**Web appendix**

**Model checking – useful diagnostic plots**

1. *Scatter plot of deviance residuals vs time.*

A plot of residuals will ideally show an even band of points with no particular pattern over time, and around 95% of values lying between the values ±1.96 (See Figure A1 for such a plot from the London data analysis). Deviations from this ideal may indicate, for example, that important variables are missing from the model, or that long-term patterns have been inadequately controlled for.

1. *Partial autocorrelation function plot of the deviance residuals*

This plot (available in major statistical packages) estimates the degree of correlation between nearby days (controlling for other nearby days). Ideally there should be little remaining autocorrelation in the deviance residuals, because we hope that the variables in the model explain most of the autocorrelation in the original outcome series. However it is common for some autocorrelation to remain, particularly at short lags. This represents a technical violation of the Poisson model assumptions, and as a sensitivity analysis or routinely, the residual autocorrelation can be explicitly modelled – for details, see Brumback et al.[1](#_ENREF_1) In the London data, there appeared to be a high degree of autocorrelation at a lag of 1 day (Figure A2a). Adding the 1-day lagged deviance residual to the model is one simple modelling approach. Here, this reduced autocorrelation (Figure A2b) and, reassuringly, the estimated exposure-outcome association and its standard error changed very little. In applications where residual autocorrelation is a serious concern, more sophisticated modelling methods exist, based on conditional (random effects) or marginal models (generalised estimating equations).[1](#_ENREF_1), [2](#_ENREF_2) If the outcome is counts of an infectious disease, residual autocorrelation can be much stronger than in our example and, with changes in numbers of susceptibles, can require special methods.

**Further relevant papers**

*General references on time series regression applications in health research:*

Peng and Dominici 2008 - Statistical Methods for Environmental Epidemiology with R - A Case Study in Air Pollution and Health[3](#_ENREF_3)

*Time series for short-term effects:*

Dominici et al 2003 - Health effects of air pollution: a statistical review[4](#_ENREF_4)

Touloumi et al 2004 - Analysis of health outcome time series data in epidemiological studies[5](#_ENREF_5)

*Splines and smoothing techniques:*

Wood 2006 - Generalized Additive Models: an Introduction with R[6](#_ENREF_6)

Ruppert et al 2003 - Semiparametric Regression[7](#_ENREF_7)

Schimek 2009 - Semiparametric penalized generalized additive models for environmental research and epidemiology[8](#_ENREF_8)

*Distributed lag models:*

Schwartz 2000 - The distributed lag between air pollution and daily deaths[9](#_ENREF_9)

Gasparrini et al 2010 - Distributed lag non-linear models[10](#_ENREF_10)

Gasparrini 2011 – Distributed lag linear and non-linear models in R: the package dlnm[11](#_ENREF_11" \o "Gasparrini, 2011 #1651)

*Model selection:*

Peng et al 2006 - Model choice in time series studies of air pollution and mortality[12](#_ENREF_12)

Touloumi et al 2006 - Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates[13](#_ENREF_13)

A

**Web appendix references**

1. Brumback B, Ryan LM, Schwartz JD, Neas LM, Stark PC, Burge HA. Transitional Regression Models, with Application to Environmental Time Series. *Journal of the American Statistical Association* 2000; **95**: 16-27.

2. Baccini M, Biggeri A, Accetta G, et al. Heat effects on mortality in 15 European cities. *Epidemiology* 2008; **19**(5): 711-9.

3. Peng RD, Dominici F. Statistical methods for environmental epidemiology with R - a case study in air pollution and health: Springer; 2008.

4. Dominici F, Sheppard L, Clyde M. Health effects of air pollution: a statistical review. *International Statistical Review* 2003; **71**: 243-76.

5. Touloumi G, Atkinson RW, Le Tertre A, et al. Analysis of health outcome time series data in epidemiological studies. *Environmetrics* 2004; **15**: 101-17.

6. Wood SN. Generalized additive models: an introduction with R: Chapman & Hall/CRC; 2006.

7. Ruppert D, Wand MP, Carroll RJ. Semiparametric regression: Cambridge University Press; 2003.

8. Schimek MG. Semiparametric penalized generalized additive models for environmental research and epidemiology. *Environmetrics* 2009; **20**(6): 699-717.

9. Schwartz J. The Distributed Lag between Air Pollution and Daily Deaths. *Epidemiology* 2000; **11**(3): 320-6.

10. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. *Stat Med* 2010; **29**(21): 2224-34.

11. Gasparrini A. Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. *J Stat Softw* 2011; **43**(8): 1-20.

12. Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollution and mortality. *J Roy Stat Soc a Sta* 2006; **169**: 179-98.

13. Touloumi G, Samoli E, Pipikou M, Le Tertre A, Atkinson R, Katsouyanni K. Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates. *Stat Med* 2006; **25**(24): 4164-78.

*Figure A1: Plot of deviance residuals over time (London data)*

****

Notes:

- the deviance residuals relate to the unconstrained distributed lag model with ozone (lag days 0 to 7 inclusive), adjusted for temperature at the same lags

- the spike in the plot of residuals relate to the 2003 European heat wave, and indicate that the current model does not explain the data over this period well; more flexible modelling of the temperature-mortality relationship (not shown) results in this spike in residuals disappearing

*Figure A2: Partial autocorrelation plots of the deviance residuals (London data)*

****

Note: the deviance residuals relate to the unconstrained distributed lag model with ozone (lag days 0 to 7 inclusive), adjusted for temperature at the same lags