Int. J. Epidemiol. Advance Access published June 12, 2013



*International Journal of Epidemiology* 2013;1–9
doi:10.1093/ije/dyt092

# Time series regression studies in environmental epidemiology

**Krishnan Bhaskaran,[1]* Antonio Gasparrini,[2] Shakoor Hajat,[3] Liam Smeeth[1] and Ben Armstrong[3]**

[1]Department of Non-Communicable Diseases Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, [2]Medical Statistics Department, London School of Hygiene and Tropical Medicine, London, UK and [3]Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, London, UK

*Corresponding author. Department of Non-Communicable Diseases Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: krishnan.bhaskaran@lshtm.ac.uk

**Accepted** 24 April 2013

Time series regression studies have been widely used in environmental epidemiology, notably in investigating the short-term associations between exposures such as air pollution, weather variables or pollen, and health outcomes such as mortality, myocardial infarction or disease-specific hospital admissions. Typically, for both exposure and outcome, data are available at regular time intervals (e.g. daily pollution levels and daily mortality counts) and the aim is to explore short-term associations between them. In this article, we describe the general features of time series data, and we outline the analysis process, beginning with descriptive analysis, then focusing on issues in time series regression that differ from other regression methods: modelling short-term fluctuations in the presence of seasonal and long-term patterns, dealing with time varying confounding factors and modelling delayed ('lagged') associations between exposure and outcome. We finish with advice on model checking and sensitivity analysis, and some common extensions to the basic model.

**Keywords** Time series, environmental epidemiology, air pollution



## Introduction

This article aims to introduce the reader to the methodological features and analytical issues involved in a study design commonly used in environmental epidemiology: the time series regression study. The design is often used in studies attempting to quantify short-term associations of environmental exposures, such as air pollution, pollen, dust and weather variables, with health outcomes.[1–3] We aim to provide the reader with an insight into some of the unique features and challenges involved in analysing time series data. It is hoped that 'consumers' of studies will gain insight into the methods and an understanding of specialist terminology used in this context, enabling more effective critical interpretation and appraisal of study reports; and also that epidemiologists who may be in a position to analyse time series datasets will find this a useful tutorial covering the key steps and important issues involved in actually carrying out a time series regression analysis. Our intention is to complement other articles which offer historical perspectives, more mathematical developments of the modelling ideas than presented here, and which cover issues uniquely relevant to specific exposures such as ambient temperature and particulate matter[4–7](further references are listed in the Supplementary appendix, available as Supplementary data at *IJE* online). It should be noted that though our focus is on time series regression, other tools for the analysis of time series data exist. Time series data occur frequently in

econometrics; some methods that are commonly used in that field aim to forecast movements in a single time series (e.g. a stock market price), and would be of limited interest to epidemiologists, but others could in principle be applied to epidemiological questions. An example is the Granger causality test, which aims to establish, via a hypothesis testing paradigm, whether movements in one time series are causally related to movements in another. We do not consider this or other methods more commonly associated econometrics further in this paper.[8]

Throughout, concepts and methods will be illustrated through an example based on a real dataset, and Stata and R code to reproduce our analyses, along with the dataset itself, are available as a Supplementary Appendix at *IJE* online.

## Data features and introduction to worked example

The illustrative example we will use is a time series regression analysis of a dataset from London. The dataset consists of a single observation for every day from 1 January 2002 to 31 December 2006, and for each day there is a measure of (mean) ozone levels that day, and the total number of deaths that occurred in the city. The question to be addressed is 'Is there an association between day-to-day variation in ozone levels and daily risk of death?', so the exposure of interest is ozone and the outcome is death. The dataset also contains daily measures of two potential confounders, temperature and relative humidity (confounding is discussed later in the paper). The first 12 rows of data are shown in Table 1. Some features worth noting are:

- Generally, a 'time series' is simply a sequence of data points recorded at regular time intervals. So in this dataset there are actually four time series (ozone, temperature, relative humidity and number of deaths), and the aim is to say something about if/how these are associated.
- The main unit of analysis (represented by a row of data) is the day and not the individual person. This will be an important point when we come to consider what the potential confounders might be in our analysis. Note however that a time series regression study does not necessarily have to be at the daily level; annual, monthly, weekly, or even hourly time series data could be analysed using the same broad methodological principles.
- The outcome is a count, which is common for time series regression studies. The denominator (the underlying population size) is not part of the dataset, which is not a concern because in these data we are usually interested in modelling variation in outcome from day to day or week to week, and population size is unlikely to change meaningfully over these timescales, so can be safely omitted from the analysis.

**Table 1** Example rows of time series data from the London dataset showing daily levels of environmental variables and daily number of deaths

| Date | Ozone ($\mu g/m^3$) | Temperature (°C) | Relative humidity (%) | *n* deaths |
|---|---|---|---|---|
| 1-Jan-02 | 4.59 | −0.2 | 75.7 | 199 |
| 2-Jan-02 | 4.88 | 0.1 | 77.5 | 231 |
| 3-Jan-02 | 4.71 | 0.9 | 81.3 | 210 |
| 4-Jan-02 | 4.14 | 0.5 | 85.4 | 203 |
| 5-Jan-02 | 2.01 | 4.3 | 93.5 | 224 |
| 6-Jan-02 | 2.4 | 7.1 | 96.4 | 198 |
| 7-Jan-02 | 4.08 | 5.2 | 93.5 | 180 |
| 8-Jan-02 | 3.13 | 3.5 | 81.5 | 188 |
| 9-Jan-02 | 2.05 | 3.2 | 88.3 | 168 |
| 10-Jan-02 | 5.19 | 5.3 | 85.4 | 194 |
| 11-Jan-02 | 3.59 | 3.0 | 92.6 | 223 |
| 12-Jan-02 | 12.87 | 4.8 | 94.2 | 201 |

### Descriptive analysis

The first step should be familiar to epidemiologists from all specialties: getting to know the data through simple plots and tables. Figure 1 shows scatter plots of both the exposure (ozone) and outcome (number of deaths) over time for the entire study period; a plot of this type can quickly reveal high-level patterns in the data. Moving average plots can also be used to supplement raw scatter plots and draw out patterns—such plots effectively smooth out the raw data by averaging over a fixed number of adjacent raw data points. In this case, the raw plots show that both ozone levels and death counts seem to be dominated by annual seasonal patterns, with ozone highest in summer and lowest in winter, and the opposite pattern for deaths. Note that one would not generally infer from this that low ozone levels in winter are a 'cause' of the higher mortality: systematic patterns over time are present in many time series, inducing correlations that are in most cases unlikely to represent causal relationships. It is for this reason that our aim is to consider associations over relatively short timescales, which are more likely to represent real causal relationships.[6]

Other informative descriptive analyses might include summary statistics, a correlation matrix for the covariates to be included in the model and an exploration of missing data. Our example dataset contains no missing values, but there are frequently missing exposure data that need to be handled in the initial data processing or in the analysis itself: dropping incomplete records is a simple strategy but may introduce bias; employing rules, algorithms or models to impute the missing data (singly or multiply) are alternatives,[9] but are not considered further here.

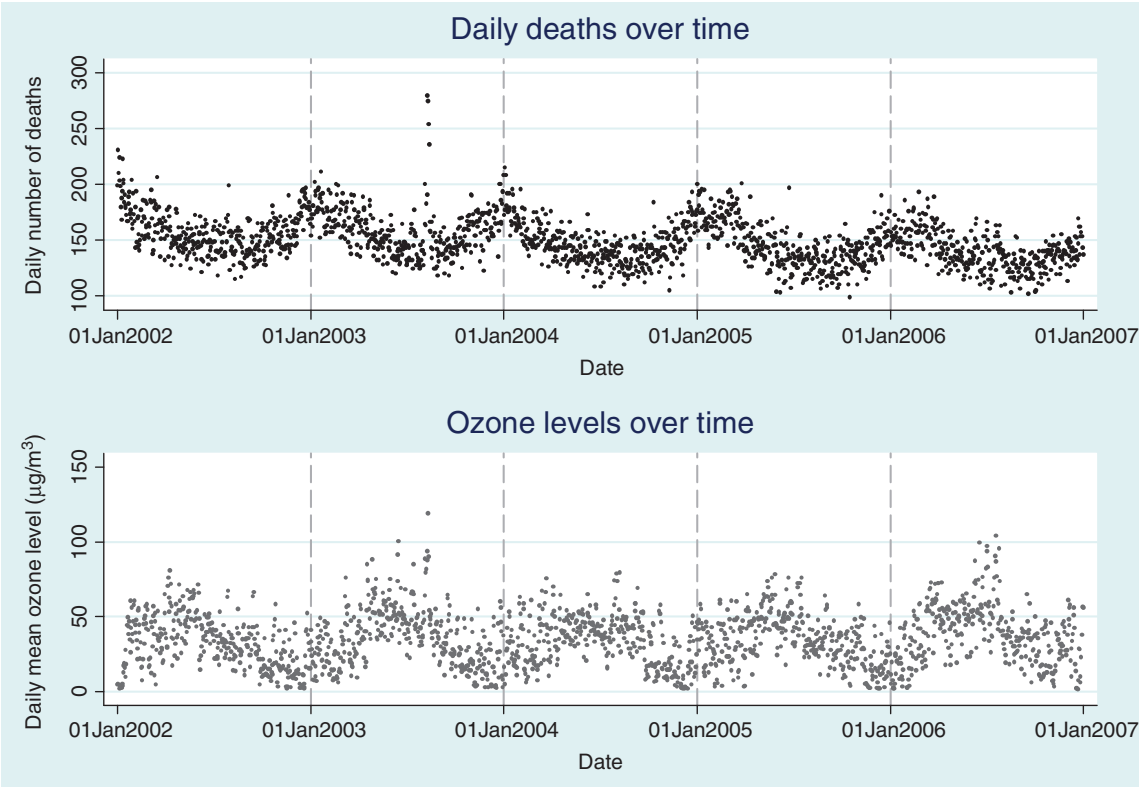


**Figure 1** Raw plots showing outcome (deaths) and exposure (ozone) data over time (London data)

## Time series regression

After carrying out some initial descriptive analyses, the next step is to begin to develop a regression model (see Analysis using Poisson regression) that will enable us to address our principal study question.

### Aim of regression analysis

The main aim of regression is to investigate whether some of the short-term variation in the outcome can be explained by changes in the main exposure; in our example, whether day-to-day changes in the number of deaths are explained in part by changes in the levels of ozone in the air. A regression approach will also allow control for multiple potential confounding factors.

### Analysis using Poisson regression

The outcome variable here is a count (the number of deaths each day). The usual regression method of choice for analysing count data is Poisson regression, but we need to bear in mind some of the unique features of time series data of this type:

- In the raw data, long-term patterns including seasonality are likely to dominate the data (as in our example). As our interest is in short-term associations, the aim is to remove (i.e. control for) these long-term patterns, and see whether the exposure of interest explains some of the remaining short-term variation. Possible strategies to control for long-term patterns are covered in detail in the next section.
- An assumption of Poisson regression is violated in the raw data: observations are unlikely to be independent, with observations close in time likely to be more similar than those distant in time (in the London dataset, this is very clear from Figure 1). However, this 'autocorrelation' is usually not intrinsic to the outcome series, but rather due to autocorrelation in the explanatory variables that are predictors of the outcome. After controlling for seasonality, long-term patterns, the exposure of interest and other explanatory variables, residual autocorrelation will tend to be much smaller than in the raw outcome data, and is usually not a major concern. Nevertheless, at the model checking stage it may be a good idea specifically to model any remaining autocorrelation and check that our conclusions do not change (see Model checking and sensitivity analysis).
- The data tend to be 'overdispersed', meaning that the variance of the outcome counts is higher than predicted under a Poisson distribution (in which variance = mean), so it is necessary to apply a simple adjustment to obtain appropriate standard errors in the model fitting (specifically, a scale parameter is applied estimated by the Pearson

chi-square statistic divided by the residual degrees of freedom[10]).

## Controlling for seasonality and long-term trends

To reiterate, the research question to be addressed is whether short-term variation in the outcome is explained by the exposure of interest; i.e. in our example, whether day-to-day changes in mortality are related to daily ozone levels. But the raw outcome data are likely to be initially dominated by seasonal patterns and long-term trends (Box 1), so it is necessary to control for these patterns in the regression model in order to effectively separate them out from the short-term associations between exposure and outcome that we are interested in. There are a number of ways to achieve this, but what they have in common is that some function of time is fitted as part of the regression model.

### *Option 1: Time stratified model (simple indicator variables)*

- A simple way of approximately modelling long-term patterns in the outcome data is to split the study period into intervals and estimate for each interval a different baseline mortality risk. In practice, this means simply including an indicator variable for each time interval in the Poisson model. One possible choice of time interval for daily data is elapsed calendar month, such that in these data there are 12 X 5 = 60 strata.
- Pros: easy to understand, and often captures main long-term patterns quite well.
- Cons: potentially large number of model parameters; implicitly assumes biologically implausible jumps in risk between adjacent time intervals.
- Figure 2a illustrates the predicted numbers of deaths from such a calendar-month stratified model applied to the London data.

**Box 1 Studying short-term associations in the presence of longer-term variation**

Seasonal and long-term patterns in both the exposure and outcome data can dominate crude associations, making the short-term associations of interest hard to detect.

By explicitly controlling for long-term patterns, the association between the exposure variable(s) of interest and the short-term variation around these long-term patterns can be explored.

### *Option 2: Periodic functions (Fourier terms)*

- Long-term patterns can be modelled more smoothly by fitting Fourier terms in the Poisson model. These are pairs of sine and cosine functions of time with an underlying period reflecting the full seasonal cycle (i.e. calendar year), and are particularly suited to capturing very regular seasonal patterns. A single sine/cosine pair will model seasonal variation in the outcome as a regular wave with a single (equally spaced) peak and trough per calendar year (the actual position of the peak and trough are guided by the data). However, harmonics (extra sine/cosine pairs with shorter wavelengths) can also be introduced which results in more flexible functions.
- Pros: models long-term patterns smoothly, using relatively few parameters.
- Cons: more mathematically complex than the time-stratified model; the modelled seasonal pattern is always forced to be the same from one year to the next, which may not reflect the data well (e.g. timing of winter peaks in deaths may vary). Fourier terms alone cannot capture long-term non-seasonal trends (this can be solved by adding a further function of calendar time).

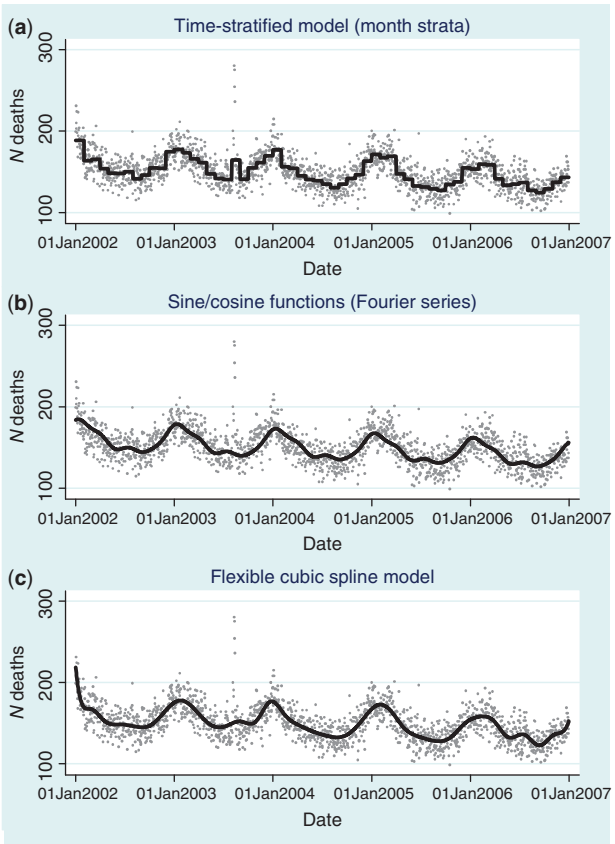


**Figure 2** Three alternative ways of modelling long-term patterns in the data (seasonality and trends)

- Figure 2b illustrates model fit for the London data using 4 sine/cosine pairs (1 fundamental plus 3 harmonics) to capture seasonality, plus a linear function of time to capture broader trends over time.

### *Option 3: Flexible spline functions*

- The third option it to fit a spline function of time; this is essentially a number of different polynomial (most commonly cubic) curves that are joined smoothly end-to-end to cover the full period. To fit a spline function in practice, we first generate a set of basis variables which are functions of the main time variable, and then include these basis variables in the Poisson model. In generating the spline basis, it is necessary to decide how many knots (join-points) there should be, which governs how many end-to-end cubic curves will be used and therefore how flexible the curve will be: too few will fail to capture the main long-term patterns closely, whereas too many will result in a very 'wobbly' function which may compete with the variable of interest to explain the short-term variation of interest, widening confidence intervals of relative risk estimates. The flexibility of the spline function is sometimes framed in terms of number of degrees of freedom rather than number of knots, where more degrees of freedom corresponds to more knots, and both imply a more flexible function.
- Pros: models long-term patterns smoothly; can capture seasonal patterns in a way that is allowed to vary from one year to the next; and will also capture long-term non-seasonal trends in the data.
- Cons: more mathematically complex than the other methods (though functions to generate the spline basis are available in major statistical packages).
- Figure 2c illustrates a spline function applied to the London data, using 34 knots [ = (number of calendar years x 7) − 1], a common choice for daily mortality data. Although there is no consensus on how many knots are optimal, 7 per year has been justified as a balance between providing adequate control for seasonality and other confounding by trends in time, while leaving sufficient information from which to estimate exposure effects).[11]

### Residual variation around the long-term pattern

If seasonality and long-term trends are controlled for using one of the above approaches, we will be left with residual variation in which the long-term patterns are no longer apparent (Figure 3). By adding the exposure of interest to this model, we can now tackle our main aim, which is to investigate whether the remaining short-term variation around the long-term pattern is in part explained by the exposure variable(s).

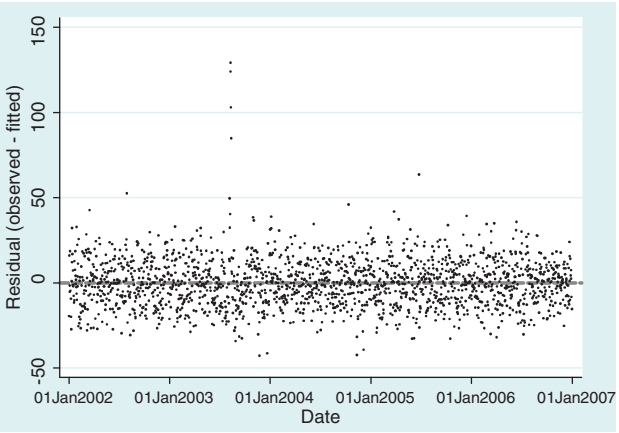


**Figure 3** Residual variation in daily deaths after 'removing' (i.e. modelling) season and long-term trend. Fitted values were from a spline model for season and long-term trend only (as illustrated in Fig 2c)

## Exposure-outcome associations and confounding

In the London data, fitting a naive Poisson model for mortality, with ozone as the only explanatory variable and no adjustment for seasonality or long-term trends, suggests that each 10 $\mu g/m^3$ increase in ozone levels is associated with a mortality risk ratio of 0.991 (95% CI 0.987 to 0.994, $P < 0.001$), i.e. higher ozone is associated with lower mortality risk. But we know that at least part of this is likely to be explained by confounding by season. After adding adjustment for season and long-term trend to the model (using a flexible spline as in Option 3 above), the direction of the estimated effect reverses (RR per 10 $\mu g/m^3 = 1.007$, 95% CI 1.003 to 1.010, $p < 0.001$; or equivalently, a 0.7% [0.3–1.0] risk increase), suggesting that in the short term, higher ozone is associated with higher mortality risk. Small effect sizes such as this are relatively commonplace in environmental epidemiology, but small effects are often still of public health importance if entire populations are exposed.

### Confounding by other time-varying factors

In the analysis adjusted only for season and long-term trend, ozone appears to be positively associated with mortality. But could there be confounding by other factors? In general epidemiology, common confounders include age, sex, body mass index, smoking status, drinking and so on, but these 'standard confounders' do not apply to our data because at the population level, the distribution of such factors does not (or is unlikely to) change from day to day, and cannot be associated with fluctuations in environmental exposures such as pollution levels. So what are the potential confounders in this kind of study? Recall that the units of analysis are the time intervals represented by single rows of data (in our case, days),

and not individuals. Therefore potential confounders should be variables that can change from day to day, and that are plausibly related to daily fluctuations in our exposure of interest (ozone), as well as the outcome (mortality). In this example, a clear candidate is ambient temperature, because temperature varies from day to day, ozone levels are related to temperature (ozone tends to be higher on hotter days due to the involvement of sunshine in the generation of ozone), and it is well established that temperature is associated with mortality risk in the short term.[3]

Adding current temperature to the model (allowing for expected non-linearity[4]) does indeed move the estimated ozone-mortality association towards the null and the adjusted effect is no longer statistically significant (adjusted mortality risk ratio per 10 $\mu g/m^3$ is 1.003, 95% CI 0.999–1.006, $P = 0.11$). This suggests that the initially estimated positive association between current ozone level and mortality risk was largely explained by confounding by temperature.

Other potential confounders of the ozone-mortality association might include further meteorological parameters such as relative humidity (included in the dataset), other pollutants and variables capturing holiday and day of the week (pollution levels are likely to be related to population-level travel behaviours, which are likely to differ over holidays and at weekends, and it is highly plausible that certain health risks also differ at such times for reasons unrelated to pollution), but these are not explored further here.

## Allowing for delayed exposure effects

In the London data, our modelling thus far has related mortality on a particular day with ozone level on the same day. But it is possible that there is a delayed (or 'lagged') association between exposure and outcome. Yesterday's ozone level may be a more important predictor of today's mortality risk than today's ozone level. Estimating the association between yesterday's ozone level and today's mortality risk (i.e. the 1-day lagged association) is simply a question of shifting the ozone series forward in time (i.e. down one row) and re-fitting the previous model (Box 2). Figure 4a shows how the estimated ozone-mortality association (adjusted for temperature) changes as we increase the lag time from 0 to 7 days. There is evidence of an association between ozone and mortality when the lag time is between 1 and 5 days. However, these different lag effects are not adjusted for each other; so far each lag has been fitted in the model one at a time. To address this, all the lagged variables (the 0- to 7-day shifted series) can be simultaneously entered in the model. This is known as a 'distributed lag model' and, applied to the London data, results in the effect estimates displayed in Figure 4b. In comparison with the individual lag models, all the effect estimates for lag days 0 to 5 inclusive have now moved towards the null, suggesting (as expected) that the estimated individual lag effects were confounded by each other. There remains evidence of independent ozone-mortality associations

**Box 2 Delayed effects and 'lags'**

A simple analysis would relate the number of outcomes on a given day to the exposure levels on that day. But we often wish to explore whether there is any delayed association.
By creating time-shifted copies of the exposure variable and including them in the model, we can explore the association between outcome today and exposure on previous days.

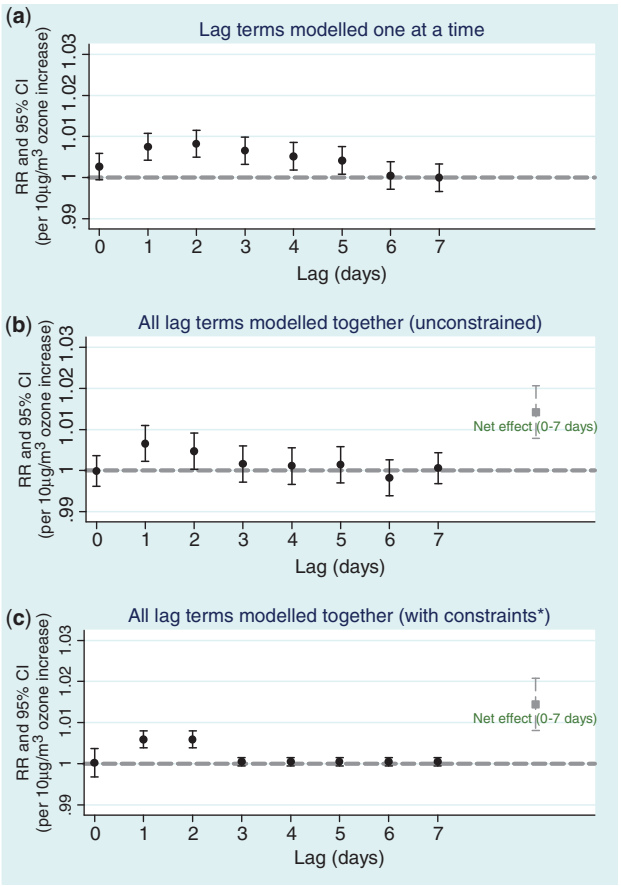


**Figure 4** Modelling lagged (delayed) associations between exposure and outcome. Asterisk indicates that the constraint applied was that the lagged associations for days 1 and 2 were the same, and for days 3–7 were the same

at lag days 1 and 2, suggesting that mortality risk on the current day is positively associated with ozone levels on the previous 2 days (or, equivalently, current ozone is positively associated with mortality on the following 2 days).

The disadvantage of the simple 'unconstrained' distributed lag model is that the lag terms are likely to be highly correlated, and collinearity in the model can result in imprecise estimates (wide confidence intervals). It is possible to overcome this by imposing some constraint on the effect estimates for the different lags (a 'constrained distributed lag model'). Figure 4c displays the results of imposing a simple constraint on the distributed lag model, namely that the effect estimates for days 1 and 2 are the same, and the effect estimates for days 3 to 7 inclusive are the same (a so-called 'lag-stratified' distributed lag model,[4] which might be justified by the broad patterns revealed in the unconstrained model of Figure 4b). Collinearity is now much reduced, fewer parameters need to be estimated, and associations at individual lags are estimated with greater precision, though a potential criticism of this approach is that the choice of constraints, if not pre-specified, could be argued to be too 'data-driven'. More complex constraints including smooth (polynomial or spline) functions of lag time can be applied.[6]

The cumulative effect of an exposure over several lag days can be calculated from a distributed lag model as the sum of the coefficients. These estimates are often similar in constrained and unconstrained models (as in the London data, shown on the right hand side in Figure 4b and c), and the similar confidence interval widths for the cumulative effect estimates from the constrained and unconstrained models have been observed before.[4]

Potentially confounding time-varying factors may also have lagged effects, which can be modelled in the same way.

### Short-term displacement, or 'harvesting'

Distributed lag models sometimes reveal an apparently odd feature: a raised risk ratio at short lags followed by an apparently protective effect at longer lags. For example, a study relating ambient temperature to hospital admissions for heart disease found that admissions increased on days with very high temperatures, but several days after the high temperature episode there were fewer admissions than expected.[12] This suggests that highly vulnerable people who were in any case within days of being admitted to hospital due to heart disease may have simply had their heart problem brought forward by a few days as a result of the high temperature episode. This is a phenomenon known as short-term displacement, or 'harvesting'. If harvesting appears to be present, the extent to which the short-term risk increase is 'cancelled out' by reductions in risk at longer lags can be ascertained by considering the cumulative association between exposure and outcome over the full lag period (estimated by summing the model coefficients, as described earlier).

## Model checking and sensitivity analysis

Having developed one or more models, before presentation it is essential to check residual plots and carry out sensitivity analyses in order to reveal any problems with the model assumptions, anomalies in the data, residual autocorrelation, or sensitivities of the main results to the decisions that have been made.[13] Useful diagnostic plots based on deviance residuals are described in the Supplementary Appendix (available as Supplementary data at *IJE* online), and other papers provide more detail.[6,14]

In addition, since the modelling process that we have outlined involves many decisions, we would recommend carrying out multiple sensitivity analyses to check that the main conclusions are robust to changes in these decisions. Sensitivity analyses might include changing amount of control for seasonality and long-term trends in the model (e.g. by changing the number of knots in the spline-based approach, or harmonics in the Fourier terms approach); specifying exposure and confounder variables in different ways (e.g. in the London analysis, we might try including relative humidity as a linear instead of categorical variable, or adjusting for maximum instead of mean daily temperature); changing the way lagged effects are included in the model; and changing other key context-specific decisions.

## Precision and power considerations

To our knowledge, there has been little formal development of power calculation methodology in this context, which may reflect the preponderance of studies using secondary or routinely collected data, where all available data are used. Nevertheless, a few broad points can be made. Factors determining the precision of a study include the length of the series (e.g. number of days) and the number of events (e.g. deaths) per day. Overdispersion (high variability in counts) can also reduce precision. For power, the size of effect that is plausible is also important. In the authors' experience, studies of pollution effects, one of the most common applications, need thousands of observation days with an average of tens of events per day, for credible precision and power.

**Box 3** Summary of key considerations and steps in a time series regression study

Explore data with simple plots and tabulations:

- Plot of exposure variable(s) against time
- Plot of outcome against time
- Correlation matrix for exposure and outcome variables
- Summary statistics for each variable
- Summary of missing data in each variable.

Methods to control for seasonality and long-term trends:

- Indicator variables for time strata (time-stratified model)
- Periodic functions of time (sine/cosine functions)
- Flexible spline functions of time.

Modelling the exposure-outcome association—immediate vs delayed effects:

- Individual lag models considering different lags one at a time
- Distributed lag models considering all lags in a single model (unconstrained, or constrained to reduce collinearity)
- Consider possible non-linear associations as in other regression contexts.

Model checking

- Diagnostic plots based on deviance residuals (see web appendix)
- Multiple sensitivity analyses changing key modelling decisions

## Further extensions

### *Non-linearity in the exposure-outcome association*

- Both the exposure of interest and other time-varying confounders might have non-linear associations with the outcome.
- This can be modelled as in other contexts: by using categorical variables, quadratic or higher order polynomials, flexible spline curves or piecewise linear 'threshold' models.[4]

### *Investigation of effect modifiers*

- Individual-level factors may still be effect modifiers (e.g. are the elderly more vulnerable to any detrimental effects of ozone?).
- This can be investigated provided it is possible to break down the overall outcome counts into stratum-specific counts based on the potential effect modifier.

### *Analysis of data from multiple locations*

- Separate analysis by location (e.g. specific cities) can increase power and provide information on heterogeneity and adaptation to environmental exposures.
- Patterns in location-specific effect estimates can be explored through techniques analogous to those used in meta-analysis,[15] or by modelling all the data in a single location-stratified model.[2]

## Summary

In this article we have outlined the key steps and complexities involved in carrying out a basic time series regression analysis (Box 3), and illustrated these in an example. Issues specific to time series regression are the presence of long-term and seasonal patterns, the possibility of delayed or non-linear associations between exposure and outcome, and the presence of autocorrelation. Aside from these, time series regression is no different from regression techniques used in other areas, and the broad steps involved (plotting and tabulating the data, controlling for confounding, presenting exposure effects appropriately and model checking) will be familiar to epidemiologists from all disciplines.

## Supplementary Data

Supplementary data are available at *IJE* online.

## Funding

No specific funding was received for this work. K.B. is supported by a National Institute for Health Research Postdoctoral Fellowship (grant number NIHR-PDF-2011-04-007). A.G. is supported by a Medical Research Council Methodology Research Fellowship (grant number G1002296). L.S. is supported by a Wellcome Trust Senior Research Fellowship in Clinical Science (grant number 082178).

**Conflict of interest:** None declared.

**KEY MESSAGES**

- Time series regression is often used in studies attempting to quantify short-term associations of environmental exposures, such as air pollution, pollen, dust or weather variables, with health outcomes.
- Time series data in these contexts may be analysed using Poisson regression models, with some extensions to deal with issues specific to time series regression, including the presence of long-term and seasonal patterns, the possibility of delayed or non-linear associations between exposure and outcome, and the presence of autocorrelation.
- Other steps involved in carrying out a time series study (plotting and tabulating the data, controlling for confounding, presenting exposure effects appropriately and model checking) will be familiar to epidemiologists from all disciplines.

## References


[1] Jimenez E, Linares C, Martinez D, Diaz J. Role of Saharan dust in the relationship between particulate matter and short-term daily mortality among the elderly in Madrid (Spain). *Sci Total Environ* 2010;**408:**5729–36.

[2] Bhaskaran K, Hajat S, Haines A, Herrett E, Wilkinson P, Smeeth L. Short term effects of temperature on risk of myocardial infarction in England and Wales: time series regression analysis of the Myocardial Ischaemia National Audit Project (MINAP) registry. *BMJ* 2010;**341:**c3823.

[3] Basu R. High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. *Environ Health* 2009;**8:**40.

[4] Armstrong B. Models for the relationship between ambient temperature and daily mortality. *Epidemiology* 2006; **17:**624–31.

[5] Bell ML, Samet JM, Dominici F. Time-series studies of particulate matter. *Annu Rev Public Health* 2004;**25:** 247–80.

[6] Schwartz J, Spix C, Touloumi G *et al*. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *J Epidemiol Community Health* 1996;**50(Suppl 1):**S3–11.

[7] Zeger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. *Annu Rev Public Health* 2006;**27:**57–79.

[8] Granger CWG. Investigating causal relations by econometric methods and cross-spectral methods. *Econometrica* 1969;**34:**424–38.

[9] Junger W, de Leon AP. Missing data imputation in time series of air pollution. *Epidemiology* 2009;**20:**S87.

[10] McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. London: Chapman & Hall/CRC, 1989.

[11] Dominici F, Samet JM, Zeger SL. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *J R Stat Soc a Sta* 2000;**163:**263–84.

[12] Schwartz J, Samet JM, Patz JA. Hospital admissions for heart disease: The effects of temperature and humidity. *Epidemiology* 2004;**15:**755–61.

[13] Katsouyanni K, Samet JM. *Air Pollution and Health: A European and North American Approach (APHENA)*. Boston, MA: Health Effects Institute, 2009.

[14] Katsouyanni K, Schwartz J, Spix C *et al*. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *J Epidemiol Community Health* 1996;**50(Suppl 1):**S12–18.

[15] Katsouyanni K, Touloumi G, Spix C *et al*. Short-term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. Air Pollution and Health: a European Approach. *BMJ* 1997;**314:**1658–63.