Model Choice in Time Series Studies of Air Pollution and Health

Roger D. Peng, PhD

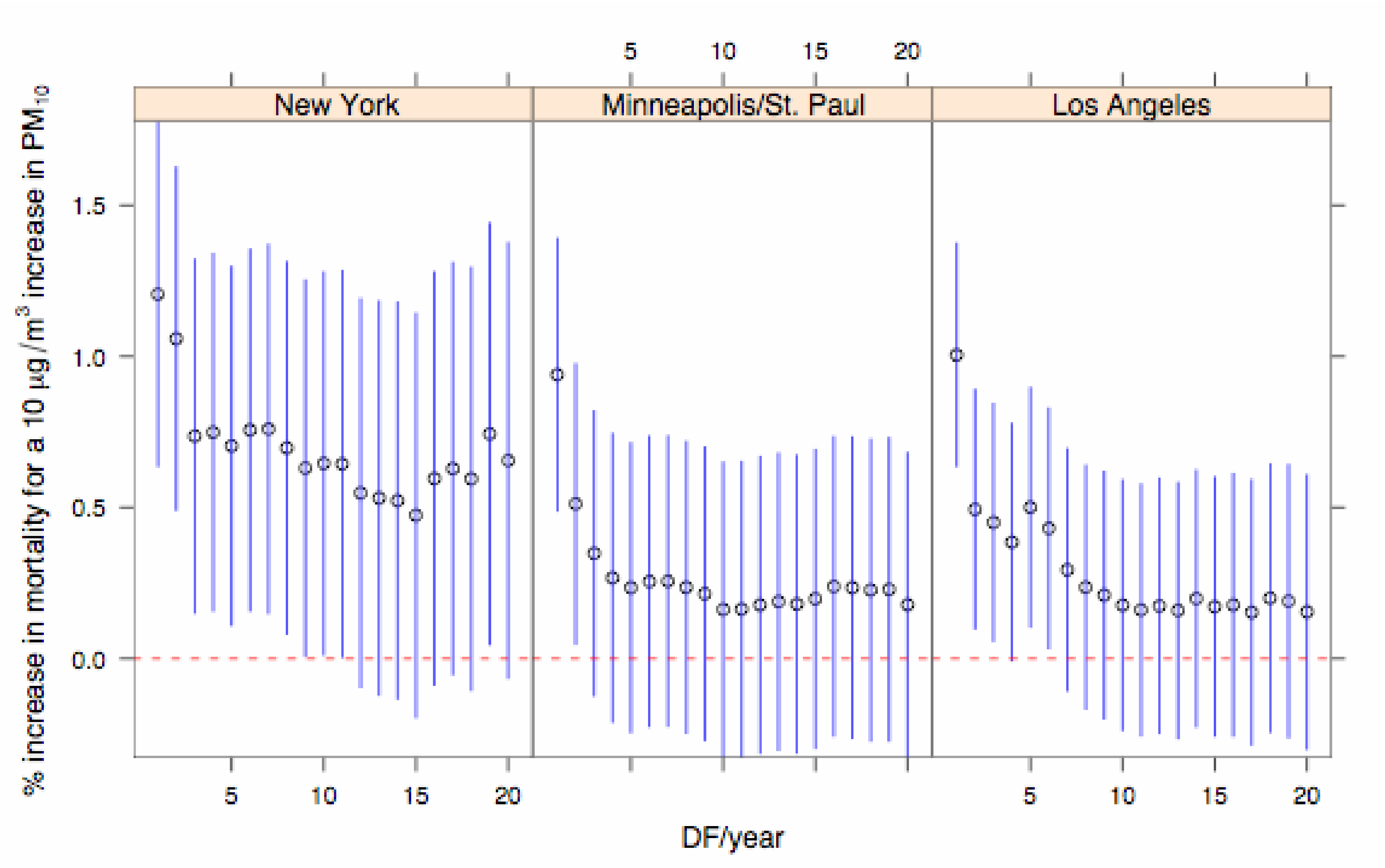
Department of Biostatistics

Johns Hopkins Blomberg School of Public Health

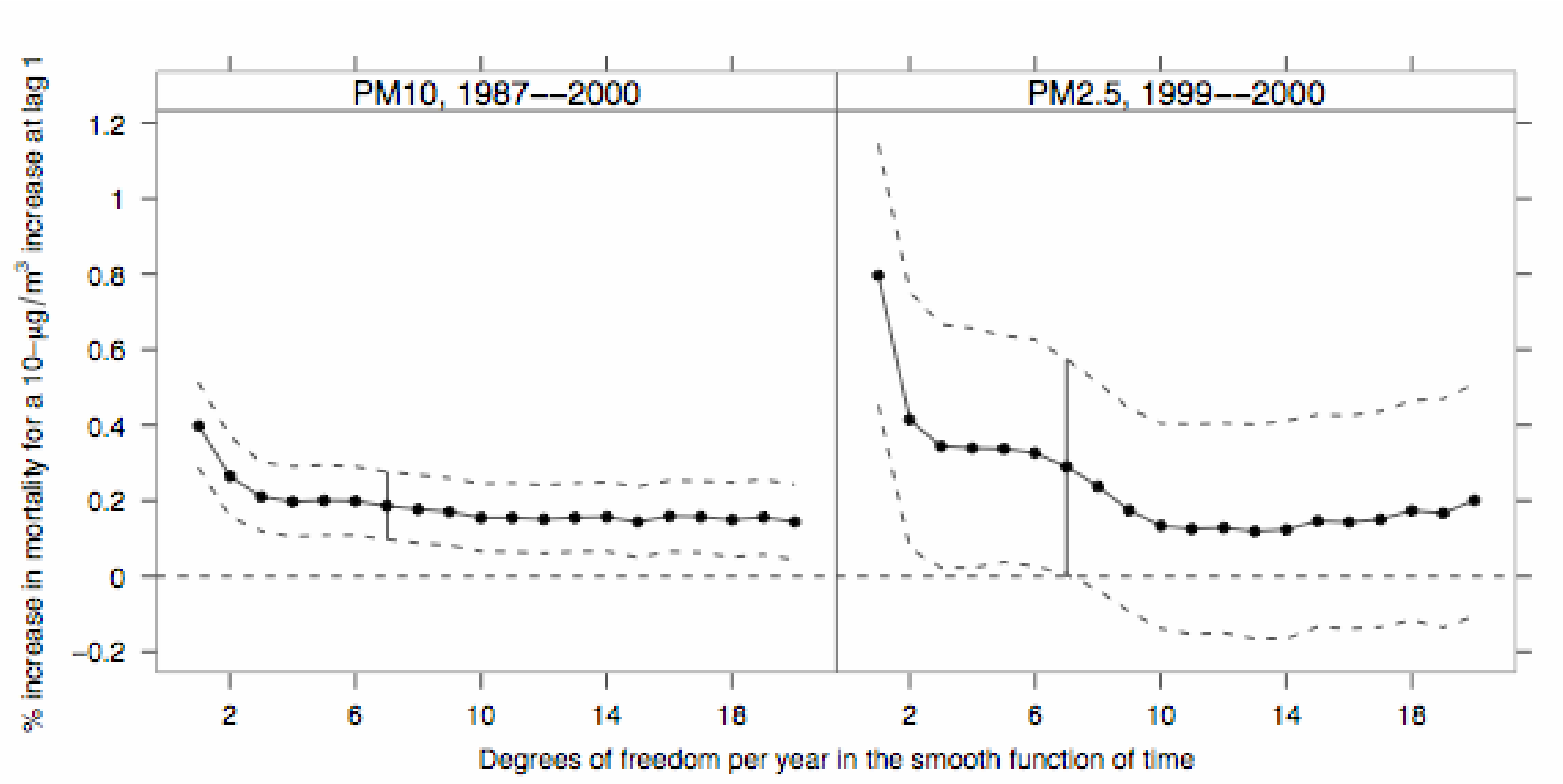
APHA 2007

Sponsors: NIEHS, EPA, Health Effects Institute

City-specific estimates, PM_{10} and mortality, 1987-2000

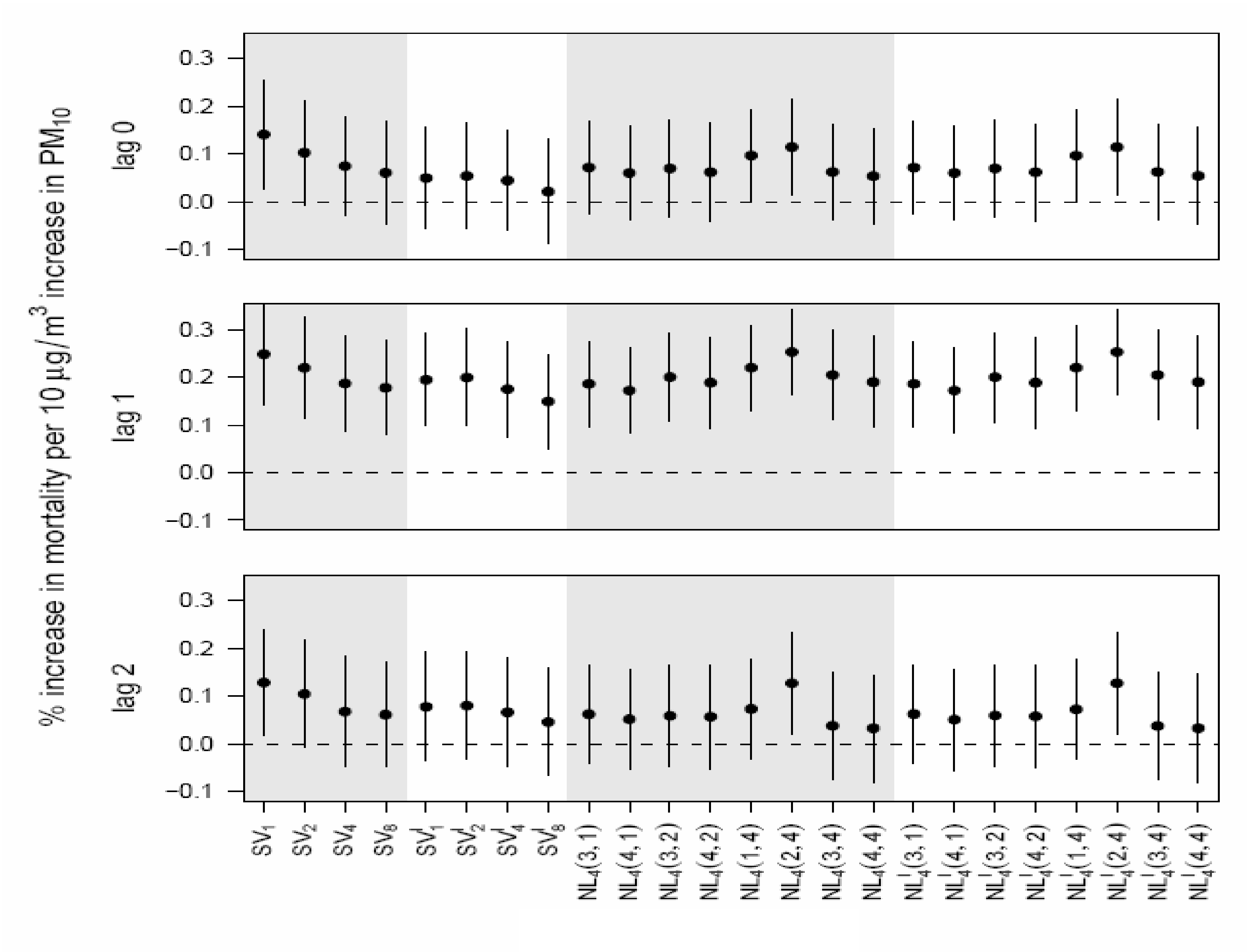


PM_{10}, $PM_{2.5}$ and Mortality, NMMAPS, 100 Cities



Dominici, *et al* 2007

PM_{10} and Mortality: Sensitivity of the National Average Estimate to Adjustment for Weather (Welty, *et al* 2005)

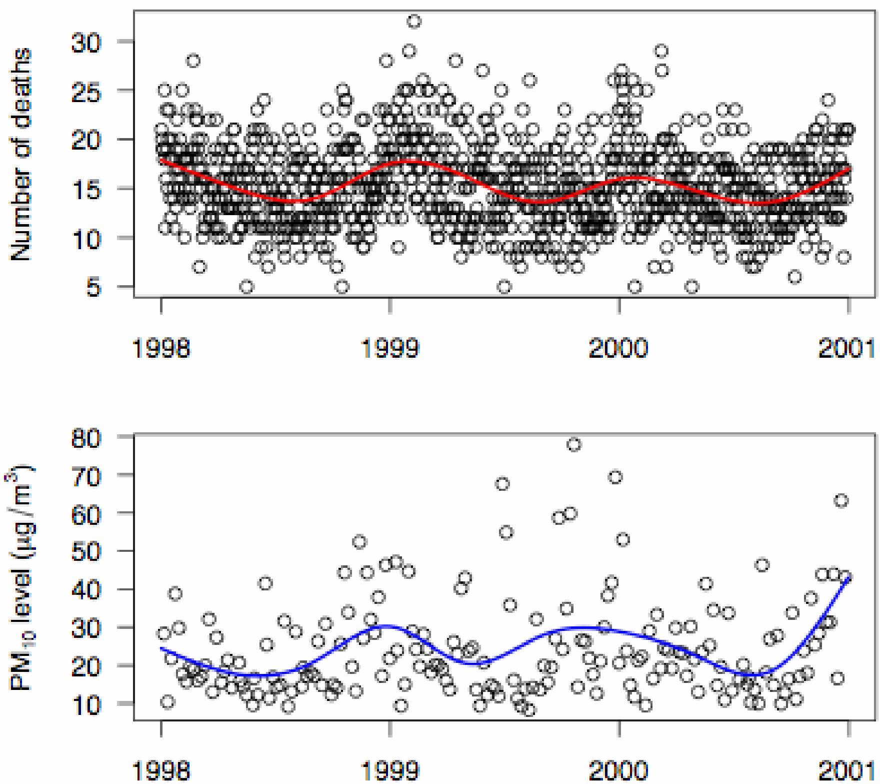


Which Estimate?

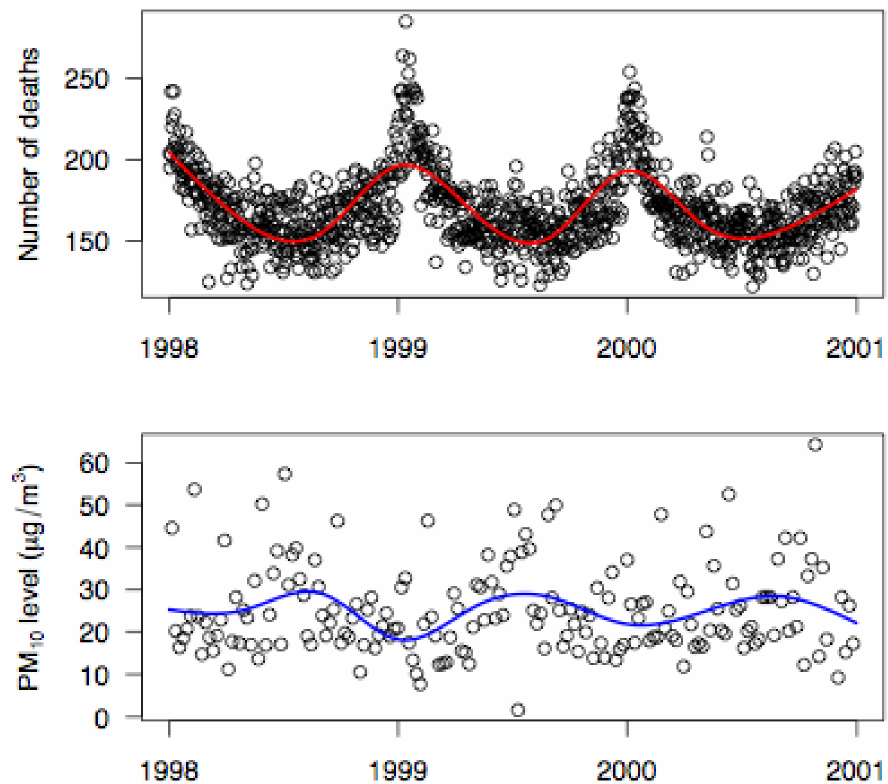
- In time series studies of air pollution and health, model choice is important
- Risk is difficult to estimate, signal-to-noise is weak
- Models are complex and risk estimates can be sensitive
- Risk estimates have substantial policy impact

Some air pollution and mortality data

Mortality and PM_{10} in San Francisco, 1998--2000



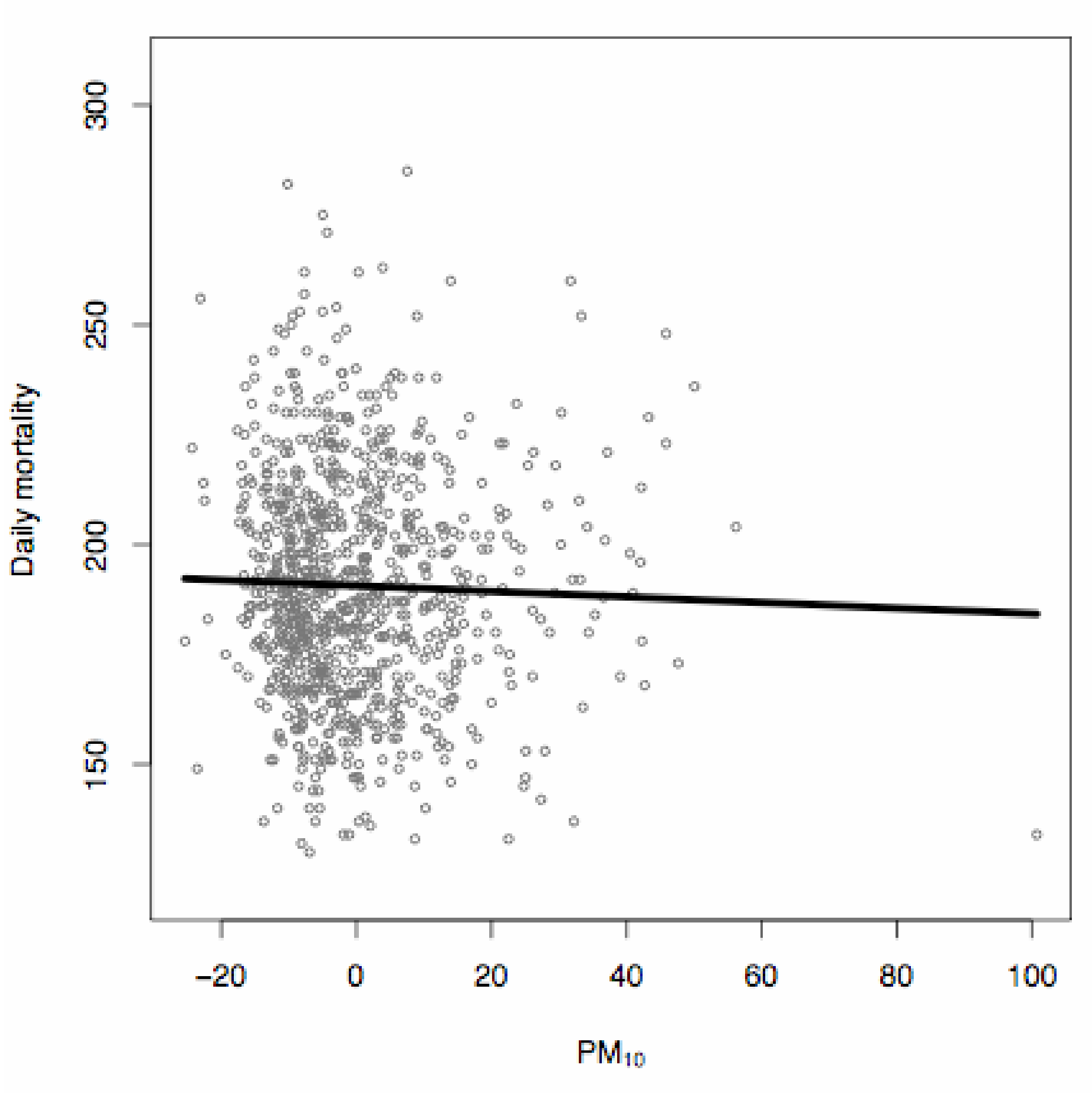
Mortality and PM_{10} in New York City, 1998--2000

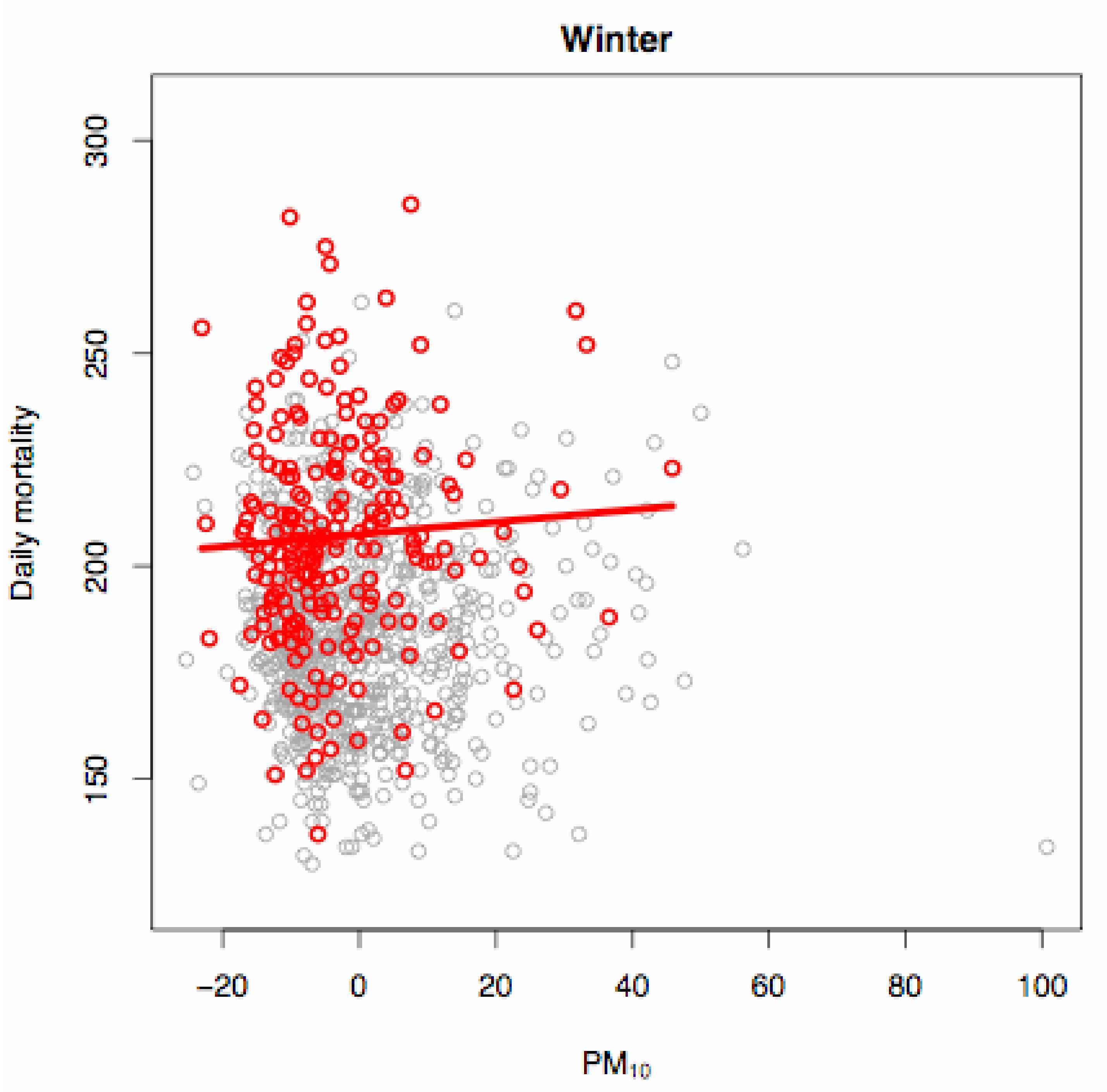


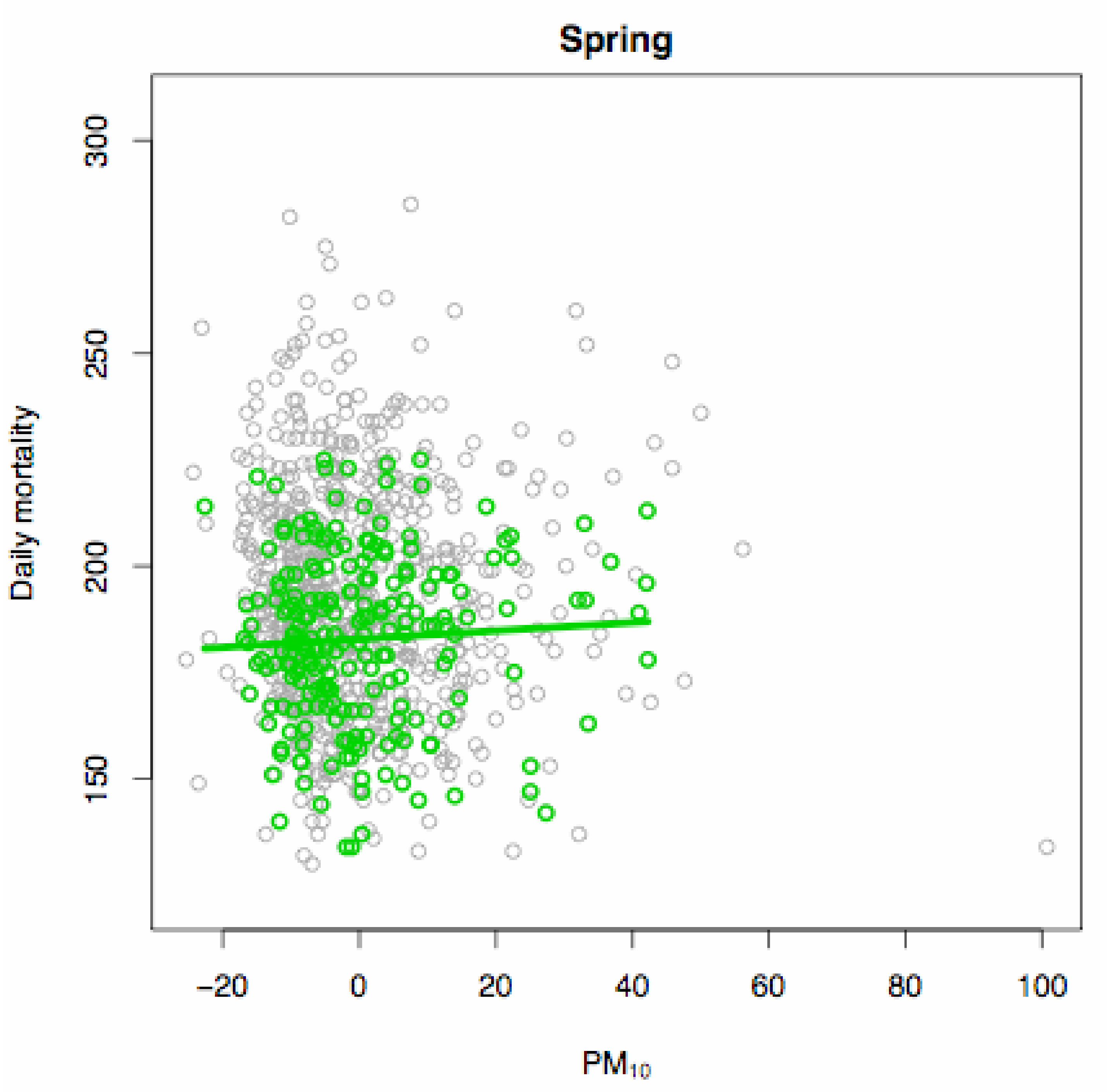
Potential Confounders in Air Pollution and Health Studies

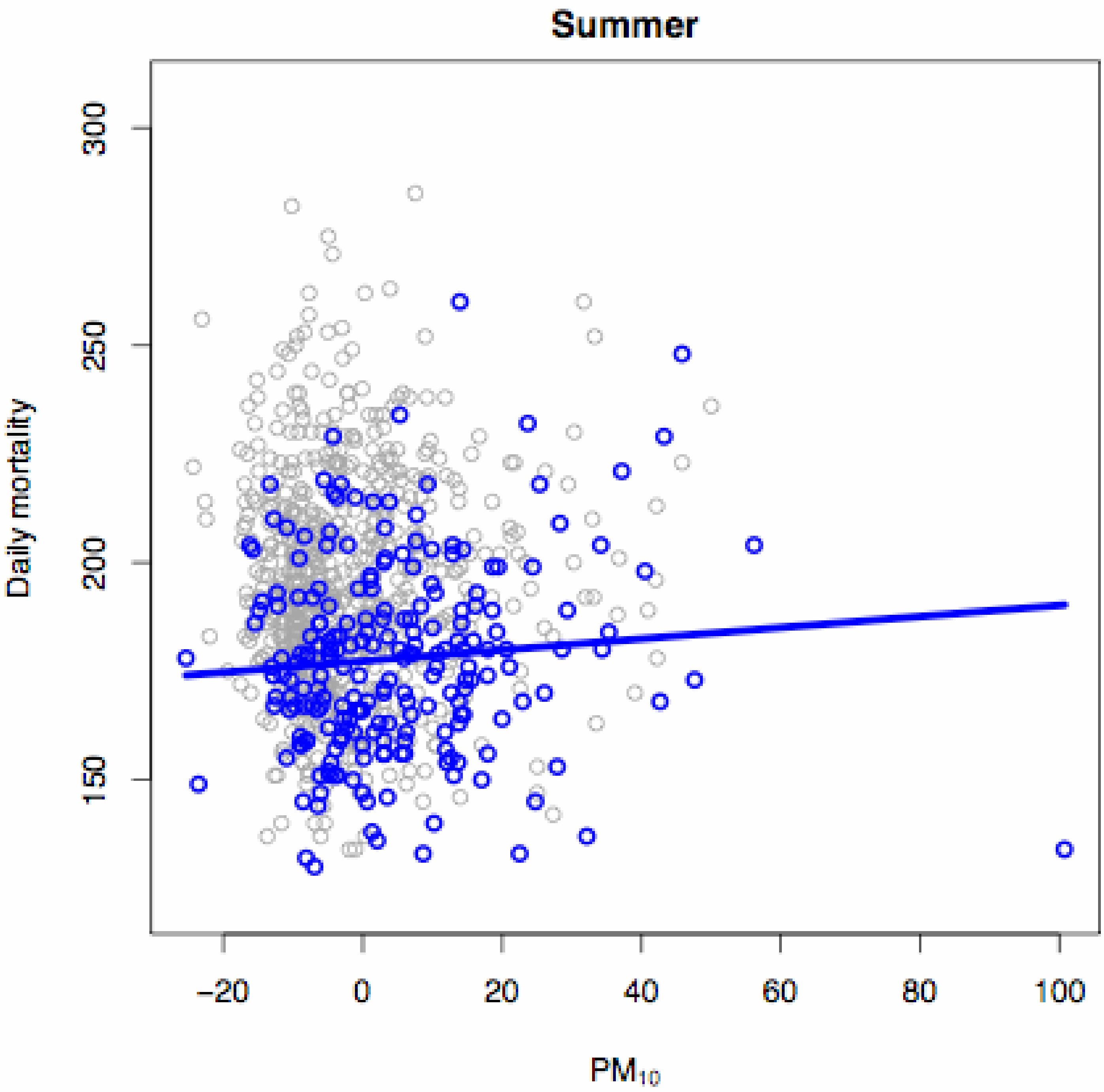
- Smoothly varying seasonal trends
- Long-term trends
  - structural changes in overall population
- Temperature
  - mortality: “J-shaped” relationship
  - PM_{10}: increasing/positive
- Humidity
- Other time-varying factors?

Confounding by Season: New York City

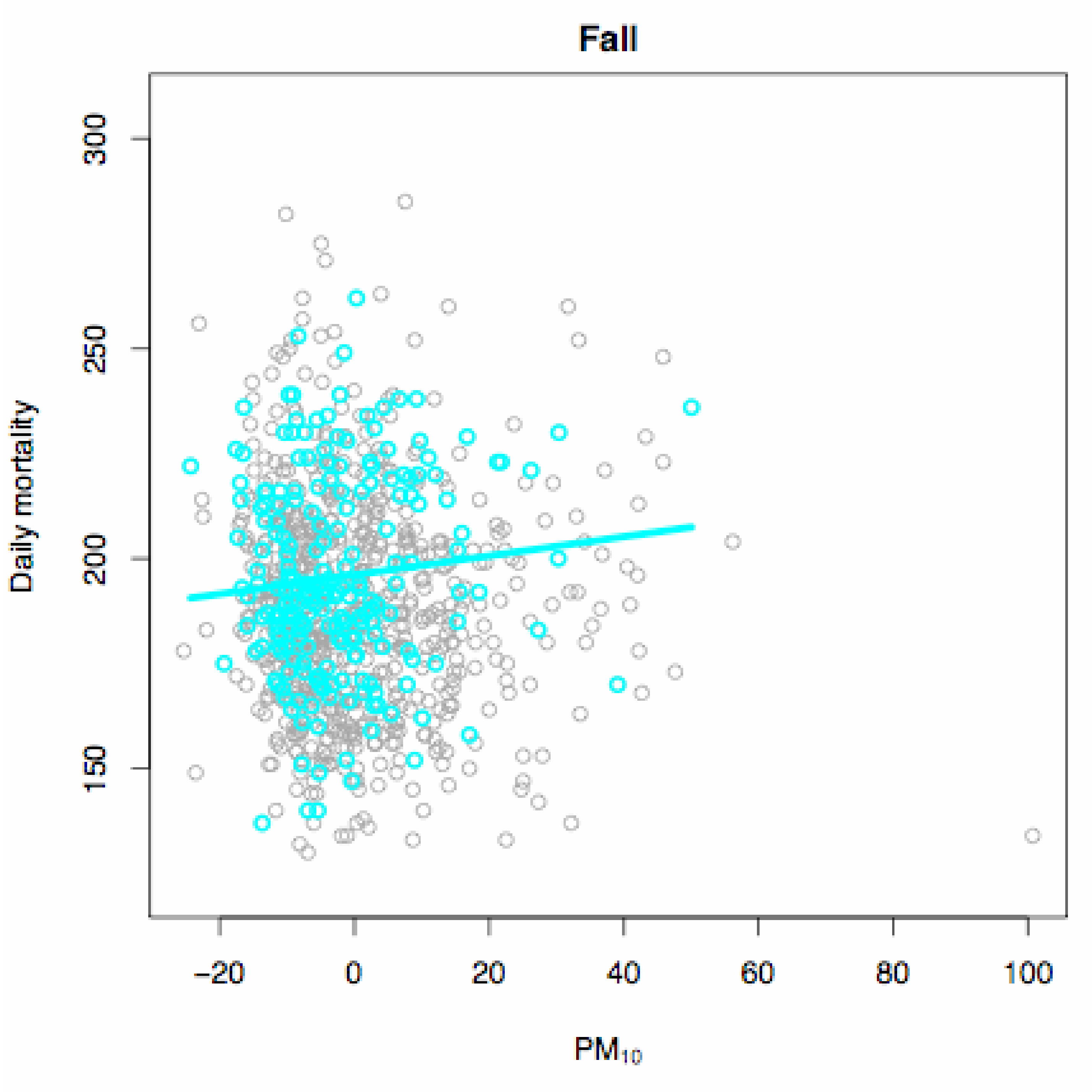



Winter
Daily mortality
300
250
200
150
−20
0
20
40
60
80
100
PM10


Spring
Daily mortality
300
250
200
150
−20
0
20
40
60
80
100
PM10

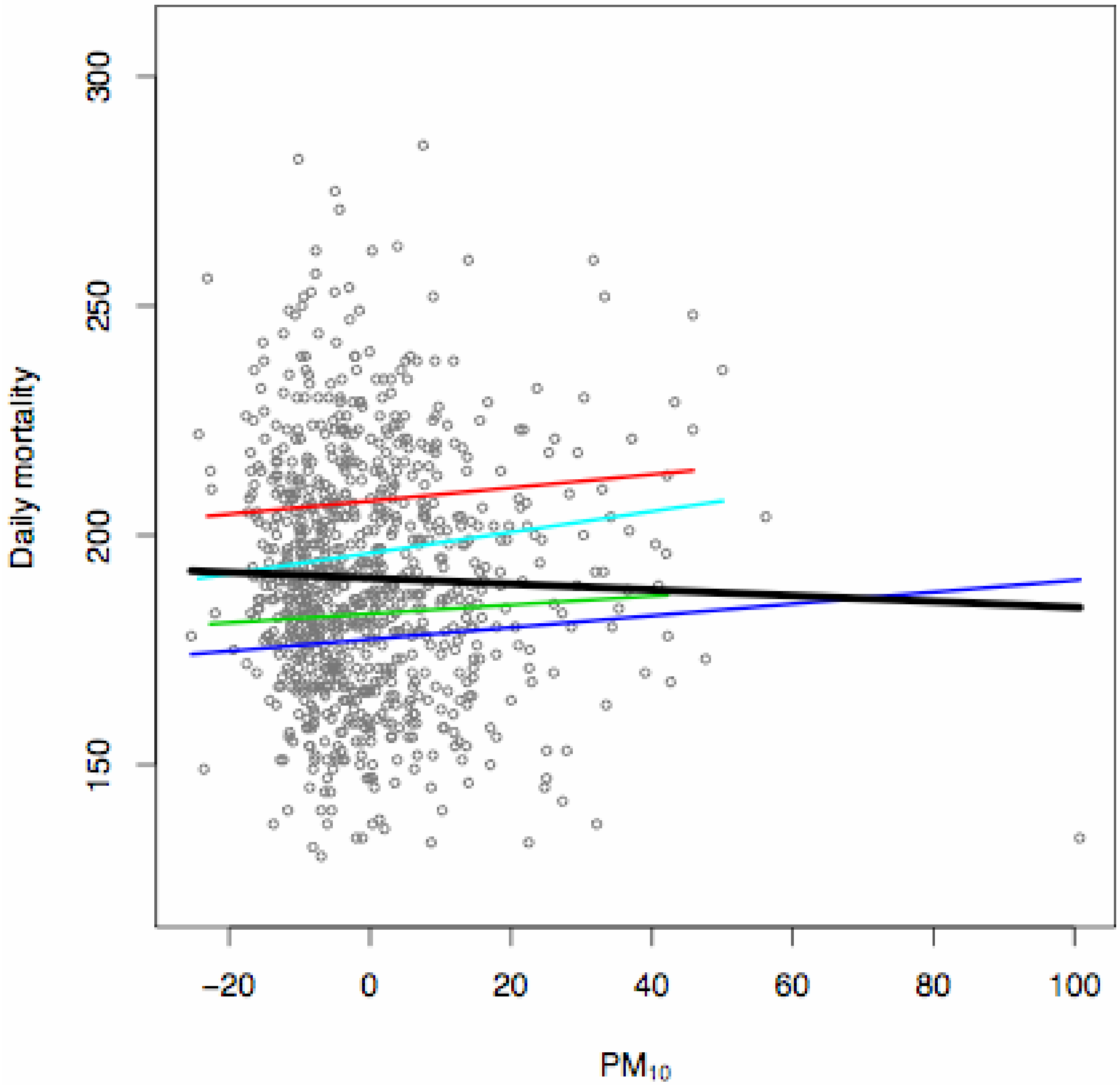


Summer
Daily mortality
300
250
200
150
−20
0
20
40
60
80
100
PM10



Fall
Daily mortality
300
250
200
150
−20
0
20
40
60
80
100
PM10

Season-specific associations are positive, overall association is negative



Stage 1: City-specific model

Poisson regression model

$$\begin{aligned} Y_t^c &\sim \text{Poisson}(\mu_t^c) \\ \log \mu_t^c &= \beta^c \boxed{x_{t-\ell}} + \boxed{s(\text{time};\, df_1)} + \boxed{s(\text{temp}_t;\, df_2)} + \cdots \end{aligned}$$

Pollutant series

Seasonal, long-term trends

Weather

Simulation Study Model of Mortality and Air Pollution

- Mortality

$$y_t \sim x_t + f(t) + \varepsilon_t$$

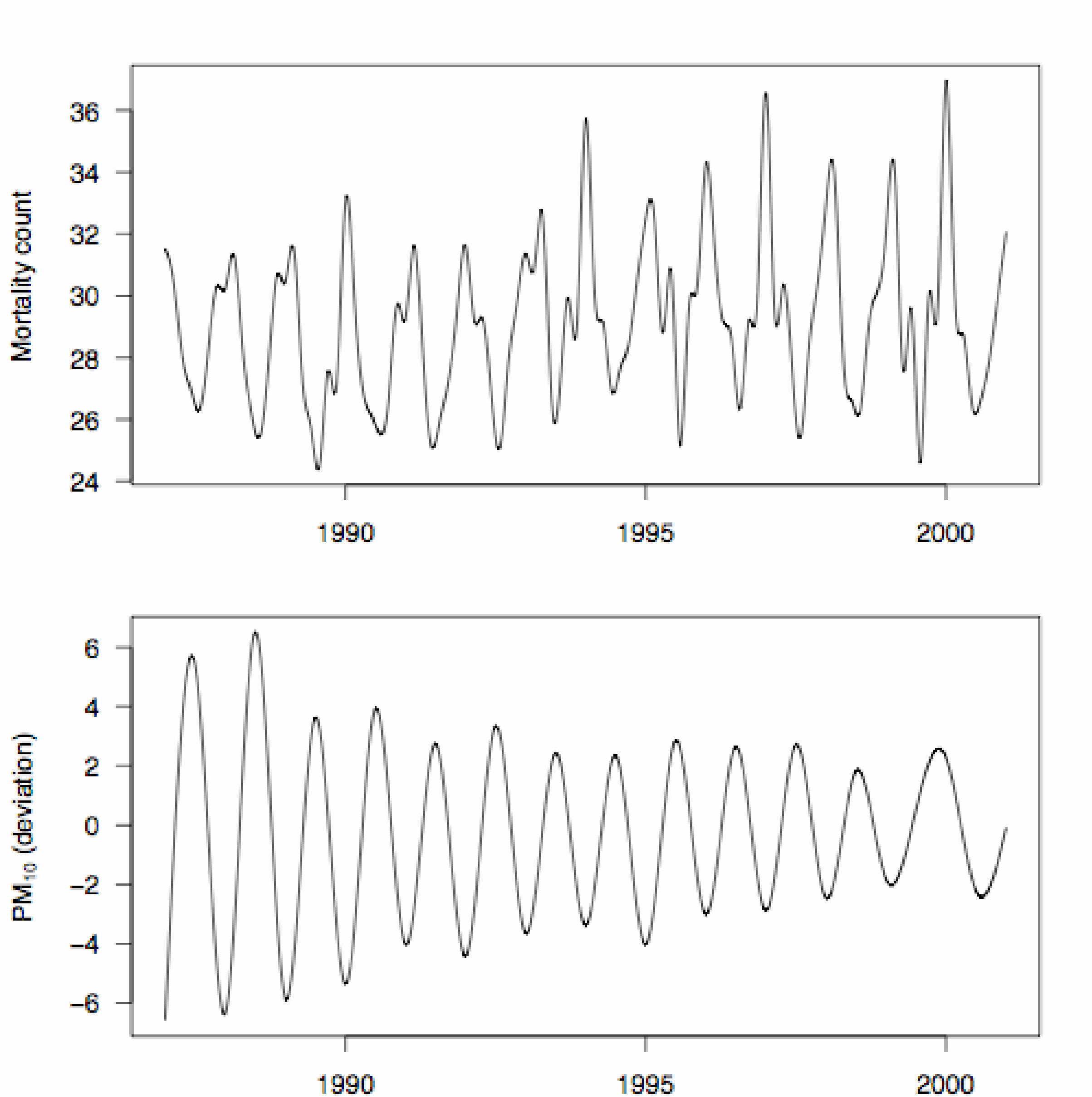
- Air pollution

$$x_t \sim g(t) + \delta_t$$

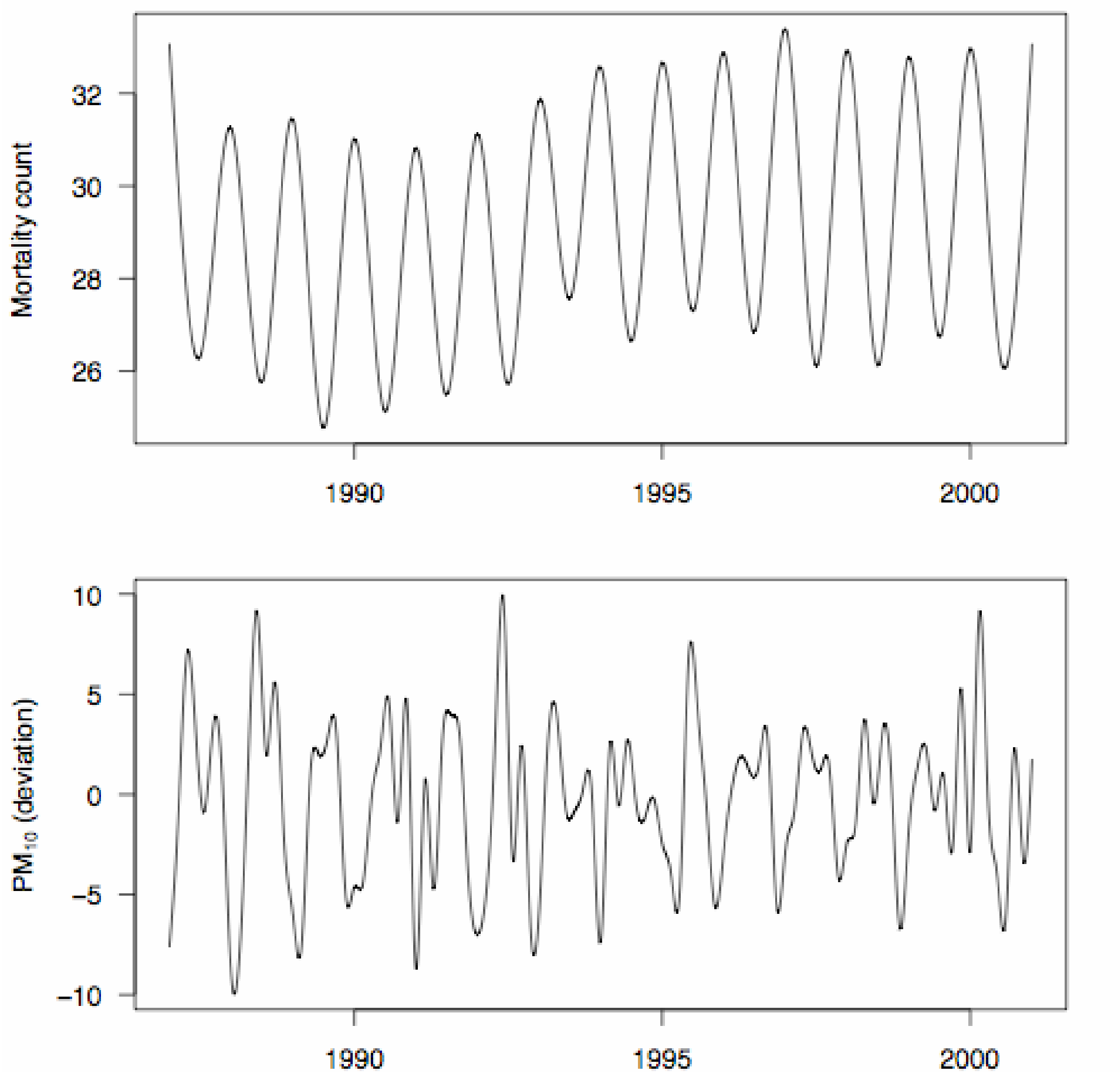
What is the relationship between *f* and *g*?

Theory indicates *f* should have smoothness required to best predict x_t

Scenario 1


Mortality count
36
34
32
30
28
26
24
1990
1995
2000
PM10 (deviation)
6
4
2
0
−2
−4
−6
1990
1995
2000

Scenario 2



Mortality count
32
30
28
26
1990
1995
2000
PM10 (deviation)
10
5
0
−5
−10
1990
1995
2000

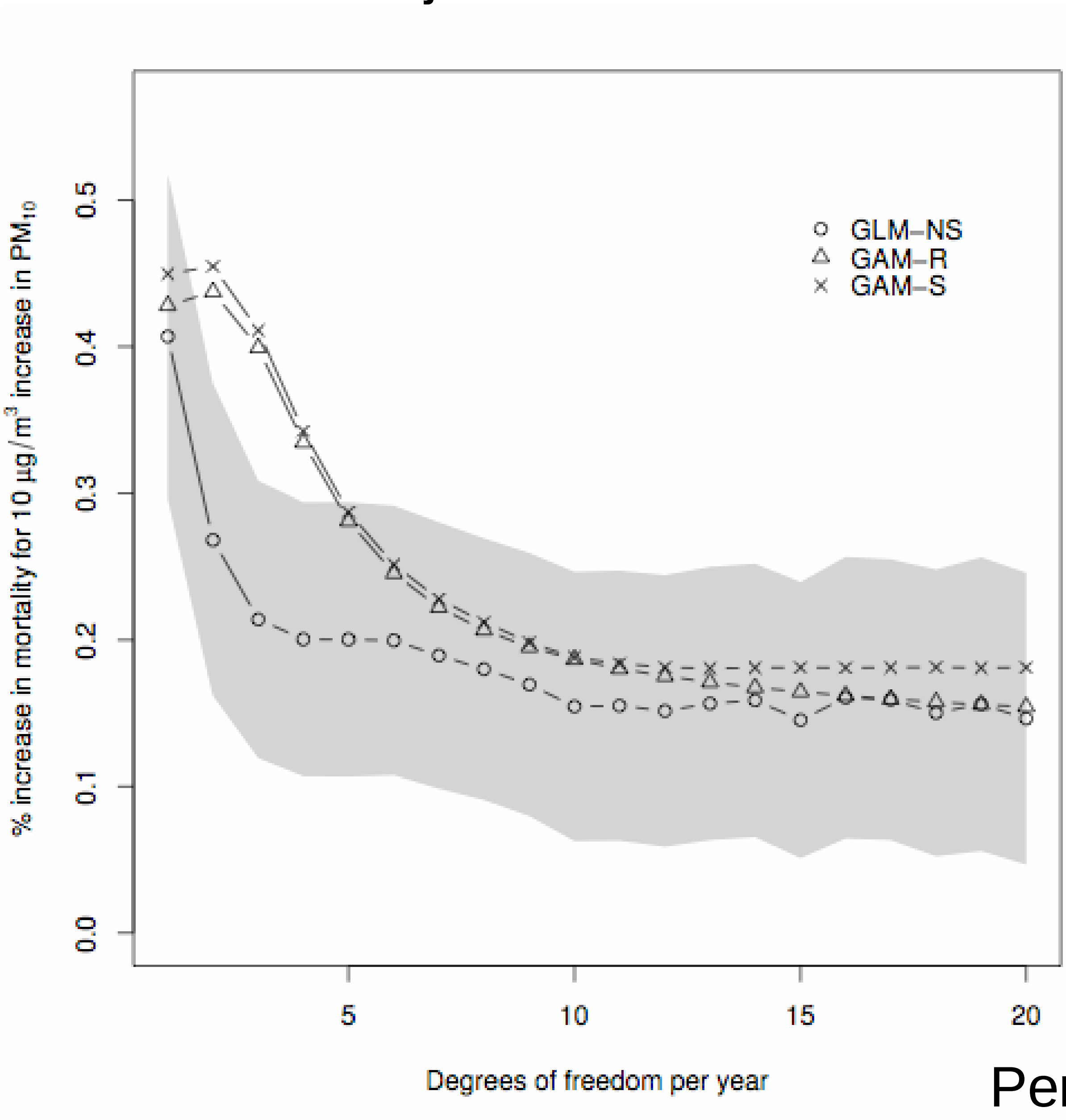
Comparison of model formulations in the literature

- Representing s()
  - smoothing splines
  - penalized splines
  - natural splines
- Choosing df
  - minimize AIC (best predict mortality)
  - minimize residual autocorrelation (via PACF)
  - best predict pollution (GCV-PM_{10})

Results of simulation study

- Bias drops off as df increases
- Variance increases a little with df
- Nonparametric smoothers require more df to remove bias
- AIC and PACF methods more biased in Scenario 2
- Predicting x_t via GCV had best performance overall (w.r.t. mean squared error)

PM_{10} and Mortality: Sensitivity of National Average Estimates to Adjustment for Seasonal Trends



Peng, *et al* 2006

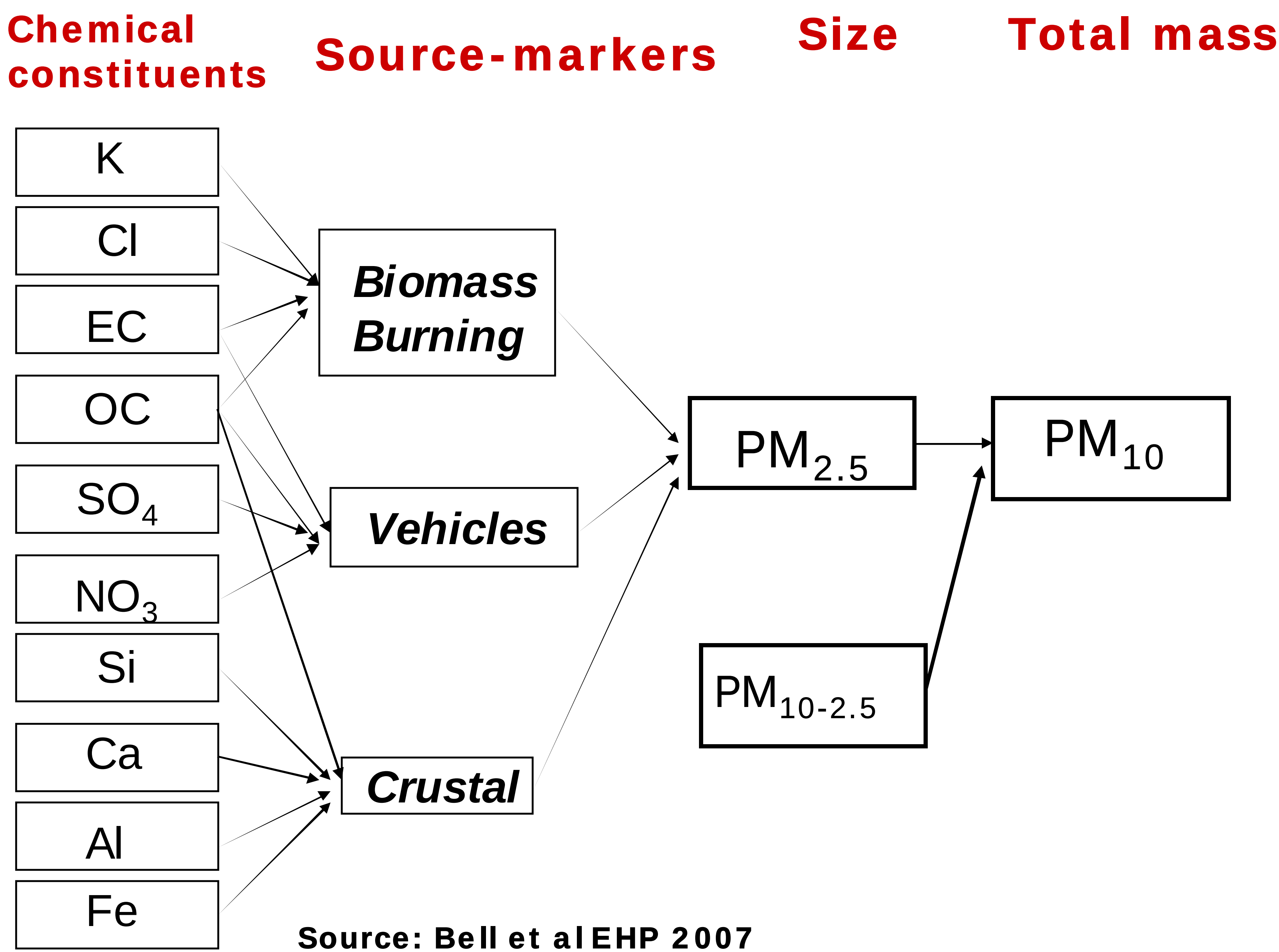
National Ambient Air Quality Standards: Statistical research has an impact

- **From US EPA NAAQS Criteria Document 1996:** ***"Many of the time-series epidemiology studies looking for associations between ozone exposure and daily human mortality have been difficult to interpret because of methodological or statistical weaknesses, including the a failure to account for other pollutant and environmental effects."***

- **From US EPA Criteria Document 2006:** ***"While uncertainties remain in some areas, it can be concluded that robust associations have been identified between various measures of daily ozone concentrations and increased risk of mortality."***

What Next?

- What are the mechanisms of PM toxicity?
  – size, chemical component, source
- Data for ~60 chemical components are available but are sparse in time and space
- Individual constituents can be highly correlated with each other
- Interactions are potentially important and difficult to identify
- Sources are not measured
- Confounding issues are more complex


Chemical constituents
Source-markers
Size
Total mass
K
Cl
EC
OC
SO_4
NO_3
Si
Ca
Al
Fe
Biomass Burning
Vehicles
Crustal
$PM_{2.5}$
PM_{10}
$PM_{10-2.5}$
Source: Bell et al EHP 2007

What Next?

- In an increasingly complex setting, we need to be able to *see the evidence* in the data
  - Avoid lumping results together and providing “the answer” (although sometimes we need a number)
- Reproduce published findings
- Allow others to examine the data, check models and test assumptions
- Enable development of new models/methods
- We need *reproducible research*

Tools for Distributing Reproducible Research

- Doing research is primary, distributing it is often considered secondary
  - Is publishing and article enough?
- We lack an infrastructure for easily distributing statistical analyses to others (we often start from scratch each time)
- Need to automate the distribution process
- Allow people to work in their natural environment

Summary

- Time series models of air pollution and health are potentially confounded seasonal trends
- Risk estimates can vary over a wide range depending on the model chosen
- National average estimates from multi-site studies are generally robust to model choices
- Reproducible research is needed to allow others to assess the evidence and provide transparency

Collaborators

- Francesca Dominici
- Tom Louis
- Aidan McDermott
- Luu Pham
- Ron White
- Jonathan Samet
- Scott Zeger
- Michelle Bell (Yale)
- Keita Ebisu (Yale)
- Leah Welty (NWU)

Stage 1: City-specific model

Poisson regression model

$$
\begin{aligned}
Y_t^c &\sim \text{Poisson}(\mu_t^c) \\
\log \mu_t^c &= \beta^c \, x_{t-\ell}^c + \text{DOW}_t + \text{AgeCat} \\
&\quad + s(\text{temp}_t;\, df_1) + s(\text{temp}_{t,1-3};\, df_2) \\
&\quad + s(\text{dew pt}_t;\, df_3) + s(\text{dew pt}_{t,1-3};\, df_4) \\
&\quad + s(t;\, df_5) + s(t;\, df_6) \times \text{AgeCat}
\end{aligned}
$$

Stage 1: City-specific model

Poisson regression model

Pollutant series

$$
\begin{aligned}
Y_t^c &\sim \text{Poisson}(\mu_t^c) \\
\log \mu_t^c &= \boxed{\beta^c\, x_{t-\ell}^c} + \text{DOW}_t + \text{AgeCat} \\
&\quad + s(\text{temp}_t;\, df_1) + s(\text{temp}_{t,1-3};\, df_2) \\
&\quad + s(\text{dew pt}_t;\, df_3) + s(\text{dew pt}_{t,1-3};\, df_4) \\
&\quad + s(t;\, df_5) + s(t;\, df_6) \times \text{AgeCat}
\end{aligned}
$$

Stage 1: City-specific model

Poisson regression model

Weather

$$
\begin{aligned}
Y_t^c &\sim \text{Poisson}(\mu_t^c) \\
\log \mu_t^c &= \beta^c x_{t-\ell}^c + \text{DOW}_t + \text{AgeCat} \\
&\quad + s(\text{temp}_t;\, df_1) + s(\text{temp}_{t,1-3};\, df_2) \\
&\quad + s(\text{dew pt}_t;\, df_3) + s(\text{dew pt}_{t,1-3};\, df_4) \\
&\quad + s(t;\, df_5) + s(t;\, df_6) \times \text{AgeCat}
\end{aligned}
$$

Stage 1: City-specific model

Poisson regression model

$$
\begin{aligned}
Y_t^c &\sim \text{Poisson}(\mu_t^c) \\
\log \mu_t^c &= \beta^c\, x_{t-\ell}^c + \text{DOW}_t + \text{AgeCat} \\
&\quad + s(\text{temp}_t;\, df_1) + s(\text{temp}_{t,1-3};\, df_2) \\
&\quad + s(\text{dew pt}_t;\, df_3) + s(\text{dew pt}_{t,1-3};\, df_4) \\
&\quad + \boxed{s(t;\, df_5) + s(t;\, df_6) \times \text{AgeCat}}
\end{aligned}
$$

Seasonal and long-term trends

Designing Tools for the Research Pipeline