6-18-2007

# IDENTIFYING EFFECT MODIFIERS IN AIR POLLUTION TIME-SERIES STUDIES USING A TWO-STAGE ANALYSIS

Sandrah P. Eckel
*Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics*, seckel@jhsph.edu

Thomas A. Louis
*Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics*

# Identifying effect modifiers in air pollution time-series studies using a two-stage analysis

Sandrah P. ECKEL and Thomas A. LOUIS

Studies of the health effects of air pollution such as the National Morbidity and Mortality Air Pollution Study (NMMAPS) relate changes in daily pollution to daily deaths in a sample of cities and calendar years. Generally, city-specific estimates are combined into regional and national estimates using two-stage models. Our two-stage analysis identifies effect modifiers of the relation between single-day lagged $PM_{10}$ and daily mortality in people age 65 and older from the 50 largest NMMAPS cities. We build on the standard approach by "fractionating" city-specific analyses to produce month-year-city specific estimated air pollution effects (slopes) in Stage I. In Stage II, we identify potential effect modifiers via weighted regression and weighted regression trees with the estimated slopes as dependent variables and predictors such as temperature, relative humidity, CO, $NO_2$, $O_3$, $SO_2$, season, year, and other city-specific characteristics.

**Key words:** hierarchical models, interaction, $PM_{10}$, regression trees

## 1 Introduction

Quantification of the health effects of air pollution guides air pollution regulation and has a far-reaching impact on both industry and human health. Current United States Environmental Protection Agency (EPA) regulations under the Clean Air Act (CAA) include primary National Ambient

Sandrah P. Eckel is a Ph.D. candidate, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA (e-mail seckel@jhsph.edu). Thomas A. Louis is Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

Air Quality Standards (NAAQS) the "attainment and maintenance of which ... are requisite to protect the public health" with an "adequate margin of safety."(CAA, Section 109. (b)) Within the context of the NAAQS, the EPA regulates the levels of a set of "criteria" pollutants including $PM_{10}$, $PM_{2.5}$, carbon monoxide (CO), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$) and ozone ($O_3$). $PM_{10}$ ($PM_{2.5}$) is defined to be particulate matter (PM) of aerodynamic diameter less than or equal to 10 (2.5) microns ($\mu$m). The health effects of PM are thought to be determined by both size and composition, with $PM_{2.5}$ considered more harmful than larger PM since $PM_{2.5}$ is small enough to penetrate deeply into the respiratory system. Each of the gaseous pollutants regulated by the EPA has been shown to have negative effects on human health. For example, the presence of CO at any level in the human bloodstream reduces the ability of the blood to carry oxygen which can lead to severe effects in those with cardiovascular disease (American Heart Association 2006).

Many studies have shown a statistically and scientifically significant relation between acute exposure to ambient particulate matter air pollution and mortality (Samet et al. 2000a; Katsouyanni et al. 2001). Statistical research on the association between acute exposure to ambient PM and mortality is challenging because we must disentangle the relatively small effects of PM from the large effects of confounding variables, such as seasonality, in the presence of other potential confounding gaseous pollutants. Previous work has shown that the positive association between $PM_{10}$ and mortality is not solely due to confounding by one of these gaseous co-pollutants (Dominici et al. 2005). A consensus has yet to be reached on how gaseous co-pollutants and other variables interact with $PM_{10}$ in the $PM_{10}$-mortality association. A recent study of 28 European cities from APHEA2 has found interactions in the $PM_{10}$-mortality relation amongst the elderly due to average level of $NO_2$, temperature, relative humidity, age standardized mortality rate, the proportion of the population older than 65 years of age and geographic region (Aga et al. 2003). Other studies of APHEA2 data found geographic region, temperature level, mean level of $NO_2$ and a city-specific age-standardized mortality rate, and percentage of population over age 65 to be effect modifiers, with $NO_2$ being the most important effect modifier as well as a confounder (Samoli et al. 2005;

Katsouyanni et al. 2001). In studies of U.S. air pollution, mean levels of $NO_2$ and ozone have not been found to change the $PM_{10}$-mortality relation (Samet et al. 2000a; Dominici et al. 2005). The discrepancy in the European and American findings may be explained by the differences in $NO_2$ concentration and sources since $NO_2$ is a better indicator for pollution due to vehicle emissions in Europe than in the U.S. (WHO 2003). A study of the $PM_{10}$-mortality relation in 10 U.S. cities found no evidence for effect modification due to season, other pollutants, or city-level socioeconomic status indicators such as unemployment rate, percentage population below the poverty line, percentage population with a college degree, or percentage nonwhite (Schwartz 2000). Two-stage models have previously been applied to 20 of the largest NMMAPS cities to identify effect modifiers and found weak evidence for effect modification due to changes in summer ozone levels (Samet et al. 2000a).

Work remains to be done to systematically study effect modifiers of the association between ambient PM and mortality to enhance our understanding of the PM-mortality relation and to produce information for the development of air pollution regulations that more effectively protect public health. In this paper, we present and apply a method to identify potential effect modifiers of the $PM_{10}$-mortality relation. In Section 2 we describe our data and summarize our two-stage analysis. Section 3 presents our Stage I analysis and summarizes the resulting data. Section 4 presents the Stage II analysis framework and results. Finally, Section 5 briefly summarizes and discusses our results and offers directions for future work. In an Appendix, we prove a sufficient condition to avoid identifying spurious effect modifiers in the two-stage analysis framework.

## 2 Data and Analysis Plan

We use data from NMMAPS, a multi-center study of 108 U.S. cities over the 14 year period from 1987-2000. Cause-specific mortality data were obtained from the National Center for Health Statistics that include information on mortality due to all-causes (non-accidental), cardiovascular

disease (CVD), respiratory disease, chronic obstructive pulmonary disease (COPD), pneumonia, and accidents. The weather information in NMMAPS obtained from the National Weather Service includes daily measurements of temperature, dew point temperature and relative humidity. Air pollution data from the EPA includes measurements of $PM_{10}$, $PM_{2.5}$, $O_3$, $NO_2$, $SO_2$, and CO. Following the EPA's regional partitioning of U.S. for the NAAQS, NMMAPS cities are grouped into eight geographic regions including the Industrial Midwest, Northeast, Northwest, Other (Alaska and Hawaii), Southern California, Southeast, Southwest and Upper Midwest. Population information for each metropolitan area was obtained from the 1990 and 2000 U.S. Census. To facilitate the reproducibility of statistical results, the data for NMMAPS are publicly available at the Internet-based Health & Air Pollution Surveillance System web site, http://www.ihapss.biostat.jhsph.edu/ and through the R package `NMMAPSdata` (Peng and Welty 2004). Previous work on 90 of the NMMAPS cities has found that, nationally, a 10 $\mu g/m^3$ increase in $PM_{10}$ is associated with a 0.21% increase in total mortality from non-external causes (Dominici et al. 2005). A weighted mean of city-specific estimates from the 50 largest NMMAPS cities (those cities to be used in our analysis) results in an estimate that, nationally, a 10 $\mu g/m^3$ increase in lag 1 $PM_{10}$ is associated with a 0.32% total increase in all-cause mortality excluding accidents.

The main purpose of our work is to conduct a two-stage statistical analysis of the NMMAPS data to more comprehensively identify potential effect modifiers in the relation between lagged $PM_{10}$ and non-accidental mortality in the U.S. population over the age of 65. We analyze data from the 50 largest NMMAPS cities with reported measurements of $PM_{10}$ since larger cities tend to have more complete pollutant data and higher daily mortality counts. The adverse health effects of air pollution are more pronounced in susceptible populations such as older adults, so we restrict our analysis to mortality in individuals over the age of 65. We consider $PM_{10}$ in lieu of $PM_{2.5}$ because $PM_{2.5}$ levels were generally not recorded from 1987-2000. Note that due to regional variation in the composition of PM, $PM_{10}$ is an "imperfect surrogate" for $PM_{2.5}$ (Samet et al. 2000a) although, by definition, studies of the health effects of $PM_{10}$ inherently include the health effects of $PM_{2.5}$ .

In the following calculations we used a lag of 1 for $PM_{10}$ measurements, although in future work, we might choose to use a lag of 0 or 2. In our analysis, we convert measured $PM_{10}$ to the units 10 $\mu g/m^3$ as is the standard in statistical analyses of the effects of PM. Throughout the rest of the paper, when we refer to PM we are implicitly referring to $PM_{10}$ .

Stage I of our analysis consists of city-specific log-linear regressions producing estimates of the month-year-city specific effects of changes in $PM_{10}$ on all-cause non-accidental mortality. In Stage II, we use the Stage I estimated month-year-city specific effects as the dependent variable in both a linear regression and a regression tree with a variety of predictor variables as possible effect modifiers of interest. Our set of possible effect modifiers include gaseous co-pollutants, weather variables, seasonality, year, city-level characteristics and functions of the co-pollutant and weather variables that serve to increase the dimensionality of our data.

# 3 Stage I Analysis

We fit the following overdispersed log-linear model for each of the 50 largest NMMAPS cities:

$$\begin{aligned} \log(E[Y_{imtj}]) &= \alpha_{0j} + \beta_{mtj}\text{PM}_{i-d,mtj} \\ &\quad + \alpha_{1j}ns_j(\text{time, df} = 7/\text{yr}) \\ &\quad + \alpha_{2j}ns_j(\text{temp, df} = 6) + \alpha_{3j}ns_j(\text{ dptp, df} = 3) \\ &\quad + \alpha_{4j}\text{dow}_{imtj} + \alpha_{5j}\text{agecat}_{imtj} \end{aligned} \tag{1}$$

where $Y_{imtj}$ is the count of non-accidental deaths on day $i$, month $m$, year $t$, city $j$ and $PM_{10}$ is measured in 10 $\mu g/m^3$ with a lag of $d = 1$. The degrees of freedom for the natural splines of time, temperature and dewpoint temperature were based on previous work on NMMAPS times series analysis (Samet et al. 2000b; Peng, Dominici, and Louis 2006; Welty and Zeger 2005).

A standard NMMAPS analysis would use the overdispersed log-linear model above, but instead

of month-year-city slopes on $PM_{10}$, we would use a city-specific slope on $PM_{10}$ (see Table 2). By including city specific intercepts, $\alpha_{0j}$, and month-year-city specific slopes on $PM_{10}$, $\beta_{mtj}$, in our models, we are "fractionating" the city-specific data into month-year-city strata. Our parameter of interest, $\hat{\beta}_{mtj}$, can be interpreted as the within month-year-city stratum effect of changes in $PM_{10}$ on expected mortality, having controlled for smooth functions of time, temperature and dew point temperature as well as day of the week and age category. More precisely, within a month-year-city stratum $mtj$, a 10 $\mu g/m^3$ increase in lag 1 $PM_{10}$ on day $i$ would result in $\exp(\hat{\beta}_{mtj})$ times the expected mortality on day $i$ without the increase in $PM_{10}$. In our Stage II analysis we use the $\hat{\beta}_{mtj}$ as dependent variables with a variety of predictor variables, each being a potential effect modifier. Some might argue that it is necessary to control for other confounding effects in the Stage I analysis, such as the confounding due to co-pollutants or other variables considered as potential effect modifiers. We wait until Stage II to address these effects to avoid overfitting and to leave more information in the $\hat{\beta}_{mtj}$. It might be the case that predictors identified in Stage II as effect modifiers are not truly effect modifiers, but spuriously appear as such in Stage II because we did not account for their confounding effect in Stage I. However, we have developed a general constraint on the type of summary statistic used for Stage II predictor variables that precludes the possibility of spurious effect modification from Stage I confounders (see Appendix for further details).

The traditional approach to studying effect modification in regression is to include interaction terms in a model and then to test whether these terms are statistically significant predictors of the outcome of interest. Stage I of our analysis produces a large number of $(\hat{\beta}_{mtj}, \mathrm{SE}_{\hat{\beta}_{mtj}})$ pairs that measure the effect of changes in $PM_{10}$ on expected mortality. Any variable that is found to be an important predictor of $\hat{\beta}_{mtj}$ in Stage II is an effect modifier of the $PM_{10}$-mortality relation. Figure 1 reveals no trend in the estimated month-year-city $PM_{10}$ effects, $\hat{\beta}_{mtj}$, as a function of $\log(\mathrm{SE}_{\hat{\beta}_{mtj}})$. As expected, the spread of the estimated month-year-city $PM_{10}$ effects increases with the standard error. Note however, that the $\hat{\beta}_{mtj}$ are highly variable, so the Stage II analysis must be weighted.

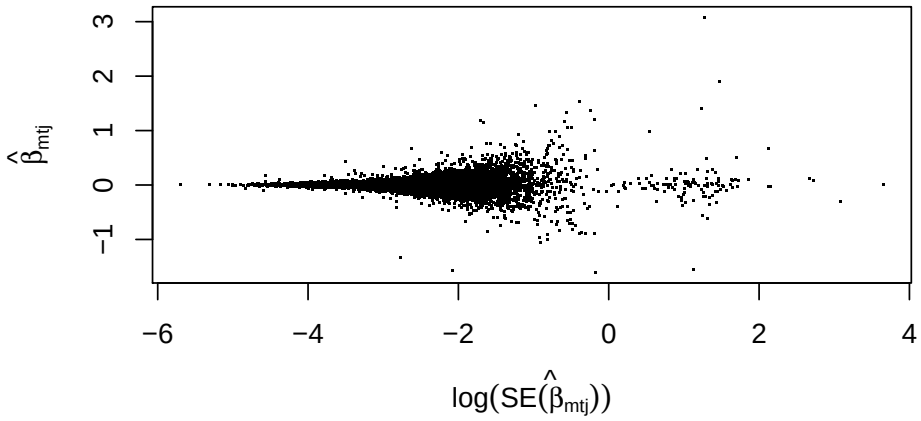


Figure 1: *The estimated month-year-city effect of $PM_{10}$, $\hat{\beta}_{mtj}$, represents the change in city-specific expected non-accidental mortality associated with a $10\mu/m^3$ increase in $PM_{10}$ in month m, year t and city j. This graph excludes 4 extreme $\hat{\beta}_{mtj}$ with values of $\hat{\beta}_{mtj}$ less than $-20$ and/or standard errors greater than* 150.

In theory, we have 8400 pairs of $(\hat{\beta}_{mtj}, \mathrm{SE}_{\hat{\beta}_{mtj}})$ since we have 14 years of monthly data for 50 cities ($14 \times 12 \times 50 = 8400$), however due to missing pollutant data we use a total of 3805 $\hat{\beta}_{mtj}$. The region specific percentage of complete month-year-city strata for data used the Stage II analysis are 48.6% for the Industrial Midwest, 60.4% for the Northeast, 19.9% for the Northwest, 77.1% for Southern California, 48.8% for the Southeast, 31.3% for the Southwest and 35.4% for the Upper Midwest.

To compare the within city variability of the $\hat{\beta}_{mtj}$ to the between city variability, while taking into account the $\mathrm{SE}_{\hat{\beta}_{mtj}} \equiv \hat{\sigma}_{mtj}$, we define and run the following hierarchical model in WinBUGS:

$$\begin{aligned}
\hat{\beta}_{mtj}|\beta_{mtj}, \hat{\sigma}^2_{mtj} &\sim N(\beta_{mtj}, \hat{\sigma}^2_{mtj}) \\
\beta_{mtj}|\mu_j, \sigma^2_{\text{within}} &\sim N(\mu_j, \sigma^2_{\text{within}}) \\
\mu_j|\theta, \sigma^2_{\text{between}} &\sim N(\theta, \sigma^2_{\text{between}})
\end{aligned}$$

We estimate the between-city variance in $(\hat{\beta}_{mtj}, \mathrm{SE}_{\hat{\beta}_{mtj}})$ to be ${}_{1.517\times10^{-6}}1.716\times10^{-4}_{3.074\times10^{-4}}$ and the within-city variance to be ${}_{8.573\times10^{-7}}1.259\times10^{-4}_{1.702\times10^{-4}}$. Hence we observe an approximately even split in the between-city versus within-city variability, with approximately 58% of the total variability in the $(\hat{\beta}_{mtj}, \mathrm{SE}_{\hat{\beta}_{mtj}})$ stemming from between-city variability.

# 4 Stage II Analysis

In Stage II, we perform a weighted analysis with the $\hat{\beta}_{mtj}$ as our outcome and a set of predictor variables we consider to be potential effect modifiers of the $PM_{10}$-mortality relation. We initially consider the potential for effect modification from: mean temperature, mean dew point temperature, mean relative humidity, trimmed mean $PM_{10}$, CO, $NO_2$, $O_3$, and $SO_2$ as well as indicators for season (warm season (May-October) vs. indoor heating season (November-April)), year and city-level characteristics such as EPA regulatory region (Upper Midwest, Industrial Midwest, Northeast, Southeast, Southwest, Northwest, Southern California, and Other) as well as, based on the 2000 census, proportion of the population ages 65 and older, proportion unemployed, proportion in poverty, proportion having earned a degree, and proportion non-white. We carefully select functions of a subset of these variables to include as potential predictors in our Stage II analysis.

## 4.1 Defining Stage II predictor variables

The dependent variables for the Stage II analysis, $\hat{\beta}_{mtj}$, are estimates of month-year-city specific $PM_{10}$ effects so our Stage II predictors, or candidate effect modifiers, must also be month-year-city specific summaries. We consider three such types of predictors: month-year-city means, mean tertiles, and covariance of tertiles indicators.

### 4.1.1 Measures of magnitude: the mean

Our first set of Stage II predictors are a standard month-year-city summary designed to measure magnitude. This set of variables consists of month-year-city means of the co-pollutant and weather variables. For example, for daily trimmed mean ozone levels, we define the month-year-city mean as $\overline{\mathrm{O}}_{3\cdot mtj} = Avg(\mathrm{O}_{3imtj})$ in month $m$, year $t$, city $j$.

### 4.1.2 Measures of deviation: tertile indicator variables

We increase the dimensionality of our data and potentially reduce excess variability due to the large margin of error in pollutant measurements by considering a "tertile" function of each co-pollutant or weather variable that measures daily level relative to its typical value for that time period. In relation to our example of ozone, tertile indicators address how the level of ozone in January 2000 in city $j$ is different from the level of ozone in all recorded Januaries in city $j$. For a given weather or co-pollutant variable, we pool daily measurements by month-city stratum, stratify into tertiles and then categorize each daily value as being either low, middle or high (-1, 0 ,1) in relation to the tertile in which the daily value falls. The tertile indicator function $T$ denotes this categorization. Hence $T_{\mathrm{O}_{3imtj}} = -1$ indicates the ozone on day $i$, month $m$, year $t$ and city $j$ falls in the lowest tertile of ozone for city $j$ in month $m$. We denote the month-year-city specific average tertile indicator of ozone by

$$\overline{T}_{\mathrm{O}_{3\cdot mtj}} = Avg(T_{\mathrm{O}_{3imtj}}) \text{ in month } m, \text{ year } t, \text{ city } j.$$

### 4.1.3 Interactions: Covariance of tertile indicators

The third form of Stage II predictors is the estimated month-year-city covariance between tertile indicators. The covariance of tertile indicators is a measure of the association between the tertile transformation of the two variables within a month-year-city period. For example, the estimated

month-year-city specific covariance of the tertile indicators for ozone and $PM_{10}$ is:

$$\hat{Cov}\left[T_{O_3}, T_{PM_d}\right]_{\cdot mtj} = Avg\left[T_{O_{3imtj}} \times T_{PM_{i-d,mtj}}\right] - Avg\left[T_{O_{3imtj}}\right] \times Avg\left[T_{PM_{i-d,mtj}}\right]$$

where we average over all days within the given month $m$, year $t$, city $j$ stratum. We omit the "hat" to simplify our notation of the estimated covariance. Due to the restriction of tertile indicators to $[-1,1]$, the covariance is bounded by -1 and 1. A relatively large positive value of $Cov\left[T_{O_3}, T_{PM_d}\right]_{\cdot mtj}$ implies that $O_{3imtj}$ tends lie in the same tertile as $PM_{i-d,mtj}$ and a negative value of $Cov\left[T_{O_3}, T_{PM_d}\right]_{\cdot mtj}$ relatively close to $-1$ implies that $O_{3imtj}$ tends to lie in the opposite tertile as $PM_{i-d,mtj}$. For example, a relatively large positive value of $Cov\left[T_{O_3}, T_{PM_d}\right]_{\cdot mtj}$ implies that if the level of ozone on a certain day $i$ in month $m$, year $t$, city $j$ is high with respect to the levels of ozone in month $m$ in city $j$ over all the reported years, the level of $PM_{10}$ on day $i$ in month $m$, year $t$, city $j$ will also be high with respect to the other days in the same month $m$ in city $j$ over all the observed years.

### 4.1.4 Final Set of Stage II predictors

Exploratory analysis of the correlations between month-year-city means of our weather and co-pollutant variables of interest, as shown in Figure 2, leads us to exclude dewpoint temperature from the Stage II analysis because of high correlation between mean dewpoint temperature and mean temperature. The remaining weather variables: mean temperature and mean relative humidity, are only slightly correlated between month-year-city strata. The two variable pairs: mean $NO_2$ and mean CO, and proportion unemployed and proportion in poverty are relatively highly correlated, but we include both pairs of variables in our Stage II analysis and take this correlation into account when interpreting our results. Instead of using regional indicators that serve as a surrogate for unmeasured variables that differentiate regions, we include weather, co-pollutant and census information that explain many of the regional differences. Our final set of 42 predictor

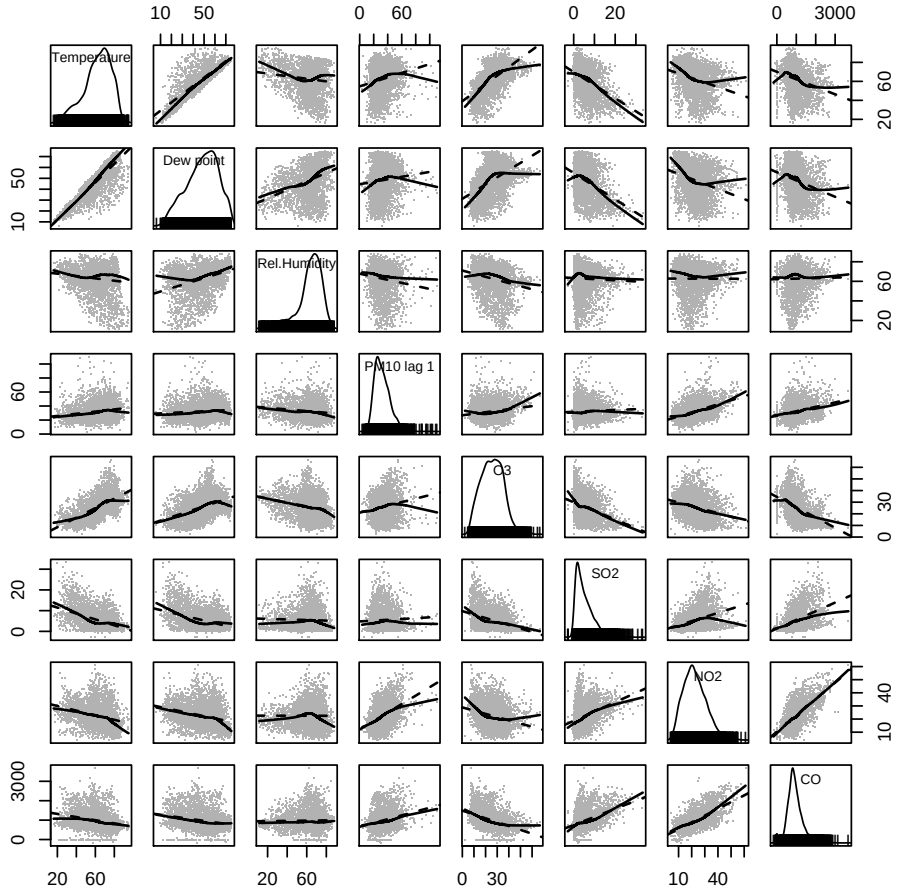


Figure 2: *Scatterplot matrix comparing month-year-city mean weather and pollutant variables for the 50 largest NMMAPS cities.*

variables in Stage II are functions of: mean temperature, mean relative humidity, trimmed mean $PM_{10}$, CO, $NO_2$, $O_3$, and $SO_2$ as well as indicators for season (warm season (May-October) vs. indoor heating season (November-April)), year and the city-level characteristics: proportion of the population ages 65 and older, proportion unemployed, proportion in poverty, proportion having earned a degree, and proportion non-white.

Our scientific question calls for a Stage II model that takes advantage of the large number of $(\hat{\beta}_{mtj}, \mathrm{SE}_{\hat{\beta}_{mtj}})$, is interpretable, and potentially allows for non-traditional forms of interaction. The two-stage framework allows for a multitude of Stage II analytic techniques. We focus on two

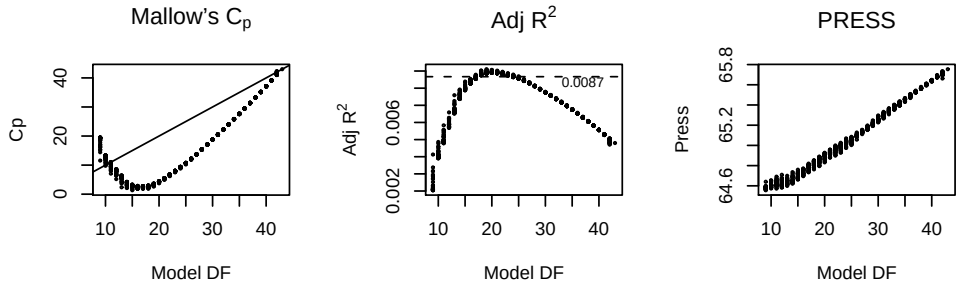


Figure 3: *Model selection criteria for all-subset Stage II weighted linear regression.* methods: weighted linear regression and weighted regression trees.

## 4.2 Weighted Linear Regression

As a candidate method for the Stage II analysis, weighted linear regression offers a straightforward interpretation of regression coefficients. However, weighted linear regression does not readily identify non-traditional higher-level interactions because the interactions must be pre-specified. The covariance of tertile indicators is one form of higher level interaction that can be pre-specified in the regression. Model selection poses a non-trivial problem with regression. We perform an all-subset inverse variance weighted linear regression with 3 model selection criteria: PRESS, Mallow's Cp and adjusted $R^2$ (see Figure 3). The adjusted $R^2$ criterion returns the most interesting models in terms of the number of predictor variables. We present a model chosen, according to maximum adjusted $R^2$, from amongst those models that included the "main effect" mean tertile indicator for any covariance of tertile indicators included in the model. This prevents a change in the basic findings of the model given a reparametrization of the predictor variables.

Figure 4 displays results from the weighted linear regression with 18 predictors. We find several predictors to be statistically significant, most notably $T_{SO_2}$, cov($T_{\text{temperature}}$, $T_{NO_2}$), $T_{\text{relative humidity}}$, mean temperature, proportion of the population age 65 and older and proportion of the population that is non-white. Both the month-year-city mean $T_{\text{temperature}}$ and the month-year-city mean temperature are chosen by the all-subset regression as important predictors

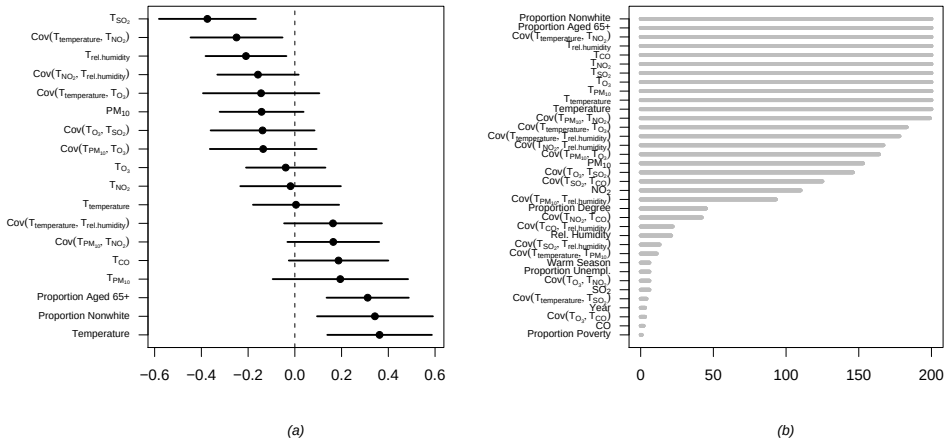


Figure 4: *(a) Results from the Stage II weighted linear regression. Estimated percent change in the effect of lag 1 $PM_{10}$ on expected mortality associated with an interquartile change in the predictor where "effect of lag 1 $PM_{10}$" is defined to be the expected estimated effect of a 10 $\mu g/m^3$ increase in lag 1 $PM_{10}$ within a month-year-city stratum on expected mortality, or $e^{E(\hat{\beta}_{mtj})}$. (b) Frequency of Stage II covariate inclusion in the 200 models with largest values of adjusted $R^2$. Cov($T_{O_3}$, $T_{relative\ humidity}$), cov($T_{PM_{10}}$, $T_{CO}$), cov($T_{temperature}$, $T_{CO}$), cov($T_{SO_2}$, $T_{NO_2}$), cov($T_{PM_{10}}$, $T_{SO_2}$) and mean $O_3$ were excluded from (b) because each of these covariates did not appear in any of the 200 models.*

for the model. Mean temperature, a measure of the magnitude of temperature, is statistically significant, whereas the $T_{temperature}$, a measure of the deviation in temperature from typical levels, is not found to be statistically significant. In our initial exploratory data analysis for the set of Stage II predictors, we found that $T_{CO}$ and $T_{NO_2}$ had a relatively high correlation of 0.39, compared to the other pairs of predictors. Both $T_{CO}$ and $T_{NO_2}$ are included in our final weighted linear regression model. Increases in $T_{CO}$ are (not statistically significantly) associated with increases in the month-year-city PM effect while increases in $T_{NO_2}$ are (not statistically significantly) associated with slight decreases in the month-year-city PM effect. Due to concerns about collinearity between these two predictors, we ran a model where we left out the $T_{CO}$ predictor, and found that $T_{NO_2}$ remains at a similar level of non-significance, although the estimated effect of $T_{NO_2}$ changes directions. We include functions of both $NO_2$ and CO in our final model since $NO_2$ has been found to be an effect modifier in previous studies and excluding functions of CO does not

lead to a dramatic change in our findings. The weighted linear regression explains a very small proportion of the variability in our data (adjusted $R^2 = 0.009$) but does capture some signal in our highly variable data (overall F-test p-value $< 0.0001$). A likelihood ratio test of the null hypothesis that the 12 predictors that are non-statistically significant in our final model have coefficients equal to zero produces a p-value of 0.02 .

We select the final weighted linear regression model to have maximal adjusted $R^2$ from an all-subset regression. Many competing models had similar values of adjusted $R^2$. To address our somewhat arbitrary model choice, we select the 200 models with largest adjusted $R^2$ (greater than 0.0087) and plot the frequency of inclusion of Stage II covariates in these models (Figure 4). Note that the main effect tertile indicators were forced in all models so that if a covariance of tertile indicators term was included, the main effect tertile indicators would also be present.

We find that, in addition to the mean tertile indicators, all 200 of these models include the covariates cov($T_{\text{temperature}}$, $T_{NO_2}$), proportion of the population age 65 and older, proportion of the population that is non-white and mean temperature, each of which was found to be statistically significant in the model with maximum adjusted $R^2$. Although $NO_2$ and CO were relatively highly correlated in our exploratory analysis, $NO_2$ is included in 110 of the 200 models, while CO is included in only 2 models. Similarly, although proportion nonwhite and proportion in poverty are relatively highly correlated, proportion nonwhite appears in all of the 200 models while proportion in poverty appears in none of the models.

## 4.3 Weighted Regression Trees

Regression trees are highly interpretable and naturally identify non-traditional interactions. For our second approach to the Stage II analysis, we use the R (R Development Core Team 2007) implementation of weighted regression trees in the `rpart`, or Recursive PARTitioning package (Therneau and Atkinson 2006). Trees tend to identify and isolate outliers by placing them in small

nodes. This ability to set apart extreme values gives the tree method an advantage over linear regression due to smaller potential outlier induced model distortion (Breiman, Friedman, Olshen, and Stone 1984). Since our data are highly variable and contain several extreme values, we specify a small minimum number of values per node (2) for attempting a split at that node and we allow the smallest node to have just one value. We weight by inverse variance, so any outliers that are set aside in terminal tree nodes with few observations are outliers in the sense of a value being extreme after having been divided by the estimated standard error of this value.

### 4.3.1 Tree with the $\hat{\beta}_{mtj}$ as the dependent variable

Analogous to the weighted linear regression, we consider the $E(\hat{\beta}_{mtj})$ as a function of Stage II candidate predictor variables, including month-year-city specific means, mean tertile indicators and the covariance of tertile indicators. We weight by inverse estimated Var $\left(\hat{\beta}_{mtj}\right)$ to account for the different levels of uncertainty in our estimates of $\beta_{mtj}$. We initially considered two sets of predictors for building our weighted regression trees. We first used the same set of 42 covariates used in the all-subset weighted linear regression (see Section 4.1.4). However, since the covariance of tertile indicators is a form of higher level interaction, we also considered using the set of predictors from Section 4.1.4 but excluding covariance of tertiles indicators. The latter approach gives the regression tree the freedom to construct higher-level interactions without beforehand specifying some of the forms of these interactions whereas the former approach provides the regression tree with more opportunities for splits, and hence potentially produces a tree with better predictive accuracy. Results from either tree can be used to identify functions of predictors of interest that could be used as effect modifiers in future modeling of the $PM_{10}$-mortality relation. We report results based on trees built using all 42 covariates as potential splitting variables because the resulting trees tend to be very similar when using either set of potential splitting variables.



In terms of the tuning parameters for tree building, we keep the `rpart` default setting of 10-

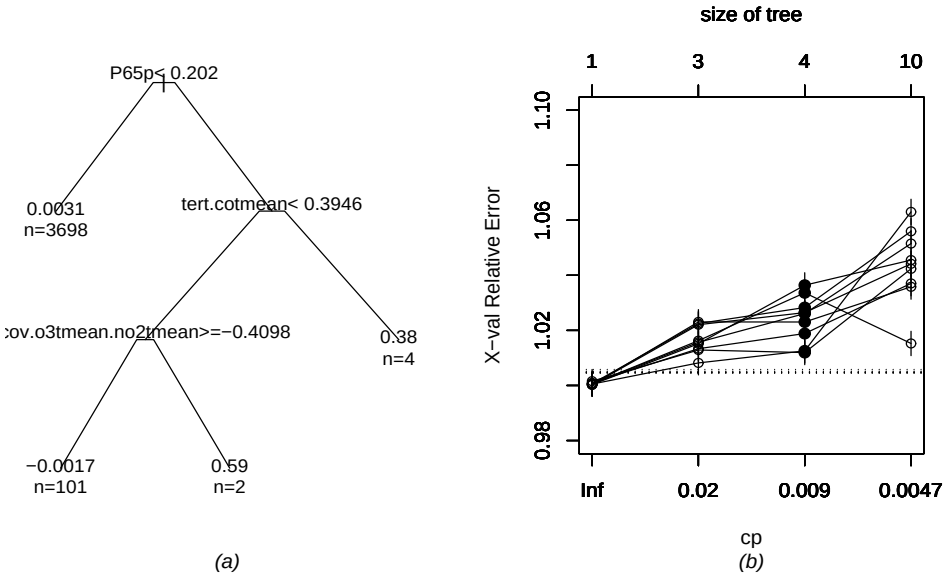


Figure 5: *(a) Weighted regression tree with* $\hat{\beta}_{mtj}$ *as the dependent variable and the* 42 *covariates used in the all-subset weighted linear regression as potential splitting variables. (b) The cross-validated relative error versus model complexity parameter (cp) under* 9 *different seeds. The solid dot represents the tree shown in (a).*

fold cross validation. We run `rpart` with 9 different seeds to inform our choice of tree size while taking into account the randomness inherent in the splitting of the data into 10 groups associated with 10-fold cross-validation (see Figure 5 (a)). The complexity parameter (cp), which determines tree size, is typically chosen to minimize the cross-validated relative error. We have not found cross-validation to be particulary informative for identifying models for our highly variable data. For example, in the all-subset weighted linear regression approach to Stage II, the cross-validation based PRESS model selection criteria did not suggest any interesting models. Part (b) of Figure 5 reveals how the high degree of variability of our data precludes us from choosing an interesting tree (i.e., consisting of many splits) based on cross-validated estimates of the tree's predictive accuracy. The nested subsequences of trees have few to no trees that abide by the +1SE cross-validation based rule for choosing the "right size" tree. Part (a) of Figure 5 shows the best tree according to cp $= 0.016$ which is slightly larger than the "right sized" trees according to standard criteria. It is not uncommon in regression tree analysis of associations to consider larger than usual trees in

order to identify a larger set of potential effect modifiers of interest.

The tree initially splits on the proportion of the population age 65 and older. The majority of the data (3698 of 3805 observations) has proportion age 65 and older less than 0.202 and hence falls in a terminal node with an assigned predicted value of 0.0031 . The only city with a proportion of older adults higher than 0.202 is St. Petersburg, Florida with a value of 0.226 . Hence in this case, the split on proportion of the population age 65 or older serves as a surrogate indicator for the city of St. Petersburg. The tree then splits on the month-year-city mean tertile indicator of CO, distinguishing 4 of the remaining observations, with $T_{CO}$ values greater than 0.3942, from the other 103 observations. The 4 observations with large mean tertile indicators of CO occur during summer months in St. Petersburg and have mean monthly temperature ranging from 80.9 to 83.7 degrees Fahrenheit. For these 4 month-year-city strata, the range in $T_{NO_2}$ is $-0.07$ to 0.30, while the range of $T_{CO}$ is much more restricted, at 0.47 to 0.55 . The final split sets aside another two outlying observations based on their values of the monthly covariance of $T_{O_3}$ and $T_{NO_2}$ .

The tree's second split is on the $T_{CO}$ . Recall that the $T_{CO}$ and $T_{NO_2}$ were both included in the final weighted linear regression model, but we did not find the $T_{CO}$ to be statistically significant. In the exploratory data analysis, the $T_{CO}$ and the $T_{NO_2}$ were correlated. We considered the possibility that this correlation induces collinearity in our linear model, masking the statistical significance of either variable when both are included in the model, although this did not appear to be the case. Collinearities do not pose a similar problem in regression trees, so the regression tree results may be identifying the relevance of the $T_{CO}$ variable that was missed in the linear regression. Of the four potential surrogate splits for the second split, none were functions of $NO_2$ .

### 4.3.2 Tree with the residuals from the linear model as the dependent variable

We next combine the parametric and non-parametric approaches by pouring residuals from the Stage II weighted linear regression down a weighted regression tree. By initially using the strengths

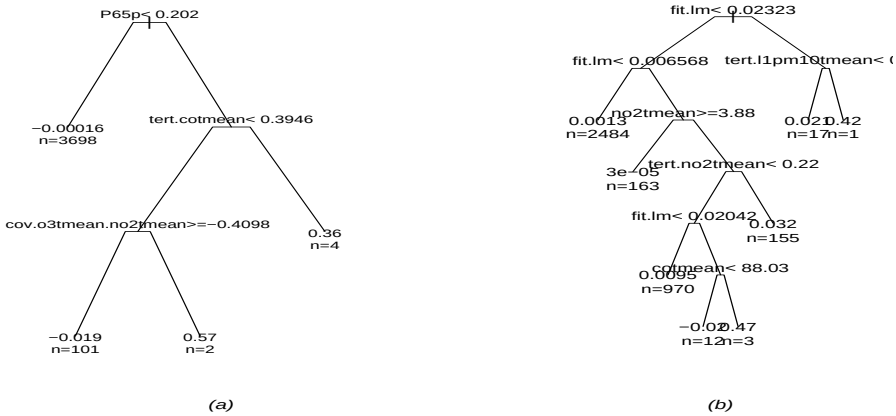


Figure 6: *(a) Weighted regression tree with weighted linear model residuals as the dependent variable and the set of* 42 *Stage II covariates as potential splitting variables. (b) Weighted regression tree with* $\hat{\beta}_{mtj}$ *as the dependent variable and the set of* 42 *Stage II covariates in addition to the predicted values from the weighted linear regression model as potential splitting variables.*

of the parametric approach in Stage II, i.e., the ability to pick out signals from potential effect modifiers in an interpretable manner, we can then use the strengths of the non-parametric regression tree to pick up any remaining signals in a non-traditional, flexible form. The weighted regression tree, displayed in Figure 6 (a), uses the residuals from the final Stage II weighted linear regression as the dependent variable and considers the same set of covariates used in the all-subset weighted linear regression as potential splitting variables. We weight by the inverse variance of the $\hat{\beta}_{mtj}$ .

The predicted values in the terminal nodes of the tree are noticeably smaller than those from Section 4.3.1 because, in this case, the outcome variable is the residuals from the linear regression of $\hat{\beta}_{mtj}$, not the $\hat{\beta}_{mtj}$ itself. Except for the predicted values at each node, the tree is identical to the tree obtained using $\hat{\beta}_{mtj}$ as the dependent variable. The non-parametric weighted regression tree identifies a different signal in the data than the parametric weighted linear regression.

### 4.3.3 Tree including the predicted values from the linear model as predictor variables

Finally, we consider a tree building approach where we use $\hat{\beta}_{mtj}$ as the dependent variable along with the same set of predictor variables we used in the Stage II top-level linear regression model in addition to the predicted values from the Stage II weighted linear regression. As displayed in Figure 6 (b), the tree splits first on the fitted values from the weighted linear regression. This confirms the utility of fitting the Stage II weighted linear regression model. However since the tree splits on variables besides the linear regression fit, we have evidence that some signal remains in the data after performing the weighted linear regression. The split on $\mathrm{T}_{\mathrm{PM}_{10}}$ singles out a specific month-year-city strata, September 1991 in St. Petersburg, Florida.

## 4.4 Comparison of Stage II analysis techniques

We compare two Stage II models, the weighted linear regression and weighted regression trees, that used $\hat{\beta}_{mtj}$ as a dependent variable and were weighted based on the uncertainty of $\hat{\beta}_{mtj}$. Our comparison criteria are root weighted mean squared error ($\mathrm{RMSE}_W$), a criterion that considers the uncertainty in our original estimates of $\beta_{mtj}$, as well as root unweighted mean squared error ($\mathrm{RMSE}_U$). We define the $\mathrm{RMSE}_W$ for each method and for a given outcome $Y_i$, $i = 1, \ldots, n$ to be

$$\mathrm{RMSE}_W = \left( \frac{\sum_{i=1}^{n} \frac{1}{Var(Y_i)} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} \frac{1}{Var(Y_i)}} \right)^{\frac{1}{2}}.$$

The weighted regression tree minimizes the $\mathrm{RMSE}_W$ and the $\mathrm{RMSE}_U$ criteria for predictive accuracy for the given data. While $\mathrm{RMSE}_U$ may appear to be an inappropriate criterion for model comparison because our models are weighted and the $\mathrm{RMSE}_U$ is not, but $\mathrm{RMSE}_U$ can be used as a model selection criteria in this case since both the linear regression and regression trees have already been selected as useful models based on criteria at the model selection stage for each type of model. An alternate approach could use model comparison criteria that use cross-validation or

give penalties for model complexity. The weighted residuals from the weighted linear regression and the weighted regression tree show similar patterns in the predicted values and outliers.

| | Method | $\text{RMSE}_W \times 10^3$ | $\text{RMSE}_U \times 10^3$ |
|---|---|---|---|
| 1 | Overall Mean | 38.5 | 130.0 |
| 2 | Weighted Linear Regression | 38.2 | 129.9 |
| 3 | Weighted Regression Tree | 37.2 | 129.7 |

Table 1: Comparing Stage II model fit for models using $\hat{\beta}_{mtj}$ as the dependent variable. In this case the weighted regression trees is that from Section 4.3.1, with the $\hat{\beta}_{mtj}$ as the dependent variable and the same set of potential predictor variables as we used in the weighted linear regression.

We perform Stage II model selection based on maximizing predictive ability, so we should use the Stage II results as a heuristic guide to understanding the data (Breiman et al. 1984). The two Stage II approaches provide complementary heuristic insights into the effect modifiers in the $PM_{10}$-mortality relation. Both the weighted linear regression and the regression tree identified the proportion of the population ages 65 and older and mean $T_{CO}$ as effect modifiers, although $T_{CO}$ was not significant in the linear regression. The regression tree results are more intuitively interpretable and produced a smaller $\text{RMSE}_W$ than the weighted linear regression, however they do not identify as large a set of effect modifiers. Due to the extreme variability of the data, the final regression trees were not amongst the allowable trees based on the predictive criteria of 10-fold cross-validated error and the linear model presented was not one of the better models according to the N-fold cross-validation based PRESS criteria. The reduced information provided by the regression trees may be due in part to the loss of power associated with using a non-parametric method as opposed to a parametric method. Since we are attempting to detect a signal of relatively small magnitude from a highly variable data set, the fact that we obtain similar results using both a parametric and a non-parametric technique gives more credence to our aforementioned predictors as effect modifiers. Regression trees are a valid approach to the analysis at hand, but should be presented in conjunction with a parametric analysis.

# 5 Discussion

This paper presents a modification of the standard two-stage method for identifying effect modifiers. In Stage I, we fractionate a parameter of interest from a city specific effect to month-year-city specific effect to increase the dimensionality of our data for Stage II. In Stage II, we use two standard modeling techniques to identify covariates that play an important role in predicting the fractionated effect estimates from Stage I, and hence identify a set of effect modifiers. We consider a set of covariates for Stage II that are functions of the standard set of variables.

We identified a set of potential effect modifiers the $PM_{10}$-mortality relation. In the final weighted linear regression model, we found a statistically significant association between $\hat{\beta}_{mtj}$ and $T_{SO_2}$, cov($T_{\text{temperature}}$, $T_{NO_2}$), $T_{\text{relative humidity}}$, month-year-city mean temperature, proportion of the population age 65 and older and proportion of the population that is non-white. All of the 200 weighted linear regressions with largest adjusted $R^2$ included the covariates cov($T_{\text{temperature}}$, $T_{NO_2}$), proportion of the population age 65 and older, proportion of the population that is non-white and mean temperature. By design, each of these 200 models also included the month-year-city mean tertile indicators of our weather and pollutant variables. Weighted regression trees, which singled out patterns in the data beyond those identified with the parametric weighted linear regression, found the proportion of the population age 65 and older and $T_{CO}$ to be predictive of the $\hat{\beta}_{mtj}$. These results serve as a heuristic to guide further research into modifiers of the $PM_{10}$-mortality relation.

The main idea behind our two-stage method is to use linear regression or regression trees to identify effect modifiers after allowing traditional times-series models do the rest of the work. Our focus is to understand modifiers of the relation between $PM_{10}$ and mortality so we value interpretability over predictive ability. One might suggest a Poisson form of CART as developed by Chaudhuri, Lo, Loh, and Yang (1995) instead of our two-stage approach. The resulting tree would

not build on standard NMMAPS models and hence would split on confounders such as weather, seasonality and time, making it difficult to pick out traditionally defined effect modifiers. Our two-stage model reduces the problem of interaction identification within the standard NMMAPS model to the simpler problem of variable selection in Stage II. In place of either weighted linear regression or weighted regression trees, we could use other standard analysis techniques for Stage II. We could gain additional information on predictor variable importance by using boosting or random forests. However there is currently no ready implementation of a regression-type weighting in the R function `randomForest` for random forests. There is an option to include appropriate weights in the `gbm` implementation of boosting in R. Alternatively, we could use Bayesian variable selection or any of the other multitude of methods applicable to Stage II.

In future methodological work, we will develop a set of guidelines for an appropriate level of fractionation (here we fractionated from city specific to month-year-city specific $PM_{10}$ effects). We will quantify the tradeoff between the increase in inherent variability due to fractionated estimates from smaller samples with the increase in structure that becomes visible with having a larger sample size of fractionated estimates from which to draw inferences. We could use an approach like Janes, Dominici, and Zeger (2007) to partition the Stage II covariate effects on nested timescales. For future work related to the air pollution application of our methodology, we will use different lags of $PM_{10}$ or a distributed lag approach and use hospitalizations for respiratory and cardiovascular reasons as outcomes.

# Acknowledgements

The work of Sandrah P. Eckel was supported by NIA grant number T32 AG00247, the Johns Hopkins Training Program in the Epidemiology and Biostatistics of Aging. We are grateful to Roger D. Peng, Holly Janes and the members of the Environmental Biostatistics and Epidemiology Group at the Johns Hopkins Bloomberg School of Public Health for their valuable comments and insight.

# References

Aga, E., Samoli, E., Touloumi, G., Anderson, H. R., Cadum, E., Forsberg, B., Goodman, P., Goren, A., Kotesovec, F., Kriz, B., Macarol-Hiti, M., Medina, S., Paldy, A., Schindler, C., Sunyer, J., Tittanen, P., Wojtyniak, B., Zmirou, D., Schwartz, J., and Katsouyanni, K. (2003), "Short-term effects of ambient particles on mortality in the elderly: results from 28 cities in the APHEA2 project." *European Respiratory Journal. Supplement*, 40, 28s–33s.

American Heart Association (2006), "Air pollution, heart disease and stroke." URL `http://www.americanheart.org/presenter.jhtml?identifier=4419`.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth, Inc.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., and Yang, C.-C. (1995), "Generalized regression trees," *Statistica Sinica*, 5, 641–666.

Dominici, F., McDermott, A., Daniels, M., Zeger, S. L., and Samet, J. M. (2005), "Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: mortality among residents of 90 cities." *Journal of Toxicology and Environmental Health A*, 68, 1071–1092, URL `http://dx.doi.org/10.1080/15287390590935932`.

Janes, H., Dominici, F., and Zeger, S. (2007), "Trends in particulate matter and mortality: an approach to the assessment of unmeasured confounding," *Johns Hopkins University, Dept. of Biostatistics Working Papers.*, Working Paper 104, URL `http://www.bepress.com/jhubiostat/paper104`.

Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., Tertre, A. L., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A., Anderson, H. R., Wojtyniak, B., Paldy, A., Braunstein, R., Pekkanen, J., Schindler, C., and Schwartz, J. (2001), "Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project." *Epidemiology*, 12, 521–531.

Peng, R. D., Dominici, F., and Louis, T. A. (2006), "Model choice in time series studies of air pollution and

mortality," *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 169, 179–203.

Peng, R. D. and Welty, L. J. (2004), "The NMMAPSdata package," *R News*, 4, 10–14, URL http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org. ISBN 3-900051-07-0.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., and Zeger, S. L. (2000a), "Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994." *New England Journal of Medicine*, 343, 1742–1749.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J., and Zanobetti, A. (2000b), "The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and mortality from air pollution in the United States." *Research Report/ Health Effects Institute*, 94, 5–70; discussion 71–9.

Samoli, E., Analitis, A., Touloumi, G., Schwartz, J., Anderson, H. R., Sunyer, J., Bisanti, L., Zmirou, D., Vonk, J. M., Pekkanen, J., Goodman, P., Paldy, A., Schindler, C., and Katsouyanni, K. (2005), "Estimating the exposure-response relationships between particulate matter and mortality within the APHEA multicity project." *Environmental Health Perspectives*, 113, 88–95.

Schwartz, J. (2000), "Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths." *Environmental Health Perspectives*, 108, 563–568.

Therneau, T. M. and Atkinson, B. (2006), *rpart: Recursive Partitioning*, URL http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm. R package version 3.1-32, R port by Brian Ripley.

Welty, L. and Zeger, S. (2005), "Are the acute effects of pm10 on mortality in nmmaps the result of inadequate control for weather and season? a sensitivity analysis using flexible distributed lag models." *American Journal of Epidemiology*, 162, 80–88.

WHO (2003), "Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Report

on a WHO Working Group. Bonn, Germany." URL http://www.who.dk/document/e79097.pdf.

# Appendix

## Sufficient conditions for no Stage II confounder effects

In the two-stage approach to identifying effect modifiers, we use the $\hat{\beta}_{mtj}$ produced in Stage I as dependent variables in Stage II, regressing on $j$-specific summaries of potential effect modifiers. Here, we find conditions that ensure a confounder that was not included in the Stage I regression will not be an effect modifier in the Stage II regression.

Consider the case where a covariate $Z$ confounds the relation between a predictor $X$ (a fixed, non-random vector) and the response $Y$. For example, consider the Stage I linear system for day $i$ and month $j$ :

$$\begin{aligned} \left[Y_{ij}|X_{ij},Z_{ij}\right] &= \alpha+\beta X_{ij}+\gamma Z_{ij}+\varepsilon_{ij} \quad (2)\\ \left[Z_{ij}|X_{ij}\right] &= \mu+\delta X_{ij}+\eta_{ij} \quad (3)\end{aligned}$$

where $\varepsilon_{ij} \sim$ i.i.d. $(0,\sigma^2)$ and $\eta_{ij}$ are exchangeable and independent of the $\varepsilon_{ij}$. Denote the vector of confounders for month $j$ by $\mathbf{Z}_j = (Z_{1j},\ldots,Z_{n_jj})$ and let $\overline{Z}_j = n_j^{-1}\sum_{i=1}^{n_j} Z_{ij}$ and $\eta_j = (\eta_{1j},\ldots,\eta_{n_jj})$.

The regression of $Y_{ij}$ on $X_{ij}$ without accounting for $Z_{ij}$ is:

$$\begin{aligned} E(Y_{ij}|X_{ij}) &= \alpha+\beta X_{ij}+\gamma(\mu+\delta X_{ij})\\ &= (\alpha+\gamma\mu)+(\beta+\gamma\delta)X_{ij} \equiv \alpha^*+\beta^* X_{ij}\end{aligned}$$

where $\alpha^* \equiv \alpha+\gamma\mu$ and $\beta^* \equiv \beta+\gamma\delta$ . When $\gamma\mu = 0 = \gamma\delta$ , $Z_{ij}$ is not a confounder because $Z_{ij}$ is either unrelated to $X_{ij}$ (when $\delta = 0$) or unrelated to $Y_{ij}$ (when $\gamma = 0$), and $E(Y_{ij}|X_{ij}) = \alpha^*+\beta^* X_{ij} = \alpha+\beta X_{ij}$ .

The fractionated Stage I regression of $Y_{ij}$ on $X_{ij}$ produces $\hat{\beta}_j^* \equiv \frac{\sum_i (Y_{ij}-\overline{Y}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2}$. From Equations 2 and 3 we have as inputs to Stage II :

$$\hat{\beta}_j^* = \beta + \gamma\delta + \gamma \frac{\sum_i (\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2} + \frac{\sum_i (\varepsilon_{ij}-\overline{\varepsilon}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2}.$$

The true model for the Stage II linear regression is determined by the conditional expectation of $\hat{\beta}_j^*$ with respect to $T(\mathbf{Z}_j)$, a scalar month-specific summary of the confounder :

$$\begin{aligned} E(\hat{\beta}_j^* \mid \mathbf{X}_j, T(\mathbf{Z}_j)) &= (\beta+\gamma\delta) + \gamma E\left[ \frac{\sum_i (\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2} \,\middle|\, \mathbf{X}_j, T(\mathbf{Z}_j) \right] \\ &\quad + E\left[ \frac{\sum_i (\varepsilon_{ij}-\overline{\varepsilon}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2} \,\middle|\, \mathbf{X}_j, T(\mathbf{Z}_j) \right] \\ &= (\beta+\gamma\delta) + \gamma E\left[ \frac{\sum_i (\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2} \,\middle|\, \mathbf{X}_j, T(\mathbf{Z}_j) \right] + 0 \qquad (4) \\ &= (\beta+\gamma\delta) + \gamma E\left[ R(\eta_j, \mathbf{X}_j) \mid \mathbf{X}_j, T(\mathbf{Z}_j) \right] + 0 \end{aligned}$$

with

$$R(\eta_j, \mathbf{X}_j) = \frac{\sum_i (\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)}{\sum_i (X_{ij}-\overline{X}_j)^2}.$$

The 0 in Equation 4 results from the independence of the $\varepsilon_{ij}$ and the $Z_{ij}$. Spurious confounding results from the second term depending on $j$. The following theorem gives a sufficient condition for $j$-independence in $E(\hat{\beta}_j^* \mid \mathbf{X}_j, T(\mathbf{Z}_j))$.



**Theorem 1:** If $E\left(\eta_{ij} - \overline{\eta}_j \mid \mathbf{X}_j, T(\mathbf{Z}_j)\right)$ does not depend on $i$, then $E[R \mid \mathbf{X}_j, T(\mathbf{Z}_j)] = 0$.

**Proof:**

$$\begin{aligned} E\left[R|\ \mathbf{X}_j, T(\mathbf{Z}_j)\right] &= E\left[\left.\frac{\sum_i(\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)}{\sum_i(X_{ij}-\overline{X}_j)^2}\right| \mathbf{X}_j, T(\mathbf{Z}_j)\right] \\ &= \frac{\sum_i E\left[\left.(\eta_{ij}-\overline{\eta}_j)(X_{ij}-\overline{X}_j)\right| \mathbf{X}_j, T(\mathbf{Z}_j)\right]}{\sum_i(X_{ij}-\overline{X}_j)^2} \\ &= \frac{\sum_i E\left[\left.(\eta_{ij}-\overline{\eta}_j)\right| \mathbf{X}_j, T(\mathbf{Z}_j)\right](X_{ij}-\overline{X}_j)}{\sum_i(X_{ij}-\overline{X}_j)^2} \end{aligned}$$

and by the condition of the theorem

$$\begin{aligned} &= \frac{E\left[\left.(\eta_{1j}-\overline{\eta}_j)\right| \mathbf{X}_j, T(\mathbf{Z}_j)\right]\sum_i(X_{ij}-\overline{X}_j)}{\sum_i(X_{ij}-\overline{X}_j)^2} \\ &= \frac{E\left[\left.(\eta_{1j}-\overline{\eta}_j)\right| \mathbf{X}_j, T(\mathbf{Z}_j)\right]\cdot 0}{\sum_i(X_{ij}-\overline{X}_j)^2} = 0. \end{aligned}$$

**Theorem 2:** Assume that for a fixed $j$ the $\eta_{ij}$ are exchangeable and that $T(\mathbf{Z_j})$ is permutation invariant, e.g., $T\left(Z_{\pi(1)j},\ldots,Z_{\pi(n_j)j}\right)$, is constant for all permutations $\pi$ of $1,\ldots,n_j$. Then, $E[R|\ \mathbf{X}_j, T(\mathbf{Z}_j)] = 0$.

**Proof:** It is straightforward to show that $[\eta_{1j},\ldots,\eta_{n_jj}|T(\eta_j)]$ is an exchangeable distribution and so, $\left[\eta_{1j},\ldots,\eta_{n_jj}|T(\eta_j)\right]$ does not depend on $i$. It follows that $E\left[\eta_{ij}|\mathbf{X}_j, T(\mathbf{Z_j})\right]$ and $E\left[\overline{\eta}_j|\mathbf{X}_j, T(\mathbf{Z_j})\right]$ do not depend on $i$ and Theorem 1 applies.

We have shown for a two-stage linear regression that using a permutation invariant $j$-specific scalar summary of a Stage I confounder as a Stage II predictor precludes the possibility that it will be a spurious effect modifier in Stage II. For example, we use the month-specific mean, $T(\mathbf{Z_j}) = \overline{\mathbf{Z}}_\mathbf{j}$, and so will not induce a spurious association between $T(\mathbf{Z_j})$ and $\hat{\beta}_j$. However, if we consider a confounder summary that is not permutation invariant, we may induce $j$-dependence.

For example, if $T(Z_j) = Z_{1j}$, then we have $i$-dependence and possibly $j$-dependence:

$$\begin{aligned} E\left(\eta_{ij} - \overline{\eta}_j |\, \mathbf{X}_j, Z_{1j}\right)\right) &= E\left(\eta_{ij} |\, \mathbf{X}_j, Z_{1j}\right) - E\left(\overline{\eta}_j |\, \mathbf{X}_j, Z_{1j}\right) \\ &= \begin{cases} Z_{1j} - \mu - \delta X_{1j} - \left[\frac{1}{n_j} Z_{1j} + \frac{1}{n_j}\sum_{i=2}^{n_j}(\mu + \delta X_{ij})\right] & \text{if } i = 1 \\ 0 - \left[\frac{1}{n_j} Z_{1j} + \frac{1}{n_j}\sum_{i=2}^{n_j}(\mu + \delta X_{ij})\right] & \text{if } i \neq 1\,. \end{cases} \end{aligned}$$

Theorem 2 extends to use of Generalized Linear Models (e.g., log-linear) in Stage I and linear regression in Stage II. Other sufficient conditions are available. For example, if Stage I is a weighted regression, then it is sufficient that $T$ be a weighted mean using the same weights as in Stage I. We leave a full consideration of these issues to future work.

## City-specific characteristics

| | City | $1000\times$ PM Estimate$_{SE}$ | % 65+ | % Non white |
|---|---|---|---|---|
| 1 | Los Angeles | $3.1_{2.3}$ | 9.7 | 51.4 |
| 2 | New York | $3.2_{3.4}$ | 12.0 | 52.6 |
| 3 | Chicago | $2.7_{1.1}$ | 11.7 | 43.7 |
| 4 | Dallas/Fort Worth | $3.5_{4.5}$ | 7.8 | 34.2 |
| 5 | Houston | $7.3_{3.2}$ | 7.4 | 41.4 |
| 6 | Phoenix | $0.3_{3.9}$ | 11.7 | 22.7 |
| 7 | Santa Ana/Anaheim | $2.4_{5.3}$ | 9.8 | 35.3 |
| 8 | San Diego | $6.0_{5.1}$ | 11.1 | 33.6 |
| 9 | Miami | $6.3_{7.1}$ | 13.3 | 30.3 |
| 10 | Detroit | $5.2_{1.6}$ | 12.1 | 48.4 |
| 11 | Seattle | $3.0_{2.6}$ | 10.5 | 24.4 |
| 12 | San Bernardino | $3.8_{4.9}$ | 8.5 | 41.3 |
| 13 | San Jose | $4.5_{4.2}$ | 9.5 | 46.4 |
| 14 | Minneapolis/St. Paul | $5.6_{2.6}$ | 11.1 | 20.4 |
| 15 | Riverside | $-2.1_{4.3}$ | 12.6 | 34.5 |
| 16 | Philadelphia | $1.8_{5.1}$ | 14.1 | 54.9 |
| 17 | Atlanta | $13.6_{8.1}$ | 8.3 | 57.3 |
| 18 | Oakland | $7.8_{8.1}$ | 10.2 | 51.3 |
| 19 | Denver | $4.2_{3.3}$ | 9.5 | 26.6 |
| 20 | Cleveland | $4.0_{1.8}$ | 15.6 | 32.6 |
| 21 | San Antonio | $-4.0_{9.3}$ | 10.4 | 31.1 |
| 22 | Las Vegas | $3.8_{2.7}$ | 10.7 | 28.3 |
| 23 | Pittsburgh | $3.2_{1.6}$ | 17.8 | 15.7 |
| 24 | Sacramento | $-1.0_{6.2}$ | 11.1 | 36.1 |
| 25 | Columbus | $7.6_{5.0}$ | 9.8 | 24.5 |
| 26 | Tampa | $13.3_{11.4}$ | 12.0 | 24.9 |
| 27 | Buffalo | $4.2_{9.9}$ | 16.0 | 17.7 |
| 28 | Milwaukee | $-0.2_{6.4}$ | 12.9 | 34.2 |
| 29 | St. Petersburg | $9.0_{9.4}$ | 22.6 | 14.1 |
| 30 | Kansas City | $2.3_{9.2}$ | 11.9 | 23.7 |
| 31 | Salt Lake City | $-2.5_{2.5}$ | 8.1 | 13.8 |
| 32 | Memphis | $0.9_{9.9}$ | 10.0 | 52.7 |
| 33 | Orlando | $2.4_{15.2}$ | 10.0 | 31.3 |
| 34 | Honolulu | $13.4_{12.5}$ | 13.5 | 78.8 |
| 35 | Indianapolis | $8.7_{5.7}$ | 11.1 | 29.6 |
| 36 | Cincinnati | $-4.9_{3.6}$ | 13.5 | 27.1 |
| 37 | Tucson | $-2.2_{4.6}$ | 14.2 | 25.0 |
| 38 | Austin | $19.0_{13.0}$ | 6.7 | 31.8 |
| 39 | Fresno | $7.0_{5.1}$ | 9.9 | 45.9 |
| 40 | Newark | $8.2_{8.1}$ | 11.9 | 55.5 |
| 41 | Jacksonville | $-7.1_{10.8}$ | 10.4 | 34.2 |
| 42 | San Francisco | $7.5_{7.4}$ | 13.8 | 50.4 |
| 43 | Worcester | $12.4_{11.3}$ | 13.0 | 10.4 |
| 44 | Rochester | $6.7_{12.8}$ | 13.0 | 21.1 |
| 45 | Tacoma | $7.3_{4.7}$ | 10.2 | 21.7 |
| 46 | Charlotte | $1.5_{12.2}$ | 8.5 | 36.0 |
| 47 | Louisville | $6.9_{10.0}$ | 13.5 | 22.7 |
| 48 | Boston | $13.9_{8.9}$ | 11.1 | 42.3 |
| 49 | El Paso | $0.3_{2.4}$ | 9.8 | 25.9 |
| 50 | Birmingham | $4.8_{3.1}$ | 13.7 | 41.8 |

Table 2: City-specific estimates for the effect of a 10 $\mu$g/m$^3$ increase in $PM_{10}$ along with city-specific information from the 2000 Census. The heading % 65+ refers to the percentage of the population age 65 or older and % Non white refers to the percentage of the population that is not white.