Assessing Geographic Heterogeneity and Variable Importance in an Air Pollution Data Set

S. Stanley Young* and Jessie Q. Xia

National Institute of Statistical Sciences, RTP, NC 27709, USA

Received 30 April 2013; revised 4 July 2013; accepted 8 July 2013
DOI:10.1002/sam.11202
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: In this article, we examine data on the relationship between air quality and mortality in the United States using a published observational data set. Observational studies are complex and open to various interpretations. We show that there is geographic heterogeneity for the effect of air pollution on longevity. We also show that the relative importance of air pollution on longevity is much less than that of income or smoking. Most often authors do not address the relative importance of variables under consideration, choosing instead to concentrate on specific claims of significance. Yet good policy decisions require knowledge of the magnitude of relevant effects. Our analysis uses three methods for determining variable importance, showing how this puts predictor variables into a context that supports sound environmental policymaking. In particular, using both regression and recursive partitioning, we are able to confirm a spatial interaction with the air quality variable PM2.5; there is no significant association of PM2.5 with longevity in the west of the United States. We also determine the relative importance of PM2.5 in comparison to other predictor variables available in this data set. Our findings call into question the claim made by the original researchers. © 2013 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 6: 375–386, 2013

Keywords: observational studies; variable importance; PM2.5; mortality; air quality

1. INTRODUCTION

Galileo's Revenge is an ironically titled book by Peter W. Huber [1] about junk science in the courtroom. Huber makes the point that it is relatively easy to fool juries and even judges. Here is the irony. Galileo (1564–1642) was a leader in the scientific revolution and championed the heliocentric universe of Copernicus. For his trouble, the Roman Inquisition, speaking for the society of the time, showed him the rack and asked him to recant the heliocentric universe. Huber seems to be saying that the tables are now turned and it is junk science that is punishing society. Cope and Allison [2] worry that White Hat Bias is now a threat to the integrity of science reporting. White Hat Bias is defined as 'bias leading to distortion of information in the service of what may be perceived to be righteous ends'. Many decisions need to be made in the analysis of an observational data set; there is rarely a simple path from data, through analysis, to a claim. Guided analysis and selective citation are examples of White Hat Bias.

* *Correspondence to:* S. Stanley Young (young@niss.org)

The current policy paradigm is that air pollution, as measured by small particles (those less than 2.5 μm or PM2.5), is killing people and it needs to be brought under further regulatory control. At one point or another the Environmental Protection Agency (EPA) and the California Air Resources Board (CARB) speak of thousands or more than 160,000 deaths attributable to PM2.5; see ref. 3. The EPA and CARB base their case on statistical analysis of observational data. But if that analysis is not correct, and small-particle air pollution is not causing excess statistical deaths, then the faulty science is punishing society through increased costs and unnecessary regulation.

Pope *et al*. [4] cited eight studies (references 4–11 in their paper), saying, 'Associations between long-term exposure to fine particulate air pollution and mortality have been observed ... more recently, in cohort-based studies.... all support the view that relatively prompt and sustained health benefits are derived from improved air quality'. These citations appear unbalanced. For example, Enstrom [5], after citing papers supporting an association says, 'Other cohort studies have also examined mortality associations with PM2.5 and other pollutants ... with somewhat different findings.' There were eight papers

that Pope *et al*. referred to supporting an association between pollution, PM2.5, and statistical deaths, and four papers that Enstrom referred to that cast doubt on the claim. Peng *et al*. [6] commented, 'For example, in air pollution epidemiology, the national relative risk of increased mortality is estimated to be 1.005 per 10 parts per billion of 24-hour ozone. Remarkably, an integrated analysis of mortality in 95 metropolitan areas can detect this signal, which translates into thousands of excess deaths per year given the universality of ozone exposure. Nevertheless, the potential for unexplained confounding cannot be denied for such a small relative risk'. For a review of animal and human studies see ref. 7.

When this controversy was breaking out in the early 1990s, the EPA asked the National Institute of Statistical Sciences to evaluate data from two cities, see ref. 8. They commented on some of the difficulties, saying 'The data used in the analyses (meteorological conditions, particulate levels, death counts) are observational; that is, data that are measured and recorded without control or intervention by researchers. Deducing causal relationships from observational data is perilous. A practical approach described by Mosteller and Tukey involves considerations beyond regression analysis. In particular, consideration should be given to whether the association between particulate levels and mortality is consistent across "settings," whether there are plausible common causes for elevated particulate levels and mortality, and whether the derived models reflect reasonable physical relationships.' They then concluded, '. . .that the reported effects of particulates on mortality are unconfirmed'. Essentially noting the same and additional difficulties, Smith *et al*. [9] agreed that the case for a significant association of low-level air pollution with statistical deaths was unproven, 'In summary, it is our view that estimates of the association between ozone and mortality, based on time-series epidemiologic analyses of daily data from multiple cities, reveal important still-unexplained inconsistencies and show sensitivity to modeling choices and data selection. These inconsistencies and sensitivities contribute to serious uncertainties when epidemiological results are used to discern the nature and magnitude of possible ozone–mortality relationships or are applied to risk assessment' [10].

It was noted by Krewski *et al*. [11], who support the view that a relationship exists between air pollution and statistical deaths, that if there are effects, they are heterogeneous, i.e. varying across the United States as shown in their Figure 21, which is reproduced here as Fig. 1(a). Smith *et al*. [9], using complex methods for ozone levels, also noted that the effects were not constant across the United States. We present their geo-map for the 8-hours and all-year measurements as Fig. 1(b). In both geo-maps, there are hot spots and vast areas where any affect of air pollution on mortality appears minimal to nonexistent; i.e. there is geographic heterogeneity. From a statistical point of view, the story could stop right here. There is interaction. Reliance on a main effect of air pollution, PM2.5, and/or ozone is not supported by the statistical analysis.

We were fortunate to obtain from Dr. C. Arden Pope III the data used in his 2009 *New England Journal of Medicin*e paper. That data is comprehensive and allows us to address two important questions. The first question is the relative importance of air pollution relative to other factors with respect to statistical deaths. In all of air pollution literature that we have surveyed there is essentially no presentation of such information. Yes, it is of interest to scientists to determine if there is any effect of air pollution, but it should also be important to decision makers to know its relative importance with respect to other factors so that possible tradeoffs can be considered. The second question concerns regional differences in the air pollution and statistical death relationship. Given Fig 1(a) and 1(b), it would be of interest to know if there is evidence for differential effects in the western United States as opposed to the East. Enstrom [5] finds no effect in California, with a relative risk of 1.00 and confidence limits of 0.98–1.02. His results are confirmed by CARB consultant Professor Jerrett, with a relative risk of 1.00 and confidence limits of 0.97–1.03. We compute multiple analyses sweeping across the county from west to east and show that one can 'cut' along the longitude passing just west of Chicago and find no effect of PM2.5 to the west and a small effect of PM2.5 on statistical deaths to the east. Both Styer *et al*. [8] and Smith *et al*. [9] make the point if the effect of the pollutant is not consistent, then it is unlikely that you have a causative agent. We agree.

Pope *et al*. [4] suggest that PM2.5 is a statistically significant cause of death uniformly across the United States, so reducing PM2.5 will lead to improved life expectancy. They tacitly take the position that PM2.5 should be reduced without regard to other variables that impact mortality. We take the position that their reporting and analysis is in support of the righteous end of saving lives; they do not cite contrary papers, e.g. Enstrom [5], and they ignore geographic heterogeneity, which they note in Krewski *et al*. [11] and so their paper is consistent with White Hat Bias. Based on our analysis, an alternative interpretation is that PM2.5 exhibits different associations with mortality in the eastern and western United States, suggesting that a single national policy is not appropriate across the entire country. In any case, the relative importance of PM2.5 to statistical mortality, as compared to other factors, should be taken into account by decision makers. Following the methods used by Krstić [12] we will provide estimates of days of lives extended for changes in PM2.5 and income.

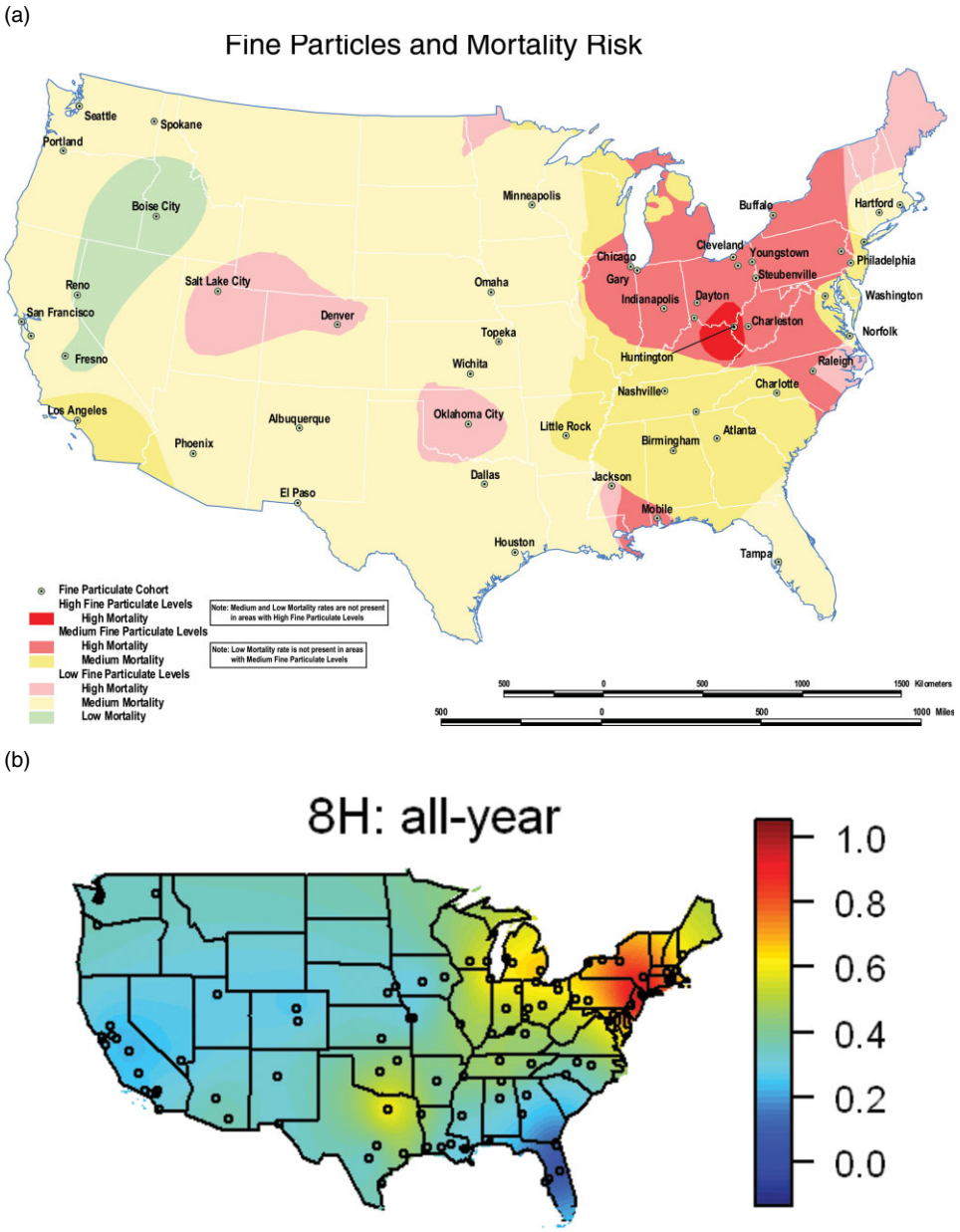

Fig. 1 (a) The risk of mortality due to fine particles varies by location (*Source:* Krewski *et al*. [11].). (b) The risk of mortality from ozone varies by location (*Source:* Personal communication from R. L. Smith *et al*. [9]). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In this article, we first describe the Pope *et al*. [4] data set. Next we describe two analysis methods, regression and recursive partitioning (RP), that can be used to assess the importance of predictor variables. Next we give a series of results: evidence of geographic heterogeneity, variable importance using three methods, and partial correlations to help understand the predictors in the Pope data set. Finally, we discuss our results and a number of literature thoughts on the nature of science inference from complex observational data.

2. DATA

Pope *et al*. [4] started with 2068 county units from which 215 county units in metropolitan areas were selected that had matching PM2.5 data available. Four New York areas were consolidated into one, so ultimately there were 211 records for 51 metropolitan areas within the United States. Note that there are only 51 distinct PM2.5 measurement stations; these were replicated as necessary and assigned to 211 metropolitan areas. The response variable was the

Table 1. Variables reported and use by Pope *et al*. for regression analysis. NB: All variables are given as change from years ∼1980 to ∼2000.

Variable	Comment
Life Expectancy, life-table methods	Response variable (Change LE)
Per capita income (in thousands of $)	Inflation adjusted to the year 2000 (Income)
Lung Cancer (Age standardized death rate)	Surrogate for smoking (LCan)
COPD (Age standardized death rate)	Surrogate for smoking. COPD denotes chronic obstructive pulmonary disease
High-school graduates (proportion of population)	(hs)
PM2.5 ($\mu g/m^3$)	Particulate matter, aerodynamic diameter ≤ 2.5 μm
Black population (proportion of population)	Self reported (black)
Population (in hundreds of thousands)	(pop)
5-Year in-migration (proportion of population) (mig)	Five-year in-migration refers to the proportion of the population who did not reside in the county 5 years earlier.
Hispanic population (proportion of population)	Self reported (hisp)
Urban residence (proportion of population)	(urban)

Table 2. Means and standard deviations for the 11 variables in the Pope data set.

	Variable	Mean	SD
1	Change LE	2.7312	0.9167
2	Lcan_d	2.3455	2.7726
3	copd_d	4.4397	2.4358
4	Change Income	8.5069	3.1608
5	Change PM	6.5477	2.9151
6	hs_d	0.1872	0.1453
7	black_d	0.0176	0.0565
8	hisp_d	0.0333	0.0431
9	Pop_d	0.9948	2.2599
10	urban_d	0.2002	0.1800
11	mig_d	−0.0063	0.0613

change in age-adjusted mortality from the early 1980s to the late 1990s. And there were 10 predictor variables; see Table 1. The predictor variables are also the change over time. For example, change in income is 'income 2000 minus income 1979'. Means and standard deviations for these variables are given in Table 2. Partial correlations among the variables are given in Table 3. The change in PM2.5 is the same for each unit within a metropolitan area. The data used in this article were obtained from Professor Pope and will be posted to http://www.datadryad.org. We also provide the data set consolidated to 51 metropolitan areas.

3. METHODS

3.1. Introduction

We use two methods of model fitting: linear regression and RP. Some of the many potential problems with regression methods are covered by Glaeser [13]. A review of regression variable importance measure is given by Nathans *et al*. [14]. RP is also used as it is robust to nonlinear relationships. The Golden Helix implementation of single and multiple tree RP is described in their user manual, HelixTree manual [15] available at http://www.goldenhelix.com/pdfs/HelixTreeManual.pdf.

3.2. Regression

The linear regression model of the following form is considered:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + e, \qquad (1)$$

where Y represents the change of life expectancy from the early 1980s to the late 1990s; X_1 to X_{10} represent the 10 covariates considered in Pope *et al*. [4], including the changes of PM2.5, income, high school graduate rate, and two proxy indicators for smoking, and so on, which are listed in Table 1. The residuals e are assumed to have an independent, identical Gaussian distribution with mean 0 and variance σ^2.

3.3. Step-Wise Regression

For the purpose of either variable selection or variable importance assignment, step-wise regressions are often conducted. NB: the order of entry is important and later we use a method that averages all orders. In the forward selection mode, the simplest model without any regression variables is first fitted:

$$Y = \beta_0 + e$$

Then one regression variable is added to the model, forming a second model:

$$Y = \beta_0 + \beta_i X_i + \varepsilon$$

The decrease of the residual sum of squares r_i is assigned to the regression variable X_i. This procedure is repeated, until all 10 variables enter the 11th model. Eventually, there will be a vector of residual sum of squares $(r_1, r_2, \ldots, r_{10})$ for the 10 regression variables. If all regression variables are independent, the vector will be unique regardless of the sequence of the variables entering the linear model in

Table 3. Partial correlations among the variables in the data set adjusted over all pairs of other variables. PM2.5 partial correlations were not significantly associated with any of the other variables, Bonferroni adjusted.

Variable	Change LE	Lcan_d	copd_d	Change Income	Change PM	hs_d	black_d	hisp_d	Pop_d	urban_d	mig_d
Change LE	1.000	**−0.263**	**−0.237**	**0.412**	0.147	0.000	−0.090	0.021	0.058	0.001	0.001
Lcan_d	**−0.263**	1.000	**0.291**	−0.027	−0.074	**−0.183**	−0.062	0.005	−0.163	0.001	0.082
copd_d	**−0.237**	**0.291**	1.000	−0.032	−0.028	**−0.257**	−0.079	−0.067	−0.015	**0.182**	0.013
Change Income	**0.412**	−0.027	−0.032	1.000	−0.005	**0.423**	−0.016	−0.002	−0.009	0.167	0.020
Change PM	0.147	−0.074	−0.028	−0.005	1.000	−0.002	0.001	−0.000	−0.112	−0.000	0.001
hs_d	0.000	**−0.183**	**−0.257**	**0.423**	−0.002	1.000	0.089	0.009	0.077	−0.057	**−0.328**
black_d	−0.090	−0.062	−0.079	−0.016	0.001	0.089	1.000	0.012	−0.001	−0.069	−0.006
hisp_d	0.021	0.005	−0.067	−0.002	−0.000	0.009	0.012	1.000	**0.395**	−0.149	−0.000
Pop_d	0.058	−0.163	−0.015	−0.009	−0.112	0.077	−0.001	**0.395**	1.000	−0.010	−0.097
urban_d	0.001	0.001	**0.182**	0.167	−0.000	−0.057	−0.069	−0.149	−0.010	1.000	−0.062
mig_d	0.001	0.082	0.013	0.020	0.001	**−0.328**	−0.006	−0.000	−0.097	−0.062	1.000

Eq. (1). However, if there are correlated variables, their r_i values will depend on the order that the variable enter the linear model.

3.4. Regression Variable Importance

Variable importance estimates are achieved by decomposing var(Y) into the parts attributable to the individual X_is. There are several methods of variable importance assignment based on the linear regression models shown in Eq. (1), as described in Lindeman *et al*. [16], Grömping [17–19], and Pratt [20] among others, Nathans *et al*. [14]. In our article, we use the method proposed by Lindeman, Merenda, and Gold (LMG). It considers all the 10! = 3,628,800 permutations of the stepwise regression using the 10 regression variables. The method is computationally intensive, but there is free code to do the analysis in R, see ref. 17. Let $(r_1^{(k)}, r_2^{(k)}, \ldots, r_{10}^{(k)})$ represent the kth variable importance assignment for the regression variables; the final variance importance assignment is just the average importance over all the permutations:

$$\overline{r}_i = \frac{1}{10!}\sum_{k=1}^{10!} r_i(k), i = 1, \ldots, 10. \quad (2)$$

For two correlated variables, a single stepwise regression will diminish the relative importance of the variable that enters the regression model at a later time; the LMG method averages across all possible full-term stepwise regressions and LMG claim their method assigns a more balanced value of importance to both variables.

We also compute the method of Pratt [20], following his example, which gives the fraction of the standard deviation of the response attributable to each of the predictors. We chose to omit those variables that were not significant (denoted by 'NS' in the results) at the significance level of 0.05, without multiplicity adjustment.

3.5. Partial Correlations

A partial correlation coefficient quantifies the correlation between two variables when adjusted for the linear effects of one or more other variables; see ref. 21. We give the partial correlations among each of the pairs of variables, conditioned on all possible pairs of other variables, in Table 3. The significance of each partial correlation was determined using the method by Fisher [22]. A Bonferroni adjusted p-value, 0.05/55 = 0.0091, was used to highlight larger partial correlations. A p-value plot, as in Schweder and Spjøtvoll [23], was used to help evaluate the multiple results.

3.6. Single Tree

RP is a data mining method useful for uncovering complicated relationships in large, complex data sets. These relationships may involve thresholds, interactions, and nonlinearities. Any or all of these relationships impede an analysis based on the standard assumptions in multiple linear regression. RP was originally designed for automatic interaction detection; see ref. 24. The method has been subject to much development and is widely used for complex modeling situations; see ref. 25. The basic analysis strategy of recursive partitioning is simple and easily understood with an example. Consider an analysis of the Pope data set for the eastern United States; see Fig. 2. The 185 observations from the Eastern United States are in the top node, denoted by N. Also given within a node are summary statistics that show the mean (u), standard deviation (s), and multiplicity adjusted p-values used in the splitting process. All potential predictor variables are examined and the variable with the smallest adjusted p-value is used to split the node into two or more daughter nodes. In this case, Change Income is the variable with the smallest adjusted p-value. Segmentation is used to find the optimal 'cut points', making in this case three daughter

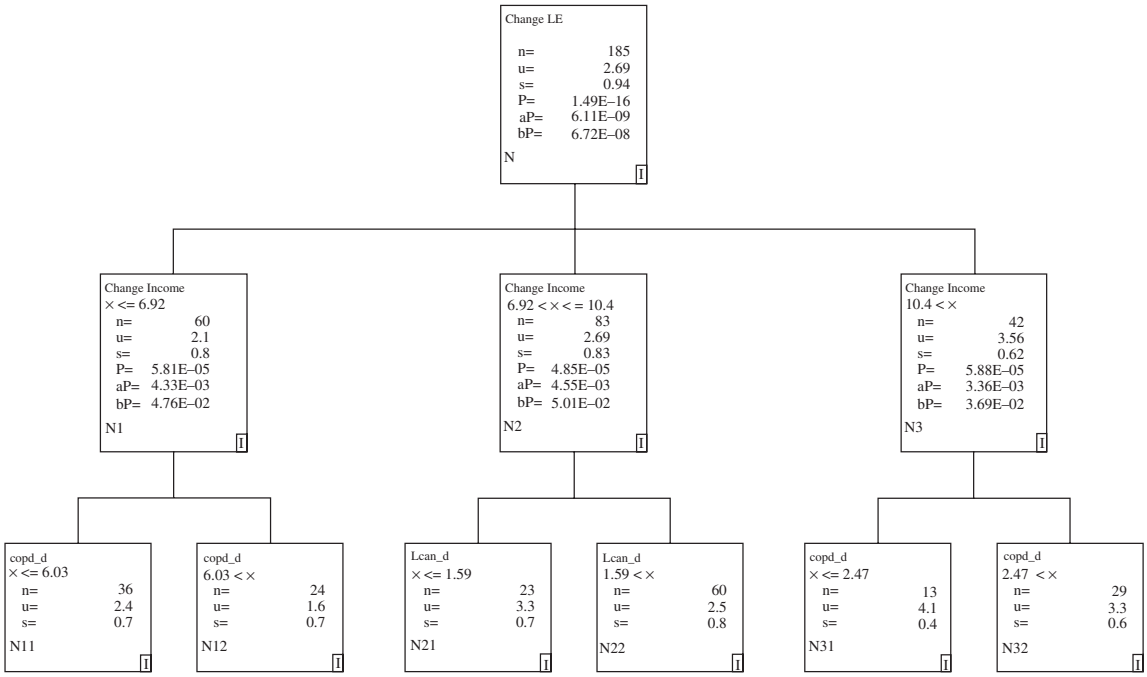

Fig. 2 Recursive partitioning analysis selects the best predictor, Change in Income, and makes two 'cuts' splitting the predictor into three groups with Life Expectancy increasing with increased income. Each of the three nodes is split in turn by variables that are surrogates for smoking, Lung Cancer and COPD. The difference of Life Expectance from the node with lowest increase in income to the highest is about 1.5 years. Lung Cancer and COPD confer about 0.8 years in increased life expectancy. Three p-values are given: the raw p-value, P, is unadjusted; aP is adjusted for the number of ways to cut the predictor into categories; bP is adjusted for cuts and variables available for making a cut.

nodes, denoted by $N1$, $N2$, and $N3$, respectively. It is a user option to control the maximum number of cut points. We set the number of cut points to a maximum of two. The p-value for this cut is adjusted to reflect the number of variables available and the number of ways the segmentation can be done, as well as the number and placement of the cuts. Each of the daughter nodes is examined in turn and is split if significant. Nodes $N1$ and $N3$ use COPD to split and Node $N2$ is split using Lung Cancer. Pope *et al*. [4] used both COPD and Lung Cancer as surrogates for smoking. Each node is split in turn and the recursive splitting stops when there are no statistically significant splits to be made. Notice that at each level of the tree building the standard deviation in each node gets smaller as splitting progresses. Tracing from N to $N1$ to $N11$ we see the standard deviations decrease as 0.94, 0.8, and 0.7, respectively.

3.7. Multiple Trees

There are advantages (more accurate predictions and the ability to assess variable importance) to computing and using multiple trees in the analysis of a data set; see refs. 26,27. Multiple trees can be computed by sampling with replacement multiple random samples from the data set and computing a tree for each such sample; see ref. 28. Alternatively, at a split, from among the valid split variables, one can randomly sample one variable to make the split; see refs. 26,27. Once there are multiple trees, they can be used to determine variable importance in two ways. One can compute how often a variable is used over all the multiple trees. Alternatively, the split variable controls all the samples below it so, across the multiple trees, the fraction of the observations controlled by a variable can be computed. The latter method is used by HelixTree [15] from Golden Helix (Bozeman, MT) and we report its results.

4. RESULTS

It is perhaps not appreciated by the general scientific community, but it is well-known among experts that air quality has a differential effect on mortality in eastern and western United States with essentially no effect in the West; see refs. 5,9,11,29. As these results are based on several data sets with analyses done by several teams of investigators, the no-detectable-effect on mortality of PM2.5 in the West appears to be real. One explanation is that PM2.5 is based on physical particle size, not specific chemical composition. Bell *et al*. [30] report that there

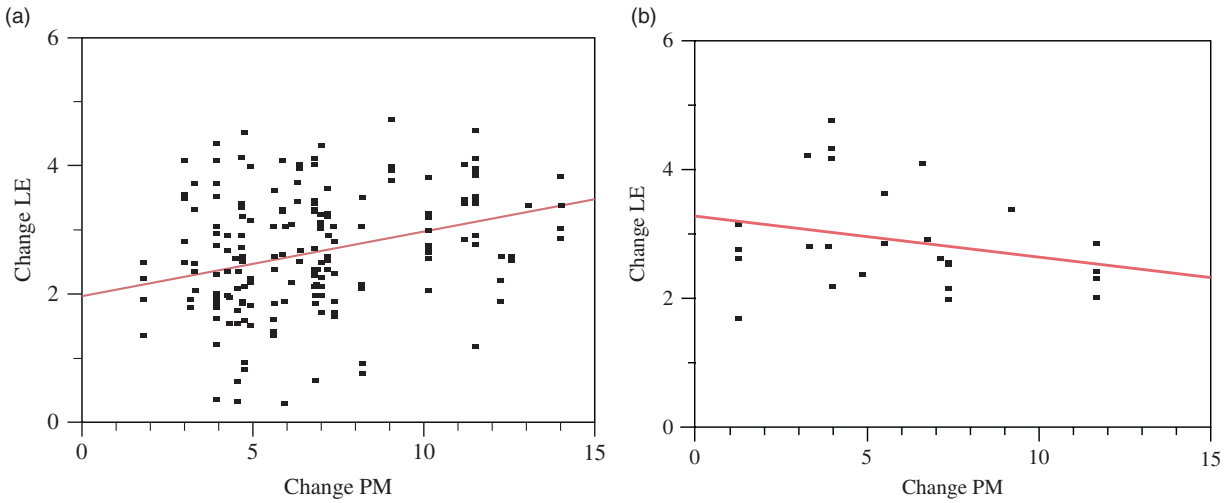

Fig. 3 Change in Life Expectancy(Change LE), in (a) East United States is positive with respect to Change PM2.5 whereas it is not statistically significant in (b) West United States. The East/West was taken as Denver. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

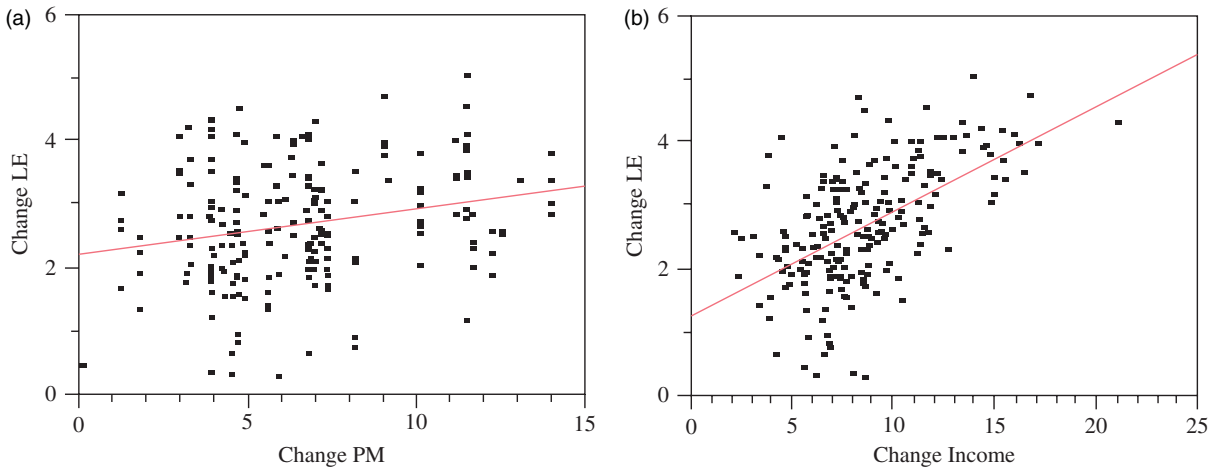

Fig. 4 Change in Life Expectancy(Change LE), versus (a) change in PM2.5 and (b) Change in Income. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

is both temporal and spatial variation in the chemical composition of PM2.5. With the Pope *et al*. [4] data set we confirm the geographic heterogeneity of PM2.5 health effects, and that there is no detectable effect in the western United States. Figure 3 gives scatter plots of change in Life Expectancy versus PM2.5 for the eastern and western United States. A linear regression for the eastern and western subsets finds a significant increase in mortality for the East, but not for the West; the slopes for the two regression lines are significantly different from one another, with p-value 0.0063. We selected Denver as the division of West/East (Figs 4 and 5). The choice of Denver was arbitrary. To better understand the effect of PM2.5 across the United States, we computed the regression of longevity on PM2.5, stepping across the United States from west to east and we give the slope of the regression line as we go; see Fig. 6. Start in the West and do a regression for western most k points. Also compute the regression for the points to the east of the k points. Now move to $k + 1$, then $k + 2$, and so on moving eastward. We add more points so we should have more power for the West and decreasing power for the East. But if there is East/West interaction, then any effect is attenuated. So the Pope *et al*.'s claim that life expectancy increases with a decrease in PM2.5 is supported in eastern United States, but not in the western United States.

Variable importance for the eastern and western United States is computed using the regression method of LMG and the RP method in HelixTree [15]. The variable importance results are given in Table 4. The predictor variables are given in order of their importance in multiple linear

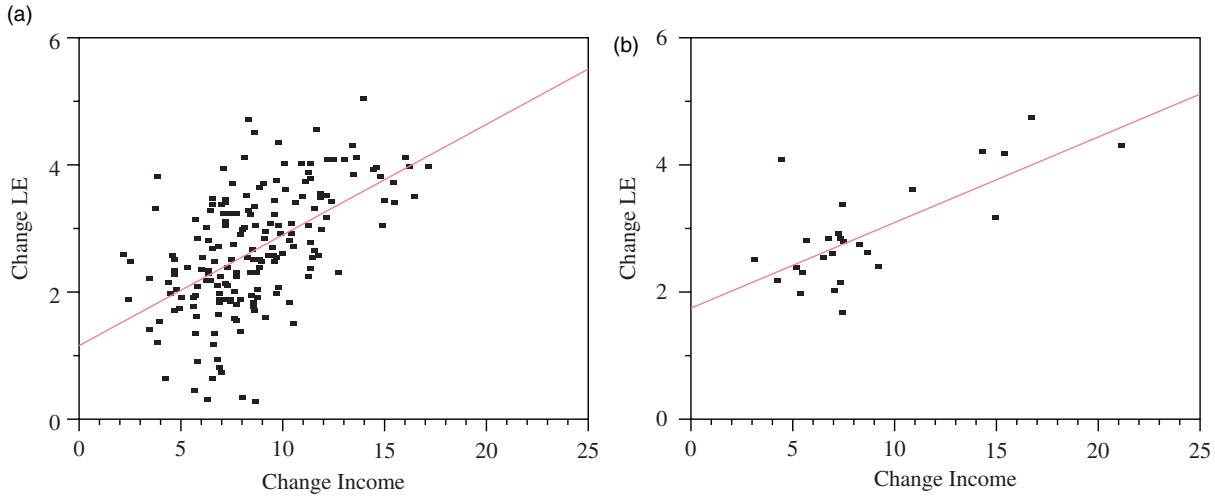

Fig. 5 Change in Life Expectancy in (a) eastern and (b) western United States increases with increased income. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

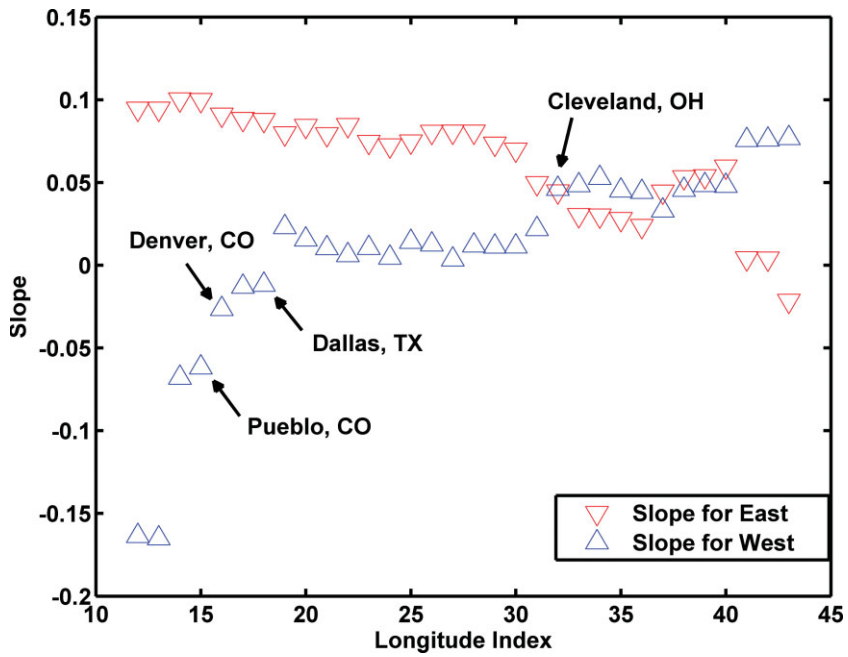

Fig. 6 Slopes of PM2.5 regression line cutting the country into West and East along a line crossing a city. Start in the West on the left of the figure. Two regression coefficients are given for each cut point. The blue points are for the city and for the points to the west of the cut point. The red points are for the points to the east of the cut point. Initially there are few cities that make up the blue points and many for the red points. When you get to Cleveland the regression coefficients are equal West and East. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

regression in the eastern United States. Increase in income is the most important variable for predicting improved mortality, in both eastern and western United States, and for both the regression and the RP variable importance methods. Lung Cancer and COPD are about equally important in the eastern United States. COPD and PM2.5 are relatively unimportant in the western United States. The Percent Graduating from High School and PM2.5 are about equally important in the eastern United States. Regression analysis indicates that %Black and Population Density are important in the western United States, but not very important in the eastern United States. Both regression and RP put the importance of PM2.5 in fourth place among the predictors, and roughly equal in importance to a high school education. Both linear regression and RP indicate that PM2.5 is unimportant in the western United States (Fig. 7).

Table 4. Variable importance. The rows are sorted by importance in East United States. 'Regression' importance by variance explained using linear regression over all 10! permutations of the order of the variables. 'Recursive Partitioning' is importance by the proportion of the samples controlled by a variable using 1000 trees. Note, in bold, the differences in regression importance for a number of the predictor variables between East and West. For Recursive Partitioning in the West, there was only one significant split, on Income; see Fig. 6.

	Regression			Recursive Partitioning			Pratt
Variable	East	West	United States	East	West	United States	United States
Income	0.2792	0.3996	0.3390	0.2865	1.0000	0.2108	21.70
COPD	**0.1789**	**0.0216**	0.1621	0.2298	0.0000	0.1199	9.83
LungCancer	0.1697	0.1806	0.1768	0.2385	0.0000	0.1467	9.40
PM2.5	**0.1095**	**0.0299**	0.0732	**0.1118**	**0.0000**	0.1302	3.78
HighSchool	0.1013	0.0859	0.0997	0.1097	0.0000	0.1066	NS
%Black	0.0620	0.1250	0.0537	0.0000	0.0000	0.0319	2.12
PopDensity	0.0370	0.1171	0.0418	0.0000	0.0000	0.0793	2.95
%Hispanic	0.0281	0.0065	0.0177	0.0237	0.0000	0.0136	NS
Migration	0.0240	0.0120	0.0228	0.0000	0.0000	0.0202	0.12
Urban	0.0103	0.0217	0.0133	0.0000	0.0000	0.0105	NS

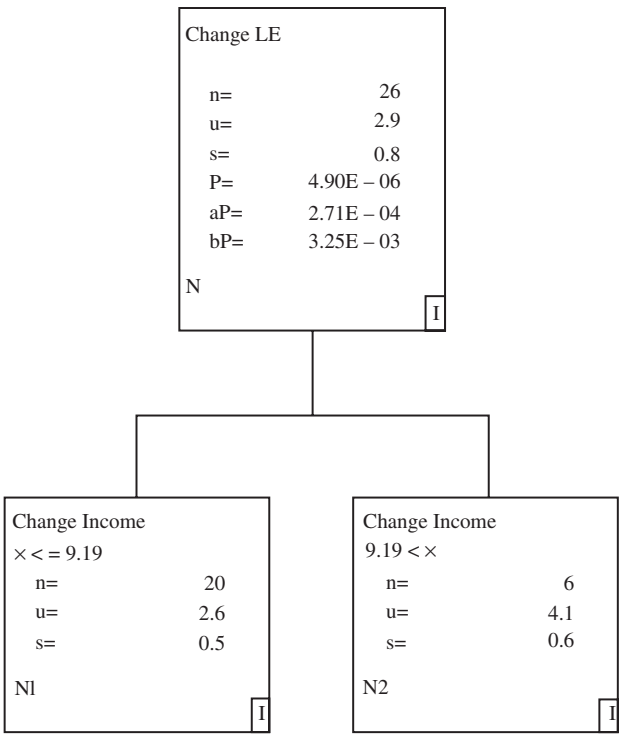

Fig. 7 Recursive Partitioning analysis of observations in the western United States. The only significant predictor of Change in Life Expectancy is Change in Income. Those that had an increase in income of over $9.19k from approximately 1980 to approximately 2000 had an increase in life expectancy of 4.1 years versus 2.6 years for those that increase their income by less than or equal to $9.19k.

The second-order partial correlations, as described by de la Fuente *et al*. [31], were computed between the pairs of the 11 variables in the data set and tabulated in Table 3. Among the partial correlations there are a number of relatively large and expected correlations; e.g. change in income and change in high school graduates; change in life expectancy and change in income; change in percent Hispanic and change in population; and so forth. Notable is the lack of partial correlation between change in PM2.5 and any of the other variables at the multiplicity-adjusted 0.0091 level.

In complex studies, the *p*-value plot is often helpful to get a sense of 'Is the observed effect of PM2.5 larger than chance?' see Fig. 8. Here we plot the ranked *p*-values for the partial correlations against the integers. A 45° line indicates that there is nothing of interest, whereas points off the line, on the blade of the hockey stick, indicate real effects.

5. DISCUSSION

The problems of observational studies have been well-known for many years; see refs. 32,33 for discussion. But there has been little or no progress in adopting better methods; see refs. 34,35. The end result is that most claims that are based on observational data fail to replicate on retesting; see refs. 36,37.

The association between PM2.5 with mortality, when compared to the associations between other variables and mortality, shows that the importance of PM2.5 is relatively small. There is no measurable association in the western United States, although it accounts for about 11% of the variance in the eastern United States. The Pratt regression analysis across the entire United States has PM2.5 explaining about 4% of the standard deviation. The partial correlations in Table 3 are given primarily to get a general sense of the complex correlation structure of the data set, but they too indicate that PM2.5 is relatively unimportant.

The examination of partial correlations is useful for examination of relationships among predictor variables.

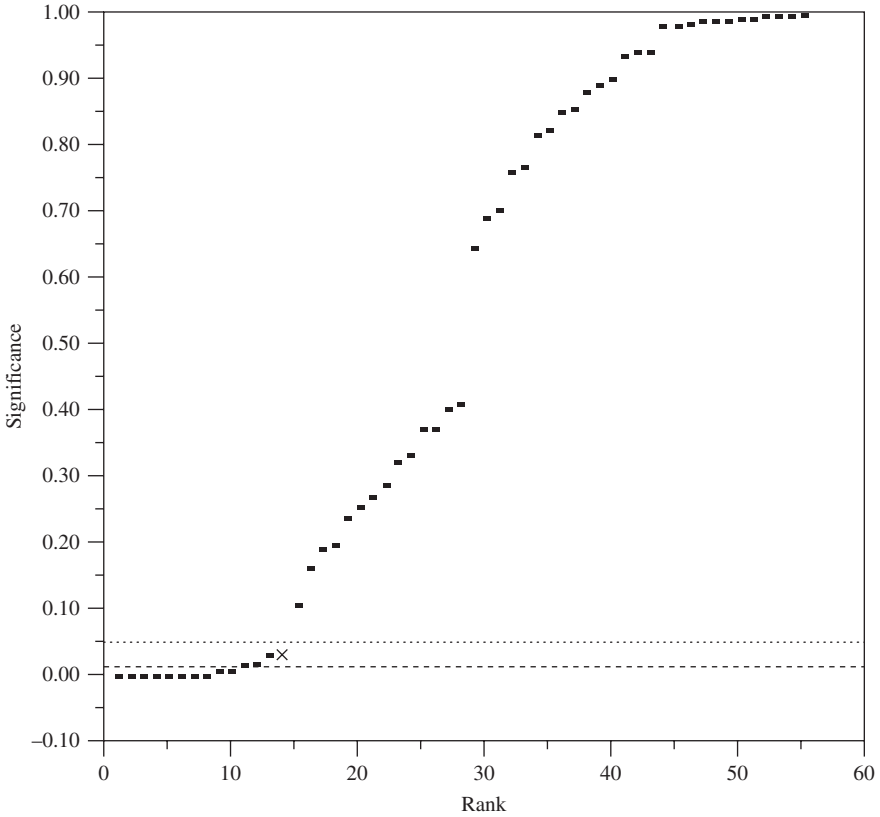

Fig. 8 *P*-value plot of the 55 *p*-values for the partial correlations. Given as a dotted line is the nominal significance level of 0.05. Given as the dashed line is the Bonferroni significance level of 0.0091. An 'x' marks the *p*-value for the partial correlation of change in life expectancy and change in PM2.5. The gap appears unusual and it unexplained.

Consider Fig. 8, where we plot the ranked *p*-values for the partial correlations against the integers. The partial correlation of PM2.5 with mortality, marked as an 'x' is not significant in the context of all the partial correlations, but it is significant if multiple testing issues are ignored. There are many other partial correlations better supported by the data. Very curiously, there is a large unexplained gap in the points falling on a 45° line. We provide a partial correlation diagram, Fig. 9, to visualize the links among the variables. This diagram suggests a straightforward strategy. One could increase education efforts with the idea of increasing income and thereby increasing longevity.

While conducting our reanalysis of the Pope data set a pertinent publication appeared, Krstić [12]. Krstić focused his reanalysis on the data set of 51 metropolitan regions. He noticed that the statistical significance of PM2.5 failed if he removed what he considered an outlier observation, Topeka, Kansas. We follow the calculations given in Krstić to estimate expected change in longevity for a change in PM2.5 or income. Change in longevity is computed as the regression coefficient times the change in either PM2.5 or Income. Krstić also adjusts this number by multiplying the result by R^2. The results of these calculations in terms of standard deviations are given in Table 5. For both PM2.5 and income, the days saved depends on the value of the change in the variable. For both PM2.5 and income a change of one standard deviation seems representative of policy goals. Again, income is more important than PM2.5.

Table 5. Change in days of life for changes in PM2.5 and Income computed as the regression coefficient times the change in PM2.5 or Income. Following Krstić, these changes are weighted by the R2 for PM2.5 and Income to give Days K.

SD	Days PM2.5	Days Income	Days K PM2.5	Days K Income
0.5	38.5	95.5	2.0	30.4
1	77.1	190.9	4.0	60.7
2	154.2	381.9	8.0	121.4

All analysis indicates that changes in income and several other variables are more influential than PM2.5, so policymakers might better focus on improving the economy, reducing cigarette smoking, and encouraging people to pursue education. With no adjustment to the significance level for multiple testing, for example, the partial correlation of change in life expectancy and PM2.5 is 0.147 and this is significant at the 0.05 (unadjusted) level. Since we provide

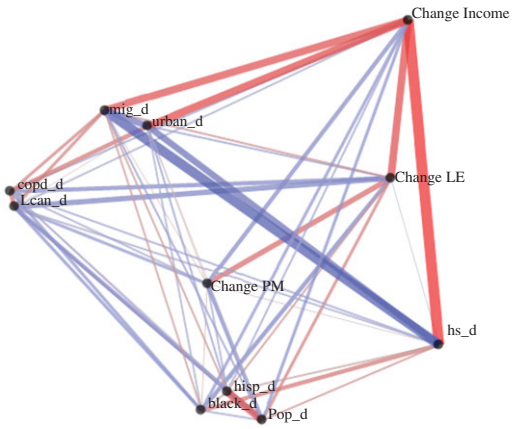

Fig. 9 Partial correlation diagram. The thicker the line the stronger the partial correlation. Positive partial correlations are red and negative are blue. Focus on Change LE, the change in life expectancy. Income is the most important variable, then smoking proxies, COPD and Lung cancer. High-school is positively associated with increased income, as expected. People migrate to metropolitan area with increasing income. Urbanization increases where income increases. The strongest path to increased life expectancy is from high-school to income to increase life expectancy. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the data set, an interested reader can produce variations of this partial correlation table as well as other analyses.

The classic view of science is expressed by Feynman[38]: 'In summary, the idea is to try to give all the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another'. In contrast, the scientist as advocate is expressed in the Stanford News headline of August 11, 2011, 'Scientists must leave the ivory tower and become advocates, or civilization is endangered, says Stanford biologist Paul Ehrlich' The sub-headline continued, 'Scientists, especially ecologists, have to be more active in explaining the meaning of their research results to the public if human behavior is going to change in time to prevent a planetary catastrophe' [39].

Complex modeling presents its own problems. Friedrich Hayek [40] in his Nobel Prize lecture of 1974 described the situation of complex modeling outside the area of the physical sciences where theory offers guidance on which variables need to be measured. In nonphysical sciences, one might simply use available measurements. In physical sciences the number of relevant variables can be small and the relationships simple, whereas in complex biological systems both the number of variables and how they are related can be very complex, which Hayek called essential complexity. With a large number of variables and complex relationships, laymen have essentially no ability to discern the validity of the model. Even experts will have trouble evaluating claims based on models. Debunking invalid models is difficult because the models are complex and because people and institutions tend to become invested in those models. We summarize the Hayek argument: There are multiple factors that are likely to impinge on a phenomenon of interest and many of these factors may not be measured or even measurable. Outside of the physical sciences we have little theory to guide us on what needs measuring. These unavailable factors can lead to biases that may be on the same order of magnitude as the phenomenon under study. In our case, the study of factors associated with mortality, the mechanism is one of *essential* complexity. The statistical modeling process is not simple, so even experts find it difficult to judge the validity of the analysis. We think the link between air pollution and longevity lines up nicely with Hayek's essential complexity. Pope *et al.* claim that PM2.5 is killing people. Krstić is on the other side, 'The observed loss of statistical significance in the correlation between the reduction of ambient air PM2.5 concentrations and life expectancy in metropolitan areas of the United States, after removing one of the metropolitan areas from the regression analysis, may raise concern for the policymakers in decisions regarding further reductions in permitted levels of air pollution emissions.' Given the lack of effect in the West and the greater importance of other predictors, we agree with Krstić that this data set does not support the claim that decreasing PM2.5 will increase longevity. If the cost of decreasing PM2.5 is high enough there could well be a net loss in longevity.

REFERENCES

[1] P. W. Huber, Galileo's Revenge, Junk Science in the Court room. New York, BasicBooks, 1991.
[2] M. B. Cope and D. B. Allison, White hat bias: examples of its presence in obesity research and a call for renewed commitment to faithfulness in research reporting. International Journal of Obesity December 1 (2009), 1–5.
[3] News Release—Air, March 1, 2011. EPA Report Underscores Clean Air Act's Successful Public Health Protections. Landmark law saved 160,000 lives in 2010 alone, http://yosemite.epa.gov/opa/admpress.nsf/6424ac1caa800aab85257359003f5337/f8ad3485e788be5a8525784600540649!OpenDocument, 2011.
[4] C. A. Pope III, E. Ezzati, and D. W. Dockery, Fine-particulate air pollution and life expectancy in the United States, N Engl J Med 360 (2009), 376–386.
[5] J. E. Enstrom, Fine particulate air pollution and total mortality among elderly Californians, 1973–2002, Inhalation Toxicology 17 (2005), 803–816.
[6] R. D. Peng, F. Dominici, and S. L. Zeger, Commentary: Reproducible epidemiologic research, American Journal of Epidemiology 163 (2006), 783–789.
[7] J. M. Schwartz and S. F. Hayward, Air Quality in America, Chapter 7, Air Pollution and Health, AEI Press, 2007.
[8] P. Styer, N. McMillan, F. Gao, J. Davis, and J. Sacks, Effect of outdoor airborne particulate matter on daily death counts, Environ Health Perspect 103 (1995), 490–497.

[9] R. L. Smith, B. Xu, and P. Paul Switzer, Reassessing the relationship between ozone and short-term mortality in U.S. urban communities, Inhal Toxicol 29(S2) (2009), 37–61.
[10] Science and Decisions: Advancing Risk Assessment, Washington DC, National Academies, http://books.nap.edu/catalog/12209.html, 2008.
[11] D. Krewski, R. T. Burnett, M. S. Goldberg, K. Hoover, J. Siemiatycki, M. Jerrett, M. Abrahamowicz, and W. H. White, Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Part II: Sensitivity Analysis, HEI Publications. http://pubs.healtheffects.org/view.php?id=6, See Figure 21, in particular, 2000.
[12] G. Krstić, A reanalysis of fine particulate matter air pollution versus life expectancy in the United States, J Air Waste Manage Assoc 62(9) (2012), 989–991.
[13] E. L. Glaeser, Researcher Incentives and Empirical Methods. www.economics.harvard.edu/pub/hier/2006/HIER2122.pdf, 2006. [Last accessed on July 4, 2013].
[14] L. L. Nathans, F. L. Oswald, and K. Nimon, Interpreting multiple linear regression: a guidebook of variable importance. Pract Assess Res Eval 17(9) (2012). http://pareonline.net/getvn.asp?v=17&n=9.
[15] HelixTree manual. http://www.goldenhelix.com/pdfs/HelixTreeManual.pdf. [Last accessed on July 4, 2013].
[16] R. H. Lindeman, P. F. Merenda, and R. Z. Gold, Introduction to Bivariate and Multivariate Analysis, Glenview, IL, Scott, Foresman, 1980.
[17] U. Grömping, Relative importance for linear regression in R: the package relaimpo, J Stat Software 17 (2006), 1. http://www.jstatsoft.org/v17/i01/. [Last accessed on July 4, 2013].
[18] U. Grömping, Estimators of relative importance in linear regression based on variance decomposition, The American Statistician 61 (2007), 139–147.
[19] U. Grömping, Variable importance assessment in regression: linear regression versus random forest, Am Stat 63 (2009), 308–319.
[20] J. W. Pratt, Dividing the indivisible: using simple symmetry to partition variance explained. In T. Pukkila, S. Puntanen (eds.), Proceedings of Second Tampere Conference in Statistics, University of Tampere, 1987, 245–260.
[21] M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol 2 (3rd ed.), Section 27.22, 1973.
[22] R. A. Fisher, The distribution of the partial correlation coefficient, Metron 3 (3–4) (1924), 329–332.
[23] T. Schweder and E. Spjøtvoll, Plots of p-values to evaluate many tests simultaneously, Biometrika 69 (1982), 493–502.
[24] J. A. Morgan and J. N. Sonquest, Problems in the analysis of survey data, and a proposal, J Am Stat Assoc 58 (1963), 415–434.
[25] D. M. Hawkins, Recursive partitioning, Comput Stat 1 (2009), 290–295.
[26] D. M. Hawkins and B. J. Musser, One tree or a forest? Alternative dendrographic models, Comput Sci Stat 30 (1999), 534–542.
[27] D. M. Hawkins and B. J. Musser, Feature selection with nondeterministic recursive partitioning, In Proceedings of the American Statistical Association [CD-ROM] Alexandria, VA, ASA, 2001.
[28] L. Breiman, Random forests, Mach Learn 45 (2001), 5–32.
[29] M. Jerrett, California-specific Studies on the PM2.5 Mortality Association. See slides 12 and 13, no increase in "All causes" death rate, http://www.arb.ca.gov/research/health/pm-mort/jerrett.pdf, 2010. [Last accessed on July 4, 2013].
[30] M. L. Bell, F. Dominici, K. Ebisu, S. L. Zeger, and J. M. Samet, Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies, Environ Health Perspect 115 (2007), 989–995.
[31] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes, Discovery of meaningful associations in genomic data using partial correlation coefficients, Bioinformatics 20 (2004), 3565–3574.
[32] L. C. Mayes, R. I. Horwitz, and A. R. Feinstein, A collection of 56 topics with contradictory results in case-control research, Int J Epidemiolgy 17 (1988), 680–685.
[33] A. R. Feinstein, Scientific standards in epidemiologic studies of the menace of daily life, Science 242 (1988), 1257–1263.
[34] S. J. Pocock, T. J. Collier, K. J. Dandreo, B. L. de Stavola, M. B. Goldman, L. A. Kalish, L. E. Kasten, and V. A. McCormack, Issues in the reporting of epidemiological studies: a survey of recent practice, BMJ 329 (2004), 883–888.
[35] P. Boffetta, J. K. McLaughlin, C. La Vecchia, R. E. Tarone, L. E. Lipworth, and W. J. Blot, False-positive results in cancer epidemiology: a plea for epistemological modesty, J Natl Cancer Inst 100 (2008), 988–995.
[36] J. P. A. Ioannidis, Contradicted and initially stronger effects in highly cited clinical research, JAMA, 294 (2005), 218–228.
[37] S. S. Young and A. Karr, Deming, data and observational studies, Significance 8 (2011), 116–120.
[38] R. P. Feynman, Surely You're Joking, Mr. Feynman! Norton paperback: p. 341, 1997.
[39] L. Bergeron, Scientists must leave the ivory tower and become advocates, or civilization is endangered, says Stanford biologist Paul Ehrlich. Stanford News 11 Aug 2011. http://news.stanford.edu/news/2011/august/ehrlich-scientist-advocates-081111.html, 2011.
[40] F. A. Hayek, Nobel prize lecture: the pretence of knowledge, http://nobelprize.org/nobel_prizes/economics/laureates/1974/hayek-lecture.html, 1974.