Model Uncertainty in Time Series Studies of Air Pollution and Health

Francesca Dominici
Department of Biostatistics
Bloomberg School of Public Health
Johns Hopkins University

joint work with Aidan McDermott, Thomas Louis, Giovanni Parmigiani, Scott L. Zeger, Trevor Hastie
Jonathan M. Samet

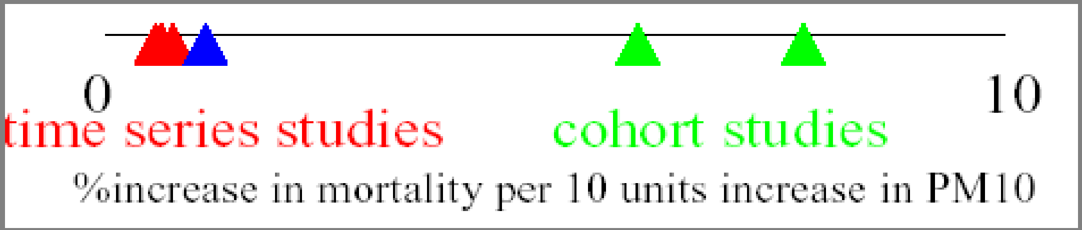
Outline

- **The Social, Political, and Regulatory Context**
- **Model Choice in Time Series Studies**
  - **Adjustment for Confounding bias**
  - **Model Uncertainty**
  - **Model Averaging**
- **Time Series and Cohort Studies**
- **ESTimating Chronic and Acute Pollution Effects Study (ESCAPES): a new Spatio-Temporal epidemiological design**

Time series and cohort studies

- Time Series studies estimate association between *probability of death* and the level of air pollution *shortly* before death (*shorter-term effects*)
- Exposure: day-to-day variations in air pollution
- *temporal variability is used to estimate acute effects associated with shorter-term exposure*

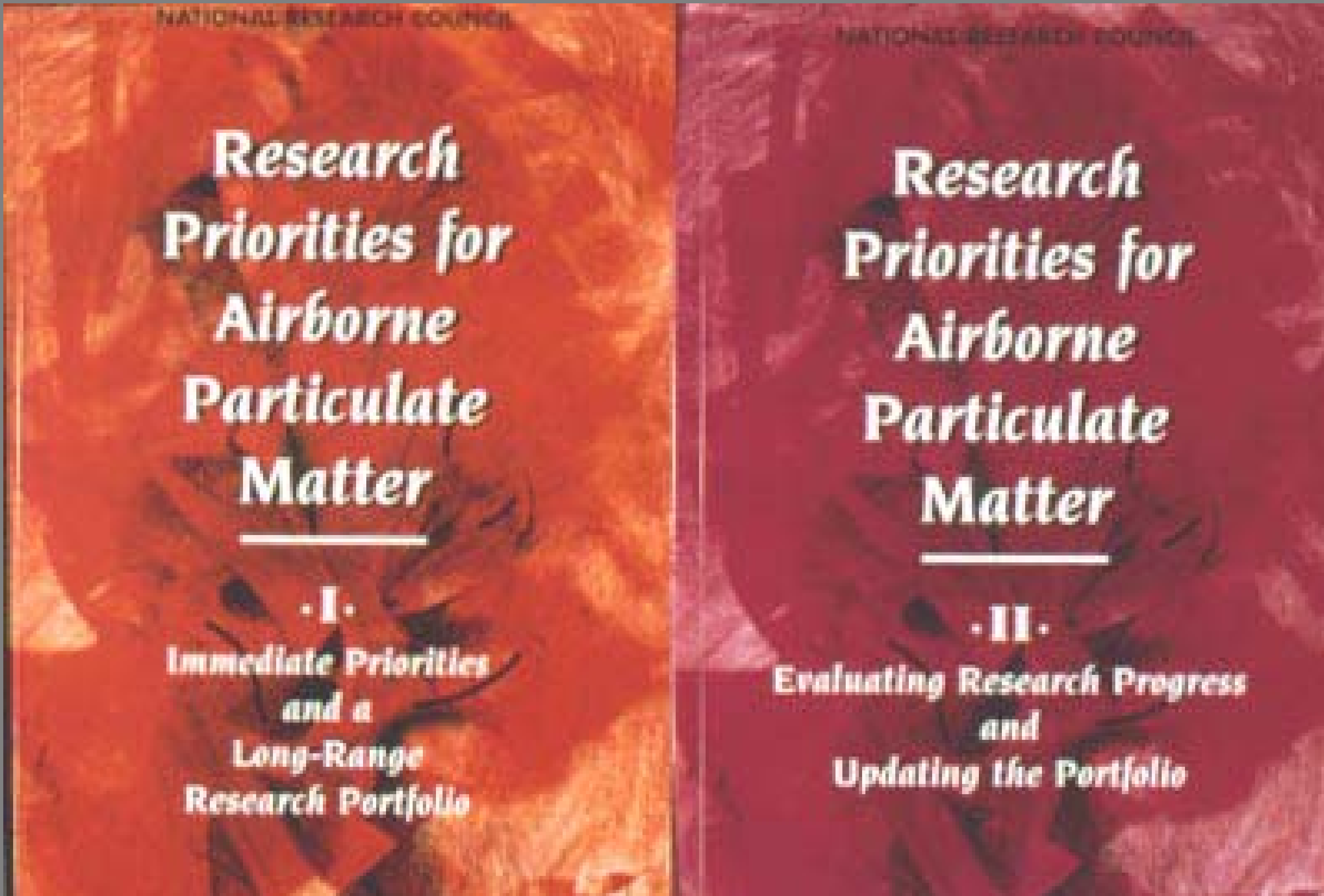
- Cohort studies estimate association between *time-to-death* and exposure to air pollution in a *lifetime (longer-term effects)*
- Exposure: city-to-city long-term average
- *spatial variability is used to estimate chronic effects associated with longer-term exposure*



Social, Political, and Regulatory Context

- EPA's process for review of the National Ambient Air Quality Standard (NAAQS) creates a sensitive political context
- Approx $100 million annually are spent in the United States alone to address uncertainties in the understanding of the health effects of particulate matter
- Several expert committees are created such as: Clean Air Act Scientific Advisory Board (CASAC) of the EPA, committees of National Academy of Sciences, World Health Organization and many others

Science to reduce uncertainty



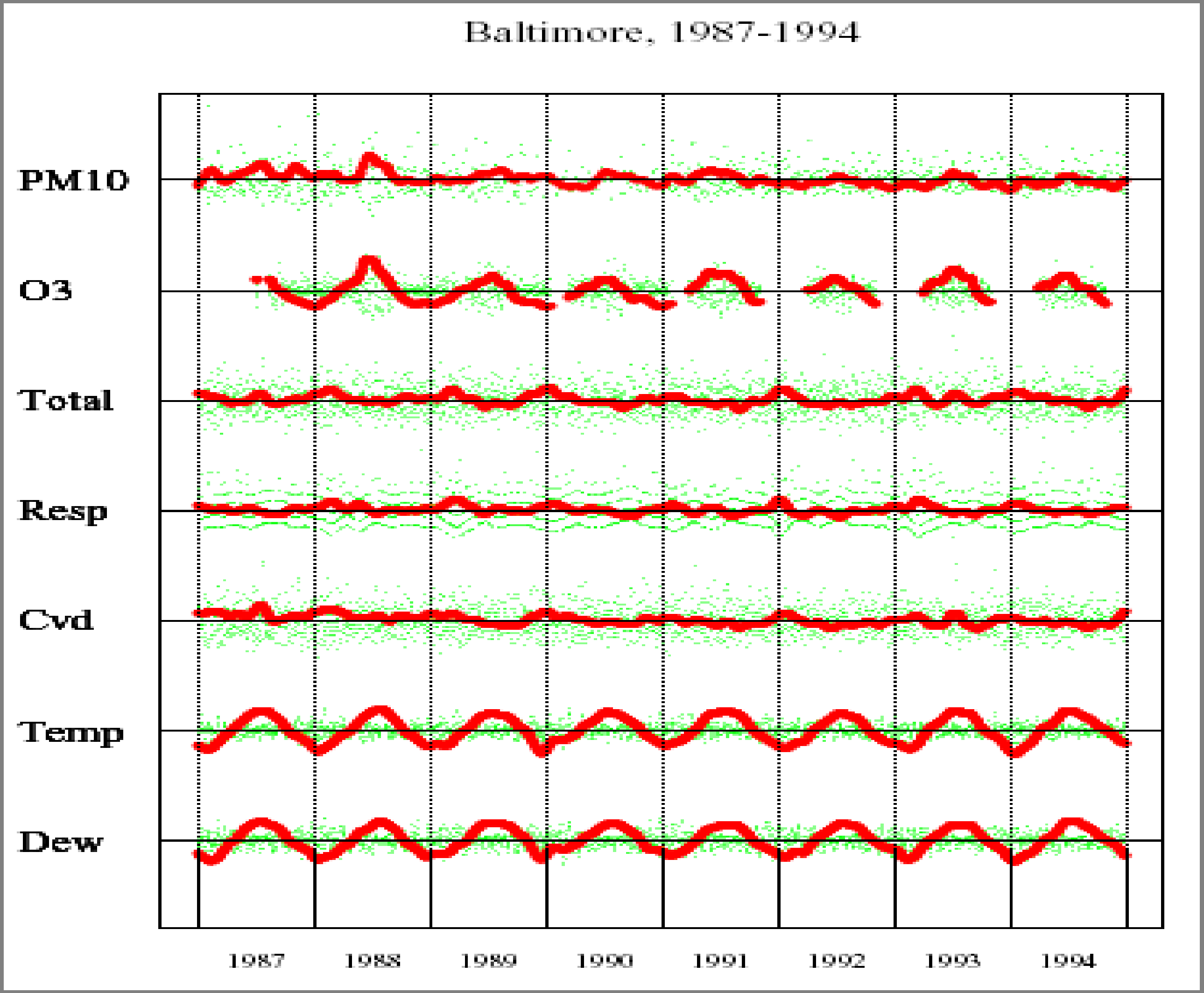
The National Mortality Morbidity Air Pollution Study

The National Morbidity Mortality Air Pollution Study (NMMAPS)

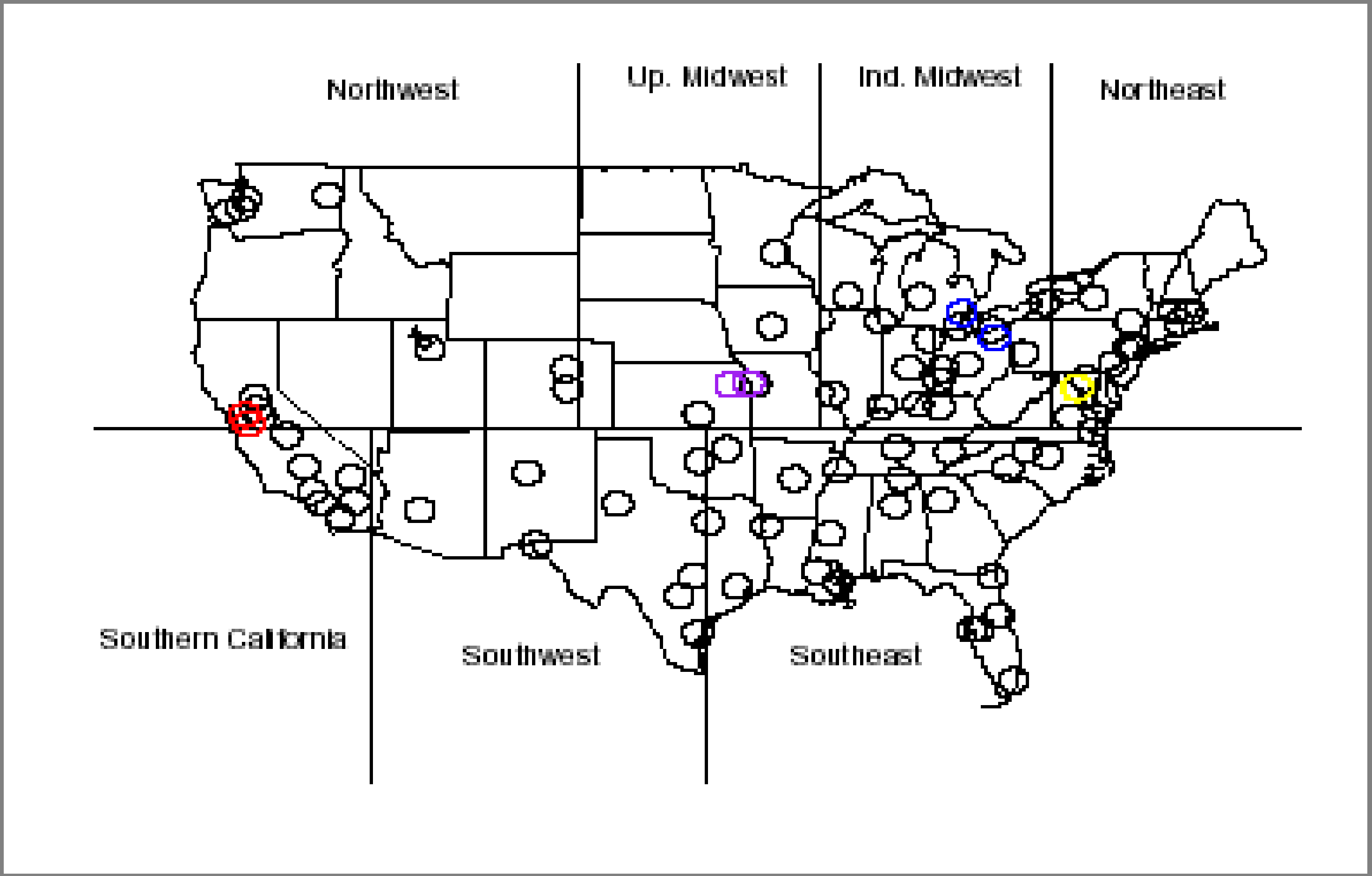
NMMAPS is a multi-site time series study assessing short-term effects of air pollution on mortality/morbidity comprising:

1. a national data base of air pollution and mortality;
2. statistical methods for estimating associations between air pollution and mortality for the 90 largest US cities, and on average for the entire nation.

Daily time series of air pollution, mortality and weather in Baltimore 1987-1994



90 Largest Locations in the USA



Statistical Methods

- **Within city.** Semi-parametric regressions for estimating associations between day-to-day variations in air pollution and mortality controlling for confounding factors
- **Across cities. Hierarchical Models for estimating:**
  - national-average relative rate
  - national-average exposure-response relationship
  - exploring heterogeneity of air pollution effects across the country

Confounding

- The association between air pollution and mortality is potentially confounded by:
  - **Weather:** mortality is higher at low and high temperature
  - **Seasonality:** mortality generally peaks in winter because of influenza epidemics
  - **Long-term trend:** improvement in medical practice lower mortality over time
- **All these phenomena cannot be attributed to air pollution**

Estimating city-specific relative rates

Semi-parametric regression model

- Y_t^c is the mortality count on day t at location c
- PM_{10t}^c is the level of particulate matter on day t at location c
- β^c is the percentage increase in mortality for 10 $\mu g/m^3$ increase in PM_{10}
- $s(\text{temp}, df)$ are smooth functions

$$
\begin{aligned}
\log E[Y_t^c] &= \text{age-specific intercept} + \beta^c PM_{10t}^c + s(\text{time}, 7/\text{year}) + \\
&+ \; s(\text{temp}, 6) + s(\text{dewpoint}, 3) + \text{age} \times s(\text{time}, 8) + \ldots.
\end{aligned}
$$

$$
\log E[Y_t^c] = \beta^c PM_{10t}^c + \text{confounders}
$$

- estimate $\hat{\beta}^c$ and its statistical variance v^c within each location
- (GAM) with smoothing splines or
- (GLM) with natural cubic splines are methods of choice

Estimating a national-average relative rate

Two-stage normal-normal hierarchical model

$$\hat{\beta}^c \sim N(\beta^c, v^c)$$

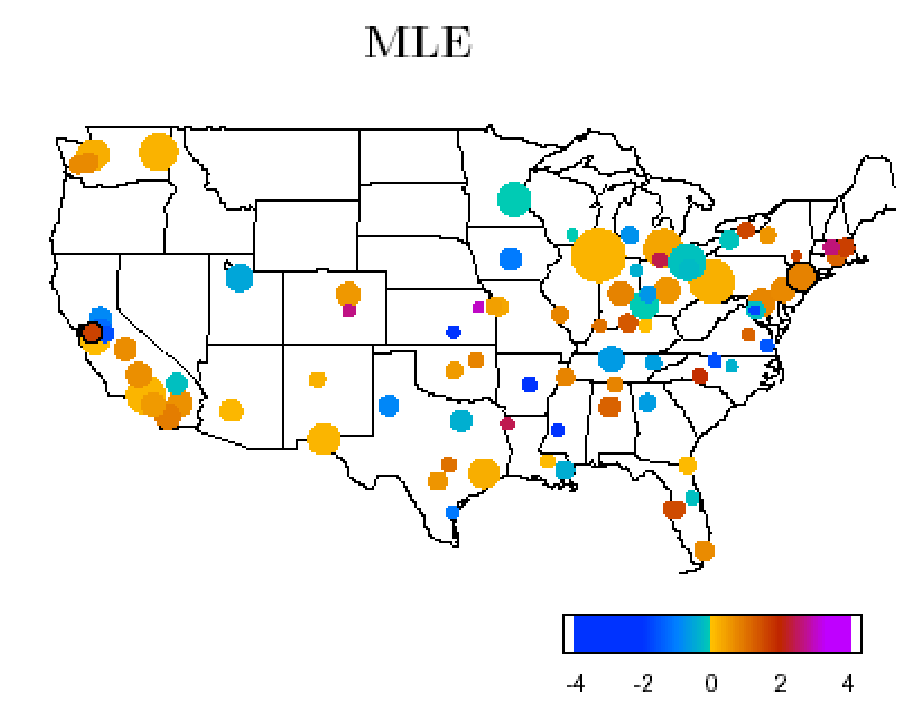
$$\beta^c = \beta^\star + \sum_j \beta_j^\star (x_j^c - \bar{x}^c) + N(0, \tau^2)$$

where

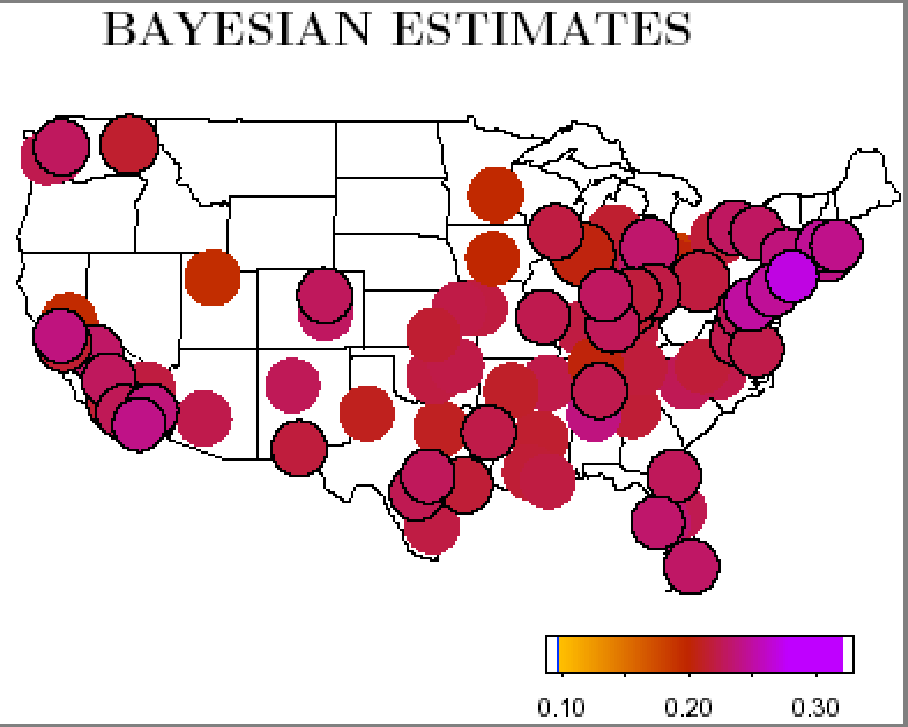
- $\beta^\star$ is the national-average relative rate
- τ^2 is the variance across cities of the true city-specific relative rates β^c (also called heterogeneity)
- $\mathrm{corr}(\beta^c, \beta^{c'}) = \exp\left(-\phi \,\mathrm{distance}(c, c')\right)$
- Goal: marginal posterior distributions of $\beta^\star$ and τ^2

Maximum likelihood and Bayesian estimates of air pollution effects

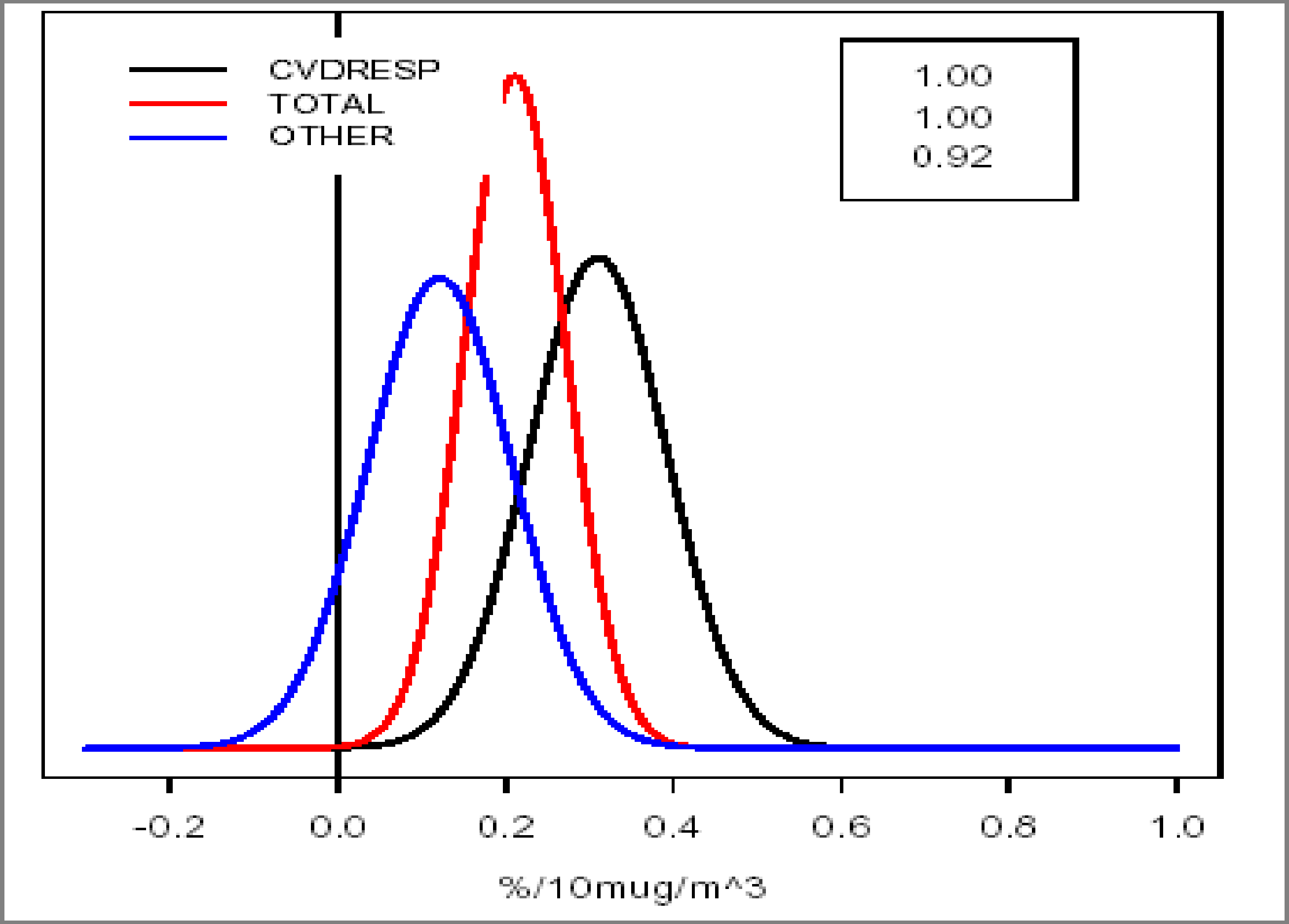
Use only city-specific information



Borrow strength across cities



National-average estimates for Cvdresp, Total and Other causes mortality



Dominici McDermott Zeger Samet 2002 EHP

Software sensitivity versus model uncertainty

- We discovered a sensitivity of the NMMAPS national average estimate to the default parameters in the Splus GAM software
- This captured the attention of the industry and the press
- However, in the environmental epidemiological community not enough attention is generally paid to the sensitivity of time series results to model choice

NMMAPS in Science July 2002

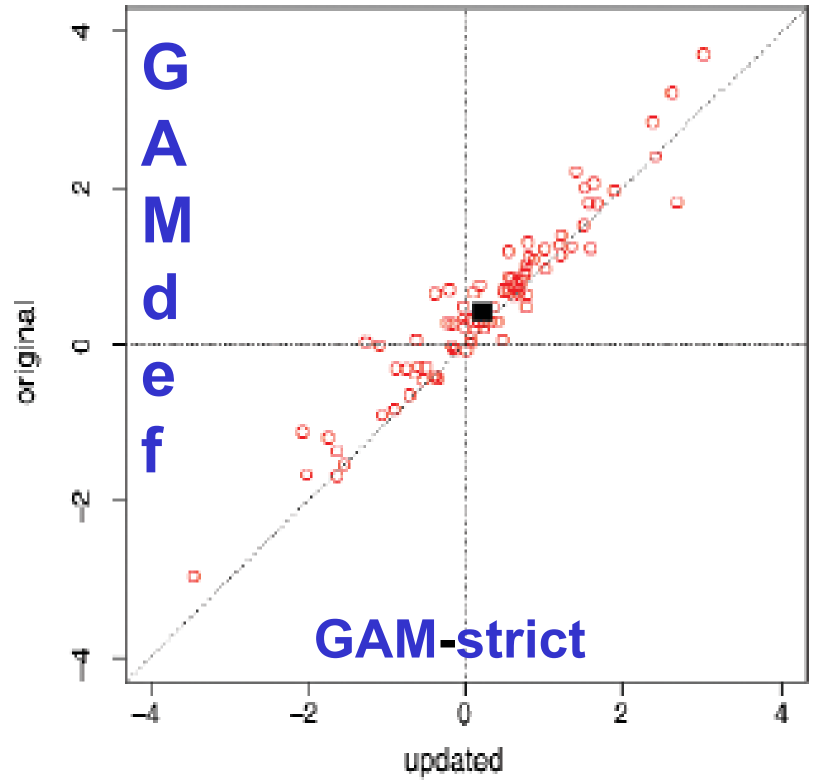
"(A)lthough many questions remain about how fine particles kill people, the HEI study shows there's no mistaking that PM is the culprit…?

Science June 10 2002

AIR POLLUTION RISKS

Software Glitch Threw Off Mortality Estimates

The authors of a landmark air pollution study have found a problem with their software application that means they overestimated the risks of fine particles, or soot. The overall conclusions of the group at Johns Hopkins University in Baltimore linking soot and death haven't changed, but the discovery is providing fresh ammunition to industry groups that have criticized the science behind federal air pollution rules issued 5 years ago. The Environmental Protection Agency (EPA) says it will examine whether the rules need to be modified to reflect the new results.



Recalculating the risk. In this reanalysis of air pollution data, the vertical distance of dots from the diagonal line shows how much the estimated excess death rate was off for each of 90 cities. Black square represents updated (0.27% per 10 µg/m³ of PM_{10}) and original (0.41%) pooled estimates. Diesel exhaust (*right*) is one source of fine particles at center of debate.

The Devil is in the Details!

October 31 2002
Revised GAM software with asymptotically exact standard errors and stricter convergence parameters posted on my web site

GAM consequences to regulatory process

- Revision of the NAAQS have been delayed for one year at least
- Initiated a major effort in a peer-reviewed re-analysis of all major time series studies in the world that used GAM
- Conclusion of the HEI Special Review Panel was: *These revised analyses have renewed the awareness of the uncertainties present in the estimates of short-term air pollution effects…*

Model Uncertainty:

Selecting Degrees of Freedom to Reduce Confounding Bias

Uncertainty in Model Choice

- How to reduce confounding bias in the estimated pollution effect is among the most discussed statistical issues in time series analyses of pollution and health
- In particular, the choice of the number of degrees of freedom in the smooth function of time is critical
- This choice determines the residual temporal variability in the daily deaths and pollution levels used to estimate the pollution effect

Measured and Unmeasured Confounding

- Measured confounders: time-varying covariates that are associated with pollution and mortality, as for example weather variables
- Unmeasured confounders: other time-varying factors, such us influenza epidemics and trends in survival that affect mortality and are temporally associated with variations in air pollution
- Goal: estimate associations between day-to-day variations in air pollution and day-to-day variations in mortality taking into account measured and unmeasured confounders

Statistical Formulation

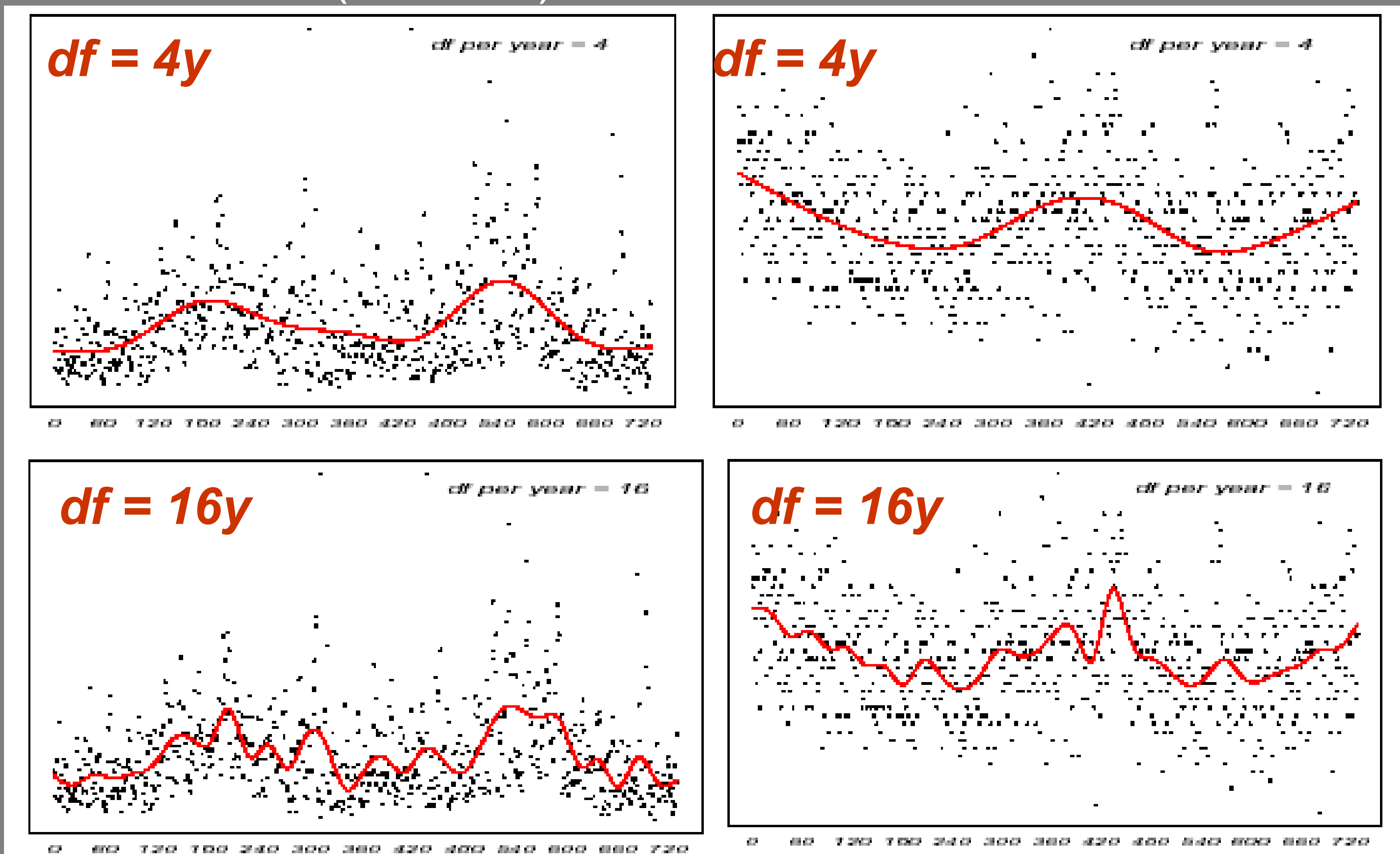
- Y_t is the number of deaths on day t
- $x_{t-\ell}$ is the pollution level on day $t-\ell$
- temp_t, dew_t are the temperature and dew point temperature level on day t.

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \\ \log(\mu_t) &= \beta_{\mathbf{d}} x_t + s(t, \mathbf{d}) + \textstyle\sum_{j=1}^{J} q_j(\text{temp}_t, \text{dew}_t) \\ \text{Var}(Y_t) &= \phi\mu_t \end{aligned}$$

- $q_j(\text{temp}_t, \text{dew}_t)$ are non-linear functions of current and past temperature
- $s(t, \mathbf{d})$ is a smooth function of calendar time
- Goal: determine $q_j(\text{temp}_t, \text{dew}_t)$ and d that sufficiently reduce confounding bias of $\widehat{\beta}$.

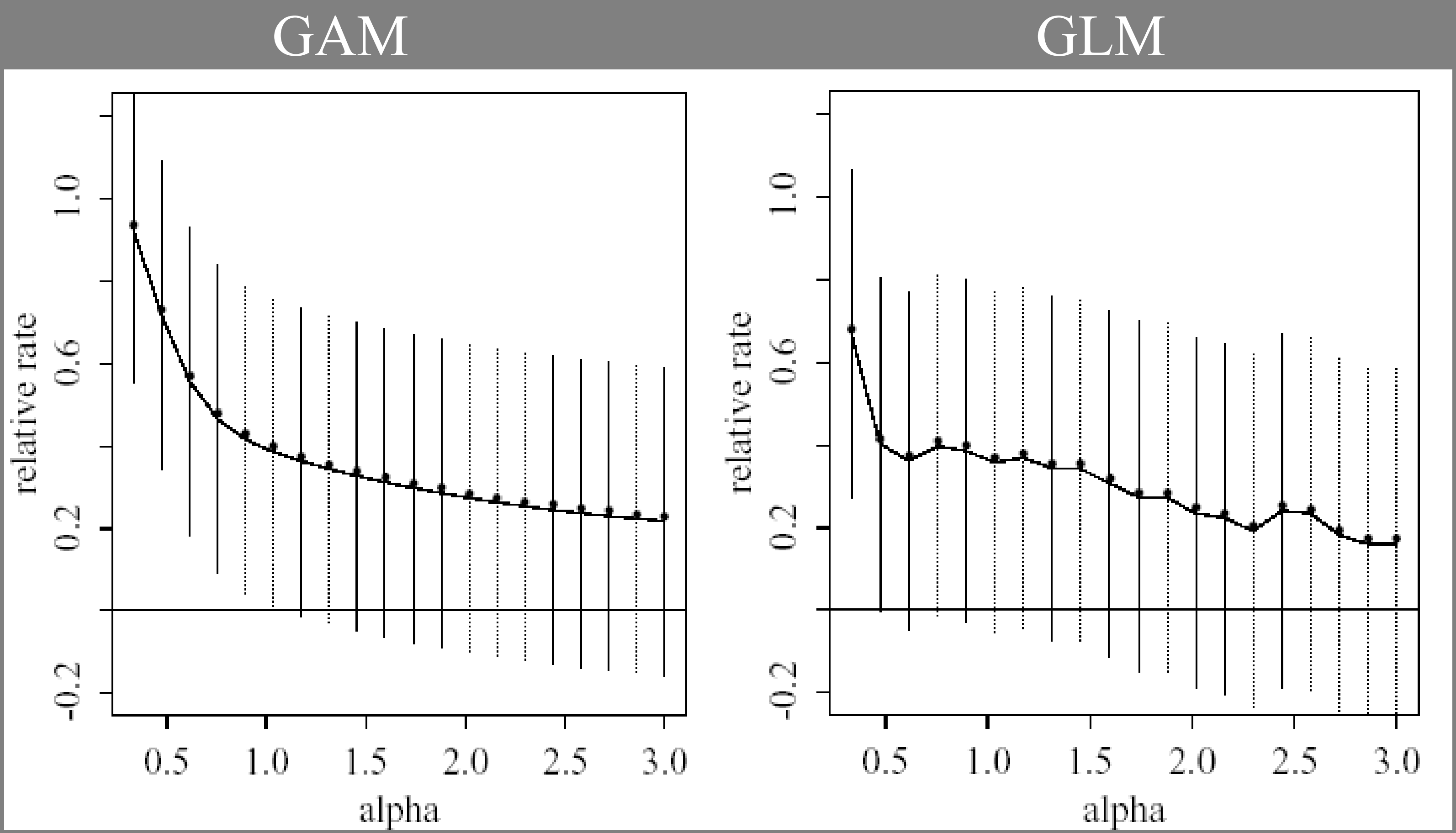
Adjusting for unmeasured confounding

PM10 (1987-88) Mortality (1987-88)



Pittsburgh NMMAPS data

Relative rates estimates as function of the degree of adjustment for confounding factors
(Pittsburgh 1987-1994)



α multiplies the NMMAPS default choice of 8 df per year
So α=2 implies that d = 16 df per year

Choosing the number of df in the smooth function of time

- Choosing too small a *df*
  - Over-smoothing
  - Leaves temporal cycles in the residual in the residual
  - Confounding bias might occur
- Choosing too large a *df*
  - under-smoothing
  - Removes all temporal variability in residuals
  - Wash out the pollution effect

Assessing Model Uncertainty with calibration

- Take a baseline choice for *df* in the smooth functions of time, temperature, dew-point temperature..
- Multiply these *df* for a parameter α
- Fit the NMMAPS hierarchical model for 20 values of α between 1/3 and 3
- Report the NMMAPS national average estimate as function of α

Assessing Sensitivity of the Results to Model Choice

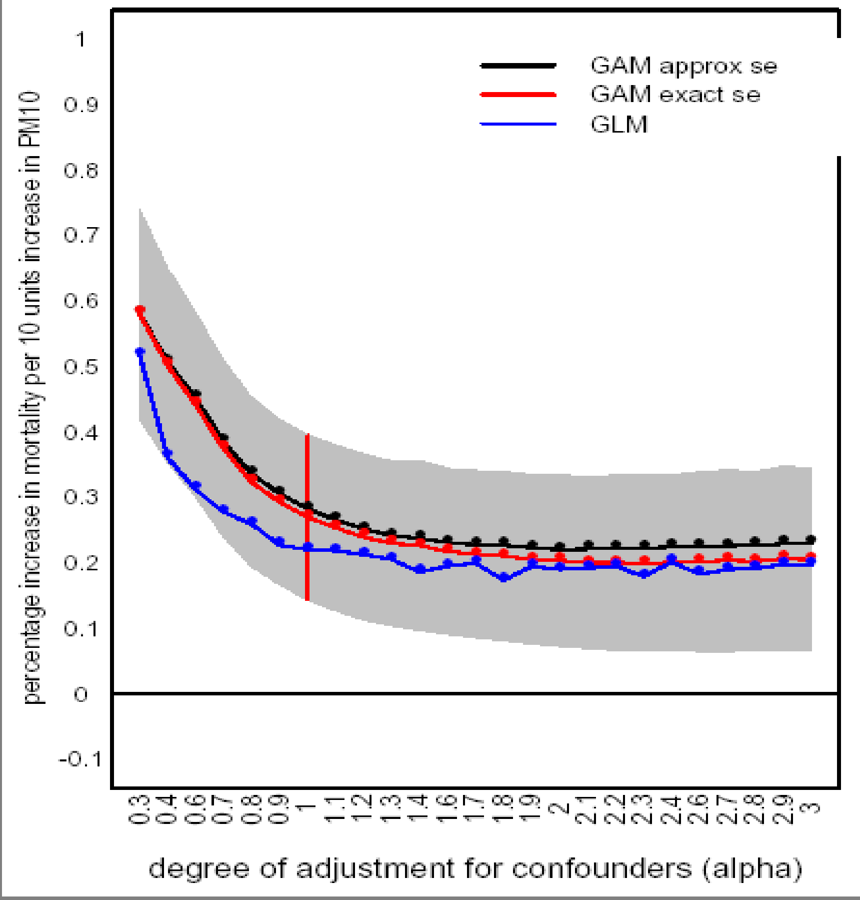
$$\log(\mu_i^c) = \beta^c(\alpha)x_i^c + s(t, \alpha \times 8\ \text{year}) + \sum_{j=1}^J s_j(\mathrm{w}_{tj}, \alpha \times 3)$$

- α are 20 equally spaced points between $1/3$ and 3
- w_{tj} are weather variables
- Goal: estimate $\widehat{\beta}^c(\alpha)$ and $v^c(\alpha)$ within each city c
- develop hierarchical models to combine city-specific estimates across cities as function of α

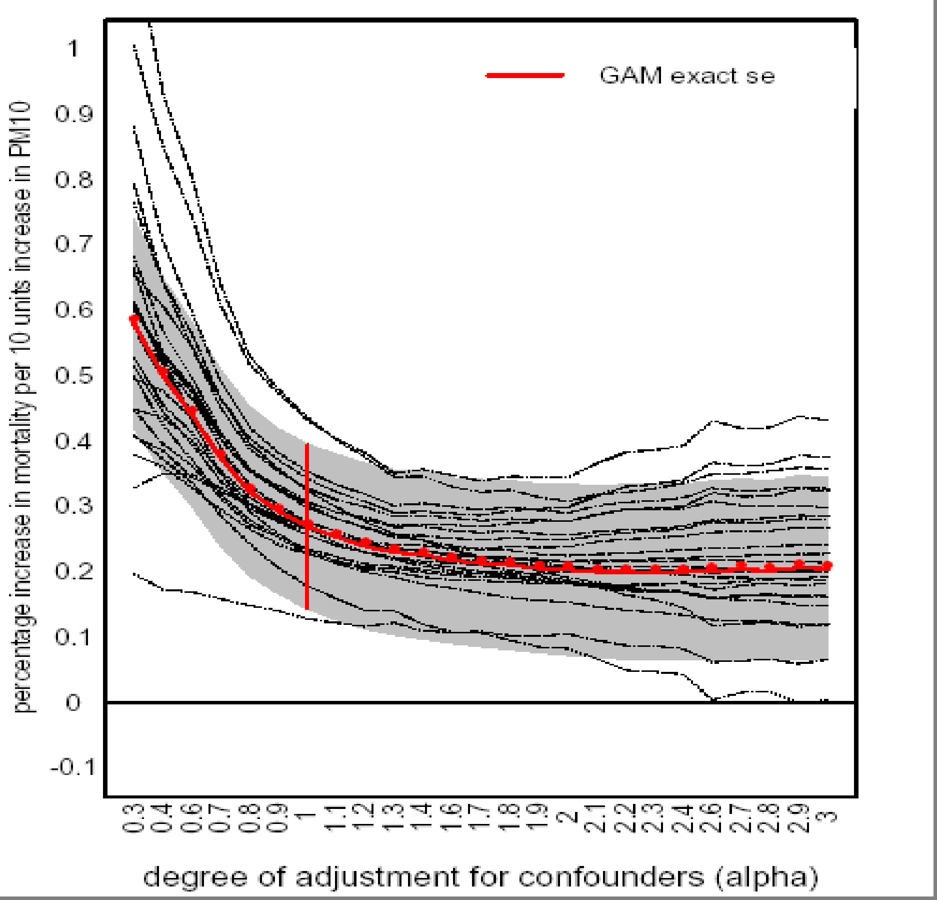
Dominici McDermott Hastie Technical Report (2003)

National average estimates versus degrees of adjustment for confounders

National-average estimates



Bayesian city-specific estimates



Regulatory Policy amidst model uncertainty

- EPA needs a single number for the entire country
- NMMAPS national-average estimate used for mortality impact estimates such:
  - **the excess number of deaths attributable to short-term exposure to air pollution**
- Looking for an important evidence for regulatory assessment and environmental policy
- To meet environmental policy expectations in presence of statistical uncertainty is difficult

Model Uncertainty

- There is no gold standard for estimating an optimal number of degrees of freedom that remove confounding bias
- Because of the complexity and non-verifiable identifiability inherent in selecting an appropriate model to adjust for unmeasured confounders, account for model uncertainty is a necessary step in air pollution risk estimation

Estimating the degrees of freedom to remove confounding bias

$$\log(\mu_i^c) = \beta(\lambda_1, \lambda_2)x_t + s(t, \lambda_1) + s(\mathbf{temp}_t, \lambda_2)$$

- Goal: estimate λ_1 and λ_2 that minimizes the MSE of $\widehat{\beta}$
- Controlling for measured confounders - that is choosing λ_2 - is mainly a model selection problem

  The goal is to identify weather variables and their functional form that best predict mortality
- Controlling for unmeasured confounders - that is choosing λ_1 - is partially identified.

  It is possible identify confounding scenarios where data-driven methods for estimating λ_1 fails, and these confounding scenario cannot be identified by the data (Ritov and Bickel 1990, and Robins and Ritov 1997)

Model Uncertainty and Model Choice

- Our strategy for assessing model uncertainty treats measured and unmeasured confounders differently
- We approach adjustment for measured confounders as model selection
- We approach adjustment for unmeasured confounders as assessment of model uncertainty in the context of prior opinions

Bayesian Model Averaging

- BMA as a natural approach for adjusting for unmeasured confounders:
  - Easy to incorporate prior information about cyclic variations in pollution and mortality where confounding is more likely to occur
  - Small number of competing models
  - Easy to implement
  - Produce a policy-relevant air pollution effects estimate that takes into account model uncertainty and biological knowledge on confounding

BMA on the degrees of freedom to remove confounding bias

$$M_{\mathrm{d}}: \log(\mu_t) = \beta_{\mathrm{d}} x_t + s(t, \mathrm{d}) + \sum_{j=1}^{J} q_j(\mathbf{temp}_t, \mathbf{dew}_t)$$

where

$$p(\beta \mid \mathbf{data}) = \sum_{\mathrm{d}} p(\beta_{\mathrm{d}} \mid M_{\mathrm{d}}) p(M_{\mathrm{d}} \mid \mathbf{data})$$

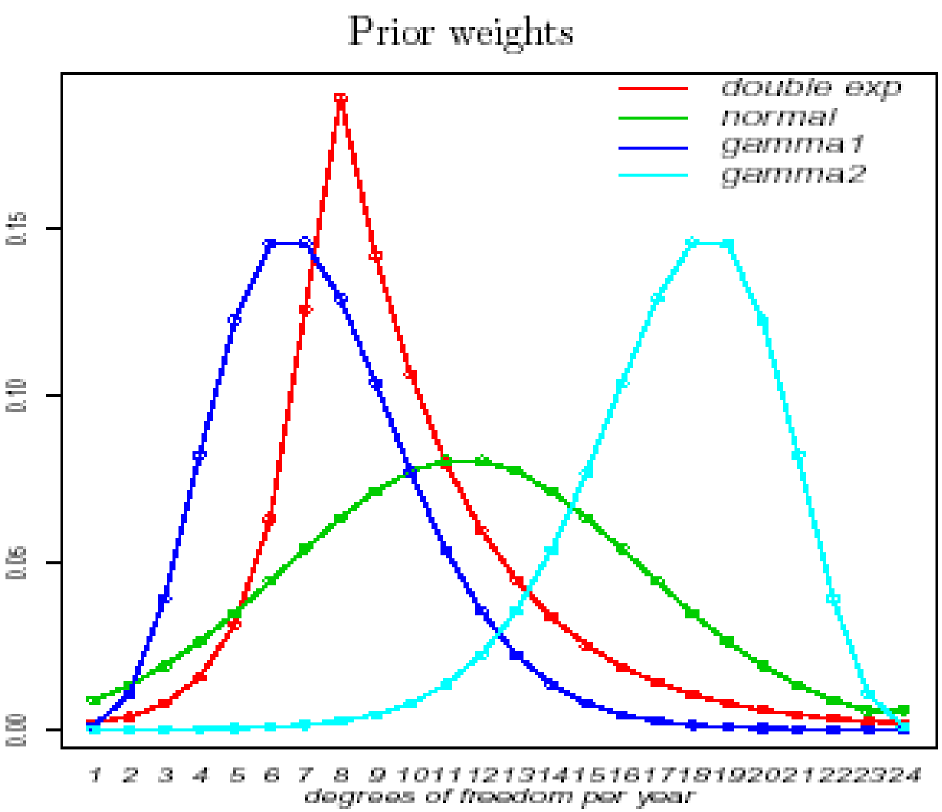
$$p(M_{\mathrm{d}} \mid \mathbf{data}) \propto \exp\left(-\tfrac{1}{2} BIC(M_{\mathrm{d}})\right)$$

Dominici Louis Parmigiani Samet (work in progress)

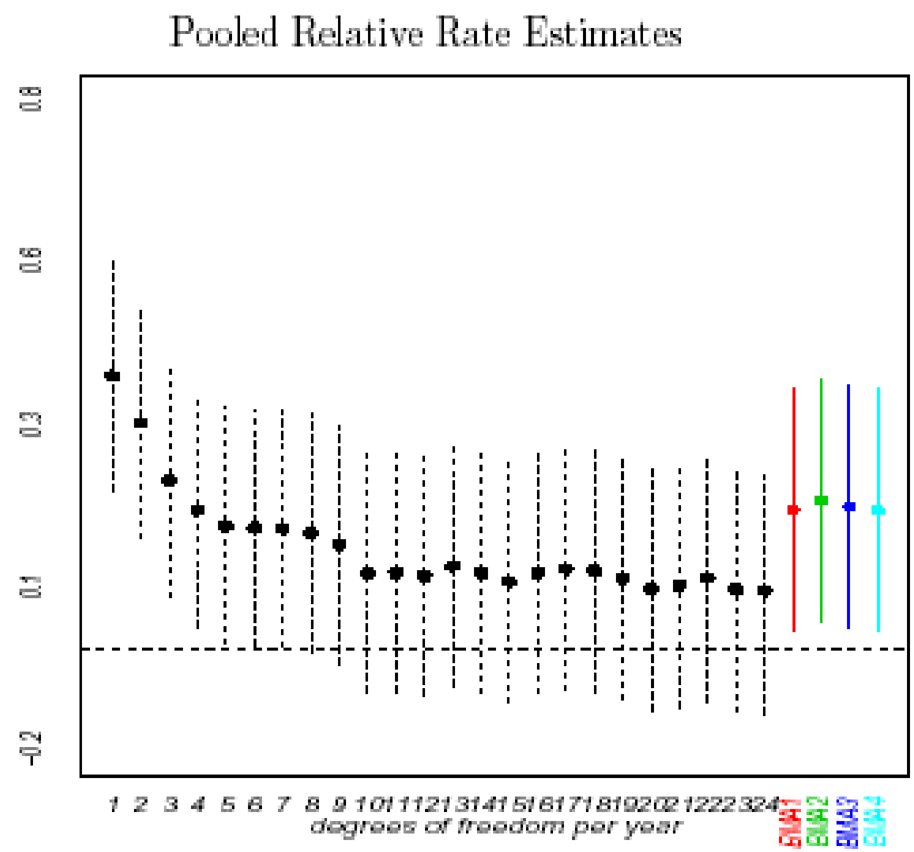
Clyde (1999): Previous work on BMA in time series studies of air pollution

BMA with respect to confounding adjustment

Prior Model Probabilities



BMA Pooled estimates



Is BMA a good idea?

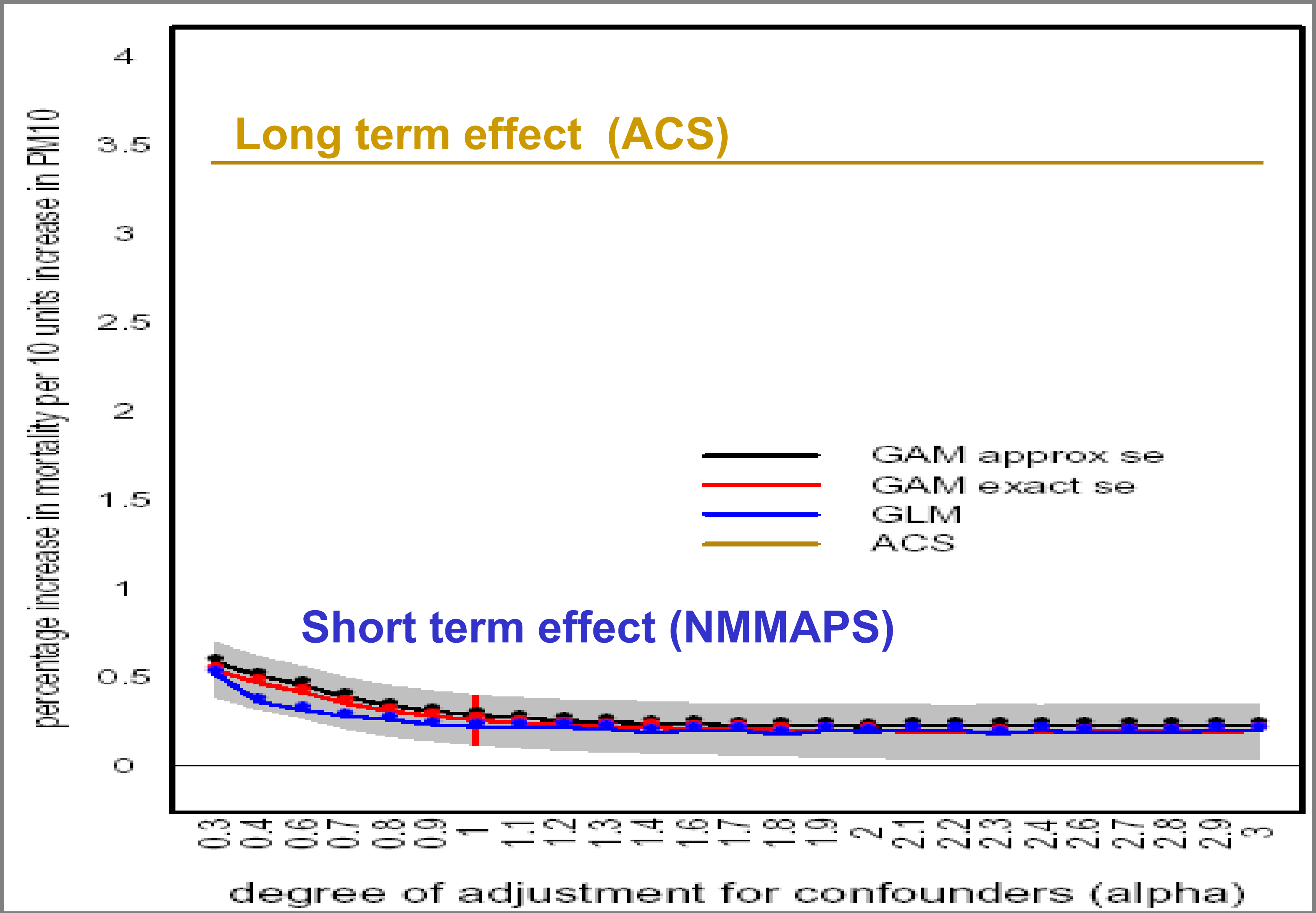
- Accounting for model uncertainty is of central importance when the goal is to estimate a policy relevant quantity (such as a national average pollution effect) which can be sensitive to model choice
- Properly informed policy decisions should depend on the point estimate and on model-selection uncertainties
- BMA as a general approach for removing confounding bias in environmental epidemiology

Epidemiological findings and number of deaths attributable to air pollution

- Time series and cohort studies provide mortality risks estimates from shorter-term and longer-term air pollution exposure
- Currently, regulators rely upon estimates of population burden of illness and premature deaths to set pollution standards

How should we use these two relative risk estimates to determine the attributable number of deaths?

Need to reconcile estimates from time-series and cohort studies



EStimating Chronic Acute Pollution Effects Study (ESCAPES)

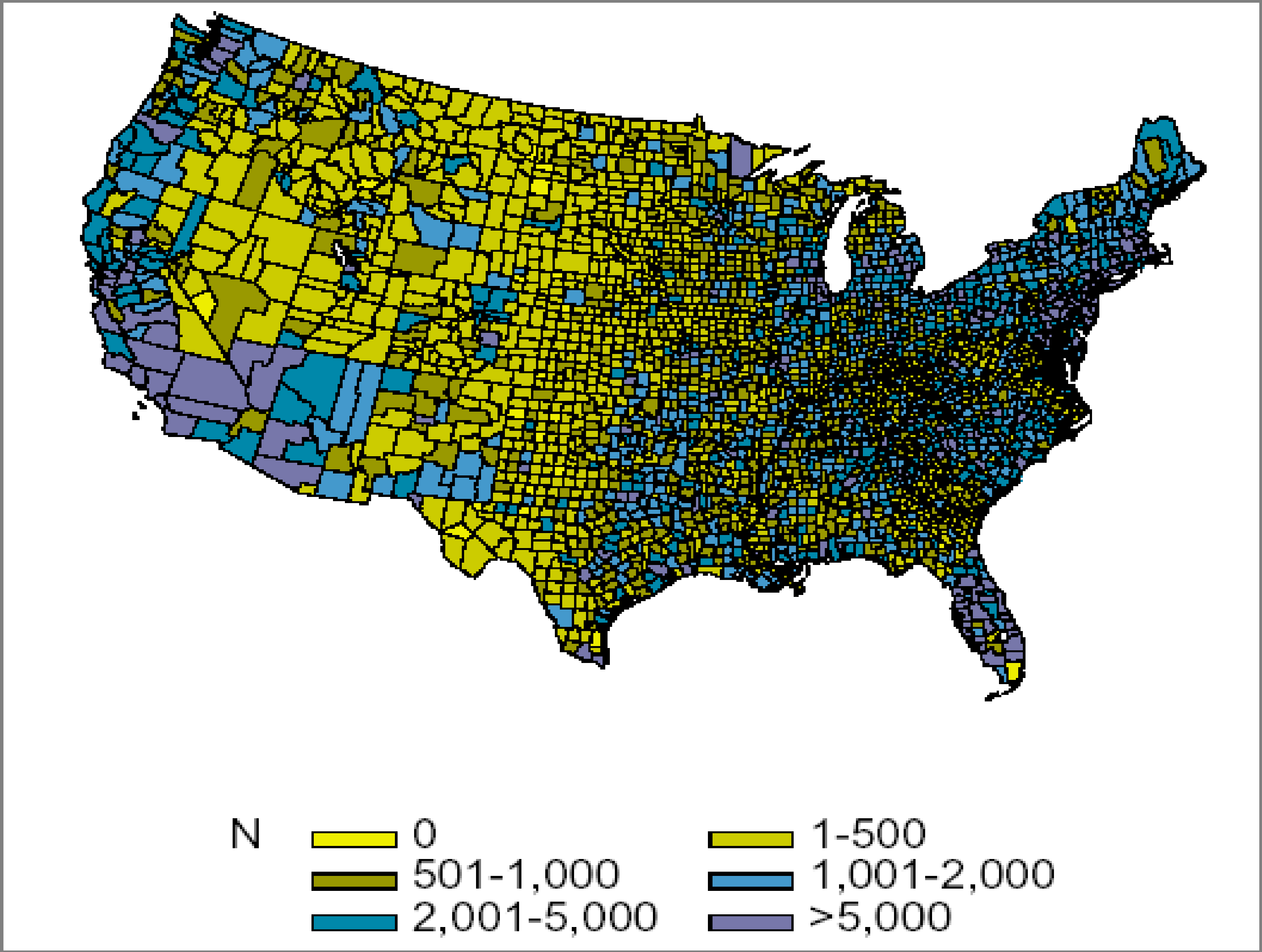
an ongoing project at
Johns Hopkins
Bloomberg School of Public Health

- **ESCAPES** is a joint time-series and cohort study aimed at estimating **health effects** associated with **acute** and **chronic** air pollution exposure for the National Medicare Cohort

ESCAPES: Data Sources

- National Medicare Cohort: 1999-2001 follow up of approx 40 million people for whom we have:
  - Zip code of residence
  - time-of-entry into the cohort
  - time-of-event of hospitalization and deaths
  - Morbidity history
  - Socio-demographic covariates
- National Air Pollution Monitoring Network: 1999-2001 daily time series of:
  - Several pollutants, including PM2.5 for over a thousand monitoring stations
  - Weather variables

Number of People enrolled in Medicare by County in 2000



National Medicare Cohort and National Air Pollution Monitoring Network

These two data sources can be linked by:

1. Counties: 327 already identified
2. Zip codes and air pollution monitoring stations:
   - More than 1,000 monitoring stations are currently available

Summary Statistics: ESCAPES Data: 2000 only

	Number of People enrolled in Medicare	Approximate Number of Deaths
All Counties in USA	41 million	2 million
Counties with PM2.5 data	18 million	800 thousand

ESCAPES (1999-2001)

Exposures:

- daily time series and yearly averages for PM2.5, PM10, other pollutants, and weather

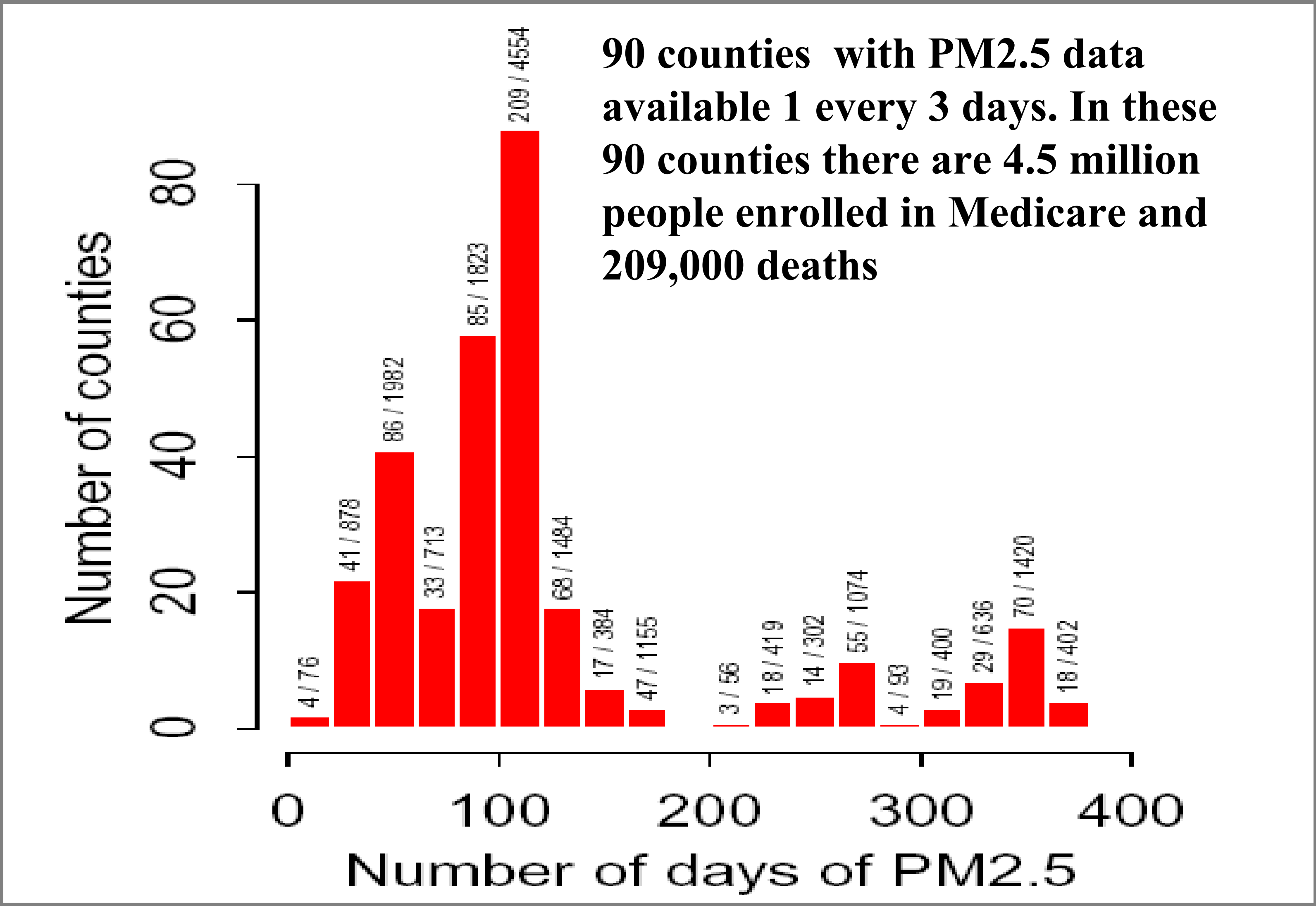
Outcomes:

- diagnosis for any morbidity indicator
- death for any cause-specific mortality

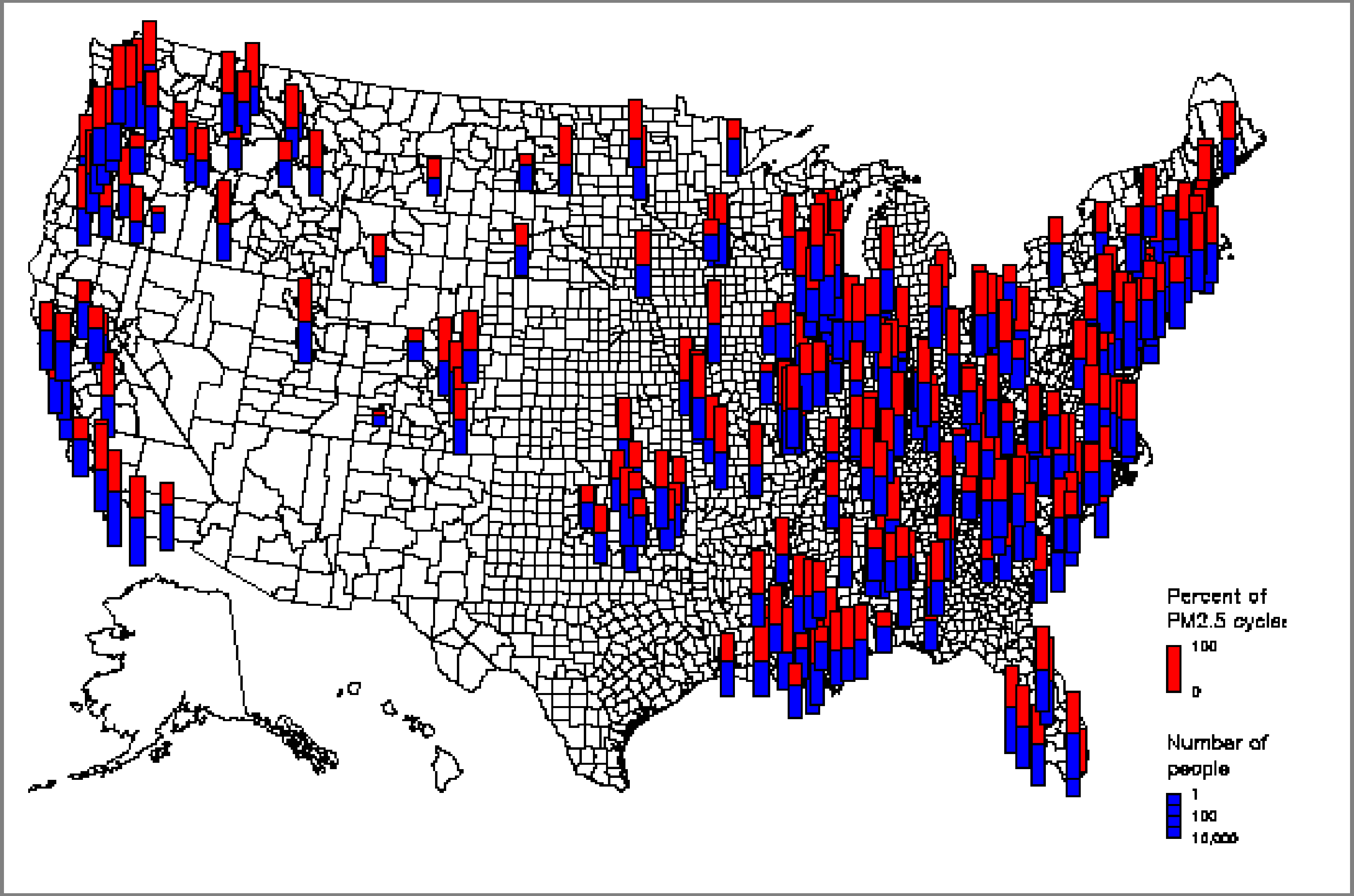
Confounders:

- individual level socio-economic variables
- individual level medical history
- individual level risk factors (smoking included) for a 20% sample of the Medicare cohort
- location-specific characteristics (data from US Census)

Number of counties with PM2.5 and number of deaths in the Medicare population



327 US counties with individual level data and daily air pollution levels for 2000



Model Uncertainty in Environmental Policy

- Air pollution effect estimates are sensitive to confounding adjustment and model choice
- Estimation of policy-relevant quantities should rely upon the integration of
  - **Good-ness of fit measures**
  - **Prior knowledge**
  - **Assessment of model uncertainty**
- NMMAPS has provided a national-average estimate of the short term effect of particulate matter on mortality, a key piece of evidence for regulatory policy
- ESCAPES is a new epidemiological design aimed at jointly estimating health effects associated with short-term and long-term exposure to fine particles and other pollutants

NMMAPS on the web

iHAPSS
Internet-based Health & Air Pollution Surveillance System

Google

Search
Search WWW Search jhsph.edu

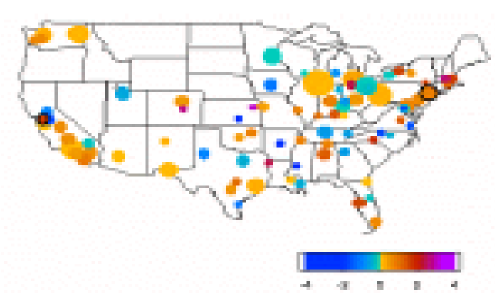
About the iHAPSS

iHAPSS is an internet-based system for monitoring the effects of air pollution on **mortality and morbidity** in US cities.

ABOUT iHAPSS

iHAPSS is funded by the Health Effects Institute (HEI)http://www.healtheffects.org/.
It provides published material, software and data to monitor the association between air pollution and mortality and morbidity.

iHAPSS is currently maintained at the Johns Hopkins Bloomberg School of Public Health Biostatistics Department http://www.jhsph.edu/Dept/Biostats/



Maximum Likelihood estimates of the log relative rates of mortality from exposure to PM_{10} (% increase in mortality per 10 unit increase in PM_{10}).

PUBLICATIONS

Current and past year publications and reports.

SOFTWARE

Tools for Data manipulation and analysis.

DATA

Mortality, air pollution and met data for 90 U.S. cities 1987-1994.

RWEB

Web based interface to R that takes the submitted code, runs R and returns the output

CONTACT

Email the admin to give your suggestion and comments.

Papers, NMMAPS data, and software posted on the web

- http://www.biostat.jhsph.edu/~fdominic
- IHAPPS (Internet-based Health Air Pollution Surveillance System, Dr Aidan McDermott)
- http://www.ihapss.jhsph.edu/
- Environmental Biostatistics Working Group (EBWG)
- http://www.ihapss.jhsph.edu/~Ebwg