CHAPTER 11: SPATIAL DATA

So far we have concentrated on longitudinal data; we now consider spatial data. Two important distinctions are between lattice and non-lattice data. and between point data, and data aggregated over space.

Pure spatial data are fundamentally different from longitudinal data from multiple individuals, since they are a single realization. In this sense they are closer to time series data, though such data are often regularly spaced (which under assumptions of stationarity allows simplification of estimation), whereas with spatial data it is usual to have non-lattice data.

This lack of replication aspect of non-lattice data dictates that sandwich estimation is not available. Hence we concentrate on likelihood-based methods.

Space-time data do offer replication.

Motivating Examples

Childhood Asthma in Anchorage, Alaska

Study details:

- Study PI is Dr Mary Gordian. Data were collected on first grade children in Anchorage, with questionnaires being sent to the parents of children in 13 school districts (the return rate was 70% which has implications for interpretation).
- We analyze data on 905 children, with 885 between the ages of 5 and 7. There were 804 children without asthma, and 101 with asthma.
- The exposure of interest is exposure to pollution from traffic. Traffic counts were recorded at roads throughout the study region and a 50m buffer was created at the nearest intersection to the child's residential address and within this buffer traffic counts were aggregated (for confidentiality reasons the exact locations were not asked for in the survey).

Figure 42 shows the residential location of the cases and non-cases in Anchorage.

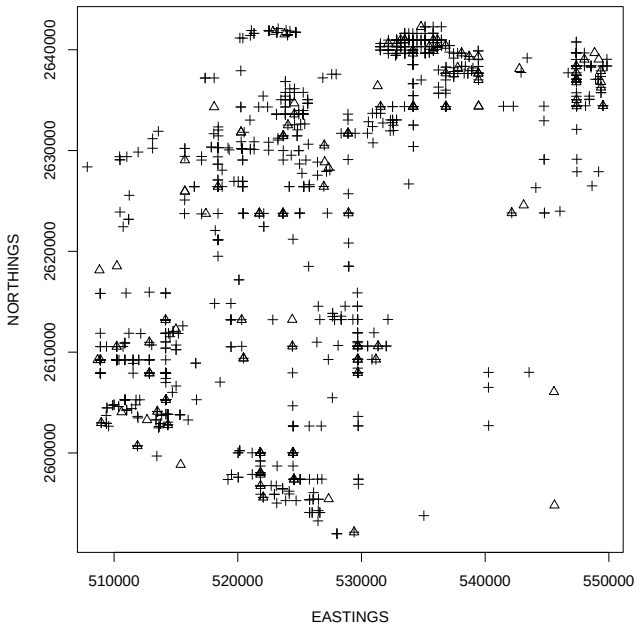

Figure 42: Asthma cases (△) and non-cases (+) in Anchorage.

Naive non-spatial logistic modeling

- Initially we may ignore confounding and the spatial nature of the data and fit a logistic regression of asthma incidence on exposure (with the exposure variable scaled to lie between 0 and 10).
- Such an analysis gives an odds ratio of 1.09 with a 90% confidence interval of 1.00–1.18.
- This analysis assumes that, given exposure, the Bernoulli 0/1 labels are independent. Due to unmeasured variables with spatial structure this is dubious and will result in inappropriate standard errors.

Scottish Lip Cancer Data

Incidence rates of lip cancer in males in 56 counties of Scotland, registered in 1975–1980. These data were originally reported in the mapping atlas of Kemp, Boyle, Smans and Muir (1985).

The form of the data is:

- Observed and expected number of cases (based on the county age populations),
- A covariate measuring the proportion of the population engaged in agriculture, fishing, or forestry (AFF),
- The standardized morbidity ratio,
- The projections of the longitude and latitude of the area centroid, and the "position" of each county expressed as a list of adjacent counties.

County No. i	Obs Cases Y_i	Exp Cases E_i	Prop AFF	SMR	Project N (km)	Projext E (km)	Adjacent Counties
1	9	1.4	0.16	6.43	834.7	162.2	5,9,19
2	39	8.7	0.16	4.48	852.4	385.8	7,10
3	11	3.0	0.10	3.67	946.1	294.0	12
4	9	2.5	0.24	3.60	650.5	377.9	18,20,28
5	15	4.3	0.10	3.49	870.9	220.7	1,12,19
6	8	2.4	0.24	3.33	1015.2	340.2	Island
7	26	8.1	0.10	3.21	842.0	325.0	2,10,13,16,17
8	7	2.3	0.07	3.04	1168.9	442.2	Island
9	6	2.0	0.07	3.00	781.4	194.5	1,17,19,23,29
...							
47	2	5.6	0.01	0.36	640.8	277.0	24,31,46,48,49,53
48	3	9.3	0.01	0.32	654.7	282.0	24,44,47,49
49	28	88.7	0.00	0.32	666.7	267.8	38,41,44,47,48,52,53,54
50	6	19.6	0.01	0.31	736.5	342.2	21,29
51	1	3.4	0.01	0.29	678.9	274.9	34,38,42,54
52	1	3.6	0.00	0.28	683.7	257.8	34,40,49,54
53	1	5.7	0.01	0.18	646.6	265.6	41,46,47,49
54	1	7.0	0.01	0.14	682.3	267.9	34,38,49,51,52
55	0	4.2	0.16	0.00	640.1	321.5	18,24,30,33,45,56
56	0	1.8	0.10	0.00	589.9	322.2	18,20,24,27,55

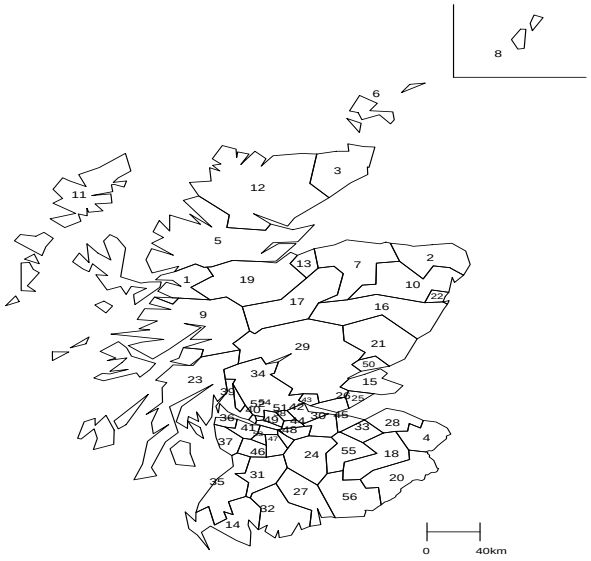

Figure 43: Labels for 56 counties of Scotland.

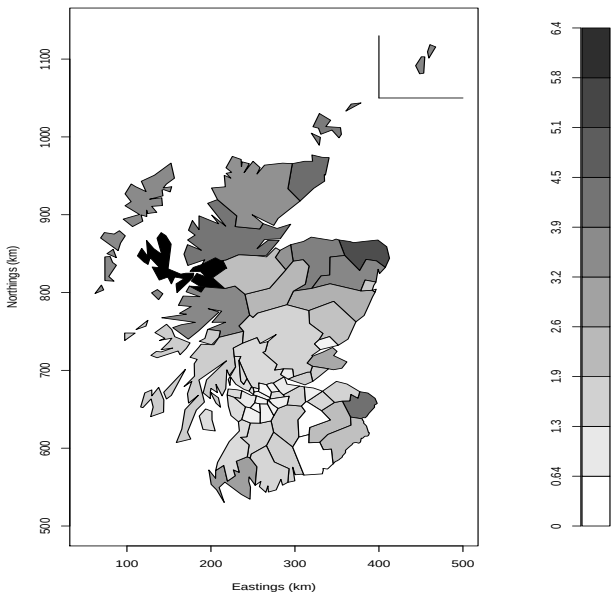

Figure 44: SMRs for male lip cancer in 56 counties of Scotland.

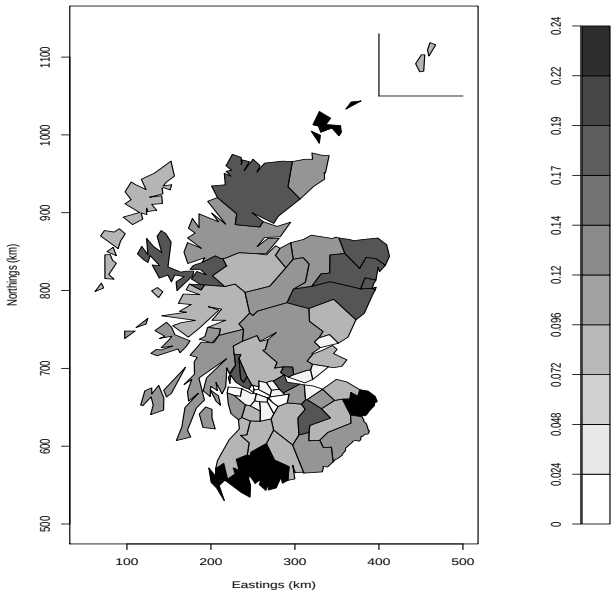

Figure 45: Map of proportion of individuals in agriculture, fishing and farming, for 56 counties of Scotland.

Inference for Spatial Regression

We may extend the GLMM approaches of Chapters 9 and 10 to the spatial setting.

Inference may then proceed via likelihood or Bayesian methods.

Suppose we have data Y_i with spatial location $\boldsymbol{s}_i$ (latitude and longitude, for example), along with covariates $\boldsymbol{x}_i$, $i = 1, ..., m$. An obvious GLMM is then

Random Component: $Y_{ij}|\theta_{ij}, \alpha \sim p(\cdot)$ where $p(\cdot)$ is a member of the exponential family, that is

$$p(y_i|\theta_i, \alpha) = \exp[\{y_i\theta_i - b(\theta_i)\})/a(\alpha) + c(y_i, \alpha)],$$

for $i = 1, ..., m$ locations.

Systematic Component: If $\mu_i = \mathrm{E}[Y_i|\theta_i, \alpha]$ then we have a link function $g(\cdot)$, with

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta} + b_i,$$

so that we have introduced random effects into the linear predictor. The above defines the *conditional* part of the model. The random effects are then assigned a distribution, and in a spatial setting it is natural to assume

$$\boldsymbol{b} = (b_1, ..., b_m)^{\mathrm{T}} \sim_{iid} N_m(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\alpha})$ is an $m \times m$ covariance matrix We also have

$$\mathrm{var}(Y_i|\theta_i, \alpha) = \alpha v(\mu_i).$$

Covariance Models

Many choices are available for the variance-covariance model $\boldsymbol{\Sigma}$. A simple choice has

$$\Sigma_{ij} = \alpha_1 \exp(-\alpha_2 \mid \boldsymbol{s}_i - \boldsymbol{s}_j \mid),$$

for $i, j = 1, ..., m$, with $\alpha_1, \alpha_2 > 0$. This model is *isotropic* since the covariance only depends on the distance between points.

Obvious links with the AR(1) models we considered in longitudinal data analysis.

It is sometimes useful to think of the b_i's as arising from a Gaussian Random Field (GRF). Specifically, a random field $b(\boldsymbol{s}) \in R^2$ is a (bivariate) Gaussian random field if $b(\boldsymbol{s}_1), ..., b(\boldsymbol{s}_n)$ is multivariate normal for any $\boldsymbol{s}_i \in R^2$, $i = 1, ..., n$. This allows us to do prediction to arbitrary locations.

An obvious extension to the linear predictor is to add independent and identically random effect also. For example, we might fit the model

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta} + b_i + v_i,$$

with $v_i \sim_{iid} \mathrm{N}(0, \sigma_0^2)$, $i = 1, ..., n$.

Childhood Asthma

We carried out a number of regression models from a Bayesian perspective. Table 12 summarizes these analyses for the odds ratio of interest. We consider the model:

- *Stage 1:*
$$Y_i|\boldsymbol{x}_i, b_i \sim \text{Bernoulli}\{p(\boldsymbol{x}_i, b_i)\},$$
for $i = 1, \ldots, n$, where Y_i is the case/non-case status, $\boldsymbol{x}_i$ is a vector containing the exposure of interest and confounders, and b_i represent unmeasured spatial effect.
- *Stage 2:*
$$\text{logit}\left\{p(\boldsymbol{x}, b_i)\right\} = \boldsymbol{x}_i\boldsymbol{\beta} + b(\boldsymbol{s}_i),$$
where $b_i = b(\boldsymbol{s}_i)$ is a realization of a GRF.
- *Stage 3:* Priors on $\boldsymbol{\beta}$ and the parameters of the GRF.

Adjustment for spatial dependence was carried out with covariances between points d apart being given by $\alpha_1 \exp(-d\alpha_2)$.

The results are almost identical across the different models, all giving evidence of an association between asthma and exposure to traffic.

The prior on α_2 was uniform on (0,50) and the posterior median was 25 with 95% credible interval was 2–49, indicating that the prior has hardly been changed.

The spatial variance parameter α_1 has posterior median 0.002, again providing evidence that there is no evidence of spatial dependence in the residuals.

Adjust for confounders	Adjustment for spatial	Odds Ratio $\exp(\beta)$	95% Interval
No	No	1.08	0.99–1.20
Yes	No	1.08	0.98–1.20
No	Yes	1.08	0.98–1.19
Yes	Yes	1.08	0.97–1.19

Table 12: Odds ratio summaries (posterior medians and credible intervals) under various models.

Scottish Lip Cancer Data

The individual-level model $Y_{ij} \sim_{iid} \text{Bernoulli}(\text{e}^{\beta_0+\beta_1 x_{ij}})$ for individual i, $j = 1, ..., N_i$, leads to

$$Y_i \sim_{iid} \text{Poisson}(N_i\{(1-\overline{x}_i)\text{e}^{\alpha_0} + \overline{x}_i\text{e}^{\alpha_0+\alpha_1}\}),$$

$i = 1, ..., m$.

Random effects can be added to the linear predictor, spatial or non-spatial (or both). Here there is a large change in $\text{e}^{\widehat{\alpha}_1}$, because the exposure has spatial structure.

Model	Relative risk	St. Err.
Quasi-likelihood	22.7	7.0
Non-spatial r.e.'s	22.5	7.8
GRF model	6.3	3.0

Table 13: Estimates and standard errors for individual relative risk, e^{α_1}.

For a Bayesian analysis we require a proper prior on α_1.

Assigning an improper uniform prior to α_0 we integrate this parameter from the model to give

$$p(\alpha_1|\boldsymbol{y}) \propto \prod_{i=1}^{n} \left(\frac{N_i[(1-x_i) + x_i\text{e}^{\alpha_1}]}{\sum_{i=1}^{n} N_i[(1-x_i) + x_i\text{e}^{\alpha_1}]} \right)^{y_i},$$

which tends to the constant

$$\prod_{i=1}^{n} \left(\frac{N_i(1-x_i)}{\sum_{i=1}^{n} N_i(1-x_i)} \right)^{y_i} \tag{48}$$

as $\alpha_1 \to -\infty$, showing that a proper prior is required. The constant (48) is non-zero unless $x_i = 1$ in any area with $y_i \neq 0$.

The reason for the impropriety is that $\alpha_1 = -\infty$ corresponds to a relative risk of zero, so that exposed individuals cannot get the disease, which is not inconsistent with the observed data unless $x_i = 1$ in an area (all individuals are exposed), and $y_i \neq 0$, in which case clearly the cases are due to exposure. A similar argument holds as $\alpha \to \infty$ with replacement of $1 - x_i$ by x_i in (48) providing the limiting constant. Figure 46 illustrates this behavior for the Scottish lip cancer example, for which $x_i = 0$ in five areas.

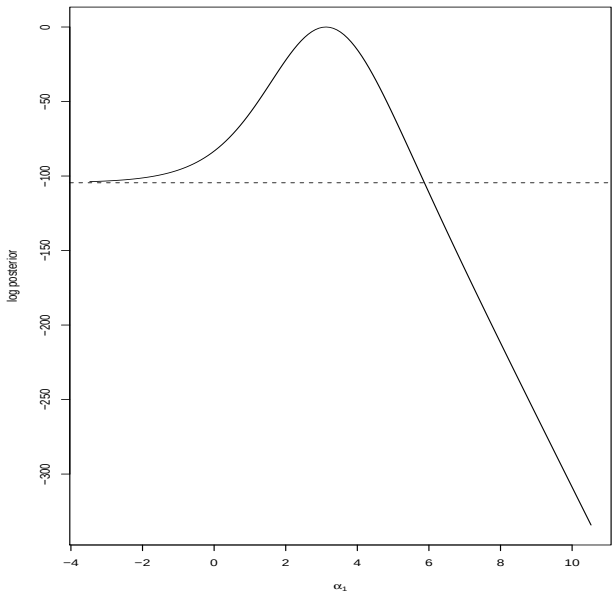

Figure 46: Log posterior for α_1 for the Scottish data; the horizontal line is the constant to which this function tends to as $\alpha_1 \rightarrow \infty$.

GeoBUGS has spatial models, and allows maps to be drawn.

```
model
{
  for (i in 1 : N) {
      Y[i]  ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]) +
      log( (1-X[i])*exp(alpha0) + X[i]*exp(alpha0+alpha1)) + U[i]
      mean[i] <- 0
  }
# Multivariate prior distribution for spatial random effects:
  U[1:N] ~ spatial.exp(mean[], xm[], ym[], tau.U, phi, 1)
#
# The following prior is derived by assuming that there is a 5% chance that the
# correlations die to 0.5 in less that 5km, and a 95% chance that they die
# to 0.5 in less than 100km.
  dhalf ~ dlnorm(3.107,0.9106)
  phi <- 0.6931/dhalf
  alpha0 ~ dflat()
# Prior says 50% RR less than one, 95% less than 50.
  alpha1 ~ dnorm(0.0,0.1768)
# Parameters of interest
  base <- exp(alpha0)
  RRx <- exp(alpha1)
}
```

```
list(N = 56, Y = c( 9, 39, 11, 9, 15, 8, 26, 7, 6, 20, 13, 5, 3, 8, 17, 9, 2, 7,
  9, 7, 16, 31, 11, 7, 19, 15, 7, 10, 16, 11,5, 3, 7, 8, 11, 9, 11, 8,
  6, 4, 10, 8, 2, 6, 19, 3, 2, 3, 28, 6, 1, 1, 1, 1, 0, 0), E = c(
  1.4, 8.7, 3.0, 2.5, 4.3, 2.4, 8.1, 2.3, 2.0, 6.6, 4.4, 1.8, 1.1,
  3.3, 7.8, 4.6, 1.1, 4.2, 5.5, 4.4, 10.5,22.7, 8.8, 5.6,15.5,12.5,
  6.0, 9.0,14.4,10.2, 4.8, 2.9, 7.0, 8.5,12.3,10.1,12.7, 9.4, 7.2,
  5.3, 18.8,15.8, 4.3,14.6,50.7, 8.2, 5.6, 9.3,88.7,19.6, 3.4, 3.6,
  5.7, 7.0, 4.2, 1.8),X = c(0.16,0.16,0.10,0.24,0.10,0.24,0.10, 0.07,
  0.07,0.16, 0.07,0.16,0.10,0.24, 0.07,0.16,0.10, 0.07, 0.07,0.10,
  0.07,0.16,0.10, 0.07, 0.01, 0.01, 0.07, 0.07,0.10,0.10,
  0.07,0.24,0.10, 0.07, 0.07, 0,0.10, 0.01,0.16, 0, 0.01,0.16,0.16, 0,
  0.01, 0.07, 0.01, 0.01, 0, 0.01, 0.01, 0, 0.01, 0.01,0.16,0.10),
xm = c(
162.1894, 385.7761, 293.9555, 377.9338, 220.6786,340.1739, 324.9915, 442.2445, 194.5176, 367.6924,
112.8916, 247.7566, 289.5922, 227.9563, 342.3574,351.3505, 280.4916, 341.6081, 249.6855, 359.5902,
348.7138, 388.7655, 180.4228, 295.4908, 333.1159,312.0605, 290.1701, 359.4153, 291.3727, 303.4219,
257.4402, 264.9711, 336.4464, 258.0319, 227.1801,234.5294, 218.3428, 279.1010, 235.0805, 254.1736,
250.8301, 287.1202, 292.3773, 288.0333, 320.5682,257.8758, 276.9737, 281.9644, 267.8444, 342.226,
274.8713, 257.8069, 265.5934, 267.8921, 321.4991,322.1780),
ym =c(834.7496, 852.3782, 946.0722, 650.501,870.9356, 1015.154, 842.0317, 1168.904, 781.3746,
828.219, 903.1592, 924.9536, 842.3052, 561.1628,713.0808, 792.1617, 801.0356, 628.6406, 825.8545,
610.6554, 760.2982, 812.7655, 699.6693, 635.7658,701.8189, 691.102, 586.6673, 669.4746, 746.2605,
670.1395, 605.9585, 568.3428, 658.671, 716.452,598.2521, 668.0481, 641.4785, 670.285, 697.044,
677.589, 657.4675, 680.7535, 699.3761, 665.2905,671.6064, 631.046, 640.8285, 654.6629, 666.7073,
736.4561, 678.8585, 683.7104, 646.5754, 682.2943,640.1429, 589.9408))
```