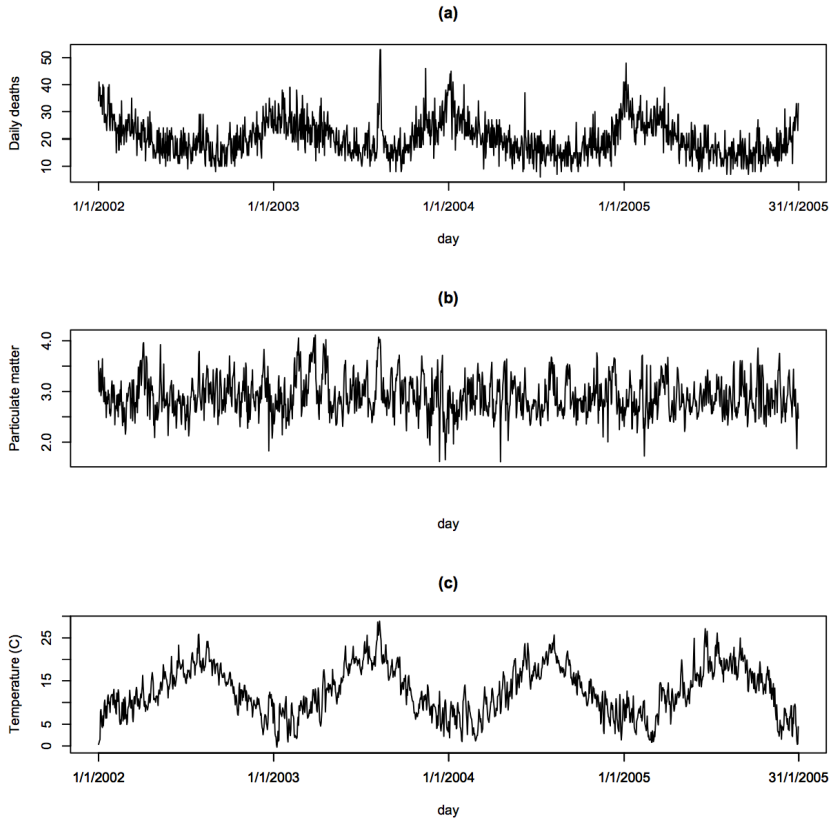
Time series epidemiology

Lecture 19

Methodological aspects of modelling temporal relationships between air pollution, temperature and mortality

- Chronic studies; long terms effects of air pollution on health
- Acute studies; short term effects



Chronic Mortality Studies

Six Cities Chronic Mortality Study

Dockery et al, NEJM 1993

- 8,096 adults followed for up to 15 yrs
- Increased relative risk of death associated with fine particles (18.6 μg/m^{-3}), adjusted for age, sex, smoking, BMI
  - Total Mortality, 26% (8%, 47%)
  - CardioPulmonary Mortality, 37% (11%, 68%)
  - Lung Cancer Mortality, 37% (-19%, 131%)
- Cox Proportional Hazards
- Semi-ecological design

Acute Mortality Studies

Summary of selected daily time series studies

Percentage Increase in Mortality per 10 μg/m^3 increase

Study area	Particular Measure	Mean PM_{10}	Total	Respiratory
Santa Clara, CA	CoH	35	0.8 (0.2,1.5)	3.5 (1.5,5.6)
Philadelphia, PA	TSP (2d m)	40	1.2 (0.7,1.7)	1.7 (1.0,2.4)
Utah Valley, UT	PM_{10} (5d m)	47	1.5 (0.9,2.1)	3.7 (0.7,6.7)
Birmingham, AL	PM_{10} (3d m)	48	1.0 (0.2,1.9)	1.5 (-5.8,9.4)
Cincinnati, OH	TSP	42	1.1 (0.5,1.7)	2.7 (-0.9,6.6)
St. Louis, MO	PM_{10} (p day)	28	1.5 (0.1,2.9)	NA
Kingston, TN	PM_{10} (p day)	30	1.6 (-1.3,4.6)	NA
Detroit, MI	TSP	48	1.0 (0.5,1.6)	NA
Steubenville, OH	TSP	61	0.7 (0.4,1.0)	NA
Athens, Greece	Smoke	NA	0.9 (0.7,1.2)	NA
Los Angeles, CA	PM_{10}	58	0.5 (0.0,1.1)	NA
Santiago, Chile	PM_{10}	115	0.6 (0.4,0.9)	NA
Chicago, IL	PM_{10}	NA	0.5 (0.1,1.1)	NA
Amsterdam, NL	PM_{10}	NA	0.6 (-0.1,1.4)	NA
Erfut, Germany	SP	NA	0.6 (0.1,1.1)	NA
Sao Paulo. Brazil	PM_{10} (2d m)	82	1.3 (0.7,1.9)	NA
APHEA (per 50)				
Western (n=5)	PM_x (p day)	NA	3.1 (2.2,4.0)	NA
Central Eastern	TSP (p day)	NA	4.9 (-0.4,10.5)	NA

Interpretation / comparison

- Interpretation and comparison of results can be difficult because of differences in
  - Measurements of pollutants
  - Statistical methodologies
  - Potential confounders that are adjusted for
  - Lagged values that are used

Issues

- Counts from underlying Poisson process
- Likely to be over-dispersion
- Unit of observation is the day
- Auto-correlation
- Underlying age distribution and smoking history don't vary day-to-day

Model Framework

- Parametric, GLMs
- Semi-parametric, GAMs

Modelling approach

- Long term trends
- Seasonality / Cyclical events
- Short term / Daily variation

Parametric approach, GLMs

- Sum of sine waves with different frequencies (together with cosine to account for phase of the pattern)
- $Y_t = \alpha \sin(wt) + \beta \cos(wt)$

  Where the frequency ($\omega = 2 \times \pi / p$)
  The period is an expression of the proportion of the year,
  i.e. a cycle happening twice a year will have period 365/2 days
- Assumes the seasonal peak is the same each year

Semi-Parametric approach, GAMs

- Outcome is assumed to depend on a sum of smooth functions of the predictor variables
  (i.e. time, temperature, PM10)
- $\log(\text{Expected Daily Deaths}) = \alpha + S_1(\text{time}) + S_2(\text{temp}) + \beta \text{PM}_{10}$
- Note the linear effect of PM10

Traditional Approach to model selection

- Comparison of models M_k and M_q, with r_k and r_q parameters.
  - M_k is nested within M_q, so that $r_k \leq r_q$
  - Forwards / backwards / stepwise selection
  - Possible criteria include deviance, AIC,

    BIC=$l(\hat{\theta}) - (r_k/2)\log n$
- Choice between which variables to include can be very difficult, especially if they are highly correlated.

Worked example with multiple exposure measurements over space

Short-term time series studies relate to a fixed region and have been conducted for periods of between 1 and 14 years.

The following daily data are required.

Health Counts of mortality from the population living within the study region, denoted by $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

Meteorology Summaries of meteorological conditions, such as temperature, denoted by $\mathbf{M} = (M_1, \ldots, M_n)$.

Air pollution (Potentially noisy) ambient air pollution concentrations, $W_t(\mathbf{s}_r)$, for day t at spatial location $\mathbf{s}_r$. A set of K locations are monitored.

Concentrations and exposures

- Note that the measure of air pollution relates to concentrations rather than actual exposures being experienced by individuals
- Exposures are generally lower than concentrations due to the time that people spend indoors away from the majority of pollution sources
- For a discussion of the relationship between ambient concentrations and personal exposures see Zeger et al. (2000) and Shaddick et al. (2008)
- However, it is ambient concentrations that are subject to regulation and are almost universally used in health studies, meaning that these health effects of of interest in there own right.

Standard modelling approach

Quasi-likelihood log-linear models are typically used rather than a Poisson model. The daily mortality data are modelled by

$$E(Y_t) = \mu_t \qquad \text{var}(Y_t) = \phi\mu_t$$

where ϕ is a dispersion parameter and the mean function is given by

where X_{t-l} is the average air pollution concentration on day $t - l$ across the study region and β_x is its relationship with health.

Further details

- In this model, the pollution levels are lagged by l days, which typically ranges between 1 and 5 days
- The non-pollutant covariates typically include
  - smooth functions of time modelled using parametric or non-parametric functions, e.g. splines
  - indicator variables, such as ‘day of the week effects’
  - lagged values of meteorological variables

Estimating X_t

The pollution and health data are often spatially misaligned.

- The health data relate to the entire study region.
- The pollution data are measurements at K distinct (point) locations across the region.

Gelfand et al (2001) call this the *change of support problem*, and argue that the desired measure of pollution is

$$X_t = \frac{1}{|\mathcal{R}|} \int_{\mathbf{s} \in \mathcal{R}} X_t(\mathbf{s}) \mathbf{ds}$$

where $\mathcal{R}$ denotes the study region.

Estimating X_t

However X_t is unknown, so the general problem is

how best to estimate the average of a spatially continuous surface from data at a small number of locations.

The majority of studies estimate X_t with

$$W_t = \frac{1}{K} \sum_{r=1}^{K} W_t(\mathbf{s}_r)$$

i.e. average of the concentrations over all the monitors in the region.

As such they estimate

- β_w - the effects of the pollution estimate W_t, and not
- β_x - the effects of the true pollution levels X_t.

Is W_t a good estimate of X_t?

The accuracy with which W_t estimates X_t (and hence β_w estimates β_x) will depend on

Spatial variation The higher the spatial variation in the pollution surface the less accurate W_t is likely to be as an estimate of X_t

Monitor placement If the monitors are clustered or intentionally positioned at locations with high (or low) pollution levels, then W_t is likely to be biased.

Measurement error If the monitors measure with error W_t may also be biased.

2. Spatio-temporal modelling

Instead of estimating average daily pollution levels by W_t can we,

- model the spatio-temporal variation in the pollution surface.
- use the model to predict the pollution levels on a regular lattice.
- estimate X_t by averaging the predicted values on the lattice.
- use the estimate, $\hat{X}_t$, in a health model to estimate β_x.

There are a number of ways of doing this.

Stage (i) - Pollution model

Model

$$
\begin{aligned}
\ln[\mathbf{W}_t(\mathbf{s}_l)] &\sim \mathsf{N}(\ln[\mathbf{X}_t(\mathbf{s}_l)]\ ,\ \sigma^2_\epsilon),\\
\ln[\mathbf{X}_t(\mathbf{s}_l)] &= \mathbf{b}_{tl}^{\mathsf{T}}\boldsymbol{\alpha} + \theta_t + \phi_{\mathbf{s}_l},\\
\theta_t &\sim \mathsf{N}(\rho_\theta\theta_{t-1}\ ,\ \sigma^2_\theta),\\
(\phi_{\mathbf{s}_1},\ldots,\phi_{\mathbf{s}_k}) &\sim \mathsf{N}(\mathbf{0}\ ,\ \sigma^2_\phi\Sigma(\rho_\phi)).
\end{aligned}
$$

Priors

$$
\begin{aligned}
\boldsymbol{\alpha} &\sim \mathsf{N}(\boldsymbol{\mu}_\alpha\ ,\ \Sigma_\alpha),\\
f(\sigma_\epsilon,\sigma_\theta,\sigma_\phi,\rho_\theta) &\propto 1,\\
\rho_\phi &\sim \text{Discrete Uniform}(\nu_1,\ldots,\nu_r).
\end{aligned}
$$

Model characteristics

- Separable space-time dependence structure, so that the same spatial structure holds for each time point.
- First order autoregressive structure in time.
- Stationary and isotropic spatial structure given by the Matern class of correlation functions, $C(h) = \sigma^2 \frac{1}{\Gamma(\kappa)} \left(\frac{\rho h}{2}\right)^{\kappa} 2K_{\kappa}(\rho h)$, where ρ governs the range of spatial dependence and the smoothness of the process increases with v.
- Setting the smoothness parameter, v, to be 0.5 gives an exponential model, $\text{Corr}[\mathbf{s}, \mathbf{s}'] = \exp(-||\mathbf{s} - \mathbf{s}'||)/\rho_{\phi}$
- The model can handle relatively large quantities of missing data due to its separable space-time dependence structure.

Stage (ii) - Estimating X_t

Average pollution levels are estimated from

$$f\left[\mathbf{X}_t|\mathbf{W}\right] = \int_{\mathbf{\Theta}} f[\mathbf{X}_t|\mathbf{\Theta}, \mathbf{W}] f[\mathbf{\Theta}|\mathbf{W}] \mathrm{d}\mathbf{\Theta},$$

the posterior predictive distribution given the pollution data $\mathbf{W}$. Here $\mathbf{\Theta}$ denotes all parameters from the stage (i) model.

Stage (iii) - Health model

Consider two different health models.

Two-stage - Estimate X_t by the median of its posterior predictive distribution, and plug the values into a quasi-likelihood health model.

Fully Bayesian - combine the stage (i) and (ii) models with the health model

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t), \\ \mu_t &= \exp(\beta_1 + f_1(t|\kappa_1) + f_2(M_t|\kappa_2) + X_t\beta_x), \end{aligned}$$

where X_t is a random variable that is updated at each iteration of the MCMC simulation algorithm.

Fully Bayesian model

- This combined approach correctly allows the variation in the posterior predictive distribution of $\mathbf{X}$ to be correctly fed 'upwards' into the health model
- To ease the computational burden, the feedback between the health and pollution models ('downwards') is cut, i.e. removing the dependence between $\mathbf{Y}$ and $\mathbf{X}$
- This can be done using the `cut` function in `WinBUGS`

3. Simulation study

To assess the effects of these different approaches, we conduct a simulation study to compare the performance of modelling exposures and feeding through the uncertainty with the standard approach (using W_t).

- The study is set on a 100 $\times$ 100 square grid (10,000 points in total) over 1095 days (3 years).
- The pollution data are generated from the spatio-temporal model in stage (i).
- The health data are generated from a Poisson model with mean function

$$\mu_t = \exp(\beta_1 + f_1(t|\kappa_1) + f_2(M_t|\kappa_2) + X_{t-l}\beta_x)$$

- 200 data sets are generated under different scenarios.

More details

- The vector of mortality counts are generated from a Poisson distribution with the previously given mean function where the non-pollutant covariates are taken from the Greater London data used in the case study
- Two values of relative risk ($\exp(\beta_x)$) are used:
  - RR=1.02, which is similar to that observed in previous studies
  - RR=1.5, which although possibly unrealistic in this context will enable any differences to be more easily observed
- Simulations were performed with 30 monitoring sites

Results

Baseline - a relatively flat pollution surface.

	Standard	**Two stage**	**Fully Bayesian**
True RR=1.02	1.020	1.020	1.020
True RR=1.5	1.500	1.500	1.500
Coverage probability (95%)	96.0	96.5	95.0

30 monitors are used, and the estimated risks are averages over 200 simulated data sets.

High spatial variation - independent of monitor locations.

	Standard	**Two stage**	**Fully Bayesian**
True RR=1.02	1.020	1.020	1.020
True RR=1.5	1.494	1.498	1.500
Coverage probability (95%)	88.5	87.0	94.5

High spatial variation - Locations with high pollution levels are more likely to be monitored.

	Standard	**Two stage**	**Fully Bayesian**
True RR=1.02	1.016	1.017	1.017
True RR=1.5	1.416	1.450	1.451
Coverage probability (95%)	45.0	68.0	75.0

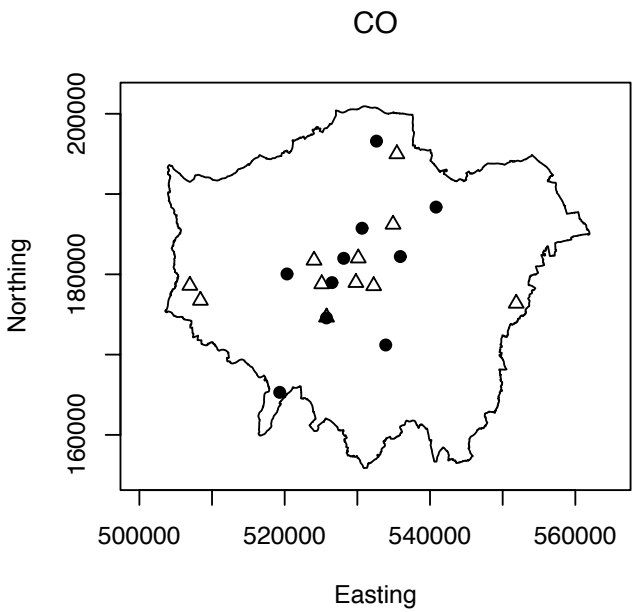
High measurement error - Error in the pollution monitors.

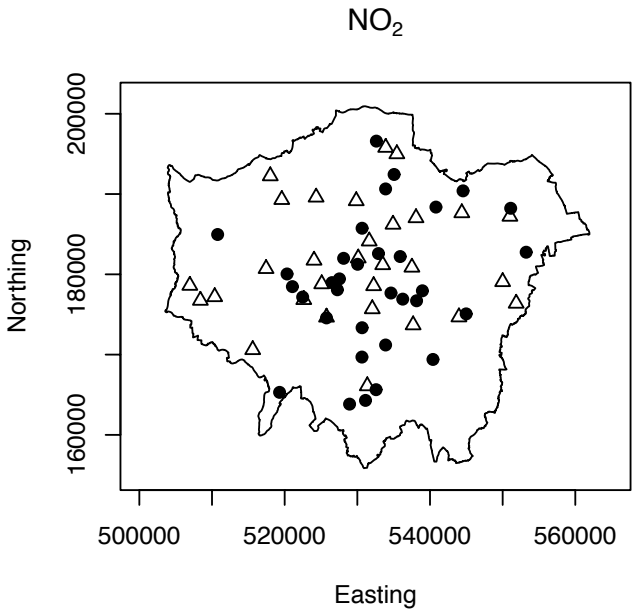
	Standard	**Two stage**	**Fully Bayesian**
True RR=1.02	1.019	1.020	1.020
True RR=1.5	1.457	1.501	1.489
Coverage probability (95%)	89.0	94.0	93.5

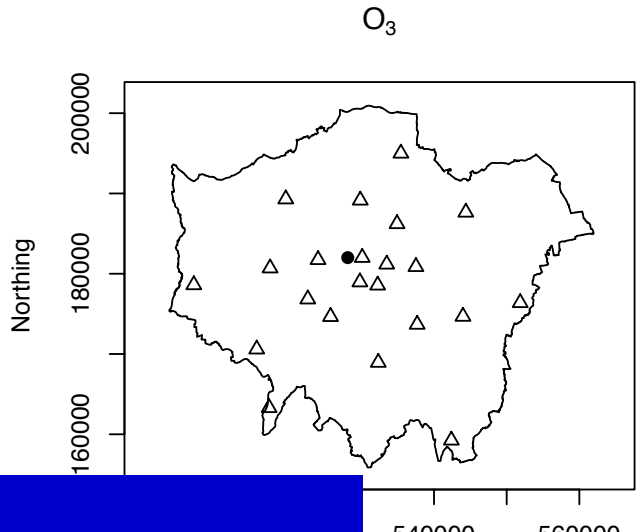
Applying the models to data from Greater London

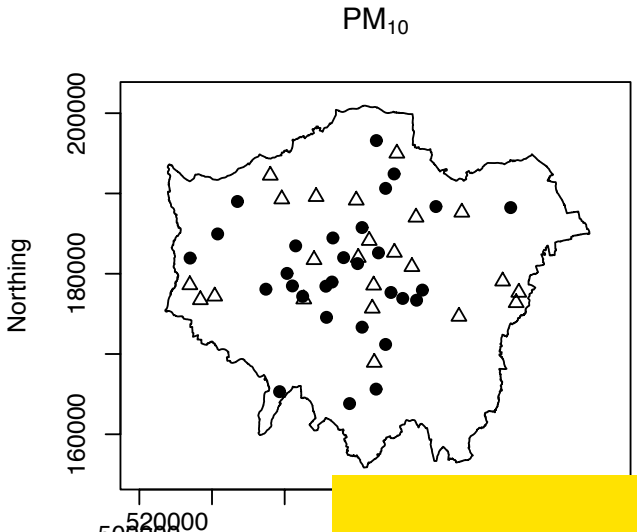
- Daily data were available for Greater London for 2003-2005.
- The response comprised counts of respiratory mortality in the over 65's.
- The results for four pollutants are presented, CO, NO_2, O_3 and PM_{10}.
- Other covariates in the health model included temperature, a smooth function time trend (using a natural cubic splines with 10 df per year) and an indicator for weekday / weekend.

Monitor locations









Pollution summary

Table: Summary of the spatial and temporal variation in the pollution data. In both cases the summaries relate to the average (mean) standard deviation over all days or monitoring sites.

Data	**Monitors**	**Missing data**	**Spatial variability**	**Temporal variability**
CO	21	6.1%	0.29	0.28
NO_2	63	6.3%	20.3	19.1
O_3	23	4.6%	9.80	19.4
PM_{10}	53	7.9%	8.52	10.9

Pollution effects

Table: Summary of the estimated relationships between pollution and mortality from each of the three models. The results are presented as relative risks for an increase in pollution of one temporal standard deviation at a lag of one day.

Pollutant	**Relative risk and uncertainty interval**	
	Standard (i)	**Fully Bayesian (iii)**
CO	1.012 (0.997, 1.028)	1.012 (0.998, 1.033)
NO_2	1.010 (0.995, 1.025)	1.013 (0.998, 1.027)
O_3	1.033 (1.015, 1.051)	1.033 (1.016, 1.052)
PM_{10}	1.017 (1.001, 1.032)	1.020 (1.004, 1.035)

5. Discussion

- High spatial variation results in the quasi-likelihood analysis having confidence intervals that are too narrow.
- Locating monitors at sites with high pollution results in bias and poor coverage for the quasi-likelihood approach.
- Modelling the pollution surface generally reduces the bias (where there is any) and increases the coverage probability.
- Significant associations observed between pollutants and respiratory mortality of similar magnitude to those observed in previous studies.