Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions

J Schwartz, C Spix, G Touloumi, L Bachárová, T Barumamdzadeh, A le Tertre, T Piekarksi, A Ponce de Leon, A Pönkä, G Rossi, M Saez, J P Schouten

Department of Environmental Health, Harvard School of Public Health, Boston, USA
J Schwartz

GSF-Forschungszentrum Umwelt und Gesundheit, Germany
C Spix

Department of Hygiene and Epidemiology, University of Athens Medical School, Greece
G Touloumi

National Center for Health Promotion, Bratislava, Slovakia
L Bachárová

Faculté de Médecine, University of Grenoble, France
T Barumamdzadeh

Laboratoire d'Hygiène de la ville de Paris, France
A le Tertre

National Institute of Hygiene, Poland
T Piekarksi

Department of Public Health Sciences, St Georges Hospital Medical School, UK
A Ponce de Leon

Helsinki City Centre of the Environment, Finland
A Pönkä

Institute of Clinical Physiology, National Research Council, Pisa, Italy
G Rossi

Institut Municipal d'Investigacio Meica de Barcelona, Spain
M Saez

Department of Epidemiology and Statistics, University of Groningen, The Netherlands
J P Schouten

Correspondence to:
Dr J Schwartz, Department of Environmental Health, Harvard School of Public Health and Channing Laboratory, Brigham and Women's Hospital Harvard Medical School, 65 Huntington Avenue, Boston MA 02115-6096, USA.

Abstract

***Study objective* – To review the issues and methodologies in epidemiologic time series studies of daily counts of mortality and hospital admissions and illustrate some of the methodologies.**

***Design* – This is a review paper with an example drawn from hospital admissions of the elderly in Cleveland, Ohio, USA.**

***Main results* – The central issue is control for seasonality. Both over and under control are possible, and the use of diagnostics, including plots, is necessary. Weather dependence is probably non-linear, and adequate methods are necessary to adjust for this. In Cleveland, the use of categorical variables for weather and sinusoidal terms for filtering season are illustrated. After control for season, weather, and day of the week effects, hospital admission of persons aged 65 and older in Cleveland for respiratory illness was associated with ozone (RR = 1·09, 95% CI 1·02, 1·16) and particulates (PM_{10} (RR = 1·12, 95% CI 1·01, 1·24), and marginally associated with sulphur dioxide (SO_2) (RR = 1·03, 95% CI = 0·99, 1·06). All of the relative risks are for a 100 μg/m³ increase in the pollutant.**

***Conclusions* – Several adequate methods exist to control for weather and seasonality while examining the associations between air pollution and daily counts of mortality and morbidity. In each case, care and judgement are required.**

(*J Epidemiol Comm Health* 1996;50(Suppl 1):S3–S11)

In 1952, an episode of high air pollution in London was associated with approximately 4000 excess deaths.[1] Other air pollution disasters in the Meuse Valley[2] and Donora, PA, USA[3] clearly established that high concentrations of air pollution could result in substantial increases in daily deaths. Analyses of data collected in London from 1958 to 1972[4,5] indicated that lower concentrations of air pollution were associated with smaller increases in daily deaths. Given the frequency with which those concentrations occur, the attributable risk from air pollution was not trivial. More recently, daily death counts and daily hospital admissions have been associated with air pollution at relatively low levels in studies from a number of groups on three continents.[6–40] Similar results have been seen in cohort studies.[41,42] Various methods have been used in these analyses. This paper reviews the analytical issues involved in these studies, discusses the problems that can occur and methods for addressing them, and provides a detailed example, using data on air pollution and hospital admissions of the elderly for respiratory illness in Cleveland, Ohio.

Methods

The study of the relationship between daily counts of health events and daily air pollution raises distributional issues and modelling issues. These are discussed in turn.

DISTRIBUTIONAL ISSUES

On any given day, only a small portion of the population dies or is admitted to hospital. The number that do is a count; that is, it can only take on values limited to the non-negative integers. This suggests that a Poisson process is the underlying mechanism being modelled. In a Poisson process, a homogeneous risk to the underlying population is assumed. Given that underlying risk, the expected number of deaths on any day is λ. Then the probability of y deaths occuring on a given day is given by

$$prob(y/\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad (1)$$

The Poisson process may not be stationary over time, that is, the underlying risk λ varies with time varying predictor variables $X_1 \ldots X_p$. In these analyses, the unit of observation is the day. Hence, while the underlying risk varies with some factors such as age or cigarette smoking, since the age distribution and smoking history of the population do not vary from day to day, these factors will not influence λ. This is a key advantage of the time series approach. One feature of the Poisson process is that even if all the covariates predictive of λ were known and measured without error, there would still be considerable unexplained variability in daily mortality. That is because the explanatory variables can at best predict λ. But even if λ is known with certainty, the Poisson process ensures stochastic variability around that expected count, as shown in the equation (1). In a classic stationary Poisson process, the variance of y is equal to λ. Many actual count processes are overdispersed, with a variance proportional to λ. Hence R^2 is unlikely to be

high in a Poisson process, and is inappropriate as a measure of goodness of fit.

Poisson regression analysis is now available on standard statistical packages, and most of the studies of mortality and hospital admission counts published in recent years have used this approach. The canonical Poisson regression is a relative risk model. It seems reasonable to suppose that if the population being studied doubled, while keeping its characteristics and all other risk factors constant, the number of increased cases due to air pollution and/or weather (if any) would also double. Since the baseline number of cases would also double under that scenario, this implies a relative risk model. In that model we assume:

$$Log(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_p X_p \qquad (2)$$

where Y is the count of deaths or hospital admissions on a given day, $E(Y)$ is the expected value of Y on that day (corresponding to λ in equation 1), $X_1..X_p$ are the predictors of daily counts, and $\beta_1..\beta_p$ are the regression coefficients for those predictors.

Recently, some papers have explored the sensitivity of results to whether the mortality counts were modelled as Poisson or Gaussian processes.[22 37] They have not shown much difference in either effect estimates or t statistics. This result is understandable in these analyses. The relative risk for most of the factors examined in these models (season, weather, or air pollution) is not large. The principal difference between a least squares regression of the logarithm of death counts and a true Poisson regression is the heteroscedasticity in the variance built into the Poisson model. Across a narrow range of expected values, this will not be large, and hence one would not expect much difference in these daily time series models.

MODELLING ISSUES

Means model – season and trend

In all of epidemiology, a basic issue in modelling is to control properly for potential confounding. Time series studies have some unique features in this regard. Many variables show systematic variation in time. For example, the number of AIDS cases in the world and the value of most stock market indices increased over the 1980s. This does not mean that we believe they are causally associated. Since any two variables that show a long term trend must be correlated, searches for correlations that are more likely to be causal must exclude these trends. These trends may not be linear. For example, the world population is increasing exponentially with time. The decline in cardiovascular death rates may be levelling off. Nevertheless, for short intervals, this may be adequately approximated with a linear time trend variable. As the number of years studied grows, the need for a non-linear trend model probably increases.

A second common attribute of many variables that evolve over time is seasonality. Many health, weather, and pollution variables show systematic variation over the course of the year. These variations would be present even if these factors were not causally related, and will induce correlations among them. And many seasonal variations in health outcomes may be due to more general factors, such as people spending more time indoors, rather than weather per se. Again, to focus on possibly causal associations with acute effects, it is necessary to remove these patterns. They are often described as long wavelength patterns, because the interval (in days) describing the pattern is long. This should not be interpreted as indicating that the patterns are sinusoidal. While the annual pattern is roughly periodic, it need not be sinusoidal or even symmetric – for example, there can be long winters and short summers, step inclines and flat declines, etc.

A final systematic component that may bias time series regressions involves calendar specific days. Day of week or holiday effects fall in this category. These patterns are not necessarily present in all data, but they occur often enough that they should be checked.

MODELLING APPROACHES FOR SEASON AND TREND

Several methods exist for dealing with these issues. These are discussed in turn.

Smoothing

For Gaussian data, one possibility is to filter these patterns out of the data before analysing them. For example, in analyses of daily counts of deaths in London in the 1960s[4] a 15 day moving average was computed. That is, the mean of daily deaths on the current day, the previous week, and the next week was computed. This moving average represents a good estimate of the expected number of deaths on that day, based on the long wavelength patterns in the data. Here long wavelength is taken to mean any patterns of 15 days or longer. The deviation from that moving average represents the shorter term fluctuations about that longer term pattern. They represent the excursions of a few days' duration, about the general pattern. The authors of several papers on this data set used this approach to focus the investigation on the correlation between short term changes in daily deaths and short term changes in air pollution and weather.[4 5]

The 15 day moving average filter has an unattractive feature. In estimating the number of deaths to expect today, it gives equal weight to all deaths in the preceding and next week, and zero weight to deaths more distant in time. Intuitively, the weight given to the information on other days should decrease as the days become further separated in time. Moreover, the sudden decrease in weight from one to zero between the seventh and eighth prior day creates distortions in the filtered data. These distortions occur in the short wavelength patterns that we seek to keep. They can be effectively reduced, and better predictions obtained by computing weighted moving averages. In

the statististics literature, this is referred to as kernal smoothing.[43] The question of whether 15 days was the appropriate number of days to choose will be addressed below.

The mean number of deaths in London winters in the 1960s was almost 300, which allows a Gaussian approximation to the count data. Most studies today involve fewer counts, where a Poisson analysis is called for. While filtered Gaussian data is still Gaussian, filtered Poisson data is not Poisson. Hence prefiltering is not possible. The goal is still the same, however, and the solution is to put the filter in the regression. This is analogous to controlling for confounders in the regression model, and can be done in a number of ways. The most direct approach is merely the generalisation of the above process. The moving average can be incorporated into Poisson regressions in a straightforward way. Since Poisson regressions are usually relative risk models, Burnett *et al*[25] estimated a Poisson regression where the relative risk due to weather and air pollution was relative to the running moving average of daily hospital admissions. This can be accomplished by modifying equation (2):

$$Log(E(Y)) = \beta_0 + \beta_1 X_1 + .. + \beta_p X_p + log(WMA(Y)) \quad (3)$$

where *WMA(Y)* denotes the weighted moving average of Y.

An alternative approach that generalises on the use of moving averages is the generalised additive model.[44] The generalised additive model allows a Poisson regression to be fitted which controls for a smooth function of time using moving averages or other types of smoothing functions. This is also quite a useful approach to control for time trend and season. In addition, it allows generalisation to multivariate smoothing. The model above uses the average to control for long wavelength patterns in the data. The other covariates, such as temperature and humidity, are treated as linear factors predicting the daily counts. However, the dependence of mortality on temperature is usually non-linear, with raised death rates on both very hot and very cold days. In some cases, a V or U shaped dependence with two linear or quadratic terms has sufficed.[36] In other cases, a curvilinear dependence was found.[22 32] Of course, if one believes that the true dependence on temperature is quadratic, one can use the square of temperature as one of the predictor variables. However, that requires knowledge not merely that the dependence may be non-linear, but that it is parabolic. While the dependence of mortality on temperature may be U shaped, the parabola is not the only U shaped function. In the absence of any theory to guide us, a flexible approach to covariate control is appropriate. A more flexible approach available in the generalised additive model is to treat temperature, and the other continuous variables, as smooth functions in the same manner as the time covariate.

Hence in the generalised additive model, we assume:

$$Log(E(Y)) = \sum_{1}^{p} S_i(X_i) \quad (4)$$

Where the X_i are the predictor variables, which include time, and the S_i are the smooth functions of those variables. The data determine the shape of the smooth functions. There are many different smoothing algorithms in addition to the weighted moving average, although they are mostly based on variants of that principal.[44] The principle issue in the use of these smooth non-parametric functions is the choice of the fraction of the data (called the smoothing parameter) that will be included in the running smooth. Intuitively, a small window is necessary to fit the time trend in the data adequately, since we expect a somewhat periodic pattern within each of the years of the study. Conversely, for weather variables, we expect that while there may be non-linearities, these will be more global, and a large number of "wiggles" on top of that pattern are unlikely to be causal. More formal approaches exist to choosing the fraction of the data to include. These include Akaike's information criteria and cross validation.[45–47] The generalised additive model approach has been used successfully in analyses of counts of both daily deaths and daily hospital admissions.[21 22 27 30 32]

Semiparametric approaches

Smoothing is not the only approach. An alternative, flexible approach is to use regression spline functions. Here, we divide each variable (time, temperature, etc) up into intervals. A cubic polynomial is fitted to each interval, and they are required to join smoothly at the boundaries of the intervals. For example, a cubic polynomial can be fitted to each three month interval of time, or to each 10°C interval of temperature. Since separate polynomials are fitted in each interval, the approach can capture local patterns. The central issue here, parallel to the choice of window above, is the choice of number of intervals. This choice can be based on a priori or theoretical considerations, or such data-driven approaches as cross validation and the information criteria sited above.

Parametric approaches

A completely different approach involves the use of sinusoidal terms to fit the long wavelength pattern in the data. Clearly, the annual cycle is relatively periodic. While the pattern of mortality over time is unlikely to be a pure sine wave, the sum of sine waves of increasing frequency can fit more complex functions. Hence trigonometric filtering has also been used in studies of air pollution and daily morbidity or mortality.[22 48] Here, the question equivalent to the choice of window is the choice of for which frequencies trigonometric terms should be included in the model. Again, a priori considerations can lead one to choose to filter out all patterns above a given wavelength, or statistical considerations can be used to choose the frequencies to fit. Not all sinusoidal terms

begin at zero on 1 January. Hence if a sine wave of a given frequency is fitted in a regression, a cosine wave of the same frequency must be fitted to account adequately for the phase of the pattern. One concern with such a model is that it assumes that the seasonal peak is the same height and occurs at the same time each year. However, there are patterns in the intensity of, for example, influenza epidemics, with two year cycles both observed in practice and justified by mathematical modelling. Improvements in heating and air conditioning may result in declining peak to trough ratios over time. The maximal week of the epidemic varies, with less pattern, over time. To some extent these issues can be dealt with by fitting two year cycles (or longer) to the data as well. However, in some instances, different sinusoidal terms will have to be fitted for different periods. For longer time series, this requires increasing attention.

A classic epidemiological approach to the model specification problem is to divide continuous data into categories. This is often considered to be model free, however it in fact fits a histogram to the data. Hence, dummy variables for season,[15] month of year,[39] or month of study[26] have been used to control for long wavelength patterns. The use of seasonal dummy variables risks underspecifying and the use of month of study dummy variables risks over specifying the model.

How much filtering

The choice of 15 days in the original London studies seems to be an artifact of wanting to have a week on either side of the day whose expectation was being computed. What is the best choice? The best choice may in fact vary from study to study depending on other local characteristics such as meteorology, etc, and is likely to be larger than 15 days. The objective criteria seem to have a natural appeal. However, the problems with these approaches (for example Akaike's information criteria) is that they are objective answers to the wrong question. The question we want to answer is, "Seasonal and other long wavelength patterns are too common in most time varying risk factors, including those that were not measured in the study, to allow them to contribute to the correlation between air pollution and daily mortality or morbidity. At what wavelength do we believe this is no longer true?" The question we can answer objectively is, "The daily variation in death counts contains both pattern and noise. At what point in fitting a predictive model to the data, using time as the predictor, do we loose confidence that we are fitting pattern and not noise?" If in noisy data the objective criteria suggest leaving in fluctuations with wavelengths of five months, should we feel comfortable with that decision or worry that there would be less risk of confounding if more filtering were done? If in a data set with high counts, pattern is detected down to wavelengths of 10 days, should we throw away so much information or allow air pollution the opportunity to explain the observed short wavelength pattern? These are questions which require epidemiological judgment. Our goal is to filter out not all pattern in the data, but only that pattern where we believe the risk of confounding by an omitted variable is high.

What is the risk of overfiltering? One risk is loss of power. By removing shorter and shorter wavelength components from the data we are effectively reducing the sample size, since less and less variation in the data is kept. In principle this can be resolved by obtaining longer time series. There is also a potential for bias. If cumulative exposure over several days is necessary to produce mortality effects, too much filtering risks throwing away precisely those patterns of exposure whose effects we wish to examine. Several methods exist to help with the epidemiologic judgment that is required.

Diagnostics plots

No matter which method is used to deal with long wavelength patterns, diagnosic plots are critical to evaluating the success of the approach. A plot of the residuals versus time, particularly if a smooth curve is fitted to the residual plot, can often identify long wavelength patterns that remain. Spectral density or periodogram plots show the amount of the variation in the data that occurs in a given frequency range. These are also helpful. However, there are certain patterns that are hard to detect when resolved into trigonometric spectra. A basic issue with frequency domain plots is that deviations from perfect sinusoidal shapes of long wavelength components show up as short wavelength components. It may be easier to determine if these are adequately dealt with (without over filtering) by examining time plots. For example, aperiodic patterns, such as a three month excursion that occurs only in one year, can more readily be detected in a residual time series plot. Variations in the magnitude of the seasonal pattern may also be easier to see in the time plot. Other approaches, such as the use of wavelets[49] as basis functions instead of trigonometric functions, may prove useful here, but have not yet been explored.

In addition to looking at residual plots, plots of the predicted outcome over time based on the seasonal model are quite useful. If one compares a series of seasonal models that represent increasing filtering (of whatever method), one can readily see when the additional filtering is beginning to predict shorter term patterns that one wishes to leave for the explanatory variables in the model. By applying this approach to different filtering schemes, one can also determine, in a given data set, whether one approach has advantages over another. It is possible that one method (for example smoothing, splines, trigonometric filters, monthly dummies) can do a better job on fitting the shape of the seasonal pattern before beginning to pick up short term patterns.

One alternative approach that has been suggested[50] is to rely on a Durbin-Watson statistic to determine whether seasonality has been controlled for. However, a Durbin-Watson statistic is a measure of first order autocorrelation for

Gaussian data, and a value near two neither indicates that autocorrelation or seasonality is satisfactorily accounted for. For example, Pope has reported that before control for season, daily cardiovascular mortality in Zurich had a Durbin-Watson statistic of 1·94 (Pope C A, personal communication).

Modelling approaches for weather

After season and trend, weather terms are the most important covariates to enter the model. Since season and other long wavelength patterns are being removed by other methods, these weather variables will be there to carry information about the effects of short term variations in weather on mortality or morbidity. Two types of weather variable can be used. One is to use the primary variables that are likely to be plausibly associated with mortality. These are widely agreed to be temperature and humidity. Whether minimum, maximum, or mean temperature is the best predictor, and whether there is any additional information in one measure after control for the other is not resolved. An alternative approach that has occasionally been tried is to use factor analysis or judgmental classification into weather patterns to come up with categoric variables for different types of weather patterns to put into the regression model. Little systematic evaluation has been done to evaluate the relative performance of the two approaches. The weather pattern category approach defines natural patterns, but whether those natural patterns are the best predictors of mortality or whether certain continuous characteristics that cut across them may better predict mortality or morbidity is not certain.

As mentioned previously, the dependence of morbidity and mortality on weather is probably non-linear, or at least piecewise linear. A number of methods exist for modelling non-linear dependencies of daily counts of mortality and morbidity on weather. These techniques include smoothing, multiple dummy variables, splines, or non-linear polynomials, and combinations of them. Whichever is chosen, some attention must be paid to the potential non-linearities, and again, diagnostic plots are critical in helping to determine if the model was well specified. Interactions between temperature and humidity also need to be examined.

Day of week, holiday, and influenza epidemics

Whatever the benefits of reducing pollution concentrations, it is clear that Monday is bad for people's health.[51 52] Weekends, in contrast, are often healthful. These variables should be considered in time series studies, although a number of published reports indicate that their inclusion does not have a major impact on the coefficients of air pollution.[16] Holidays may have similar effects. Influenza epidemics were a major concern as a potential confounder in studies done in the 1960s. However, an adequate seasonal model reduces the level of concern. Nevertheless, if influenza data are available, they should be used. Approaches have included dummy variables for influenza epidemic weeks, as well as some modelling approaches for the shape of the epidemic.

Lag structure (delayed effects)

The effects of all the explanatory variables may be immediate, or may occur with some lag. In addition, there may be a disturbed lag structure. That is, cold weather could effect mortality both on the concurrent days and on the next day. In that case, the effect of cold weather on any day's mortality would be the sum of the effect on that day and the previous day. It is probable that the impact of today's temperature and yesterday's temperature on today's death count will differ in magnitude. One approach to this is to examine models with multiple lags of the explanatory variables simultaneously included. Because those variables are serially correlated, this will often produce unstable estimates. Instead, a constraint is often put on the system. The simplest constraint is the moving average. If the exposure variable is defined to be the mean of a two or three day moving average of the explanatory variable, the contribution of multiple days can be estimated, subject to the constraint that the effect of each day is identical. This is helpful in at least identifying the existence of multiple day effects, but the constraint is not very realistic. A more realistic constraint would allow the influence of exposure to decline with time. One such approach is the geometrically distributed lag model. In that model, we would assume:

$$\log(E(Y_i)) = covariates + \beta(X_i + \alpha X_{i-1} + \alpha^2 X_{i-2} + \alpha^3 X_{i-3} + \ldots)$$

Such a model can be fitted iteratively. Polynomial distributed lag models[53] fit a polynomial function to the pattern of the lagged effects. For example, a second order polynomial centred at lag 1 would fit a model where the largest impact of temperature was from the previous day's temperature, and the impact of the temperature two day's before, and on the concurrent day was the same, and reduced from the lag 1 day by a factor defined by the parabolic function that was fitted. Programs for estimating these models are common in econometrics, but they have not been widely explored in epidemiology. Standard statistical packages such as *SAS* do include these programs, although not for Poisson models. As with geometric distributed lag models, they can be fitted to Poisson regression using iterative approaches, for example using *proc NLIN in SAS*. Greater attention needs to be paid to these approaches in the future.

One issue is how far back to go in exploring the lags. Exploring too many lag structures risks identifying non-causal relationships that have occured by chance. If the seasonal model is relied upon to deal with long wavelength patterns, then restrictions to a week or less seem most reasonable for the exploratory variables.

Interactions

One question that is sometimes explored is whether the association between mortality and/

Table 1 Centile points of the health and environmental data for Cleveland, Ohio during the years 1988–90

Variable	*10%*	*25%*	*50%*	*75%*	*90%*	*Mean*
Admissions	13	16	21	26	32	22
Temperature (°F)	27	37	51	67	74	51
Dew point (°F)	17	27	42	56	64	41
SO_2 (ppb)	13	20	31	45	61	35
O_3 ppb*	30	40	53	68	88	56
PM_{10} (μg/m³)	19	26	39	56	72	43

* Hourly maximum, other pollutants are 24 hour average

or morbidity and air pollution varies with season or other factors. This raises several issues. The major one is epidemiological. To avoid increased risks of drawing false conclusions, standard epidemiological practice requires that hypotheses be generated first, and then tested. The first question that must be addressed is what is the hypothesis that is being tested? In some cases, this is clear. Ozone concentrations are very low during cold weather, and people spend less time outdoors during cold weather, reducing exposure. While ozone does penetrate indoors, it also reacts with fabric and other surfaces indoors. During cold weather, when windows and doors are shut, the indoor concentrations are a very small fraction of the (low) outdoor concentrations. Hence, overall, we expect that the ratio of population average personal exposure to outdoor monitored exposure will be lower in the winter. Therefore, one hypothesises a different, and smaller, ozone slope in the colder weather. Having specified this rationale, a test of the hypothesis is justified.

The next issues are statistical. Given the small overall effect size for air pollution, there is limited power to detect interactions. Care must be taken to maximize that power. Dividing the sample into four calendar quarters or even shorter intervals is probably a recipe for instability, rather than inference. Given the basic hypothesis outlined above, a division into two six month periods, when weather is warmer and colder, seems most reasonable. Dividing into more categories than justified by prior specific hypothesis risks the use of a postiori hypotheses which may just be attempts to rationalise patterns that are merely instability. The published reports on stepwise regression are relevant here. Next, we must determine the specific hypothesis being tested. If the hypothesis is whether the slope is different from zero in each season, then separate regressions for each season are in order. However, if there is a significant association overall, then it is hard to see why exposure would produce no effect in some seasons. Rather, it seems more reasonable to hypothesise that the slope may be different, because the relationship between outdoor and personal exposure is different. In that case, an interaction term is more appropriate. It tests whether the slope in one season is different than in the other.

Multipollutant models

One occasionally sees studies that have fitted regression models using four or even more collinear pollutants in the same regression, and sought to draw inferences about which variables were causal from that model. Other studies have used stepwise procedures, again with many pollutants in the candidate list. Some times, more than one proxy for particulate air pollution has been included in the same regression model. Given the non-trivial correlation of the pollutant variables, and the relatively low explanatory power of air pollution for mortality or hospital admissions, such procedures risk letting the noise in the data choose the pollutant. The APHEA project has taken a more cautious approach. Single pollutant models are fitted initially. If more than one pollutant seems to be associated with the outcome, then attempts are made to separate the pollutants by examining associations with one pollutant stratified by the level of the other pollutant, etc. Greater reliance in APHEA is being placed on examination of associations across study centres as a way to help separate the effects of individual pollutants.

Covariance model

Measurements connected in time and/or space, such as repeated measurements of the same population on consecutive days or measurements of persons from nearby geographical areas, are likely to be correlated and not independent. In the case where two observations closer together in time are more alike than two randomly chosen observations, this is referred to as serial correlation. If the serial correlation in the outcome is due to omitted covariates (for example, epidemics) or imperfectly controlled for covariates (for example, weather), then that omission or imperfect control will leave serial correlation in the residuals of the model. Serial correlation will not bias the regression coefficients, but will bias the estimated standard errors. As when modelling the expected value, the specification of a model for the covariance with several parameters can allow a parsimonious description of this correlation. In the Gaussian case, this model assumes, as before, that:

$$E(Y_i) = X_i\beta$$

but

$$COV(Y_iY_j) = V_{ij}$$

If the structure of V (defined above) is known, then a generalised least squares approach yields the maximum likelihood estimate, that is:

$$\beta = (X'V^{-1}X)^{-1}X'V^{-1}$$

and

$$COV(\beta) = \sigma_y^2(X'V^{-1}X)^{-1}$$

In general, the correlation structure of V is not known a priori and must be estimated from the data. In order to avoid using up degrees of freedom, it is usually necessary to remodel the correlation in the covariance as a function of one or more parameters. For serially correlated data, autoregressive and moving average models represent efficient schemes. In biomedical applications, autoregressive processes

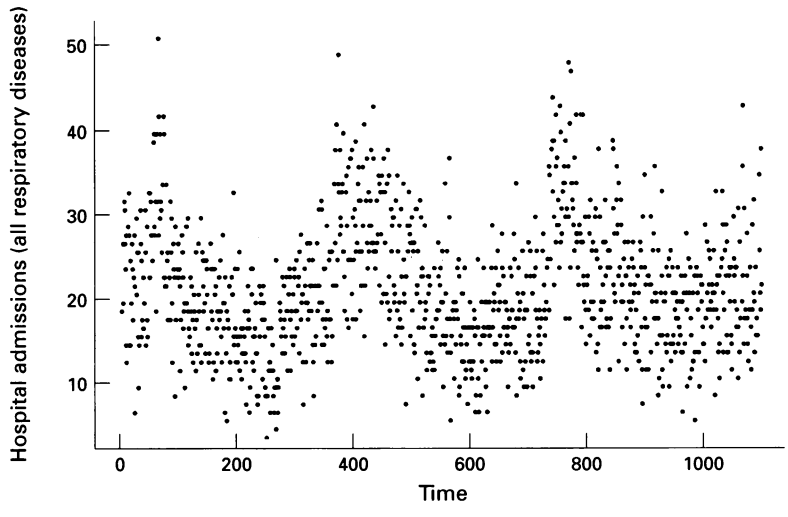

Figure 1 A plot of daily counts of hospital admissions of persons aged 65 and older in Cleveland, Ohio for respiratory illness (ICD 9 460-519).

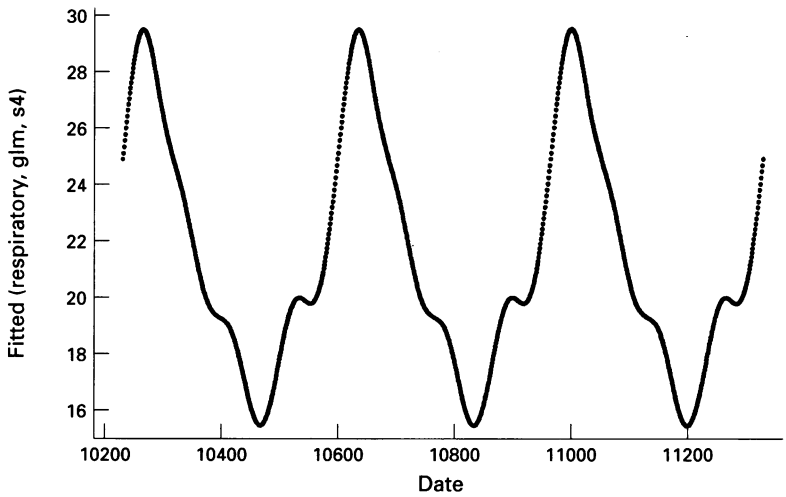

Figure 2 A plot of the predicted number of daily hospital admissions of persons aged 65 and older in Cleveland, Ohio for respiratory illness based on a model including sinusoidal terms with periods 1 year, six months, four months, and three months.

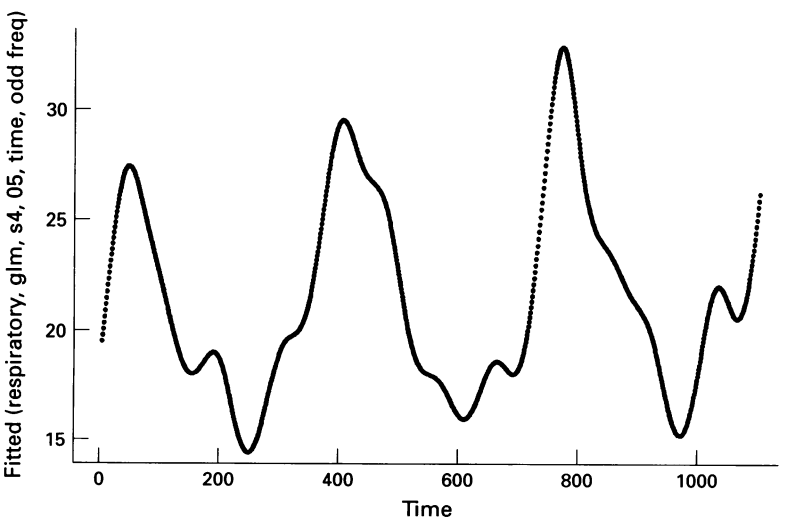

Figure 3 A plot of the predicted number of daily hospital admissions of persons aged 65 and older in Cleveland, Ohio for respiratory illness including sinusoidal terms with periods of two years, one year, six months, 4·53 months, four months, and three months plus a linear quadratic time trend.

are more usually found. This estimation process generally entails an iterative process.

A similar approach exists for Poisson regression models. As in a classic Poisson regression, the model assumes

$$\log[E(Y_i)] = X_i\beta$$

where X_i is the matrix of covariates on day i, Y_i the mortality counts on day i, and E denotes expected value. The covariance matrix is assumed to be of the form:

$$\alpha A^{1/2} R A^{1/2}$$

where $A_{ij} = E(Y_i)\ \delta_{ij}$, the classic Poisson covariance, α is the overdispersion parameter, $\delta_{ij} = 1$ when i-j and 0 otherwise, and R is an autoregressive matrix. The order of R is estimated empirically from the data. α is estimated from the residual χ^2 using the method of McCullagh and Nelder.[54]

Liang and Zeger[55] have shown that even if R is poorly estimated, it is possible to compute robust variance estimators, using the sandwich estimator, in order to obtain assymptotically unbiased estimates of the standard errors. However, this relies on having multiple time series. While this is true in panel studies, in a study of daily mortality there is only one time series and robust variance estimates cannot be computed. If the number of years examined is large, each year can be treated as a replicate and robust estimates than computed, but this is often not the case. Hence, as in classic time series, it is necessary to pay attention to modelling R adequately. That is, the order of the autoregressive process must be determined. The usual method for determining this is to examine partial autocorrelation functions of the residuals. Again, it should be noted that the Durbin-Watson statistic only measures first order autocorrelation and will be inadequate for this task. Once the order is specified, R can be specified and an autoregressive Poisson regression estimated.

In general, after control for season and trend, the magnitude of the serial correlation in mortality and hospital admissions data is low (for example, in the order of 0·50–0·20) and the estimates will be little changed by incorporating serial correlation.

Example and illustration

To illustrate these issues, we have used data on hospital admissions for all respiratory disease in persons aged 65 and older in Cayahoga County, which includes the city of Cleveland, Ohio. The data are for the years 1988–90. Table 1 shows some summary statistics on the health and environmental data from Cleveland. Daily monitoring was available for sulphur dioxide (SO_2). For ozone (O_3), monitoring was only done during the warm season. For particulates (PM_{10}), daily monitoring began in the autumn of 1988, with occasional monitoring earlier in the year.

Figure 1 shows a plot of hospital admissions for respiratory disease during the study period.

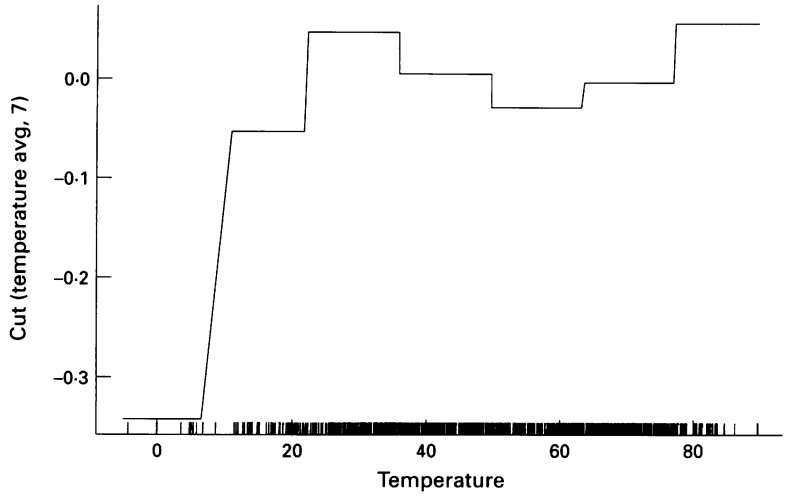

Figure 4 The mean residual number of hospital admissions of persons aged 65 and older in Cleveland, Ohio for respiratory illness from the seasonal model (fig 3) in relation to seven categories of mean daily temperature.

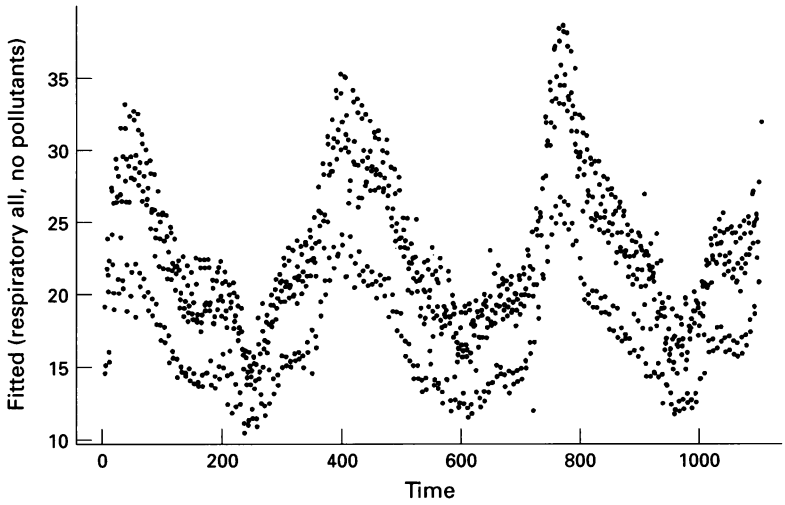

Figure 5 The predicted number of daily hospital admissions of persons aged 65 and older in Cleveland, Ohio for respiratory illness. The model included sinusoidal terms with periods two years, one year, six months, 4·53 months, four months, and three months, plus a linear and quadratic time trend. In addition, dummy variables for seven categories each of temperature and humidity and daily dummy variables were included.

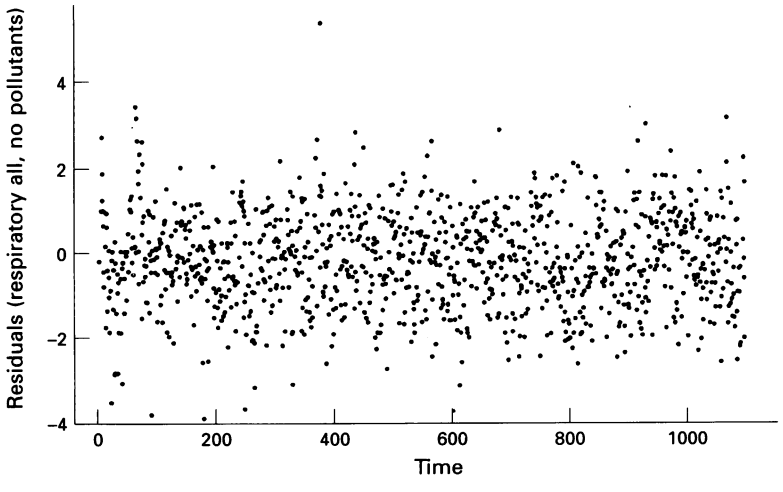

Figure 6 The residuals of the final model without air pollution plotted against day of study.

Table 2 Regression results from Poisson regression models using trignometric filtering. Results are for a 100 μg/m³ increase in each patient

Pollutant	*RR*	*(95% CI)*
O_3*	1·09	(1·02, 1·16)
PM_{10}†	1·12	(1·01, 1·24)
SO_2‡	1·03	(0·99, 1·06)

* Average of the hourly maximum for the two days prior to admission.
† Average of two days prior to admission.
‡ Average of day of admission and day before admission.

A clear seasonal pattern is evident. This pattern can be roughly predicted using a model including sine and cosine terms for periods of 12 months, 6 months, 4 months, and 3 months (fig 2). While that model captures the basic seasonal pattern, it has some disturbing features. The magnitude of the winter peaks is the same each year, which does not appear to be true in figure 1. And the shoulders on the seasonal patterns are the same each year. Since these probably result from respiratory epidemics, this seems an unlikely model. When the model is expanded to include sinusoidal terms with a two year period and linear and quadratic time trend terms, it shows annual peaks which differ in magnitude, somewhat less simple shapes to the peaks, and slightly different shaped shoulders. Examination of the spectral density function of the residuals of this model identified a cycle with period of 2·65 cycles per year. Incorporating sine and cosine terms for this frequency produced a predicted curve (fig 3) that showed different shaped peaks for the different years, including major differences in the shoulders. This picture seems more realistic than the previous ones, and no other spectral frequencies were significantly elevated in magnitude.

Temperature and day of the week terms were then added to that model. Figure 4 shows the pattern of admissions by seven categories of daily temperature. Over most of the temperature range, a classic U shaped dose response was seen, with increases on cold and hot days. However, as the mean daily temperature fell below 20°F, hospital admissions actually decreased in Cleveland. This may reflect the nature of the population being studied. In very cold weather, they simply do not go outside and and are not exposed to the weather. It certainly illustrates the advantage of taking a flexible approach to modelling the dependence on temperature. The pattern of admissions by day of the week was as expected – low on the weekends. Figure 5 shows the predicted number of admissions each day from the final model without air pollution, and figure 6 shows the residuals from that model. No seasonal pattern is evident in that plot, although one possible outlier is apparent. Once this model was completed, individual pollutants were tested. The results of those regressions are shown in table 2. Significant associations are seen for PM_{10} and O_3, with somewhat weaker evidence for SO_2. Excluding the possible outlier had a trivial impact on all of the associations.

Supported in part by a John D and Catherine T MacArthur Fellowship, and National Institute of Environmental Health Sciences grant ES-00002.

1 Her Majesty's Public Health Service. *Mortality and morbidity during the London fog of December 1952.* London: HMSO, 1954. Report No 95 on Public Health and Medical Subjects.
2 Firket M. *Fog along the Meuse Valley. Transactions of the Faraday Society* 1936;**32**:1192–7.
3 Shrenk HH, Heimann H, Clayton GD *et al. Air pollution in Donora PA: epidemiology of the unusual smog episode of October 1948. Preliminary report.* Washington, DC: US Public Health Service, 1949. Public Health Bulletin no 306.
4 Mazumdar S, Schimmel H, Higgins ITT. Relation of daily mortality to air pollution: an analysis of 14 London winters, 1958/59–1971/72. *Arch Environ Health* 1982;**37**:213–20.
5 Schwartz J, Marcus A. Mortality and air pollution in London: a time series analysis. *Am J Epidemiol* 1990;**131**: 185–94.
6 Hatzakis A, Katsouyanni K, Kalandidi A, *et al.* Short term effects of air pollution on mortality in Athens. *Int J Epidemiol* 1986;**15**:73–81.
7 Bates DV, Szito R. Hospital admissions and air pollutants in Southern Ontario: the acid summer haze effect. *Environ Res* 1987;**43**:317–31.
8 Pope CA III. Respiratory disease associated with community air pollution and a steel mill, Utah valley. *Am J Public Health* 1989;**79**:623–28.
9 Fairley D. The relationship of daily mortality to suspended particulates in Santa Clara County, 1980–1986. *Environ Health Perspect* 1990;**89**:159–68.
10 Katsouyanni K, Hatzakis A, Kalandidi A, *et al.* Short term effects of atmospheric pollution on mortality in Athens. *Archives of Hellenic Medicine* 1990;7:126–32.
11 Schwartz J. Particulate Air pollution and daily mortality in Detroit. *Environ Res* 1991;**56**:204–13.
12 Sunyer J, Anto JM, Murillo C, Saez M. Effects of urban air pollution on emergency room admissions for chronic obstructive pulmonary disease. *Am J Epidemiol* 1991;**134**: 277–86.
13 Kinney PI, Ozkaynak H. Associations of daily mortality and air pollution in Los Angeles County. *Environ Res* 1991; **54**:99–120.
14 Pope CA III. Respiratory hospital admissions associated with PM_{10} pollution in Utah, Salt Lake, and Cache valleys. *Arch Environ Health* 1991;**46**:9–97.
15 Schwartz J, Dockery DW. Particulate air pollution and daily mortality in Steubenville, Ohio. *Am J Epidemiol* 1992;**135**: 12–20.
16 Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. *Am Rev Respir Dis* 1992;**145**:600–4.
17 Dockery DW, Schwartz J, Spengler JD. Air pollution and daily mortality: associations with particulates and acid aerosols. *Environ Res* 1992;**59**:362–73.
18 Pope CA, Schwartz J, Ransom M. Daily mortality and PM_{10} pollution in Utah Valley. *Arch Environ Health* 1992;**42**: 211–17.
19 Thurston GD, Ito K, Kinney PL, Lippman M. A multiyear study of air pollution and respiratory hospital admissions in three New York State metropolitan areas results for 1988 and 1989 summers. *J Expos Analy Environ Epidemiol* 1992;**2**:429–50.
20 Schwartz J, Koenig J, Slater D, Larson T. particulate air pollution and hospital emergency visits for asthma in Seattle. *Am Rev Respir Dis* 1993;**147**:826–31.
21 Schwartz J. Air pollution and hospital admissions for the elderly in Birmingham, Al. *Am J Epidemiol* 1994;**139**: 589–90.
22 Schwartz J. Air pollution and daily mortality in Birmingham Al. *Am J Epidemiol* 1993;**137**:1135–47.
23 Spix C, Heinrich J, Dockery D, *et al.* Air pollution and daily mortality in Erfurt, East Germany from 1980–1989. *Environ Health Perspect* 1993;**101**:518–26.
24 Sunyer J, Saez M, Murillo C, Castelsaque J, Martinez F, Anto JM. Air pollution and emergency room admissions for chronic obstructive pulmonary disease: A 5-year study. *Am J Epidemiol* 1993;**137**:701–5.
25 Burnett RT, Dales RE, Raizenne ME, *et al.* Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to Ontario hospitals. *Environ Res* 1994;**65**:172–94.
26 Schwartz J. Air pollution and hospital admissions for the elderly in Detroit, Michigan. *Am J Resp Crit Care Med* 1994;**150**:648–55.
27 Schwartz J. The use of generalized additive models in epidemiology. In: *International Biometric Society XVII International Conference; Proceedings.* Invited Papers; 55–80. Hamilton, Ontario, 1994.
28 Schwartz J. Pm10, ozone, and hospital admissions for the elderly in Mineapolis – St Paul, Minnesota. *Arch Environ Health* 1994;**49**:366–74.
29 Walters S, Griffiths RK, Ayres JG. Temporal association between hospital admissions for asthma in Birmingham and ambient levels of sulphur dioxide and smoke. *Thorax* 1994;**49**:133–40.
30 Schwartz J. Non-parametric smoothing in the analysis of air pollution and respiratory illness. *Canadian Journal of Statistics,* 1994;**4**:471–87.
31 Schwartz J. Particulate air pollution and daily mortality in Cincinnati, Ohio. *Environ Health Perspect* 1994;**102**:186–9.
32 Schwartz J. Air pollution and daily mortality: a review and meta-analysis. *Environ Res* 1994;**64**:36–52.
33 Schwartz J. What are people dying of on high air pollution days? *Environ Res* 1994;**64**:26–35.
34 Saldiva PHN, Pope CA, Schwartz J, *et al.* Air pollution and mortality in elderly people: a time series study in Sao Paulo, Brazil, *Arch Environ Health* 1996 (in press).
35 Schwartz J. Short term fluctuations in air pollution and hospital admissions for respiratory disease. *Thorax* 1995 (in press).
36 Touloumi G, Pocock SJ, Katsouyanni K, Trichopoulos D. Short term effect of air pollution on daily mortality in Athens: A time series analysis. *Int J Epidemiol,* 1994;**23**: 957–67.
37 Kinney PL, Ito K, Thurston GD. A sensitivity analysis of mortality/PM10 association in Los Angeles. *Inhalation Toxicology* 1995;7:59–70.
38 Ito K, Thurston GD. An investigation of sensitive subpopulations in daily PM10/mortality associations. *Journal of Exposure Assessment and Environmental Epidemiology* 1996 (in press).
39 Ostro BD, Sanchez JM, Aranda C, Eskeland GS. Air pollution and mortality: Results from a study of Santiago, Chile. *Journal of Exposure Assessment and Environmental Epidemiology* 1996 (in press).
40 Burnett RT, Krewski D, Vincent R, *et al.* Associations between ambient particulate sulfate and Admissions to Ontario Hospitals for cardiac and respiratory disease. *Am J Epidemiol* 1995:**142**:15–22.
41 Dockery DW, Pope CA, Xu X, *et al.* An association between air pollution and mortality in six US cities. *N Eng J Med* 1993;**329**:1753–9.
42 Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Health CW. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *Am J Resp Crit Care Med* 1995;**151**:669–74.
43 Gasser T, Muller HG, Mammitsch V. Kernals for nonparametric curve estimation. *Journal of the Royal Statistical Society B,* 1985;**47**:238–52.
44 Hastie T, Tibshirani R. *Generalized additive models.* London: Chapman and Hall, 1990.
45 Akaike H. Information theory and an extension of the maximum likelihood principle. In: EN Petrov, F Czaki eds. *Second international symposium on information theory.* Budapest: Akademiai Kiado, 1973;267–81.
46 Stone M. Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* 1974;**36**:111–47.
47 Golub G, Heath M, Wahba G. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* 1979;**21**:215–24.
48 Schwartz J, Spix C, Wichmann HE, Malin E. Air pollution and acute respiratory illness in five German communities. *Environ Res* 1991;**56**:1–14.
49 Daubechies I. *Ten lectures on wavelets.* Philadelphia: SIAM, 1992.
50 Thurston GD, Kinney PL. Air pollution epidemiology: considerations in time series modelling. *Inhalation Toxicology* 1995;7:71–84.
51 Bates DV, Baker-Anderson M, Sizto R. Asthma attack periodicity: a study of hospital emergency visits in Vancouver. *Environ Res,* 1987;**43**:317–31.
52 Schwartz J, Koenig J, Slater D, Larson T. Particulate air pollution and hospital emergency visits for asthma in Seattle. *Am Rev Respir Dis* 1993;**147**:826–31.
53 Almon S. The distributed lag between capital appropriations and expenditures *Econometrica* 1965;**33**:178–96.
54 McCullagh P, Nelder JA. *Generalized linear models* London: Chapman and Hall, 1989.
55 Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;**73**:13–22.