

Data Analytic Procedures for Monitoring Specific Pollutants in Epidemiological Studies

Richard L. Smith*

December 31, 2002

Abstract

This note describes some statistical procedures relevant to studying the effect of a specific pollutant, such as emissions from diesel fuel, in epidemiological studies. After some initial remarks about the formulation of the problem in statistical terms, we consider two methodologies in detail. The first concerns exposure modeling via spatio-temporal interpolation of a pollutant based on a limited network of monitors. The second is about the assessment of health effects when data on multiple constituents of air pollution data are available; an example is given based on data from Phoenix, where data on $PM_{2.5}$ broken into individual elements were available. However the discussion also highlights the inadequacy of many existing data bases for this kind of study.

1 Introduction: What is the Question?

Time series studies of air pollution and health, such as the NMMAPS study (Samet *et al.* 2000a, 2000b), are designed to answer questions about the short-term impact (e.g. excess numbers of deaths per year) of an atmospheric pollutant, such as PM_{10} or $PM_{2.5}$, on some measure of human health. Similarly, prospective studies (Krewski *et al.* (2000)), which attempt to measure long-term as well as short-term effects, are ultimately designed to assess the total impact on the human population of some aspect of air pollution health effects. It seems to me that the ultimate objective of research in diesel fuel emissions, or any other specific component of air pollution, is to answer similarly broad-brush questions of total impact on human health, but restricted to that component of the total pollution. I emphasize this point because the effective design of an epidemiological study depends on a clear articulation of its objectives. By focussing on these objectives and the kind of statistical analysis needed to achieve them, we gain insight into the data that need to be collected, and the sample sizes required to answer the questions of interest with a satisfactory degree of precision.

In the case of diesel pollution, a key aspect seems to be that it is not possible to measure directly the emissions of diesel fuel: realistic methods of measurement depend on proxies, such as the concentrations of certain chemical markers in the ambient air, from which we can infer the diesel emissions with some measurement error.

The effect of measurement error in epidemiological studies of air pollution is one of the known difficulties of this area of research. It is widely acknowledged that pollution measured by ambient air

*Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260; rls@email.unc.edu. Presented at the HEI Workshop to Improve Estimates of Diesel and Other Emissions in Epidemiologic Studies, Baltimore, December 4–5 2002. This paper is based in part on research supported by EPA Cooperative Agreement CR-827737-01-0 and by NSF grants DMS-9971980 and DMS-0084375.

monitors differs substantially from that measured by personal monitors, and some limited attempts have been made to account quantitatively for the effect this has on regression relationships (e.g., Dominici *et al.* (2000)). If proxy variables are used for diesel emissions, then the question of how to deal with measurement error is central to the analysis. A typical analysis may use the following variables:

- X : Measured air pollution variables (e.g. criteria pollutants, PM_{2.5}, specific chemical markers)
- Y : Response, e.g. mortality, hospital admissions due to asthma,
- Z : True variable of interest, e.g. diesel emissions

We would like to fit a model of the form

$$f(Y) = \beta Z + \text{other fixed effects} + \text{random error} \quad (1)$$

where $f(Y)$ is some function of the response Y , and the other fixed effects include confounders such as meteorology, terms to represent time trends and seasonal effects, etc. The measurement error problem arises because in epidemiological studies, we are not able to measure Z directly but must use X as a set of proxy variables.

However, we also have available training data (e.g. from supersites) that can allow us to study in detail the joint distributions of X and Z . With this information, the probability distributions we need to specify are

1. $p(X|Z)$, the distribution of X given Z — this may come from detailed chemical analyses
2. $p(Z)$, the prior distribution of Z — this is needed for Bayesian analyses

Using Bayes' Theorem, the last two distributions may be combined to produce $p(Z|X)$, the conditional distribution of Z given X . This in turn may be combined with (1) to produce a posterior distribution of β .

This approach, being explicitly Bayesian and taking into account the uncertainty of our knowledge about Z , is different from the conventional non-Bayesian approach to measurement error in regression. While there is much room for discussion about details, I believe that an approach of this broad structure is needed to take account of the different sources of uncertainty.

Two questions are:

1. What should X be? — In other words, what are the best variables to use as proxies for diesel fuel? Possibly that question is already answered by other presentations at this workshop, but a balance may need to be struck between using a good chemical marker and one that is easy and cheap to measure on the extensive scale needed for a good epidemiological study.
2. What kind of information should be use to assess $p(Z)$? A crude measure might be based on traffic patterns, but perhaps something more sophisticated is needed.

2 Spatial Statistics

One way of reducing measurement error would be if we had better methods of spatial interpolation between monitors. Patrick Kinney, at this workshop, has highlighted some of the difficulties with diesel particulate matter (DPM) in comparison with the more extensively studied fine particulate matter (PM_{2.5}). The spatial-temporal variability of PM_{2.5} is dominated by the temporal component — at any fixed point in time, the field tends to be fairly homogeneous spatially, and this makes it

relatively easy to interpolate based on a limited set monitors. With DPM, it appears, the situation is reversed — the variability is dominated by the spatial component. Limited studies with personal monitors have highlighted the difficulty of interpolating DPM.

There is no ready answer to this problem, but in a recent study, Smith *et al.* (2003) have proposed a new approach to interpolating PM_{2.5}, and this method may provide some clues as to how to deal with the harder problem of DPM.

This particular analysis was based on one year’s PM_{2.5} data from 74 monitors in three southern states (NC, SC, GA). Weekly averages were computed based on data available at each monitor. Also available were the latitude-longitude coordinates of the monitor, and a “land use” variable (agricultural, commercial, forest, industrial or residential). Preliminary analysis of the data suggested showed a strong weekly variability common to all sites, but with upward or downward shifts depending on the characteristics of the sites (Fig. 1).

A model was fitted of the form

$$y_{st} = \omega_t + \psi_s + \ell_s + \eta_{st} \quad (2)$$

where y_{st} is the square root of PM_{2.5} for week t at site s (the square root transformation was found to improve the fit of the model), ω_t is a week effect, ψ_s represents a fixed spatial component of the variability, ℓ_s is a land-use effect (one value for each of the five possible land uses) and η_{st} is a random component. The “spatial interpolation” aspect of this only applies to η_{st} , the other components being regarded as fixed. The parameters ω_t and ℓ_s were treated as fixed effects, similar to an analysis of variance, while ψ_s was modeled as a smooth function of s through a thin-plate spline representation.

As an example of the output of this analysis, a map of overall mean PM_{2.5} was produced, and compared with the (much rougher) map produced by simply averaging and interpolating the raw values (Fig. 2). Another feature of the spatial analysis, not possible with non-statistical interpolation methods, is the possibility of creating a standard error map (Fig. 2(b)). This would be needed to pursue the kind of analysis discussed in Section 1: for example, if Z represents the PM_{2.5} at a specific location, and X represents a set of PM_{2.5} measurements from monitors, then the conditional distribution $p(Z|X)$ involves some measure of variability and not just the conditional mean of Z given X .

It is conjectural whether a similar technique could be applied to the interpolation of DPM. However, some features of the method should be noted. It has been pointed out that for DPM, although the spatial variability is typically much higher than for PM_{2.5}, the *pattern* of spatial variability seems to be fairly constant in time (for example, because of consistent traffic patterns). This suggests that a model of the form (2), in which there is a fixed term ψ_s to represent the consistent spatial pattern of observations, would do much better than simple geostatistical interpolation of the spatial field at each time point. The method is also well adapted to missing data, since typically (even with PM_{2.5} data) observations are not available from all monitors at all time points.

3 PCA, Empirical Bayes and Related Methods

We now turn to a different aspect of the analysis, using examples based on Kim *et al.* (1999), Smith *et al.* (2000), in which PM₁₀ and PM_{2.5} data were available from the Phoenix site (1995–1997), along with data on the contributions of individual chemical elements to the PM_{2.5} total. Daily mortality and meteorological data were also available. The challenge in this study was how to incorporate a large number of possible covariates into a regression analysis.

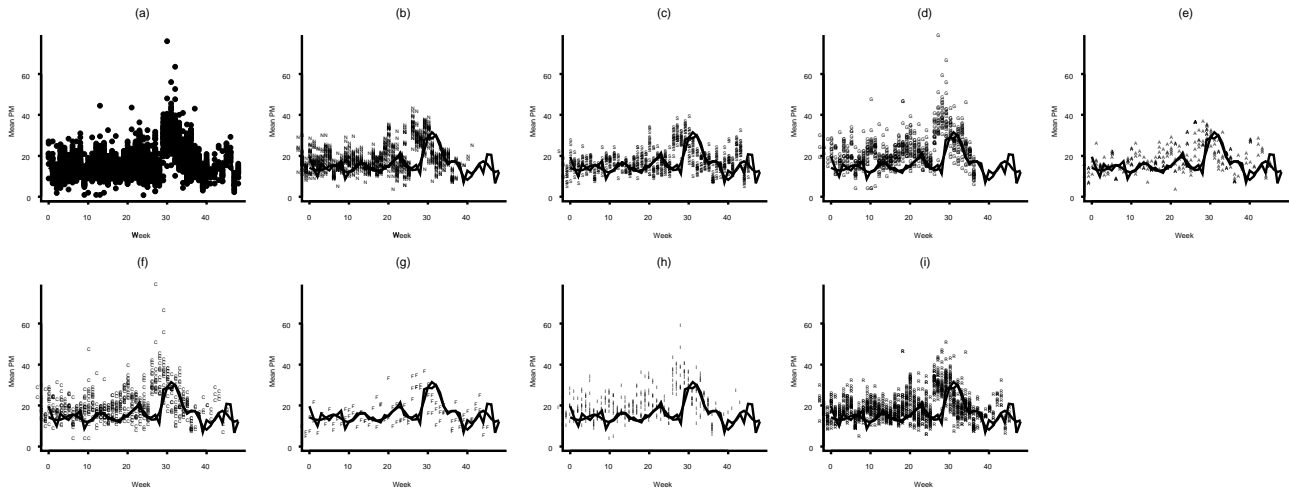


Figure 1. Comparison of fitted trend and raw data for subpopulations. (a) All data combined, (b) NC, (c) SC, (d) GA, (e) Agricultural sites, (f) Commercial, (g) Forest, (h) Industrial, (i) Residential. A single common weekly trend is shown (same on all plots), with superimposed data points from the individual stations in each subpopulation.

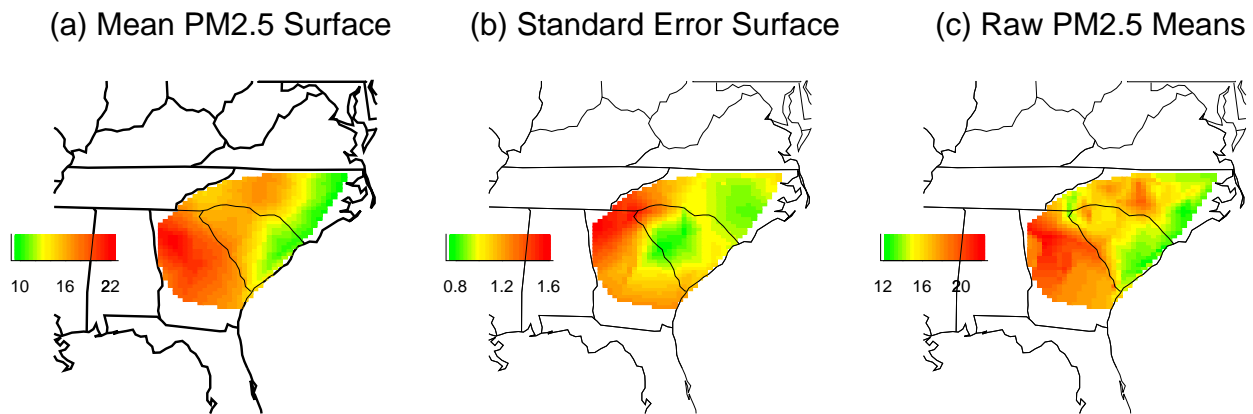


Figure 2. Reconstructed surface for overall annual mean $PM_{2.5}$. (a) Mean surface. (b) Standard error. (c) Plot of raw data with linear interpolation on a triangulation (S-PLUS “interp” routine).

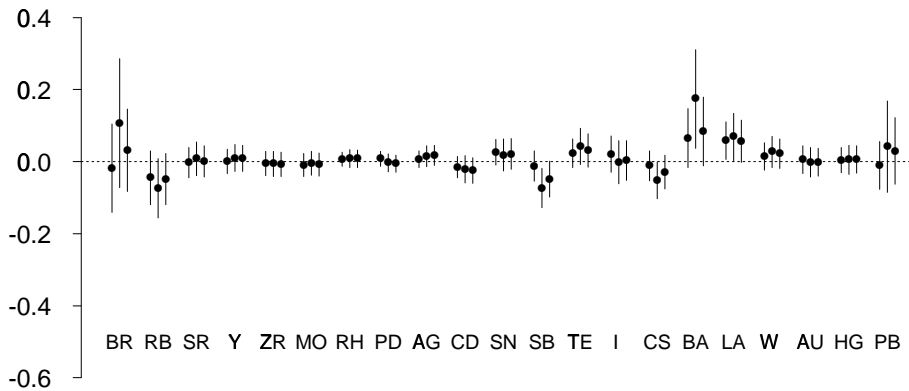
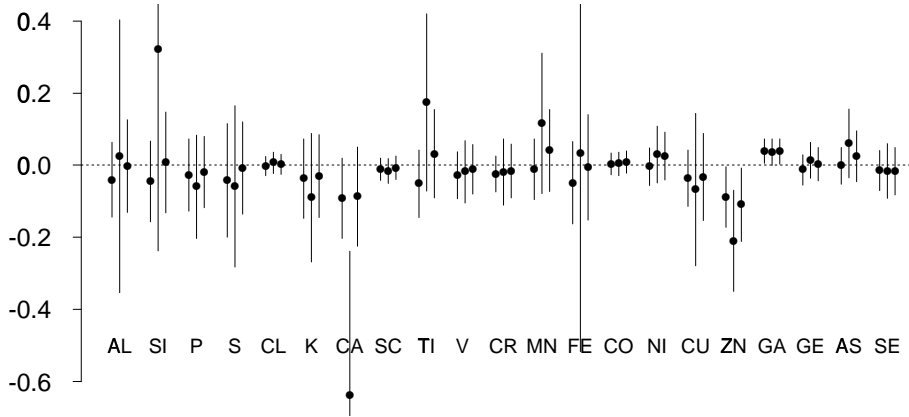


Figure 3. Estimates and 95% confidence bands obtained for each of 42 elements. Three estimates for each element: (left) ordinary least squares (OLS) estimates when elements are added one at a time; (middle) OLS estimate when elements are added all together; (right) Ridge regression estimates.

One conclusion that contradicted conventional wisdom on air pollution health effects was that for this data set, it appeared that the effect due to coarse particles (difference between PM_{10} and $PM_{2.5}$) was stronger than that due to $PM_{2.5}$, though the $PM_{2.5}$ effect was significant above a threshold in the region of $20\text{--}25 \mu\text{g}/\text{m}^3$ (Smith *et al.* (2000)).

In further analysis, we considered regressions in which the covariates of a pollution-mortality relationship included the measured concentrations (for 300 days of data) of 42 chemical elements. In Fig. 3, the regression coefficients and 95% confidence limits are plotted, computed three different ways: (a) putting in the 42 elements one at a time (the left-hand member of each group of three estimates), (b) including all 42 elements at once (middle), (c) including all 42 elements but reducing the variability of the estimates by ridge regression (right). Method (a) produces reasonable-looking estimates, but the difficulty is that we do not believe that any single element is responsible for

the mortality effect: the intention behind the analysis is to identify groups of elements, or specific combinations of elements, that might act as some kind of marker. The analysis (b) includes all the elements, but many of the individual estimates have very large standard errors as indicated by the length of the vertical bars. Ridge regression, method (c), is a method described in many texts on regression analysis, that aims to reduce the variability of the individual estimates at the cost of introducing a small amount of bias. In this case, the ridge estimates are quite similar to the single-variable estimates in (a), and apparently superior to those in (b).

Ridge regression is just one of a number of so-called shrinkage methods that have been designed to improve the performance of regression estimates when there are very many regressors or a high degree of multicollinearity. Other methods include the lasso method (Tibshirani 1996), various methods developed in the context of calibration (Brown 1993), and empirical Bayes methods (Carlin and Louis 1996). As an example of the latter, Kim *et al.* (1999) adapted the triple-goal estimates of Shen and Louis (1998) to this regression setting.

For this particular data set, none of the methods was successful in identifying any particular element or group of elements that was especially strongly associated with pollution health effects, a conclusion that was scarcely surprising given the overall weak association between PM_{2.5} and mortality in this data set.

For an alternative analysis, Smith *et al.* (2000) looked at seasonal effects with a view to examining whether seasonal patterns in the PM-mortality effect could in any way be associated with seasonal pattern in the pollutants. Table 1 shows seasonal coefficient estimates for the mortality effect of coarse PM (the corresponding results for fine PM showed no evidence of any seasonal effect). The effects are strongest in the spring and summer.

Season	Mean coarse PM ($\mu\text{g}/\text{m}^3$)	Regression Coefficient	Standard Error	t statistic	p value
Winter	33.6	0.0036	0.0023	1.5	0.13
Spring	28.9	0.0139	0.0026	5.3	0.0001
Summer	31.6	0.0063	0.0026	2.4	0.018
Fall	39.3	0.0023	0.0022	1.0	0.3

Table 1. Effect of coarse PM on mortality, measured by season.

A subset of the single-element data (Al, Si, S, Cl, K, Ca, Ti, Mn, Fe, Cu, Zn, Pb) was further classified using a principal components analysis which showed that the crustal elements (Al, Si, K, Ca, Ti, Mn, Fe) explained 55% of the variation of coarse PM, the anthropogenic elements (Fe, Cu, Zn, Pb) explained 30%, and the elements of marine origin (Cl in NaCl; Na was not measured) explained 5%. Table 2 shows a breakdown by season of the means of three principal components corresponding to each of these groups.

Season	Crustal	Anthropogenic	Marine
Winter	-.144	.503	-.589
Spring	-.278	-.323	.073
Summer	.004	-.483	.41
Fall	.245	.222	.03

Table 2. Breakdown by season of mean level of each of the three principal groups of elements (standardized to overall mean 0 for each component)

It appears that the seasons in which the anthropogenic effect is low are also the ones with highest effect due to PM. This is, of course, contrary to the common understanding that the effect

of naturally occurring particulate matter in the atmosphere is much less important than the effect of industrial and other anthropogenic sources. There may be some natural explanation of this phenomenon in the case of Phoenix, or the whole result may just be an artifact of the relatively small scale of the study, but either way, it suggests questions for further study.

4 Summary and Conclusions

Any meaningful attempt to integrate measures of diesel emissions into large-scale epidemiological studies will have to deal up front with the issue of bias induced by measurement error. Section 1 of this discussion has outlined a possible conceptual approach and some of the main issues that need to be addressed in order to apply it.

One aspect of measurement error bias comes from the fact that any measure of air pollution is only available at ambient monitoring sites and these are not typically in good agreement with measurements derived from personal exposure monitors. This is an issue with studies based on $PM_{2.5}$, but is expected to be a much more serious issue with measures of diesel fuel emissions because of the much greater spatial variability of the latter compared with $PM_{2.5}$. In Section 2, I have outlined some current thinking about the spatial interpolation of $PM_{2.5}$, which improve substantially on simple geostatistical methods, and may be applicable to other pollutants as well.

Section 3 of this discussion has highlighted some of the issues involved in regression analysis of pollution-mortality relationships when there are a large number of pollution-related covariates, represented here by the individual elements in chemical analyses of $PM_{2.5}$. Ordinary least-squares regression estimates do not perform well when there are many covariates and/or a high degree of multicollinearity. Alternative methods such as ridge regression or empirical Bayes analysis are available, and may be expected to improve the simultaneous estimation of many regression coefficients. These methods will not help with very limited data sets. The biggest obstacle to this whole field of research, as I see it, is designing and collecting data for epidemiological studies on a sufficiently large spatial-temporal scale to make meaningful health effects estimates.

5 References

- Brown, P.J. (1993), *Measurement, Regression and Calibration*. Oxford University Press.
- Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.
- Dominici, F., Zeger, S.L. and Samet, J.M. (2000), A measurement error model for time series studies of air pollution and mortality. *Biostatistics* **1**, 157–175.
- Kim, Y., Spitzner, D., Zhang, Z., Smith, R.L. and Fuentes, M. (1999), Accounting for multiple pollutants in pollution-mortality studies. *Proceedings of the ASA Biometrics Section*, 1–10.
- Krewski, D., Burnett, R.T., Goldberg, M.S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M. and White, W.H. (2000), *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality*. A Special Report of the Institute's Particulate Epidemiology Reanalysis Project. Health Effects Institute, Cambridge, MA.
- Samet, J.M., Dominici, F., Zeger, S.L., Schwartz, J. and Dockery, D.W. (2000a), National morbidity, mortality and air pollution study. Part I: methods and methodologic issues. Research Report 94, Health Effects Institute, Cambridge, MA.
- Samet, J.M., Zeger, S.L., Dominici, F., Curriero, F., Coursac, I., Dockery, D.W., Schwartz, J. and Zanobetti, A. (2000b), National morbidity, mortality and air pollution study. Part II:

morbidity, mortality and air pollution in the United States. Research Report 94, Health Effects Institute, Cambridge, MA.

Shen, W. and Louis, T.A. (1998), Triple-goal estimates in two-stage hierarchical models. *J.R. Statist. Soc. B* **60**, 455–471.

Smith, R.L., Kim, Y., Fuentes, M. and Spitzner, D. (2000), Threshold dependence of mortality effects for fine and coarse particles in Phoenix, Arizona. *Journal of the Air and Waste Management Association* **50**, 1367–1379.

Smith, R.L., Kolenikov, S. and Cox, L.H. (2003), Spatio-temporal modeling of PM_{2.5} data with missing values. Tentatively accepted for *Journal of Geophysical Research–Atmosphere*.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso. *J.R. Statist. Soc. B* **58**, 267–288.