NISS

The Effect of Airborne Particulate Matter on Daily Death Counts

P. Styer, N. McMillan, F. Gao, J. Davis
and J. Sacks

Technical Report Number 28
October, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Although the information in this document has been funded wholly or in part by the United States Environmental Protection Agency under assistance agreement #CR819638-01-0 to the National Institute of Statistical Sciences, it may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

THE EFFECT OF AIRBORNE PARTICULATE MATTER ON DAILY DEATH COUNTS

Patricia Styer[1], Nancy McMillan[1], Feng Gao[1], Jerry Davis[2] and Jerome Sacks[3]

National Institute of Statistical Sciences

October 12, 1994

KEY WORDS: Poisson regression, PM_{10}

1 Introduction

In order to determine if airborne particulates contribute to excess mortality, researchers have adopted multiple regression techniques to measure the effects of particulates on daily death counts (1,2). Other factors, such as extreme temperatures, can affect mortality, and the regression techniques are used in an effort to adjust for these other known influences.[4] The regression coefficient corresponding to a measure of particulate level is then interpreted as the effect of particulate pollution on mortality, accounting for stress from the other influences. If this coefficient is a statistically significant positive number, the conclusion is that mortality increases with increasing levels of particulates. This association is then elevated to a causal interpretation: particulates cause death, and researchers estimate that soot at levels well below the maximum set by federal law "kills up to 60,000 in U.S. each year" (3,4), and similar calculations "put the annual toll in England and Wales at 10,000." (5)

Studies vary on the particulate measures that are used and on locations analyzed. In the analyses presented here, we use PM_{10}, which specifies "particulate matter with an aerodynamic diameter less than or equal to a nominal 10 micrometers" (6). The current U.S. EPA standard is based on this measure. The locations we analyze, Cook County, Illinois and Salt Lake County, Utah, both have

[1]National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0 and National Science Foundation Grant DMS-9208758.

[2]National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162, and Department of Marine, Earth and Atmospheric Sciences, N.C. State University, Raleigh, NC 27695-8208. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

[3]National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162. Research supported in part by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

[4]Though many factors could be involved, research has generally limited attention to meteorological sources such as temperature and humidity. In some cases, other air pollution measures such as sulphur dioxide and ozone are included.

relatively long records of PM_{10} monitoring. The monitoring data is discussed in more detail in the *Data* section below.

The data used in the analyses (meteorological conditions, particulate levels, death counts) are observational — that is, data that are measured and recorded without control or intervention by researchers. Deducing causal relationships from observational data is perilous. A practical approach described by Mosteller and Tukey (7) involves considerations beyond regression analysis. In particular, consideration should be given to whether the association between particulate levels and mortality is consistent across "settings", whether there are plausible common causes for elevated particulate levels and mortality, and whether the derived models reflect reasonable physical relationships.

There is a high degree of association of PM_{10} with meteorology, and a high degree of association of mortality with weather. For example, in the summer in Cook County the correlation coefficient between the daily average of PM_{10} and daily mean temperature is 0.52 and the correlation between daily elderly (age 65 or greater) mortality and mean temperature is 0.25. The confounding effects of weather as a partial cause of both particulate levels and mortality may not be controllable by standard regression methods; the appearance of an effect for particulates, i.e., a positive coefficient for the PM_{10} term, may, as a result, be spurious[5](see Appendix B).

The results for Cook County and Salt Lake County (Sections 4 and 5 below) show that the appearance and size of a PM_{10} effect is quite sensitive to model specification. In particular, there is a pronounced seasonal effect. In Cook County, PM_{10} only appears as a factor in the spring and the fall, but not the winter nor the spring. A detailed semi-parametric analysis indicates that only the months of May and September exhibit a particulate effect. The models used are discussed in Section 3. An alternative analysis, using generalized additive models, supports these results (8). In Salt Lake County there is a similarly isolated effect of particulates limited to the month of June and no evidence of a PM_{10} effect in any season. Consistency of association of particulates with mortality is not found here.

Several studies carried on at various locations in the United States have reported small yearly increases in mortality resulting from increases in particulates. In our Cook County analyses the effect of PM_{10} in the spring and fall (or in May and September) induces a similar positive yearly increase in mortality from increases in particulates, but the increase is from 1/2 to 1/3 the size usually reported in other studies depending on the analyses performed; a similar factor is reported in (8). In Salt Lake County the size of the yearly effect is far smaller and statistically insignificant.

[5]We have not addressed errors-in-variables issues which can also be a cause for spurious relationships. The errors-in-variables concern arises from the differences between measured PM_{10} and the actual PM_{10} exposure experienced by the population; there are similar concerns for the meteorological variables.

What remains unexplained is why, in Cook County, effects should appear in the spring but not in the summer, the fall but not the winter. Neither is it clear why the effect of particulates on mortality should not appear in any season in Salt Lake County.

The appearance of a PM_{10} effect in May and September in Cook County led to the speculation that pollen may be implicated, but no such evidence was found using pollen data monitored in the city of Chicago, the major population component of Cook County. Other analyses carried out for the fall season in Cook County on different subgroups of the population produced no definitive differences among subgroups (see Table 7 below).

The inconsistency of the regression analyses, the unresolved status of plausible common causes of particulate levels and mortality, the confounding effects of weather, and the unavailability of plausible biophysical mechanisms to explain the empirical analyses prevent concluding that there is an effect between "today's" mortality and "yesterday's" particulates. The question appears to us to be unresolved.

2 Data

The data used for the statistical studies have three main components: mortality counts, particulate levels, and meteorology. The sources of the data are described in this section along with some summary statistics.

Mortality Data

Daily death counts for the period 1985 through 1990 come from death certificate records for Cook and Salt Lake County residents, collected by the National Center for Health Statistics, and made available to us by John Creason, EPA. While mortality data are available for longer periods, PM_{10} data are unavailable before 1985. Each death record contains a cause of death code and some basic demographic information. In compiling daily death counts, we excluded all deaths from accidental causes, as well as deaths of county residents occurring in other locations. We refer to the remaining number of deaths as total deaths. The main analyses were performed with total deaths among the population aged 65 or greater (elderly deaths). We carried out additional analyses for total deaths, unrestricted by age, for deaths classified by specific causes, and for selected population subgroups such as elderly Blacks and elderly males. We classified the disease-specific causes of death by the International Classification of Diseases (ICD) codes that appear on the mortality records. We adopted the classification scheme detailed in Fairley (1990) (9), extracting cancer deaths (ICD categories 140-209), circulatory deaths (ICD categories 390-459), and respiratory deaths (ICD categories 11, 35, 472-519,710.0, 710.2,710.4).

In Cook County, there was an average of 117 nonaccidental deaths per day for all ages. Among

Table 1: Mean Daily Mortality for Nonaccidental Causes of Death. Total mortality indicates the mean number of daily deaths of county residents of all ages, excluding accidental deaths, homicides, and suicides. Elderly mortality indicates the subset of these deaths among county residents aged 65 and older. Circulatory, cancer, and respiratory deaths are classified by the primary cause of death code listed on residents' death certificates.

	Cook County					Salt Lake County	
	Elderly	Total	Circulatory	Cancer	Respiratory	Elderly	Total
Winter	90.4	126.7	62	29	12	7.4	10.2
Spring	82.3	116.7	56	28	10	6.8	9.2
Summer	77.0	110.6	53	28	9	6.3	8.5
Fall	81.5	115.6	55	29	10	6.6	8.9

residents aged 65 and over, there was an average of 83 deaths per day. Death counts vary by time of year, with higher numbers in winter and fewer deaths in summer. In Salt Lake County, there was an average of 9 nonaccidental deaths for all ages and 7 nonaccidental deaths for residents 65 and over. As in Cook County, there are slightly more deaths in the winter. Table 1 displays some summary statistics for both Cook County and Salt Lake County mortality.

Particulate Data

In current monitoring efforts, particulates are measured throughout the United States. There are both 24-hour and annual ambient air quality standards for particulate matter (6). In the first case, the standard is attained when the "expected number of days per calendar year with a 24-hour average concentration above 150 μgm^{-3} is equal to or less than one." In the second case, the standard is attained when the "expected annual arithmetic mean concentration is less than or equal to 50 μgm^{-3}." To comply with these standards, it is sufficient to collect samples from each monitoring site only once every six days, though there are a few locations with monitors that operate on a daily basis. For Cook County, the particulate data comes from a network of PM_{10} monitors reported in the EPA Aerometric Information Retrieval System (AIRS) for the period 1985 through 1990. During this time, there were 20 separate monitors in operation, though several monitors were operated for only a brief period of time. The Cook County network includes one daily station where PM_{10} samples are collected on a daily basis. The remaining stations collected samples every sixth day. The daily station observations are frequently missing, with 69 percent of the values recorded once the monitoring station began operation in April 1985. To fill in some of the missing values, we used the daily means of all available monitoring data as the basis for constructing our measures of PM_{10}. With all available data, there are observations for 75 percent of the days after April 1, 1985.

Table 2: Summary of PM_{10} Values. Statistics listing the minimum, 25^{th}, median, 75^{th}, and maximum PM_{10} values for daily stations in Cook and Salt Lake Counties, and the corresponding network averages for all available monitoring data. The number of days with observations over 150 $\mu g m^{-3}$ is listed in the final column.

	Min	25^{th} P	Median	75^{th} P	Max	# over 150
	(a) Cook County					
Daily Station	3	28	38	51	365	3
Network Mean	4	27	37	50	365	2
	(b) Salt Lake County					
Daily Station No. 12	9	33	48	67	194	13
Daily Station No. 1001	4	18	26	38	487	10
Network Mean	6	24	35	50	320	14

Since many of the 20 monitoring stations were in operation for a short period, there is a maximum of 12 observations on any single day. Furthermore, the six-day monitoring stations tend to operate on the same schedule, so many of the days have only the single daily monitor contributing to the daily mean.

In Cook County, PM_{10} levels are generally highest in the summer. Figure 1 (top) shows the distribution of daily PM_{10} values by month. It is also clear from this picture that mean levels are generally well below the EPA standard of 150 $\mu g m^{-3}$. In Table 2(a), the daily means from all available stations are compared with the values from the single daily monitoring station. These show close agreement, with three observations over the EPA standard for the daily station and two observations over 150 for the daily means.

In Salt Lake County, there were six PM_{10} monitors operating between June 1985 and December 1990. The monitoring network includes two daily stations. Figure 1 (bottom) shows the distribution of of the network means of daily PM_{10} values by month. The distribution of PM_{10} in Salt Lake County differs slightly from the distribution in Cook County. The overall levels are similar, though there are more days in Salt Lake County with PM_{10} levels over 150 $\mu g m^{-3}$. Unlike Cook County, there is an increase in overall levels in winter (December through February), though isolated occurrences of high particulate levels occur throughout the spring and summer. In Table 2(b), we present some summary statistics from the two daily stations and the daily means compiled from all available data from the six station network. Particulate levels at daily station number 12 are generally higher than at daily station number 1001, with the exception of the maximum value which is due to just one large observation. We performed additional analyses using data from station

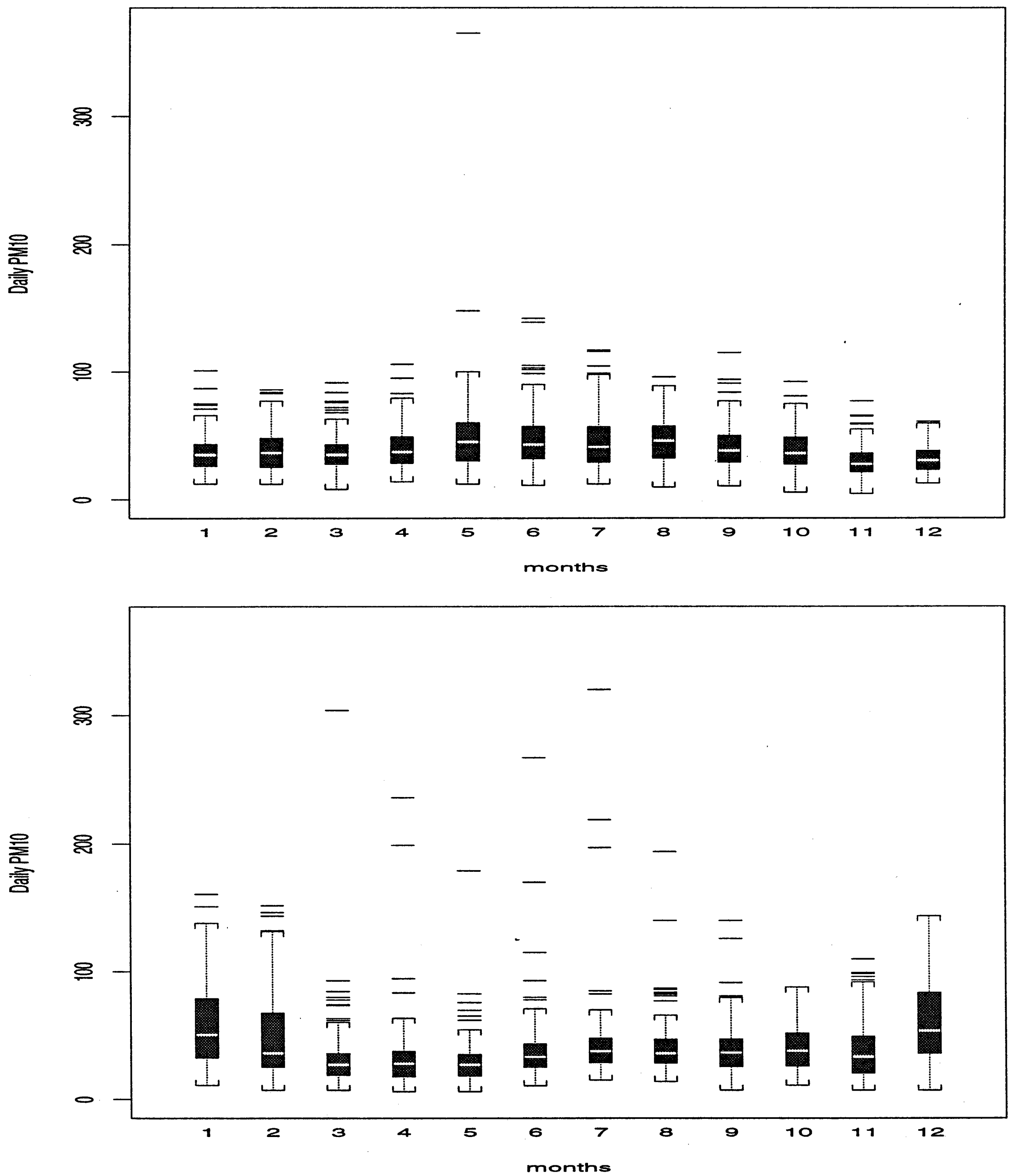

Figure 1: Daily PM_{10} by month for Cook County (top) and Salt Lake County (bottom). Box plots by month showing the distribution of the daily network averages of PM_{10} observations.

number 12 and station number 1001 individually to confirm that averaging these stations to create our PM_{10} covariate was not masking an effect.

Meteorological Data

The meteorological data used in this study are based on hourly surface observations taken at O'Hare International Airport (Cook County) and Salt Lake City International Airport (Salt Lake County). We extracted the data from the National Climatic Data Center's National Solar and Meteorological Surface Observation Network (1961-1990) data base, which contains hourly surface observations in addition to solar radiation data. Our primary analyses concentrated on three meteorological variables: temperature, specific humidity, and barometric pressure. We excluded other variables such as solar radiation, cloud cover, wind speed and wind direction. Cloud cover and wind variables were omitted to make our primary analyses more directly comparable with other research and because factors like wind may have more direct connection with PM_{10} than those included. For each variable we did include, we calculated the daily mean, based on hourly values. And, because weather may have a lagged effect on mortality, we also included the values of temperature, humidity and pressure from the two previous days. In other analyses, we considered the effect of wind chill in the winter and solar radiation and a heat index in the summer. These variables did not improve the prediction of mortality; the analyses are not included. The inclusion of wind speed and lagged wind speed in Cook County did not change the results from any of the models fit without wind.

Table 3 presents a summary of the meteorological data considered in various analyses. The data set containing the original hourly observations for these variables had only a few scattered missing values. We filled in the missing hourly observations by assigning the value from the previous hour, and then computed the daily mean values based on twenty-four observations.

Pollen Data

Pollen data was obtained from the pulmonary unit at Grant Hospital, Chicago, Illinois, courtesy of Judith Young. During the study period, pollen counts were recorded on a daily basis, except for weekends and holidays when cumulative samples were taken. To fill in daily pollen values from the cumulative values, we employed a model to predict daily pollen from local meteorological conditions and then distributed the total pollen amounts to the individual days based on this model. We considered pollen from trees, mold spores, and ragweed.

3 Model Formulation

Our primary analyses model daily death counts as a Poisson process. For most analyses, we split the data by three-month seasons and fit separate models within each season. Winter is taken as

Table 3: Description of Meteorological Variables.

Variable	Description
tmean	average daily temperature(C) from hourly observations
tlag1	average temperature from 1 day before
tlag2	average temperature from 2 days before
qmean	average daily specific humidity(gkg^{-1}) from hourly observations
qlag1	average specific humidity from 1 day before
qlag2	average specific humidity from 2 days before
pr	average daily station pressure (mb)from hourly observations
prlag1	average station pressure from 1 day before
prlag2	average station pressure from 2 days before

the the three months December, January, February; spring as March through May, *et cetera*. All season-by-season models include a yearly factor and a within-season trend (day) component. The specification of the trend component differs by season. For each season, we considered either a polynomial or a piecewise linear trend component and selected the shape that fit the data best. Although the covariates differ for different analyses, the basic model assumes that the daily death counts (Y) are Poisson-distributed with

$$\log(EY) = X\beta,$$

where X contains terms corresponding to a yearly factor, a within-seasonal trend component, relevant meteorological covariates, and a measure of particulates. The parameters of the model were fit by the iterative reweighted least squares algorithm in the statistical software package Splus (see, for example (10)).

To account for a possible lagged effect of PM_{10}, we focus primarily on pmmean3, the average of the current day's PM_{10} together with the values for the two preceding days. Missing values were ignored, so the mean values were based on any available observations. We compare the results from these models with models that incorporated each of the three single day values. We also did analyses using only the current day, pmmean2 (today and yesterday), and pmmean5. In essence, the results using pmmean3 are consistent with these other choices of PM_{10} measures, we only report a typical result from Cook County using pmmean5 in the fall.

Auxiliary to the Poisson regression models used is a semi-parametric model which, through its nonparametric character, avoids the necessity of specification of special forms while allowing a reasonably accurate selection of important covariates. The details of the model as it was used are

Table 4: Candidate Covariates for Poisson Regression Analyses Based on Results from Semi-parametric Modeling on Elderly Mortality. Active variables appearing in the month-by-month analyses using the semi-parametric model described in Appendix A. The variable *day* is the day of month (1-31), *pmmean3* is the simple average of the observed network daily means for the concurrent day and two previous days. The meteorological variables are described in Table 3.

Month	Cook County	Salt Lake County
Jan	day	pr,tmean,day
Feb	qmean,qlag1,pr,prlag2	day,prlag1,prlag2
Mar	day, [pmmean3]	
Apr		
May	tmean,qlag2, [pmmean3]	
Jun	pr	day, [pmmean3]
Jul	tmean	qlag1,prlag2, [pmmean3]
Aug	tlag1,pr,qlag1	
Sep	qlag2,pr,prlag2, [pmmean3]	day,tmean
Oct		pr
Nov	qlag2	
Dec	day,qlag1,prlag1	day,tlag1

given in Appendix A. This model is used in several ways. Primarily, it was used to select relevant meteorological covariates and to focus on potentially important interactions as well as nonlinear functional forms for some of the covariates. Models selected in this fashion tend to be more parsimonious than models selected with standard stepwise procedures, with no loss of explanatory power. In addition, a month-by-month analysis using the semi-parametric model revealed that PM_{10} was usually an inactive factor.

By focusing on the months where PM_{10} does appear active, a possible connection with pollen was suggested. Accordingly we obtained pollen data from the City of Chicago and introduced it in the analyses of May and September as well as in additional analyses covering August 15 to September 15, the ragweed season. In no case did any pollen variables appear. The observed PM_{10} effect in May and September is not explained by the presence of pollen particles.

With the focus on pmmean3, the meteorological covariates that were considered at the first stage include the current day's values as well as the preceding 2 days' values. The particular covariates included for a season's analysis incorporated those found in the monthly analyses by the semi-parametric model. Table 4 shows the set of active factors for each month in both Cook

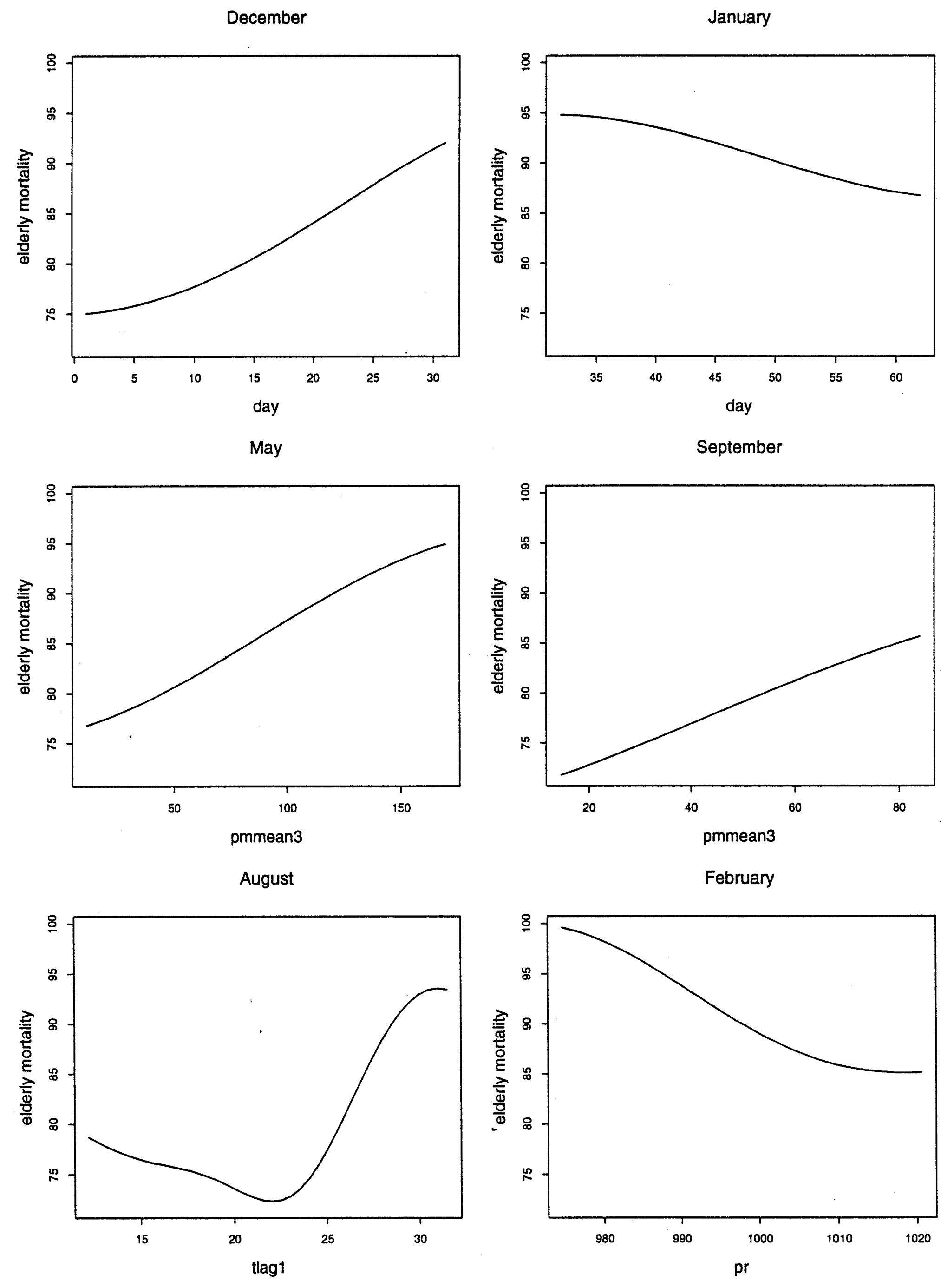

Figure 2: Some of the estimated effects for Cook County from the semi-parametric model. Predictions of elderly mortality holding all other variables constant at their median levels. The top two plots show the day-of-month effect for December and January, highlighting the peak in the number of deaths around January 1; the middle plots show the relationship between PM_{10} and mortality; the bottom plots show the potentially nonlinear dependence of mortality on meteorology.

and Salt Lake County in the semi-parametric model. We considered each of these covariates as the candidate variables for inclusion in the Poisson regression models, along with the functional forms and interactions suggested by the fitted response surfaces from the semi-parametric model. To illustrate this use of the semi-parametric model, we include some plots of estimated effects[6] of pmmean3, temperature, pressure, and day-of-year for some selected months (Figure 2). These plots show the so-called Christmas effect on mortality, with a spike in the number of deaths around the beginning of Januray, the linear effect of PM_{10} in May and September, and the nonlinear effects of temperature and pressure. Using the combined list of covariates from the months comprising each season, we used a stepwise variable selection technique to obtain a model without any measure of PM_{10}. Typically, this led to two or three meteorological covariates selected for each season to predict daily mortality. As a final step, we include the measure of PM_{10} and examine the direction and size of the corresponding coefficient.

To illustrate the importance of considering a season-by-season analysis, we also present results from an analysis combining the full year of observations for both Cook County and Salt Lake County. In this analysis, we fit a yearly factor, a cubic time trend for each season, the meteorological covariates that were significant predictors of mortality in the season-by-season models, and seasonal interaction terms for selected meteorological covariates. We then compare the estimation of the PM_{10} effect from the models with and without PM_{10}-by-season interaction terms.

4 Empirical Evaluation in Cook County, IL

There are several sets of results for Cook County. We first present full year and season-by-season analyses using the Poisson regression model estimating daily death counts for individuals 65 and older (elderly mortality).[7] The linear predictors are detailed in Tables 5(a) and 6(a). As discussed in the previous section, the covariates other than the yearly factor and the PM_{10} variable were chosen using stepwise selection techniques based on the list of candidate covariates in Table 4. Other models and results for Cook County are summarized in Table 7.

In our full year analysis of Cook County, we conclude that it is necessary to estimate a separate PM_{10} effect for each season. Since the effect of meteorology differs by season, for example, increasing temperature acts as a stress factor in summer but decreasing temperature creates stress in winter, we began by considering models for the full year which permitted separate estimates of the effect of weather within each season. Our final full year model to predict elderly mortality from meteorology includes separate seasonal terms for the yearly factors, the day-of-year effect,

[6]These effects are computed by conditioning the remaining variables on their median values.

[7]Because daily death counts are high here, an ordinary (normal) regression model will give the same results.

Table 5: Full Year Poisson Regression Models – Elderly Mortality. The left-hand side shows the models fit to predict daily mortality, using shorthand notation where an asterisk (*) indicates the interaction terms are included and poly(variable, n) indicates a polynomial term for the given variable of order n. Specification of an interaction implies inclusion of all lower order terms. The right-hand side shows the estimated effects of the PM_{10} variable, along with estimated standard errors in parenthesis.

(a) Cook County

Linear Predictor log(EY)		PM_{10} Coeff
season*(year+poly(day,3)+poly(tlag1,2))+ q+qlag2+pr+poly(prlag1,2)+pmmean3		0.00054(0.00020)
season*(year+poly(day,3)+poly(tlag1,2)+pmmean3)+ q+qlag2+pr+poly(prlag1,2)	Winter	-0.00001(0.00047)
	Spring	0.00083(0.00034)
	Summer	-0.00028(0.00036)
	Fall	0.00195(0.00047)

(b) Salt Lake County

Linear Predictor log(EY)		PM_{10} Coeff
season*(year+poly(day,3)+pr+tlag1)+pmmean3		0.00014(0.00046)
season*(year+poly(day,3)+pr+tlag1+pmmean3)	Winter	0.00006(0.00067)
	Spring	-0.00159(0.00125)
	Summer	0.00112(0.00092)
	Fall	0.00043(0.00126)

Table 6: Seasonal Poisson Regression Models – Elderly Mortality. Models for mortality estimated separately within each season are listed on the left-hand side. Estimated coefficients and standard errors are shown on the right.

(a) Cook County		
	Linear Predictor log(EY)	PM_{10} Coeff
Winter	year+day+Jan+Feb+pr+poly(prlag1,2)+pmmean3	0.00024(0.00046)
Spring	year+qlag2+pmmean3	0.00088(0.00030)
Summer	year+poly(day,2)+poly(tlag1,2)+pr+prlag+pmmean3	-0.00024(0.00035)
Fall	year+qlag2+pmmean3	0.00138(0.00040)

(b) Salt Lake County		
	Linear Predictor log(EY)	PM_{10} Coeff
Winter	year+day+Jan+Feb+tlag1+pr+pmmean3	0.00007(0.00069)
Spring	year+pmmean3	-0.00096(0.00116)
Summer	year+poly(day,2)+pmmean3	0.00082(0.00091)
Fall	year+pr+pmmean3	-0.00006(0.00123)

and temperature lagged one day. This permits the estimation of separate coefficients within each season for these terms. Other covariates whose effect do not vary significantly by season for Cook County include specific humidity for the concurrent day, 2-day lagged specific humidity, and station pressure for the concurrent day and previous day. We added the three-day mean PM_{10} variable, and compared the results from fitting a single estimate for the entire year with fitting separate estimates by season. The estimate for the single PM_{10} effect is 0.00054 with a standard error of 0.00020. Hence, an increase of 10 μgm^{-3} of PM_{10} corresponds to approximately .54 percent more deaths, given constant levels of all other covariates. When the season-by-PM_{10} interaction term is added, the PM_{10} effect remains significant only in the spring and fall (Table 5(a)). The estimated effects for the winter and summer are essentially zero.[8] To compare the overall effect of PM_{10} from this model, we calculated the predicted increase in the number of deaths in each season if PM_{10} were increased by 10 units. We then took the overall effect of PM_{10} to be the average of seasonal effects weighted by the length of season. The overall predicted increase in mortality is .63 percent. A nonparametric analysis in (8) produces similar results. A similar calculation, based on independent analyses of each month using the semi-parametric model, produces a .41 percent increase.

A finer tuned season-by-season analysis is obtained by fitting a separate model for each season.

[8]The chi-square test for the difference in deviance caused by inclusion of separate seasonal estimates for PM_{10} supports this inclusion with a p-value of approximately 0.001.

Table 7: Summary of Regression Models, Cook County, Fall Season. The population subgroups for each analysis are listed in the left-hand column. The final models are indicated in the middle column, and the corresponding coefficients and standard errors for the PM_{10} variable are listed in the right-hand column.

Population	Linear Predictor log(EY)	PM_{10} Coefficient
Total Mortality	year+tmean+qlag2+poly(prlag2,3)+pmmean3	0.00080(0.00040)
Males 65+	year+poly(qlag2,2)+tmean+pmmean3	0.00159(0.00069)
Females 65+	year+qlag2+pmmean3	0.00087(0.00054)
Blacks 65+	year+poly(qlag2,3)+pmmean3	0.00166(0.00089)
Whites, Others 65+	year+qlag2+pmmean3	0.00134(0.00045)
Circulatory Deaths	year+poly(qlag2,2)+poly(prlag2,3)+pmmean3	0.00064(0.00052)
Respiratory Deaths	year*tlag2+poly(qmean,2)+pmmean3	0.00220(0.00125)
Cancer Deaths	year+poly(qmean,3)+poly(tlag2,2)+pmmean3	0.00162(0.00071)
Elderly Mortality	year+poly(qlag2,3)+pmmean5	0.00158(0.00047)
Wed., Thurs., Fri.	year+qlag2+pmmean3	0.00075(0.00061)

Here, we used the variables suggested by the semi-parametric models for the corresponding months to choose a parsimonious model predicting mortality from meteorology. The results for the separate seasonal analyses are presented in Table 6(a). The covariates included in the seasonal models vary significantly between seasons, suggesting that a separate model for each season may be more realistic than one full year model. The PM_{10} coefficients and standard errors, however, are similar to the full year analysis with the season-by-PM_{10} interaction terms. There is a significant effect in spring and fall, and no significant effect in the winter and summer.

The reported standard errors are calculated assuming independent observations. To check this assumption, we examined the autocorrelation structure of the standardized residuals for the full year analysis. We computed the first seven lagged autocorrelations and found no correlations greater than 0.03[9]. We conclude that there is no evidence of significant serial correlation. Other diagnostic plots of the residuals confirm that the modeling assumptions are reasonable.

To investigate the consistency of the PM_{10} effect for different populations, we modeled daily death counts from several subgroups within Cook County and for different measures of PM_{10}, like a 5-day mean (pmmean5) instead of a 3-day mean. Since the largest estimated PM_{10} effect for elderly mortality is in the fall, we restricted attention to this season. These analyses include total mortality (nonaccidental deaths, all ages), elderly males and females, elderly Blacks and non-Blacks, and total mortality classified by disease categories, including circulatory disease, respiratory disease, and cancer. For each group, we refit the semi-parametric model by month to obtain the list of candidate covariates for the Poisson regression analysis. Table 7 shows the results from the final models selected.

To address concern over potential weekday, weekend effects in both PM_{10} and mortality, we also investigated the effect of fitting the fall, elderly mortality model, detailed in Table 6(a), to subsets of the data determined by day of week. We first subsetted to weekdays Wednesday, Thursday, and Friday since for these days the pmmean3 covariate is unaffected by the decline in PM_{10} over the weekend. The resulting pmmean3 coefficient is given in Table 7; It is approximately half of the coefficient when all the data is used. We also analyzed each day of the week individually. Although all the pmmean3 coefficients were positive, only the coefficient based on the Sunday data was significant. The average of the seven daily coefficients was 0.00135, comparable to the coefficient from our original fall, elderly mortality analysis. Similar effects were observed in the spring. We interpret these results as inconclusive, neither supporting nor denying a weekday effect.

While there appear to be inconsistencies in Table 7 for example, a significant effect of PM_{10} on males but not on females, the difference of the two effects may be insignificant. Similarly, for

[9]These values are all less than the approximate critical value of $2/\sqrt{N} = 0.045$.

circulatory and cancer deaths. The greater effect for cancer deaths than for circulatory deaths is in contrast to the opposite numbers reported in (1) for Philadelphia. The lack of significance for Blacks is due to the greater standard error resulting from the smaller size of the black population in Cook County. The distinction between using pmmean5 rather than pmmean3 is to reduce the size of the effect somewhat (from .00195 to .00158) but it remains significant.

5 Empirical Evaluation in Salt Lake County, UT

The analyses for Salt Lake County were carried out in similar fashion to those carried out in Cook County. The semi-parametric model was used on transformed (square-root of) mortality in order to ameliorate the effect of non-normality and non-constant variances in the presence of small counts. The analyses proceeded as before from the variables in Table 4 to the models found in Table 6(b).

The semi-parametric model identified PM_{10} as active in June and July. An estimated effect plot for July indicated that the effect of PM_{10} in July was oscillatory (as in March in Cook County) rather than monotone as in June (or as in May and September in Cook County). See Figure 3. The Poisson regression analysis, however, did not find a PM_{10} effect in the summer. In fact, for the full-year and seasonal models, PM_{10} was never a significant predictor of elderly mortality in Salt Lake County.

For the full-year analysis, the single estimate of the PM_{10} effect is 0.00014 with a standard error of 0.00046 (Table 5(b)). The full-year model including the season-by-PM_{10} interaction term fails to indicate a significant PM_{10} effect in any single season. Additional analyses using pmmean3 calculated using only the data from station number 12 and station number 1001 individually yielded similar results. Furthermore, unlike Cook County, the chi-square test for the difference in residual deviance does not support the inclusion of a season-by-PM_{10} interaction term. Even after fine-tuning the analysis by fitting a separate meteorological model for each season, PM_{10} does not emerge as a significant predictor of mortality (Table 6(b)).

6 Summary

We analyzed data from Cook County, Illinois and Salt Lake County, Utah in order to assess the connections among mortality, particulates (PM_{10}) and weather. We found that season plays a strong role. We found inconsistent results: no effect of PM_{10} was found in Salt Lake County in any season; no effect was found in Cook County in winter and summer; small, positive PM_{10} effects were found in Cook County in the spring and fall, and perhaps more specifically, in the months of May and September. Reported effects of particulates on mortality are not confirmed by these analyses;

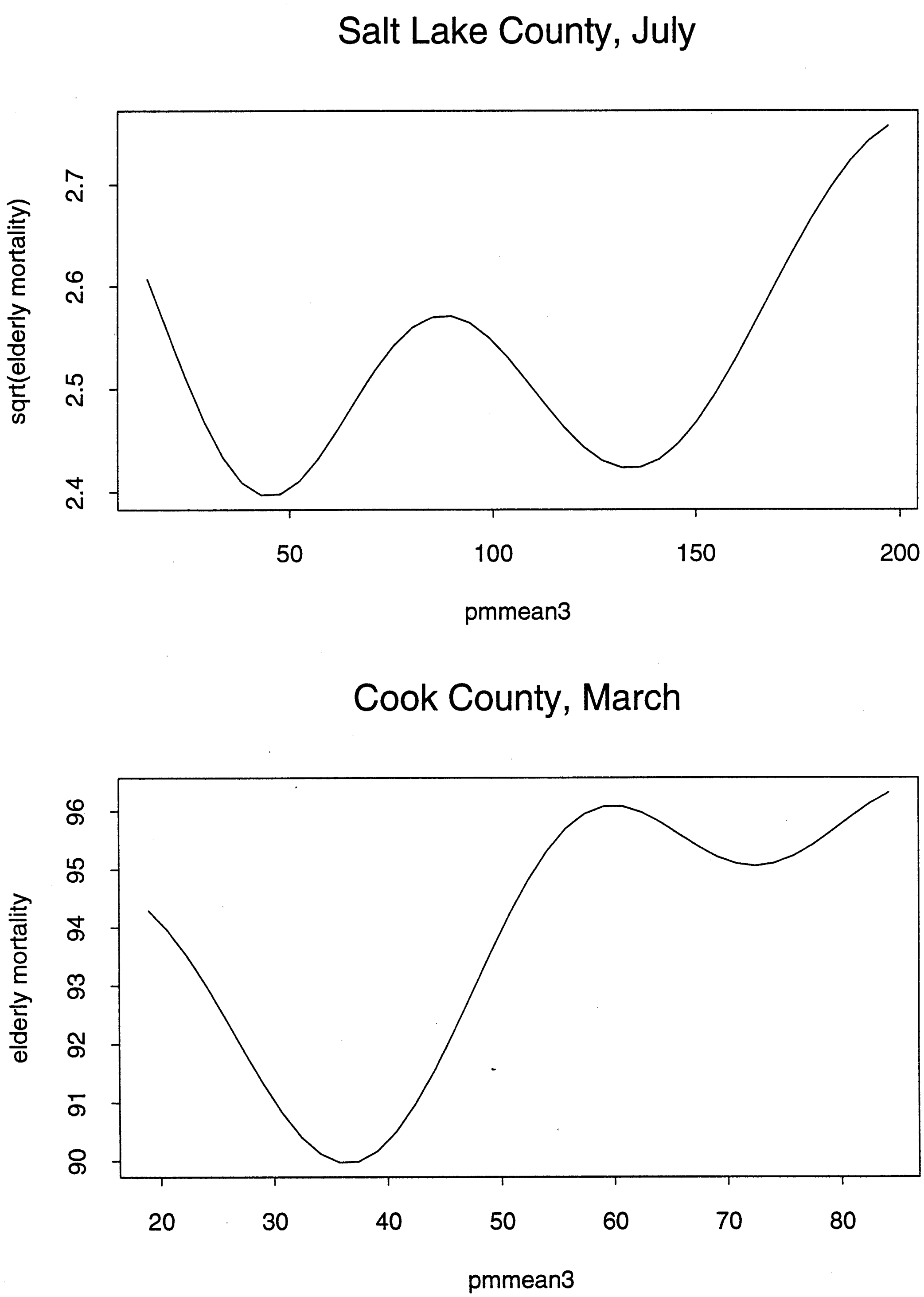

Figure 3: Estimated oscillatory effects from the semi-parametric model. In some months, PM_{10} appears as an active factor in the semi-parametric model, but the effect appears to be spurious.

whether increases in PM_{10} cause increases in mortality remains unresolved.

Appendix A. Semi-parametric Model

On day i, in year j, with meteorological condition met and PM_{10} value pm, where met is a 9-dimensional vector of the meteorological variables listed in Table 3 and pm could be any of the PM_{10} measures used in the analyses, let $x = (pm, met, i, j)$. The vector $x = (\zeta_1, \ldots, \zeta_{12})$ is 12-dimensional. The response $y(x)$ (mortality) is assumed to be a realization of a stochastic process, $Y(x)$:

$$Y(x) = \beta_j + Z(x) + \varepsilon_{ij}$$

where β_j are constants, $j = 1, 2, \ldots, 6$, $Z(x)$ is a zero mean Gaussian process with covariance function $\text{Cov}(Z(x), Z(x')) = \sigma_Z^2 R(x, x')$ to be specified later, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$. For more discussion on the use of this technique for modeling response surfaces, see (11), and the references cited there.

Assume, as in (11), that the covariance between $Z(x)$ and $Z(x')$ is

$$\sigma_Z^2 R(x, x') = \sigma_Z^2 \exp(-\sum_{k=1}^{12} \theta_k |\zeta_k - \zeta_k'|^{p_k})$$

where $x = (\zeta_1, \ldots, \zeta_{12})$, $x' = (\zeta_1', \ldots, \zeta_{12}')$, $\theta_k \geq 0$; $k = 1, \ldots, 11$, $\theta_{12} = 0$ and $1 \leq p_k \leq 2$; $k = 1, \ldots, 12$. θ_{12} corresponds to the year variable. This class of stationary processes provides us with a wide range of functions.

Given the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for q consecutive years starting from year 1 (1985) with n_j data points in year j and $n_1 + \cdots + n_q = n$ and, provided σ_Z, σ_ε and $R(\cdot, \cdot)$ are known, the best linear unbiased predictor (BLUP) $\hat{y}(x)$ at a new point x in year j can be written as

$$\hat{y}(x) = \hat{\beta}_j + \hat{Z}(x) = \hat{\beta}_j + r'(x) C^{-1}(y - F\hat{\beta})$$

where $y = (y_1, y_2, \ldots, y_n)$, $C = \text{Corr}(y) = (\sigma_Z^2/\sigma^2)R + (\sigma_\varepsilon^2/\sigma^2)I$, where $\sigma^2 = \sigma_Z^2 + \sigma_\varepsilon^2$, and $R = \{R(x_i, x_j), 1 \leq i \leq n; 1 \leq j \leq n\}$, the $n \times n$ matrix of correlations among Z's at the data points, $r(x) = (\sigma_Z^2/\sigma^2)[R(x_1, x), \ldots, R(x_n, x)]'$,

$$F = \begin{pmatrix} \vec{1}_{n_1 \times 1} & \vec{0} & \cdots & \vec{0} \\ \vec{0} & \vec{1}_{n_2 \times 1} & \cdots & \vec{0} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{0} & \vec{0} & \cdots & \vec{1}_{n_q \times 1} \end{pmatrix}_{n \times q},$$

and $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_q)' = (F'C^{-1}F)^{-1}F'C^{-1}y$, which is the usual generalized least-squares estimate of $\beta = (\beta_1, \ldots, \beta_q)'$.

The parameters σ_Z, σ_ε, θ's and p's are fit by maximum likelihood. Cross validation is used to assess variability of estimates. Values of p indicate smoothness of the response surface as a function of the corresponding variables. Larger values of θ usually indicate greater importance of the corresponding variables if the variables are on normalized scales. During the covariate selection procedure, those coefficients (θ's) which are zero are the factors not included; the others are selected.

Appendix B. The Problem of Confounding

To examine the confounding relationship between PM_{10} and the meteorological variables, a forward-selection ordinary least squares regression analysis was performed with log PM_{10} (the natural logarithm of today's PM_{10}) serving as the response variable and the meteorological variables serving as the covariates. The meteorological variables in the PM_{10} analysis were those included in the mortality analysis. The same seasonal structure was maintained for the PM_{10} analysis as for the mortality analyses.

Cook Co. As mentioned earlier, PM_{10} levels were highest in the spring and summer while fall and winter levels were depressed. R^2 values from the final models based on the forward-selection ordinary least squares regression analyses ranged from a low of 20 percent in the winter to a high of 50 percent in the summer. Thus the relationship was strongest during the season with the highest PM_{10} levels. With the exception of the 2-day lag temperature term (tlag2) in the fall, the regression coefficients for the various temperature terms were positive. Today's temperature (tmean) showed up in all seasons with the exception of summer, while the square of today's temperature (tmean) showed up in all seasons. All seasons except winter exhibited a strong rise in PM_{10} with increasing temperature. The coefficients on the specific humidity terms were negative. Yesterday's specific humidity (qlag1) was important in all seasons, while today's specific humidity (qmean) showed up in spring and fall. A quadratic term (qlag2^2) showed up in the summer. These main effect results are consistent in the sense that warmer drier conditions contribute to increased levels of particulate matter. Interaction plots generally indicated that at low temperatures PM_{10} levels increased with increasing specific humidity while the reverse was true at higher temperatures. Station pressure (2-day lagged variable, prlag2) showed up only in the fall and then with a positive sign.

Salt Lake Co. The amount of variation in PM_{10} explained by the meteorological covariates ranged from 29 percent in summer (a time of low PM_{10} levels) to a high of 58 percent in the winter (a time of high PM_{10} levels). In contrast to Cook Co., station pressure was a significant variable in all seasons in addition to temperature and specific humidity variables. Station pressure lagged one day (prlag1) was the first variable to enter the forward selection process in fall and winter where it added 19 and 42 percent to the R^2 value, respectively. The sign of the regression coefficients on the pressure terms were positive except during the summer. In spring and summer, temperature terms were the first to enter the forward selection process. The signs on the temperature terms varied with the season and within the season for different terms. Specific humidity terms entered for all season in a negative manner except for winter. It is unlikely that there was a direct relationship between pressure and PM_{10} levels; rather, station pressure appears to have served as a surrogate for those meteorological conditions which have a direct physical/chemical relationship with PM_{10}. Typically, in winter high pressure is accompanied by colder, dryer conditions. In spring and summer PM_{10} levels generally increase as temperature increases; in winter PM_{10} levels decrease as temperatures rise. In fall an initial decrease in PM_{10} levels as temperatures rise turns to an increase in PM_{10} levels as temperatures move above 7C. In winter, summer and fall PM_{10} levels initially increase with rising humidity levels and then begin to drop as humidity continues to rise. In spring PM_{10} levels decrease as humidity increases.

Results on fitting mortality to weather variables alone, without PM_{10}, indicated that temperature, humidity and pressure are all implicated (Tables 5 and 6); this was confirmed also by Smith's analyses (8).

References

1. Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am Rev Respir Dis 145:600–604(1992).

2. Pope CA III, Schwartz J, Ransom MR. Daily mortality and PM_{10} pollution in Utah Valley. Arch Environ Health 47:211–217(1992).

3. Slipp D. Bad things come in small particles. Wall Street Journal April 24, 1991.

4. Hilts PJ. Studies say soot kills up to 60,000 in U.S. each year. The New York Times July 19, 1993.

5. Brown W. Dying from too much dust. New Scientist March 12, 1994; 12–13.

6. US EPA. National Primary and Secondary Ambient Air Quality Standards. Federal Register 36:22384 (25 November 1971) and amendments.

7. Mosteller F, Tukey JW. Data Analysis and Regression: A Second Course in Statistics. Reading, Massachussetts: Addison-Wesley, 1977.

8. Smith RL. A Spline regression approach to the particulate mortality problem. Working paper (1994).

9. Fairley D. The relationship of daily mortality to suspended particulates in Santa Clara County, 1980-1986. Environ Health Persp 89:159–168(1990).

10. Chambers J, Hastie T. Statistical Models in S. New York: Chapman & Hall, 1993.

11. Sacks J, Welch WJ, Mitchell TJ, Wynn HP. Design and Analysis of Computer Experiments. Statistical Science 4:409–435(1989).