ABSTRACT

CHEN, ZHIDONG. An Analysis of the Effects of PM10 on Mortality in Pittsburgh, PA for the Period 1992 to 2000. (Under the direction of Roger H. von Haefen.)

My thesis examines the relationship between PM10 and non-accidental mortality by focusing on one location, Pittsburgh, PA for the period of 1992 to 2000. The standard Ordinary Least Squares (OLS) method and Generalized Linear Models (GLM) are the two methods that are mainly employed. Quantile regression techniques have also been used to further examine the effects of PM10 on different levels of the quantiles of mortality. The results indicate that PM10 is, in some cases, strongly related to excess mortality especially for elder people (older than 74); although, the results vary from year to year and season to season.
An Analysis of the Effects of PM10 on Mortality in Pittsburgh, PA for the Period 1992 to 2000

by
Zhidong Chen

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Economics

Raleigh, North Carolina

2013

APPROVED BY:

_______________________________
Roger H. von Haefen
Committee Chair

_______________________________
Steven Sexton

_______________________________
Donald Martin
DEDICATION

To my beloved father, Zhongqi Chen, mother, Li Yang and younger brother, Shufeng Chen.
BIOGRAPHY

Zhidong Chen was born in Luan, Anhui Province, China in 1989. She majored in Mathematics and Applied Mathematics and earned her Bachelor’s Degree from Southwestern University of Finance and Economics in Chengdu, China in 2008. With the support of her parents she came to the United States in the fall of 2010 and began pursuing an MS degree in economics at North Carolina State University.
ACKNOWLEDGMENTS

I would like to thank the following people who gave freely of their time to help make this thesis possible.

Dr. Professor Roger H. von Haefen and Dr. Professor Steven Sexton (Department of Economics, North Carolina State University)

Dr. Judy Wang, Dr. Donald Martin, Dr. Brian Reich, Dr. John Monahan, Dr. Ana-Maria Staicu (Department of Statistics, North Carolina State University)

Dr. Jerry Davis, Dr. Kristen Foley and Dr. Adam Reff (U.S. Environmental Protection Agency)

Dr. Richard Smith (Statistical and Applied Mathematical Sciences Institute (NSF))

Dr. Jenise Swall (Department of Statistical Sciences and Operations Research, Virginia Commonwealth University)
# TABLE OF CONTENTS

List of Tables .................................................................................................................. vi

List of Figures .................................................................................................................. vii

1. Introduction .................................................................................................................. 1

2. Literature Review ....................................................................................................... 3

3. Data Section ................................................................................................................ 8

4. Variable Selection ..................................................................................................... 11

5. Statistical Methods ................................................................................................... 12
   a. Generalized Linear Models ..................................................................................... 13
   b. Model Estimation ................................................................................................... 15
   c. Quantile Regression ............................................................................................... 15

6. Exploratory Data Analysis .......................................................................................... 17

7. Collinearity .................................................................................................................. 22

8. Results ....................................................................................................................... 24

9. Relative Risk .............................................................................................................. 31

10. Thresholding ............................................................................................................ 32

11. Prediction ................................................................................................................... 32

12. Conclusion ............................................................................................................... 33

13. References ............................................................................................................... 35

14. Figures and Tables ................................................................................................... 42
LIST OF TABLES

Table 1: Meteorological variables that were considered for use as covariates.................42

Table 2: Selected summary statistics ..................................................................................42

Table 3: Correlation matrix for the variables in the statistical models..............................43

Table 4: Results for the Variance Inflation Factor for the covariates..............................43

Table 5: Statistics for avg3pm10 and the square root of death3 for the OLS model.............44

Table 6: Statistics for avg3pm10 and the square root of death3 for the GLM model ..........45

Table 7: Statistics for avg3pm10 and the square root of total death..................................46

Table 8: Individual coefficients and standard errors from the AIC procedure for the OLS model for the nine years taken together.................................................................47
LIST OF FIGURES

Figure 1: Boxplot for PM10..................................................................................................................48
Figure 2: Boxplot for Death3..................................................................................................................49
Figure 3: Boxplot for Death2..................................................................................................................50
Figure 4: Boxplot for Death1..................................................................................................................51
Figure 5: Plot for Jan. to Apr. 1994 .......................................................................................................52
Figure 6: Plot for Jun. to Sep. 1994 .......................................................................................................53
Figure 7: Mortality Plot for death3 with a cubic spline fit .................................................................54
Figure 8: Mortality (counts) and PM10 (ppb).......................................................................................55
Figure 9: GLM for all nine years ............................................................................................................56
Figure 10: Residuals for the 1992 to 2000 Linear Fit ..........................................................................59
Figure 11: Residuals Autocorrelation Function for the 1992 to 2000 Linear Fit .........................60
Figure 12: Plot of Measured Mortality (R) and Predicted (B) for the Year 2000 .......................61
Figure 13: Quantile plot for avg3pm10 ...............................................................................................62
1. Introduction

As economic development reaches a certain threshold, societies shift their focus from maximizing economic growth to other quality of life considerations. Among them, the health effects of air pollution are becoming a major concern. Brunekreef and Holgate (2002) point out that although air pollution from the combustion of traditional fossil fuel is now present in much lower concentrations than several decades ago, society’s concerns about air pollution are showing no tendency to decline. Concern continues about the prominence of pollution problems arising from ground-level ozone, airborne particles, oxides of nitrogen, and other pollutants. The severe smog conditions in Beijing, China since January 2013, the subsequent online discussion and a grass-roots lobbying effort for stricter environmental regulations has attracted worldwide attention. The latest Global Burden of Diseases, Injuries, and Risk Factors Study, for the first time lists air pollution along with high blood pressure, tobacco and alcohol as a top 10 risk factor for disease (Lancet, 2012). Therefore, it remains an important policy question whether and how much certain types of pollution cause excess death.

The relationship between outdoor air pollution levels of PM10 (particles less than or equal to 10 micrometers in diameter) and adverse health effects has been well-documented in the United States, and similar work is now beginning in East Asia (Huang et al., 2009 and Cheng et al., 2013). In China air pollution levels in many of the major cities (e.g., Beijing, Shanghai and Guangzhou) have been very high; to further explore the environmental issues facing
China will be one of the main aims of my future study and research, and I hope this thesis can lay the foundation for that. Its main goal is to examine the relationship between adverse health effects and air pollution levels in the atmosphere for the period from 1992 to 2000. As done in Smith et al. (2000) for Birmingham, AL, I will focus on one location, Pittsburgh, PA. At one time Pittsburgh was notorious for its severe air pollution. The wide use of cheap and easily obtainable bituminous coal and its history of large scale steel production made the problem difficult to control1. As the most visible by-product of coal consumption, smoke’s controls were not enacted in Pittsburgh until more than a century after travelers described it as “hell with the lid off”2. A brief natural gas boom during the 1880s and early 1890s provided the city with a short clean air period and led to a smoke control movement. In 1941, following the example of St. Louis in its success of improving air quality by passing ordinances requiring the use of clean fuel, Pittsburgh adopted similar ordinances; however, it was believed that what eventually reduced most of the smoke was the piping of clean natural gas into the city from the southwest. The shift from coal-burning to diesel-electric locomotives by the Pennsylvania Railroad and other regional railroads in the 1950s helped further improve air quality significantly in the region. The efforts of working to encourage local enforcement of the Clean Air Act in the 1970s and 1980s by the advocacy organization

---


“Group Against Smoke and Pollution” also brought about some improvement. The collapse of the iron and steel industry in the 1980s was believed to have resulted in the most substantial air quality improvements. Today the greatest source of air pollution is from the by-products of coking (converting coal into coke) facilities and automobiles. This paper focuses on the period from 1992 to 2000 after Pittsburgh’s industrial decline.

In addition, Pittsburgh is a U.S. EPA Supersite for the ambient monitoring of particulate matter, and thus has detailed information (e.g. hourly PM10 data) on the chemical species that make up particulate matter in that city. In this paper I will mainly focus on identifying the effects of an important pollutant, PM10, on daily mortality by examining whether the results are consistent across different statistical models. This topic is of interest in part because of the improved quality of life that results from reduced air pollution levels and the benefits of having a healthy population. Many research studies have been devoted to trying to relate adverse health effects to air pollution levels. The following are some examples of some of those research studies.

2. Literature Review.
A number of epidemiology papers were published in the early 1990s relating PM10 to adverse health effects. Schwartz and Marcus (1990) associated particulate air pollution with daily mortality (after controlling for temperature and humidity) in London, England, both in

the smog periods of the 1950s and the lower pollution levels of the late 1960s and early 1970s. After that, Schwartz and his colleagues attempted to replicate these findings in the United States. For example, Schwartz and Dockery (1992a and 1992b) confirmed a significant positive relationship between total suspended particles\textsuperscript{4} and increased mortality in both Philadelphia and Steubenville, Ohio. Pope, Schwartz, and Ransom (1992) concluded that an increase in the 5-day moving average PM10 levels (equal to 100µg/m\textsuperscript{3}) was associated with an estimated increase in deaths per day equal to 16\% in the Utah Valley. Evidence was also found by Schwartz (1993 and 1994) for a significant association between daily mortality or hospital admissions due to pneumonia and chronic obstructive pulmonary disease and inhalable particles (PM10) in Birmingham, Alabama. Also, Schwartz (1995) concludes that air pollution concentrations were associated with increased respiratory hospital admissions of the elderly; the strongest evidence for this association was for PM10 followed by ozone. The author found that “the magnitude of the effect was small (relative risk 1.06 in New Haven and 1.10 in Tacoma for a 50 µg/m\textsuperscript{3} increase in PM10, for example) but, given the ubiquitous exposure, this has some public health significance.”

In addition to the papers mentioned above, there are some other papers that are interesting and relevant. For example, Thurston (1996) reviewed available time-series studies on PM10 and mortality and concluded that an acute pollution-mortality association can occur at routine

ambient levels of PM10. Chay and Greenstone (2003) found that a 1 mg/m3 reduction in particulates results in 4-8 fewer infant deaths per 100,000 live births at the county level (a 0.35-0.45 elasticity). Semat et al. 2000 provides a detailed discussion of morbidity and mortality from air pollution in the U.S. They provided a detailed discussion of model development and provide concentration-response curves and thresholds for many of the largest cities (including Pittsburgh) in the U.S. Their studies provide us with evidence about the consistency of the link between PM10 and adverse health outcomes, including morbidity and mortality for cardiovascular and respiratory diseases.

There are also other types of pollutants that are of interest such as ground-level ozone and sulfur dioxide (SO2). Extensive epidemiological and toxicological studies have been done in order to determine the health impact of ambient ozone level. Among these are Schwartz (1995), Bell et al. (2004), Bell et al. (2005), Ito et al. (2005), Medina et al. (2006) and Smith, Xu and Switzer (2009). Schwartz (1995) pointed out that the PM10 associations were little changed by controlling for either ozone or SO2. The ozone association was likewise independent of the other pollutants. The SO2 association, however, was substantially attenuated by controlling for ozone in some cases, and by control for PM10 in others. Therefore in our model we will only include PM10 as our primary pollution measure.

Most recent research on the effects of particulate matter on human health focuses on PM2.5, particles with a diameter less than or equal to 2.5 µm. Dockery et al. (1993) stated in their
famous Six Cities Study that statistically significant and robust associations between air pollution and mortality were observed. Air pollution was positively associated with death from lung cancer and cardiopulmonary disease but not with death from other causes considered together. Mortality was most strongly associated with fine particulate air pollution, including sulfates. Francine, Schwartz, Speizer and Dockery (2006) extended the Harvard Six Cities Study and concluded that total, cardiovascular, and lung cancer mortality and ambient PM2.5 concentrations were positively associated respectively, and that reduced PM2.5 concentrations were associated with reduced mortality risk. Fann et al. (2012) in their paper on the public health burden associated with exposure to ambient PM2.5 and ozone estimated that 130,000 PM2.5-related deaths and 4,700 ozone-related deaths resulted from 2005 air quality levels. They also estimated that for populations aged 65 to 99, nearly 1.1 million life years were lost from PM2.5 exposure and approximately 36,000 life years were lost from ozone exposure. This is an important contribution to the air quality literature and is a strong motivation for air pollution control. The reason that we are using PM10 in our analysis is because our database for mortality ends in the year 2000; PM2.5 data were not generally available at that time.

However, Styer, et al. (1995), showed that for two counties, Cook County, Illinois, and Salt Lake County, Utah, the evidence was mixed. In Salt Lake County, Utah no evidence was found that PM10 contributes to excess mortality; although, for the Cook County, Illinois there is some evidence of a positive PM10 effect in spring and autumn, but not in winter and
summer. Styer et al. (1995) did their analysis using the current day value of PM10, a two-day average value of PM10 (today and yesterday), and a 5-day average value of PM10 (today and 4 previous days). Smith et al. (2000) found that lagged values of the weather covariates (particularly specific humidity) and pollutants were often significant predictors of mortality in Birmingham, AL.

Only one article was found about the effect of air pollution on mortality or morbidity in Pittsburgh, PA. Mazumdar and Sussman (1983) mentioned a possible association between heart disease mortality/morbidity and same day particulate levels in Pittsburgh. The evaluation of adverse health effects resulting from exposure to relatively low levels of regulated air pollutants is currently of major concern. The determination of short-term or acute effects is necessary for this evaluation. By using methodology that directly addressed the time series nature of the data, Sussman investigated acute health effects of daily levels of air pollution in Allegheny County, Pennsylvania, using both mortality and morbidity events as the adverse health response to ambient pollution. Health effects were determined using the air quality data for sulfur dioxide (SO2) and particulates as measured by the Coefficient of Haze from three monitoring stations located within the county. The mortality analysis provided a replication of a previous study performed in the New York area. The analysis was limited to the investigation of same day effects. Results indicated a possible association between heart disease mortality/morbidity and same day particulate levels. No association between SO2 and mortality/morbidity was seen at the present level of SO2. These findings
were in agreement with those obtained in the New York City study and in a re-analysis of London winter data. The need for investigations of delayed effects and weather-mortality/morbidity relationships, as well as evaluation of hospital discharge record information, was stressed.

3. Data Section.

The health endpoint data came from the National Morbidity, Mortality and Air Pollution Study (NMMAPS) documented in the book by Peng and Dominici (2008). These data are available in the NMMAPSlite library, which is available through the R statistical package.

Many researchers have used this data set, see for example Smith et al. (2009) and Bell et al. (2004, 2005). The data consists of meteorological variables, air pollution variables and mortality data that include total non-accidental deaths, respiratory and cardiovascular deaths and chronic obstructive pulmonary disease plus deaths from other causes collected from 108 locations from 1987 through 2000. Three age categories are available in the data set: under 65, 65 to 74 and older than 74. Our focus in this thesis is on total non-accidental deaths. All three age categories will be considered, although we will mainly concentrate on the third category (older than 74). Mortality data for the period 2001 to 2005 has been developed, but is not generally available except to certain research groups. Hospital admissions data were originally part of the NMMAPS data set, but have been withdrawn because of privacy concerns. These data are available to selected research groups.
There were concerns about how some of the meteorological data in NMMAPS were arrived at, especially in regard to how the mean values of some of the data were obtained. For example, often the mean value of a meteorology covariate was arrived at by taking the average of the maximum value and the minimum value, and in other case, the average was the 24-hour average. To alleviate this concern, our meteorological data was obtained for the Pittsburgh airport (40.4914N, 80.2328W, elevation 367m) from the U.S. National Climatic Data Center in Asheville, NC. These data consist of temperature (daily maximum, minimum and the daily mean), mean daily values for relative humidity, dew point temperature, wind speed, station pressure and visibility. Other variables that have been considered as covariates, such as specific humidity, water vapor deficit, were calculated based on the variables from the data set. Table 1 gives all the meteorological variables considered as confounding covariates. The three variables that were used as covariates in this study are marked with an asterisk. The dew point temperature, or simply the dew point, is the temperature to which air must be cooled to bring about saturation with water vapor. The specific humidity is the ratio of the mass of water vapor to the mass of water vapor plus the mass of dry air. Because mass is not a measured atmospheric variable, the equation for specific humidity is converted to a ratio of pressures times the ratio of molecular weights using the equations of state for dry air and water vapor. Therefore specific humidity was calculated from the temperature, relative humidity and station pressure. Relative humidity tells one the degree of saturation of the atmosphere with water vapor, but not the actual amount of water vapor. Changes in relative humidity can be caused by a change in the water vapor content of the atmosphere or a change
in temperature or both, which limits its use for many applications. We also derived the water vapor pressure deficit, which is the actual vapor pressure minus the saturation vapor pressure (a function of temperature). Hourly data are also available for each of the meteorological parameters. These meteorological data will be more useful in this study compared to the meteorological data in the NMMAPS data base. For Pittsburgh, much of the PM10 data were missing in the NMMAPS data set. Thus the air pollution data (PM10) used in this study came from the U.S. EPA Air Quality System (AQS) data achieve. The AQS is EPA’s repository of ambient air quality data. Pittsburgh hourly (from which we calculated summary statistics) PM10 data are collected using the Tapered-Element Oscillating Microbalance (TEOM) instrumentation system (Kulkarni et al., 2011) for the years 1992 to 2000. The monitor was located at 40.4136N, 79.9414W. For the period 1 January 1992 to 31 December 2000, 3077 observations were available, only 6.3% of the data were missing.

Lagged values up to three days were available for all covariates. Table 2 provides some summary statistics for the covariates and the response variables used in the main model of this paper. It is worth noting that the median number of deaths decreases dramatically when passing from death3 (22) to the two lower categories (8 and 6, respectively). The maximum number of daily deaths during the nine year period was 49, which occurred for the oldest group of people. The maximum PM10 reading was $114 \mu g/m^3$. More will be said about these readings in a later section.

Counting all the potential covariates (meteorological variables and PM10) and their lags there are many sets of covariates that could be used in the regression models. In some studies (Smith et al., 2000), the initial approach was to find the best model for mortality using only the meteorology covariates and then to add PM10 to the model. This was also my initial approach to variable selection based on the AIC\(^5\) procedure; however, it was found that there was really no single best model. The results using the AIC procedure were, therefore, inconclusive in this situation. The second approach was to use covariates that had been used in the past so that the present results could be compared to those results. In most recent papers (e.g., Smith et al., 2009 and Bell et al., 2004) the temperature and the dew point temperature were the main meteorology covariates. Those two variables were used in this study along with wind speed. The temperature and dew point are important because they are used to model the heat stress index in the summer. In addition, in the summer wind provides an evaporative cooling mechanism for the human body. Temperature and wind speed are used to model the wind chill index in the winter. Some concern was expressed by Styer et al. (1995) about using wind speed because of its potential interaction with PM10. That interaction, however, was not statistically significant in our case. It is true that wind speed will affect the dispersion of particulate matter in the atmosphere; however higher wind speeds have contributed to the re-suspension of particles into the atmosphere. Table 3 shows a correlation matrix for the variables used in the statistical models. The information in the table indicated that temperature and dew point are highly related. The high correlation is

\(^5\) AIC stands for Akaike information criterion.
expected because at high temperature the atmosphere can hold more water vapor. But these two covariates are essential to modeling mortality as has been shown in numerous publications (e.g. Smith, 2009), so we chose to keep these in our analysis. Thus in the end, mean daily temperature, dew point and wind speed were the meteorology covariates used in model. Three forms of the PM10 covariate were considered for use in the various statistical models. One form was the average of today’s PM10 and the two previous days. The second form was based on the average of yesterday’s value plus the two previous days. Finally today’s PM10 plus three lags were used as individual covariates. The decision was made to use the first form based on t-statistics that were obtained from an OLS model for all nine years that used temperature, dew point temperature and wind speed as covariates. This form was also used in Cook County by Styer et al., (1995).

5. Statistical Methods.

The statistical models that have been developed have health endpoints, (e.g., mortality), as the response variable and PM10 levels and meteorological parameters as covariates. All statistical models were developed using the R statistical programming language. Meteorology represents a confounding effect in assessing mortality since meteorologically-caused death is also a possibility. For example, during the summer months heat stress resulting from high temperatures and dew points can cause heat-stress related deaths. In addition, PM10 levels may be affected by meteorological conditions. There is a detailed discussion of the confounding effects of meteorology in Appendix B of Styer et al. (1995).
Most statistical research involving health endpoints and count data uses Poisson regression techniques, although classical regression techniques (ordinary least squares, OLS) have also been used (Smith et al., 2000). In the present case when OLS methods were used, the square root of mortality was used as the response variable to help stabilize the variance. Nonlinear modeling was done using natural cubic splines within the Generalized Linear Model (GLM) procedure in R. The Poisson family can be selected in the GLM package. In addition to these techniques quantile regression methods (Hao and Naiman, 2007 and Koenker, 2005) have been used to assess the relationship between mortality and the meteorology and PM10 covariates. It is of interest, based on the coefficients for the covariates, to see how these coefficients change as one moves from the 0.1 quantile to the 0.9 quantile. Quantile regression was done for at least three levels: 0.1, 0.5 and 0.9. Because the mortality data are count data, the estimation of the conditional quantiles requires that the count data be adjusted in some manner. This is done smoothing the count data. To generate this smoothness, the technique referred to as jittering was used, which is described in Machado and Silva (2005). The section on quantile regression will provide a more detailed discussion of jittering.

a. Generalized linear models. In addition to OLS, statistical models for mortality were developed using non-parametric regression models, which provide a distribution-free basis for predicting the response variable over the range of the data and are helpful in dealing with the non-linearities that are frequently present in the relationship between mortality, pollution and meteorology (Smith et al., 2000). The framework for the models is based on Generalized
Linear Model (GLM) theory (McCullagh and Nelder, 1989). Previous statistical analyses of environmental variables have been successful using this class of models (Zheng et al., 2007; Camalier et al., 2007). Generalized Additive Models (GAM) (Hastie and Tibshirani, 1990; Wood, 2006) is an extension of the GLM approach and can be run within the R GLM function. GAMs allow for covariates like those used in simple regression models ($x_i^T \beta$) as well as for functional forms of the covariates (indicated by f()), which provides greater flexibility when modeling complex non-linear processes. Below is the general form of the equation used in this work:

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \text{dowf} + \text{seaf} + \text{yrf}$$

(1)

The g() in Equation (1) is referred to as the link function (McCullagh and Nelder, 1989; Dobson and Barnett, 2008), which specifies the relationship between the linear portion of the model on the right hand side of the equation and the expected response $\mu_i$. The natural log-link function has been found to be the most appropriate link function to stabilize the variance of the mortality data, and so is used in this work. The parameter $\beta_0$ represents the overall mean and $f_j(x_{i,j})$ is the value of the smoothing function associated with the $i$th value of the explanatory variable $j$, where $j = 1\ldots p$. The term dowf is a factor that represents the effect of the day of the week (Sunday to Saturday, respectively). The term seaf is a factor that represents the effect of a given season. The term yrf is a factor that represents the effect of a given year. A natural cubic spline (Green and Silverman, 1994; Hastie and Tibshirani, 1990; Wood, 2006) with three degrees of freedom was used as the smoothing function in our work to allow for a non-linear response between each covariate (PM10 and meteorology) and the
mortality counts. The basic regression model also contains a covariate for the trend, which is based on days, and is fitted using either the loess (Cleveland et al., 1992) function in R or with natural cubic splines.

When a log-link function is used in a GLM, the effect of each covariate is defined as the percent change in the response variable for a unit change in the given covariate assuming that all other covariates remain constant.

b. Model estimation. For models based on OLS, standard least squares procedures were used. Within the GLM framework, the Poisson family was specified. Model runs were also made with the quasi family, which refers to the quasi-likelihood model or to quasi-likelihood inference (McCullagh and Nelder, 1989). This is the approach taken in Smith et al. (2009). This family allows one to estimate the degree of over-dispersion in the analysis. Estimation employing the maximum likelihood procedure requires that we specify the form of the distribution of the observations. However, to use a quasi-likelihood estimation procedure, we only have to specify the relationship between the mean and the variance of the observations. For the GLM models we have used both the Poisson and quasi families.

c. Quantile regression. The discussion of quantile regression is based on Hao and Naiman (2007). Median regression has been used to represent the relationship between the central location of the response and a set of covariates. The median describes the central location of
the distribution. In conditional-median regression the conditional 0.5 quantile is modeled as a function of a family of covariates. Other quantiles can be used to describe the non-central positions of a distribution. The $p^{th}$ quantile denotes the value of the response below which the proportion of the population is $p$. Any position of a distribution can be specified by quantiles, thus allowing for a complete description of the distribution (Koenker, 2005; Hao and Naiman, 2007).

Quantile regression, which models conditional quantiles as functions of predictors (e.g., covariates), is due to Koenker and Bassett (1978). Given that many quantiles can be modeled in quantile regression, one can obtain a more complete understanding of how the distribution of the response variable is affected by the covariates, in addition to obtaining information about the shape change of the distribution. A set of equally spaced conditional quantiles can characterize the shape of the conditional distribution, and in addition obtain information on its central location.

The quantile regression model developed by Koenker and Bassett (1978) estimates the effects that a given covariate has at differing quantile levels in the conditional distribution. The regression lines for the median and off-median quantiles capture the location shift (given by the line for the median), as well information on the scale and shape shifts (provided by the other quantiles). Thus the differential effect of a given covariate on the full distribution can
be assessed. The authors provide the following fundamental equation which corresponds to the basic OLS equation.

\[ y_i = \beta_0^{(p)} + \beta_1^{(p)}x_i + \varepsilon_i^{(p)} \quad (2) \]

where \(0 < p < 1\) indicates the proportion of the population having values below the quantile at \(p\). In a manner similar to OLS, the \(p\)th conditional quantile is given by the following equation.

\[ Q^{(p)}(y_i/x_i) = \beta_0^{(p)} + \beta_1^{(p)}x_i \quad (3) \]

The \(p\)th conditional quantile is determined by the specific quantile coefficients and the specific value of \(x\). As Hao and Naiman (2007) indicate, numerous equations can be expressed in the form of equation (2). If the analysis specifies 12 quantiles, these 12 equations yield 12 coefficients for \(x_i\).

6. Exploratory Data Analysis.

The purpose of this section is to take a closer look at the raw data before we look at the results from the regression analysis. Figures 1-4 show box plots for PM10 and the three levels of mortality (death1 (<65), death2 (65 to 74) and death3 (≥ 75)). The higher PM10 levels and the higher mortality counts are out of phase with each other. The highest PM10 levels occur in the summer while the highest mortality counts occur in the winter months. The mortality peaks in the winter are probably due to infectious diseases such as influenza. The box plots show the median value, the upper and lower quartiles and the quantiles at 0.05 and 0.95 for each month by year. Values outside those ranges are plotted as black dots. The median number of mortality counts (considering all nine years) for death1 is 6, for death2 it
is 8 and for death3 it is 22 (see Table 2). The largest count for mortality considering all years and all age ranges was 49 for mortality category death3 which occurred in January 1999. For death1 the maximum value is 17, which appears to have occurred three times (April 1992, January 1993, and April 2000), while for death2 this value is 24 in January 1994. The interquartile range increases from 3 for death1 to 7 for death3. The largest season to season variability in mortality appears to occur for the oldest group of people (death3) and decreased for the younger groups of people. Looking across all years, there appears to be a slight downward trend with advancing years for the two younger groups, while for the oldest group of people there appears to be a slight increase over the years.

The highest PM10 value reported over the nine year period was 114.08 µg/m³ in March 1995. Values above 80 µg/m³ occurred in all years up through 1998. In 2000, no values above 60 µg/m³ were reported. The median PM10 value for the entire time period was 24.46 µg/m³. The U.S. EPA mandated standard for PM10, at that time, as reported in Styer et al. (1995), consisted of both a 24-hr and an annual standard. The 24-hr standard is attained when the expected number of days per calendar year with a 24-hr average concentration above 150 µg/m³ is equal to or less than one. The annual standard is attained when the annual arithmetic mean concentration is less than or equal to 50 µg/m³. As mentioned in an earlier section, this study is based on hourly PM10 data not on data collected from monitoring locations that sample every third or sixth day. Our results show that both PM10 and meteorology are statistically significant predictors of mortality for death3; however, for death
levels one and two only the meteorology covariates were statistically significant. Please see the results section.

The previous figures clearly indicate that mortality levels are highest in winter at the same time that temperatures are at their lowest and the amount of water vapor in the atmosphere, as indicated by the dew point temperature, is also very low. Many authors state that the high number of deaths in winter can be attributed to infectious diseases like influenza, which is often the case. But equally often it is assumed that neither meteorological conditions nor particulate matter are the cause of these infectious diseases or at least contribute to them in some way.

In the recent article by Deschênes and Greenstone (2008), the authors discuss how temperatures far outside of normal ranges can be dangerous to human health and can result in premature mortality. The authors point to recent work (see their article for a list of the research articles they cited) that indicates that extremely high temperatures can cause excess deaths related to cardiovascular, respiratory, and cerebrovascular diseases (diseases related to the brain and its blood vessels). The need for the regulation of body temperature causes additional stress on the cardiovascular and respiratory systems under extreme temperature conditions. These high temperatures are also related to decreases in the viscosity of the blood and higher cholesterol levels. Deschênes and Greenstone also cite literature that indicates that deaths that are caused immediately after a period of high temperatures are often followed by
a reduction in the number of deaths in the days following the high temperatures. This pattern is often referred to by the disagreeable term “harvesting”, meaning that people who are in this category are ones who would have died in the near term anyway.

Cold days are also responsible for excess mortality. Deschênes and Greenstone state that these cold conditions can cause cardiovascular stress due to changes in blood pressure, vasoconstriction, and an increase in blood viscosity (which can lead to clots). In addition, levels of red blood cell counts, plasma cholesterol, and plasma fibrinogen are affected. Susceptibility to pulmonary infections may increase because breathing cold air can lead to bronchoconstriction.

In an earlier paper, Deschênes and Moretti (2007) provide the most comprehensive evidence on the impacts of cold days on mortality. They found “evidence of a large and statistically significant effect on mortality within a month of the cold wave. This effect appears to be larger than the immediate effect, possibly because it takes time for health conditions associated with extreme cold to manifest themselves and to spread.” Thus, in the case of cold weather, it may be that there are delayed impacts and that the full effect of a cold day takes a few weeks to manifest itself. Further, they find that the impact is most pronounced among the young and elderly and concentrated among cardiovascular and respiratory diseases. In this thesis it seems likely that these additional stresses imposed on the human body by these meteorological conditions make people more susceptible to epidemics like influenza.
Figure 5 shows the seasonal transition in death3, temperature, dew point temperature, and PM10 for the period January to April 1994. The count value for mortality has a decreasing trend during this period. There are many peaks in the PM10 values over the entire time period, but the overall trend is fairly flat. The large increase in PM10 values will occur later in the spring and in the summer. There are some time periods when higher PM10 values are associated with higher mortality values, but this is not always the case. Time series techniques may be useful in determining the relationship between these two variables. As expected temperature and dew point rise through this time period. The amount of water vapor present in the atmosphere is very temperature dependent and is indicated by the dew point temperature. The correlation coefficient in Table 2 for the temperature and dew point temperature variables was 0.953. Almost all parameters that provide information about the amount of water vapor in the atmosphere are highly correlated with atmospheric temperature. Figure 6 is similar to Figure 5 except it is for the period June through September in the same year. During this period mortality shows no long term trend; PM10 values are generally high throughout the period with the largest value occurring in June. There is a slight downward trend in temperature and dew point during this period. There is an interesting event after the 10th of June (which is near day 10 in Fig.6) when PM10, mortality and temperature levels are all high. The peak in mortality occurs several days after the peak of PM10. A somewhat similar event occurred near the end of June (about day 30 on the plot) when all three terms were simultaneously high. Again the peak in mortality occurred a short time after the peak in PM10.
7. Multicollinearity.

When working with the observational meteorological and chemical data used in the current model, multicollinearity is always a potential problem. A very simple measure of collinearity can be assessed using the correlation matrix similar to Table 2. As shown in the table, the correlation between temperature and dew point temperature is 0.953. Multicollinearity is the name given to the phenomenon when independent variables are highly correlated. Although the estimation procedure does not break down when multicollinearity exists, severe problems may arise. The major undesirable consequence resulting from multicollinearity is that the variances of the OLS estimates of the parameters of the collinear variables are quite large. That will cause the parameter estimates to not be precise and the hypothesis testing not powerful.

Correlation matrix is, however, only helpful in detecting near-linear dependence between pairs of regressors. To assess the degree of collinearity, we have taken the singular value decomposition of the design matrix, X, as recommended by Belsley (1991) and examined the ratio of the largest singular value of the design matrix to all the other singular values of the design matrix. These ratios are defined as the condition indices of the design matrix. It should be noted that the squares of the singular values of the design matrix are the eigenvalues of $X'X$. The ill-conditioning in $X$ is indicated by the size of the singular values. For each near linear dependency there will be a small singular value of $X$. The extent of the ill-conditioning is based on the calculated conditioning indices. There are as many
dependencies among the columns of the matrix X as there are high ratio values. Based on the
guidance provided by Belsley (1991), weak dependencies are indicated by condition indices
of around 5 to 10, while moderate to strong relationships are indicated by condition indices
of 30 to 50. Based on the recommendations by Belsey the columns of the X matrix have been
scaled to have equal length; however, no centering has been done. The ratios obtained using
the meteorological and PM10 covariates in our regression model were found to be:
1.000, 1.867, 4.968, 6.587, 8.436, 10.014, 11.130, 12.735, 13.370, 16.373, 21.822, 30.939, 44.373
Two ratios fall in the 30 to 50 interval mentioned by Belsley.

In addition, the Variance Inflation Factor (VIF) was calculated for each covariate (see Table
4). To obtain the VIF for the regression coefficients for each covariate required that each
covariate be regressed on all the other covariates and that the ratio of one over one minus R
squared to be calculated. A VIF above 10 may indicate that the given regression coefficients
are poorly estimated because of collinearity (Montgomery et al., 2012).

The two methods above that we have used to assess the presence of collinearity also seem to
indicate that collinearity is a potential problem. As mentioned earlier in the thesis, I chose to
leave the temperature and dew point temperature in the regression model because the two are
considered essential confounding covariates in estimating the relationship between air
pollution and mortality as has been shown in many other research papers in this area.
Dropping one of these two covariates was considered, but it also possibly leads to specification errors. More exploration of this multicollinearity problem may be worth doing in future research work.

8. Results.

Figures 1-4 provided a detailed look at how PM10 and mortality varied over the entire nine-year period for all three age groups. Figure 5 showed how the change in the death3 category for mortality, PM10, temperature and dew point change during the transition period from winter to spring of 1994. The spring to summer transitions are shown in Figure 6. Figures 7 and 8 showed further evidence of the seasonal variability of PM10 and mortality. Figure 9 is based on a GLM run for all nine years using the basic model (temp, dewp, wdsp, plus lags and avg3pm10). This figure clearly shows that a linear approach to modeling mortality is possible given the 95% confident bands provided in that figure. As a result, the use of OLS in addition to non-parametric techniques seems reasonable. The results obtained in this thesis have been compared to two older papers (Styer et al., 1995 and Smith et al., 2000) that employ a similar approach taken in this paper. Initially we examined the relationship between PM10 and the meteorology variables using OLS techniques. The analysis used the meteorological covariates from the basic model with all their lags. Because of the concern expressed earlier about having PM10 and wind speed in the model, we also included an interaction term for them in the regression analyses that was done. This term was not statistically significant and was therefore dropped from further consideration. About 27% of
the variation in PM10 could be explained by the meteorology variables alone. Different sets of meteorological variables will produce different levels of explained variance. In one run it was possible to explain over 40% of the variation in PM10.

Table 5 shows the results for the OLS model. Results are presented for the entire nine years taken together, and for the individual years and seasons. The values of the coefficients are consistent with those found by Smith et al. (2000) and Styer et al. (1995). The PM10 covariate (avg3pm10) was found to be statistically significant when considering the entire nine years and for the years 1993, 1994, 1995 (at the 0.10 level), 1996 (at the 0.10 level) and 1999 (at the 0.1 level) and for the winter months (at the 0.1 level) and the spring months. The results from Styer et al. (1995) for Cook County found somewhat similar inconsistencies in seasonal results: PM10 was important in the spring but not the summer, and in the fall but not the winter. Those authors were, as the present author is, unable to explain why PM10 is a significant predictor of mortality in some years and seasons, but not in other years and seasons.

Levels of PM10 were at their highest in the period from 1993 to 1995; Mortality levels were also high in 1994 and 1995. If one examines the more detailed plots (Figure 5, spring 1994 and 6, summer 1994) there is some correspondence between spikes in the PM10 levels and spikes in the mortality counts (e.g., around days 22, 48, 63, 72, 93, and 115) in Figure 5, and around days 15, 95, and 102 in Figure 6. However, there are just as many cases of spikes in
the PM10 levels with no corresponding spikes in the mortality counts in this time period. The temperature and dew point show some correspondence with the events listed above. The unusually low temperature and dew point on about day 15 (Figure 5) coincides with a spike in the mortality count. The lower temperatures on succeeding days were followed by peaks in the PM10 levels, but not in mortality levels. Perhaps an increased demand for heating resources led to the higher PM10 levels. The very high peak in PM10 on day 49 was preceded by rising temperatures and dew points.

Table 8 gives the coefficients, standard errors, t-statistics and p-values for the AIC selected meteorology covariates from the OLS model using all nine years of data. Despite the fact that the meteorology covariates and PM10 covariate are statistically significant of mortality, they only account for about 20% of the variation in death3.

Figure 10 is a residual plot for the OLS model for the entire nine years. The plot indicates that the basic OLS model is adequate. Figure 11 is a plot of the autocorrelation function for the residuals from the same model. There is some evidence of weak statistically significant autocorrelation, but this was not considered to be strong enough to be a problem in this study.

Table 6 shows the results for statistical models that were done using natural cubic splines within the GLM framework. The results for the nine years taken together indicate that the PM10 variable (avg3pm10) was not a significant predictor of mortality. For year 1993 and
the spring season, PM10 was statistically significant. With the exception of 1994, these results are similar to those obtained using OLS; although, many of the p-values were much higher in the GLM model. Evidently the OLS model provided a better fit to the data than the GLM model. Table 7 shows the result for total death (the sum of mortality levels 1, 2, and 3). The results are similar to those obtained for root_death3 using OLS (Table 5). The results for the entire nine years are similar in Tables 5 and 7. Very low p-values were found in all three cases for the year 1993 and for the spring season. The results from Tables 5-7 indicate that PM10 is a statistically significant predictor of death for the older group of citizens.

Another issue that may arise is over-dispersion. If the response variable has a Poisson distribution, then the variance of the square root of the response variable is approximate one-fourth (Smith et al., 2000). Based on the OLS model for all nine years, which is representative of all the models we have run the standard unbiased estimator of $\sigma^2$ is $s^2 = 0.268$ with 3047 degrees of freedom. The over-dispersion is found to be approximately 6.7 percent. This level of over dispersion is statistically significant based on a chi-squared test even though the difference (0.268-0.25) is very small. However, it is also possible to calculate the over-dispersion parameter using the following formula that is found in Gelman and Hill (2007) and Tutz (2012).

$$\Phi = 1/(n-p) \sum (y - \hat{u})^2 / \hat{u}$$

The summation is from 1 to n. The parameter p is the number of covariates, y is the response variable and $\hat{u}$ are the fitted values. For our data, the dispersion parameter using this formula
was 0.98. The value of the dispersion can also be obtained using GLM with the family quasi (link = log, variance = μ). This value was 1.01. It appears that over-dispersion is not a problem for my work.

The analysis of mortality categories 1 (death1) and 2 (death2) were handled in a slightly different manner than was used for category 3 (death3). In addition to the usual covariates plus factors for year, season and day of the week, covariates were added for the interactions between $avg3pm10$ and year and between $avg3pm10$ and season. For death1, $avg3pm10$ was not statistically significant (p value of 0.953), and the two interaction terms with $avg3pm10$ were not statistically significant. As a final test, an F statistic was calculated according to the formula:

$$\frac{(SSE(\text{reduced}) - SSE(\text{full}))/\Delta df}{SSE(\text{full})/df} \sim F(\Delta df, df)$$

The full model contained the two interaction terms with $avg3pm10$, while the reduced model did not have those terms. Based on the F statistic the p value was 0.4684, which indicates that the reduced model was to be preferred over the full model and further confirms that the interaction terms were not statistically significant. The same procedure was used to examine death2. The p-value for $avg3pm10$ was 0.7295. The F test p-value was 0.8940, again confirming that the interaction terms were not statistically significant. These results indicate that PM10 was not a statistically significant predictor of mortality for people in the groups 1 and 2.
However, Styer et al.’s (1995) summary of their analysis likely applies to ours as well: “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 us from concluding that there is an effect between “today’s” mortality and “yesterday’s” particulates. The question appears to be unresolved.”

In our initial quantile analysis, we consider the same basic model that we used for the OLS and GLM work. The count data do not follow a continuous distribution, thus it is necessary to adjust the count values to give them the characteristics of continuous data. As mentioned in an earlier section of the thesis, we used a jittering procedure (Machado and Santos Silva, 2005) to try to simulate continuous data. In that procedure, a random sample from the uniform distribution (runif(n,0,1)) is added to the count data, where n is the number of observations. The procedure entails using 50 iterations and then providing a summary of those iterations. If a comparison is made between the original count data and the fitted values, only the integer portion of the fitted values is used for the comparison. The theoretical approach to the jittering process is discussed in the paper cited above. Three values of quantiles were considered (0.1, 0.5, and 0.9). The goal was to investigate the statistical significance of PM10 at those quantile values. At 0.1<sup>th</sup> quantile, the p-value for the covariate avg3pm10 is 0.1126, at 0.5<sup>th</sup> quantile, the p-value is 0.9948 and at 0.9<sup>th</sup> quantile, the p-value is essentially zero. Thus the PM10 is a significant predictor of mortality at a quantile value of
0.9, but not at the other levels. This would seem to indicate that for mortality counts below a quantile of 0.9, PM10 may not be a good predictor of death, while above that level it’s a strong predictor of death.

Another quantile run was done for death3 and avg3pm10. The raw count data were used in this case. Quantile runs were done for the PM10 covariate at quantile levels from 0.1 to 0.9 at 0.1 intervals. Figure 13 shows a plot for avg3pm10 (mean of the current day PM10 value and the two preceding days) for quantile values from 0.1 to 0.9 by 0.1. The slopes of the lines for the mean and the median are almost identical. The lines for the mean and median indicate the central location shifts. All the off-median lines for quantile values are basically parallel. The median clearly indicates that as PM10 levels decrease, as happens when moving from the summer PM10 peaks to the lower levels in winter, the mortality increases as the box plots presented earlier clearly show. The similar slopes for the off-median quantile values indicate that location shifts at all these quantile values are basically the same. No shape shift is indicated. Thus, as one moves along a given quantile value line the change in the death count for a given change in PM10 will be the same for all quantile values.

Based on the quantile runs when all four covariates and their lags are in the model, we can use a variation of the Wald test (Koenker and Bassett, 1982) to test for significant differences in the slope values for these models. The anova.rq package in R was used for these tests. Slope comparisons were made for quantile levels at 0.1 and 0.5, 0.1 and 0.9, 0.5 and 0.9. The
p-values were 0.9658, 0.5033, and 0.2546, respectively, which indicated that the null hypotheses cannot be rejected. There are no statistically significant differences in the slopes.


We have calculated the relative risk using the data from our OLS model for mortality. For the PM10 covariate in that model the estimated regression coefficient for the PM10 term is $5.017 \times 10^{-3}$ with a standard error of $1.032 \times 10^{-3}$. Recall that we have applied a square root transformation to the mortality data. Because we have used this transformation, it is somewhat difficult to arrive at the relative risk. However, we can arrive at an approximation using the following procedure (see Smith et al., 2000). The mean number of deaths per day for the age category greater than or equal to 75 is 22.8. If we assume a 10 µg/m$^3$ increase in PM10, then based on the coefficient (0.0050171), this should lead to an increase of 10 X 0.0050171 in the mean square root of death. Thus the estimated relative risk is $(22.8^{0.5} + 0.050171)/22.8 = 1.021125$. An approximate 95% confidence interval is (1.013,1.030) obtained after applying the same calculation to both endpoints of a 95% confidence interval for the coefficient.

10. Thresholding.

In this thesis, we approach this issue by assuming a threshold value of 50µg/m$^3$, which is the National Ambient Air Quality Standard (NAAQS) for PM10. We project the increased
mortality rate for PM10 levels above that quantity. Our guidelines are the ones discussed in Smith et al., 2000. The covariates are the same ones used in previous models. The original GLM equation is given by:

$$\text{glm.out} = \text{glm}(\text{root_death3} \sim x + \text{meteorological covariates}, \text{family} = \text{poisson})$$

where the x stands for PM10. This equation is modified to look like:

$$\text{glm.out} = \text{glm}(	ext{death3} \sim x_1 + x_2 + \text{meteorological covariates}, \text{family} = \text{poisson})$$

where $x_1 = (\text{pm10} \leq \text{threshold value})$

$$x_2 = (\text{pm10} > \text{threshold})*(\text{pm10}-\text{threshold})$$

For a PM10 level 50µg/m3 above the threshold level of 50µg/m3 mortality levels would increase by 0.24% in this case.

11. Prediction.

It was of interest to see how well level-three mortality could be predicted using the GLM natural cubic splines model that we had been using in the earlier parts of the thesis. The year 1992 to 1999 were used to develop a prediction equation for mortality levels in the year 2000. The null deviance was 84163 on 2757 degrees of freedom; the residual deviance was 67447 on 2718 degrees of freedom. The covariate for PM10 (avg3pm10) was the most statistically significant predictor ($p$-value = 0.00002). Figure 12 shows the actual mortality levels for the year 2000 and the predicted levels for that year. The seasonal variation in mortality levels is captured very well by the model; however, the large swings in the actual mortality are not well predicted, which is not a surprise given the poor fit for mortality obtained with the 1992
to 1999 statistical model. Being able to predict the seasonal variation in mortality coupled with the importance of PM10 in the model is a first step in an attempt to actually develop a more advanced prediction model for mortality.

12. Conclusions

The purpose of this thesis has been to determine if PM10 is a statistically significant predictor of mortality in Pittsburgh. Given the complexity of the issue as exemplified by the differences in the results from the various regression models, a definitive statement about that relationship is not possible. The following observations are relevant.

1. The results shown in Figures 5 to 7 indicate that particulate pollution is, in some cases, strongly related to mortality even though there is no consistency in the results from year to year or season to season as Styer et al. (1995) also found. Both OLS and GLM models agree that this relationship is strong in the spring and in the year 1993.

2. The relationship between meteorology variables, PM10 and mortality is very complicated. One group of meteorological covariates may lead to results that indicate that PM10 is a significant predictor of mortality, while a different group of meteorological covariates may indicate that PM10 is a statistically significant predictor of mortality.
3. Some authors have questioned the need for having a measure of atmospheric humidity in these regression models. The work by Smith et al. (2000) demonstrated that humidity should be in these regression models. However, as discussed in this thesis the form of the humidity measurement is important. The dew point temperature and the specific humidity are the best forms that should be used in the regression models because they measure the actual amount of water vapor in the atmosphere and can be related to assessing the level of heat stress on the human body.

4. Quantile regression using the jittering method indicates that for lower values of mortality counts, PM10 may not play a significant role, however, for the higher levels of mortality counts it appears to be very significant.

5. The final form of PM10 to be used in these models is an open question. The one selected here has been used by other authors (Styer et al., 1995); however this does not mean that it’s the best form.
REFERENCES


Part I: Methods and Methodological Issues
Part II: Morbidity and Mortality from Air Pollution in the United States
Part III: PM10 Concentration-Response Curves and Thresholds for the 20 Largest US Cities.


Figures and Tables

Table 1. Meteorological variables that were considered for use as covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>minimum</th>
<th>1st quarter</th>
<th>median</th>
<th>mean</th>
<th>SD</th>
<th>3rd quarter</th>
<th>maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Maximum Temperature (max, C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Minimum Temperature (min, C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Temperature (temp, C)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Dewpoint Temperature (dewp, C)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Specific Humidity (sphum, g/kg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Relative Humidity (relhum, %)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Vapor Pressure (vpres, hPa)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Vapor Pressure Deficit (vapdef, hPa)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Station Pressure (stp, hPa)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Visibility (visib, km)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hr Mean Wind Speed (wdsp, m/s)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Meteorological covariates used in the final statistical models. Lags of up to three days were calculated and used in the statistical models.

Table 2. Selected summary statistics*

<table>
<thead>
<tr>
<th>Variable</th>
<th>minimum</th>
<th>1st quarter</th>
<th>median</th>
<th>mean</th>
<th>SD</th>
<th>3rd quarter</th>
<th>maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp (C)</td>
<td>-25.22</td>
<td>2.67</td>
<td>11.83</td>
<td>1.78</td>
<td>9.98</td>
<td>19.56</td>
<td>30.83</td>
</tr>
<tr>
<td>dewp (C)</td>
<td>-31.28</td>
<td>-2.78</td>
<td>5.06</td>
<td>4.84</td>
<td>9.78</td>
<td>13.11</td>
<td>23.39</td>
</tr>
<tr>
<td>wdsp (m/s)</td>
<td>0.41</td>
<td>2.63</td>
<td>3.58</td>
<td>3.75</td>
<td>1.56</td>
<td>4.63</td>
<td>10.71</td>
</tr>
<tr>
<td>pm10 (ug/m3)</td>
<td>0.90</td>
<td>16.91</td>
<td>24.46</td>
<td>27.77</td>
<td>15.04</td>
<td>34.79</td>
<td>114.08</td>
</tr>
<tr>
<td>death1 (count)</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>6.64</td>
<td>2.61</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>death2 (count)</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>8.71</td>
<td>3.13</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>death3 (count)</td>
<td>9</td>
<td>19</td>
<td>22</td>
<td>22.78</td>
<td>5.56</td>
<td>26</td>
<td>49</td>
</tr>
</tbody>
</table>

* temp, dewp, wdsp and pm10 are 24-hr averages
Table 3. Correlation matrix for the variable in the statistical models

<table>
<thead>
<tr>
<th></th>
<th>temp</th>
<th>dewp</th>
<th>wdsp</th>
<th>pm10</th>
<th>death1</th>
<th>death2</th>
<th>death3</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>1.000</td>
<td>0.953</td>
<td>-0.302</td>
<td>0.433</td>
<td>-0.097</td>
<td>-0.159</td>
<td>-0.306</td>
</tr>
<tr>
<td>dewp</td>
<td>0.953</td>
<td>1.000</td>
<td>-0.283</td>
<td>0.365</td>
<td>-0.087</td>
<td>-0.152</td>
<td>-0.345</td>
</tr>
<tr>
<td>wdsp</td>
<td>-0.302</td>
<td>-0.283</td>
<td>1.000</td>
<td>-0.304</td>
<td>0.109</td>
<td>0.103</td>
<td>0.091</td>
</tr>
<tr>
<td>pm10</td>
<td>0.433</td>
<td>0.365</td>
<td>-0.304</td>
<td>1.000</td>
<td>-0.003</td>
<td>0.017</td>
<td>-0.138</td>
</tr>
<tr>
<td>death1</td>
<td>-0.097</td>
<td>-0.087</td>
<td>0.109</td>
<td>-0.003</td>
<td>1.000</td>
<td>0.073</td>
<td>0.037</td>
</tr>
<tr>
<td>death2</td>
<td>-0.159</td>
<td>-0.152</td>
<td>0.103</td>
<td>0.017</td>
<td>0.073</td>
<td>1.000</td>
<td>0.072</td>
</tr>
<tr>
<td>death3</td>
<td>-0.360</td>
<td>-0.345</td>
<td>0.091</td>
<td>-0.138</td>
<td>0.037</td>
<td>0.072</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4. Results for the Variance Inflation Factor for the covariates

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Variance Inflation Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>25.345</td>
</tr>
<tr>
<td>temp1</td>
<td>43.601</td>
</tr>
<tr>
<td>temp2</td>
<td>42.995</td>
</tr>
<tr>
<td>temp3</td>
<td>27.700</td>
</tr>
<tr>
<td>dewp</td>
<td>19.468</td>
</tr>
<tr>
<td>dewp1</td>
<td>29.446</td>
</tr>
<tr>
<td>dewp2</td>
<td>29.197</td>
</tr>
<tr>
<td>dewp3</td>
<td>18.780</td>
</tr>
<tr>
<td>wdsp</td>
<td>1.335</td>
</tr>
<tr>
<td>wdsp1</td>
<td>1.883</td>
</tr>
<tr>
<td>wdsp2</td>
<td>1.915</td>
</tr>
<tr>
<td>wdsp3</td>
<td>1.652</td>
</tr>
<tr>
<td>pm10</td>
<td>1.475</td>
</tr>
</tbody>
</table>
Table 5. Statistics for avg3pm10 and the square root of death3 for the OLS model.

|                           | Coefficients | Std. Error | t-value | Pr(>|t|) |
|---------------------------|--------------|------------|---------|----------|
| (a) Results for the nine years (1992 to 2000) take together | 0.0050171    | 0.0010320  | 4.862   | 1.22e-06 |
| (b) Results for Individual years |              |            |         |          |
| 1992                      | 0.003977     | 0.002609   | 1.525   | 0.1283   |
| 1993                      | 0.006778     | 0.002393   | 2.833   | 0.00488  |
| 1994                      | 0.008087     | 0.003060   | 2.643   | 0.00872  |
| 1995                      | 0.003960     | 0.002281   | 1.736   | 0.08339  |
| 1996                      | 0.005559     | 0.002889   | 1.924   | 0.0551   |
| 1997                      | 0.000226     | 0.000874   | 0.258   | 0.796    |
| 1998                      | 0.005049     | 0.003135   | 1.611   | 0.1081   |
| 1999                      | 0.009659     | 0.005217   | 1.852   | 0.06506  |
| 2000                      | 0.005700     | 0.006046   | 0.943   | 0.3465   |
| (c) Results for the seasons |              |            |         |          |
| Winter                    | 0.004303     | 0.002534   | 1.698   | 0.0899   |
| Spring                    | 0.007425     | 0.002117   | 3.507   | 0.00048  |
| Summer                    | 0.003199     | 0.001967   | 1.626   | 0.10432  |
| Fall                      | 0.001199     | 0.002086   | 0.575   | 0.56567  |
Table 6. Statistics for avg3pm10 and death3 for the GLM model.

| Pr(>|t|)          |                  |
|-------------------|------------------|

(a) Results for the nine years
(1992 to 2000) take together 0.16975

(b) Results for Individual years

| Year | Pr(>|t|) |
|------|---------|
| 1992 | 0.77184 |
| 1993 | 0.04912 |
| 1994 | 0.56870 |
| 1995 | 0.50277 |
| 1996 | 0.99007 |
| 1997 | 0.78264 |
| 1998 | 0.44093 |
| 1999 | 0.24277 |
| 2000 | 0.91167 |

(c) Results for the seasons

| Season | Pr(>|t|) |
|--------|---------|
| Winter | 0.56480 |
| Spring | 0.00955 |
| Summer | 0.37568 |
| Fall   | 0.94347 |
Table 7. Statistics for avg3pm10 and the square root of total death

|                        | Coefficients | Std. error | t-value | Pr(>|t|) |
|------------------------|--------------|------------|---------|----------|
| (a) Results for the nine years (1992 to 2000) taken together | 0.003778     | 0.001079   | 3.502   | 0.00047  |
| (b) Results for individual years                           |              |            |         |          |
| 1992                   | 0.002395     | 0.002606   | 0.919   | 0.35870  |
| 1993                   | 0.006537     | 0.002517   | 2.598   | 0.00978  |
| 1994                   | 0.005544     | 0.003078   | 1.801   | 0.07286  |
| 1995                   | 0.002097     | 0.002248   | 0.933   | 0.35164  |
| 1996                   | 0.009053     | 0.002914   | 3.106   | 0.00205  |
| 1997                   | 0.003784     | 0.002760   | 1.371   | 0.17125  |
| 1998                   | 0.003152     | 0.003653   | 0.863   | 0.3887   |
| 1999                   | 0.003185     | 0.005059   | 0.630   | 0.52941  |
| 2000                   | 0.003246     | 0.006049   | 0.537   | 0.59193  |
| (c) Results for the seasons                                  |              |            |         |          |
| Winter                 | 0.002951     | 0.002684   | 1.099   | 0.27110  |
| Spring                 | 0.006259     | 0.002183   | 2.867   | 0.00425  |
| Summer                 | 0.003324     | 0.001699   | 1.957   | 0.05069  |
| Fall                   | -0.00017     | 0.002026   | -0.084  | 0.9335   |
Table 8. Individual coefficients and standard errors from the AIC procedure for the OLS model for the nine years taken together.

| Coefficients: | Estimate    | Std. Error  | t value | Pr(>|t|)    |
|---------------|-------------|-------------|---------|------------|
| (Intercept)   | 4.6939478   | 0.0600439   | 78.175  | < 2e-16    |
| temp          | -0.006210   | 0.0037256   | -1.667  | 0.095618   |
| temp3         | -0.011454   | 0.0019930   | -5.748  | 9.95e-09   |
| dwp           | 0.0103844   | 0.0038818   | 2.675   | 0.007509   |
| dewp1         | -0.012278   | 0.0025730   | -4.772  | 1.91e-06   |
| wdsp          | 0.0114283   | 0.0069037   | 1.655   | 0.097949   |
| avg3apm10     | 0.0043830   | 0.0009852   | 4.449   | 8.94e-06   |
Figure 1. Boxplot of the seasonal PM10 levels: 1 (winter, December to February), 2 (spring, March to May), 3 (summer, June to August), 4 (fall, September to November)
Figure 2. Boxplot of the seasonal mortality levels for the death 3 category: 1 (winter, December to February), 2 (spring, March to May), 3 (summer, June to August), 4 (fall, September to November)
Figure 3. Boxplot of the seasonal mortality levels for the death 2 category: 1 (winter, December to February), 2 (spring, March to May), 3 (summer, June to August), 4 (fall, September to November)
Figure 4. Boxplot of the seasonal mortality levels for the death1 category: 1 (winter, December to February), 2 (spring, March to May), 3 (summer, June to August), 4 (fall, September to November)
Figure 5. Plot of the observed data (mortality level death3, PM10, temperature, and dew point temperature) for the period January to April, 1994.
Figure 6. Plot of the observed data (mortality level death3, PM10, temperature, and dew point temperature) for the period June to September, 1994.
Figure 7. Mortality level death3 for all nine years (1992 to 2000) of data with a cubic spline fit.
Figure 8. Mortality level death3 (counts) and PM10 (µg/m³) for all nine years (1992-2000) of data.
Figure 9. Model output with 95% confidence bands for the GLM run using all nine years (1992 to 2000) of data
Figure 10. Residuals from an OLS model run using all nine years (1992 to 2000) of data.
Figure 11. Autocorrelation plot based on the residuals for an OLS model run using all nine years (1992 to 2000) of data.
Figure 12. Observed mortality level death3 (red) data for the year 2000 and the predicted mortality level (black) for the same year.
Figure 13. Quantile regression plot for mortality level death3 and PM10 (covariate: avg3pm10) for tau values 0.1 to 0.9 by 0.1.