# Fine Particulate Matter Does Not Detectably Affect Mortality Rates in One Hundred United States Cities

Louis Anthony (Tony) Cox, Jr., tcoxdenver@aol.com

Douglas A. Popken, dpopken@systemsview.com

DRAFT FOR DISCUSSION

## ABSTRACT

Fine particulate matter (PM2.5) concentrations (C) in air are often considered to contribute causally to increased human mortality rates (R), with the shape of the C-R relation being approximately linear. We tested this causal hypothesis in 100 United States cities, using the large, publicly available, NMMAPS database and US Census data. As expected, multiple linear regression models do show a strong, significant, approximately linear statistical C-R association, consistent with many previously reported results. However, closer analysis of residuals and confounding factors suggests that this statistical association is not causal and that applying linear models to quantify the aggregate C-R relation is misleading: other variables, such as month of the year and mean daily temperature, create strong, nonlinear confounding that explains the statistical association between PM2.5 and mortality rates. Such confounding is hard to detect or control using linear statistical models. Conditional independence tests for potential causation, non-parametric classification tree analysis, Bayesian Model Averaging (BMA), and Granger-Sims causality testing, show no evidence that PM2.5 concentrations have any causal impact on increasing mortality rates. This apparent absence of a causal C-R relation, despite the statistical association, has potentially important policy and economic implications for managing health risks associated with, but not necessarily caused by, PM2.5 exposures.

## INTRODUCTION

Recent policy discussions of air pollution regulation in the United States have raised an apparent trade-off between improving health and increasing jobs: lower levels of air pollution are projected by the EPA and others to save lives (or, more accurately, to reduce mortality rates), but are projected by some economists to slow economic productivity and jobs growth. This perceived trade-off has led to a slow-down in the implementation of proposed new environmental regulations intended to protect human health. But is it real? Would further reducing levels of air pollution really extend lives?

Opinions about the answer have been sharply divided. On the one hand, many epidemiological studies report statistically significant associations between average ambient concentrations of pollutants such as ozone and fine particulate matter (PM2.5) and all-cause (non-accidental) mortality rates[1-4]. This statistical relation between exposure concentrations and mortality rates, often called the *concentration-response* (C-R) function, is typically modeled as being approximately linear, all the way down to zero concentration. On the other hand, experimental studies and analyses of clinical data have been hard-pressed to find clear evidence of a non-zero C-R relation for excess mortalities at current ambient concentration levels for specific pollutants such as PM2.5[5, 6]. Experts in statistical and epidemiological methodology have worried that frequent reports of positive, approximately linear, C-R relations may be driven by the expectations and modeling assumptions of analysts, rather than by data alone[7-9]. Bolstering this concern, some re-analyses of data and analyses of new data that attempt to account for *model uncertainty* (typically, by using many alternative sets of

modeling assumptions and weighting them based on agreement with the data, via Bayesian Model Averaging (BMA), rather than selecting any single set of modeling assumptions that might drive conclusions) have not confirmed the existence of previously reported clear, positive C-R relations between PM2.5 and mortality rates[6, 10,11]. Thus, the most important current scientific question about the C-R relation remains: *Do changes in current ambient concentrations of pollutants cause changes in mortality rates?* For risk analysts, a possibly even more important question is: How can we best use facts and data to address such questions about potential causal relations between exposures and health effects? Ideally, risk analysts would be able to provide neutral information and analysis (not driven by preferences or assumptions favoring any particular policy conclusion) to clarify the nature of probabilistic causal relations, such as the C-R relation, to inform risk management decision-making and policy-making.

The real-world data available for analysis is less than ideal, largely because most C-R data are available at the level of cities and metropolitan areas, but not at the level of accurately measured individual exposures and responses. Analysis of such aggregate data runs a risk of potential ecological biases, leading to false-positive associations (e.g., due to unmeasured confounders or unmodeled measurement errors); or to distortions (false negatives as well as false positives) in estimated statistical associations due to unmodeled differences between aggregate exposure concentrations (based on measurements at a fixed set of monitoring locations) and the spatially resolved true exposures of individuals[13]. Yet, city-level C-R data seem too informative about potentially important causal relations to simply ignore; thus, the question arises of how best to use them to address the fundamental scientific question

of how changing C (e.g., by reducing average ambient concentrations of PM2.5) would change R (e.g., by reducing the average age-specific mortality rates) in different locations, while bearing in mind the limitations of such data. This paper presents new analyses of a large data base on C-R relations in over 100 U.S. cities, with the goals of (a) Independently testing/verifying conclusions about C-R relations for PM2.5 drawn from other data sets (such as the Harvard Six Cities Study and the American Cancer Society survey data emphasized in EPA's assessment of health benefits from further reducing PM2.5[3]); and (b) Performing additional analyses and tests emphasizing potential causation between changes in PM2.5 and changes in mortality rates.

**The "Hundred Cities" Data Set for C-R Relations in U.S. Cities**

A large, publically accessible, data base (NMMAPS) for studying city-level C-R relations is the iHAPSS (internet-based Health and Air Pollution Surveillance System) data base of pollutant levels and mortality rates for U.S. cities made available on-line by Johns Hopkins at www.ihapss.jhsph.edu/. This data base provides historical daily data (from January 1,1987 through December 31, 2000), meteorological data (temperature and humidity), and mortality counts for 108 U.S. cities, of which 101 are currently populated with at least some PM2.5 data. (PM2.5 data was not collected in all years and days in all cities and often had several-day gaps between data points.) The mortality data include all-cause mortality (excluding accidents) and cause-specific mortality counts, are as follows:

- accident – accidental death

- copd – Chronic Obstructive Pulmonary disease
- cvd – cardiovascular deaths
- death – all non-accidental death
- inf - influenza
- pneinf – pneumonia and influenza
- pneu - pneumonia
- resp –respiratory deaths

We divided the original mortality count values for the above variables (from NMMAPS) by the population base for each city, year, and age category (from US Census data) to obtain corresponding annual mortality rates by cause, city, year, and age category; see Appendix A for details.) Since most deaths occur among people over 75, we focus on the exposure-mortality association in this age group. For completeness, the sections on Bayesian Model Averaging and Granger-Sims causality also consider the two younger age categories in the NMMAPS data set: people under 65 (*agecat* = 1 in NMMAPS) and between 65 and 75 (*agecat* = 2 in NMMAPS).

Because the data span over a decade for multiple cities, they are well suited for examining how historical changes in PM2.5 concentrations, C, (from day to day, month to month, and year to year) have affected changes in mortality rates, R, for different cities and time scales. Because they are collected from many cities, the NMMAPS data also provide an excellent opportunity to study heterogeneity in city-level C-R relations. Moreover, by joining the NMMAPS data to freely available census data on population sizes and demographics, it is possible to examine other correlates of changes in mortality rates and PM2.5 levels, and thus to identify and control for some specific potential confounders of the C-R relation observed in city-level data. We will refer to the resulting combined data set as the “Hundred Cities” data set (although, to be precise, it

contains data from 101 cities.) We used a derived variable, *pm25Reconstruct*, to estimated PM2.5 concentration levels. As explained at the ihapps web site (http://www.ihapss.jhsph.edu/data/FAQ.html):

> *The median of the trends is stored in a variable with suffix "mtrend". Adding a variable ending in "tmean" with its corresponding "mtrend" variable should get you something resembling the original averaged values. Adding the "tmean" and "mtrend" variables adds the average detrended series with the median of the long term trends from each monitor. It is not an exact reconstruction of any particular series.*

Accordingly, we computed *pm25Reconstruct* = *pm25tmean* + *pm25mtrend* from the original data, to facilitate cross-city comparisons.

**Research Questions and Methods**

The remainder of this paper examines the following research questions by applying a variety of statistical methods (exploratory, descriptive, hypothesis-testing, and estimation) to the Hundred Cities data set. All data files used can be retrieved from the cited web sites (or, if desired, from the authors as .xls/.csv files. R scripts used in data preparation and analysis are also available upon request).

1. *Is there evidence of a statistically significant positive association between PM2.5 levels and mortality rates?* A striking statistical critique of the large literature that concludes that further reducing PM2.5 levels will save lives (or, more accurately, will postpone deaths) is the counter-claim that Bayesian Model Averaging (BMA) models (which attempt to control for model uncertainty) find little or no evidence of a positive association between PM2.5 levels and mortality rates, at least in Canada[11,12]. We therefore first attempt to reproduce in the Hundred Cities data set the common

finding[1-4] that many statistical models do indicate a significant positive linear association between PM2.5 and mortality. To this end, we fit multivariate linear regression models for mortality rates vs. PM2.5 and other explanatory variables in the Hundred Cities data set, to test whether the regression coefficients for PM2.5 are significantly positive. Next, we test whether any such associations disappear when automated variable selection is used.

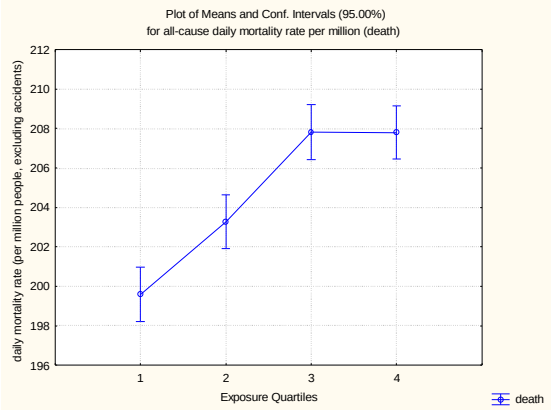
2. *Is there evidence of a causal relation between changes in PM2.5 concentrations, C, and changes in all-cause or cause-specific mortality rates, R?* To address this question, we will borrow from time series econometrics the Granger test for potential causation between two time series. This test is based on the intuitive criterion that there is evidence that exposure concentrations C might contribute to causing adverse responses R if and only if the future time series of R values can be predicted better from past and present values of R and C than from past and present values of R alone. We will also apply conditional independence tests, which test whether observed associations between R and C are fully explained away by other variables (e.g., confounders) that are associated with both of them.

## Results for Aggregate C-R Statistical Associations

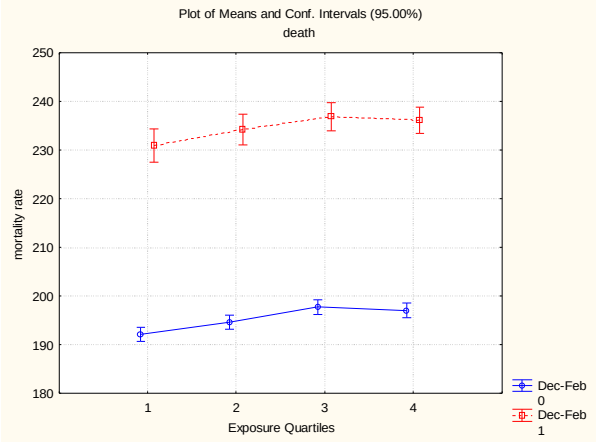
Figure 1 shows the empirical C-R relation for all-cause mortality rates (per million people per day, excluding accidents), pooled across all 101 cities. The horizontal axes in Figures 1a-c show quartiles of the estimated PM2.5 frequency distribution, and the y axes show mortality rates. Vertical bars around each mean value indicate 95% confidence intervals.

Figure 1. Exposure, month, and temperature help to predict daily mortality rates

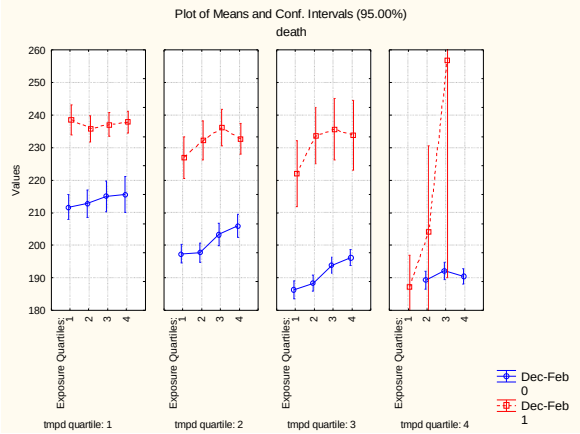
a. All-cause mortality vs. PM2.5 exposure (C-R)




b. C-R relation by season (Dec-Feb vs. other)




c. Death rate decreases with mean temperature (tmpd)




d. C-R association depends on season

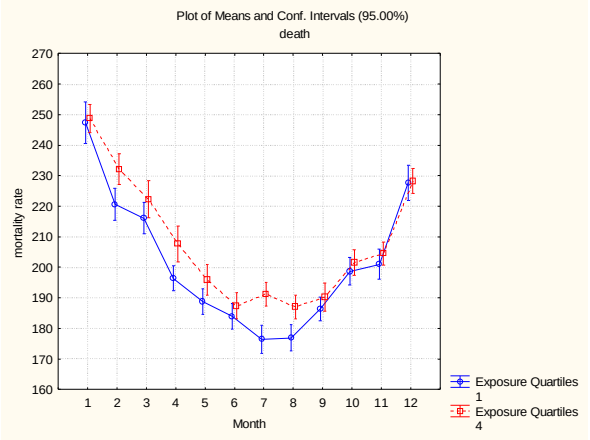


Figure 1a shows that the crude C-R relation is strong, positive, and apparently linear down to the lowest exposures (first quartile) examined. (The same holds when resolution is increased to show deciles.) Although there is apparent saturation at higher exposure levels (quartiles 3 and 4), one would expect a regression model to show a strong, significant, C-R relation. Table 1, discussed below, confirms this expectation: estimated PM2.5 level (*pm25Reconstruct*) is indeed a statistically significant predictor of mortality rate.

Figure 1b shows that the mortality rate depends on more than just PM2.5 exposure: season of the year, indicated by a distinction between the coldest winter months (*Dec-Feb*) (upper curve) and the rest of the year (lower curve), has a large impact on mortality rates. As shown in more detail in Figure 1d, which plots mortality rates against month of year for both the lowest quartile and the highest quartile of estimated PM2.5 exposure, both month of year and PM2.5 exposure help to predict mortality rate. (The curves for the other two quartiles of exposure fall between these other two curves, but are not shown in Figure 1d to avoid crowding the diagram.) It is clear that mortality rates are elevated in December through February, compared to the rest of the year; the season indicator variable *Dec-Feb*, with a value of 1 for these three months and 0 for the other nine months, was constructed to indicate this high-mortality season of the coldest winter months. Figure 1c is an interaction plot showing that mortality rate decreases with temperature (mean temperature, *tmpd*), but generally increases with PM2.5, suggesting that cold, polluted air is most associated with increased mortality rates.

Table 1. Multiple linear regression of non-accident mortality rates (death) against estimated pm2.5 levels, controlling for other variables, shows a statistically significant linear C-R association between them

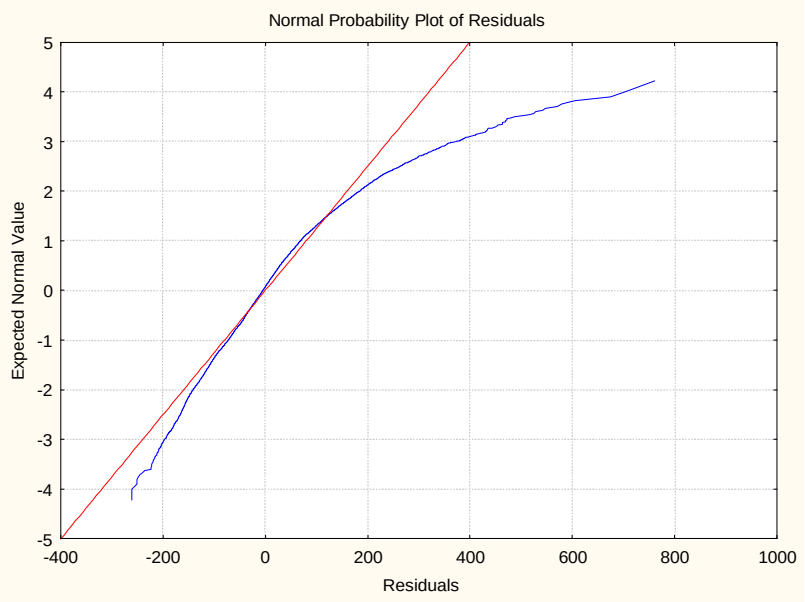
| N=54089 | Regression Summary for Dependent Variable: death<br>R= .26579752 $R^2$= .07064832 Adjusted $R^2$= .07039052<br>(All cities, Age > 75) | | | | | |
|---|---|---|---|---|---|---|
| | b* | Std.Err. of b* | b | Std.Err. of b | t(54073) | p-value |
| Dec-Feb | 0.14 | 0.01 | 27.24 | 1.00 | 27.3 | 0.000 |
| Phighschool2000 | 0.13 | 0.01 | 169.99 | 10.48 | 16.2 | 0.000 |
| tmin | 0.12 | 0.02 | 0.56 | 0.09 | 6.5 | 0.000 |
| dptp | 0.09 | 0.02 | 0.43 | 0.08 | 5.3 | 0.000 |
| pm25Reconstruct | 0.04 | 0.00 | 0.32 | 0.03 | 9.2 | 0.000 |
| pHisp | 0.04 | 0.01 | 29.19 | 4.88 | 6.0 | 0.000 |
| medianIncome | 0.03 | 0.01 | 0.00 | 0.00 | 3.9 | 0.000 |
| pBlack | 0.02 | 0.01 | 14.13 | 3.80 | 3.7 | 0.000 |
| Year | -0.03 | 0.01 | -0.81 | 0.14 | -5.7 | 0.000 |
| Month | -0.05 | 0.00 | -1.13 | 0.10 | -11.1 | 0.000 |
| pOther | -0.05 | 0.00 | -44.65 | 4.10 | -10.9 | 0.000 |
| Purban2000 | -0.05 | 0.00 | -70.62 | 6.34 | -11.1 | 0.000 |
| mnrh | -0.06 | 0.01 | -0.27 | 0.04 | -7.1 | 0.000 |
| Pdegree2000 | -0.13 | 0.01 | -149.41 | 8.92 | -16.7 | 0.000 |
| tmpd | -0.30 | 0.02 | -1.42 | 0.11 | -12.4 | 0.000 |
| Intercept | | | 1838.17 | 284.36 | 6.5 | 0.000 |

Key: mnrh = mean relative humidity; pOther = proportion of ethnic groups other than black, white, or Hispanic; dptp = dew point temperature; see NMMAPS documentation for further definitions.

Table 1 summarizes a multiple linear regression model for mortality rates (*death*) (i.e., deaths per million people per day) regressed against other variables, including estimated PM2.5 levels (*pm25Reconstruct*). (Regression was performed using the commercial statistical software environment *Statistica 9.0* with backward stepwise variable selection. The *b* coefficients are the ordinary least squares regression coefficients, and the *b** coefficient are their standardized values.) PM2.5 has a highly statistically significant positive association with mortality rate, even after correcting for other variables using linear regression. This is consistent with previous reports of a significant positive PM2.5 C-R association, e.g., in the American Cancer Society (ACS) survey data[15].

Before accepting or interpreting regression model results, the assumptions of the model should be critically assessed in light of the data. Figure 2 shows that the assumptions of the multiple linear regression model are not perfectly satisfied in the Hundred Cities data set, since the residuals (the differences between model predictions and observations) depart systematically from the normal distribution that would be expected if the model correctly described the data. Thus, claims about associations or lack of associations based on multiple linear regression modeling are limited by the potential biases due to model misspecification.

Figure 2. Regression diagnostics reveal model misspecification

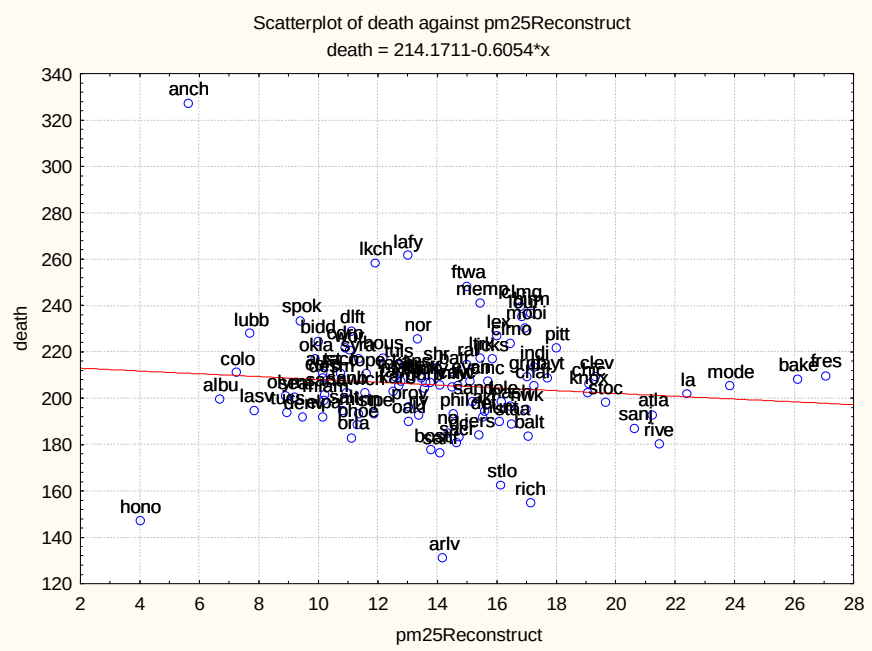



## Results for City-Specific C-R Associations

Figure 3 shows a scatter plot of average PM2.5 and mortality rate values for the cities in the Hundred Cities data base, illustrating the substantial heterogeneity in exposure levels and death rates. Average PM2.5 levels differ by more than a factor of 4

across cities. Average all-cause mortality rates among inhabitants over 75 years old also differ by more than 2-fold across cities.

Figure 3. Different cities have very different PM2.5 exposure levels and death rates




Such heterogeneity across cities suggests that fitting any single model to the pooled data from many cities may fail to capture differences in C-R relations in different locations. To explore this possibility, while avoiding assuming a linear relation between predictors and mortality rates, we applied the nonlinear, nonparametric modeling technique of classification and regression tree (CART) analysis[17] to data for each city separately. Classification tree analysis successively conditions on the levels of statistically significant predictors to quantify how the dependent variable (mortality rate) varies with different combinations of predictor values. In classification trees (generated by the commercial data mining program *KnowledgeSeeker*™[18]) for all-cause mortality rate, PM2.5 (*pm25Reconstruct*) is a significant, though weak (compared to other

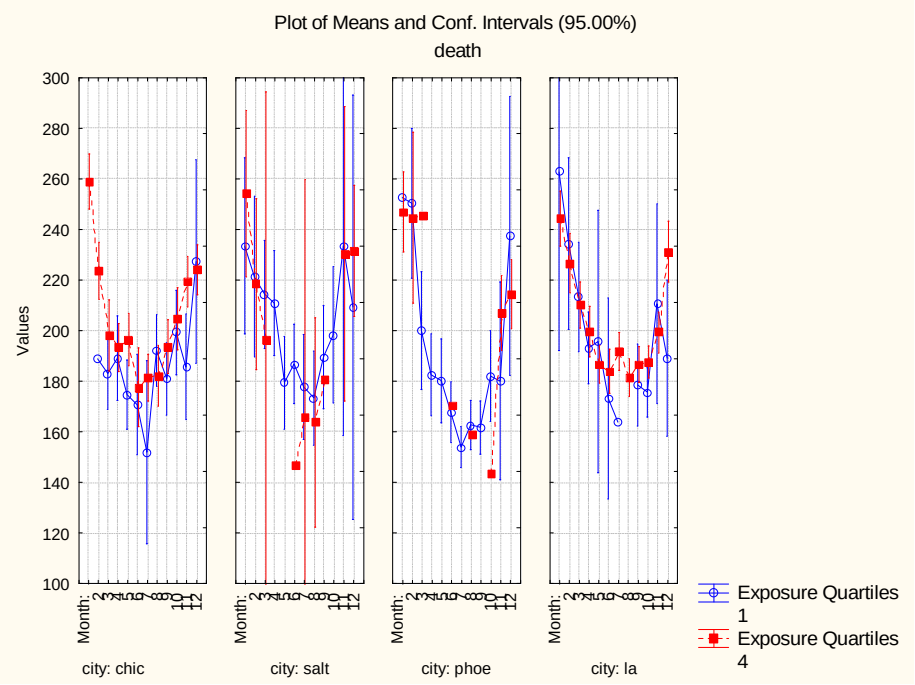
variables) predictor of all-cause-mortality when considered alone. However, in most cities, it provides no additional predictive power after conditioning on *Month* (the first split in the tree); that is, mortality rate is (at least approximately) *conditionally independent* of PM2.5 exposure levels, given the month[19]. Such conditional independence suggests that PM2.5 exposure level is not a cause of mortality rate[19-22], despite their statistical association, in most cities.

More specifically, 21 of the hundred cities – Bakersville, Birmingham, Chicago, Denver, El Paso, Fresno, Los Angeles, Las Vegas, Louisville, Minneapolis, New Orleans, New York, Phoenix, Sacramento, San Diego, San Jose, Salt Lake City, Santa Anna, Stockton CA, Topeka, and Tucson – showed significant associations ($p < 0.05$) between PM2.5 and mortality rate when the only candidate predictor considered was *pm25Reconstruct*. In Birmingham and Louisville, the associations are significantly negative (lower mortality risks at higher exposure levels); in Los Angeles and Topeka they are J-shaped (significantly lower at intermediate concentrations than at the low and high ends); at other locations, they are generally positive.

However, as suggested by Figure 1, and confirmed in Figure 5, month-of-year is an important confounder of the empirical C-R association. Indeed, after conditioning on *Month* in classification tree analysis, only 6 of the 21 cities still showed a significant C-R association with PM2.5: Chicago, El Paso, Los Angeles, New York, Stockton CA, and Topeka (with a J-shaped C-R curve for Topeka). Six out of the 100 cities might be expected to have a positive association due to chance, as we used a $p \leq 0.05$ level for the classification tree analysis; and/or due to other confounders not yet controlled for, such as income level and mean daily temperature.

To help visualize and interpret the findings from the classification tree analysis, Figure 4 shows average mortality rates by month for four individual cities, for both the highest PM2.5 exposure quartile (solid dots) and the lowest exposure quartile (open circles). In Chicago, which was one of six cities that still showed a positive association between PM2.5 exposure and mortality rate after conditioning on month, the majority of month-specific mortality rates are higher for days with high estimated exposure levels (quartile 4) than for days with low estimated exposure levels (quartile 1). (Classification tree analysis confirms this pattern without restricting the comparison to quartiles.)

Figure 4. Conditioning on *Month* eliminates the C-R relation in most cities




By contrast, in Phoenix and Salt Lake City, month-specific mortality rates are not systematically higher on high-exposure days than on low-exposure days; hence, these cities no longer show a clear positive C-R association after conditioning on Month.

Figure 5. *Month* is a strong, nonlinear confounder of PM2.5-mortality rate C-R relations

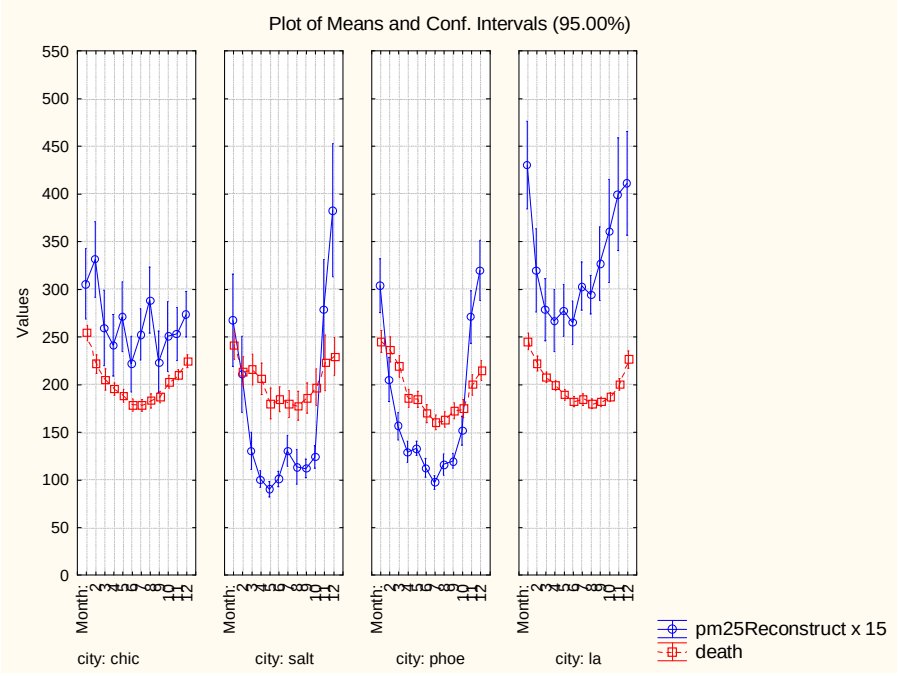



Figure 5 illustrates a general finding that helps to explain why only six cities showed a significant positive C-R relation after conditioning on *Month:* month-of-year (especially, the coldest winter months December-February) is a strong, non-linear confounder of the C-R relation. It is significantly associated with PM2.5 levels (which are multiplied by 15 in Figure 5 to display them on the same vertical scale as mortality rates) and also, independently, with mortality rates. The associations are U-shaped, with nadirs occurring in different months for different cities; hence they cannot be fully control for using linear modeling of aggregate data, as in Table 1. Comparing daily mortality rates for high- and low-exposure concentration days within the same month shows that the association between month and mortality rate is not mediated by the association between month and PM25 exposure: neither daily PM2.5, nor its lagged values, are significant predictors of mortality rates within a month.

Conditioning on month eliminates the C-R association between PM2.5 and mortality rates in most (all but six) cities. Conditioning on mean daily temperatures as well as on month eliminates all significant C-R associations, strongly suggesting that there is no detectable causal relation between them, since causal associations typically remain even after conditioning on other factors[19-22].

To facilitate independent reproduction of these conditional independence results by investigators without access to classification tree analysis programs, we used thae KnowledgeSeeker™ program to identify two clear confounders – *months*, as already illustrated, and also mean daily temperature (*tmpd*) – that jointly account for most or all of the association between estimated PM2.5 (*pm25Reconstruct*) levels and mortality rates. (Additional confounders, such as high income, which is associated with low exposure and low mortality rates, could also be included in a more complete causal model, but correcting for *months* and *tmpd* suffices to establish the main conclusion.) To test the alternative hypothesis that PM2.5 exposure is independently associated with increased mortality rates against the null hypothesis that it is not, we generated a cross-tabulation of mortality rates broken by the following combinations of variables: *city* (all), *month* (all), *tmpd* ($4^{th}$ quartile *vs.* $1^{st}$ quartile) and PM2.5 (*pm25Reconstruct*) ($4^{th}$ quartile *vs.* $1^{st}$ quartile). We quantified the fraction of all combinations for which the upper ($4^{th}$) quartile of exposure has a higher mortality rate than the lowest ($1^{st}$) quartile of exposure, for the same combination of *city*, *months* and *tmpd* quartile. (Thus, we controlled for these variables by conditioning on their levels, rather than by using linear regression to estimate, and subtract out, their statistical effects. Conditioning has the advantage that it allows for possible non-linearities and interactions.) The null hypothesis of no

independent effect of PM2.5 on mortality rate corresponds to a fraction of 0.5. The result of this non-parametric test was that 0.51 (95% CI [0.47, 0.55]) of the combinations had a greater mortality rate for the high-exposure days than for the matched low-exposure days; thus the evidence does not allow the null hypothesis to be rejected. (The mean difference between mortality rates on high-exposure vs. low-exposure days matched for city, month, and mean daily temperature also has a 95% confidence interval that includes zero, and hence leads to the same conclusion.) Any statistical program that performs multi-way cross-tabs can be used to reproduce these findings, even without classification tree analysis.

Interpretively, if these results are accepted at face value, it appears that the highly statistically significant association between PM2.5 and all-cause mortality in aggregate data (Table 1) is largely or wholly explained by the confounding effects of temperature and month at the individual city level. (The converse is not true: differences in PM2.5 levels do not explain away the effect of *Month*, as shown in Figure 1d.) These conclusions also hold for cardiovascular and other cause-specific mortality risks. Current and lagged PM2.5 values have no statistically significant positive associations with all-cause or cause-specific mortality rates, after controlling for weather-related and month variables. Thus, the significant statistical associations between PM2.5 and mortality rates do not appear to reflect a causal relation (as revealed by conditional independence tests) but rather reflect confounding by variables such as *Month* (Fig. 5) and temperature. (Other confounders, such as income, may also play a role, but are not needed to show that the statistical C-R association is already largely or wholly explained away just by minths, or by months and temperature.) Although it may be impossible to

prove a negative (no causal relation between PM2.5 and mortality rates), the large sample sizes and many cities in the NMMAPS data set suggest that any causal relation is too weak to easily detect, and that past reports of positive C-R associations between PM2.5 and mortality rates do not necessarily indicate that PM2.5 causes excess mortalities.

**Results for City-Specific Bayesian Model Averaging (BMA) of C-R Associations**

To independently test whether the strong positive C-R relation in Table 1 disappears when Bayesian Model Averaging is used to address model uncertainty, as some previous investigators have claimed[11,12], we applied a BMA algorithm, bicreg (Bayesian Model Averaging for Linear Regression Models) available in the R library – BMA (http://cran.r-project.org/web/packages/BMA/BMA.pdf) – to 200 randomly selected records from each of 101 cities and each age category (600 total for each city). All analysis was performed within the R 2.13 programming environment.

The BMA algorithm determines the fraction of models that fit the data – here, linear regression models – in which each of the candidate independent variables is a significant predictor of the dependent variable (here, mortality rates). Table 2 below shows the fraction of models in which the coefficient of *pm25Reconstruct* exceeded zero, averaged over all cities. (These only count cases where the posterior mean coefficient was nonnegative; it was significantly negative for several cities). A common, though, rough, interpretation of such fractions is that they indicate the probabilities that *pm25Reconstruct* would be included as a significant predictor of each mortality rate in

the correct model, if it were known. Consistent with the results of previous BMA investigations[11,12], Table 2 confirms that it is unlikely (probabilities range from 2% to 12%) that PM2.5 is a predictor of all-cause or cause-specific mortality rates when BMA is used to try to account for model uncertainty.

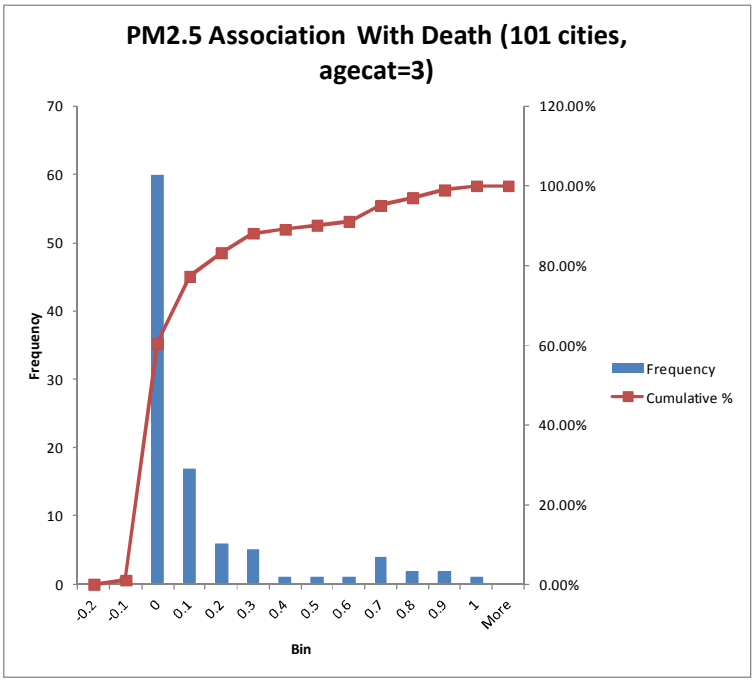
Table 2. BMA-derived probabilities of significant PM2.5 association averaged over 101 individual NMMAPS cities (nonnegative associations only)

| Indicator Variable | Age Category 1 | 2 | 3 | Average |
|---|---|---|---|---|
| accident | 0.07 | 0.06 | 0.04 | **0.06** |
| copd | 0.05 | 0.04 | 0.07 | **0.05** |
| cvd | 0.06 | 0.05 | 0.08 | **0.06** |
| death | 0.10 | 0.08 | 0.12 | **0.10** |
| inf | 0.02 | 0.02 | 0.02 | **0.02** |
| pneinf | 0.07 | 0.04 | 0.07 | **0.06** |
| pneu | 0.07 | 0.04 | 0.06 | **0.06** |
| resp | 0.03 | 0.04 | 0.07 | **0.05** |
| **Average** | **0.06** | **0.04** | **0.07** | **0.06** |

Figure 6 shows the histogram of probability values across 101 cities when *death* is the dependent variable and age category = 3 (i.e., over-75, which exhibited the highest overall average of 12.41%). Figure 6 also displays the significant negative associations (indicated as negative values of the inclusion probabilities), to better illustrate the spectrum of association. Over half the cities exhibit a zero (or negative) association between estimated PM2.5 (*pm25Reconstruct*) levels and all-cause mortality rate (*death*), even in this relatively elderly group (see cumulative distribution scale on the right side of Figure 6); and for over 90% of cities, it is more likely than not that

PM2.5 is not a predictor of mortality rate, consistent with previous claims about BMA analyses of PM2.5 and mortality[11,12].

Figure 6. BMA analysis (linear) shows that a positive coefficient for PM2.5 concentrations is unlikely in all but 10 cities.




(The ten cities with a probability of significant association greater than 50% in this analysis were: ny, roch, corp, stoc, oakl, buff, tamp, salt, tope, sanf.) Considering that BMA is based on linear regression modeling, whose limitations for this data set (e.g., incomplete control of strong nonlinear confounding) have already been discussed (see Figures 4 and 5), these BMA results are consistent with the hypothesis of no significant positive relation between PM2.5 concentrations and mortality.

**Testing for Time Series Causation**

The conclusion that conditional independence tests show no evidence of a causal relation between PM2.5 levels and mortality rates is certainly potentially controversial, and invites skeptical scrutiny, given that a large technical literature and important regulatory actions are predicated on the opposite conclusion[1-4]. A different analytic approach can be applied to cross-check whether the same conclusion is supported by more detailed time series analysis.

The intuition that genuine causes should (a) precede their effects; and (b) help to predict their effects, provides the conceptual basis for statistical tests of potential causation between two time series variables, such as PM exposure and mortality rate. Of these tests, one of the best developed and most widely applied is the Granger-Sims test, developed in econometrics for investigating potential causal relations in time series

To apply the Granger-Sims test to test for a possible causal C-R relation between PM2.5 and mortality rates, we first extracted from the NMMAPS city data sets all sequences of days for which *pm25mean* values were provided daily for at least 50 consecutive days. This produced 190 data sequences, distributed over 56 cities and various time periods within 1998-2000 (when regular data collection became more common). For each sequence, we performed a Granger-Sims statistical test for potential causality between *pm25Reconstruct* and the all-cause (except accidents) and seven cause-specific mortality rates calculated from the NMMAPS data set and Census data, using the granger.test (bivariate Granger causality testing for multiple time series) algorithm available in the R library – MSBVAR (http://rss.acs.unt.edu/Rdoc/library/MSBVAR/html/granger.test.html). (We modified the algorithm to also report the direction – positive or negative – of the gradient relationship

between changes in PM2.5 and changes in mortality rates.) We ran the test for 7 different maximum lags (the test includes each independent variable at each lag up to a maximum) for each age category in NMMAPS (a total of 3990 tests or 1330 for each age category). For each test, we recorded the p-value and the gradient direction.

Table 3 summarizes the fraction of these tests in which the p-value was less than 0.05 and the gradient relation between PM2.5 and mortality rate was positive. The fraction of such significant positive associations is what might be expected to occur by chance with a p-value of 0.05 for each individual test: around 5% or less, with an average value of 0.037. There is no evidence of a significant positive causal relation between PM2.5 exposures and mortality rates.

Table 3. Granger-Sims analysis shows no significant causal association between PM2.5 and mortality rates in daily time series

| | Age | | | |
|---|---|---|---|---|
| **Relation** | **< 65** | **65-74** | **75+** | **Average** |
| pm25Reconstruct -> accident | 0.044 | 0.058 | 0.048 | **0.050** |
| pm25Reconstruct -> copd | 0.029 | 0.038 | 0.050 | **0.039** |
| pm25Reconstruct -> cvd | 0.009 | 0.044 | 0.041 | **0.031** |
| pm25Reconstruct -> death | 0.023 | 0.036 | 0.051 | **0.037** |
| pm25Reconstruct -> inf | 0.002 | 0.022 | 0.014 | **0.013** |
| pm25Reconstruct -> pneinf | 0.052 | 0.041 | 0.036 | **0.043** |
| pm25Reconstruct -> pneu | 0.055 | 0.039 | 0.038 | **0.044** |
| pm25Reconstruct -> resp | 0.037 | 0.039 | 0.045 | **0.040** |
| **Average** | **0.031** | **0.040** | **0.041** | **0.037** |

(Fraction of sequences with p-value $\leq$ 0.05 and positive correlation)

## Discussion and Conclusions

One of the most fundamental scientific questions about air pollution health effects is whether changes in ambient levels of pollutants cause changes in mortality rates. A substantial body of literature has fit statistical models to various data sets, with different designs, and much of this literature concludes that there is a significant positive, approximately linear, statistical relation between ambient levels of PM2.5 and mortality rates in multiple cities. Other studies have concluded that much of this apparent association disappears when Bayesian Model Averaging is used to try to account for model uncertainty. We have independently tested and confirmed both findings in a publically available data set.

Multiple linear regression in the Hundred Cities data set strongly confirmed the presence of statistically significant positive correlations and linear regression coefficients for the C-R association between estimated daily levels of PM2.5 and estimated daily mortality rates (and cause-specific mortality rates, specifically cardiovascular disease mortality rates). However, such correlation is not necessarily evidence of causality. A new finding is that testing these significant statistical associations for potential causation using conditional independence tests yielded no evidence of a potential causal relation between PM2.5 and mortality rates. Instead, it appears that, in this large data set, the statistical association between PM2.5 and mortality rates is due to confounding by variables such as month and average daily temperature, with the winter months having both higher pollution levels and, independently of pollution level, higher death mortality.

Such confounding is hard to detect and control for in linear modeling (since the mortality rates are U-shaped functions of month), and regression diagnostics indicate

that linear models do not correctly describe the data-generating process. However, we confirmed that Bayesian Model Averaging (BMA) shows little association between PM2.5 and mortality rates (significant positive association is unlikely in 90% of cities). Moreover, the nonlinear, nonparametric technique of classification tree analysis confirmed that statistically significant C-R associations disappear after conditioning on city, time of year, and other variables, without requiring any linearity assumptions.

To further test this apparent lack of evidence for a causal relation, we applied Granger-Sims tests to examine whether changes in daily PM2.5 levels preceded, and helped to explain, changes in daily mortality rates. The results indicated no more than a random association between PM2.5 changes and changes in mortality rates, supporting the conclusion that, in these cities, the data give no reason to believe that changing PM25 levels would cause any effect on mortality rates, despite the statistical associations between them.

These results raise the obvious risk management and policy question of whether regulations intended to promote longer and healthier lives by further reducing PM2.5 levels would actually cause these desired effects. The analyses presented here find no evidence that they would do so. To the contrary, a simple, direct interpretation of the Hundred Cities data set is that it is consistent with previous findings that PM2.5 levels are statistically associated with mortality rates, but provides no indication that reductions in PM2.5 levels cause any reduction in mortality rates. If true, this conclusion has potential practical significance for informing policy debates over the likely costs and public health benefits of further reducing PM2.5 pollution levels in U.S cities.

## REFERENCES


1. Daniels MJ, Dominici F, Samet JM, Zeger SL. 2000. Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time-series for the 20 largest US cities. Am J Epidemiol. Sep 1;152(5):397-406.

2. Dominici F, Daniels MJ, Zeger SL, Samet JM. 2002. Air Pollution and Mortality: Estimating Regional and. National Dose-Response Relationships. Journal of the American Statistical Association. 97(457): 100-111

3. EPA (U.S. Environmental Protection Agency). 2011. The Benefits and Costs of the Clean Air Act from 1990 to 2020. Final Report – Rev. A. Office of Air and Radiation, Washington D.C.

4. Franklin M, Zeka A, Schwartz J. 2007. Association between PM2.5 and all-cause and specific-cause mortality in 27 US communities. J Expo Sci Environ Epidemiol. May;17(3):279-87.

5. Green LC, Armstrong SR. 2003. Particulate matter in ambient air and mortality: Toxicologic perspectives. Regul Toxicol Pharmacol. Dec;38(3):32

6. Schwartz J. 2007. Burying evidence: The Union of Concerned Scientists' Unscientific Claims About Air Pollution and Health. Available at http://joelschwartz.net/pdfs/BuryingEvidence011707.pdf\

7. Clyde M . Model uncertainty and health effect studies for particulate matter. Environmetrics. 2000 Nov-Dec; 11(6):745-63

8. NRC (National Research Council), Estimating the Public Health Benefits of Proposed Air Pollution regulations, Washington, D.C., Nat. Academy Press, 2002

9. GAO (Government Accountability Office), Particulate Matter: EPA has Started to Address the National Accademies' Recommendations on Estimating Health Benefits, but More Progress Is Needed, 2006, GAO-06-780, Washington, DC

10. Fraas, AG. The Treatment of Uncertainty in EPA's Analysis of Air Pollution Rules: A status report, 2011, J. Benefit-Cost Analysis, (2) 2:1-25

11. Koop G, McKitrick R, Tole L. 2007. Does air pollution cause respiratory illness? A new look at Canadian cities. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.164.6048

12. Koop G, Tole L. 2004. Measuring the health effects of air pollution: To what extent can we really say that people are dying from bad air? Journal of Environmental Economics and Management 47:30-54. See also: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.164.6048

13. Sheppard L, Burnett RT, Szpiro AA, Kim S-Y, Jerrett M, Pope CA, Brunekreef B. Confounding and exposure measurement error in air pollution. Air Quality, Atmosphere & Health 2011.

14. Oberdörster G Pulmonary carcinogenicity of inhaled particles and the maximum tolerated dose. Environ Health Perspect. 1997 Sep;105 Suppl 5:1347-55

15. Stoeger T, Reinhard C, Takenaka S, Schroeppel A, Karg E, Ritter B, Heyder J, Schulz H. Instillation of six different ultrafine carbon particles indicates a surface area threshold dose for acute lung inflammation in mice. Environ Health Perspect. 2006 Mar;114(3):328-33.

16. Pope CA 3rd, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. Circulation. 2004 Jan 6;109(1):71-7.

17. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med.* 2003 Dec;26(3):172-81.

18. Angoss Software, 2005. KnowledgeSeeker 5.0 Program. http://www.angoss.com/

19. Frey L, Fisher D, Tsamardinos I, Aliferis C, Statnikov A. Identifying Markov Blankets with Decision Tree Induction. 2003. http://citeseer.ist.psu.edu/frey03identifying.html

20. Ellis B, Wong WH. Learning causal Bayesian Network structures from experimental data. *Journal of the American Statistical Association.* 2008 June 103(482):778-789.

21. Shipley B. (2000). *Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations and Causal Inference.* Cambridge University Press. New York, NY. http://callisto.si.usherb.ca:8080/bshipley/my%20book.htm

22. Wood F, Griffiths TL, Ghahramani Z. (2006) A Non-Parametric Bayesian Method for Inferring Hidden Causes. In *Uncertainty in Artificial Intelligence* (UAI-2006): 536-543. http://www.sigmod.org/dblp/db/conf/uai/uai2006.html; http://citeseer.ist.psu.edu/758370.html

# Appendix A. Data Sources

### NMMAPS Data and Functions

- NMMAPS = National Mortality, Morbidity, and Air Pollution Study
- NMMAPSData R package version 0.4-2 downloaded from http://www.ihapss.jhsph.edu/publications/publication.htm
- Run on R version 2.9.2 (note NMMAPSData R package will not run on later versions of R)

### Specification of Relevant Counties and Aggregation into Cities

- Counties.txt file from NMMAPS: http://www.ihapss.jhsph.edu/data/NMMAPS/Data/data_download_gz.htm

### Poverty and Income Data

- By state and by county of the US
- Years available: 1989, 1993, 1995, 1997, 1998, 1999, 2000
- US Census State and Small Area Poverty Estimates downloaded from http://www.census.gov/did/www/saipe/data/statecounty/data/index.html
- Fields used:
    - 35- 38 Estimated percent of people of all ages in poverty
    - 134-139 Estimate of median household income
- Years 1990-1992, 1994, and 1996 are computed via linear interpolation between the nearest endpoint years.
- Years 87-88 used the values from 1989

### Hispanic Origin

- 1987-89
    - By Age (category), Year, and State of US
    - Race/Hispanic Origin Codes: 5 + 6 (White Hispanic + Black Hispanic)
    - US Census Estimates of the Population of States by Age, Sex, Race, and Hispanic Origin: 1981 to 1989 downloaded from http://www.census.gov/popest/archives/1980s/80s_st_detail.html.
    - To obtain county level data, we used the distribution by counties within each state in 1990 and applied the same distribution to state totals for 87-89. This was performed separately within each of the 3 age categories.
- 1990-99
    - By Age (category), Year, and US County
    - Hispanic Status as a separate code

    - Intercensal State and County Characteristics Population Estimates with 1990-Base Race Groups downloaded from http://www.census.gov/popest/datasets.html (each year in a separate file)
- 2000
    - By Age (category) and US County
    - Hispanic Status in combination with each race category
    - Modified Race Data Summary File 2000 Census of Population and Housing downloaded from http://www.census.gov/popest/archives/files/MRSF-01-US1.html.

### Race (Black and Other)

- 87-89
    - By Age (category), Year, and US County
    - Intercensal County Estimates by Age, Sex, Race: 1980-1989 downloaded from http://www.census.gov/popest/archives/1980s/PE-02.html.
- 90-99 (same as Hispanic Origin above)
- 2000 (same as Hispanic Origin above)

#### Educational Attainment and Urban vs. Rural

- Year 2000 only for each city
- Extracted from NMMAPS database