Hierarchical Bayesian spatio-temporal modelling of air pollution in the UK for estimating long term exposure

Sabyasachi Mukhopadhyay[1], Sujit Sahu[1], Lucy Davis[2]

[1]University of Southampton [2]Met Office, Exeter

EPSRC funded project

UNIVERSITY OF Southampton

Outline

1. Importance of monitoring air-pollution
2. Our goal
3. New spatio-temporal models
4. Results
5. Discussion and future research

Importance of air-pollution monitoring

- Outdoor pollution has long-term effects on human health in terms of shorter life expectancy and greater medical expenses.

Headlines

Measures taken in UK

- Pollution directives enacted according to UK, European and WHO guidelines.
- "The National Survey of Air Pollution" established in 1961 to measure pollution.

Importance of air-pollution monitoring

- Outdoor pollution has long-term effects on human health in terms of shorter life expectancy and greater medical expenses.

Headlines

Measures taken in UK

- Pollution directives enacted according to UK, European and WHO guidelines.
- "The National Survey of Air Pollution" established in 1961 to measure pollution.

Difficulties in modelling air-pollution data

- Collection of data is expensive and hence monitoring sites are sparse.
- High proportion of data absent in monitoring sites due to discontinuation of existing sites and/or addition of new sites.
- There is high variability in the data even within small distances, especially in urban areas (Shaddick et al., 2014).
- Modelling such data is a challenge.

Recent Statistical modelling of UK data on air pollution

1. Large number of research articles. We mention two most recent work on UK data.
2. Pirani, Gulliver, Fuller, Blangiardo (2014) did spatio-temporal modelling on short-term affect of PM_{10} in London.
    1. Used data on PM_{10} for 728 days during 2002–2003. Covariates are output of numerical model on a 1km grid, data on emission, temperature etc.
    2. Fitted and compared 5 different regression models.
    3. Did not incorporate spatio-temporal interaction term.
3. Shaddick et al. (2013) used data on annual average of NO_2 concentration from parts of Europe including UK in 2001.
    1. Spatial model includes various covariates affecting air pollution.
    2. No temporal modelling, hence cannot be used for measuring long term exposure.

Recent Deterministic modelling on UK data

1. Computer simulation model named Air Quality Unified Model (AQUM) (Savage et al., 2013).
    1. Use atmospheric variables like teperature, humidity, wind speed, wind direction etc.
    2. Use data on emission from various sources.

Output of AQUM

Hourly concentration of air pollutants over corners of a square grid (1km or 12km) for a specified period of time.

3. The AQUM output (like those from other similar models, e.g. CMAQ) is not very accurate.
4. Adjusted to correct bias using observed hourly data. We will denote this by AAQUM.
5. AAQUM values are recommended for prediction at an unobserved locations and also for forecasting.

Aims and objectives of our work

1. To model daily levels of four major pollutants namely, $PM_{2.5}$, PM_{10}, Ozone and NO_2 for the period 2007–2011.

2. To build up a process based suitable spatio-temporal model that
   1. can handle highly variable air pollution data.
   2. is more accurate than recently developed methods.
   3. is based on a spatial process which allows us to interpolate at any unobserved location.

3. To incorporate output of our model (along with their uncertainties) into the model measuring the impact of pollution on human health.

Preliminaries

- We have
  1. Observed daily data from 166 AURN sites for five years (1826 daily values in each site).
  2. Daily AQUM model output, generated on corners of 12km grid cells of size 79$\times$ 80 over UK, supplied by the Met Office.
- Square-root transformation is taken to stabilize variance of the data (like many other authors, e.g., Berrocal et al., 2010).
- We also have information on the types of the monitoring sites, like Kerbside, Rural, Urban etc.
- AQUM output can be used as a covariate in the model since it is generated using other important covariates like emission and meteorological variables e.g., temperature, humidity, wind speed and direction.

Mean of different air pollutants categorised by type of sites

Type	Number	$PM_{2.5}$	PM_{10}	Ozone	NO_2
Urban Background (UB)	75	12.9	19.2	59.4	46.6
Roadside (RS)	49	14.2	21.1	50.3	74.6
Rural (RL)	23	8.4	14.6	68.4	19.1
Urban Centre (UC)	24	13.8	20.1	50.3	59.4
Urban Industrial (UI)	10	11.3	19.9	54.9	50.3
Suburban (SB)	19	15.1	22.8	57.4	45.6
Kerbside (KS)	6	20.0	30.1	28.3	134.1
Remote (RM)	5	NA	13.8	71.7	11.2
Airport (AP)	1	13.3	19.0	53.1	63.0
Overall average pollution	—	12.97	20.27	58.50	55.77

Note: All measurements are in $\mu g/m^3$ unit.

Spatio-temporal models

- General form of spatio-temporal model (Cressie and Wikle, 2011; Banerjee, Carlin and Gelfand, 2004):

$$\begin{aligned} \mathbf{Z}_t &= \mathbf{O}_t + \boldsymbol{\epsilon}_t, \\ \mathbf{O}_t &= \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t, \end{aligned} \tag{1}$$

- $\mathbf{Z}_t$ is the square-root of observed data from n sites.
- $\boldsymbol{\beta}$ is the regression parameter, $\mathbf{X}_t$ design matrix of covariates at time t.
- $\boldsymbol{\epsilon}_t$ follows multivariate normal with parameters $(0, \sigma_\epsilon^2 \mathbf{I}_n)$ independent of $\boldsymbol{\eta}_t$.
- $\boldsymbol{\eta}_t$ is space-time interaction term, modeled by a suitable Gaussian Process model.

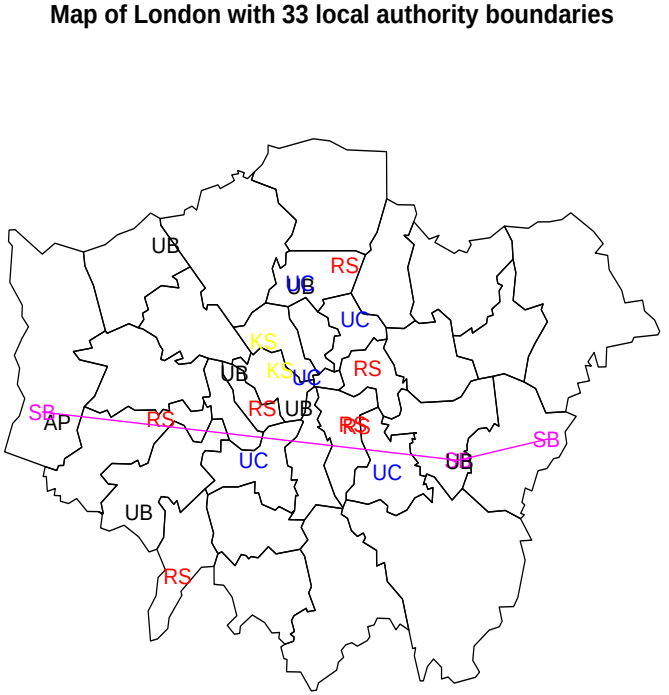
Anisotropic model

Not isotropic Correlation between sites not only distance dependent.

Wayout Correlation function needs to be anisotropic.

Option 1 Model based on Gaussian Predictive Process (GPP) (Banerjee et al., 2008) provides one such option.

Option 2 Spatial deformation using elliptical distance function leads to anisotropy (Schmidt and O'Hagan, 2003).



Space-time GPP model

- Same form as used by Sahu and Bakar (2012).
- The model is:

$$\mathbf{Z}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{A}\mathbf{w}_t + \boldsymbol{\epsilon}_t$$

- $\mathbf{A} = \mathbf{C}\mathbf{S}_w^{-1}$, $\mathbf{C}$ denoting the $n \times m$ cross-correlation matrix between the random effects at n observation locations and m knots, $s_1^*, \ldots, s_m^*$, $\mathbf{S}_w$ is the $m \times m$ correlation matrix of random affects $\mathbf{w}_t = \{w(s_1^*, t), \ldots, w(s_m^*, t)\}$.
- $\mathbf{w}_t$ is specified as :

$$\mathbf{w}_t = \rho\mathbf{w}_{t-1} + \boldsymbol{\eta}_t \tag{2}$$

- $\boldsymbol{\eta}_t \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ independently, $\Sigma_{\boldsymbol{\eta}} = \sigma_{\boldsymbol{\eta}}^2 S_{\boldsymbol{\eta}}$.
- $\Sigma_{\boldsymbol{\eta}}$ has dimension $m \times m$, can be chosen to be of much lower dimension than the same for two previous models GP and AR.
- $\mathbf{w}_0 \sim N(0, \sigma_0^2 \mathbf{S}_{\boldsymbol{\eta}})$.

spTimer model contd.

What we gain by using spTimer

1. A suitable space-time model for sparse data.
2. Easily implemented in the `R` package `spTimer` (Bakar and Sahu, 2014).
3. Properties of the model and forms of the posterior distributions developed in Sahu and Bakar (2012).

What we suffer from

- Knot locations $(s_1^*, \ldots, s_m^*)$ are fixed throughout.
- Cannot have point process models for knot locations.

Random knots

Previous attempts of modelling the knots

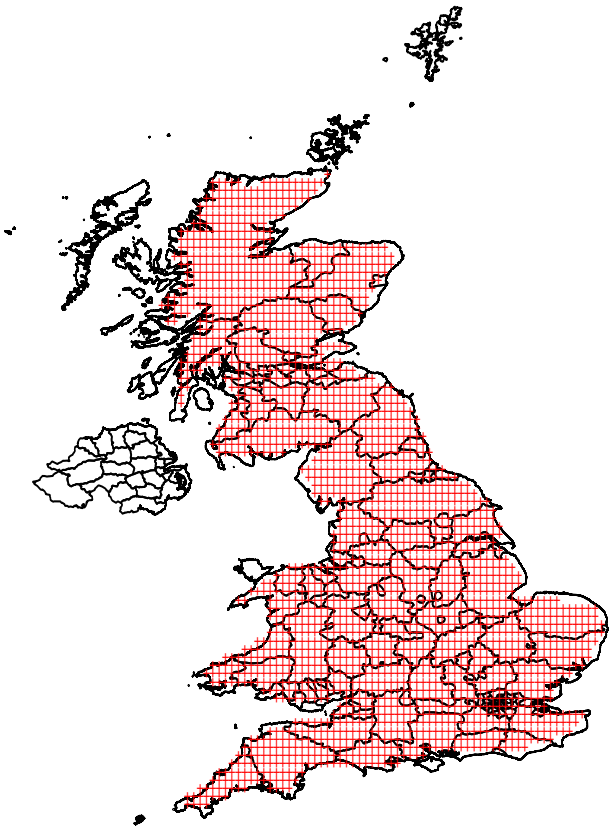
- Guhaniyogi et al. (2011) worked with fixed number of knots. But knot locations were assumed random according to an intensity function $f(s) = \exp(\lambda(s))$, where $\lambda(s)$ is taken as a suitable mixture model.
- Katzfuss (2012) used RJMCMC assuming the number of knots and their locations as random.
- Both the articles are based on spatial data only – not spatio-temporal.

Our extension

- We assume number of knots is fixed and domain of the knot locations lies on the set of grid locations for which AQUM output have been generated.
- Probability of a particular grid location being selected as knot is proportional to the population size of the local authority in

Why such a choice?

* AQUM values are available on the corners of the grid cells. Using them as knots adds more information to the model.
* Densely populated places should get more knots, since our main interest is to measure impact of pollution on human health.



Hierarchical space-time model

- Denoting Θ as set of hyperparameters, model $\mathbf{Z}_t$ is written as,

$$[\mathbf{Z}_t \mid S^*, \Theta] \quad = \quad \mathbf{X}_t\boldsymbol{\beta} + \mathbf{A}\mathbf{w}_t + \boldsymbol{\epsilon}_t$$

- $\boldsymbol{\epsilon}_t$ follows multivariate normal with parameters $(0, \sigma_\epsilon^2 \mathbf{I}_n)$ independently of $\boldsymbol{\eta}_t$.
- Exponential correlation function with decay ϕ is used to calculate Σ_η.
- At S^* we define $\mathbf{w}_t = \rho \mathbf{w}_{t-1} + \boldsymbol{\eta}_t$.
- $\boldsymbol{\eta}_t \sim N(0, \boldsymbol{\Sigma}_\eta)$ independently, $\Sigma_\eta = \sigma_\eta^2 S_\eta$.
- $\mathbf{w}_0 \sim N(0, \sigma_0^2 \mathbf{S}_\eta)$.
- Denoting D as sample space of all the grid cells on which AQUM is generated, N being total number of such grid cells, $p(\bar{s})$ as the population density of the local authority in which $\bar{s}$ lies, $\mathbf{p} = (p(\bar{s}_1), \ldots, p(\bar{s}_N))$,

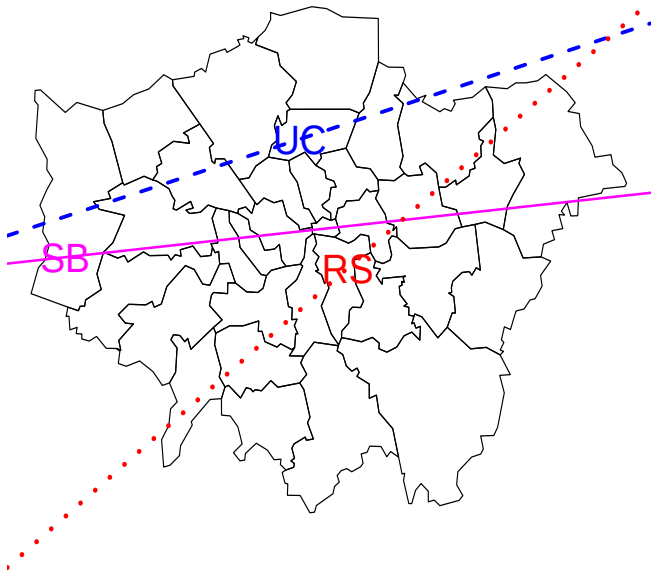
$$[S^* \mid \mathbf{p}] = \prod_{i=1}^{m} \frac{p(s_i^*)}{\int_D p(s) ds}$$

Linear Part

- We allow sitewise regression lines. If there are r many type of sites then $\mathbf{X}_t\boldsymbol{\beta}$ can be modeled as

$$\mathbf{X}_t\boldsymbol{\beta} = \sum_{k=0}^{r} \delta_k(s_i)(\gamma_{0k} + X(s_i, t)\gamma_{1k}),$$

  where $\delta_0(s_i) = 1$ for all s_i, $\delta_k(s_i) = 1$, if s_i is of k-th type of site, $k = 1, \ldots, r$, $\delta_k(s_i) = 0$, otherwise. $X(s_i, t)$ is AQUM value.
- Different regression lines can be obtained from this general form.
- When r=1, $\delta_1(s_i) = 1$, for all s_i, it leads to the same simple linear regression model for all site types.



- Thus the model is allowed to be different for different site types, i.e. one model for Urban sites, another for Kerbside etc.

Posterior distributions and simulation

- The joint posterior distribution is

$$\begin{aligned}
&\log\left([S^*, \Theta, \mathbf{w}, \mathbf{Z}]\right) \propto \\
- &\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathbf{O}_t - \mathbf{A}\mathbf{w}_t)'(\mathbf{Z}_t - \mathbf{O}_t - \mathbf{A}\mathbf{w}_t) \\
- &\sum_{t=1}^{T}\frac{1}{2\sigma_{\boldsymbol{\eta}}^2}(\mathbf{w}_t - \rho\mathbf{w}_{t-1})'\Sigma_{\boldsymbol{\eta}}^{-1}(\mathbf{w}_t - \rho\mathbf{w}_{t-1}) \\
- &\frac{1}{2\sigma_0^2}\mathbf{w}_0'\Sigma_{\boldsymbol{\eta}}^{-1}\mathbf{w}_0 - (T+1)\log(|\Sigma_{\boldsymbol{\eta}}|) + \log(\pi(S^*)) + \log(\pi(\Theta))
\end{aligned}$$

- Number of knots m is taken to be fixed. Chosen using validation mean square error.
- Posterior distributions of Θ, $\mathbf{w}$ given S^* are same as those in the spTimer model. Can be updated using `spTimer`.
- Updated S^* using Metropolis-Hastings algorithm.

Incorporating anisotropy through deformation

- We have taken distance function like $d(s_i, s_j) = (s_i - s_j)' B (s_i - s_j)$, where B is Positive Definite.
- B can be written as TT'.
- Form of T taken as
$$\begin{pmatrix} 1 & 0 \\ \varphi & 1 \end{pmatrix}$$
- φ is assumed follow uniform distribution in (0, 1).
- Posterior of φ is of the same form as S^* with $\pi(S^*)$ replaced by $\pi(\varphi)$. Can be updated in the same way as S^*.

Catalogue of fitted models

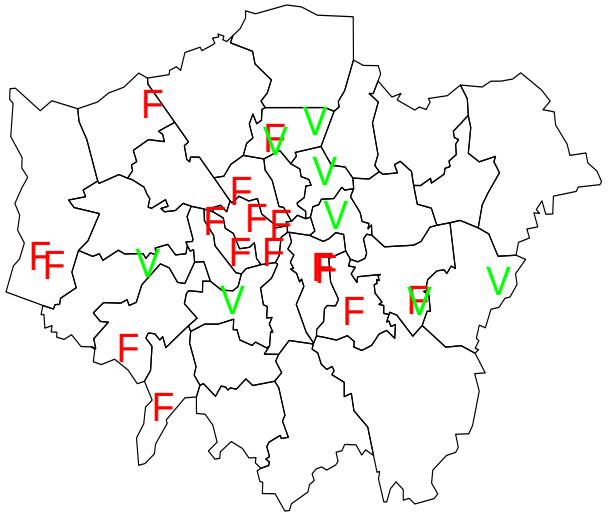
Model	Linear Part	Knots	Time series	Spatial Process
Model-1	AQUM	not required	Independent	GP
Model-2	AQUM	not required	AR process	AR
Model-3	AQUM	fixed	AR process	GPP
Model-4	Sitewise Linear	fixed	AR process	GPP
Model-5	AQUM	random	AR process	GPP
Model-6	Sitewise Linear	random	AR process	GPP
Model-7	AQUM	fixed	AR process	GPP (anisotropic)

Choice of prior of hyperparameters

- The parameter representing mean, e.g, $\boldsymbol{\beta}$, ρ, assumed to follow normal distribution with mean $(\mu_{\boldsymbol{\beta}}, \mu_\rho)$ and variance $(\delta^2_{\boldsymbol{\beta}}, \delta^2_\rho)$.
- We choose $\mu_{\boldsymbol{\beta}} = \mu_\rho = 0$; variance equal to 10^4.
- Variance parameters like σ^2_ϵ are assumed to follow Inverse Gamma distribution with shape parameter as 1 and scale parameter as 2 to have proper prior specification.
- The prior distribution for ϕ is taken as Gamma(a, b) where a and b estimated by empirical Bayes (EB).

Validation of the models

1. Among the n many monitoring sites we choose at least 10% at random. Denote those as validation sites.
2. Pretend that data at validation sites have not been observed and need to be predicted.
3. Use rest of the data for fitting a model.
4. Predict the pollution values at the validation sites and calculate the RMSE by comparing with the observed data of those sites.
5. The model with the least RMSE is the most suitable one.



- All measurements are in μgm per m^{-3}.

Results for validating data from whole of UK

Table: Root Mean Square Error (18 validation sites and 148 fitting sites)

	$PM_{2.5}$	PM_{10}	Ozone	NO_2
SD	9.55	12.0	21.82	38.06
AQUM (raw)	8.03	14.27	19.49	36.55
kriging	5.36	9.18	18.98	38.28
Model-1	5.21	8.77	16.27	34.5
Model-2	5.26	9.10	16.9	45.3
AAQUM (raw)	5.0	8.02	13.63	32.36
Model-3	4.78	7.67	12.56	24.99
Model-4	4.79	7.59	12.58	25.3
Model-5	4.79	7.59	12.58	25.3
Model-6	4.77	7.55	12.65	26.91
Model-7	4.81	7.60	12.49	27.15

Results for validating data from London

Table: Root Mean Square Error (8 validation sites and 21 fitting sites)

	$PM_{2.5}$	PM_{10}	Ozone	NO_2
SD	9.82	13.40	23.97	46.84
kriging	9.69	16.75	18.74	39.07
AQUM (raw)	8.46	13.65	18.53	33.37
Model-1	5.71	6.90	14.77	35.18
Model-2	3.95	4.90	14.11	32.37
AAQUM (raw)	4.16	5.29	13.51	27.45
Model-3	3.47	3.80	12.75	24.99
Model-4	3.62	3.73	12.66	21.97
Model-5	3.47	3.79	10.05	21.22
Model-6	3.62	3.73	10.63	21.92
Model-7	3.47	3.78	12.78	24.21
Pirani et al.	4.75	—	—	—

Further validation results.

- We also compare the best statistical Model-6 with the AAQUM outputs using:
  - one site at a time leave-out cross-validation RMSE.
  - But we only validate the sites with at least 30% observations to have stable RMSE.

Table: 115 RMSEs for NO_2 in the UK

Models	Minimum	Mean	SD	Maximum
RMSE AQUM	8.07	33.60	21.24	134.59
RMSE AAQUM	6.19	28.23	21.09	131.13
RMSE model-6	10.28	22.69	12.55	87.43

Summary of Cross-validation RMSEs for London data

17 RMSEs for NO_2				
Models	Minimum	Mean	SD	Maximum
RMSE AQUM	20.22	43.66	32.18	134.59
RMSE AAQUM	14.81	37.36	33.21	131.02
RMSE Model-6	15.60	33.44	25.32	97.53
12 RMSEs for Ozone				
RMSE AQUM	15.71	19.02	7.91	39.02
RMSE AAQUM	9.17	12.53	7.39	32.05
RMSE model-6	6.36	11.02	7.69	31.60
8 RMSEs for PM_{10}				
RMSE AQUM	11.36	15.95	5.82	28.05
RMSE AAQUM	4.03	8.18	5.83	19.35
RMSE model-6	3.70	6.98	4.97	16.87
7 RMSEs for $PM_{2.5}$				
RMSE AQUM	7.56	5.80	0.85	9.94
RMSE AAQUM	3.17	3.91	0.58	4.53
RMSE model-6	3.07	3.59	0.36	4.97

Improvement achieved

Improvement over the AAQUM.

Pollutant	London	UK
NO_2	10.5%	21%
$PM10$	14.7%	–
Ozone	12.05%	–
$PM_{2.5}$	8%	–

Comparison with *Pirani et al.* (2014)

Improvement is about 27%, although for different data sets of PM_{10}.

Conclusions

1. AQUM outputs are accurate but clearly need to be adjusted on the basis of observed data.
2. AQUM outputs are improved by site-wise and pollutant-wise adjustments as detailed in the previous talk.
3. These adjustments also use 1 kilometer background pollution map and further information, which implies that actual AURN observations may have been used many times over which in turn makes it harder to assess uncertainty in the predictions.
4. Our process based statistical models further improve the AAQUM outputs (by about 8-21% reduction in RMSE) as shown by the leave one out crossvalidation study.
5. Statistical models have the added advantage of producing the correct prediction uncertainty in air pollution estimates, which are required for the health outcome model.

Discussion

1. We have proposed a number of nonstationary, anisotropic models which worked well for all four important pollutants.
2. Introduced distribution for knot locations using population density surface.
3. Parsimonious model enabled by AQUM.
4. We are able to measure long term exposure since we have modelled daily data for 5 year period.
5. Our prediction uncertaintites are exactly correct as those have been asessed by a statistical model where all our assumptions regarding the data and the computer model output are explicitly stated.

THANK YOU