Estimating Overall Air Quality and its Effect on Human Health

Powell, Helen
University of Glasgow, School of Mathematics and Statistics
15 University Gardens
Glasgow (G12 8QQ), United Kingdom
E-mail: h.powell.1@research.gla.ac.uk

Lee, Duncan
University of Glasgow, School of Mathematics and Statistics
15 University Gardens
Glasgow (G12 8QQ), United Kingdom
E-mail: Duncan.Lee@glasgow.ac.uk

1. Background

1.1 Air pollution and health studies

The adverse health effects associated with ambient air pollution can be estimated using ecological time series studies, which comprise daily data for the population living with an extended urban area. The response data, $\mathbf{y} = (y_1, \ldots, y_n)_{n\times 1}$, are daily counts of mortality or morbidity outcomes, which are realter to air pollution concentrations and m other explanatory variables, $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)_{n\times m}$, the latter of which can include a smooth function of time and a measure of temperature.

The concentrations of a number of pollutants are measured daily by a network of q monitors located within the study region. Thus for a given pollutant i, there is an $n \times q$ matrix of observations, $W_i = (\mathbf{w}_{i,1}, \ldots, \mathbf{w}_{i,n})$. The observations for day t are denoted by $\mathbf{w}_{i,t} = (w_{i,t}(\mathbf{s}_1), \ldots, w_{i,t}(\mathbf{s}_q))$, where $(\mathbf{s}_1, \ldots, \mathbf{s}_q)$ are the coordinates of the monitoring sites. Most studies only consider the health impact of a single pollutant, and as the health data relate to the whole region, a representative value for the entire region is required for each day. This representative value is typically estimated by

$$\bar{X}_{i,t} = \frac{1}{q}\sum_{l=1}^{q} w_{i,t}(s_l), \tag{1}$$

the average value from the q monitoring sites (see for example Lee and Shaddick (2008)). The health effects of this simplistic single day estimate are typically estimated from a Poisson generalized linear model. A general form of this model is

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \quad \text{for } t = 1, \ldots, n, \\ \ln(\mu_t) &= \mathbf{z}_t^{\mathrm{T}}\boldsymbol{\alpha} + g(\bar{X}_{i,t-k}), \end{aligned} \tag{2}$$

where the regression parameters for the non-pollution covariates are denoted by $\boldsymbol{\alpha}$. The representative value of pollution has been lagged by k days and, the shape of the relationship between air pollution and health is represented by the function $g(\cdot)$. Although the majority of studies use an approach similar to that described above, it has a number of deficiencies, and in this paper we develop a more spatially representative measure of air pollution than that given by (1).

1.2 A spatially representative measure of air pollution

Considering a single pollutant i, the monitor average has a number of deficiencies as a spatially representative measure of air pollution. Firstly, it is likely that some of the locations of the pollution monitors may have been chosen by preferential sampling, meaning that they are purposely placed at sites with high pollution concentrations (Loperfido and Guttorp (2008)). It is therefore likely that equation (1) will overestimate the true average concentration. This overestimation will be compounded by the fact that a number of the monitors will be located next to main roads, which will lead to higher concentrations being recorded compared to the non-roadside monitors (where most people live). The monitor average also does not take into account the population density across the study region, therefore, (1) may not directly relate to a sizeable proportion of the population.

Instead, for a single pollutant i, we believe that the appropriate exposure measure is the average level of pollution to which the population were exposed. On day t, this can be approximated by

$$X_{i,t} \approx \sum_{j=1}^{N} P(\mathbf{s}_j^*) X_{i,t}(\mathbf{s}_j^*), \tag{3}$$

where $\mathbf{s}^* = (\mathbf{s}_1^*, \ldots \mathbf{s}_N^*)$ form a regular grid covering the entire study region, and to preserve scale $\sum_{j=1}^{N} P(s_j^*) = 1$.

Finally, the desired pollution measure for a region is inherently an unknown quantity, and hence the uncertainty in any estimate should be allowed for when estimating its health effects. Therefore, the aim of this paper is to: (i) produce a spatially representative measure of overall air quality; and (ii) incorporate it into a health model, taking account of the uncertainty in the model.

2. Methods

We propose a three stage approach for estimating the overall effects of air quality on human health, which addresses the limitations of the standard approach outlined in the previous section. The first stage describes the estimation of (3) for a single pollutant, the second combines these spatially representative values into an overall index of air quality, while the third estimates its effects on health.

2.1 Pollution model (single pollutant)

We propose estimating the approximation of $X_{i,t}$, given by equation (3), separately for each day, using a Bayesian geostatistical model. Inference is implemented using the geoR, (Ribeiro Jr. and Diggle (2001)), add on package for the statistical programme R, (R Development Core Team (2011)), which uses direct simulation rather than Markov chain Monte Carlo (MCMC) methods. For day t, we model the concentrations of a generic pollutant, $\mathbf{w}_t = (w_t(\mathbf{s}_1), \ldots, w_t(\mathbf{s}_q))$ (the subscript i has been dropped for notational simplicity) at q monitoring sites, as

$$\begin{aligned}
\ln(\mathbf{w}_t) &\sim \mathrm{N}(\mathbf{B}_t\boldsymbol{\gamma}, \sigma^2 V(\phi, \nu^2)), \\
\boldsymbol{\gamma} &\sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}), \\
f(\sigma^2) &\propto 1/\sigma^2, \\
\phi &\sim \text{Discrete Uniform}(a_1, \ldots, a_r), \\
\nu^2 &\sim \text{Discrete Uniform}(b_1, \ldots, b_f).
\end{aligned} \tag{4}$$

The data are modeled on the log scale as suggested by Ott (1990). The spatial trend is represented by $\mathbf{B}_t\boldsymbol{\gamma}$, where $\boldsymbol{B}_t$ is an $n \times p$ matrix of covariates, and $\boldsymbol{\gamma}$ are the associated regression parameters. These parameters are assigned a weakly informative multivariate Gaussian prior, with a large variance and a diagonal correlation matrix, i.e. $\Sigma_\gamma = \sigma^2_\gamma I$.

The covariance structure of the model is represented by $\sigma^2 V(\phi, \nu^2) = \sigma^2(R(\phi) + \nu^2 I)$, which combines spatially structured correlation (via $R(\phi)$) with measurement error (via $\nu^2 I$). The variance parameter σ^2 is given a functional flat prior on the log scale (Diggle and Ribeiro Jr (2007)), i.e. $f(\log(\sigma)) \propto 1$ which is equivalent to $f(\sigma^2) \propto 1/\sigma^2$. The spatial correlation matrix is denoted by $R(\phi)$, and is modeled by the Matern class of functions with smoothness parameter $\kappa = 1.5$. The parameter ϕ represents the range of spatial correlation, and is given a discrete uniform prior for computational efficiency. Finally, ν^2 is the noise to signal ratio, and is also assigned a discrete uniform prior distribution for the same computational reasons as described for ϕ.

Using direct simulation we can generate J samples $\Theta^{(j)} = (\boldsymbol{\gamma}^{(j)}, \sigma^{2(j)}, \phi^{(j)}, \nu^{2(j)})$ from the posterior distribution of (4). Conditional on each set of samples $\Theta^{(j)}$, Bayesian Kriging can be used to predict the pollution surface at a set of prediction locations, $\mathbf{s}^* = (\mathbf{s}^*_1, \ldots \mathbf{s}^*_N)$, which form a regular lattice of points over the study region $\mathcal{R}$. Each set of predictions can then be exponentiated back to their original scale and weighted by the associated population density, thus producing J samples from the posterior predictive distribution of (3).

2.2 Air quality indicator

Air pollution is made up of a complex mixture of numerous pollutants, and it is more realistic to estimate the health effects of a measure of overall air quality, rather than that of a single pollutant. The Bayesian geostatistical model can be applied separately to each of F pollutants, providing that each one is measured at enough locations to make a geostatistical analysis feasible. These posterior predictive distributions can then be combined by constructing an air quality indicator (AQI, see for example Bruno and Cocchi (2002) and Lee et al. (2011)), which is a synthetic index of overall air quality. However, each pollutant will have a different level of temporal variation, so must be transformed onto a common scale. Therefore, we apply a simple linear re-scaling to the J estimates of (3) for each pollutant, so that each one has mean zero and standard deviation one. From these standardized values the posterior distribution of the AQI can be constructed as

$$\text{AQI}_t^{(j)} = \frac{1}{F}\sum_{i=1}^{F} \frac{X_{i,t}^{(j)} - \mu_i}{\sigma_i} \qquad \text{for } \; j = 1, \ldots, J, \tag{5}$$

where μ_i and σ_i are the pollutant specific mean and standard deviations used in the re-scaling. Thus, the above equation produces J samples from the posterior predictive distribution of the air quality indicator, conditional on each set of observed pollution data $(\mathbf{w}_{1,t}, \ldots, \mathbf{w}_{F,t})$.

2.3 Health model

The health model we propose differs from (2), because it allows for the uncertainty in the posterior predictive distribution of the AQI given by (5). This can be implemented by utilizing a Bayesian approach, where the air pollution concentration is randomly generated from its posterior predictive distribution at each MCMC iteration. The model we propose is given by

$$
\begin{aligned}
Y_t &\sim \text{Poisson}(\mu_t) \qquad \text{for } t = 1, \ldots, n, \\
\ln(\mu_t) &= \mathbf{z}_t^T \boldsymbol{\alpha} + \text{AQI}_t \beta, \\
\alpha_j &\sim \text{N}(0, 10) \qquad \text{for } j = 1, \ldots, m, \\
\beta &\sim \text{N}(0, 10), \\
\text{AQI}_t &\sim f(\text{AQI}_t | \mathbf{w}_{1,t}, \ldots, \mathbf{w}_{F,t}). \qquad (6)
\end{aligned}
$$

The regression parameters $(\alpha_1, \ldots, \alpha_m, \beta)$ are assigned diffuse Gaussian priors, with mean zero and a large variance of 10. Inference for this part of the model is based on MCMC simulation, where the parameters are updated in three steps, namely: $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, β and $\{\text{AQI}_t\}_{t=1}^n$. The vector $\boldsymbol{\alpha}$ is updated in blocks via a Metropolis step, using a random walk proposal distribution with a diagonal variance matrix. The coefficient β is updated singularly, but also by a Metropolis step. The overall pollution concentrations, $\{\text{AQI}_t\}_{t=1}^n$, are generated by randomly sampling from its posterior predictive distribution given the observed pollution data.

3. London application

3.1 Data

The data relate to the area of Greater London (roughly the area within the orbital M25 motorway), and comprise daily measurements of air pollution, population health, for the over 65s, and meteorology for the 3 year period spanning 2001 to 2003.

The pollution data comprise daily measurements of carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3) and particulate matter (PM_{10}). In addition, we also have a number of potential covariates for the geostatistical model including, the local environment in which the monitor is located, the locations of the pollution monitors and, a modeled estimate of the yearly average concentrations of CO, NO_2 and PM_{10}, on a 1 kilometer square grid across Greater London (similar data for O_3 were not available). Finally, we have the population density associated with each of these 1km locations. The health data comprise the daily counts of the total numbers of respiratory related deaths, in Greater London. Aggregate daily temperature levels were also obtained.

3.2 Model building

We specified the spatial trend for the pollution model to include an indicator term, for whether the monitor had been placed at a roadside or background locale, and the modeled pollution estimates which were closest to each monitoring site. We predicted concentration levels for each of our single pollutants at every 2km location and specified that all of these locations would be at non-roadside environments.

In addition to a measure of pollution the covariates in the health model also included the average daily temperature and a smooth function of time. We specified a quadratic relationship between temperature and respiratory related deaths as there were some very hot days in Greater London. The inclusion of the quadratic rather than linear term also produced a smaller AIC value. We represented the prominent seasonal pattern in the residuals by a natural cubic spline of time (day of the study). Our choice of 7 knots per year, 21 in total, was made by minimizing AIC. The pollution concentration estimates were also lagged by one day.

3.3 Results

We fitted both a standard generalized linear model, given by equation (2), which included the monitor average as the measure of pollution, and the Bayesian health model given by (6), which was updated at each iteration with one of the posterior predictive AQI estimates. We compared the relative risk for a one standard deviation increase in the pollutant and associated 95% confidence and credible intervals for both models. Under the Bayesian model the credible intervals are wider as the uncertainty in the pollution concentrations has been accounted for. In the case of CO this wider interval has meant that the relative risk for a one standard deviation increase in CO is no longer significant. The widest interval occurs for the AQI as the uncertainty in the four pollutants has been cumulated into the one measure.

Table 1. The relative risk (RR) and associated 95% confidence and credible intervals for both the monitor average and our proposed spatial average.

Monitor Average			*Spatial Average*		
Pollutant	**RR**	**95% CI**	**Pollutant**	**RR**	**95% CI**
CO	1.013	(1.000,1.027)	CO	1.009	(0.995,1.025)
NO_2	1.012	(0.999,1.026)	NO_2	1.011	(0.996,1.027)
O_3	1.018	(1.000,1.037)	O_3	1.023	(1.004,1.043)
PM_{10}	1.009	(0.995,1.023)	PM_{10}	1.014	(0.999,1.034)
AQI	1.020	(1.005,1.034)	AQI	1.022	(1.006,1.043)

4. Future Work

In the future we hope to extend the Bayesian geostatistical model to a multivariate model. This would allow us to predict pollution concentrations based on the current observations of many other pollutants as opposed to only those of the pollutant we are trying to predict.

REFERENCES

Bruno, F. and D. Cocchi (2002). A unified strategy for building simple air quality indices. *Environmetrics 13*, 243 – 261.

Diggle, P. and P. Ribeiro Jr (2007). *Model-based geostatistics* (1st ed.). Springer Series in Statistics.

Lee, D., C. Ferguson, and E. Scott Marian (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society, Series A 174*, 109 – 126.

Lee, D. and G. Shaddick (2008). Modelling the effects of air pollution on health using Bayesian dynamic generalised linear models. *Environmetrics 19*, 785–804.

Loperfido, N. and P. Guttorp (2008). Network bias in air quality monitoring design. *Environmetrics 19*, 661 – 671.

Ott (1990). A physical explanation of the log normality of pollutant concentrations. *Journal of Air Waste Management Association 40*, 1378 – 1383.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ribeiro Jr., P. and P. Diggle (2001). geoR: a package for geostatistical analysis. *R-NEWS 1*(2), 15–18.