ENVIRONMETRICS
Environmetrics (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.889

Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach

T. R. Fanshawe[1*,†], P. J. Diggle[1], S. Rushton[2], R. Sanderson[2], P. W. W. Lurz[2], S. V. Glinianaia[3], M. S. Pearce[3,4], L. Parker[5], M. Charlton[6] and T. Pless-Mulloli[3]

[1]*Department of Medicine, Lancaster University, U.K.*
[2]*Institute for Research on Environment and Sustainability, Newcastle University, U.K.*
[3]*Institute of Health and Society, Newcastle University, U.K.*
[4]*School of Clinical Medical Sciences, Newcastle University, U.K.*
[5]*Community Health and Epidemiology/Pediatrics, Dalhousie University, Halifax, Canada*
[6]*National Centre for Geocomputation, National University of Ireland, Ireland*

SUMMARY

Studies investigating associations between air pollution exposure and health outcomes benefit from the estimation of exposures at the individual level, but explicit consideration of the spatio-temporal variation in exposure is relatively new in air pollution epidemiology. We address the problem of estimating spatially and temporally varying particulate matter concentrations (black smoke = BS = PM_4) using data routinely collected from 20 monitoring stations in Newcastle-upon-Tyne between 1961 and 1992. We propose a two-stage strategy for modelling BS levels. In the first stage, we use a dynamic linear model to describe the long-term trend and seasonal variation in area-wide average BS levels. In the second stage, we account for the spatio-temporal variation between monitors around the area-wide average in a linear model that incorporates a range of spatio-temporal covariates available throughout the study area, and test for evidence of residual spatio-temporal correlation. We then use the model to assign time-aggregated predictions of BS exposure, with associated prediction variances, to each singleton pregnancy that occurred in the study area during this period, guided by dates of conception and birth and mothers' residential locations. In work to be reported separately, these exposure estimates will be used to investigate relationships between maternal exposure to BS during pregnancy and a range of birth outcomes. Our analysis demonstrates how suitable covariates can be used to explain residual spatio-temporal variation in individual-level exposure, thereby reducing the need to model the residual spatio-temporal correlation explicitly. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: dynamic linear model; environmental epidemiology; exposure estimation; particulate matter; spatio-temporal process

1. INTRODUCTION

Links between long-term or short-term exposure to particulate matter and morbidity or mortality in both children and adults are now well established (Pope III and Dockery, 2006). In particular, there is growing evidence of an association between air pollution exposures during pregnancy and adverse birth outcomes

*Correspondence to: T. R. Fanshawe, Department of Medicine, Lancaster University, U.K.
†E-mail: t.fanshawe@lancaster.ac.uk

Received 16 July 2007
Accepted 9 August 2007

(Glinianaia *et al*., 2004b; Sram *et al*., 2005) or infant survival (Glinianaia *et al*., 2004a; Ritz *et al*., 2006), especially for respiratory-related causes (Woodruff *et al*., 2006). In order to test hypotheses related to these associations using observational data, estimates of pollution levels to which each mother was exposed during different periods of pregnancy are needed. Some previous studies have either assumed homogeneity of exposure at any one time across large geographical areas (Woodruff *et al*., 1997; Samet *et al*., 2000), or estimated exposure using a crude average (Bobak, 2000). Only recently has modelled city-wide variation in exposure and its impact on health outcomes been considered (Jerrett *et al*., 2005).

In this paper, we use data from the UK Particulate Matter and Perinatal Events Research (PAMPER) study to demonstrate a method for estimating a spatio-temporal exposure surface of black smoke (BS), equivalent to PM_4, concentrations over the city of Newcastle-upon-Tyne for the time period 1961–1992.

We consider data in the form of a set of time series, one for each of a number of monitoring locations within the spatial region of interest, and not necessarily providing data at a common set of times. Various approaches have been suggested in the statistical literature for analysing environmental spatio-temporal data of this kind; for reviews, see Kyriakidis and Journal (1999) and Sahu and Mardia (2005). Key approaches to such analysis include: directly modelling the joint space-time distribution of the observations, treating time as an additional dimension (e.g. Brown *et al*., 2001), modelling the data as a set of spatial processes correlated in time (e.g. Bogaert and Christakos, 1997) or, more commonly, as a set of time series correlated in space (e.g. Meiring *et al*., 1998).

Most of this work has used Gaussian processes as models for the underlying spatio-temporal phenomenon, $S(x, t)$ say, with a consequent focus on the specification of valid, appropriate and computationally tractable covariance functions for $S(x, t)$ (Gneiting *et al*., 2007). An exception is Higdon (2007), who describes a non-Gaussian kernel convolution approach. Stroud *et al*. (2001) extend state-space models of time series to the space-time domain in order to avoid making assumptions of stationarity and separability of the covariance function. In some examples, the relatively weak dependence between observations either in space or in time has enabled the modelling process to be simplified: for example, Handcock and Wallis (1994) found a lack of temporal dependence in annual winter average temperatures in northern U.S.A. In contrast, other examples exhibit long-term temporal dependence, such as the Irish wind speed data of Haslett and Raftery (1989).

In the field of air pollution, several authors have addressed the simultaneous consideration of spatial and temporal variations of exposure. Carroll *et al*. (1997) modelled ozone exposure in Texas, U.S.A. by splitting the spatio-temporal variation into two components: a deterministic, spatially constant component and a stationary, zero-mean Gaussian random field. Zidek *et al*. (2002) modelled the spatial covariance between residuals using a space deformation approach (Meiring *et al*., 1998) after first fitting an AR(3) model to hourly PM_{10} levels in Vancouver, Canada (Li *et al*., 1999). Sahu *et al*. (2006) illustrated one way in which available covariates may be used by modelling $PM_{2.5}$ monitoring data using two random spatio-temporal processes, corresponding to urban and rural areas respectively, and weighted by population density.

In this paper, we demonstrate a pragmatic, two-stage modelling strategy. We first estimate the seasonally varying temporal trend using a dynamic linear model, then account for remaining spatio-temporal variation using temporally and/or spatially varying covariates. We demonstrate that for our data, residual spatio-temporal correlation is not significant. In principle, we could include a spatio-temporally correlated residual term, at the cost of a substantial increase in computational complexity. However, in our view explicit models of spatio-temporal correlation should be used only when the possibility of obtaining an adequate explanation of spatio-temporal variation using covariate information has been exhausted. In our application, the key step was not to rely on routinely available covariate information but instead to construct a suitable surrogate using a combination of land-use information and digital

DOI: 10.1002/env

images of domestic chimneys which, for the area and time-period in question, constituted a major source of BS exposure for pregnant women.

2. THE UK PAMPER STUDY

The UK PAMPER study is a historical cohort study to investigate the relationship between adverse pregnancy outcomes and a range of socio-economic, meteorological and pollution-related factors. In this paper, we model levels of weekly BS using data routinely, albeit spasmodically, recorded at 20 air pollution monitoring stations within the city of Newcastle-upon-Tyne (the 'study area') between October 1961 and December 1992 (the 'study period'). The data are available from the UK Air Quality Archive (http://www.airquality.co.uk/archive/data_and_statistics_home.php). Figure 1 shows the locations of the 20 monitoring stations within the study area, and the locations of five further monitoring stations that we will use for model validation. Pless-Mulloli *et al*. (personal communication) provide more details of the study's background and setting.

Figure 2 shows the period of time over which each monitor was in operation ('active'). Over the whole study period, the number of monitors active during any single week varied from three to ten.

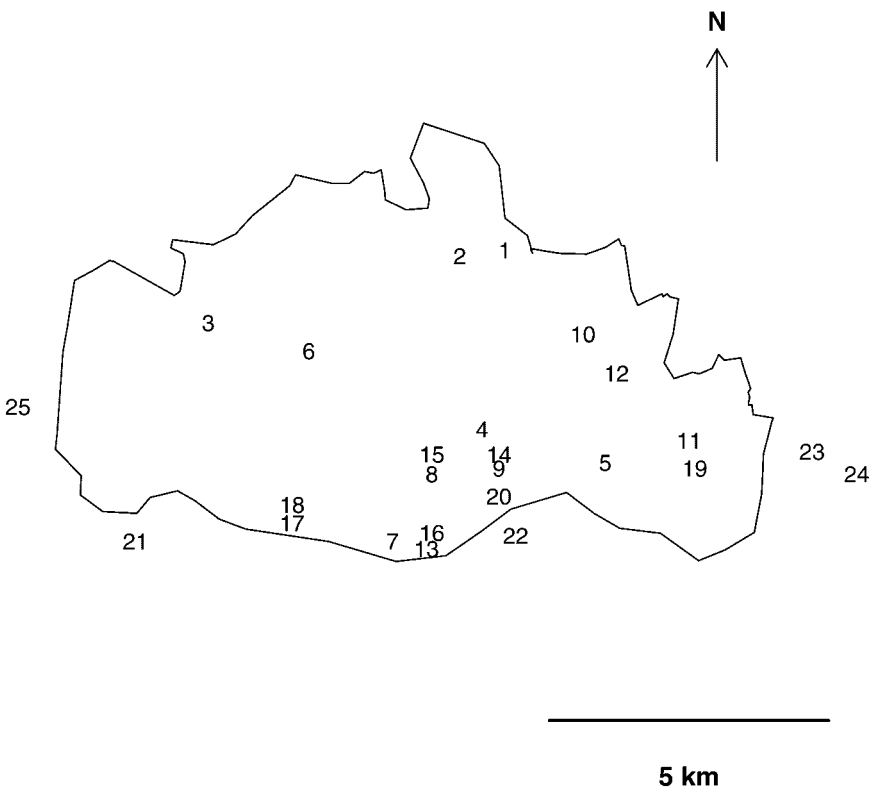

Figure 1. Outline of the PAMPER study area, Newcastle-upon-Tyne. Locations of black smoke monitoring stations used for modelling are numbered 1–20; those used for validation are numbered 21–25

DOI: 10.1002/env

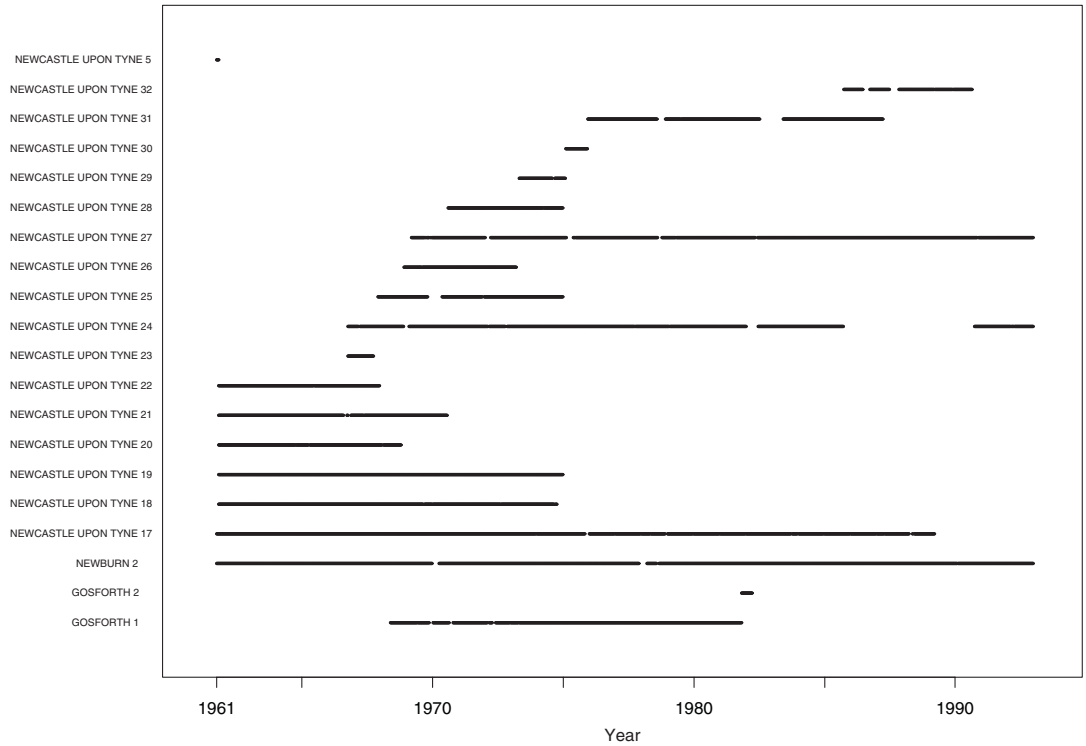

Figure 2. Diagram showing PAMPER monitoring station activity. Periods of activity are indicated by a black line

In our experience, the relatively sparse spatial coverage of the study area by monitors is typical, and strongly influenced our approach to the prediction problem.

Our aim is to attach to each of the 109 086 singleton births that occurred in the study area during the study period, a predicted BS exposure level and associated prediction variance, both for individual weeks of the pregnancy and time-aggregated over months, trimesters and over the whole pregnancy period. Each birth is characterised by the date of birth, the estimated date of conception (for births with available gestational age) and the mother's residential location (grid reference) at which BS levels are to be estimated. In future work, we will investigate associations between this modelled exposure and a range of adverse birth outcomes, including birthweight, low birthweight, preterm birth, stillbirth, infant mortality and congenital abnormality.

3. MODELLING THE EXPOSURE SURFACE

3.1. Exploratory analysis

In common with other environmental applications (e.g. Brown *et al*., 2001; Zidek *et al*., 2002), we found that a log-transformation approximately stabilises the variance of BS and gives a roughly linear time trend, i.e. city-wide average BS levels have experienced an approximately exponential decline over the study period. We therefore model the log-transformed values of BS recorded at each monitoring station.

DOI: 10.1002/env

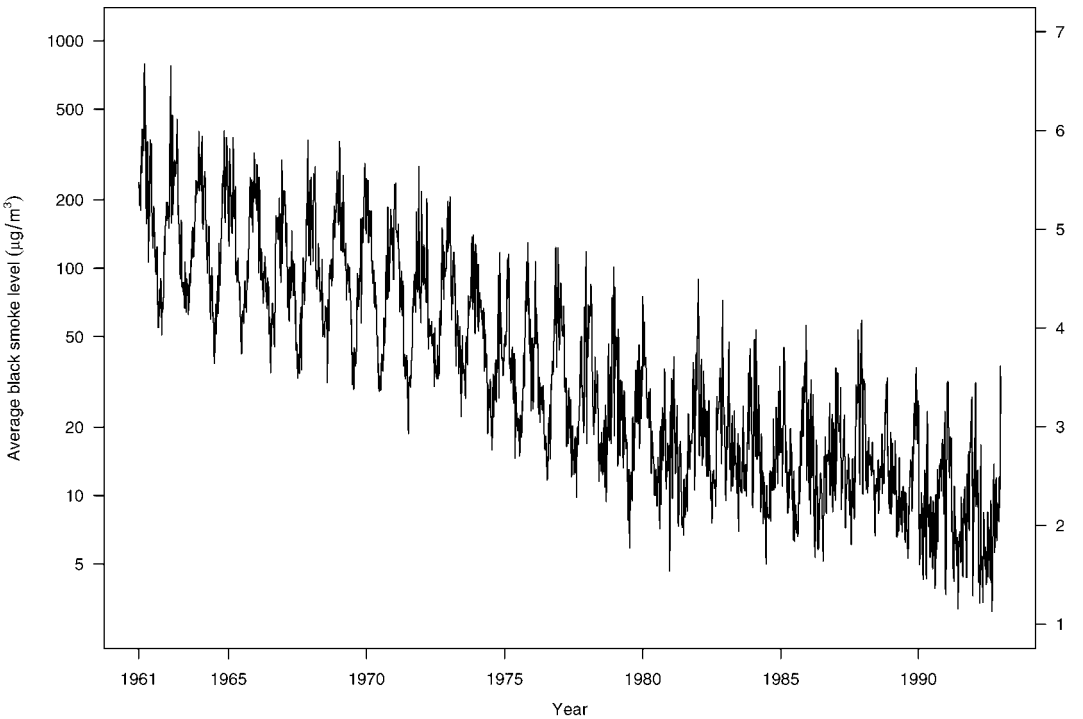

Figure 3. Area-wide weekly average black smoke levels, plotted as a time series. The original scale is shown on the left vertical axis and the logarithmic scale on the right vereical axis

Let Y denote log-transformed BS. Figure 3 shows the area-wide average, $\bar{Y}_t$ say, in each of the 1631 weeks of the study period, in each case calculated as the average of the observed log-BS levels at all monitoring stations that were active during the week in question. The scale of the overall temporal variation in $\bar{Y}_t$ is much larger than is the spatial variation between different monitors at any given time, which is typically of the order of 1 unit on the logarithmic scale, although occasional recorded values fall much further than this from the corresponding city-wide average. For the subsequent modelling, we use all available data. Re-fitting the final model excluding 74 recorded values (out of 10 174, i.e. around 0.7%) of log-BS more than 1.5 units away from the area-wide average has only a small impact on parameter estimates and predictions, and as we have no basis for treating these values as recording errors, we retain all of the data in the analysis presented below.

Figure 3 also shows that there is a strong seasonal component to average BS levels. Annual peaks and troughs occur each winter and summer respectively, albeit with some variation from year to year. This seasonal pattern is also evident from inspection of the data from individual monitors.

3.2. The modelling strategy

Our strategy is first to model the expectation of the area-wide weekly average log-transformed BS levels, $\mu_t = \mathrm{E}[\bar{Y}_t]$, ignoring any spatial variation. This results in an estimate $\hat{\mu}_t$. We then use spatio-temporally referenced covariates $\mathbf{w}$ to account for residual variation between monitors. Hence, if t denotes week

DOI: 10.1002/env

and x geographical location, we model log-transformed BS, $Y_t(x)$, as

$$Y_t(x) = \hat{\mu}_t + \mathbf{w}^T \boldsymbol{\beta} + Z_t(x) \tag{1}$$

where $Z_t(x)$ is a residual term which may or may not exhibit temporal and/or spatial correlation, and $\hat{\mu}_t$ is treated as an offset, provided that its associated prediction variance is negligible. Note that in Equation (1), time is treated as discrete, with a resolution of 1 week, whereas x is treated as a spatial continuum, and that $\mathbf{w}$ depends implicitly on t and x. This framework acknowledges that, although our data are confined to a discrete set of monitor locations, our aim is to predict BS at every maternal residence within the study area.

Our two-stage modelling strategy is informed by two considerations. Firstly, the exploratory analysis showed that the temporal variation in $Y_t(x)$ dominates the residual spatio-temporal variation. Secondly, and not untypically (cf de Luna and Genton, 2005), our data are temporally rich but spatially sparse. Together, these features enable relatively precise estimation of the spatially constant component μ_t. Other authors have preferred to fit different models to the individual time series obtained from each monitor, treating periods of inactivity as missing data (Haslett and Raftery, 1989; Meiring *et al.*, 1998). For our data, the extent of the incompleteness of the time series from individual monitors, as shown in Figure 2, makes this a less attractive strategy. Finally, construction of the spatio-temporal part of the model is greatly helped by the availability, at both monitor and residential locations, of a set of spatio-temporal covariates that are predictive of BS levels. Hence, anticipating the results in Sub-section 3.4, we do not necessarily need to build an elaborate spatio-temporal stochastic model for the residual component $Z_t(x)$.

3.3. Stage 1: modelling area-wide average BS levels

To model μ_t, we note from Figure 3 the approximately linear decline in log-BS levels over the study period, and the clear seasonal pattern, with higher levels occurring during the winter months. We anticipated that the seasonal pattern might be partially attributable to seasonal variation in temperature. We therefore obtained daily temperature readings from nearby weather recording stations for the whole study period, and calculated a value d_t as the average of the daily minimum temperature readings over the 7 days in week t. Finally, we set $\omega = 2\pi/52$ as the frequency corresponding to an annual cycle.

A first, static regression model for the spatial average $\bar{Y}_t$ is

$$\bar{Y}_t = \alpha + \beta t + \gamma d_t + A\cos(\omega t) + B\sin(\omega t) + U_t \tag{2}$$

where α, β, γ, A and B are parameters and the U_t are mutually independent $\mathrm{N}(0, \sigma_U^2)$ residuals.

Figure 4a shows that the model (2) captures much of the seasonal variation in $\bar{Y}_t$; for clarity, the diagram shows only representative results from years 1984 to 1992. However, the residuals show strong evidence of short-term and long-term autocorrelation, with small peaks corresponding to one- and two-year lags indicating that a static seasonal component is inadequate (Figure 4b). Re-examination of Figure 4a suggests that the lack of fit is primarily due to year-by-year variation in the phase and amplitude of the seasonal pattern. We therefore consider instead a dynamic regression model (West and Harrison, 1997),

$$\bar{Y}_t = \alpha + \beta t + \gamma d_t + A_t\cos(\omega t) + B_t\sin(\omega t) + U_t \tag{3}$$

DOI: 10.1002/env

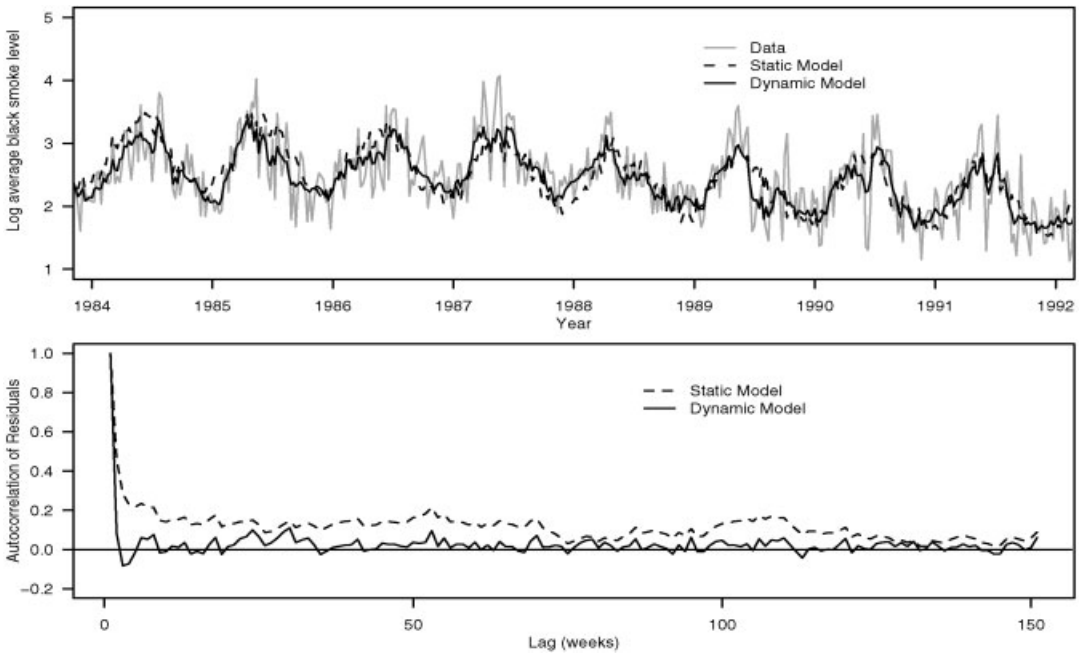

Figure 4. a. Fit of static (2) and dynamic (3) regression models for area-wide average black smoke levels, 1984–1992; b. Autocorrelation of residuals from static and dynamic models for area-wide average black smoke levels

where α, β, γ and U_t are as before, but now the static parameters A and B have been replaced by independent random walks, hence

$$A_t|A_{t-1} \sim \mathrm{N}(A_{t-1}, \sigma_A^2)$$
$$B_t|B_{t-1} \sim \mathrm{N}(B_{t-1}, \sigma_B^2)$$

Given initial values A_0 and B_0, the dynamic model (3) can be fitted either by direct maximisation of the likelihood function, or via a Kalman filter followed by Kalman smoothing using, for example, functions `kfilter` and `smoother` in the contributed R package `sspir` (`www.R-project.org`).

The estimated parameter values $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ differ little between models (2) and (3), but the estimated residual variance $\hat{\sigma}_U^2$ drops from 0.14 to 0.08 and the corresponding R^2-value increases slightly from 0.88 to 0.93. More importantly, the dynamic model provides a qualitative improvement in fit compared to the static model; the residual autocorrelation largely disappears (Figure 4b) as a result of the more flexible fit to the seasonal pattern (Figure 4a).

3.4. Stage 2: modelling residual spatio-temporal variation

If we now apply the area-wide fitted values, $\hat{\mu}_t$ say, from Equation (3) to the values of log-transformed BS at individual monitors, the residuals show a strong spatial pattern, with monitors towards the south of the study area tending to have large, positive residuals. This is consistent with the fact that early in the study period this part of the study area was dominated by areas of heavy industry.

DOI: 10.1002/env

As described in Sub-section 3.2, we seek to explain this effect by treating $\hat{\mu}_t$ as an offset in a linear model for monitor-specific log-transformed BS levels $Y_t(x)$ that includes spatio-temporally referenced covariate. Note that to achieve our aim of predicting BS exposure at every residential location, any covariates in the model must be available not only at monitor locations, but throughout the study area.

3.4.1. Covariates. To account for residual spatio-temporal variation, we constructed the following candidate covariates:

w_1: domestic chimney count within 500 m;
w_2: distance to nearest industrial area;
w_3: binary indicator of land use, either residential ($w_3 = 1$) or non-residential ($w_3 = 0$);
w_4: binary indicator of whether the 1956 Clean Air Act (CAA) had ($w_4 = 0$) or had not ($w_4 = 1$) been implemented;
w_5: area of industry within 500 m.

Covariates w_1, w_2 and w_5 were derived at a time resolution of 1 year from digitised annual images of the study area.

Covariate w_3 was derived as follows. For any monitoring location x within the study area and a given value of $r > 0$, we counted the number of births in each year within a radius r of location x. We then identified, by trial-and-error, a range of values of r for which the resulting count distribution was strongly bimodal, suggesting a classification of monitoring locations with high counts as residential and locations with low counts as non-residential. Using this criterion with $r = 150$ m provided the clearest distinction between residential and non-residential locations. Moreover, by this criterion the residential status of each monitoring location did not appear to change over time. We therefore considered w_3 to be time constant, and defined a residential area to be one for which at least 50 births occurred within a 150-m radius throughout the study period. The majority of monitors classified in this way as non-residential were in known industrial areas, although one was in a known commercial area.

Covariate w_4 was obtained from local government records. The CAA was implemented in stages across administrative sub-areas of the city between 1959 and 1978. The assumption that implementation within a sub-area took place at a fixed date, rather than gradually over a longer period of time, is questionable. However, in the absence of more detailed information, we took the pragmatic decision to define w_4 as a binary factor, changing from 1 to 0 at the nominal implementation date for the sub-area in question.

For a preliminary assessment of the importance of each candidate covariate, we compared monitor-specific average residuals and covariates as follows. For each monitor, at location x say, we defined the average residual as a time average of $Y_t(x) - \hat{\mu}_t$ over those weeks t in which the monitor was active, and the average covariate as the corresponding time average of the covariate at the same location. For the binary covariates, w_3 and w_4, we compared the two distributions of average residuals corresponding to $w = 0$ and $w = 1$. For w_1, w_2 and w_5, we examined scatterplots of monitor-specific average residuals against average covariate values.

On this basis, we discarded the industry variable w_5 because it showed a relatively weak relationship with monitor-specific average residuals and a strong relationship with the other covariates. The other covariates all showed a potentially useful relationship with the monitor-specific average residuals, and were therefore retained. Figure 5 shows the plot for the chimney count variable, w_1. Each point represents a monitor, and is labelled according to its residential status. The plot shows a positive relationship with chimney count for monitors in residential areas, and a negative relationship in non-residential areas, suggesting a strong interaction effect between chimney count and residential status.

DOI: 10.1002/env

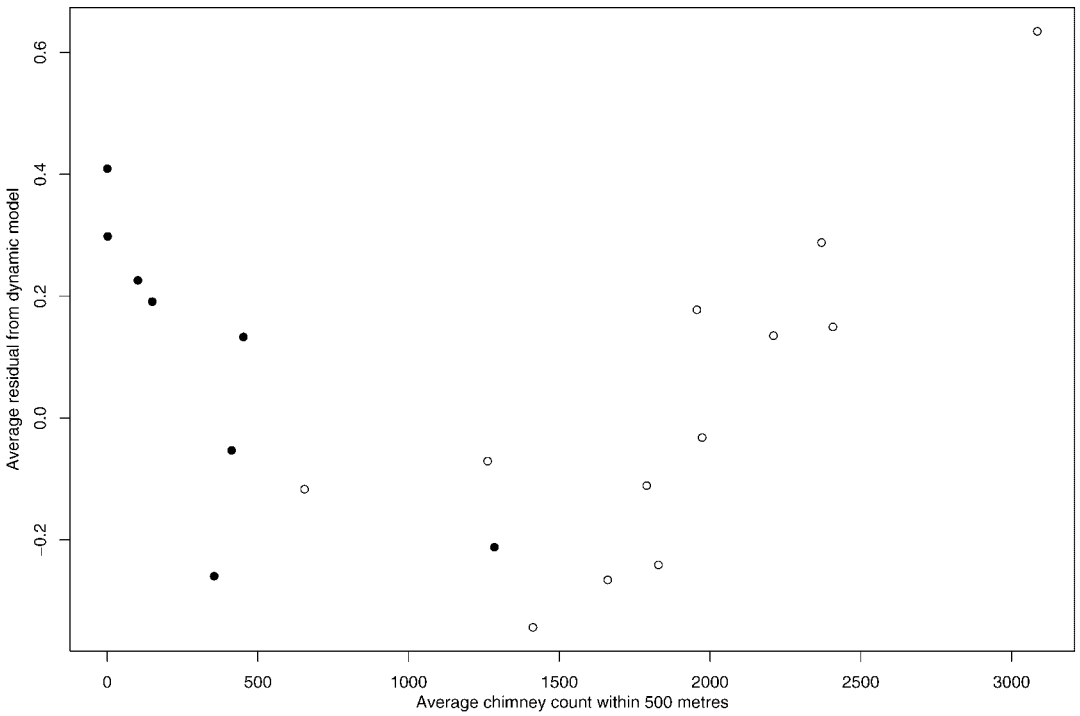

Figure 5. Monitor-specific average residual from dynamic model (3), plotted against average chimney count within 500 metres. Points are labelled according to the monitor's residential status (open circle = residential, filled circle = non − residential).

A possible explanation for this is that within industrial areas, very few domestic chimneys would be found close to the most heavily polluting industries, whereas rather more would be found close to the lighter industries. In residential areas, there is relatively little variability between levels of emission per chimney, and pollution levels therefore show a direct relationship with chimney count.

Another important interaction is between chimney count and date of implementation of the CAA. After the CAA was implemented, the emission of black smoke from any building was prohibited. Thus, as a surrogate for local levels of black smoke emission, the chimney count could be considered as being effectively zero after CAA implementation. However, as discussed below, care is needed to interpret correctly the combined effect of CAA implementation and the estimated area-wide temporal trend, $\hat{\mu}_t$.

3.4.2. Model formulation. We now consider a single linear model for the data from all monitors. Taking into account the above remarks, we assume the following model:

$$Y_t(x_k) = \hat{\mu}_t + \beta_{10} w_{10} + \beta_{11} w_{11} + \beta_2 w_2 + \beta_3 w_3 + \beta_4 w_4 + Z_t(x_k) \tag{4}$$

where x_k is the location of monitor k, w_2, w_3 and w_4 are as defined above, and $w_{1i} = w_1 \mathrm{I}(w_3 = i)$ where $\mathrm{I}(\cdot)$ is the indicator function. We also assume that $Z_t(x_k) \sim N(0, \sigma_Z^2)$ independently for all k and t. However, to preserve the interpretation of $\hat{\mu}_t$ as the area-wide average of S_t, we also need to centre each covariate appropriately. We therefore require that, for any given t, the fitted value from

DOI: 10.1002/env

the spatio-temporal model (4), averaged over all monitors active at t, should equal $\hat{\mu}_t$. To satisfy this condition, for each covariate w at each time t we calculate $\bar{w} = (\sum w)/m_t$, where the sum is over the m_t monitors active at time t, and subtract each value of $\bar{w}$ from the corresponding value of w before entering into Equation (4).

3.4.3. Assessment of model fit. Including monitor-specific fixed effects would be incompatible with our goal of spatial prediction. However, as a part of the assessment of the model fit, we did consider the effect of adding monitor-specific levels α_k to the right-hand side of Equation (4). This resulted in only a small increase in the R^2 value, from 0.84 to 0.86, and we therefore reverted to model (4).

To test the assumption of independent residuals $Z_t(x_k)$, we calculate a standardised average residual for each monitor k as

$$\bar{Z}_k = n_k^{-0.5} \sum_t \{S_t(x_k) - \hat{Y}_t(x_k)\}$$

where n_k is the number of weeks in which monitor k was active. Under the assumed model, $\bar{Z}_k^* \sim N(0, \sigma_Z^2)$, for all k. Figure 6 shows the standardised residuals plotted at their corresponding monitor locations. The visual impression is of a concentration of large, negative residuals close to the southern boundary of the study area. However, visual impressions from sparse spatial data can be misleading. For a formal test, we compute for each distinct pair (i, j) of monitors $u_{ij} = \|x_i - x_i\|$ and $v_{ij} = (\bar{Z}_i^* - \bar{Z}_j^*)^2$. We then use the sample correlation between u_{ij} and v_{ij} as a measure of the spatial dependence and

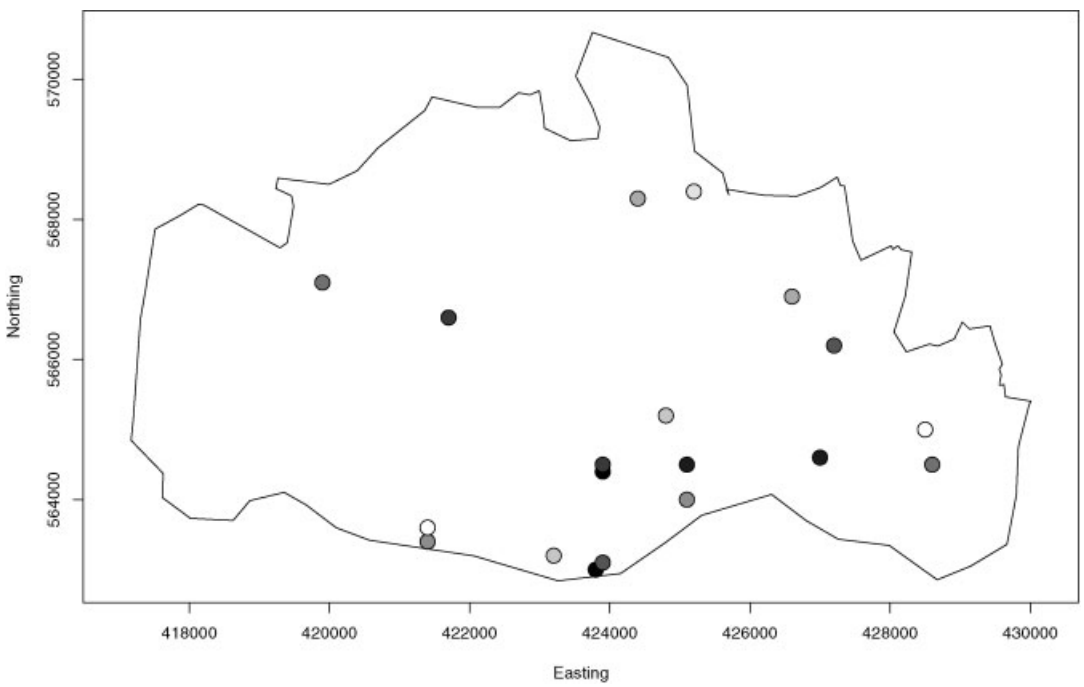

Figure 6. Map of standardised monitor-specific residuals from model (4). Darker shades indicate larger negative residuals, and lighter shades larger positive residuals

DOI: 10.1002/env

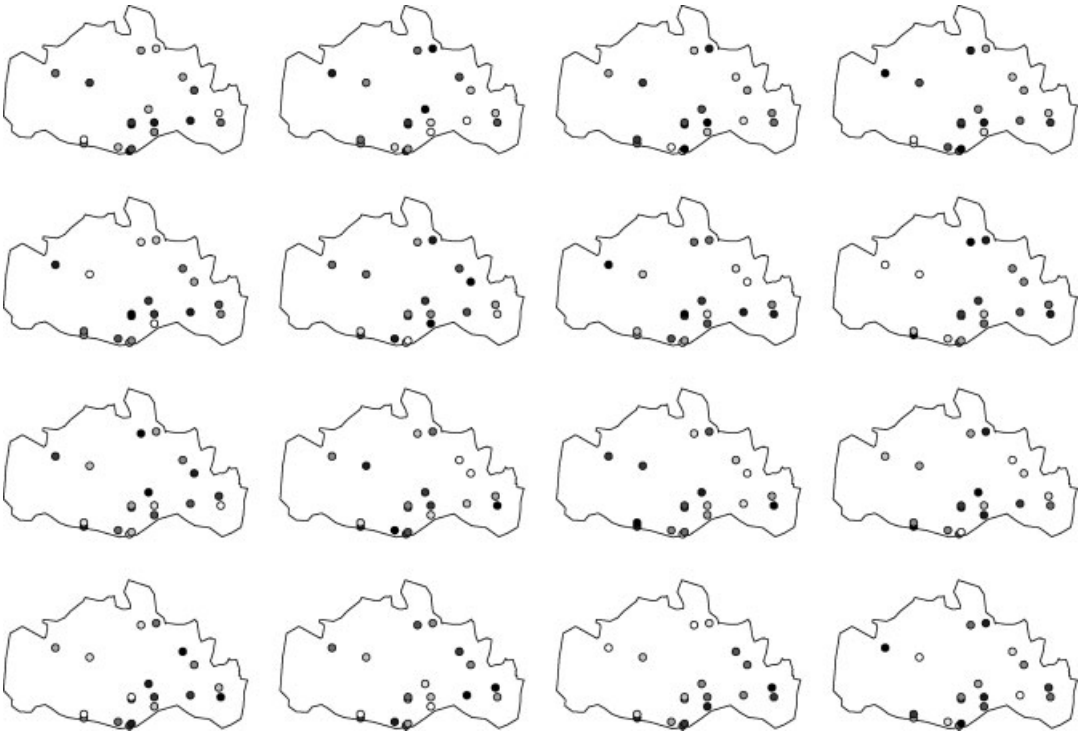

Figure 7. 16 replicates of a map of standardized monitor-specific residuals from model (4) with monitor locations randomly reassigned. Darker shades indicate larger negative residuals, and lighter shades larger positive residuals. The observed map (Figure 6) appears in the top left

compare the observed value with that obtained after randomly re-labelling the monitoring locations. The resulting Monte Carlo test, based on 999 independent re-labellings, gives a p-value of 0.7, corresponding to no significant evidence of spatial structure. Consistent with the result of the formal test, maps of re-labelled residuals (Figure 7) show chance spatial concentrations of large and small residuals comparable to those seen in Figure 6. We conclude that the assumption of spatially independent residuals $Z_t(x_k)$ is reasonable, and that any differences between monitors are likely to reflect properties of the monitors themselves, rather than of their locations.

We also examined the temporal pattern of residuals at individual monitoring stations. Time plots of residuals, shown in Figure 8, reveal clear lack of fit for some monitors over some time periods. Table 1 summarises the fit of the model to individual monitors, including the five validation monitors located outside the study area. The R^2-values for the 20 monitors within the study area vary between 0.21 and 0.87, but the smaller values of R^2 are generally associated with monitors for which we have relatively little data.

3.4.4. Validation. To assess the model's external validity, we used five additional monitors situated just outside the study area. The locations of these monitors are shown in Figure 1. Historical records were less readily available for locations outside the study area, for example the aerial photographs needed to construct the chimney count variable were available only for the years 1966 and 1974. For this reason,

DOI: 10.1002/env

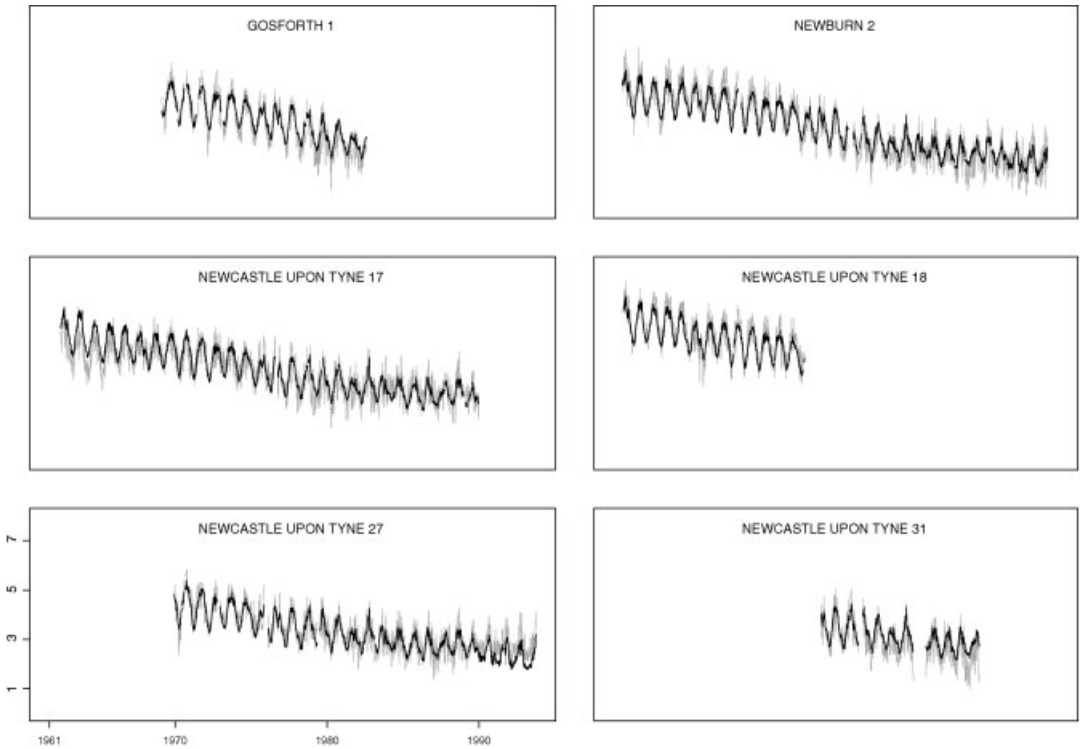

Figure 8. Observed (grey lines) and fitted (black lines) values from model (4) for six monitors used for model fitting

we consider only data from these years in our assessment of validity. Table 1 summarises the fit for these five monitors. The fit is rather poor for some monitors, notably Hebburn 3, and we would not recommend extrapolating the model beyond the study area. Inevitably, imposing a common model on all available monitor locations within the study area compromises the fit to any individual monitor's data, but is a necessary simplification in order to address our goal of spatio-temporal prediction at arbitrary locations. Extrapolation beyond the study area is likely to exacerbate this effect, for example although the locations of the validation monitors are geographically close to the boundary of the study area, they differ in their historical pattern of land use.

3.4.5. Interpretation of the spatio-temporal model coefficients. An alternative interpretation of Equation (4) is obtained by re-casting the dynamic model (3) to allow different area-wide average log-transformed BS levels before and after CAA implementation. We denote these by $\mu_{d,t}$ ('dirty') and $\mu_{c,t}$ ('clean'), respectively, and let p_t be the proportion of active monitors at week t that are dirty. Then, μ_t is a weighted average of log-transformed BS levels in dirty and clean areas,

$$\mu_t = p_t \mu_{dt} + (1 - p_t)\mu_{ct}$$

DOI: 10.1002/env

Table 1. Summary of spatio-temporal model fit for each of the 20 monitors used for model fitting and five monitors used for validation (see Figure 1). n refers to the number of weeks for which the monitor was active. $R^2 = 1 -$ (residual variance/raw variance)

Monitor		n	Mean residual	Standardised mean residual	Residual variance	Raw variance	R^2
1	Gosforth 1	676	0.14	3.74	0.17	0.92	0.81
2	Gosforth 2	22	0.22	1.03	0.20	0.34	0.42
3	Newburn 2	1598	−0.02	−0.78	0.20	1.58	0.87
4	Newcastle 17	1399	0.08	2.84	0.24	0.81	0.71
5	Newcastle 18	670	−0.16	−4.09	0.13	0.66	0.80
6	Newcastle 19	688	−0.09	−2.44	0.19	0.70	0.73
7	Newcastle 20	360	0.08	1.59	0.26	0.66	0.60
8	Newcastle 21	445	−0.22	−4.66	0.25	0.66	0.63
9	Newcastle 22	321	0.24	4.27	0.20	0.61	0.67
10	Newcastle 23	52	0.09	0.68	0.22	0.53	0.59
11	Newcastle 24	1064	0.20	6.38	0.26	1.61	0.84
12	Newcastle 25	339	−0.08	−1.52	0.14	0.55	0.74
13	Newcastle 26	224	−0.32	−4.86	0.09	0.44	0.79
14	Newcastle 27	1198	−0.12	−4.05	0.15	0.65	0.76
15	Newcastle 28	229	−0.23	−3.55	0.08	0.38	0.78
16	Newcastle 29	89	−0.23	−2.18	0.20	0.32	0.39
17	Newcastle 30	44	−0.06	−0.37	0.08	0.30	0.73
18	Newcastle 31	527	0.23	5.34	0.19	0.54	0.65
19	Newcastle 32	224	−0.09	−1.40	0.14	0.34	0.58
20	Newcastle 5	5	−0.31	−0.69	0.05	0.06	0.21
21	Blaydon 3	4	−1.27	−2.55	0.28	0.25	−0.11
22	Gateshead 5	77	−0.26	−2.27	0.11	0.48	0.76
23	Hebburn 3	52	−1.27	−9.17	0.27	0.33	0.18
24	Hebburn 4	52	0.62	4.51	0.26	0.43	0.38
25	Newburn 1	87	0.63	5.85	0.09	0.28	0.67

Now suppose that $\mu_{dt} = \mu_{ct} + \lambda_t$ for some function λ_t, so that

$$\begin{aligned}\mu_t &= \mu_{ct} + p_t\lambda_t \\ &= \mu_{dt} + p_t\lambda_t - \lambda_t\end{aligned}$$

The estimated contribution to the right-hand side of Equation (4) for a clean monitor is $\hat{\mu}_t - \hat{\beta}_4 p_t$, whilst the estimated contribution for a dirty monitor is $\hat{\mu}_t + \hat{\beta}_4 - \hat{\beta}_4 p_t$. These quantities estimate μ_{ct} and μ_{dt}, respectively. Hence, $\hat{\beta}_4$ can be interpreted as an estimate of the difference in average log-transformed BS levels between dirty and clean areas, on the assumption that this difference is constant over time. The estimate of this difference is $\hat{\beta}_4 = 0.33$, with standard error 0.013. Figure 9 shows the observed average difference in log-transformed BS between dirty and clean monitors at each time, and supports the assumption that λ_t is approximately constant.

Direct interpretation of the other β-coefficients in Equation (4) is more difficult, owing to the necessary standardisation of the covariates w. Nevertheless, we note that in each case the parameter estimate has the anticipated sign (i.e. negative for $\hat{\beta}_{10}$, $\hat{\beta}_2$ and $\hat{\beta}_3$, positive for $\hat{\beta}_{11}$).

DOI: 10.1002/env

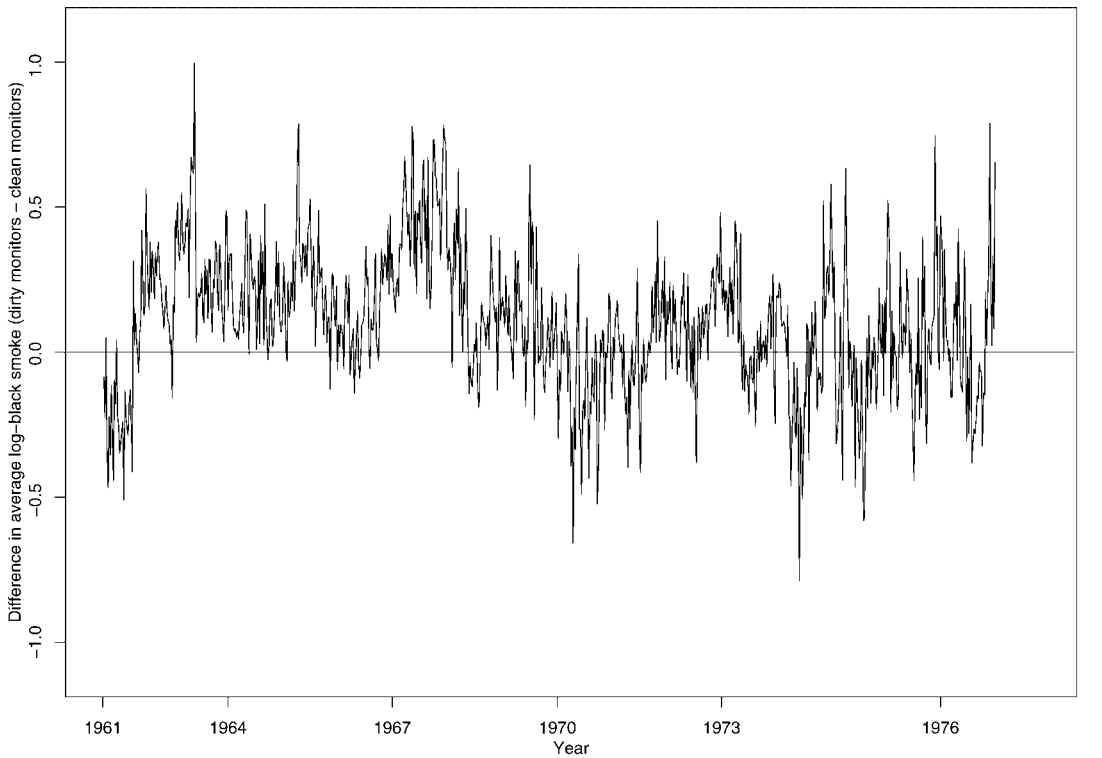

Figure 9. Difference between average log-black smoke levels in monitors operating in areas before ('dirty') and after ('clean') the implementation of the 1956 Clean Air Act

4. PREDICTION OF BS EXPOSURE AT RESIDENTIAL LOCATIONS

Our aim is to predict BS exposure at each maternal residential location, both for individual weeks and aggregated over time within the pregnancy. Thus, to compute prediction variances we need to consider not only the prediction variance for a single week, but also the covariance between predictions made for different weeks. Each birth is associated with a single residential location, x say, so in order to estimate an individual mother's exposure we need to only consider prediction at that location. To simplify notation, we therefore suppress the dependence on x and write S_t for the BS level at time t, $Y_t = \log(S_t)$, and $\mathbf{w}_t$ for the covariate vector at this location x and week t. The following discussion then holds for any location x.

Suppose that our target for prediction is the time-aggregated BS exposure over weeks $t_1, \ldots, t_n$. As the prediction variance of $\hat{\mu}(t)$ is small by comparison with that of Y_t (approximately 0.08 vs 0.27), we treat $\hat{\mu}_t$ as known and equal to μ_t. The predicted value of $Y_t - \mu_t$ is $\hat{Y}_t - \hat{\mu}_t$ where $\hat{Y}_t = \mathbf{w}_t^T \hat{\boldsymbol{\beta}}$, with associated prediction variance

$$\begin{aligned} V(\hat{Y}_t) = V(Y_t - \hat{Y}_t) &= V(Z_t + \mathbf{w}_t^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \\ &= V(Z_t) + V(\mathbf{w}_t^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \\ &= \sigma_Z^2 + \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_t \end{aligned}$$

DOI: 10.1002/env

Also, for $t \neq u$,

$$\begin{aligned}\mathrm{Cov}(\hat{Y}_t, \hat{Y}_u) &= \mathrm{Cov}(\mathbf{w}_t^T\hat{\boldsymbol{\beta}} + \hat{Z}_t, \mathbf{w}_u^T\hat{\boldsymbol{\beta}} + \hat{Z}_u) \\ &= \mathrm{Cov}(\mathbf{w}_t^T\hat{\boldsymbol{\beta}}, \mathbf{w}_u^T\hat{\boldsymbol{\beta}}) \\ &= \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_u\end{aligned}$$

Under the fitted model (4), $\sigma_Z^2 \gg \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_t$ in any week t (approximately 0.27 and 0.00006, respectively), and it follows that

$$\mathrm{Var}\left(\sum_{t=t_1}^{t_n}\hat{Y}_t\right) = \sum_{t=t_1}^{t_n}\mathrm{Var}(\hat{Y}_t) + 2\sum_{t<u}\mathrm{Cov}(\hat{Y}_t, \hat{Y}_u) \approx \sum_{t=t_1}^{t_n}\mathrm{Var}(\hat{Y}_t) \approx n\sigma_Z^2$$

We require predictions on the original scale, rather than on the log-transformed scale. At a given location, $S_t = \exp(Y_t)$ and our targets for prediction are of the form $T = n^{-1}\sum_{t=t_1}^{t_n} S_t$. Under our assumed model, each S_t follows a log-Normal distribution. Writing $\xi_t = \mathrm{E}[Y_t]$ and $\Sigma_{tu} = \mathrm{Cov}\{Y_t, Y_u\}$, it follows that

$$\mathrm{E}(S_t) = \exp(\xi_t + \Sigma_{tt}/2)$$

$$\mathrm{Var}(S_t) = \exp(2\xi_t + \Sigma_{tt})(\exp(\Sigma_{tt}) - 1)$$

and for $t \neq u$,

$$\mathrm{Cov}(S_t, S_u) = \exp(\xi_t + \xi_u + (\Sigma_{tt} + \Sigma_{uu})/2)(\exp(\Sigma_{tu}) - 1)$$

The prediction variance for the average black smoke level T, over weeks $t_1, \ldots, t_n$, follows as

$$\mathrm{Var}(T) = \mathrm{Var}\left(\frac{1}{n}\sum_{t=t_1}^{t_n} S_t\right) = \frac{1}{n^2}\left(\sum_{t=t_1}^{t_n}\mathrm{Var}(S_t) + 2\sum_{t<u}\mathrm{Cov}(S_t, S_u)\right) \approx \frac{1}{n^2}\sum_{t=t_1}^{t_n}\mathrm{Var}(S_t)$$

and approximate prediction intervals can be computed using a Normal approximation. For example, an approximate 95% prediction interval for T is

$$\frac{1}{n}\sum_{t=t_1}^{t_n}\exp(\hat{\xi}_t + \hat{\Sigma}_{tt}/2) \pm 1.96\sqrt{\mathrm{Var}(\hat{T})}$$

Figure 10 shows a grey-scale image of predicted values on the logarithmic scale for 4 weeks, corresponding to summer and winter in 1969 and 1982. Non-residential locations, for which prediction is of no interest, are shown in Figure 10 as white areas. One feature of Figure 10 is the relatively low spatial variation at any one time, by comparison with the variation either between different seasons in the same year, or between different years for the same season. This pattern is consistent with our exploratory analysis of these data as reported in Sub-section 3.1.

DOI: 10.1002/env

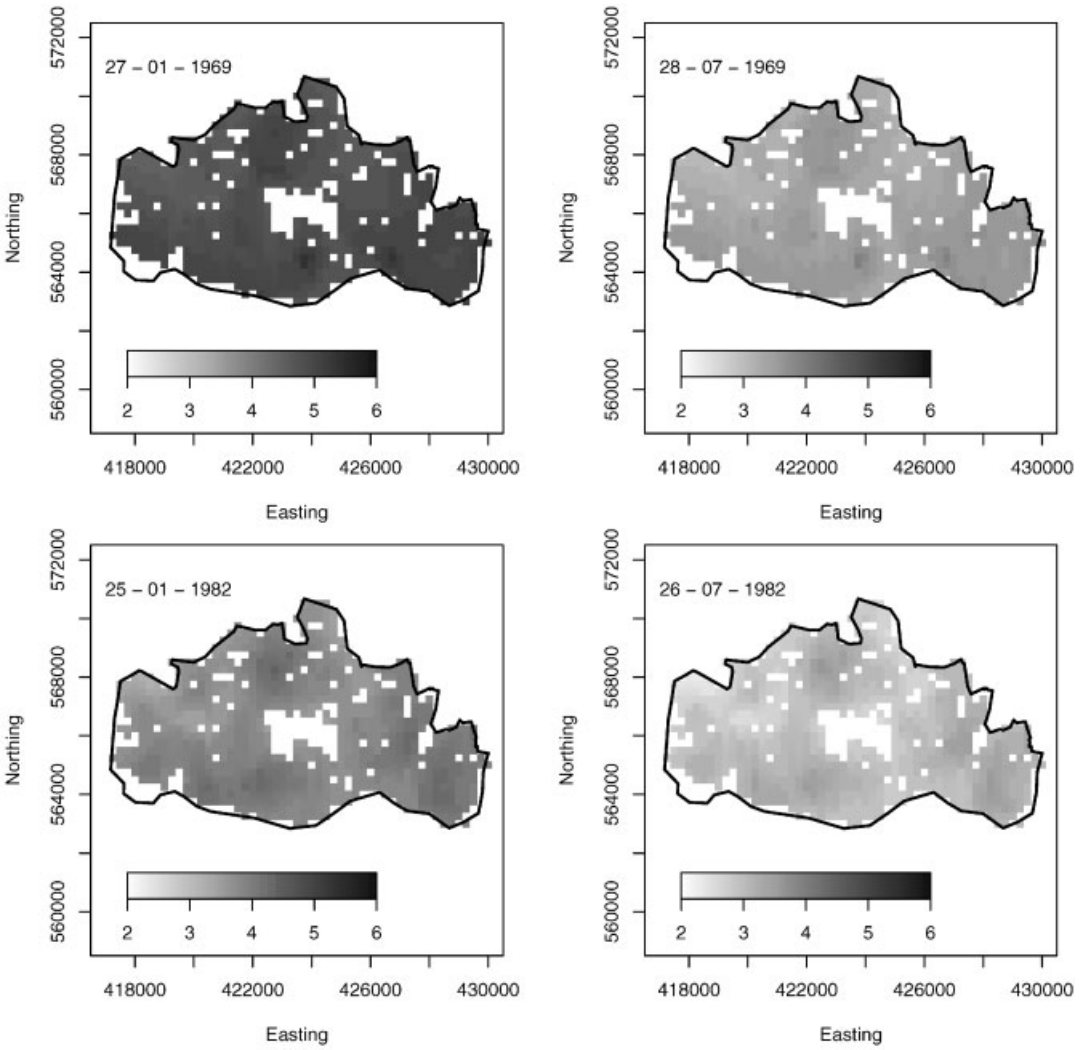

Figure 10. Point predictions for log-BS levels for four single weeks (dates inset) representing winter and summer, 1969 and 1982. White pixels correspond to non-residential areas, for which no prediction is made

The pattern of prediction variances is qualitatively similar to that of the predictions themselves, as a consequence of the log-Normal distributional assumption for untransformed BS concentrations.

5. DISCUSSION

We have demonstrated a two-stage modelling strategy for modelling spatio-temporal data using monitoring data that are temporally dense and spatially sparse, a common scenario in epidemiological studies of air pollution exposure. In the first stage, we used a dynamic model for the purely temporal

DOI: 10.1002/env

trend, while in the second we used appropriately constructed covariates to take account of remaining spatio-temporal variation. Using a dynamic model in the first stage obviates the need to consider separate models for short-term and long-term correlation between observations, and in our application resulted in a materially better fit to seasonal variation in spatially averaged pollution levels than was obtainable from a static harmonic regression model.

The area-wide average log-transformed BS levels given by the first-stage model are relatively precise, with prediction variance around 0.08 compared with predicted values ranging between 1.7 and 6.3. In contrast, the spatial sparsity of the data makes it important to take account of the uncertainty in the predictions at particular locations. Our exposure estimates will subsequently be used as covariates in an analysis of the relationship between exposure and adverse birth outcomes, in which context it will be necessary to check that conclusions are robust against the statistical error in the exposure estimates. We believe that these estimates, although only surrogates for the true levels of pollution to which mothers were exposed, indicate a more realistic pattern of exposure than would an assumption of homogeneity of exposures across a whole city. This seems likely to hold true both for particulate matter and for other pollutants, for which there is evidence elsewhere (Haas, 1995; Meiring *et al*., 1998; Zidek *et al*., 2002).

In our application, we have been able to model the spatio-temporal variation without the need to model spatio-temporal correlation in the residuals. This greatly eases the computational burden of computing predictions and prediction variances. In principle, the methodology extends directly to models with correlated residuals, provided that we are prepared to specify a spatio-temporal covariance structure for the residual process $Z_t(x)$; see, for example, Gneiting *et al*. (2007). In problems of this kind, we would always advocate the use of relevant covariate information to explain as much as possible of the spatio-temporal variation. Nevertheless, and as the results from the validation sites indicate, extrapolation beyond the area from which the model was constructed is almost certainly unwarranted. In other settings the importance of other sources of pollution, for example traffic emissions, may require the use of different covariates. The means by which suitable covariates are identified and constructed is not necessarily straightforward and may require a degree of imagination; in our application, the construction of the chimney count covariate and careful consideration of its interaction with both the residential/non-residential land-use classification and with the effect of the Clean Air Act were crucial to the implementation of the methodology.

ACKNOWLEDGEMENTS

The PAMPER study was funded by the Wellcome Trust, UK charity, grant number 072465/Z/03/Z. TRF is supported by a Doctoral Training Account studentship and PJD by a Senior Fellowship from the Engineering and Physical Sciences Research Council (EPSRC).

REFERENCES

Bobak M. 2000. Outdoor air pollution, low birth weight, and prematurity. *Environmental Health Perspectives* **108**: 173–176.

Bogaert P, Christakos G. 1997. Stochastic analysis of spatiotemporal solute content measurements using a regression model. *Stochastic Hydraulics and Hydrology* **11**: 267–295.

Brown PE, Diggle PJ, Lord ME, Young PC. 2001. Space-time calibration of radar rainfall data. *Applied Statistics* **50**: 221–241.

Carroll RJ, Chen R, George EI, Li TH, Newton HJ, Schmiedliche H, Wang N. 1997. Ozone exposure and population density in Harris County, Texas. *Journal of The American Statistical Association* **92**: 392–404.

de Luna X, Genton MG. 2005. Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica* **15**: 547–568.

DOI: 10.1002/env

Glinianaia SV, Rankin J, Bell R, Pless-Mulloli T, Howel D. 2004a. Does particulate air pollution contribute to infant death? A systematic review. *Environmental Health Perspectives* **112**: 1365–1371.
Glinianaia SV, Rankin J, Bell R, Pless-Mulloli T, Howel D. 2004b. Particulate air pollution and fetal health: a systematic review of the epidemiological evidence. *Epidemiology* **15**: 36–45.
Gneiting T, Genton MG, Guttorp P. 2007. In *Statistical Methods for Spatio-Temporal Systems*. Chapter 4 Chapman and Hall/CRC. Boca Raton, Florida, USA.
Haas TC. 1995. Local prediction of a spatiotemporal process with an application to wet sulfate decomposition. *Journal of The American Statistical Association* **90**: 1189–1199.
Handcock MS, Wallis JR. 1994. An approach to statistical spatial-temporal modelling of meteorological fields. *Journal of The American Statistical Association* **89**: 368–378.
Haslett J, Raftery AE. 1989. Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. *Applied Statistics* **38**: 1–50.
Higdon D. 2007. In *Statistical Methods for Spatio-Temporal Systems*. Chapter 6 Chapman and Hall/CRC. Boca Raton, Florida, USA.
Jerrett M, Buzzelli M, Burnett RT, DeLuca PF. 2005. Particulate air pollution, social con-founders, and mortality in small areas of an industrial city. *Social Science and Medicine* **60**: 2845–2863.
Kyriakidis PC, Journal AG. 1999. Geostatistical space-time models: a review. *Mathematical Geology* **31**: 651–684.
Li K, Le ND, Sun L, Zidek JV. 1999. Spatial-temporal models for ambient hourly PM10 in Vancouver. *Environmetrics* **10**: 321-338.
Meiring W, Guttorp P, Sampson PD. 1998. Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics* **5**: 197–222.
Pless-Mulloli T, Glinianaia SV, Rushton S, Lurz PWW, Sanderson R, Fanshawe TR, Pearce MS, Charlton M, Shirley M, Rankin J, Diggle PJ. 2007. Within-city space and time variation of black smoke exposure during pregnancy over 32 years. Submitted.
Pope III CA, Dockery DW. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association* **56**: 709–742.
Ritz B, Wilhelm M, Zhao Y. 2006. Air pollution and infant death in southern California, 1989–2000. *Pediatrics* **118**: 493–502.
Sahu SK, Gelfand AE, Holland DM. 2006. Spatio-temporal monitoring of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics* **11**: 61–86.
Sahu SK, Mardia KV. 2005. Recent trends in modeling spatio-temporal data. In *Proceedings of the special meeting on Statistics and Environment*; 69–83.
Samet JM, Dominici F, Curreiro FC, Coursac I, Zeger SL. 2000. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. *The New England Journal of Medicine* **343**: 1742–1749.
Sram RJ, Binkova B, Dejmek J, Bobak M. 2005. Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental Health Perspectives* **113**: 375–382.
Stroud JR, Muller P, Sanso B. 2001. Dynamic models for spatiotemporal data. *Journal of The Royal Statistical Society: Series B* **63**: 673–689.
West M, Harrison PJ. 1997. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag: New York.
Woodruff TJ, Grillo J, Schoendorf KC. 1997. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. *Environmental Health Perspectives* **105**: 608–612.
Woodruff TJ, Parker JD, Schoendorf KC. 2006. Fine particulate matter (PM2.5) air pollution and selected causes of postneonatal infant mortality in California. *Environmental Health Perspectives* **114**: 786–790.
Zidek JV, Sun L, Le N, Ozkaynak H. 2002. Contending with space-time interaction in the spatial prediction of pollution: vancouver's hourly ambient PM10 field. *Environmetrics* **13**: 595–613.

DOI: 10.1002/env