Statistical integration of disparate information for spatially-resolved PM exposure estimation

Chris Paciorek
Department of Biostatistics

May 4, 2006

www.biostat.harvard.edu/~paciorek

LyX - FoilTeX - pdfLaTeX

Collaborators:

- Yang Liu, Doug Dockery, Francine Laden, Joel Schwartz, Helen Suh (EH-EER)
- Brent Coull (Biostatistics)
- Dave Holland, Ana Rappold (EPA)
- Shobha Kondragunta (NOAA)
- Montse Fuentes (NCSU Statistics)

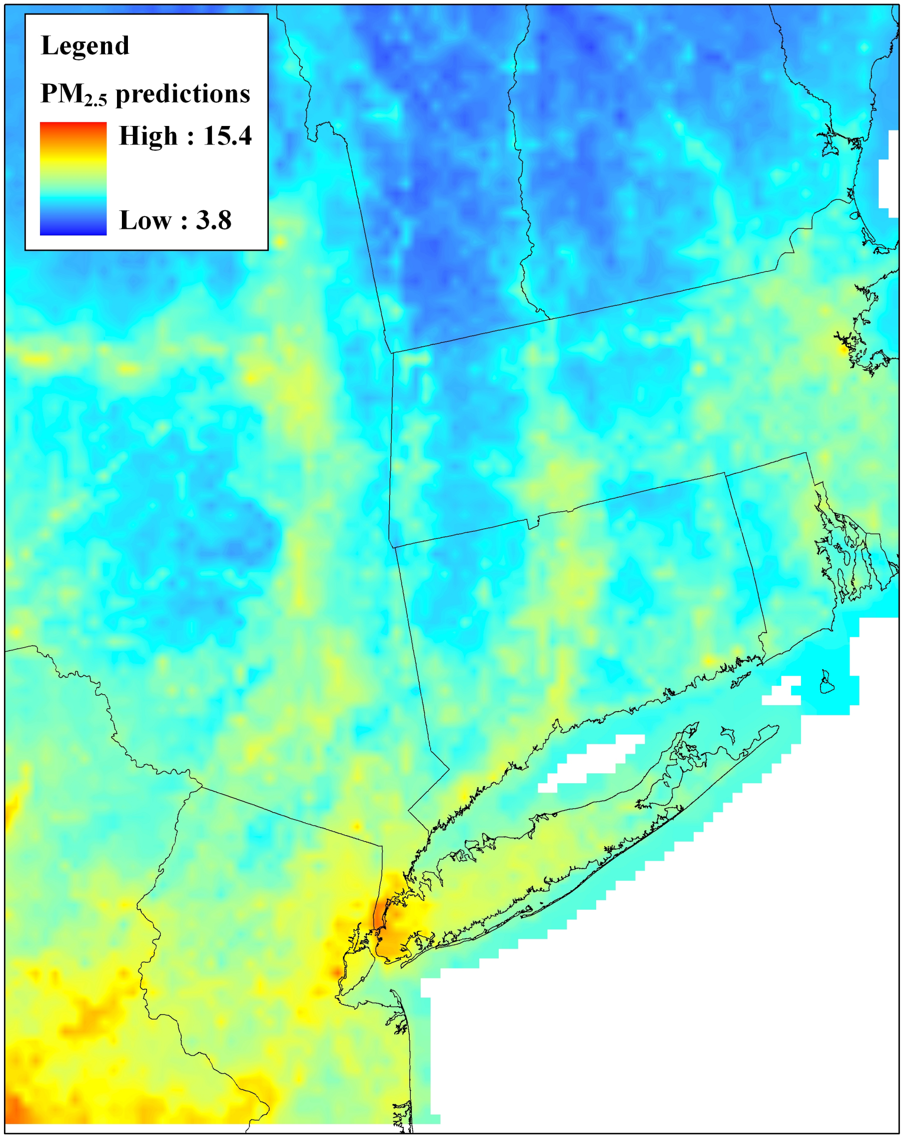
HSPH Health Studies Using Spatial Estimates of Exposure to PM

- NHS: Mortality and cardiovascular outcomes in the NHS cohort (Laden, Schwartz, Suh)
  - nationwide, chronic exposure
- NAS: Cardiovascular biomarkers in the NAS cohort (Schwartz, Suh)
  - eastern MA, acute exposure
- MA-mortality/admissions: Mortality and hospital admissions in Massachusetts based on DPH data (Schwartz, Coull)
  - MA, acute exposure
- MA-birthweights: Birthweights in Massachusetts based on DPH data (Schwartz)
  - MA, chronic exposure

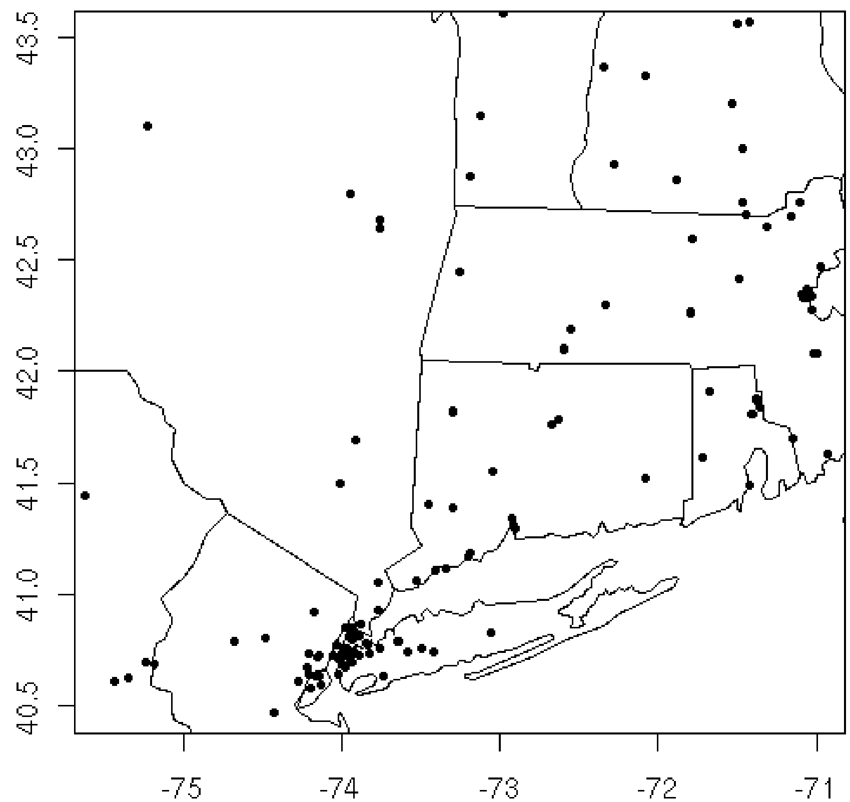
Current exposure estimation efforts and limitations

- NHS: statistical modeling of EPA monitoring data using spatial and regression techniques
  - gaps in spatial coverage
  - few PM2.5 monitors pre-1999
- NAS: central-site estimates
  - no spatial heterogeneity included yet
  - current effort with spatial model using Harvard monitoring data based on a single spatial surface estimate - no space-time interaction
- MA-mortality/admissions: case-crossover analysis based on central site data
  - no spatial heterogeneity included
  - if spatial heterogeneity included, case-crossover requires time-varying spatial estimates
- MA-birthweights: not analyzed
  - need spatially resolved chronic exposure estimates
  - current spatial model only for greater Boston

NHS modeling effort



Estimated PM for one month

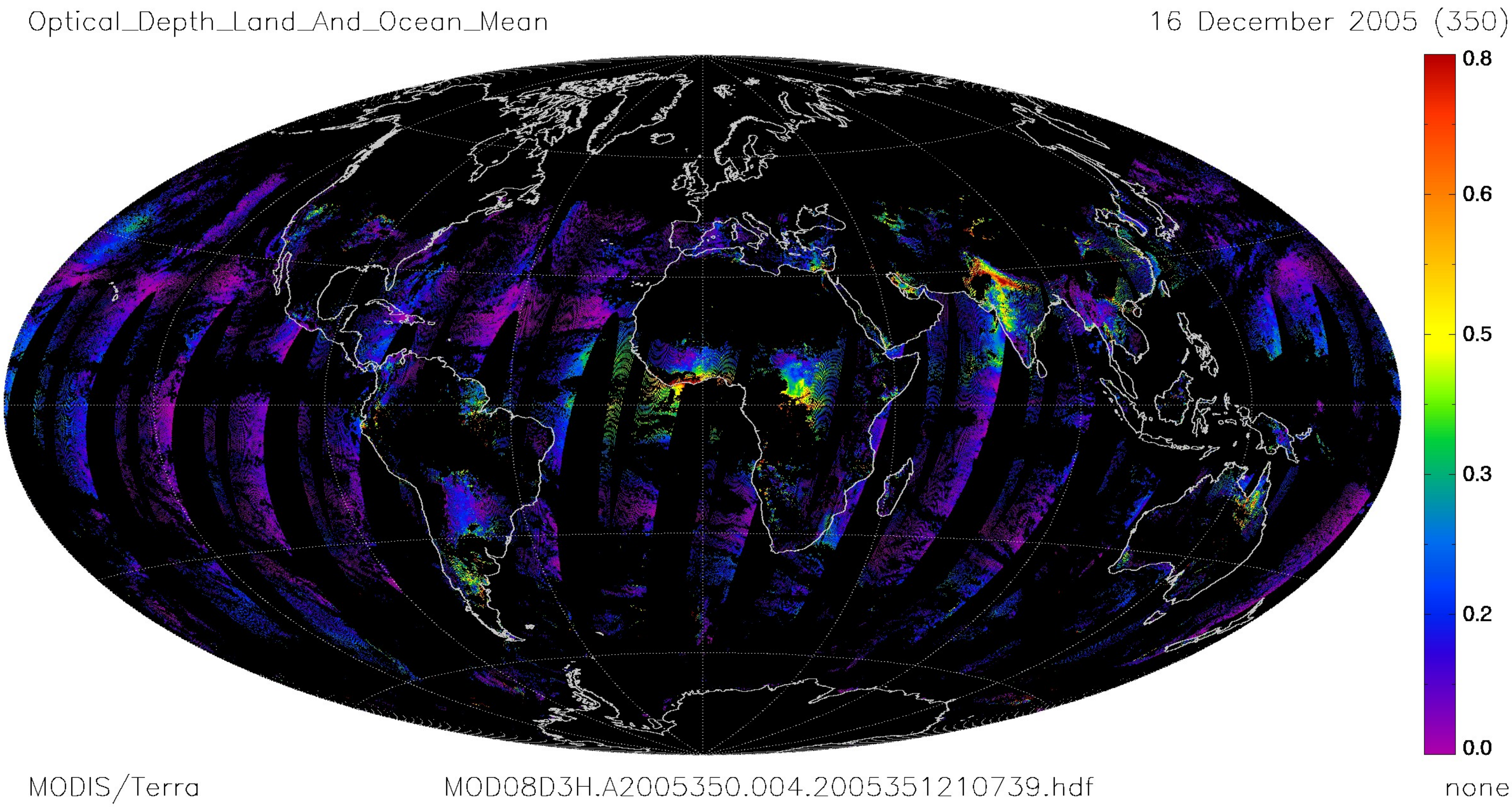


Monitor locations

Satellite and deterministic modeling information

- MODIS and MISR satellite measurements of aerosol optical depth (AOD) (NASA)
  - early 2000-ongoing, every 2-9 days, single measurement
  - 10-20 km pixels
  - missing observations due to cloud cover, surface reflectance
  - AOD measures aerosols (in PM2.5 size range) over entire atmospheric column
- GOES satellite measurements of AOD (NOAA)
  - 1995-ongoing, every 30 minutes
  - 4 km pixels
  - missing observations due to cloud cover, surface reflectance
  - AOD measures aerosols (in PM2.5 size range) over entire atmospheric column
- EPA CMAQ atmospheric chemistry model
  - PM2.5 and a few components: sulfate, nitrate, ammonium, EC, OC (degree of error may vary by component)
  - full 2001 run completed (EPA)
  - other runs for MA may be available, 1988-2002, possibly beyond (NY DEC)
  - 12 km pixels

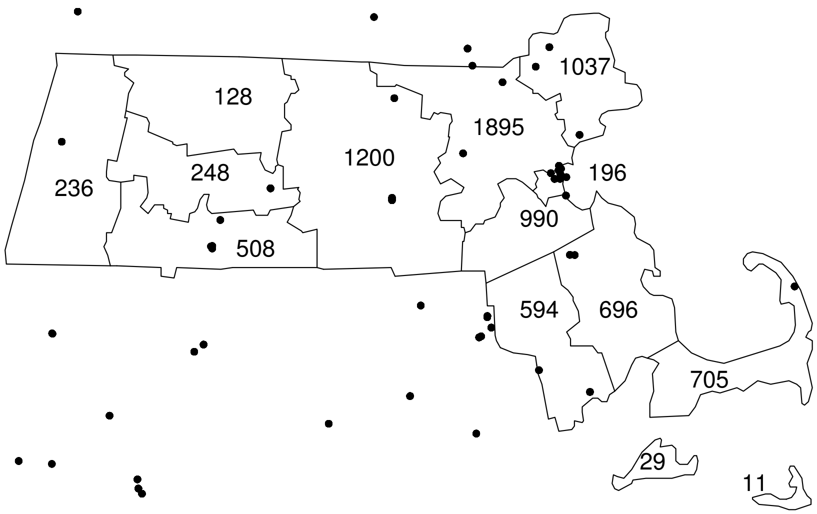
One day of MODIS observations



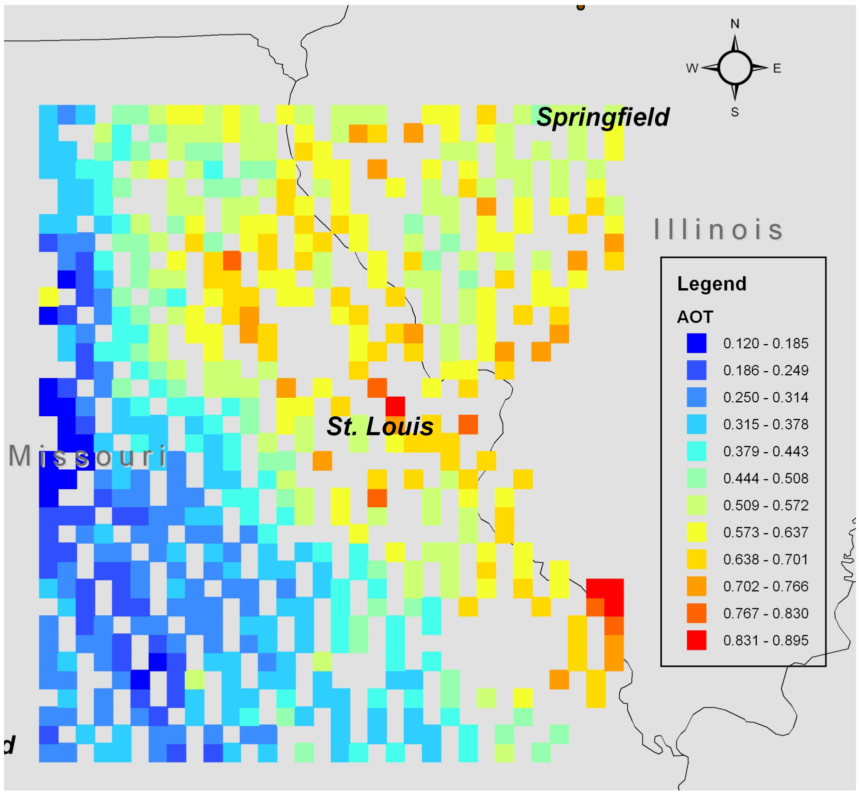
Current exposure estimation efforts and opportunities

- NHS: statistical modeling of EPA monitoring data using spatial and regression techniques
  - gaps in spatial coverage   MODIS/MISR, GOES, 2001 national CMAQ run
  - few PM2.5 monitors pre-1999   GOES
- NAS: central-site estimates
  - no spatial heterogeneity included yet
  - current effort with spatial model using Harvard monitoring data based on a single spatial surface estimate - no space-time interaction   GOES, local CMAQ runs
- MA-mortality: case-crossover analysis based on central site data
  - no spatial heterogeneity included
  - if spatial heterogeneity included, case-crossover requires time-varying spatial estimates   GOES, local CMAQ runs
- MA-birthweights: not analyzed
  - need spatially resolved chronic exposure estimates   MODIS/MISR, GOES, local CMAQ runs
  - current spatial model only for greater Boston

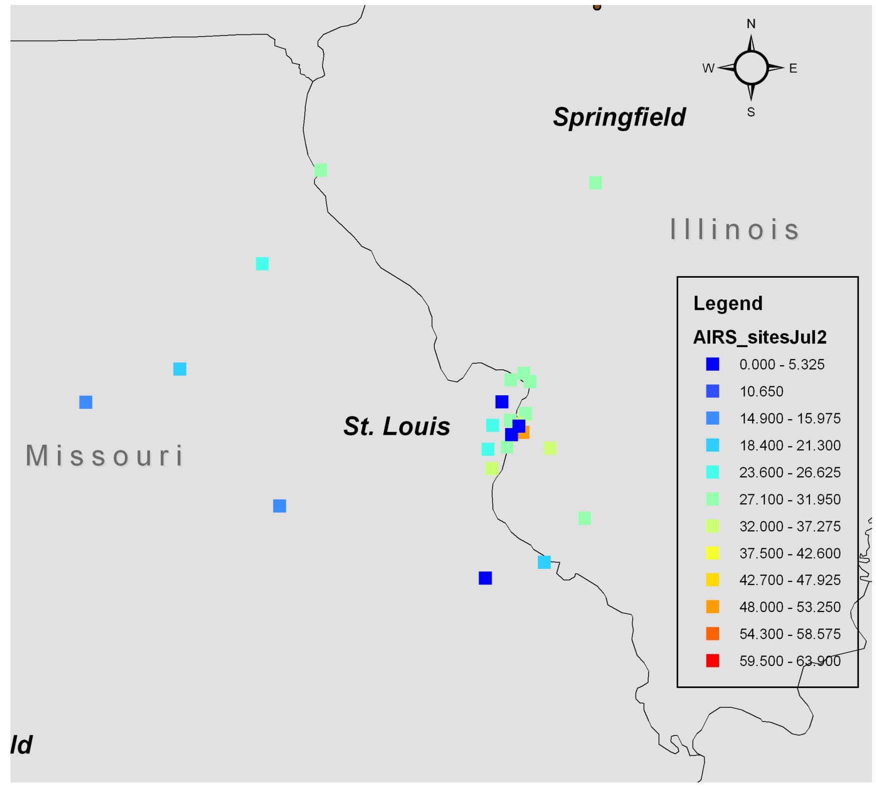
Spatial coverage in Massachusetts (AQS)



Example day of coverage of MODIS AOD



MODIS AOT



$PM_{2.5}$ monitors

courtesy of M. Franklin, Y. Liu, P. Koutrakis

Data integration for regional, chronic exposure estimation

- HEI-funded effort to estimate monthly PM2.5 exposure
- 2000-2006
- eastern U.S. at high-resolution (10 km or less)
- data sources:
  - EPA monitors
  - MODIS/MISR satellite AOD
  - GIS-derived and meteorological covariates: distance to road, population density, wind speed
- goal: produce a database of exposure estimates for use in epidemiological analyses
- future work: use GOES to extend estimates back in time (pre-2000)

Proposed statistical approach

- Fit monthly spatial surfaces of PM2.5: $g_t(s)$
- Monitor observations: $\log Y^g_{i,t} \sim \mathcal{N}(g_t(s_i), \sigma^2)$
- Satellite observations: $\log Y^{rs}_{A,t} \sim \mathcal{N}(a_{A,t} + b_{A,t} \sum_{s \in A} g_t(s), \tau^2)$
  - additive ($a_{s,t}$) and multiplicative ($b_{s,t}$) bias may vary in space and time
  - statistical methods may allow us to estimate the bias in smoothly-varying way
- Local covariate information: represent spatial surface as local and less-local structure
  - $g_t(s) = \sum_k f_k(x_k(s)) + h_t(s)$
- Constrain $h_t(s)$ to vary smoothly in space
  - ensure smooth surfaces and allow for prediction where no observations are located based on local averaging
  - one possible approach is a computationally-efficient Fourier basis representation of a Gaussian spatial process (Paciorek and Ryan, submitted; Paciorek in prep.)
- Fit a Bayesian statistical model and make predictions of PM2.5 ($g_t(s)$) at new locations, s (Fuentes and Raftery, 2005)

Strengths of statistical integration

- estimation of PM surface based on all information
  - ground data: gold standard + higher resolution in urban area
  - remote sensing: broad spatial coverage but coarse resolution
  - other information can be included:
    e.g., GIS information, possible cloud cover biases, vertical profile information from atmospheric chemistry models (Liu et al. 2004)
  - synthesis of differing resolutions of the data sources
- model structure allows for internal validation/calibration of remote sensing data
- model provides estimates of uncertainty in estimated PM at every location

Pilot study

- focus on 2001 and use GOES and CMAQ
- specific aims:
  - benefits of using GOES and CMAQ for estimation pre-2000
  - benefits of using CMAQ to calibrate total column aerosol
  - benefits of higher-resolution satellite data for post-1999

Data Integration for Local, Acute Estimation

- no funding yet but internal EPA funding proposal underway and much of the health data already in house (Schwartz, Suh) — suggestions for funding?
- high spatial resolution desirable
- daily estimates needed
- time-frame: mortality 1998-2002, birthweight: 1995-2002, NAS 2000-2003; more recent data may be obtained/geocoded
- GOES and CMAQ potentially available for 1995-2005
- birthweight requires chronic estimates: potentially just average over daily estimates or fit a simpler model for monthly average exposure

Proposed statistical approach

- Fit daily spatial surfaces of PM2.5: $g_t(s)$
- Monitor observations: $\log Y^g_{i,t} \sim \mathcal{N}(g_t(s_i), \sigma^2)$
- Satellite observations: $\log Y^{rs}_{A,t} \sim \mathcal{N}(a_{s,t} + b_{s,t} \sum_{s \in A} g_t(s), \tau^2)$
  - additive ($a_{s,t}$) and multiplicative ($b_{s,t}$) bias may vary in space and time
  - statistical approaches may allow us to estimate the bias in smoothly-varying way
- Local covariate information: represent spatial surface as local and less-local structure
  - $g_t(s) = f(x(s)) + h_t(s)$ (approach as taken in NHS analysis)
- Constrain $h_t(s)$ to vary smoothly in space and time
  - ensure smooth surfaces and allow for prediction where no observations are located based on local averaging
  - missing monitor and satellite data require borrowing strength across days: $h_t(s) = \phi h_{t-1}(s) + \epsilon_t$
  - potentially very computationally demanding
- Fit a Bayesian statistical model and make predictions of PM2.5 ($g_t(s)$) at new locations, s

Challenges for local estimation

- obtaining GOES observations: NOAA hasn't processed most years and validation is needed first
- obtaining high-quality CMAQ output for sufficient years
  - CMAQ is computationally demanding
- very high resolution available only through regression on covariates
- speciation?
  - available only at limited monitors
  - CMAQ provides limited components: sulfate, nitrate, EC, OC
  - how to get best estimates of spatial surfaces of components?
    - ∗ estimate total PM surface and decompose into components based on regression relationships?
    - ∗ combine CMAQ and monitors for limited components and coarse spatial resolution?

Additional thoughts...

- Opportunities
  - potential usefulness of satellites for exposure estimation in international context where monitoring is limited
  - satellite data for other pollutants?
    * NO2 available but at low resolution (GOME satellite, 250 km); OMI at 13 km since 2005
    * ozone measurements are taken but don't capture surface ozone well
    * BC at 13 km (OMI since 2005) or BC at 40 km (TOMS)
    * overlooked possibilities?
  - CMAQ output on other pollutants?
  - need for partnerships with atmospheric chemistry modeling groups?

- Challenges
  - is PM2.5 sufficiently heterogeneous spatially to make the proposed efforts worthwhile?
  - does noise in satellite and CMAQ output limit usefulness at scales of epidemiological interest?
  - given spatially-resolved exposure estimates, how deal with health effects confounded by unmeasured spatially-varying confounders
  - health analyses (particularly survival analysis and logistic regression) that account for measurement error (Berkson-type structure: Gryparis, Paciorek and Coull (in prep.))
  - speciated components