda_1(s)$$

modeled part:

$$\lambda_1(s) = \exp(\eta(s))$$

Conditional Logistic models

Assume that the complete vector is used for a binary label so that

$$y_i = \begin{cases} 1 & \text{if } s_i \in \{s_i\},\ i = 1,\ldots,m \\ 0 & \text{otherwise} \end{cases}$$

- Hence, y_i is 1 for case and 0 for a control

Logistic spatial models

- Then:

$$y_i \sim Bern(p_i)$$

$$p_i = \frac{\lambda_1(s_i)}{1+\lambda_1(s_i)}$$

$$If \;\; \lambda_1(s_i) = \exp(x_i'\beta)$$

$$where \; x_i'\beta \text{ is a linear predictor}$$

- This is just a logistic regression formulation
- Hence as long as you have covariate information at the locations of controls and cases you can assume a conditional logistic spatial model

Logistic Spatial models

As long as $\lambda_1(s_i) = \exp(x_i'\beta)$

then $x_i'\beta$ is just a linear predictor at the site locations

$x_i'\beta$ can be individual covariates (age, gender etc)

or

spatially specific (e.g. pollution measure, distance from a source).

The linear predictor can include random effects also.

Typical example

- Location (s), distance from a pollution source (d), age (x) as variables must be available for all cases and controls

$$\eta_i = \psi_0 \exp\{\alpha_1 d_i + \alpha_2 x_i\} = \exp\{\alpha_0 + \alpha_1 d_i + \alpha_2 x_i\}$$

$d_i = || s_i - s_0 ||$ distance from source

s_0 is the source location

Addition of Random effects

- It is easy to add various types of REs
- UH can be added via an individual level zero mean Gaussian effect:V~N(0,tau)
- CH is slightly different: A CAR model cannot be simply applied here
- Can use a CAR if you can defined neighborhoods?
- Otherwise must use a full MVN geostatistical model

Spatial.exp

- For point referenced data (i. e. measures made at locations) we can specify an effect such that :

$$u_1,.....,u_m \sim MVN(\mu, C)$$

$$C : \text{covariance matrix}$$

$$C_{ij} = \text{cov}(u_i, u_j) = \tau\rho(|| s_i - s_j ||)$$

$$d_{ij} = || s_i - s_j ||$$

$$\rho(d_{ij}) = \exp(-\alpha d_{ij}^{\beta})$$

Bayesian Geostatistical models

$$\eta_i = \exp\{A_i\}$$

$$A_i = \alpha_0 + \alpha_1 d_i + \alpha_2 x_i + v_i + u_i$$

$$v_i \sim N(0, \tau_v)$$

$$\mathbf{u} \sim MVN(\mathbf{0}, C)$$

$$C_{ij} = \upsilon \exp(-\alpha d_{ij}^{\rho})$$

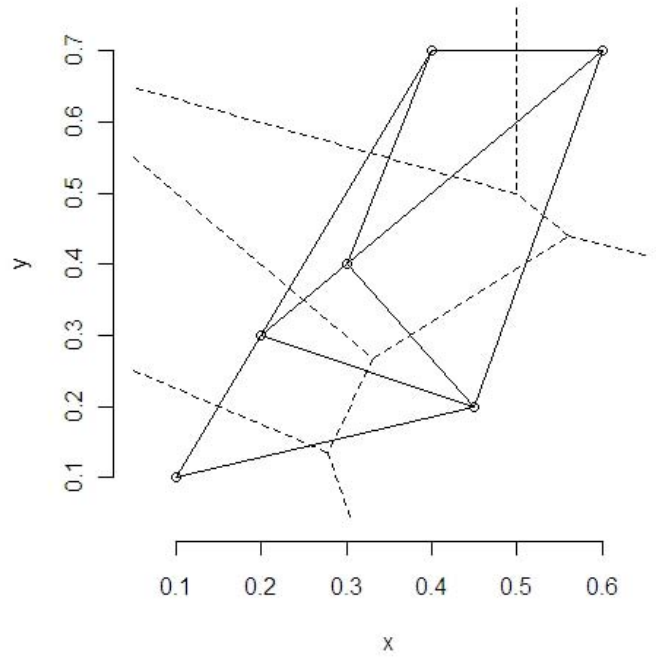
- Note: the spatial correlation effect has zero mean
- The spatial.exp model is available on WinBUGS
- Related to log Gaussian Cox processes

Alternative spatial structures

- Spatial.exp is very sensitive to sample size,
- It is very slow: inversion of NxN matrix at each iteration
- It is sensitive to structure of sampling mesh (singularities)
- Alternative: choose natural neighbors and consider intrinsic CAR? Much simpler
  - Possible via Delauney triangulation

Delauney Neighbors

X 0.1 0.2 0.4 0.45 0.6 0.3
Y 0.1 0.3 0.7 0.2 0.7 0.4
Num 2 4 3 4 3 4



Example

- Larynx and lung cancer (NW England)
- Dataset: larynx_cas_con_Idis.txt

Variables: x, y, ind, dis, age

- Code file: logistic_case_con_bern_AGE.odc
- Using Delauney neighbors to define adjacencies

Models

	DIC	pD
I D only	447.45	0.44
II D+V	439.74	41.12
III D+V+A	366.67	89.01
IV D+A	444.69	1.82
V D+V+U	447.4	5.67
VI D+V+U+A	352.94	118.10

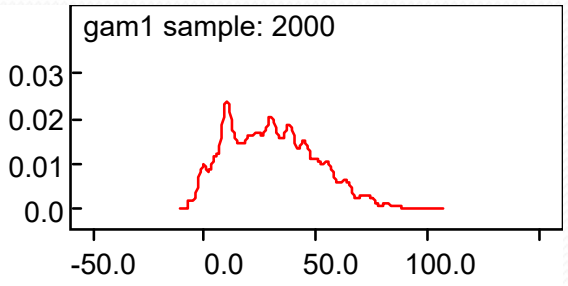
Lowest DIC is model VI

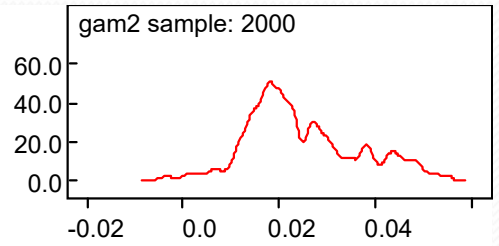
- D: distance; V: UH; U: CH; A: age

Model VI results

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
gam0	-7.623	1.475	0.2146	-10.64	-7.43	-5.53	10001	2000
gam1	30.56	19.92	1.523	-1.001	29.31	73.13	10001	2000
gam2	0.02479	0.01175	0.001538	0.003891	0.02232	0.04966	10001	2000





Reference

- Lawson, A. B. (2012) Bayesian Point Event Modeling in Spatial and Environmental Epidemiology. *Statistical Methods in Medical Research* 21, 5, 509-530