ORIGINAL ARTICLE

Applying a moving total mortality count to the cities in the NMMAPS database to estimate the mortality effects of particulate matter air pollution

S Roberts, M A Martin

Occup Environ Med 2006;**63**:193–197. doi: 10.1136/oem.2005.023317

See end of article for authors' affiliations

Correspondence to:
Dr S Roberts, School of
Finance and Applied
Statistics, Australian
National University,
Canberra ACT 0200,
Australia; steven.roberts@
anu.edu.au

Accepted
9 November 2005

Objectives: To apply a new method for estimating the association between daily ambient particulate matter air pollution (PM) and daily mortality to data from over 100 United States cities contained in the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database and to see whether the results from the 90 cities NMMAPS analysis are robust to this different modelling approach. This new method has recently been shown to provide improved estimates for the association between PM and daily mortality when every-day PM data are unavailable. It avoids the need for selecting a lag of PM at which the mortality effects of PM are to be investigated.

Methods: With the aid of analytical methods and databases developed for NMMAPS, Poisson log linear models controlling for long term trends and weather effects were used to estimate the association between PM and mortality for cities in the NMMAPS database using the new method. A two stage Bayesian hierarchical model was then used to combine city specific estimates to form a national average PM mortality effect estimate.

Results: A 10 $\mu\text{g}/\text{m}^3$ increase in PM was associated with a 0.12% increment in total mortality and a 0.17% increment in cardiovascular and respiratory mortality. These results are consistent with those found in the NMMAPS analysis.

Conclusions: There is a statistically significant association between short term changes in PM and mortality on average for the cities contained in the NMMAPS database. These findings are further evidence that this widespread pollutant adversely affects public health.

Numerous time series studies have investigated the association between daily mortality and some measure of daily ambient particulate matter air pollution (PM).^{1–7} These studies typically fit a generalised additive model⁸ or generalised linear model⁹ to concurrent time series of daily mortality, PM, and meteorological covariates. The general consensus from studies of this nature is that increases in PM are associated with increases in mortality. One limitation of these studies is that many have only used data from a single or a few locations. To address this limitation a number of recent studies have analysed data from a relatively large number of locations and combined the results to form regional and/or national average estimates for the effect of increases in PM on mortality.^{10–12} Two of the more prominent bodies of work of this nature are the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) which covers over 100 United States cities, and the Air Pollution and Health—A European Approach study which covers 22 European cities.

The United States NMMAPS analyses that have investigated the health effects of PM have typically been restricted to using a single day's PM as the measure of PM exposure. This is a consequence of PM measurements only being required every sixth day for regulatory purposes. The constraint of only being able to use a single day's PM means that the question of how the mortality effects of PM are distributed over time cannot be investigated. This is undesirable because evidence from past studies, including toxicological evidence, suggests that the effect of PM on mortality lasts for multiple days.^{13 14} In this vein, studies have shown that using a single day's PM can result in a sometimes large underestimation of the true effect of PM on mortality.^{14 15} These studies also show that the problem of underestimation

can be avoided if every-day PM data is available and a distributed lag model used to estimate the mortality effect of PM.

A new model has recently been introduced that allows inference on the information about the effect of PM on mortality that is lost when daily PM data is unavailable.¹⁶ This model uses a moving total mortality time series in place of the current day's mortality time series that is used in most time series studies of PM and mortality. It was shown that this model offers improved estimates for the effect of PM on mortality compared to the standard method of using the current day's mortality time series. In the paper by Roberts (2005) it was stated that this new model could be used to find improved estimates of the effect of PM on mortality in a large number of cities in the United States; however, the paper only analysed data from two cities. In this paper, we apply this new model to over 100 cities contained in the NMMAPS database and compare the results to those obtained using either the current, previous, or two day's previous PM concentration—that is, the lags of PM used in the recent 90 cities NMMAPS analysis.¹¹ Due to the importance of the NMMAPS analyses in the air pollution mortality debate it is essential that the robustness of these analyses to alternative model specifications is investigated.

METHODS

Data

The data used in this paper were obtained from the publicly available NMMAPS database, funded by the Health Effects Institute. The data consist of daily time series of mortality,

Abbreviations: NMMAPS, National Morbidity, Mortality, and Air Pollution Study; PM, particulate matter air pollution

weather, and PM for 109 United States cities for the period 1987–2000. Two different mortality time series were considered, the first being the number of non-accidental deaths (total mortality) and the second the number of deaths due to cardiovascular and respiratory disease. The weather time series data are 24 hour averages of temperature and dew point temperature, computed from hourly observations. The measure of PM used was the ambient 24 hour concentration of particulate matter of less than 10 μm in diameter, measured in units of $\mu\text{g}/\text{m}^3$. This data was all accessed via the NMMAPS database in R.¹⁷ Further details on the data used can be obtained at <http://www.ihapps.jhsph.edu/> and are discussed in previous NMMAPS analyses.¹⁸

Single city analyses

The basic model fit in the NMMAPS analyses to the data from a single city takes the following form:¹⁹

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \\ \log(\mu_t) &= \text{confounders}_t + \beta \text{PM}_{t,l} \\ \text{confounders}_t &= \text{DOW} + \text{Age} + s(\text{temp}_t, df = 6) \\ &\quad + s(\text{temp}_{t,1-3}, df = 6) \\ &\quad + s(\text{dew}_t, df = 3) + s(\text{dew}_{t,1-3}, df = 3) \\ &\quad + s(t, df = 7 \times \# \text{ years}) \\ &\quad + s(t, df = 0.15 \times 7 \times \# \text{ years}) \times \text{Age} \\ \text{Var}(Y_t) &= \phi \mu_t \end{aligned} \quad (1)$$

where Y_t is the number of deaths from a particular cause on day t for a particular age category, DOW is an indicator variable for the day of the week, Age is an indicator variable for the age category to which the deaths belong (<65 years, 65–74 years, and ≥ 75 years), temp_t is the average temperature on day t , $\text{temp}_{t,1-3}$ is the previous three days' average temperature, dew_t and $\text{dew}_{t,1-3}$ are similarly defined quantities for dew point temperature, ϕ is a parameter allowing for over dispersion in the observed number of deaths, and $\text{PM}_{t,l}$ is the lag l PM concentration. The $s(\cdot)$ are smooth functions of temperature, dew point temperature, and time. The smoothness of the functions is controlled by the degrees of freedom (df): smaller degrees of freedom are associated with smoother functions. The term $s(t, df = 7 \times \# \text{ years})$ controls for seasonally varying factors such as winter influenza epidemics, using 7 degrees of freedom per year. The term $s(t, df = 0.15 \times 7 \times \# \text{ years}) \times \text{Age}$ controls for age specific mortality trends, using $0.15 \times 7 = 1.05$ degrees of freedom per year. Justification for the choice of number of degrees of freedom used in the NMMAPS analyses can be found in previous studies.¹⁹ In this paper, the smooth functions are represented using natural cubic splines; other common choices are smoothing splines or penalised splines.

In the above model, the climate related variables (temperature and dew point temperature) are only included up to a lag of three days. One concern is whether three is a sufficient number of lags to adequately control for the confounding effects of weather conditions. A recent paper addressed this issue by investigating whether the results of the NMMAPS analyses were partially a consequence of inadequate control for weather and season.²⁰ Through the assessment of a number of competing models, including the use of 2 weeks of temperature lags, the study concluded that the national average PM effect estimates obtained using models similar to Model (1) were robust to model specification for weather and seasonal confounding.

Notably, in the above model, the PM exposure measure is restricted to being a single day's PM. For example, using the previous day's PM concentration would mean Model (1) was

fit using the lag 1 PM concentration. Once Model (1) is fit to the individual city data, the estimates of the mortality effect of PM obtained from each city (the estimates of β in each city) are combined to form a national average estimate. The process of combining the estimates will be described in more detail below.

The model described by Roberts (2005) relies on the fact that daily mortality counts are available for cities in the NMMAPS database irrespective of the sampling frequency used for PM. The model takes advantage of this fact by using information available in the daily mortality data to extract information about the effect of PM on mortality over a period of more than a single day, otherwise unavailable with every-six-day PM measurements. To do this, the current day's mortality count used in Model (1) is replaced by a moving total mortality count. The moving total used is a forward moving total, meaning that the current day's mortality count is replaced by the sum of the current day's mortality count, the next day's mortality, and so on for some specified number of days. We will use the term “ k day moving total” to mean the sum of today's and the subsequent $k-1$ days' mortality counts. Under this model, the k day moving total mortality counts are modeled as independent Poisson random variables with a time varying mean $\mu_{t,k}^{\text{tot}}$ on day t given by:

$$\begin{aligned} Y_{t,k}^{\text{tot}} &\sim \text{Poisson}(\mu_{t,k}^{\text{tot}}) \\ \log(\mu_{t,k}^{\text{tot}}) &= \text{confounders}_t + \beta \text{PM}_{t,0} \\ \text{confounders}_t &= \text{DOW} + \text{Age} + s(\text{temp}_t, df = 6) \\ &\quad + s(\text{temp}_{t,1-3}, df = 6) \\ &\quad + s(\text{dew}_t, df = 3) + s(\text{dew}_{t,1-3}, df = 3) \\ &\quad + s(t, df = 7 \times \# \text{ years}) \\ &\quad + s(t, df = 0.15 \times 7 \times \# \text{ years}) \times \text{Age} \\ \text{Var}(Y_{t,k}^{\text{tot}}) &= \phi \mu_{t,k}^{\text{tot}} \end{aligned} \quad (2)$$

where $Y_{t,k}^{\text{tot}}$ is the k day moving total mortality count for a particular cause on day t for a particular age category and is the current day's or lag 0 PM concentration. The lag 0 PM concentration is used because it allows the effect of PM on the current and subsequent $k-1$ days to be investigated. The reason for this is that if the mortality effects of PM last for more than a single day, a day of high PM will not only cause the current day's mortality count to be elevated but also the mortality counts on subsequent days. By using a moving total mortality count we are able to capture some of the increased mortality on subsequent days, information lost if only the current day's mortality count is used.

In this paper, Model (1) will be fit using the lag 0, lag 1, and lag 2 PM concentrations because these are the lags of PM that were investigated in the recent NMMAPS multi-city analyses.¹¹ Model (2) will be fit using a three day moving total mortality count ($k = 3$) because this allows the effect of PM on the current and subsequent two days' mortality to be explored, the same lag effects of PM that are investigated when the lag 0, lag 1, and lag 2 PM concentrations are used. Models (1) and (2) were fit using R to the cities in the NMMAPS database.¹⁹

Combining single city estimates

The estimates for the effect of PM from each city (the estimates of β) were combined using a two level Bayesian hierarchical model.²⁰ Using this model, the city specific PM effect estimates can be combined to form a pooled or national average PM effect estimate. The model assumes that $\beta_i | \beta_i, \sigma_i \sim N(\beta_i, \sigma_i)$ and that β_i are independent $N(\beta^p, \sigma^p)$. The model also assumes uniform prior distributions for β^p and σ^p . Here $\hat{\beta}_i$ is the PM effect estimate for city i obtained

Table 1 National average PM effect estimates obtained from Model (1) and Model (2)

Model	Base <i>df</i>	Double <i>df</i>	Half <i>df</i>
All cause mortality			
Model (1), lag 0	0.079 (0.050)*	0.044 (0.046)	0.107 (0.052)
Model (1), lag 1	0.180 (0.044)	0.149 (0.047)	0.254 (0.048)
Model (1), lag 2	0.059 (0.056)	0.024 (0.056)	0.143 (0.054)
Model (2), <i>k</i> = 3	0.115 (0.037)	0.107 (0.034)	0.145 (0.039)
Cardiovascular and respiratory mortality			
Model (1), lag 0	0.103 (0.068)	0.056 (0.067)	0.134 (0.066)
Model (1), lag 1	0.232 (0.060)	0.179 (0.067)	0.309 (0.059)
Model (1), lag 2	0.210 (0.078)	0.144 (0.075)	0.305 (0.079)
Model (2), <i>k</i> = 3	0.168 (0.047)	0.140 (0.044)	0.196 (0.051)

*1000 times the national average PM effect estimate. The value in parentheses is the standard error of this estimate.

from Model (1) or Model (2), β_i is the unknown “true” effect of PM on mortality for city i with standard deviation σ_i , and β^p is the pooled or national average PM mortality effect. The two level Bayesian model described here has been used in previous multi-city studies.^{20–22} The model was implemented using the freely available TLNise statistical software.²¹ An alternative method for combining the city specific estimates which used weights that were inversely proportional to the variance of the estimates was also considered but gave very similar results to the two level Bayesian hierarchical model.

RESULTS

Table 1 contains the national average PM effect estimates at lags of 0, 1, and 2 obtained from pooling the city specific estimates obtained from Model (1) and the national average effect estimate obtained from pooling the city specific estimates obtained from Model (2). This table also contains

the national average estimates obtained by simultaneously halving or doubling the degrees of freedom used in the smooth functions of time, temperature, and dew point temperature in Model (1) and Model (2). This gives some indication of how sensitive the results are to the choice of degrees of freedom in the smooth functions. The discussion that follows will focus on the base degrees of freedom estimates, that is, the estimates based on the original degrees of freedom appearing in the definitions of Model (1) and Model (2).

The estimates obtained from Model (1) for the three PM exposure measures are quite different. For example, for total mortality the effect estimate obtained using the lag 1 PM concentration is three times larger than the estimate obtained using the lag 2 PM concentration and the national average effect estimate is only statistically different from zero for the lag 1 PM concentration. This is undesirable because it means

Figure 1 City specific PM effect estimates and corresponding 95% confidence intervals for a random sample of 20 cities in the NMMAPS database using Model (1) and Model (2).

Main messages

- Our study applies a new method for estimating the association between PM and daily mortality to the over 100 cities contained in the NMMAPS database.
- The results of this study show that the conclusions of the NMMAPS analysis are robust to an alternative model specification and provide further evidence that this widespread pollutant adversely affects public health.
- Due to the importance of the NMMAPS analyses in the air pollution mortality debate it is essential that the robustness of these analyses to alternative model specifications be investigated.

the results are sensitive, in terms of statistical significance, to the choice of lag that is used for reporting the final results. In the NMMAPS analyses the results obtained using the lag 1 PM concentration have been the most widely reported. Model (2) avoids this problem because in a situation where only the lag 0, lag 1, and lag 2 PM concentrations are being considered the natural choice for Model (2) is $k = 3$, meaning only one form of Model (2) would be fit. The national average estimates obtained from Model (2) are roughly equal to the average of the estimates obtained from the three different PM lags using Model (1). The estimates obtained from Model (2) are also more precise (smaller standard error) than the estimates obtained using Model (1), a consequence of the fact that the estimates obtained using Model (2) are based on more data. For both models the estimates obtained increased when the number of degrees of freedom used halved, and decreased when the number of degrees of freedom used doubled. However, these changes were immaterial and suggest that the results are not particularly sensitive to the choice of number of degrees of freedom.

An important observation from table 1 is that the sum of the three individual lagged coefficient estimates from Model (1) is substantially larger than the estimate obtained using Model (2). Initially, one might suspect that the sum of the three estimates obtained from Model (1) should be similar to the single estimate obtained from Model (2) using a three day moving total mortality count. However, there are a number of reasons why this typically will not be the case:

- The three estimates obtained from Model (1) are highly correlated, which means that summing the three estimates is, in a sense, double or even triple counting, depending on the extent of the correlation.
- The moving total model is not a perfect substitute for having every-day PM data and will tend to produce estimates that have a negative bias, hence the estimate obtained from the moving total model should be considerably smaller than the positively biased estimate obtained by summing the three estimates obtained from Model (1).
- Even though the combined estimates obtained from Model (1) are all positive, for the individual city analyses a significant proportion of the estimates obtained from Model (1) were negative (approximately 43% in the case of all cause mortality). In cities with one or more negative estimates from Model (1), we would anticipate that the sum of the estimated coefficients from Model (1) would be closer to the Model (2) estimate than is the case for the national average case.

Figure 1 contains the city specific PM effect estimates and corresponding 95% confidence intervals for a random sample

Policy implications

- The health effects of air pollution have been documented in many studies and the results are used for regulatory decisions. It is critical that the results of these studies are checked for robustness to model specification.
- Actions to reduce the ambient concentration of particulate matter air pollution would have beneficial results on the health of the United States population.

of 20 cities in the NMMAPS database using Model (1) and Model (2). From this figure, it also appears that the city specific estimates obtained from Model (2) are roughly an average of the three city specific estimates obtained from Model (1). The large variability in the estimates obtained for the various cities illustrates one of the reasons why obtaining pooled estimates based on a large number of cities is valuable.

DISCUSSION

For time series studies that investigate the health effects of PM, the lack of daily PM measurements in the majority of United States cities is problematic. Two of the more important problems are that distributed lag models which investigate the health effects of PM on multiple days cannot be fit and a choice must be made about which lag of PM will be used for reporting of results. Assuming that PM cannot be beneficial to health and that mortality displacement is not occurring, both of which would result in negative associations between PM and mortality, the first of these two problems will result in an underestimation of the adverse health effects of PM. The second of the two problems can be avoided to some degree by estimating and reporting the health effects of PM at multiple lags. However, invariably this will lead to selection effect problems as people will tend to focus on the largest or most statistically significant estimate. In the NMMAPS analyses the health effects of PM were estimated and reported using the lag 0, lag 1, and lag 2 PM concentrations.¹¹ In these studies the pooled national average PM effect estimate based on the lag 1 PM concentration was the estimate that was focused on, perhaps because it was the largest of the three estimates and in the case of total mortality the only lag that had a pooled effect estimate that was statistically different from zero. The problem of selecting a PM lag for reporting purposes can be overcome by the use of the moving total mortality count of Model (2), which does not require multiple model fits. The results of applying this model to the cities in the NMMAPS database resulted in similar conclusions to those obtained using the lag 1 PM concentration—a positive association between PM and increases in daily mortality.

A recent study showed that the results from the NMMAPS analyses were robust to model specification for weather and seasonal confounding.²⁰ This was an important result because the importance and influence of the NMMAPS analyses means that ensuring that the results are robust to different modelling choices is essential, particularly in light of the recent problems associated with the application of generalised additive modelling software in these types of studies.¹¹ The results presented in this paper, which used a different modelling strategy to the NMMAPS analyses, provide further evidence of the robustness of the results obtained from these analyses. It is important to note that we are not advocating that the moving total count of Model (2) replace finding and

reporting the effect estimates at different lags of PM as done in the NMMAPS analyses, but rather that Model (2) is a useful means of obtaining one PM effect estimate for reporting purposes that avoids selection effect issues.

Authors' affiliations

S Roberts, M A Martin, School of Finance and Applied Statistics, Australian National University, Canberra ACT, Australia

Competing interests: none

REFERENCES

- 1 Ito K, Kinney PL, Thurston GD. Variations in PM-10 concentrations within two metropolitan areas and their implications for health effects analyses. *Inhal Toxicol* 1995;**7**:735–45.
- 2 Rahlénbeck SI, Kahl H. Air pollution and mortality in East Berlin during the winters of 1981–1989. *Int J Epidemiol* 1996;**25**:1220–6.
- 3 Ostro BD, Hurley S, Lipsett MJ. Air pollution and daily mortality in the Coachella Valley, California: a study of PM10 dominated by coarse particles. *Environ Res* 1999;**81**:231–8.
- 4 Chock DP, Winkler SL, Chen C. A study of the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. *J Air Waste Manage Assoc* 2000;**50**:1481–500.
- 5 Gouveia N, Fletcher T. Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status. *J Epidemiol Community Health* 2000;**54**:750–5.
- 6 Ballester F, Saez M, Perez-Hayos, et al. The EMECAM project: a multicentre study on air pollution and mortality in Spain: combined results for particulates and for sulphur dioxide. *Occup Environ Med* 2002;**59**:300–8.
- 7 Roberts S. Interactions between particulate air pollution and temperature in air pollution mortality time series studies. *Environ Res* 2004;**96**:328–37.
- 8 Hastie TJ, Tibshirani RJ. *Generalized additive models*. London: Chapman and Hall, 1990.
- 9 McCullagh P, Nelder JA. *Generalized linear models*. London: Chapman and Hall, 1989.
- 10 Le Tertre A, Medina S, Samoli E, et al. Short-term effects of particulate air pollution on cardiovascular diseases in eight European cities. *J Epidemiol Community Health* 2002;**56**:773–9.
- 11 Dominici F, McDermott A, Daniels M, et al. *Revised analyses of the national morbidity, mortality, and air pollution study, Part II*. Cambridge, MA: The Health Effects Institute, 2003.
- 12 Samoli E, Analitis A, Touloumi G, et al. Estimating the exposure-response relationship between particulate matter and mortality within the APHEA project. *Environ Health Perspect* 2005;**113**:88–95.
- 13 Clarke RW, Catalano PJ, Koutrakis P, et al. Urban air particulate inhalation alters pulmonary function and induces pulmonary inflammation in a rodent model of chronic bronchitis. *Inhal Toxicol* 1999;**11**:637–56.
- 14 Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology* 2000;**11**:320–6.
- 15 Roberts S. An investigation of distributed lag models in the context of air pollution and mortality time series analysis. *J Air Waste Manage Assoc* 2005;**55**:273–82.
- 16 Roberts S. Using moving total mortality counts to obtain improved estimates for the effect of air pollution on mortality. *Environ Health Perspect* 2005;doi:10.1289/ehp.7774.
- 17 Peng RD, Welty LJ. The NMMAPSdata Package. *R News* 2004;**4**:10–14 (<http://cran.r-project.org/doc/Rnews/>).
- 18 Samet JM, Zeger S, Dominici F, et al. *The National Morbidity, Mortality, and Air Pollution Study (NMMAPS), Part I. Methods and methodological issues*. Cambridge, MA: Health Effects Institute, 2000.
- 19 Peng RD, Welty LJ, McDermott A. *The National Morbidity, Mortality, and Air Pollution Study Database, Working Paper 44*. R. Johns Hopkins University, Dept of Biostatistics, 2004.
- 20 Welty LJ, Zeger SL. Are the acute effects of particulate matter on mortality in the National Morbidity, Mortality, and Air Pollution Study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *Am J Epidemiol* 2005;**162**:80–8.
- 21 Everson PJ, Morris CN. Inference for multivariate normal hierarchical models. *J R Statist Soc B* 2000;**62**:399–412.
- 22 Bell ML, McDermott A, Zeger SL, et al. Ozone and short term mortality in 95 urban communities, 1987–2000. *JAMA* 2004;**292**:2372–8.