Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2003/06

Estimates of the short term effects of air pollution in Italy using alternative modelling techniques

A. Biggeri, M. Baccini, G. Accetta,
C. Lagazio, J. Schwartz

Università degli Studi
di Firenze

Epidemiology

Estimates of the short term effects of air pollution in Italy using alternative modelling techniques

Abstract

Recently serious criticism was raised about the use of standard statistical software to fit Generalized Additive Models (GAM) to epidemiological time series data. Inappropriate settings of convergence parameters in the backfitting algorithm (implemented by Splus) results in inaccurate inference about the effect of linear covariates. Moreover standard errors are underestimated, since only the linear component of the smoother(s) is included in the variance-covariance matrix computation. A Splus macro for approximate standard errors has been recently proposed. We analysed the association between PM10 and Mortality/Hospital Admissions in the Italian Meta-analysis of Short-term effects of Air pollutants (MISA), using GAM with penalized regression spline fitted by the direct method in R software (GAM-R), which correctly computes the variance-covariance matrix. A comparison with default GAM with smoothing spline fitted via backfitting in Splus (GAM-S) and with Generalized Linear Models with natural cubic spline (GLM+NS) is provided. GLM+NS and GAM-R give similar results. For total mortality GLM+NS and GAM-R gave respectively an overall percent increase of 0.98 (95 percent confidence interval: 0.35,1.61; random effects model) and 1.04 (0.41,1.67) for 10 $\mu g/m^3$ of PM10, compared to the GAM-S estimate of 1.24 (0.63,1.86). The GLM+NS and GAM-R standard errors are consistently higher than GAM-S ones.

Introduction

Short-term effects of air pollution on health are widely documented and several meta-analyses were conducted (1-7).
Recently major concern was raised about numerical accuracy of the estimates of pollutant effect obtained fitting Generalized additive Models (GAM) (8). Ramsay et al. (9) and Dominici et al. (10) identified important critical points in the analyses of epidemiological time series using commercial statistical software which fits GAM by backfitting algorithm. In brief:

1. the estimated standard errors obtained fitting GAM in Splus or SAS are biased;
2. the default convergence criteria of backfitting algorithm defined in Splus (and, to a lesser degree, in SAS) are too lax to assure convergence and lead to biased estimates of pollutant effect.

For the point 1, these statistical software provide an approximation of the variance-covariance matrix, which takes into account only the linear component of the variable that was fit with a smooth function (11). Then, a bias is expected whenever strong non-linearity and non-orthogonality between parametric and non-parametric terms are present, such as between the smoother for time and the pollutant concentration in epidemiological time series analysis (9).
The second point is for certain aspects trivial: whenever the magnitude of the effect to be estimated is of the same order of the convergence criteria some degree of numerical instability is expected, which decreases as the effect size increases. More interestingly, if data exhibit relevant degree of concurvity ("collinearity" among parametric and non-parametric components of the model (9)), convergence of backfitting algorithm can be very slow (11-12). Dominici et al. (10) showed that, when a spline for time and a spline for weather are included in the model, the greater the degree of concurvity, the greater is the overestimation of the pollutant effect.

The present paper analyses the data of the Italian Meta-analysis of Short-term Effects of Air Pollution (MISA) (13-14), using alternative modelling approaches to GAM fitted by backfitting algorithm: GLM with natural cubic spline(s) and GAM fitted by the *gam* function implemented for R software by Wood (15). These approaches do not have the difficulties described above.

Methods

The characteristics of epidemiological time series data require statistical methods able to control for nonlinear confounding effect of temporal trend. One approach to dealing with temporal trend is to divide the time span of the study into shorter periods and fit separate polynomials within each range. This does not require the same polynomial to fit different ranges. Natural cubic splines are a form of this approach. While flexible, they can still be sensitive to the position of break points between the time periods (knots). To avoid this, many air pollution studies used more flexible semi-parametric approaches, specifying models with smoothing splines or locally weighted regressions in moving ranges of the data (loess). These models, belonging to the class of GAM (11), are those implemented in Splus and SAS by backfitting algorithm.
In the following analyses, we specified GAMs with penalized regression spline. Penalized regression splines use separate polynomials in each range (as natural splines do), but they reduce the sensitivity to knots location by using many of them and avoid excessively wiggly curves by constraining the coefficients not to change too much between one break point and another (16). The use of penalized regression splines eliminates the backfitting algorithm of GAM, while still providing the flexibility of smoothing splines. This approach is implemented in R.

MISA investigated mortality for all natural causes and for cardiovascular and respiratory diseases and hospital admissions for cardiovascular and respiratory diseases. Health data were collected from Local Health Authorities and regional files. Daily pollutant concentrations were obtained from Regional Environmental Protection Agencies or local sources. The same procedure for collecting data was used in all participating cities (Turin, Milan, Verona, Ravenna, Bologna, Florence, Rome, Palermo).
MISA used a common model for city specific analysis (13). The analysis was age-stratified (0-64; <65-74; 75+). Parametric terms for weather and only one spline for time were included in the model. The number of degrees of freedom for the splines was specified a priori (5 per year for mortality only for the third age class, since indicator variables for season were used for the first two; 6-5-6 per year for hospital admissions for cardiac diseases for the three age classes, respectively; 7-5-6 per year for hospital admissions for respiratory diseases). Residuals analysis and sensitivity analysis were performed.
This modelling strategy could apparently be reassuring given Dominici's results apply to models including at least two smoothing terms (10), but, as we show, this is not true.

We evaluate the sensitivity of our results to different modelling strategies, fitting:

- GAM by backfitting algorithm using Splus with default (of the order $\varepsilon<10^{-3}$) or stringent ($\varepsilon<10^{-14}$) convergence criteria (GAM-S);

- GAM by direct method implemented in R 1.6.1 Software (17) (GAM-R);

- GLM with natural cubic spline with fixed pre-specified knots (18) fitted by standard IRLS algorithm (GLM+NS).

The function *gam* of R allows inclusion in the model of penalized regression splines whose smoothing parameters are fixed to obtain the desired number of degrees of freedom or selected by Generalized Cross Validation method. This function maximizes the penalized likelihood by a direct method which avoids the iterative process nested in the backfitting algorithm (12, pp.69-70; 15).
The GAM implementation in R correctly calculates the variance-covariance matrix.

GLM+NS is a fully parametric alternative to GAM (10). The main drawback to using GLM+NS stands in the dependence of the fitted curve on the knots position (11). In this analysis knots were placed evenly throughout the covariate values. Comparing the two different approaches (GAM-R and GLM+NS) we checked the sensitivity of results obtained fitting a parametric natural spline to alternative specifications of the knots position.

We present below the results of the Italian Meta-analysis (MISA) on short-term effect of PM10 for the calendar period 1995-1999. For mortality (available data from 6 cities) lag 0-1 is used, while for hospital admissions (7 cities) we chose lag 0-3. The combined meta-analytic estimates were calculated using fixed and random effects models (19).

Results

As expected, we did find relevant disagreement between the standard GAM-S fitted by backfitting, with default or more stringent criteria, and the two alternative approaches, GLM+NS and GAM-R.

Figure 1 (a-b) shows the results from GAM-S with default convergence criteria (coefficient estimates and estimated standard errors) on the Y-axis versus the results from GLM+NS,

respectively, in the X-axis. Each point in the figure corresponds to a city-specific estimate; points marked as bold squares represent combined meta-analytic estimates (3 mortality outcomes, 2 hospital admission outcomes). They look quite similar to those of Dominici et al. (10), even if the model includes only one smoothing term: the GLM+NS coefficient estimates are generally lower and the estimated standard errors are greater, proportionally to their magnitude, than those obtained from GAM-S with default convergence criteria.
Figure 1 (c-d) shows the results using GAM-R on the Y-axis versus GLM+NS on the X-axis. The results are consistent. However point estimates from GLM+NS are usually lower than point estimates obtained from GAM-R.
Estimated standard errors by GAM-S did not change using the more stringent convergence criteria, while point estimates become very close to GAM-R (Fig. 1 (e-f)).

Table 1 reports the results of meta-analyses. Overall the main conclusions do not change using the different approaches. For most of the outcomes considered, effects are statistically significant, although using GLM+NS and GAM-R their magnitude is lower and confidence intervals are wider than using GAM-S with default convergence criteria. Greater uncertainty emerged with regard to the association with respiratory diseases.
Using GLM+NS, the overall estimated percent increase of total mortality for natural causes for 10 $\mu g/m^3$ increase of PM10 was 0.98 (95 percent confidence interval: 0.35,1.61; random effects model) in the calendar period 1995-1999, for a lag time of 0-1 day. Using GAM-R, the overall estimated percent increase was 1.04 (0.41,1.67). These compare with the biased estimate of 1.24 (0.63,1.86) from GAM-S with default convergence criteria.

Discussion

Modelling epidemiological time series presents difficulties due to: (a) the small order of magnitude of the effects; (b) the strong confounding effect of seasonality/time trend and weather.
Since the 1990s, the use of GAM became common, allowing flexible and local non-parametric modelling of confounders (20). The statistical software commonly used to fit GAMs is based on backfitting algorithm, which could be affected by problems of convergence and produce biased effect estimates, depending on the degree of concurvity in data (10-12).
Moreover, coherently with Ramsay et al. (9) we found that estimated standard errors from GAM fitted by Splus (and SAS), even with stringent convergence criteria, are invalid. The reason lies in the fact that the calculated variance-covariance matrix does not take into account the non-linear components of the smoother(s).

A possible alternative to GAM via backfitting is to specify GLM with parametric natural spline(s) (10). This solution is exempt of problems of convergence, but the number and position of knots must be specified a priori. This could be a limit for applicability of such approach to many situations (11-12). To reduce this problem we used only one spline for time. Since time is an equally spaced covariate (it indexes days under study) and the number of knots is not small (between 5 to 7 per years), we do not expect relevant differences from different knots specifications, but we do not exclude that, in presence of particular local variations of seasonality, results could be sensitive to knots placement. Moreover, the choice of knots positions could be a more important problem for other covariates like temperature.

A second alternative to GAM-S consists in fitting GAM by the direct method implemented for R by Wood (15). After fixing appropriate convergence criteria, the backfitting algorithm and the direct method provide similar point estimates of pollutant effect. However, the direct method is not affected by convergence problems and, being the smoothing terms penalized regression splines,

offers significant computational efficiencies over smoothing splines in large datasets and requires less computation for standard errors (16, 21). Here we found that standard errors of particles effect estimates are consistently estimated by GLM+NS and by GAM-R.

Conclusion

On the one hand GAMs with penalized regression spline(s) provides a more flexible approach in modelling epidemiological time series data than Generalized Linear Models with natural spline(s). On the other hand inference based on GAMs suffers disadvantage to be not well established. In our context the most relevant issue is the choice of the appropriate amount of smoothing for non-parametric function(s) or degrees of freedom for the parametric splines. Different formal and informal selection strategies have been proposed (7, 22-23), but further investigations are necessary. Asymptotic results about consistency of estimates in Additive Models indicates that the amount of smoothing appropriate for good estimation of parametric coefficients may be less than the amount appropriate for the optimal estimation of non-parametric part of model (24-26). Pursuing the optimal curve in adjusting for non-linear confounders could lead to biased estimate of pollutant effect, while a certain degree of undersmoothing could assure better inference. As a strictly related problem, robustness of inference based on GAM to alternative choices of smoothing parameter should be investigated (27, personal communication).

In conclusion, we could consider many epidemiological analyses of epidemiological time series as mainly exploratory analyses. Indeed, strong model selection and the use of semi-parametric models with substantial concurvity among pollutant and weather/time resulted in unstable estimates and underestimated standard errors, when standard software was used.
Major recent meta-analyses used a priori choices of smoothing parameters and more uniform criteria of model specification for city-specific analyses than single published studies. This reduced the risk of “multiple looks” at the data. Provided GLM+NS or GAM fitted by direct method have been used, point and interval estimates are valid.
Finally, the uncertainty about the size of the effects is not small. We are looking to extremely small effects, even if very important from a Public Health point of view, and some instability in the city-specific estimates has to be expected, and cannot be avoided. Combined meta-analytic estimates showed no major inconsistencies and retained statistical significance.

References

1 Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery D, Schwartz J, Zanobetti A. The National Morbidity, Mortality, and Air Pollution Study (HEI Project No. 96-7): morbidity and mortality from air pollution in the United States. Health Effects Institute, Cambridge, MA, 2000.

2 Atkinson RW, Anderson HR, Sunyer J, et al. Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project. Air Pollution and Health: a European Approach. American Journal of Respiratory Critical Care Medicine 2001; 164:1860-66.

3 Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopolis Y et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project. Epidemiology 2001; 12:521-531.

4 Burnett RT, Cakmak S, Brook JR. The effect of the urban ambient air pollution mix on daily mortality rates in 11 Canadian cities. Canadian Journal of Public Health 1998; 89:152-6.

5 Ballester Diez F, Saez Zafra M, Perez-Hoyos S, et al. The EMECAM project: a discussion of the results in the participating cities. Estudio Multicentrico Espanol sobre la Relacion entre la Contaminacion Atmosferica y la Mortalidad. Revista Espanola de Salud Publica 1999; 73: 303-14. (spanish)

6 Saez M, Ballester F, Barcelo MA, Perez-Hoyos S, Bellido J, Tenias JM et al. A combined analysis of the short term effects of photochemical air pollutants on mortality within the EMECAM project. Environmental Health Perspectives 2002; 110:221-228.

7 Hoek G, Brunekreef B, Verhoeff A, van Wijnen J, Fischer P. Daily mortality and air pollution in The Netherlands. Journal of Air Waste Management Association 2000; 50:1380-1389.

8 Kaiser J. Software Glitch Threw Off Mortality Estimates. Science 296, 14 June 2002:1945-46.

9 Ramsay TO, BurnettRT, Krewski D. The Effects of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter. Epidemiology 2003; 14:18-23.

10 Dominici F, McDermott A, Zeger SL, Samet J. On the use of Generalized Additive Models in Time-series Studies of Air Pollution and Health. American Journal of Epidemiology 2002; 156, 193-203.

11 Hastie TJ, Tibshirani RJ. Generalized Additive Models. London:Chapman & Hall, 1990.

12 Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models. London:Chapman & Hall, 1994.

13 Biggeri A, Bellini P, Terracini B. Meta-analysis of the italian studies on short-term effects of air pollution. Epidemiologia e Prevenzione 2001; 25 (suppl):1-72. (italian)

14 Biggeri A, Baccini M., Accetta G., Lagazio C. Estimates of short-term effects of air pollutants in Italy. Epidemiologia e Prevenzione 2002; 26(4):203-205. (italian)

15 Wood SN. Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. Journal of the Royal Statistical Society B 2000; 62:413-428.

16 Currie ID, Durban M. Flexible smoothing with P-splines: a unified approach. Statistical Modelling 2002; 4: 333-349.

17 the R Development Core Team. 2002. R Language Version 1.5.1. ISBN 901167-55-2 http://cran.r-project.org

18 de Boor C. A Practical Guide to Splines. New York:Springer Verlag, 1978.

19 DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7:177-188.

20 Schwartz J. The use of generalized additive models in epidemiology. Proceedings in XVII International Biometric Society Conference, Hamilton, Ontario 1994; 55-80.

21 Lumley T, Sheppard L. Time series analyses of air pollution on health: straining at gnats and swallowing camels? Epidemiology 2003; 14: 13-14.

22 Schwartz J, Spix C, Touloumi I, Bacharova L et al. Methodological issues in studies of air pollution and daily counts of deaths or hospital admission. Journal of Epidemiology and Community Health 1996; 50 (suppl. 1): S3-S11.

23 Kelsall J, Samet J, Zeger S. Air pollution and mortality in Philadelphia, 1974-1988. American Journal of Epidemiology 1997; 146: 750-762.

24 Heckman N. Spline Smoothing in partly linear model. Journal of the Royal Statistical Society B 1986; 48, 244-248.

25 Rice J. Convergence rates for partially splined models. Statist. Probabil. Letters 1986; 4, 203-208.

26 Cuzick J. Semiparametric additive regression. Journal of the Royal Statistical Society B 1992; 54: 831-843.

27 Dominici F, McDermott A, Hastie T. Issues in Semi-parametric Regression with Applications in Time-series Models for Air Pollution and Mortality. 2003; unpublished manuscript.

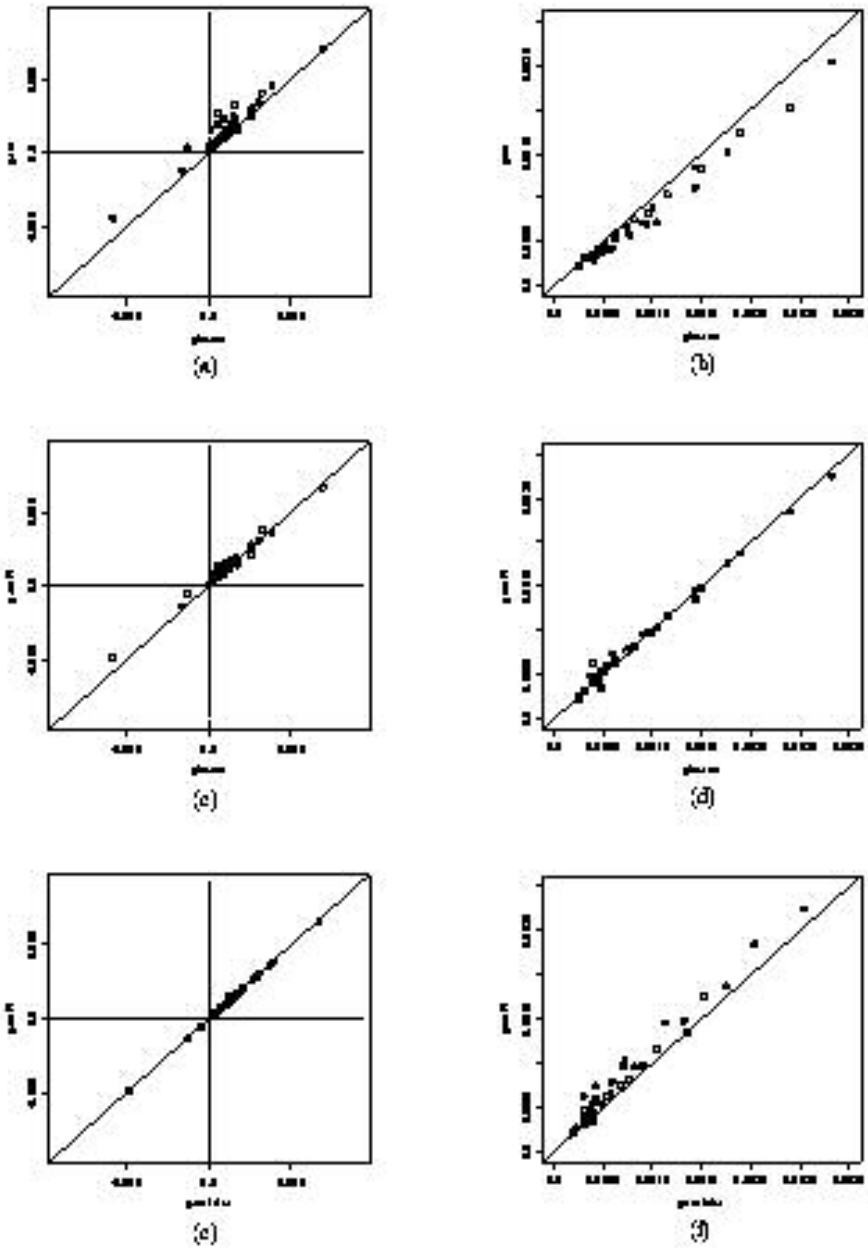

Figure 1. Italian Meta-analysis of Short-term Effects of Air Pollution. MISA 1995-1999.
(a) City specific and Meta-analytic (in square bold) Effect Estimates (log Relative Risk) for PM10 (increase of 10 $\mu g/m^3$) by fitting GAM-default settings (Y-axis) vs GLM-NS (X-axis).
(b) City specific and Meta-analytic (in square bold) Standard Error Estimates for PM10 effects by fitting GAM-default settings (Y-axis) and GLM-NS (X-axis).
(c) City specific and Meta-analytic (in square bold) Effect Estimates (log Relative Risk) for PM10 (increase of 10 $\mu g/m^3$) by fitting GAM-direct method in R Language (Y-axis) vs GLM-NS (X-axis).
(d) City specific and Meta-analytic (in square bold) Standard Error Estimates for PM10 effects by fitting GAM-direct method in R Language (Y-axis) and GLM-NS (X-axis).
(e) City specific and Meta-analytic (in square bold) Effect Estimates (log Relative Risk) for PM10 (increase of 10 $\mu g/m^3$) by fitting GAM-direct method in R Language (Y-axis) vs GAM-stringent convergence criteria (X-axis).
(f) City specific and Meta-analytic (in square bold) Standard Error Estimates for PM10 effects by fitting GAM-direct method in R Language (Y-axis) and GAM-stringent convergence criteria (X-axis).

	Mortality						Hospital Admissions			
	All Natural Causes		Cardiovascular		Respiratory		Cardiac		Respiratory	
Method	fixed	random	fixed	random	fixed	random	fixed	random	fixed	random
GAM-S default										
	1.12	1.24	1.23	1.43	2.24	1.96	1.23	1.30	2.13	2.35
	0.82;1.42	0.63;1.86	0.76;1.69	0.62;2.25	1.09;3.41	-0.69;4.68	0.93;1.53	0.83;1.78	1.76;2.50	1.52;3.18
GAM-S stringent										
	0.92	1.06	1.03	1.24	1.96	1.69	0.99	1.02	1.26	1.42
	0.62;1.22	0.46;1.66	0.57;1.50	0.43;2.06	0.81;3.13	-0.97;4.42	0.69;1.29	0.64;1.39	0.89;1.63	0.66;2.20
GAM-R										
	0.90	1.04	1.05	1.26	1.92	1.70	0.95	0.95	1.34	1.34
	0.55;1.25	0.41;1.67	0.52;1.58	0.41;2.12	0.64;3.21	-0.96;4.43	0.50;1.40	0.50;1.40	0.84;1.86	0.64;2.03
GLM+NS										
	0.85	0.98	0.97	1.21	1.74	1.41	0.77	0.82	0.73	0.91
	0.52;1.18	0.35;1.61	0.45;1.50	0.32;2.10	0.44;3.05	-1.41;4.32	0.40;1.15	0.32;1.32	0.27;1.20	-0.04;1.86

Table 1. Italian Meta-analysis of Short-term Effects of Air Pollution. MISA 1995-1999. Combined meta-analytic estimates of percentage increase in outcome (95% CI) associated to a PM10 increase of 10 $\mu g/m^3$ by fixed and random effects models. City specific estimates obtained by GAM via backfitting with default convergence criteria of Splus 2000, GAM via backfitting with stringent convergence criteria, GAM via direct method in R Software, GLM with natural cubic spline.

Abbreviations used in the text

MISA: Meta-analysis of Italian studies on Short-term effects of Air pollution
GAM: Generalized Additive Models
GAM-R: Generalized Additive Models fitted using the direct method
GLM+NS: Generalized Linear Models with natural cubic spline(s)
GAM-S: Generalized Additive Models fitted using the backfitting algorithm
ICD 9: International Classification of Diseases, Ninth Revision
IRLS: Iterative Re-weighted Least Squares