ISSN 1440-771X
ISBN 0 7326 1086 9

Statistical Methodological Issues in Studies of Air Pollution And Respiratory Disease

Bircan Erbas and Rob J. Hyndman

Working Paper 6/2001

2001

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

MONASH UNIVERSITY
AUSTRALIA

Statistical Methodological Issues in Studies of Air Pollution and Respiratory Disease

Bircan Erbas[1] and Rob J Hyndman[2]

20 September 2001

Abstract: Epidemiological studies have consistently shown short term associations between levels of air pollution and respiratory disease in countries of diverse populations, geographical locations and varying levels of air pollution and climate. The aims of this paper are: (1) to assess the sensitivity of the observed pollution effects to model specification, with particular emphasis on the inclusion of seasonally adjusted covariates; and (2) to study the effect of air pollution on respiratory disease in Melbourne, Australia.

Keywords: air pollution, autocorrelation, generalized additive models, seasonal adjustment, respiratory disease.

[1]Department of Public Health, The University of Melbourne, VIC 3010, Australia.
[2]Department of Econometrics & Business Statistics, Monash University, VIC 3800, Australia.

1 Introduction

1.1 Background

The adverse effects of air pollution on respiratory disease have been widely documented in countries of diverse populations, geography and climate. In fact, increases in respiratory admissions and respiratory mortality suggest adverse effects of air pollutants well below the recommended World Health Organization guidelines (Touloumi et al., 1997).

Recently, there has been some effort to determine the replicability of these findings across a range of exposure outcomes. For example, the APHEA (Air Pollution and Health, a European Approach) produced a standard protocol designed to assess replicability across different countries (Katsouyanni et al., 1996).

We extend this work on replicability by examining the robustness of the estimated relationships between air pollution and respiratory disease under different statistical models. The work is motivated by the idea that applications of different statistical models with varying underlying methodological assumptions may lead to different conclusions regarding the air pollution and respiratory disease relation.

1.2 Data

COPD (Chronic Obstructive Pulmonary Disease) and asthma hospital admissions from all short-stay acute public hospitals in Melbourne, registered on a daily basis by the Department of Human Services (State Government of Victoria), were used as response variables for the period 1 July 1989 to 31 December 1992. International Classification of Disease (ICD) codes for COPD (490–492, 494, 496) and asthma (493) were used to define COPD and asthma.

Melbourne is the second largest city in Australia, with the main source of air pollution emissions from motor vehicles. In Melbourne, levels of sulfur dioxide are relatively low

due to the absence of sulfur-emitting industries. Particulate pollution is highest during autumn and winter, due to the widespread use of wood fires. High levels of nitrogen dioxide (about 65% of total emissions from motor vehicles) and ozone are major constituents of air pollution in this city (Environment Protection Authority, 1999).

Air pollution data were obtained from the Environment Protection Authority (EPA). Maximum hourly values were averaged each day across nine monitoring stations in Melbourne, for nitrogen dioxide, sulfur dioxide, and ozone, all measured in parts-per-hundred-million (pphm). Particulate matter was measured by a device which detects back-scattering (B_{scat}) of light by visibility-reducing particulates between 0.1 and 1μm in aerodynamic diameter. Air particles index (API) were derived from $B_{scat} \times 10^{-4}$. Meteorological data include three hourly maximum daily levels of relative humidity, dry bulbs temperature and dew point temperature. The measures were averaged across four monitoring stations in the Melbourne area.

1.3 Statistical Methodological Issues

A key issue which arises is controlling for seasonal variation in respiratory disease and air pollution. Fourier terms of sine and cosine pairs with varying periods have been accepted as a method to control for seasonal variation in respiratory disease (Hoek et al., 1997, Simpson et al., 1997). However, few studies have controlled for possible seasonality in the covariates (Schwartz 1993, Kelsall et al., 1997, Samet et al., 2000). The sensitivity of the observed effects may change with inclusion of confounding effects for seasonality in model specification. Therefore, it is necessary to control for possible confounding that may induce spurious pollution effects.

To assess the strength and magnitude of seasonal and or cyclic variation in the pollutants and climatic variables, we utilize a method of seasonal adjustment called STL (Seasonal-Trend decomposition based on Loess smoothing) (Cleveland and Terpenning, 1982). Covariates exhibiting strong seasonality were adjusted with the STL method and the resulting seasonally adjusted series were used in subsequent analysis.

We explore the robustness of the pollution-respiratory disease relation using a variety of regression-based approaches, controlling for secular trends, seasonality, and confounding effects of climate. These models include: (1) Generalized Linear Models (GLM); (2) Generalized Additive Models (GAM); (3) Parameter Driven Poisson Regression Models (PDM); and (4) Transitional Regression Models (TRM). In each case, we consider models based on a Poisson distribution, incorporating over-dispersion and serial correlation where possible.

2 Statistical Models

2.1 Generalized Linear Models

For a Generalized Linear Model (GLM) with a log link function, we specify the expectation of a random variable Y_t as

$$\mathrm{E}(Y_t|\boldsymbol{X}_t) = \exp\left(\beta_0 + \sum_{i=1}^{r} \beta_i X_{t,i}\right). \tag{1}$$

Refer to McCullagh and Nelder (1989) for a detailed discussion of GLMs.

Here Y_t denotes daily counts of respiratory disease and air pollution and $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})'$ denotes the explanatory variables at time t. We assume an overdispersed Poisson model, estimated using a quasi-likelihood approach. Akaike's Information Criterion, AIC (Akaike, 1973) was used for variable selection.

2.2 Generalized Additive Models

A nonparametric alternative to the parametric GLM is the Generalized Additive Model (GAM). GAMs allow non-linear relationships between the response variable and each

explanatory variable (Hastie and Tibshirani, 1990). For a GAM, we assume

$$\mathrm{E}(Y_t|\boldsymbol{X}_t) = \exp\left(\beta_0 + \sum_{i=1}^{r} g_i(X_{t,i})\right) \tag{2}$$

where each g_i is a smooth, possibly non-linear, univariate function. Any of the g_i can be made linear to obtain a semi-parametric model. As with a GLM, we use quasi-likelihood estimation.

Cubic smoothing splines were used to estimate the non-parametric functions g_i. We fix the smoothing parameter to be that value for which $\hat{g}_i$ has four "degrees of freedom" (see Hastie and Tibshirani, 1990).

A step-wise model selection procedure in S-PLUS (1999) was used to determine the optimal GAM. Both linear and non-linear terms were allowed for each covariate, and the step-wise procedure automatically selected whether each covariate should be included, and if so, whether it should be linear or non-linear. The AIC was used in this algorithm for variable selection.

2.3 Parameter Driven Models

In a parameter driven model (PDM), serial correlation is set up through an unobservable latent process (Zeger, 1988). A Poisson regression model has conditional mean

$$\mathrm{E}(Y_t|\varepsilon_t, \boldsymbol{X}_t) = \exp(\boldsymbol{X}_t'\boldsymbol{\beta} + \varepsilon_t), \tag{3}$$

where $\boldsymbol{\beta}$ denotes a vector of parameters, and ε_t is a latent process allowing both overdispersion and autocorrelation in Y_t. We allow ε_t to follow a first-order autoregressive process.

2.4 Transitional Regression Models

Transitional Regression Models (TRM) were introduced by Brumback et al., (2000). In this paper we present a special case of a TRM, defined as GLM with time series errors. For a Poisson with AR(1) errors the conditional mean is defined as

$$\mu_t = \exp(\boldsymbol{X}_t'\boldsymbol{\beta}) + \psi_1 e_{t-1}\sqrt{\upsilon_t}, \tag{4}$$

where $e_t = (Y_t - \upsilon_t)/\sqrt{\upsilon_t}$ and $\upsilon_t = \exp(\boldsymbol{X}_t'\boldsymbol{\beta})$.

Here e_t is scaled to give constant variance. Note that $e_t = \psi_1 e_{t-1} + \delta_t$ where $\{\delta_t\}$ is an independent series with zero mean.

3 Results

Each of the four models was fitted to the asthma and COPD hospital admissions data. To simplify the analysis of seasonality, we excluded the leap days of 29 February 1992 in each series. The following covariates were considered for each model.

- Fourier series functions $\sin(2\pi jt/365)$ and $\cos(2\pi jt/365)$ for $j = 1, 2, \ldots, J$. The value of J was chosen using the AIC. For COPD admissions, $J = 4$ and for asthma admissions, $J = 10$.
- Time trend (a quadratic time trend was considered for GLM, PDM and TRM).
- Day of week factor.
- Covariates at time t and lags of up to 5 days.
- Seasonally adjusted climatic variables: dry bulb temperature and humidity.
- Seasonally adjusted pollutants: nitrogen dioxide ($NO_{2,t}$) and ozone ($O_{3,t}$);
- Non-seasonally adjusted pollutants: sulfur dioxide ($SO_{2,t}$) and air particles index (API_t).

For sulfur dioxide and API, there was virtually no seasonality observed. Lagged values of each of the climatic and pollutant covariates were considered up to five days previously.

To allow comparison across different statistical models we use the following three measures:

- Mean square error (MSE) = mean $\{(Y_t - \hat{Y}_t)^2\}$, where $\hat{Y}_t$ are the (inverse link transformed) fitted values.
- Mean square proportional error (MSPE) = mean $\{(Y_t - \hat{Y}_t)^2/\hat{Y}_t\}$.
- AIC = $n\log(\sigma^2) + 2p$, where σ^2 is the variance of the raw residuals (response minus fitted values), and p is the number of degrees of freedom in each model.

3.1 COPD hospital admissions in Melbourne, Australia from 1 July 1989 to 31 December 1992

Table 1 displays results from the analyses of COPD hospital admissions, using different statistical methods. Where a variable has been included in a linear function, the relative risk is shown. For the GAM, variables which were included using a smoothing spline are denoted by $g(\cdot)$.

Table 1: *Relative Risk and 95% CI of COPD hospital admissions for an increase from the 10th to 90th percentile for levels of pollutants, generated using different statistical methods.*

COPD								
	GLM		GAM		PDM		TRM	
Pollutant	RR	95%CI	RR	95% CI	RR	95%CI	RR	95% CI
$NO_{2,t}$	1.06	1.00–1.12	1.06	1.01–1.11	1.05	1.00–1.11	1.05	1.00–1.10
$O_{3,t-2}$			1.06	1.00–1.11				
API_{t-2}			0.95	0.91–1.00				
$SO_{2,t-2}$			g( )					
MSE	13.23		12.76		12.88		12.29	
MSPE	1.24		1.19		1.13		1.16	
AIC	3340.42		3292.19		3252.84		3243.09	

Daily COPD hospital admissions increased significantly with increased ambient outdoor levels of same day nitrogen dioxide (NO_2). The estimated nitrogen dioxide coefficients

from the models in Table 1 were consistent, replicated across different models, and were statistically significant ($p < 0.5$), which is indicative of a robust relationship between nitrogen dioxide and COPD hospital admissions.

The observed effects for ambient outdoor levels of ozone, particulates and sulfur dioxide were all highly sensitive to model specification.

A GAM analysis showed a nonlinear relationship between sulfur dioxide and COPD hospital admissions in Melbourne, Australia (see Figure 1). This is similar to that found in London (Schwartz and Marcus, 1990) and Europe (Touloumi et al., 1994).

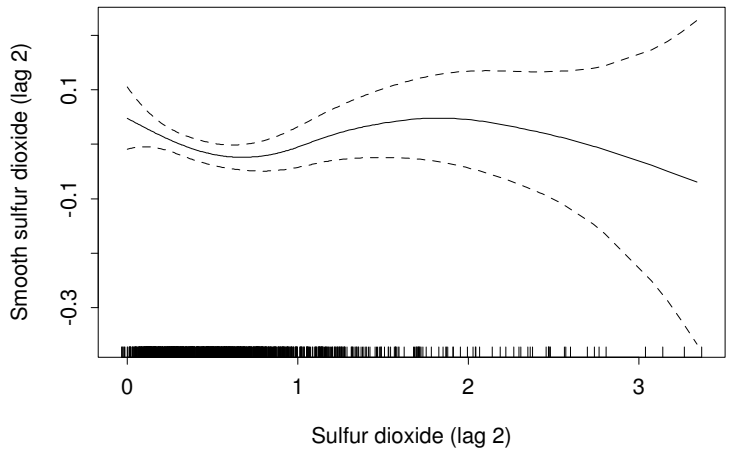

Figure 1: *The nonlinear function for sulfur dioxide (lagged 2 days). Dashed lines represent pointwise 95% confidence intervals.*

Figure 2 displays the residual autocorrelation function for each of the models in Table 1. The GLM is inadequate because of significant serial correlation. The other three models fare better, although there is some significant correlation remaining in the GAM and PDM.

Of these three models, the TRM is ranked highest on the basis of AIC and MSE, and the PDM is best on the basis of MSPE. However, the GAM has the important advantage that it explains more of the variation in COPD admissions through the structure of the covariates than through the correlation terms.

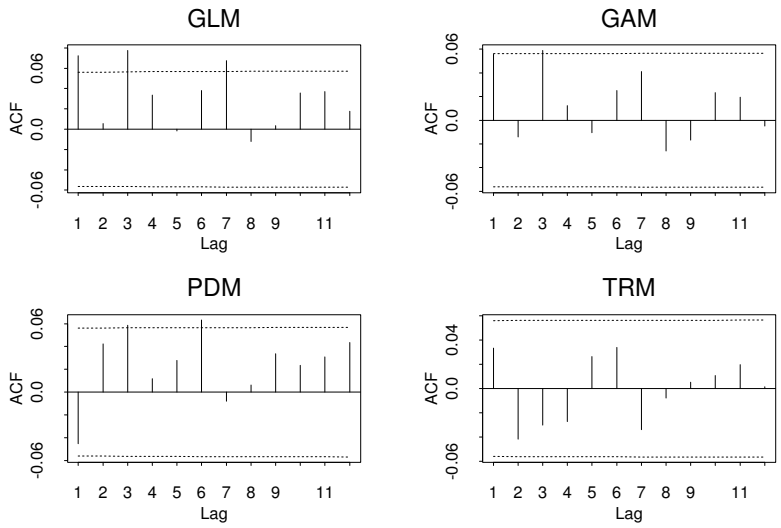

Figure 2: *The autocorrelation function for the deviance residuals from the models applied to COPD hospital admissions in Table 1.*

3.2 Asthma hospital admissions in Melbourne, Australia from 1 July 1989 to 31 December 1992

Table 2 displays results from the analyses of asthma hospital admissions, using different statistical methods.

Table 2: *Relative Risk and 95% CI of asthma hospital admissions for an increase from the 10th to 90th percentile for levels of pollutants, generated using different statistical methods.*

Asthma								
	GLM		GAM		PDM		TRM	
Pollutant	RR	95%CI	RR	95% CI	RR	95%CI	RR	95% CI
$NO_{2,t}$	1.05	1.01–1.08	1.05	1.01–1.09	1.04	1.01–1.08	1.05	1.02–1.08
$NO_{2,t-1}$			0.96	0.92–0.99				
$O_{3,t}$			0.97	0.93–1.00				
$O_{3,t-1}$	0.96	0.93–0.99			0.97	0.94–1.09	0.97	0.95–0.99
$O_{3,t-2}$			g( )					
API_t								
$SO_{2,t}$								
MSE	57.81		53.68		56.05		55.8	
MSPE	1.75		1.61		1.70		1.69	
AIC	5244.03		5153.41		5207.56		5206.92	

Similar results were found for asthma hospital admissions and same day nitrogen dioxide. The observed effects for same day nitrogen dioxide in Table 2 were robust to different model specifications, but lagged 1 day effects were not agreeable. The observed effects

for ambient outdoor levels of ozone and API were highly sensitive to model specification.

A GAM analysis of asthma hospital admissions showed a significant nonlinear effect of ozone lagged 2 days. This result is displayed in Figure 3.

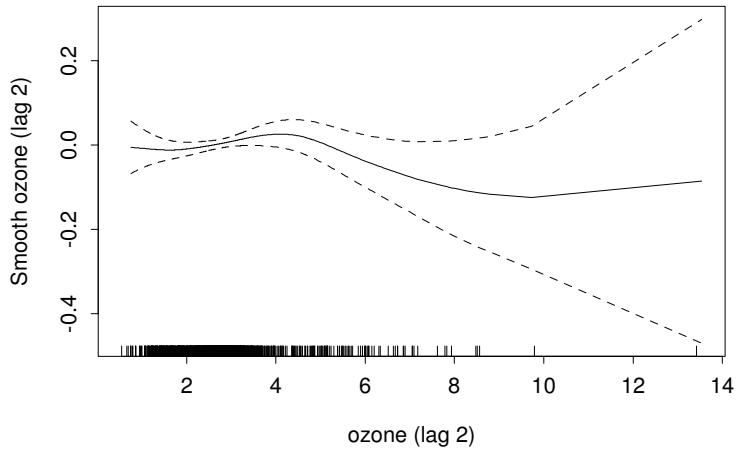

Figure 3: *The nonlinear function for ozone (lagged 2 days). Dashed lines represent pointwise 95% confidence intervals.*

Figure 4 displays the residual autocorrelation function for each of the models in Table 2. For asthma hospital admissions, the GAM is best on all three criteria, although all models were inadequate due to the strong and significant correlation pattern in the residuals.

4 Discussion and Conclusions

This study extends recent epidemiological studies by focusing on the following question: How robust is the observed pollution-respiratory disease relation to different statistical models with various underlying methodological assumptions?

The statistical methodologies adopted in this study are all variations of regression methods. They range from nonnormal methods (generalized linear and additive models), to recently developed parameter and observation driven models (Poisson regression with autocorrelation and transitional regression models).

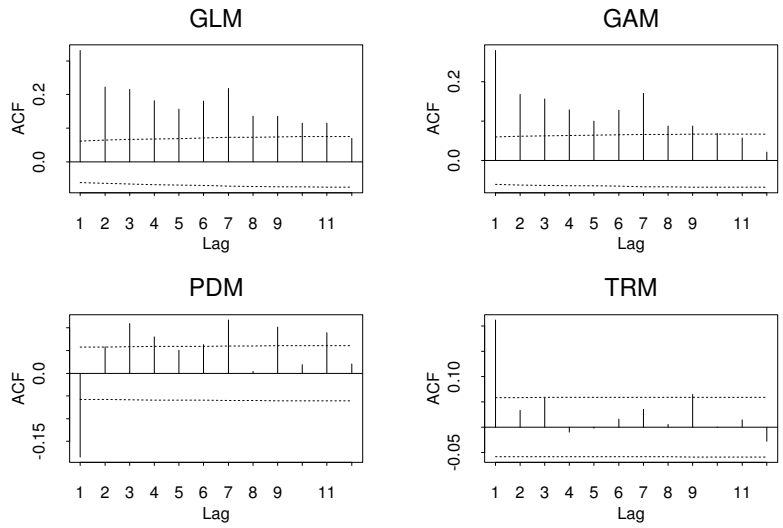

Figure 4: *The autocorrelation function for the deviance residuals from the models applied to asthma hospital admissions in Table 2.*

The findings from this study show that the relation between ambient outdoor concentrations of nitrogen dioxide and COPD hospital admissions is robust to different statistical methodology. The positive same day relationship between respiratory admissions and nitrogen dioxide reported in this study is similar to a study of air pollution and hospital admissions in Sydney, Australia by Morgan et al., (1998a). A positive result was also reported between respiratory mortality and nitrogen dioxide in a Sydney study by Morgan et al., (1998b) and a recent Melbourne EPA study (Melbourne Mortality Study, 2000).

The observed effects of ozone on both COPD and asthma hospital admissions were highly sensitive to model specification. Both the Melbourne mortality study (EPA Melbourne Mortality Study, 2000) and the Sydney study (Morgan et al., 1998b) report positive associations between ozone and respiratory mortality. A study conducted in Brisbane (Simpson et al., 1997) also confirms this association. A six-city European study (Anderson et al., 1997) established significant associations between same day and lagged 1 day ozone and daily admissions for COPD. They report these associations as the strongest and most replicable.

The relationship between particulates (API) and both COPD and asthma hospital admissions were non-robust. However, a negative association is similar to other studies in

Australia. The Sydney hospital admissions study (Morgan et al., 1998a) report negative effects of particulates (measured similar to the Melbourne study) on asthma hospital admissions. The Brisbane mortality study (Simpson et al., 1997) report a non–significant negative association between respiratory mortality and maximum 1 hour BSP (can be compared with $PM_{2.5}$). The relationship between sulfur dioxide and both COPD and asthma admissions were non-robust.

Table 3 displays the strengths and weaknesses of each model used in this study. A + indicates a strength and – indicates a weakness of the methodology.

Table 3: *Strengths and weaknesses of the statistical methods used in this study.*

	GLM	**GAM**	**PDM**	**TRM**
Methodological Issues				
Nonnormality	+	+	+	+
Overdispersion	+	+	+	+
Nonlinearity	–	+	–	–
Autocorrelation	–	–	+	+

The statistical methods presented in this study were inadequate in addressing all the methodological issues common to studies of respiratory disease and air pollution. For COPD hospital admissions a poisson regression model with AR(1) errors performed best, although a GAM was in some ways preferable because of its superior ability to explain variation through the structure of the covariates. For asthma hospital admissions, a GAM performed best, although no model was completely satisfactory in representing the strong correlation structure in the residuals.

GAM methodology is a flexible method that accounts for complex covariate effects. A recent study by Coull et al., (2000) extends the nonparametric framework, with the inclusion of a time series error structure for the residuals and including random effects to reflect population heterogenity resulting in an additive mixed models analysis for a normally distributed response variable. The extension of these models to nonnormally distributed outcomes would be of great interest and potential value in modeling respiratory disease and air pollution. Studies of respiratory disease have clearly demonstrated that the outcome (respiratory disease) is not normally distributed and further developments

in the statistical methodology should reflect this.

Acknowledgements: This study was funded by a Public Health Ph.D. Postgraduate Scholarship from the National Health and Medical Research Council.

5 References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory.* B.N. Petrov & F. Csaki (eds), Adademiai Kidao, Budapest; pp. 267–281.

Anderson, H.R., Spix, C., Medina, S., Schouten, J., Castellsague, J., Rossi, G., Zmirou, D., Touloumi, G., Wojtyniak, B., Ponka, A., Bacharova, L., Schwartz, J., & Katsouyanni, K. (1997) Air pollution and daily admissions for chronic obstructive pulmonary disease in 6 European cities: results from the APHEA project. *European Respiratory Journal*, **10**, 1064–1071.

Cleveland, W.S., & Terpenning, I.J. (1982) Graphical methods for seasonal adjustment. *Journal of the American Statistical Association*, **77**, 52–62.

EPA, "EPA Air Quality Information" (1999) Environment Protection Authority, Victoria, Australia. *http://www.epa.vic.gov.au/aq/info/*

EPA, "EPA Melbourne Mortality Study" (2000) Environment Protection Authority, Victoria, Australia. *http://www.epa.vic.gov.au/aq/info/*

Hastie, T., & Tibshirani, R.J. (1990) *Generalized additive models.* London: Chapman and Hall.

Hoek, G., Schwartz, J., Groot, B., & Eilers, P. (1997) Effects of ambient particulate matter and ozone on daily mortality in Rotterdam, The Netherlands. *Archives of Environmental Health*, **52**, 455–463.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J.M., Tobias, A., Ponka, A., Medina, S., Bacharova, L. & Anderson, H.R. (1996) Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *Journal of Epidemiology and Community Health, Supplement*, **50**, S12–S18.

Kelsall, J.E., Samet, J.M., Zeger, S.L., & Xu, J. (1997) Air pollution and mortality in Philadelphia, 1974–1988. *American Journal of Epidemiology*, **146**, 750–762.

McCullagh, P., & Nelder J.A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Morgan, G., Corbett, S., Wlodarcyzk, J. (1998a) Air pollution and hospital admissions in Sydney, Australia, 1990-1994. *American Journal of Public Health*, **88**, 1761–1766.

Morgan, G., Corbett, S., Wlodarcyzk, J. (1998b) Air pollution and daily mortality in Sydney, Australia, 1989–1993. *American Journal of Public Health*, **88**, 759–764.

Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I., & Zeger, S.L. (2000) Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. *The New England Journal of Medicine*, **343**, 1742–1749.

Schwartz, J (1993) Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology*, **137**, 1136–1147.

Schwartz, J., & Marcus, A. (1990). Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology*, **131**, 185–194.

Simpson, R., Williams, G., Pteroeschevsky, A., Morgan, G., & Rutherford, S. (1997) Associations between outdoor air pollution and daily mortality in Brisbane, Australia. *Archives of Environmental Health*, **52**, 442–454.

S-PLUS (1999). *Modern statistics and advanced graphics.* Seattle, Washington: MathSoft, Inc.

Touloumi, G., Katsouyanni, K., Zmirou, D., Schwartz, J., Spix, C., Ponce de Leon, A., Tobias, A., Quennel, P., Rabczenko, D., Bacharova, L., Bisanti, L., Vonk, J.M., & Ponka, A. (1997) Short-term effects of ambient oxidant exposure on mortality: a combined analysis within the APHEA project. *American Journal of Epidemiology*, **146**, 177–185.

Touloumi, G., Pocock, S.J., Katsouyanni, K., & Trichopoulos, D. (1994). Short-term effects of air pollution on daily mortality in Athens: a time series analysis. *International Journal of Epidemiology*, **23**, 957–967.

Zeger, S. (1988). A regression model for time series of counts. *Biometrika*, **75**, 621–629.