APPENDIX AVAILABLE ON REQUEST

Research Report 157, Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Two Indian Cities

Part 1. Short-Term Effects of Air Pollution on Mortality: Results from a Time-Series Analysis in Chennai, India

Kalpana Balakrishnan et al.

Part 2. Time-Series Study on Air Pollution and Mortality in Delhi

Uma Rajarathnam et al.

Appendix F. Single and Multiple Monitor Models (Part 1)

Note: Appendices Available on the Web appear in a different order than in the original Investigators' Report. HEI has not changed these documents. Appendices were relettered as follows:

Appendix C was originally Appendix I
Appendix D was originally Appendix II
Appendix E was originally Appendix III
Appendix F was originally Annexure 2
Appendix G was originally Annexure 3 (Figure 1)
Appendix H was originally Appendix IV
Appendix I was originally Appendix V
Appendix J was originally Appendix VII

Note: Appendices F & G are for Part 1; Appendices H–J are for Part 2.

Correspondence for Part 1 may be addressed to Dr. Kalpana Balakrishnan, Professor and Head, Department of Environmental Health Engineering, Sri Ramachandra University, Porur, Chennai 600 116, India. *kalpanasrmc@vsnl.com*. Correspondence for Part 2 may be addressed to Dr. Uma Rajarathnam, Enzen Global Solutions, 90, Madiwala, Hosur Road, Bangalore 560 068, India. *uma.r@enzenglobal.com*.

The PAPA Program was initiated by the Health Effects Institute in part to support the Clean Air Initiative for Asian Cities (CAI-Asia), a partnership of the Asian Development Bank and the World Bank to inform regional decisions about improving air quality in Asia. Additional funding was obtained from the William and Flora Hewlett Foundation. The contents of this document have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee but did not undergo the HEI scientific editing and production process.

2 Annexure-II

2.1 Single Monitor models

The single monitor model is an alternative to the more popular model using the daily mean of all monitors as a single exposure series which is representative for the entire city. The latter is useful if the readings recorded by the various monitors show a high degree of correlation over time. This is not the case for Chennai as the monitor means, variabilities and movements over time all show significant differences. An alternative means to construct a single exposure series would be to use a single monitor and assume it to be representative for the entire city. Accordingly, in this section, we describe three single monitor series corresponding to A. Nagar, V. Nagar and the industrial average.

We use the generalized additive model (Hastie and Tibshirani 1990; Dominici et al. 1999) to model the association between daily mortality and each of the three single monitor exposure series in turn while allowing for a time trend in mortality and potential confounders such as temperature and relative humidity. In each case, a quasi-Poisson link was used and over dispersion was allowed for. Statistical analysis was carried out using version 1.3-29 of the mgcv package in R (Wood 2006).

For example, the model fitted using the A. Nagar exposure series was:

$$\log(\mathrm{E}(\mathrm{Mortality}_t)) = \alpha_0 + \beta_1 \mathrm{ANagar}_{(t-1)} + f_1(t) + f_2(\mathrm{temp}_t) + f_3(\mathrm{rh}_t) \quad (2.1)$$

where

$\mathrm{Mortality}_t$ = Total all cause non-accidental deaths on day t

ANagar_t = PM_{10} reading for A. Nagar on day t

$temp_t$ = Average daily temperature on day t

rh_t = Average daily relative humidity on day t

The model in (2.1) only considers those values of *t* corresponding to which the recording for A. Nagar is available. Degrees of freedom for the smooth terms was selected following the algorithm in the Statistical Models section *Selection of Confounder Degrees of Freedom* in the report separately for each of the three exposure series. This led to a choice of 6 degrees of freedom for time, 4 degrees of freedom for temperature and 4 degrees of freedom for relative humidity per year for the model using A. Nagar data.

We also fitted the same model using the exposure series from the V. Nagar monitor and the industrial average separately. The degrees of freedom for time, temperature and relative humidity are respectively 5,6 and 4 per year for the industrial average and 8,6 and 5 per year for V. Nagar.

Previous studies (Peng et al. 2006) have raised concerns about the ability to estimate the relative risk due to PM in situations of high concurvity (i.e. when a regression of the PM series on time, temperature and relative humidity leads to a residual variance which is near zero). In our case, the adjusted R^2 values from the regressions of the A. Nagar, V. Nagar and industrial series regressions on the confounders are respectively 0.1, 0.2, and 0.2, suggesting that the situation is one of moderate concurvity.

2.2 Multiple monitor models using observed exposures only

While the three models in the previous section lead to reasonably consistent estimates, they are not entirely satisfactory. Using this approach helps circumvent the problem of combining monitors to arrive at a spurious level of concentration

distinct from what different parts of the city witness. However, owing to the limited spatial scale of monitors, none of the monitors reflect average population exposures. Moreover, the differences in the behaviour of the three series over time implies that each has a different correlation with total mortality. From a statistical perspective, the models are unsatisfactory because they do not make the most efficient use of all available data. For example, fitting the model based on only the monitor readings from A. Nagar involves throwing away 66% of the exposure data and 65% of the mortality data. None of the single monitor models adequately explains the variability in the mortality data and this is reflected by the high values of the overdispersion parameters. With a view to addressing these concerns, we next fitted a model using the observed data for both A. Nagar and V. Nagar in a single regression model. We next extended the model to also include the industrial average. Our approach does not involve combining the readings from the various monitors into a single exposure series. In the first case, we fit a generalized additive model with separate terms for the A. Nagar and V. Nagar monitors and allow the monitors to have a possibly differential impact on daily mortality. This may be justified by the differential criterion used by the CPCB for defining residential, commercial and industrial monitors discussed in the Methods section *Air Pollution Data* of the report.

To define our model specification, we first define indicator variables:

$$I_{AN_{(t)}} = \begin{cases} 1 & \text{if A.Nagar reading is recorded on day t} \\ 0 & \text{otherwise.} \end{cases}$$

$$I_{VN_{(t)}} = \begin{cases} 1 & \text{if V.Nagar reading is recorded on day t} \\ 0 & \text{otherwise.} \end{cases}$$

If both readings are available, a multi monitor formulation would be

$$\begin{aligned}\log(E(\text{Mortality}_t)) &= \alpha_0 + \beta_1 \text{ANagar}_{t-1} + \beta_2 \text{VNagar}_{t-1} + f_1(t) \\ &+ f_2(\text{temp}) + f_3(\text{rh}) \qquad (2.2)\end{aligned}$$

Since readings are available for only one of the monitors on some days, we change the above formulation to accommodate this as follows:

$$\begin{aligned}\log(E(\text{Mortality}_t)) &= \alpha_1 I_{AN_{(t-1)}} + \alpha_2 I_{VN_{(t-1)}} + \alpha_3 I_{AN_{(t-1)}} * I_{VN_{(t-1)}} \qquad (2.3) \\ &+ \beta_1 \text{ANagar}_{(t-1)} * IAN_{(t-1)} + \beta_2 \text{VNagar}_{(t-1)} * IVN_{(t-1)} \\ &+ f_1(t) + f_2(\text{temp}) + f_3(\text{rh})\end{aligned}$$

Equation (2.3) is equivalent to assuming the models:

$$\log(E(\text{Mortality}_t)) = \alpha_1 + \beta_1 \text{ANagar}_{t-1} + f_1(t) + f_2(\text{temp}) + f_3(\text{rh})$$

on days when only the A. Nagar readings are available. The parameter α_1 is an adjustment to the intercept and includes the contribution of the missing V. Nagar reading on such days. Similarly, on days when the V. Nagar readings are available only, our model assumes:

$$\log(E(\text{Mortality}_t)) = \alpha_2 + \beta_2 \text{ VNagar}_{t-1} + f_1(t) + f_2(\text{temp}) + f_3(\text{rh}).$$

The parameter α_2 similarly is an adjustment to the intercept and includes the contribution of the missing A. Nagar reading on such days. On days when both monitor readings are available, equation (2.3) is equivalent to assuming the models:

$$\begin{aligned}\log(E(\text{Mortality}_t)) &= (\alpha_1 + \alpha_2 + \alpha_2) + \beta_1 \text{ANagar}_{t-1} + \beta_2 \text{ANagar}_{t-1} \\ &+ f_1(t) + f_2(\text{temp}_t) + f_3(\text{rh}_t)\end{aligned}$$

This model allows for the possibility that $\beta_1 = \beta_2$ and, in contrast to the single

monitor formulations, also allows the user to conduct a statistical test of the hypothesis $H_0 : \beta_1 = \beta_2$. Note that interpretation of β_1 and β_2 could be problematic if the monitor readings at A. Nagar and V. Nagar were highly correlated as in this case we can expect the estimates from the single and multiple monitors to differ. However, this is not a concern for our data because the inter-monitor correlations are low. In our situation, adding in the V. Nagar exposures will reduce some of the unexplained variation from the model based on data from A. Nagar only and hence will lead to more precise estimation. For purposes of fitting, as before, an over dispersion parameter was allowed for and the model was fitted using the mgcv package in R as before. The degrees of freedom for each of the smooth terms was the maximum degrees of freedom for that term from the three single monitor models viz. 24 for time, 17 for temperature and 14 for relative humidity.

We conducted a Wald test of the hypothesis $H_0 : \beta_1 = \beta_2$ but did not find sufficient evidence to support a differential monitor effect. We thus estimated a pooled common β coefficient as a weighted average of the coefficients for the two monitors, with the inverse of the estimated variance as weights. Such a choice of weights leads to a more precise estimate(the pooled coefficient has a shorter confidence interval than the individual single monitor estimates as shown in Appendix G).

We next refitted the model specified in equation (2.3) to include the industrial average also. Let

$$I_{\mathsf{Ind}_{(t)}} = \begin{cases} 1 & \text{if the industrial average is available on day t} \\ 0 & \text{otherwise.} \end{cases}$$

$$I_{AN_{(t)}} = \begin{cases} 1 & \text{if the A.Nagar reading is recorded on day t} \\ 0 & \text{otherwise.} \end{cases}$$

$$I_{VN_{(t)}} = \begin{cases} 1 & \text{if the V.Nagar reading is recorded on day t} \\ 0 & \text{otherwise.} \end{cases}$$

We assume the model

$$\begin{aligned} \log(E(\mathsf{Mort}_t)) &= \alpha_1 I_{\mathsf{Ind}_{(t-1)}} + \alpha_2 I_{AN_{(t-1)}} + \alpha_3 I_{VN_{(t-1)}} + \alpha_4 I_{\mathsf{Ind}_{(t-1)}} * I_{AN_{(t-1)}} \\ &+ \alpha_5 I_{\mathsf{Ind}_{(t-1)}} * I_{VN_{(t-1)}} + \alpha_6 I_{VN_{(t-1)}} * I_{AN_{(t-1)}} \qquad (2.4) \\ &+ \alpha_7 I_{VN_{(t-1)}} * I_{AN_{(t-1)}} * I_{\mathsf{Ind}_{(t-1)}} + \beta_1 \mathsf{Ind}_{(t-1)}) * I_{\mathsf{Ind}_{(t-1)}} \\ &+ \beta_2 \mathsf{ANagar}_{(t-1)} * I_{AN_{(t-1)}} + \beta_3 \mathsf{VNagar}_{(t-1)} * I_{VN_{(t-1)}} \\ &+ f_1(t) + f_2(\mathsf{temp}) + f_3(\mathsf{rh}). \end{aligned}$$

As before, the α terms are adjustments to the intercept to allow for missing exposure data, β_1 is the log mortality ratio for a unit increase in PM recorded by the industrial monitors. Similarly β_2 and β_3 are the same corresponding to the monitors at A. Nagar and V. Nagar respectively. As before, we conducted a Wald test to check whether the data support the claim of a differential impact of each monitor on mortality i.e. to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$. If the data did not find sufficient evidence to reject H_0, we can pool the three estimates using a weighted average with the inverse variances as weights to arrive at a single common estimate.

2.3 Multiple monitor model with imputation of missing values

In this section, we explore an alternative means of fitting a multi monitor model. We fit the same model as in the previous section but we additionally impute the missing exposure data. The imputation based models make greater use of available data as it is no longer necessary to discard the mortality data for days on which a monitor reading is not recorded. However, weekends were exempted from this analysis as it was felt that the available data were too sparse for any imputation model to give precise results. An imputation approach which has been successfully used in the past (O'Neill et al. 2002) is to substitute fitted values from a regression of observed exposures on meteorological covariates such as visibility.

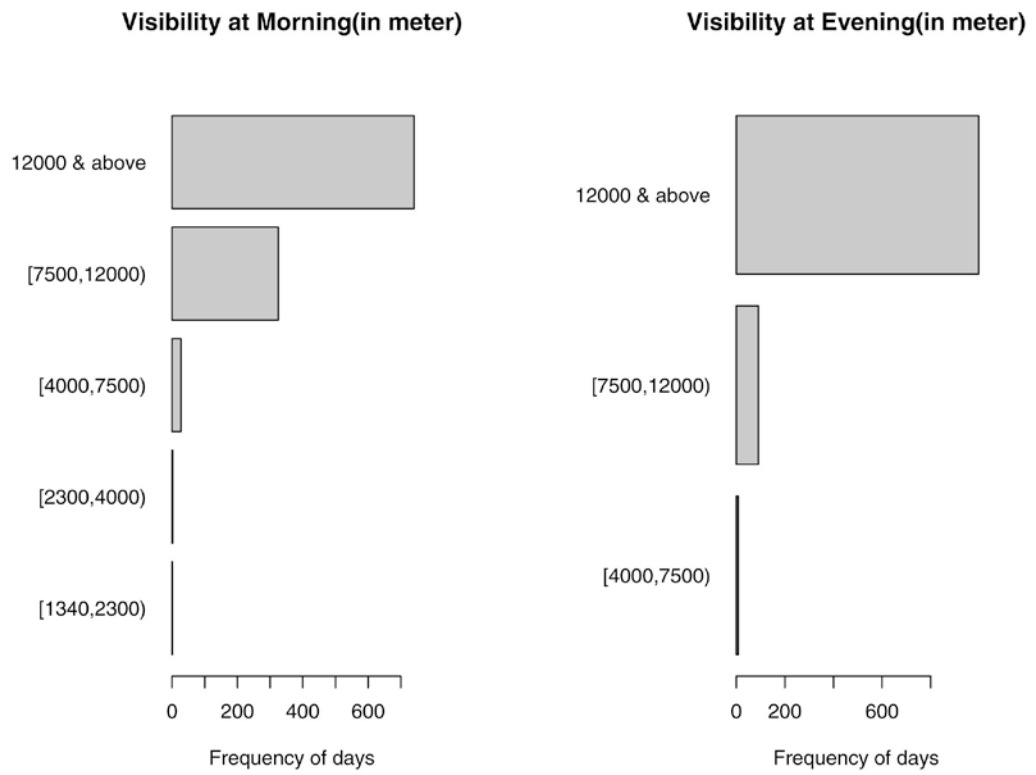

Figure F.1. Visibility data for Chennai city during the study period.

Unfortunately, in our case, data on exact visibility were not available as the meteorological office was only able to provide categorized visibility. Figure F.1 shows a bar plot based on these data. The figure shows that the majority of the readings are clustered in a single category rendering the data useless for imputation purposes. As a result alternative approaches were adopted to impute the missing data.

Missing data were imputed by predicted values from a regression of the observed exposure series on time. This is a flexible model in that it does not make any assumptions other than that each series of monitor readings can be modeled as a smooth function of time. Specifically the exposure recorded by the i^{th} monitor on day t_j is assumed to be generated by the model:

$$PM_{ij} = g_i(t_j) + e_{ij}; e_{ij} \sim \mathrm{N}(0, \sigma_i{}^2) \quad \text{independently for monitor } i, \qquad (2.5)$$
$$i = 1, \ldots, 5; j = 1, \ldots, n_i.$$

where the functions $g_i(.)$ are unknown but assumed to be smooth. We do not make any assumptions about the functional form of the $g_i(.)$*'s* but allow them to be estimated from the data. This can be accomplished by modeling the $g_i(.)$*'s* as splines and fitting the model in (2.5) as a generalized additive model again using mgcv in R.

The model in (2.5) was fitted to the observed ($PM_{ij,tj}$) values separately for each monitor. Predicted values of PM_{ij} were obtained from the fit for days on which monitor readings were not recorded.

Figure F.2 shows the predicted $g_i(.)$ functions together with prediction intervals separately for each monitor. Rug plots have also been included to give the reader a sense of the distribution of missing values. The estimated degrees of freedoms for the

$g_i(.)$ functions for A. Nagar, V. Nagar and industrial average series were 7, 9. and 9 respectively.

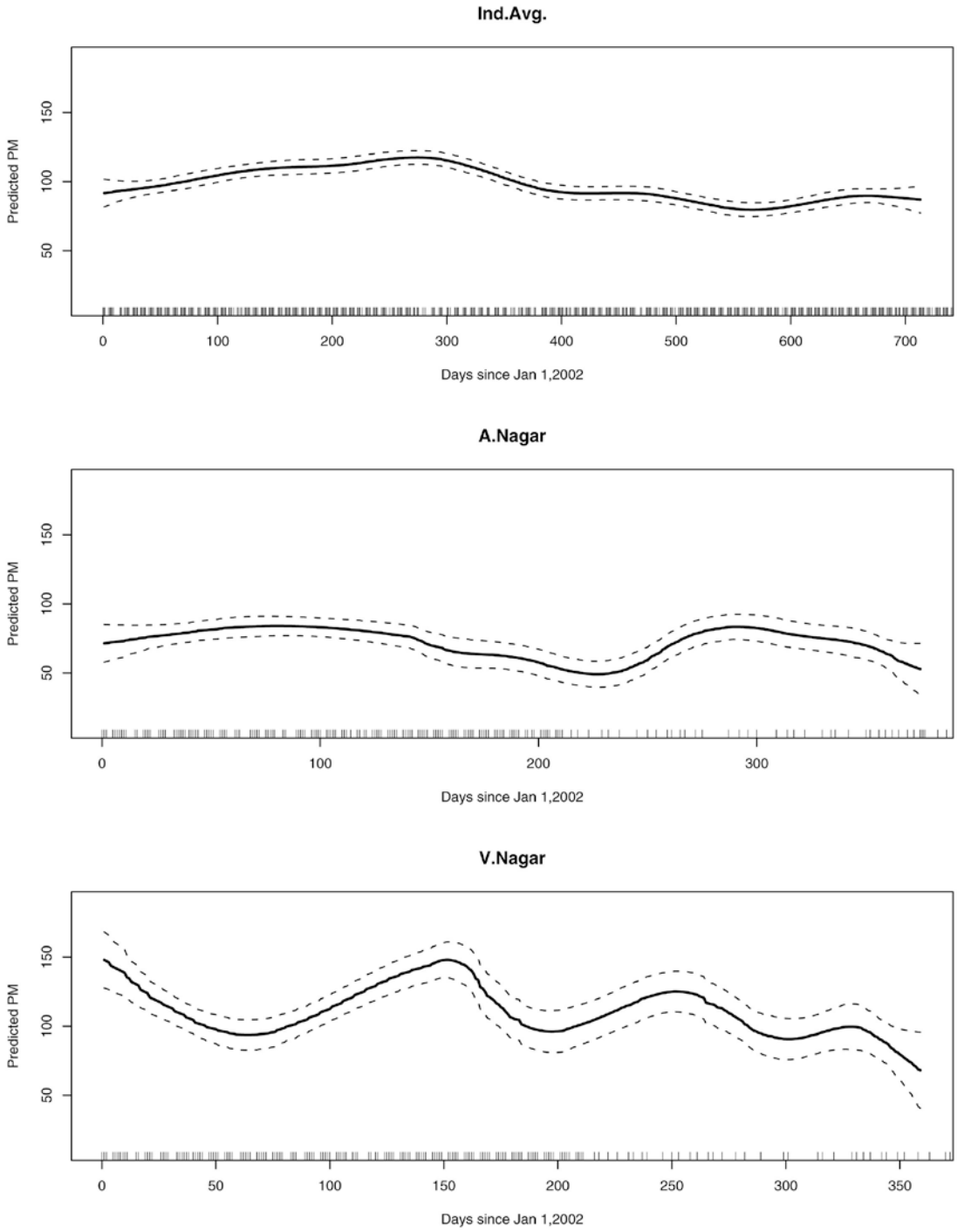

Figure F.2. Smoothed plot of daily PM_{10} concentrations by monitors with pointwise 95% confidence intervals.

Model (2.5) required fitting separate spline regressions by monitor. Rejection of the Bartlett test for equal variances as reported in previous section did not support pooling the data across monitors. Degrees of freedom were chosen separately for each of the above regressions using the conventional approach of minimizing the GCV. Also we did not use the methods of Dominici et al. (2004) as for the single monitor models because here we are in fact interested in the conventional goal of optimizing the predictive power of our model rather than in minimizing confounding error.

The estimated as well as the observed pollution data were then used to fit the following multi-monitor model:

$$\begin{aligned} \log(E(\text{Mortality}_t)) &= \alpha_0 + \beta_1 \text{ANagar}_{t-1} + \beta_2 \text{VNagar}_{t-1} \\ &+ \beta_3 \text{Industrial}_{t-1} + f_1(t) + f_2(temp) + f_3(rh). \end{aligned} \tag{2.6}$$

The degrees of freedom for the confounders were taken to be the same as that in the multi monitor model described in the previous section. Detailed results are presented in Table 10.

A general concern with imputation models in the statistical literature is that subsequently the user proceeds as if there were no missing values so that adjustments for the uncertainty in the imputed values are rarely made. However, an advantage of using a well formulated statistical model such as (2.5) for imputation rather than substitution of ad-hoc values is that this allows the user to account for the uncertainty in the imputed values while estimating the standard error or confidence interval of the *β's.* This can be done by drawing repeated bootstrap

samples from the fitted model in (2.5), estimating β from each of these bootstrap samples and using the empirical standard deviation of these β values as the appropriate standard error. We have followed this procedure and the resulting confidence intervals are given in Table 10.

As per the discussion of Peng et al. (2006) if the $g_i(t_j)$'s are smoother than $f_1(t)$ and if we are able to model $f_1(t)$ using enough basis functions to well represent the relationship between ambient air pollution and time, then the squared bias and the variance of the regression coefficients β can be made to be asymptotically negligible. As we increase the number of basis functions in the representation of $f_1(t)$, the bias diminishes but the variance increases. On the other hand if $g_i(t_j)$'s are more wiggly than $f_1(t)$, unbiased estimation of regression coefficient is possible but with inflated variance. In summary the asymptotic results suggest that modeling $f_1(t)$ with enough degrees of freedom to represent the temporal dependence of ambient air pollution adequately leads to an asymptotically unbiased estimate of the air pollution coefficient. In addition, as we increase the complexity in the representation of $f_1(t)$, the bias of the regression coefficient decreases and its standard error increases.

REFERENCES

Dominici F, Zeger SL, Samet J M. 1999. Combining Evidence on Air Pollution and Daily Mortality from the Largest 20 US cities: A Hierarchical Modeling Strategy. Royal Statistical Society, Series A, with discussion 163: 263-302.

Dominici F, McDermott A, Hastie T. 2004. Improved Semi-parametric Time Series Models of Air Pollution and Mortality. Journal of the American Statistical Association,

468:938-948.

Hastie T, Tibshirani R. 1990. Generalized additive models. Chapman & Hall, New York.

O'Neill MS, Loomis D, Meza VT, Retama A, Gold D. 2002. Estimating particle exposure in the Mexico City metropolitan area, Journal of Exposure Analysis and Environmental Epidemiology 12, 145-156.

Peng RD, Dominici F, Louis TA. 2006. Model choice in time series studies of air pollution and mortality (with discussion). Journal of the Royal Statistical Society, Series A, 169 (2), 179–203.

Wood SN. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.