

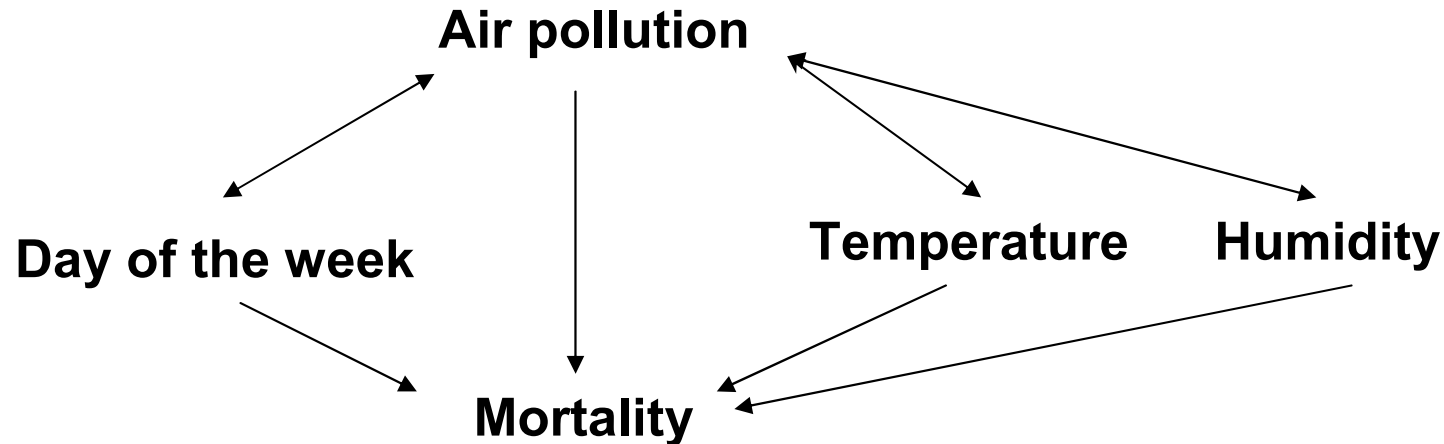
Multi-site Time Series Analysis

Controlling for Confounders

**SAMSI Spatial Epidemiology
Fall 2009**

Howard Chang
hhchang@jhsph.edu

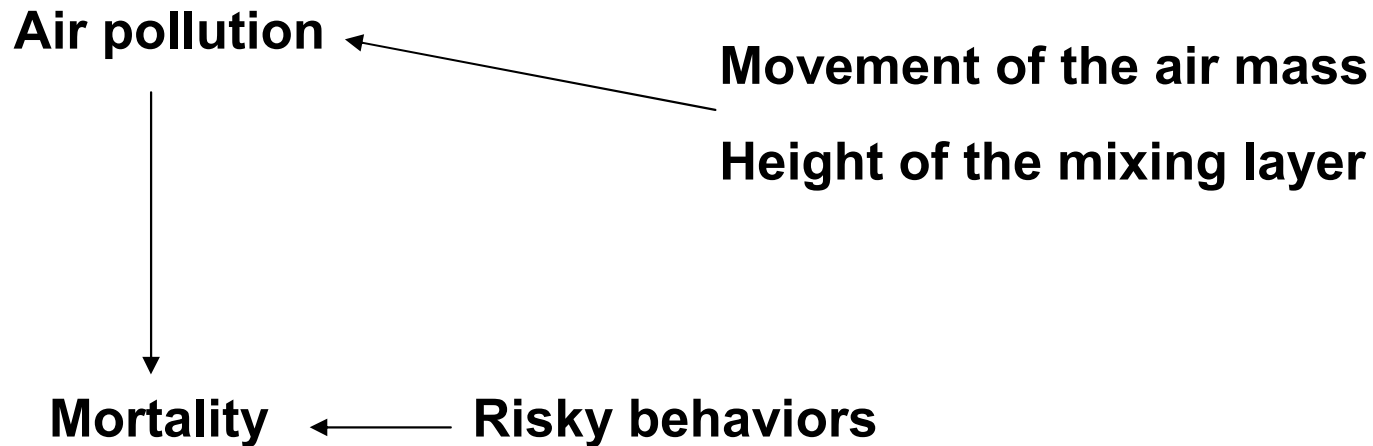
Confounders for Air Pollution and Health



Confounder Check List

- A confounder is a known risk factor of mortality
- A confounder is associated with daily level of air pollution
- A confounder is not in the causal pathway of air pollution and mortality.

Confounders for Air Pollution and Health



There are also unmeasured variables that determine pollution level but are unlikely to correlate with risk factors of mortality. These do not need to be included in the health model.

Unmeasured Confounders

- Seasonal trend: influenza and respiratory infections
- Long-term trend experienced by the population:
 - improvement in medical care
 - trends in the occurrence of major diseases
 - changes in population size

Current Approach:

Use calendar time as a surrogate and include a smooth function of time in the health model

- Under-smooth → residual confounding effects.
- Over-smooth → attenuate true pollution effect

Unmeasured Confounders

Common ways to represent the smooth function of time $f(t)$:

- (1) natural cubic splines
- (2) penalized or smoothing splines

Common ways to determine smoothness (df):

1. Maximize prediction of health outcome time series using AIC or BIC.
2. Maximize prediction of the exposure time series using generalized cross-validation.
3. Minimize autocorrelation in the residuals by minimizing the sum of absolute partial autocorrelation function (PACF).

Simulation Study (Peng 2006)

Daily mortality and PM_{10} simulated for Minneapolis 1987-1994

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$\log \mu_t = \beta_0 + \beta PM_t + f(t) + q(\text{temp}_t)$$

$$PM_t = g(t) + r(\text{temp}_t) + \xi_t \quad \xi_t \sim \mathcal{N}(0, \sigma^2)$$

$f(t)$ and $g(t)$ have natural cubic spline representations:

$$f(t) = \sum_{j=1}^{m_1} a_j B_j(t)$$

$$g(t) = \sum_{j=1}^{m_2} b_j H_j(t)$$

Simulation Study (Peng 2006)

$$f(t) = \sum_{j=1}^{m_1} a_j B_j(t)$$
$$g(t) = \sum_{j=1}^{m_2} b_j H_j(t)$$

Dominici *et al.* (2004 *JASA*) showed that if we model $f(t)$ using enough degrees of freedom to capture $g(t)$:

The effect of PM_{10} is:

- (1) Asymptotically unbiased if $g(t)$ is smoother than $f(t)$ ($m_2 < m_1$)
- (2) Unbiased if $g(t)$ is rougher than $f(t)$ ($m_2 > m_1$)

Simulation Study (Peng 2006)

Results for natural splines

	<i>Moderate concavity</i>		<i>High concavity</i>	
	<i>g(t)</i> <i>smoother</i>	<i>g(t)</i> <i>rougher</i>	<i>g(t)</i> <i>smoother</i>	<i>g(t)</i> <i>rougher</i>
<i>Bias</i> ($\times 1000$)				
AIC	0.012	0.012	0.026	0.119
PACF	0.059	0.305	0.401	1.701
BIC	0.492	0.471	3.302	2.782
GCV-PM ₁₀	0.021	0.002	0.014	0.034
df = m_1	0.013	0.005	0.018	0.024

Simulation Study (Peng 2006)

“Not how you smooth, but how much you smooth”

1. Maximizing the predictive power of calendar time on the pollutant performs best.
2. AIC performs well as it tends to select larger models as sample size increases.
3. BIC performs the worst. (Prediction versus adjustment selection?)
4. More degrees of freedom per year needed for penalized splines.
5. Minimizing PACF may be important too for obtaining correct standard error for the health effects.

Side Note (SE Under-estimation)

The pooled estimate is unaffected if the bias in SE is

1. Additive
2. Multiplicative and the SE is constant across sites

When the SE is non-constant across sites:

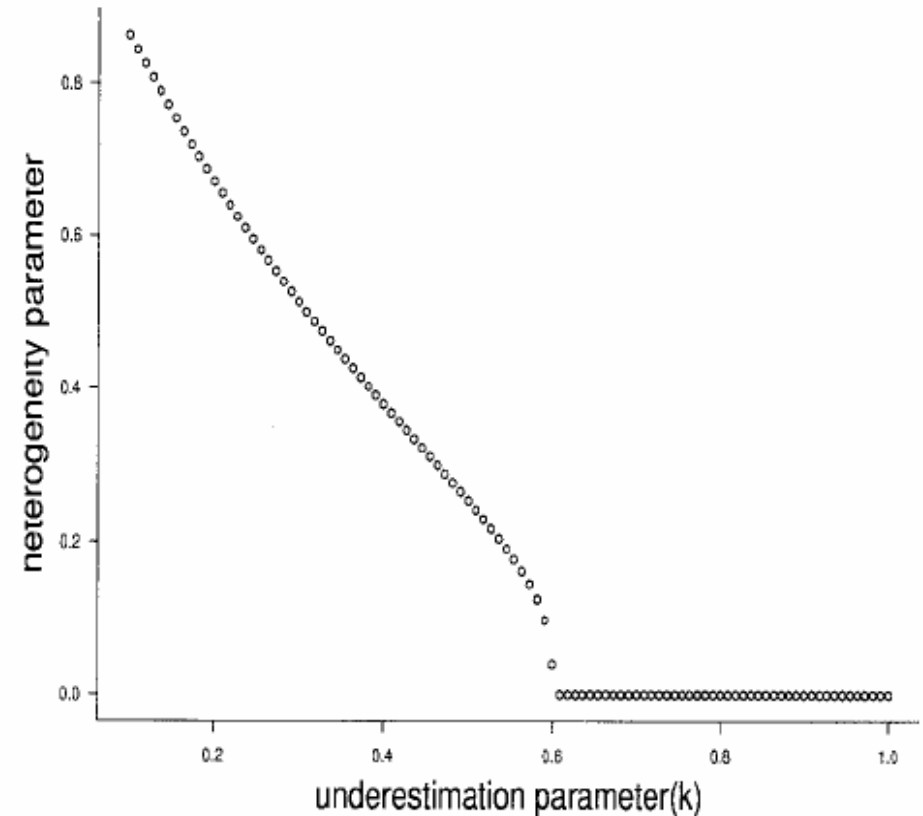
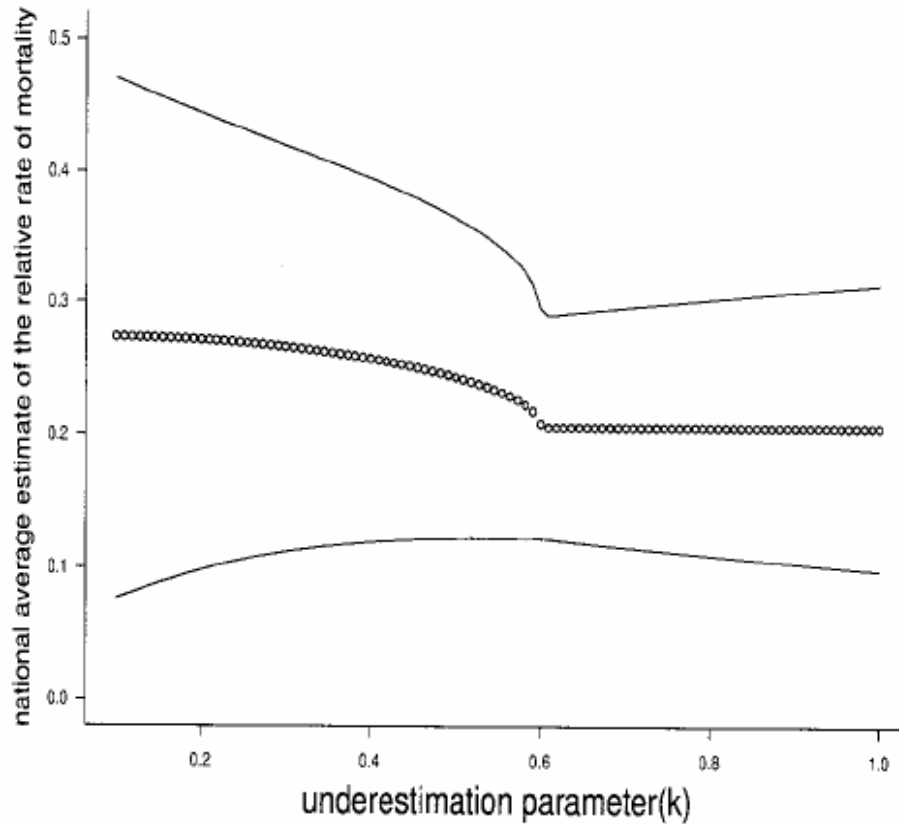
Overall pooled mean is insensitive for up to 40% underestimation in the NMMAPS study

Leads to over-estimation of the heterogeneity parameter.

Consequently, site-specific estimates are non-robust due to “under-shrinkage.”

Side Note (SE Under-estimation)

Simulation Study with NMMAPS



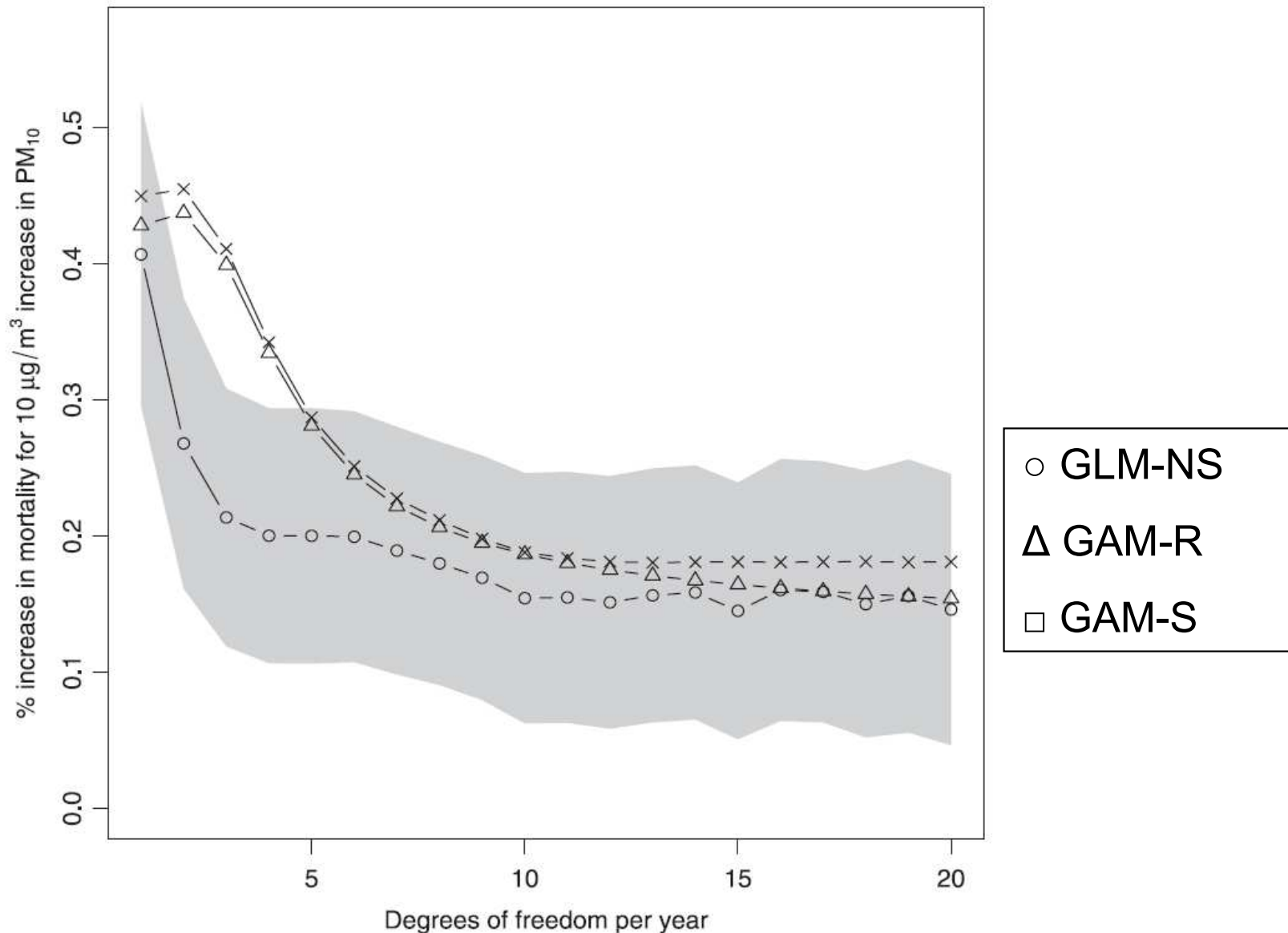
(1-k) = percent under-estimation

NMMAAPS Analysis (Peng 2006)

National pooled relative risk for daily mortality per 10 $\mu\text{g}/\text{m}^3$ increase in same-day PM_{10} level across 90 cities (1987-1994)

<i>Method</i>	<i>AIC</i>	<i>PACF</i>	<i>GCV-PM₁₀</i>
GLM-NS (natural splines)	0.20 (0.11, 0.29)	0.25 (0.14, 0.36)	0.20 (0.10, 0.29)
GAM-R (penalized splines)	0.25 (0.16, 0.34)	0.35 (0.24, 0.46)	0.26 (0.16, 0.35)
GAM-S (smoothing splines)	0.27 (0.18, 0.37)	0.35 (0.24, 0.46)	0.26 (0.16, 0.37)

NMMAPS Analysis (Peng 2006)



Peng RD, Dominici F, Louis TA (2006). "Model choice in time series studies of air pollution and mortality (with discussion)," *Journal of the Royal Statistical Society, Series A*, 169 (2), 179–203.

The Bukowski's Hypothesis

Do pollution time series studies contain uncontrolled or residual confounding by risk factors for acute health events?

Hypothesis: Time series analysis does not consider *acute triggers* on health outcomes. These triggers will confound the health effect estimates if they are associated with (e.g.) cardiovascular mortality and ambient pollution level.

Example: Traffic stress



Response (Goldberg *et al.*)

Recall that time series analysis relies on *temporal contrast*

Key Argument

Under normal society conditions, population-average of individual-level characteristics should vary very smoothly in time.

Does stress co-vary with ambient air pollution daily over the long study period?

Does the % of people suffering from traffic related stress change monthly?

Particularly, does this risk factor influence individuals (i.e. children and the elderly) most susceptible to air pollution?

Model Uncertainty for Confounders

Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air?

Koop and Tole (2004)

Typically only the health effect estimates from a single model are reported after extensive model selection and sensitivity analysis.

Sometimes the health effects can be sensitive to how the confounders (*df*) are included. Can the positive association due to multiple-testing?

One idea is to use *Bayesian model averaging* (BMA) to account for uncertainty in models with different sets of confounders.

Goal: incorporate uncertainty in confounders in the final health effect estimates.

Koop and Tole (2004)

Assuming we have K competing health models with different set of confounders, we wish to calculate the marginal posterior distribution of the health effect:

$$[\theta | data] = \sum_{k=1}^K [\theta | data, M_k] [M_k | data]$$

Koop and Tole Approach:

Carry out variable selection between 7 pollutants at lag 0~3, meteorological variables, and time splines.

Model health outcome using *Normal* regression to implement the MC³ approach of Madigan and Raftery (1996)

Data: Mortality in Toronto 1992-1997

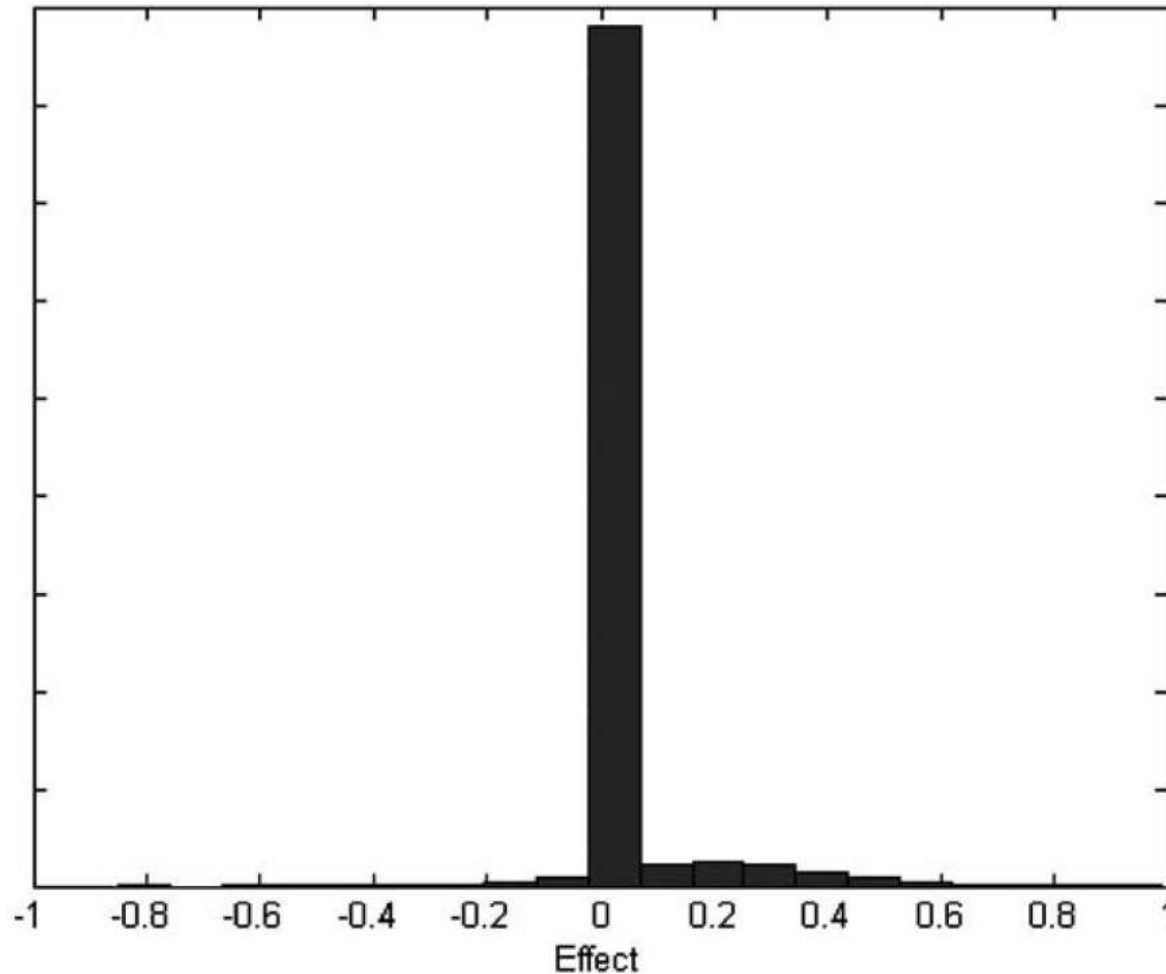
Koop and Tole (2004)

Table 1 **Linear Regression Model Without Time Splines**
Probability of including each explanatory variable

Explanatory variable	Lag	
	0	1
<i>Pollutants</i>		
SO ₂	0.032	0.027
CO	Extremely low 0.047	0.349
NO	0.023	0.042
NO ₂	Inclusion probs! 0.026	0.067
O ₃	0.142	0.026
PM _{10-2.5}	0.021	0.024
PM _{2.5}	0.040	0.070
<i>Meteorological variables</i>		
PRESSURE	0.989	0.497
TEMP	0.089	0.347
HUMIDITY	0.025	0.050
CLOUD	0.023	0.045
WIND	0.095	0.067

Koop and Tole (2004)

Cumulative effect current ~ 3-day lag



Large amount of model uncertainty.

Top 10 models only have total posterior model probability of 12%

Fig. 1. Posterior of cumulative effect of O₃.

Koop and Tole (2004)

Cumulative effect (current ~ 3-day lag) with time splines

	Posterior Mean	Posterior SD
SO ₂	0.013	0.088
CO	0.004	0.045
NO	0.001	0.018
NO ₂	0.027	0.123
O ₃	0.020	0.102
PM _{2.5}	0.017	0.098

Model versus Adjustment Uncertainty?

For estimating health effects of air pollution, we are more interested in building an *explanatory model* for the interpretation of regression coefficients.

An explanatory model is concerned with the inclusion/exclusion of confounders.

Some criticism on the Koop and Tole approach:

- (1) Is model selection that relies on *prediction power* appropriate for epidemiological studies?
- (2) Can we average models when the *interpretation* of the risk coefficient changes?
- (3) Should the exposure of interest and some covariates always be included in the model?
- (4) Should the prior model probabilities be flat if we know certain variables are confounders?

Clyde (2000)

- Model with Poisson regression.
- Devise a family of CIC prior for model selection such that posterior model probabilities are approximated by AIC and BIC.
- Use leaps-and-bounds to first identify candidate models.
- Data: Mortality in Birmingham 1985-1988
- Two-stage estimation approach:
 - (1) Estimate non-parametric time trends first using thin-plate smoothing splines (40 knots). Use the posterior mean based on BMA as a linear predictor in the health model.
 - (2) Carry out model selection again with variable selection between *cumulative* PM_{10} effect (lag 0 ~ 4) and meteorological variables.

Clyde (2000)

SUMMARIES	AIC	BIC
P(Relative Risk = 1 data)	0.03	0.72
Posterior Mean of Relative Risk	1.052	1.015
Posterior Mean of Relative Risk for Models with PM ₁₀	1.054	1.053
Relative Risk for Best Model	1.025	1.00

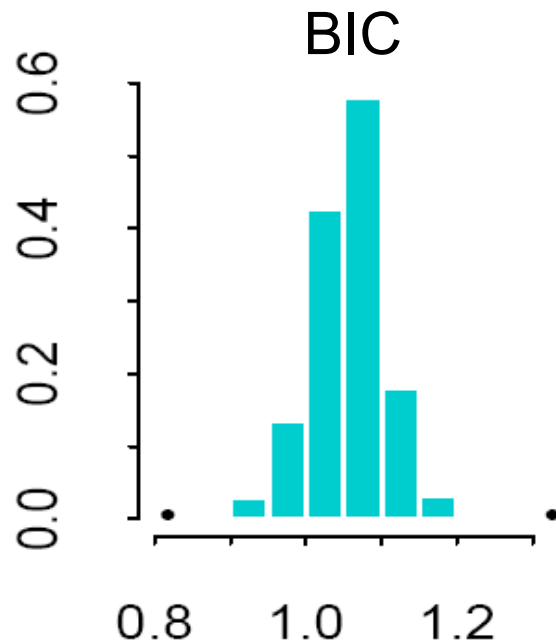
	Predictive MSE	AIC	BIC
Best model →	Model Selection	16.83	16.03
	Bayesian Model Averaging	16.31	15.98

Again we see that:

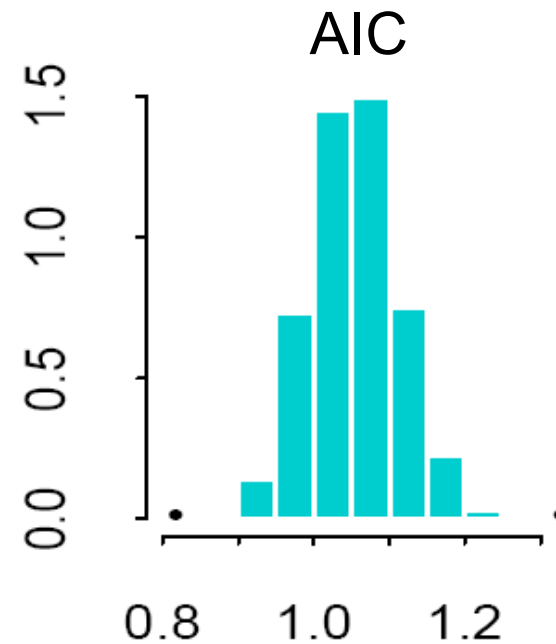
- Using BIC, PM₁₀ is less often included in the model.
- However, there are models where we obtained posterior intervals of relative risk > 1 under both criteria.
- Also, using BIC results in higher predictive power for mortality. 24

Clyde (2000)

Should we only look at models that include the exposure?



Distribution of Relative Risk



Distribution of Relative Risk

Unlike the Koop and Tole's analysis, the results appear quite robust.

Model versus Adjustment Uncertainty?

A standard model selection approach that optimizes prediction tends to choose parsimonious models. Exposure variable can be excluded if it does not contain as much predictive power as other meteorological variables (typically under-estimation).

Similarly, some confounders might not be included in the model and health effect estimates will be biased (typically over-estimation).

While BMA allows us to incorporate model uncertainty in parameter estimation, it is best suited for prediction and the “averaged” individual coefficient can be difficult to interpret.

When controlling for confounders, we like to “over-fit” to ensure our health effect estimates are not spurious due to confounders.

STEADy (Crainiceanu *et al.* 2008)

Consider the following model:

$$Y_i = \beta^\alpha X_i + \sum_{k=1}^K \gamma_k^\alpha Z_{ik} + \sum_{m=1}^M \alpha_m \delta_m^\alpha U_{im} + \epsilon_i^\alpha$$

Exposure **always**
in the model

Confounders
always in the
model

Potential confounders
(included if $\alpha_m = 1$)

Denote: $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_M)$ $\alpha \subseteq \alpha'$ if α is nested in α'

Let β^* be the true health effect and α^* the true model.

Then $\beta^\alpha = \beta^*$ if $\alpha^* \subseteq \alpha$.

We would like to identify an α_0 such that $\alpha^* \subseteq \alpha_0$.

STEADy (Crainiceanu *et al.* 2008)

Note that in a standard BMA approach where no variables are assumed in the model *a priori*, we can partition the model space into two parts:

$$\tilde{\beta}^* = \sum_{\alpha: \alpha^* \subseteq \alpha} E(\beta^\alpha \mid \alpha, D)p(\alpha \mid D) + \sum_{\alpha: \alpha^* \not\subseteq \alpha} E(\beta^\alpha \mid \alpha, D)p(\alpha \mid D)$$

Unbiased estimates of the health effects

Incorrect estimate of the health effects

STEADy (Crainiceanu *et al.* 2008)

Stage I: Fit [X | Z, U]

Identify strong predictors of the exposure X and include them in the outcome model for Y .

1. Divide the model space into $M+1$ subsets with the same number of predictors. Find the model that maximizes the likelihood in each subset.
2. By examining the deviance and the point estimates/confidence intervals of the exposure effects, select a “best” model p . Define $Z' = \{Z, U_p\}$ where U_p contains the predictors of model p .

Stage II: Fit [Y | $Z', U/\{U_p\}$]

Identify *additional* strong predictors of Y in the outcome model using the same algorithm as above.

Among all models with similar exposure effect estimates, identify one with the smallest exposure effect variance.