Predicting Long-term Exposures for Health Effect Studies

Lianne Sheppard

Adam A. Szpiro, Johan Lindström, Paul D. Sampson

and the MESA Air team

University of Washington

CMAS Special Session, October 13, 2010

Introduction

- Most epidemiological studies assess associations between air pollutants and a disease outcome by estimating a health effect (e.g. regression parameter such as a relative risk):
  - A complete set of pertinent exposure measurements is typically not available

  ➔Need to use an approach to assign (e.g. predict) exposure
- It is important to account for the quality of the exposure estimates in the health analysis

  ➔*Exposure assessment for epidemiology should be evaluated in the context of the health effect estimation goal*
- Focus of this talk: Exposure prediction for cohort studies

Outline

- Example: MESA Air
- Predicting ambient concentrations
  - Spatial and spatio-temporal statistical models
  - Incorporating air quality model output
- Evaluating predictions
  - Focus on temporal/spatial scale needed for health analyses
- Lessons learned from one year of CMAQ predictions
- Summary and conclusions

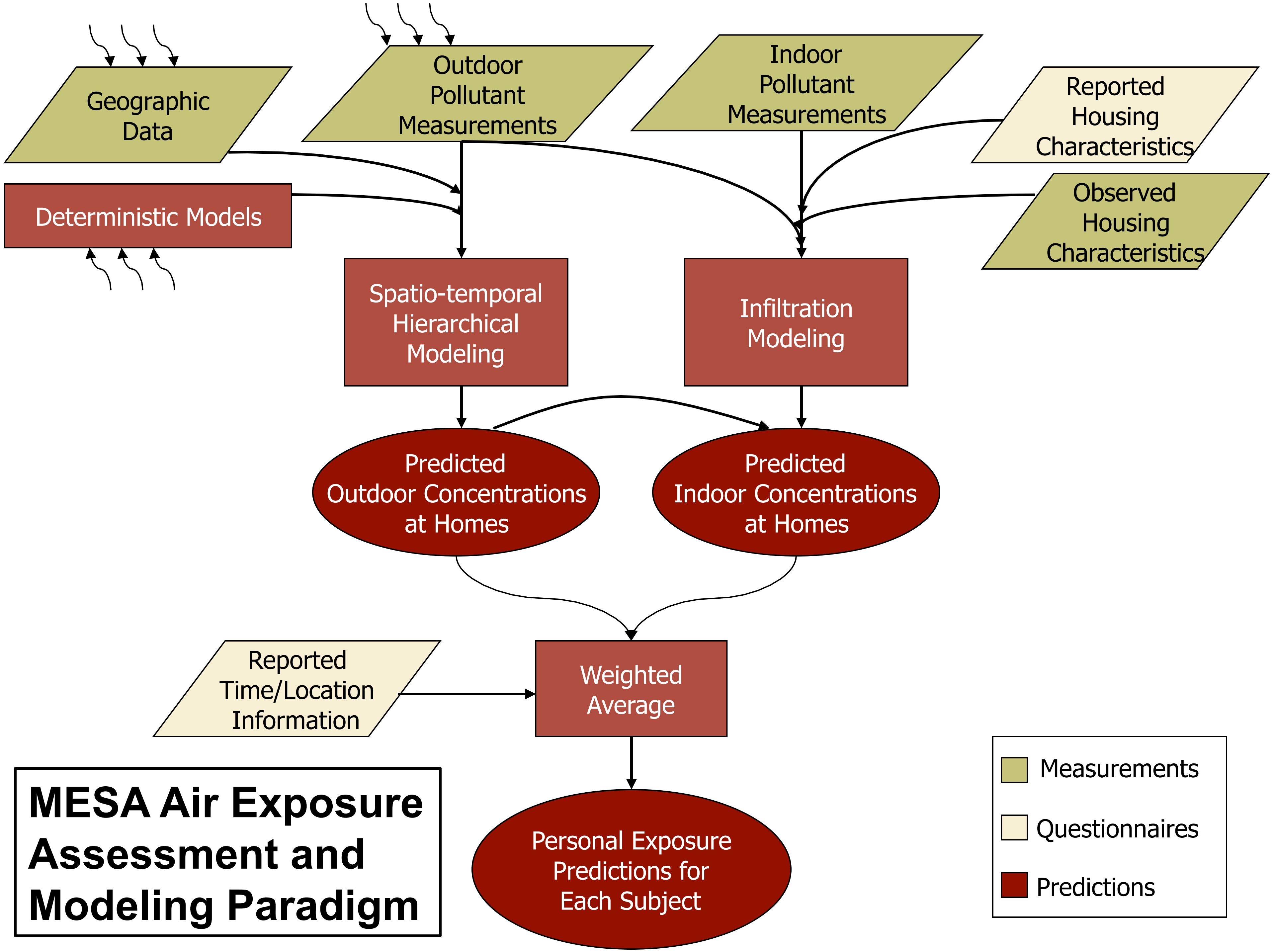
Example: MESA Air Study

- Multi-Ethnic Study of Atherosclerosis (MESA) Air Pollution Study
  - Ten-year national study funded by U.S. EPA
- Objective
  - Examine relationship between chronic air pollution exposure and subclinical cardiovascular disease progression
- Approach
  - Prospective cohort study with 6000-7000 subjects
    - 6 metropolitan areas (Los Angeles, New York, Chicago, Winston-Salem, Minneapolis-St. Paul, Baltimore)
  - Predict long term exposure for each subject
  - Longitudinally measure subclinical cardiovascular disease
  - Estimate effect of air pollution on CVD progression

Air Pollution Exposure Framework

- Personal exposure:
  E^P = ambient source (E^A) + non-ambient source (E^N)
  - E^A = ambient concentration (C^A) * attenuation (α)
    - Ambient concentration contributes to exposure both outdoors and indoors due to the infiltration of ambient pollution into indoor environments
  - Ambient exposure attenuation factor: $\alpha = [f^o + (1 - f^o)F_{inf}]$
    - Ambient attenuation is a weighted average of infiltration (F_{inf}), weighted by time spent outdoors (f^o)

- Exposure of interest: Ambient source (E^A) or total personal (E^P)



Geographic Data
Outdoor Pollutant Measurements
Indoor Pollutant Measurements
Reported Housing Characteristics
Deterministic Models
Observed Housing Characteristics
Spatio-temporal Hierarchical Modeling
Infiltration Modeling
Predicted Outdoor Concentrations at Homes
Predicted Indoor Concentrations at Homes
Reported Time/Location Information
Weighted Average
Personal Exposure Predictions for Each Subject
MESA Air Exposure Assessment and Modeling Paradigm
Measurements
Questionnaires
Predictions

Exposure Assessment Challenge

- Need to assign individual air pollution exposures to all subjects ➔ Predict from ambient monitoring and other data
  - Focus is on long-term average exposure
  - Impractical to measure individual exposure for all subjects
- Desired properties of prediction procedure
  - Minimal prediction error
  - Practical implementation (not too time consuming)
  - Good properties in health analyses
- Prediction approaches for long-term average exposures:
  - City-wide averages
    - Seminal cohort studies (6 cities, ACS) focused on variation between cities
  - Spatial models
  - Spatio-temporal models

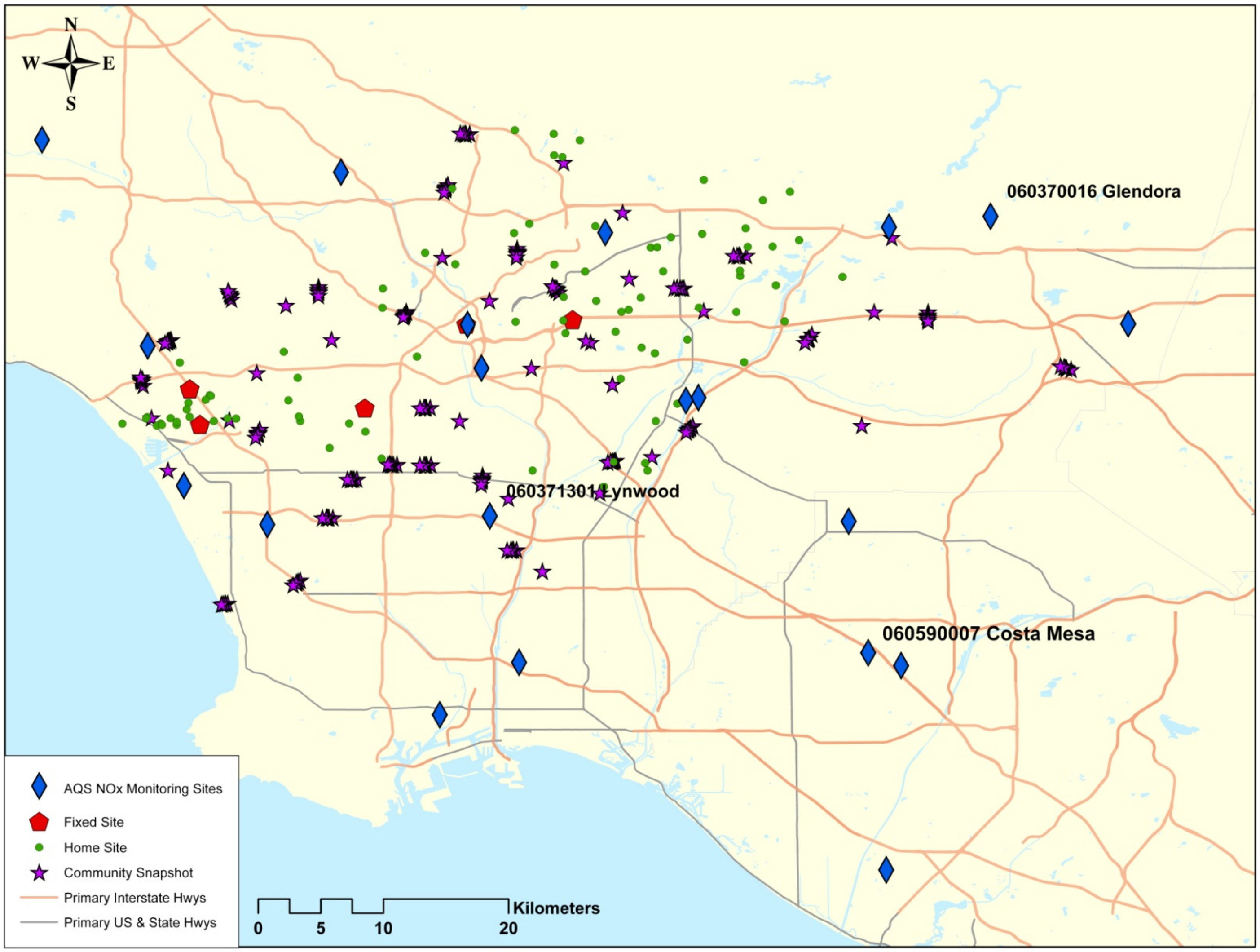
Spatial Prediction Modeling

- General approach:
  - Measure concentrations at a (relatively limited) set of monitoring locations
  - Predict concentrations at subject homes based on these monitoring data
  - Assume home concentration will be most like measured values at “similar” monitoring locations
    - Similar in terms of proximity and/or spatial covariates
- Conditions for spatial prediction to be appropriate
  - Interested in fixed time-period long-term averages
  - Monitoring data are representative of the time period of interest
    - Long-term averages or shorter but representative times
- Otherwise, need spatio-temporal predictions

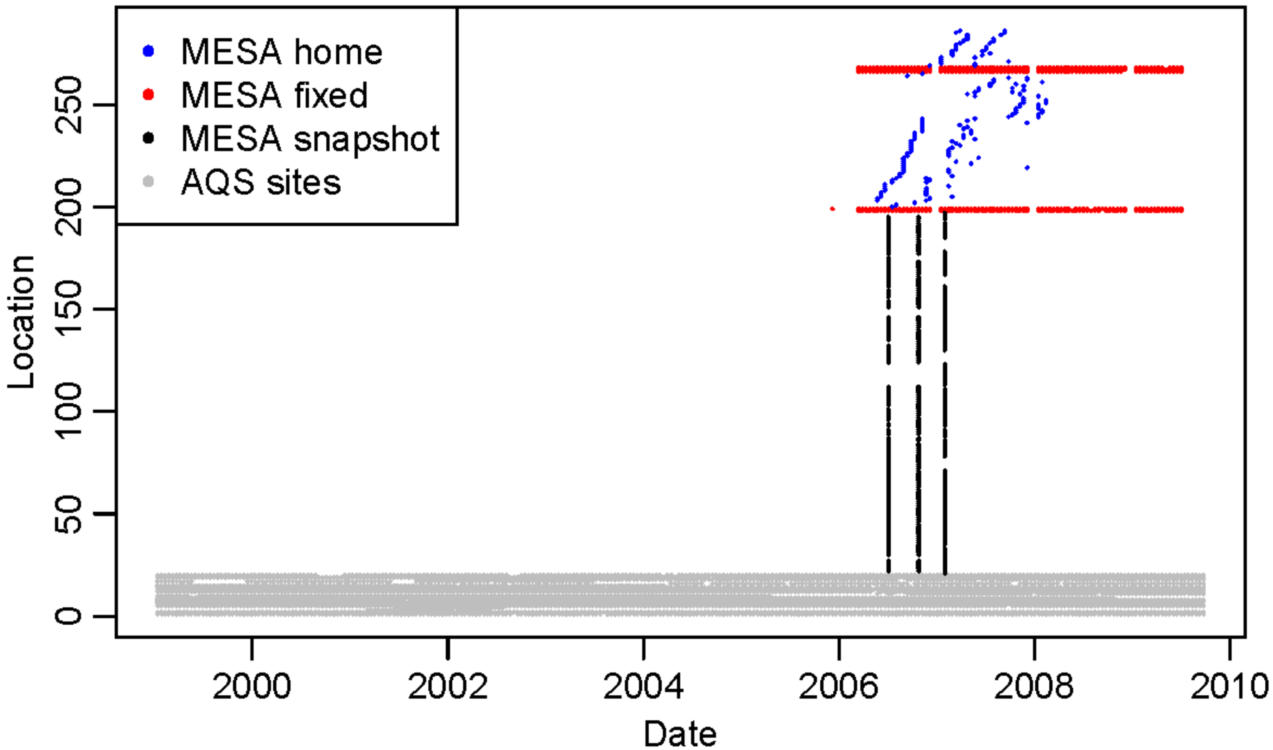
Spatial Prediction Methods

- Nearest monitor assignment
  - Assign concentration based on nearest monitoring locations
- *K*-means averaging
  - Average measured concentrations at the *K* nearest monitoring locations
- Inverse distance weighting
  - Average measured concentrations at all monitoring locations, weighted by distance
- Ordinary kriging
  - Smooth the data by minimizing the mean-squared error
- Spline smoothing
  - Theoretically equivalent to kriging; implementation details different
- Land use regression (LUR)
  - Predict from a regression model using geographic covariates
- Universal kriging
  - Predict by kriging combined with LUR

Locations of NO_x Monitors and Subject Homes in MESA Air (Los Angeles)

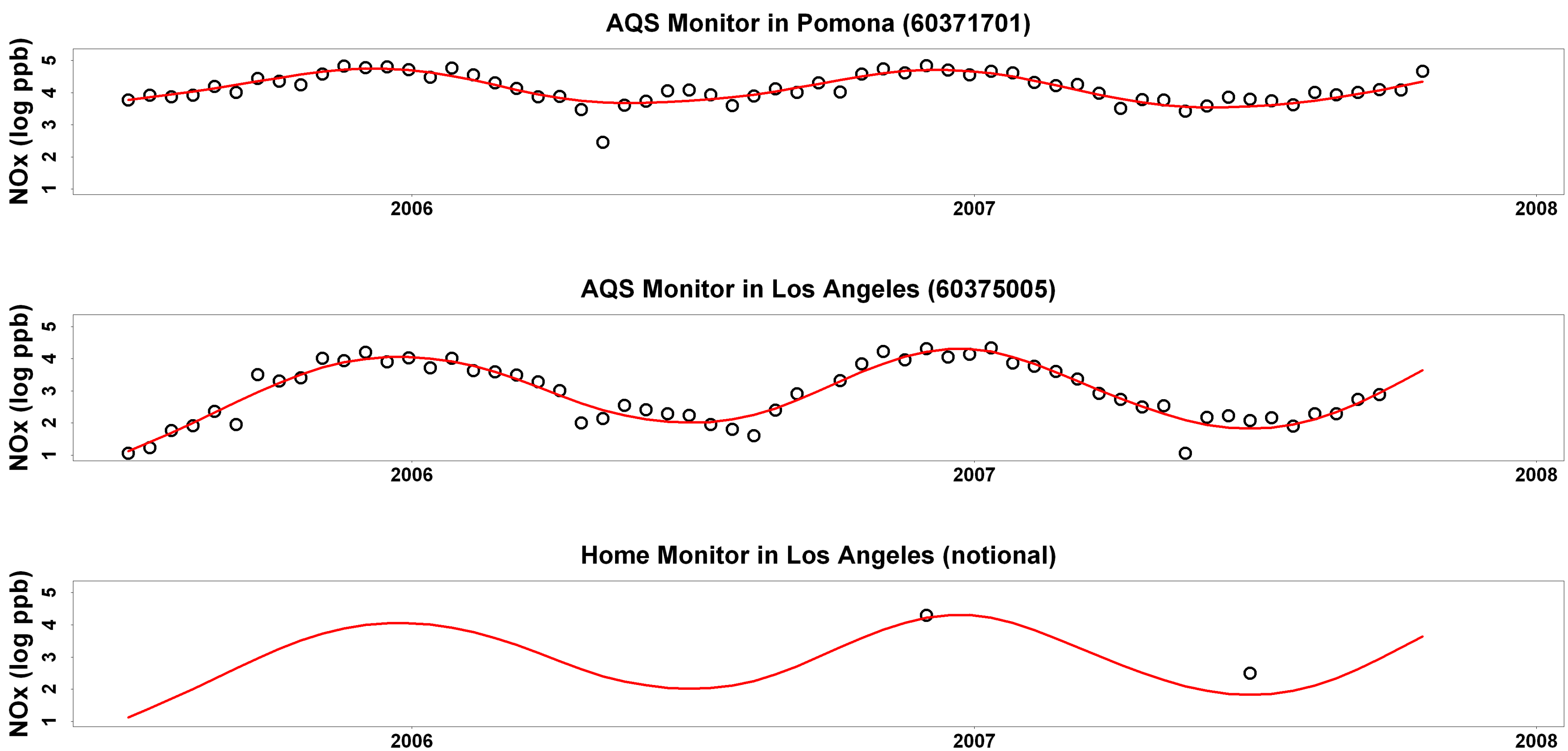


MESA Air NO_x Monitoring Data in Los Angeles

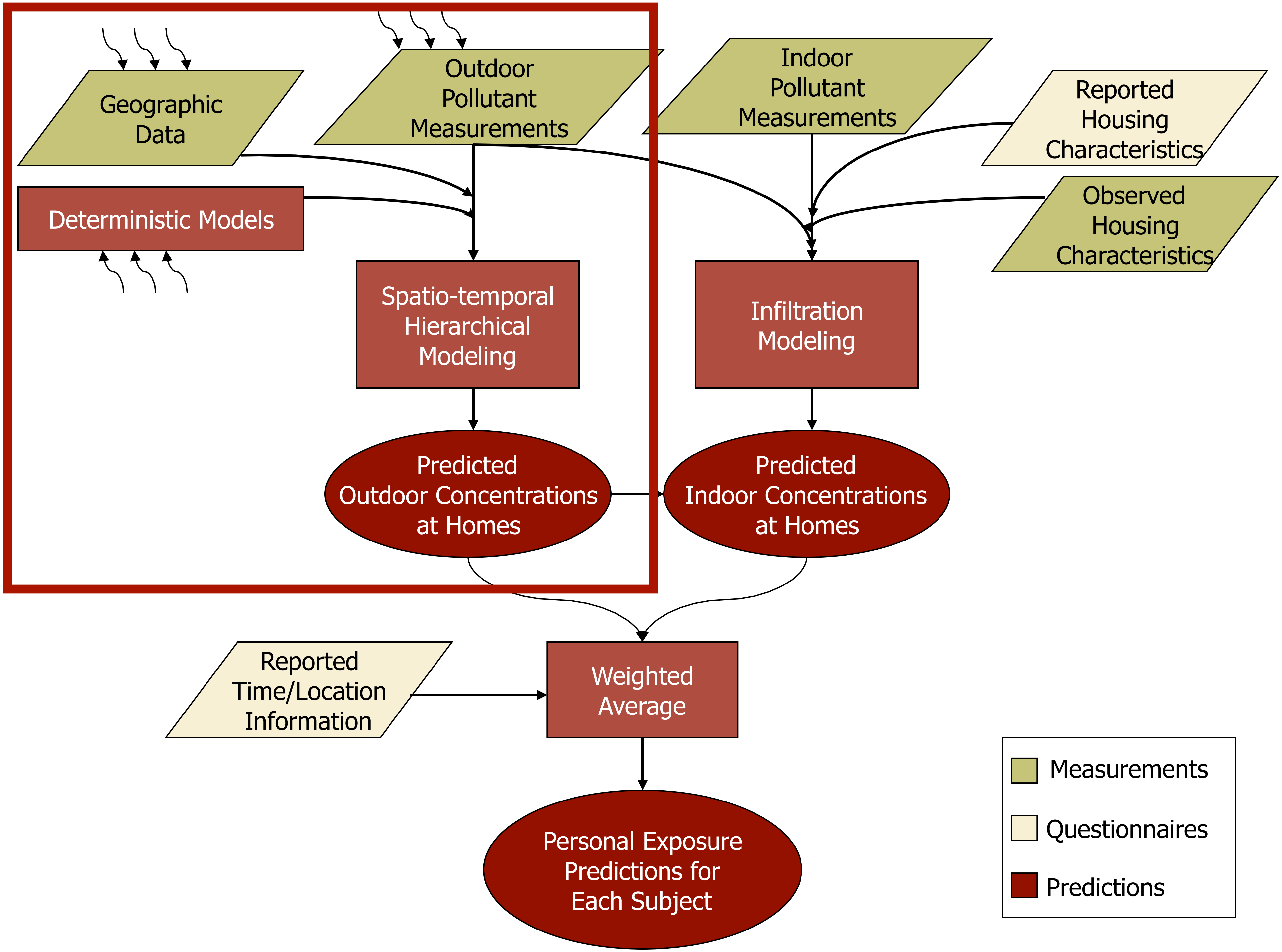


	# Sites	Start date	End date	# Obs
AQS	20	Jan 1999	Oct 2009	4180
MESA Air fixed	5	Dec 2005	Jul 2009	399
MESA Air home outdoor	84	May 2006	Feb 2008	155
MESA Air snapshot	177	Jul 2006	Jan 2007	449

Need For Spatio-Temporal Model



Space-time interaction and temporally sparse data suggest a spatio-temporal model to predict long-term averages



Geographic Data
Outdoor Pollutant Measurements
Indoor Pollutant Measurements
Reported Housing Characteristics
Deterministic Models
Observed Housing Characteristics
Spatio-temporal Hierarchical Modeling
Infiltration Modeling
Predicted Outdoor Concentrations at Homes
Predicted Indoor Concentrations at Homes
Reported Time/Location Information
Weighted Average
Personal Exposure Predictions for Each Subject
Measurements
Questionnaires
Predictions

MESA Air Spatio-Temporal Model Inputs

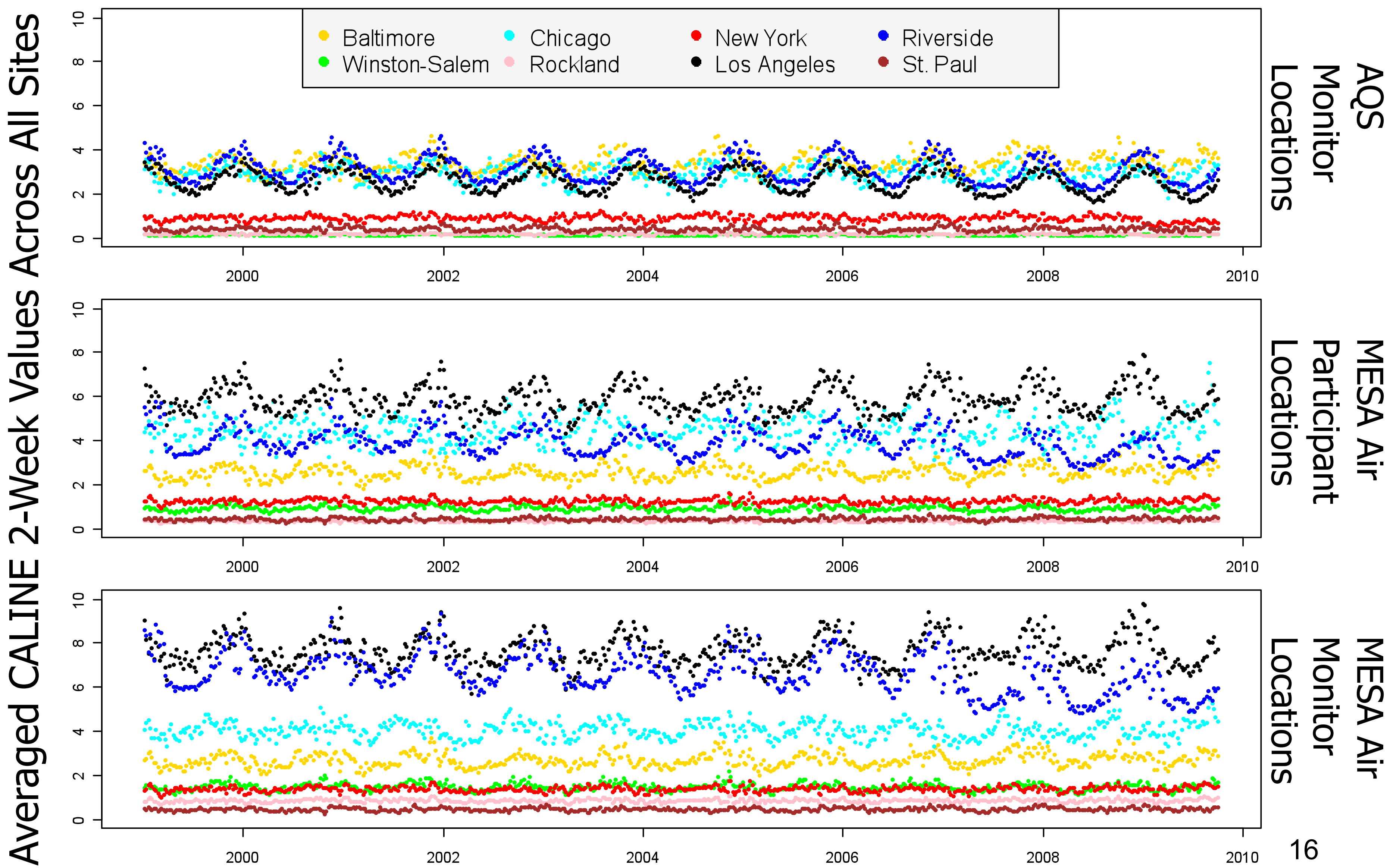
- Geographic Information System (GIS) predictors and coordinates
  - Spatial location
  - Road network & traffic calculations
  - Population density
  - Other point source and/or land use information
- Monitoring data
  - Air monitoring from existing EPA/AQS network
  - Air monitoring from supplemental MESA Air monitoring
  - Meteorological information
- Deterministic air quality model predictions
  - CMAQ: gridded photochemical model
  - AERMOD: bi-Gaussian plume/dispersion model
  - UCD/CIT air quality model: source-oriented 3D Eulerian model based on the CIT photochemical airshed model
  - CALINE: line dispersion model for traffic pollution

MESA Air GIS Covariates

Predictor Variable	Symbol	Units	Buffer radii	Functional Form
Land Use				
Population	Pop	Total people within buffer (m)	500,1000,1500,2000, 2500,3000,5000, 10000,15000	scaled by 1/10000
Intense Use Land	Int	km^2	50, 100, 150, 300, 500, 750	untransformed
Open Space Land	Open	km^2	50, 100, 150, 300, 500, 750	untransformed
Distance to Coast	D2C	meters	n/a	trunc. 15km & 25km scaled by 1/1000
Distance to industrial Source (rail road, air port, etc...)	D2V	meters	n/a	untransformed
Industrial NO_x emissions	NOx		3000,15000,30000	untransformed
Roadway				
Distance to nearest A1, A2, or A3	D2R	meters	n/a	Log10
Distance to nearest A1	D2A1	meters	n/a	Log10
Distance to nearest A2	D2A2	meters	n/a	Log10
Distance to nearest A3	D2A3	meters	n/a	Log10
Length of A1roads within buffer	A1	meters	50, 100, 150, 300, 500, 750, 1000, 5000 10000, 15000	scaled by 1/1000
Length of A2 roads within buffer	A2	meters	50, 100, 150, 300, 500, 750, 1000, 5000 10000, 15000	scaled by 1/1000
Length of A3 roads within buffer	A3	meters	50, 100, 150, 300, 500, 750, 1000, 5000 10000, 15000	scaled by 1/1000

Need variable selection to avoid overfitting!

Regional CALINE Predictions by Location Type

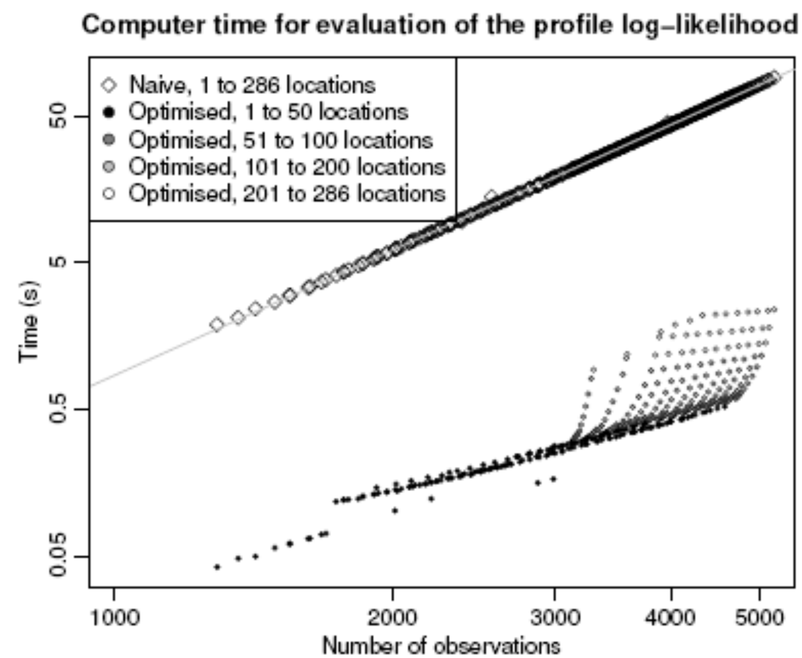


Spatio-Temporal Exposure Model

- $\boxed{\boldsymbol{C}_{s,t} = \mu_{s,t} + \nu_{s,t}}$ ⟵ **measured concentrations on log scale**
- $\boxed{\mu_{s,t} = \beta_{0,s} + \sum_{i=1,\ldots,m} \beta_{i,s} \boldsymbol{f}_i(\boldsymbol{t}) + \gamma \boldsymbol{M}(\boldsymbol{s},\boldsymbol{t})}$ ⟵ **temporal trends at location *s* + space-time covariate**
  - $\boldsymbol{f}_i(\boldsymbol{t})$ smooth temporal basis functions derived from data
  - $\beta_{i,s}$ spatial random fields distributed as $\boldsymbol{N}\left(\boldsymbol{X}_i\alpha_i, \Sigma(\phi_i, \sigma_i^2)\right)$
    - Geostatistical covariance structure with "land use regression" covariates for population, traffic, land use, etc.
  - $\boldsymbol{M}(\boldsymbol{s},\boldsymbol{t})$ space-time covariate
- $\boxed{\nu_{s,t}}$ ⟵ **variation from temporal trend (mean 0)**
  - Geostatistical spatial structure with simple temporal correlation
    - Process noise + measurement error

Estimation Methodology

- Large number of parameters and thousands of observations makes estimation challenging
  - Maximum likelihood estimation based on full Gaussian model works, but very computationally intensive
- Two approaches improve computational efficiency:
  - Reduce number of parameters to be optimized by using profile likelihood or REML
  - Reduce time for each likelihood computation by taking advantage of structure of model



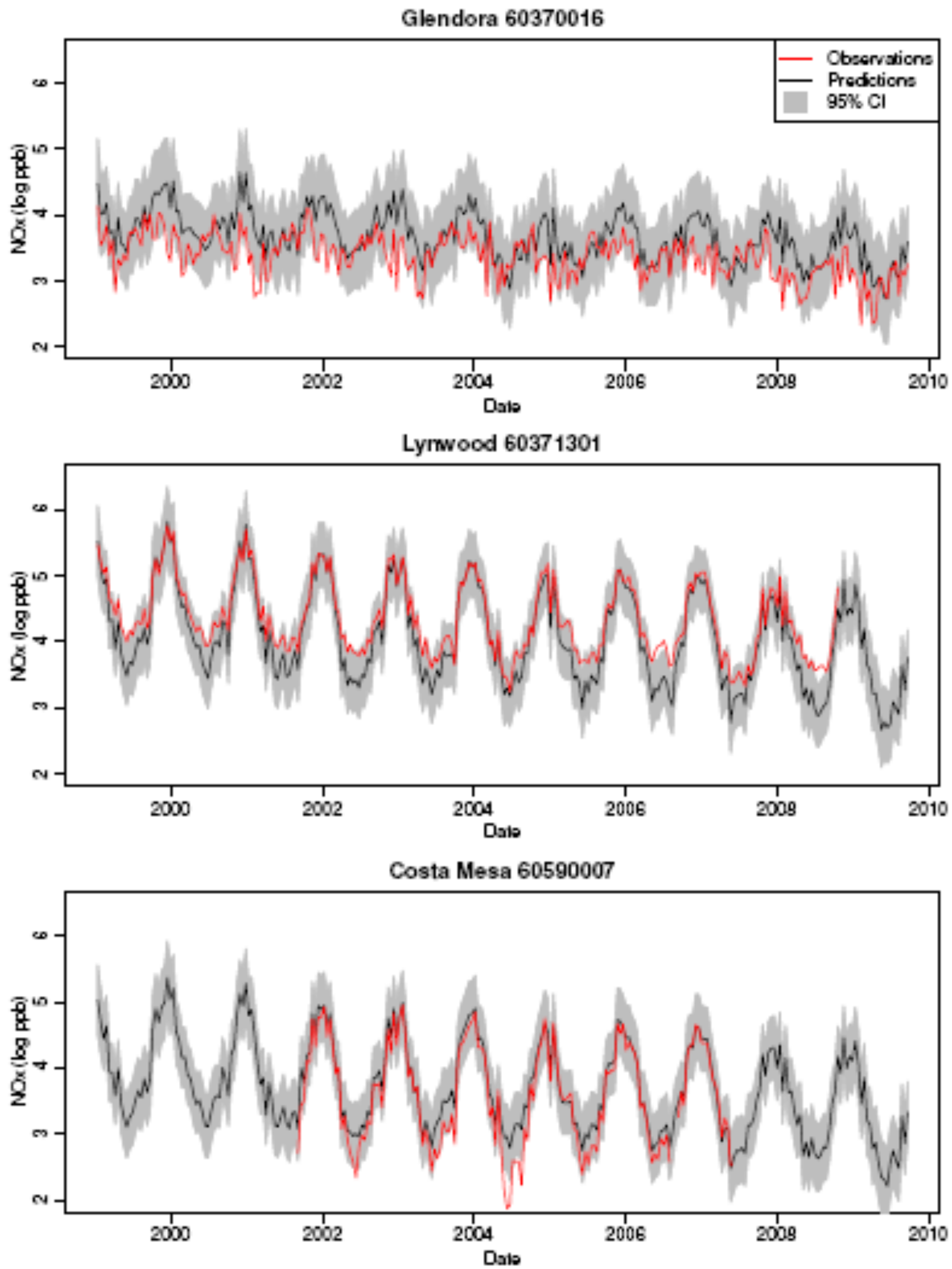
R Package

- MESA Air spatiotemporal model has been efficiently implemented in an R package
  - Johan Lindström, available on CRAN in 1-2 months
- So far, used to generate and cross-validate NO_x predictions in Los Angeles

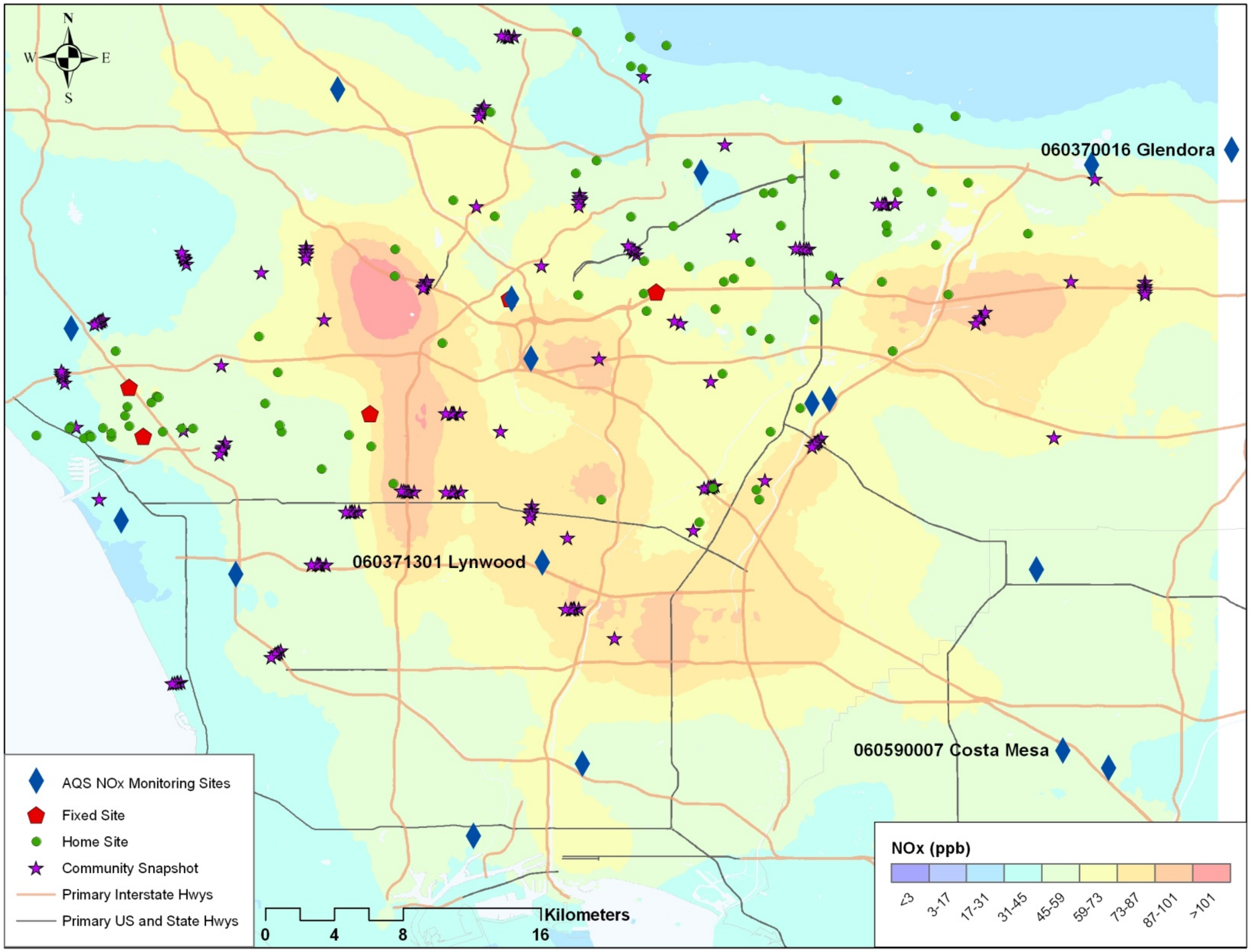
On linux: `install.packages("SpatioTemporal_0.1.0.tar.gz", repos=NULL, type="source")`

```
mesa.data.model <- create.data.model(mesa.data,list(c("log10.m.to.a1",
     "log10.m.to.a2", "km.to.coast"), "km.to.coast", "km.to.coast"))

##estimate parameters
par.est <- fit.mesa.model(x.init, mesa.data.model, type="p",
            hessian.all=TRUE, control=list(trace=3,maxit=1000))

##extract the estimated parameters
par <- par.est$res.best$par.all
##and uncertainties from the hessian
par.sd <- sqrt(diag(-solve(par.est$res.best$hessian.all)))

##Do cross-validated predictions with fixed parameters
pred.cv <- predictCV(par.est$res.best$par.all, mesa.data.model,
                     Ind.cv, silent = FALSE)
```

Predicted NOx Concentrations In Los Angeles:



Smooth Predicted Long-Term Average NO_x Concentrations in Los Angeles



Validation Strategies

- Must do some kind of validation study to test accuracy of predictions at locations not used to fit the model
  - Not sufficient to look at regression R^2 (and this is not available for kriging anyway)
- Ideally test with separate validation dataset not used in model selection or fitting
  - Typically infeasible because want to use all the data
- Cross-validation is a useful alternative
  - Fit the model repeatedly using different subsets of the data and test on the left-out locations
    - Leave-one-out, ten-fold, etc.
  - No universally best approach to cross validation, but there are some guiding principles
    - Each cross-validation training set should be similar in size to full dataset
    - Leave out highly correlated locations together

Cross-Validation of Los Angeles NO_x Predictions

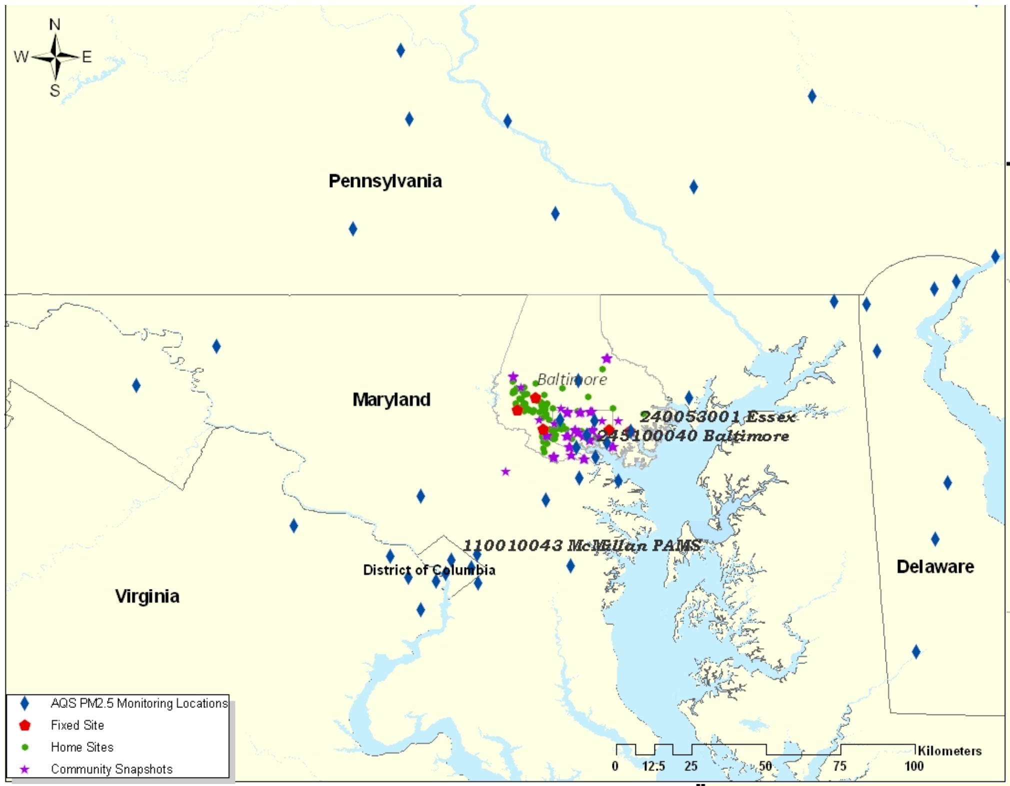
	No Caline			With Caline		
	RMSE	R^2	Cov.	RMSE	R^2	Cov.
AQS & MESA fixed						
2-week	17.90	0.80	0.91	18.12	0.79	0.90
Long-term avg.	11.97	0.58		12.26	0.56	
Snapshot						
2006-07-05	7.94	0.52	0.93	7.62	0.56	0.95
2006-10-25	13.32	0.68	0.97	13.32	0.68	0.95
2007-01-31	15.69	0.66	0.99	15.77	0.66	0.98
Home sites	9.34	0.89	0.97	9.06	0.90	0.95
Average		0.67			0.69	
Closest		0.74			0.76	
Smooth		0.74			0.76	

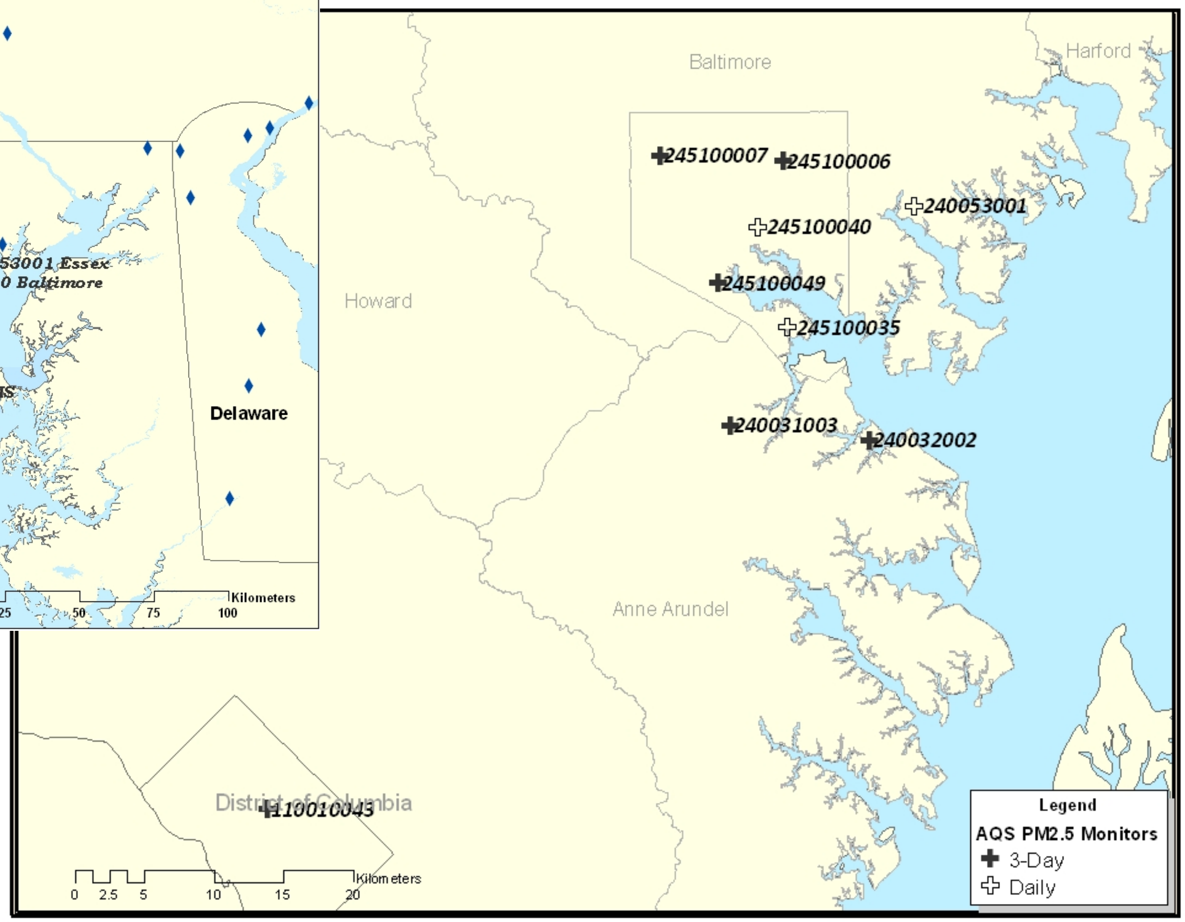
- Use cross-validation to assess accuracy of predicting long-term averages at subject homes
  - Modify R^2 at home sites so we don't "take credit" for predicting temporal variability

Initial Assessment of CMAQ for Use in MESA Air

- Approach:
  - Initial evaluation to determine how to incorporate CMAQ output into our spatio-temporal model
  - Examine scatterplots, summaries of correlations, and smooth trends
  - Focus on the effect of time scale
- Data:
  - One year (2002) of CMAQ predictions in Baltimore
    - 12 km grid
    - Interpolated to AQS locations in Baltimore City and greater metropolitan area
  - $PM_{2.5}$ data at AQS locations

Locations of the AQS $PM_{2.5}$ Sites in the Baltimore Area





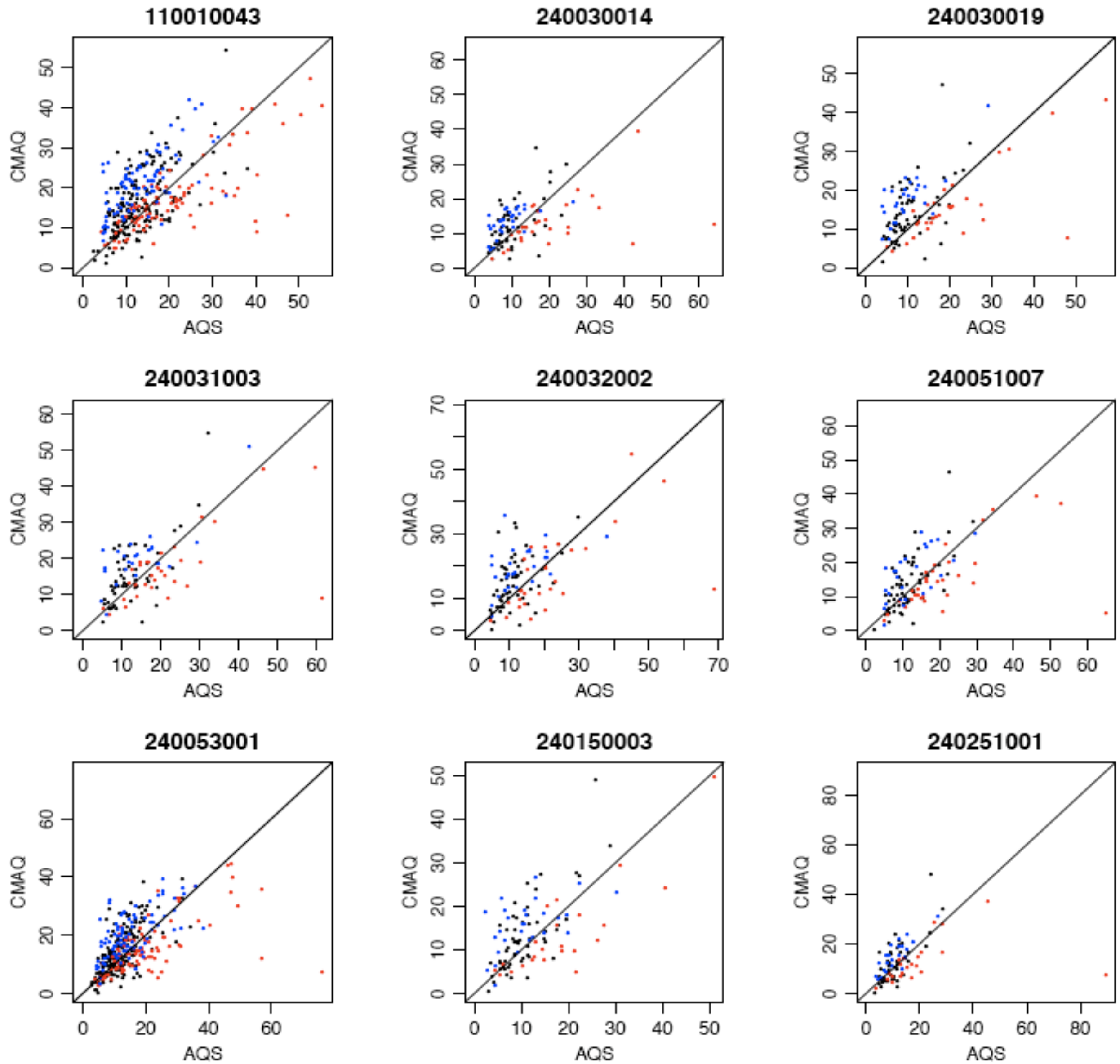
AQS sites operating in 2002

Daily Data: Interpolated CMAQ Predictions vs. AQS

Red: summer

Black: spring/fall

Blue: winter

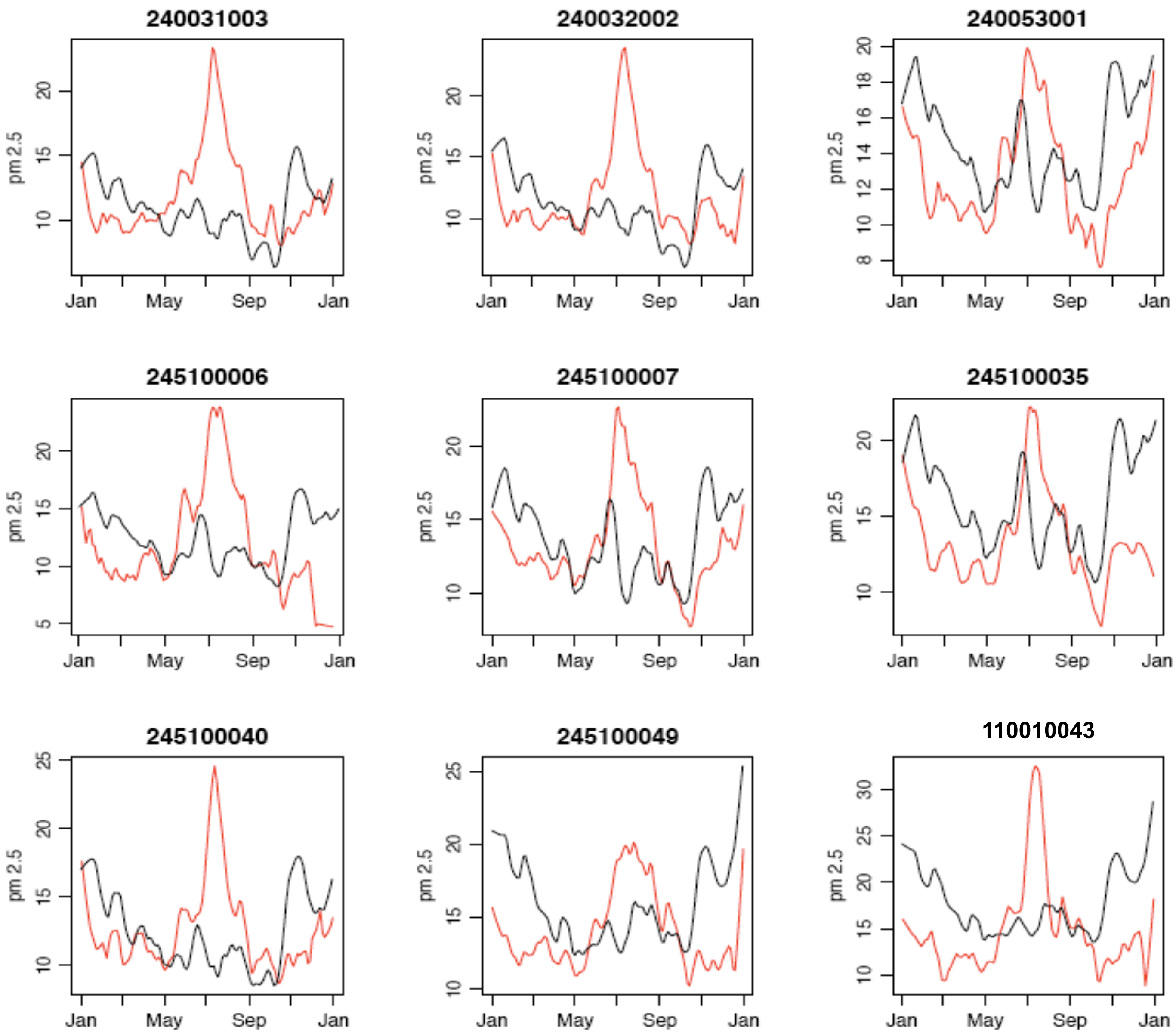


Seasonal Trends: CMAQ and AQS

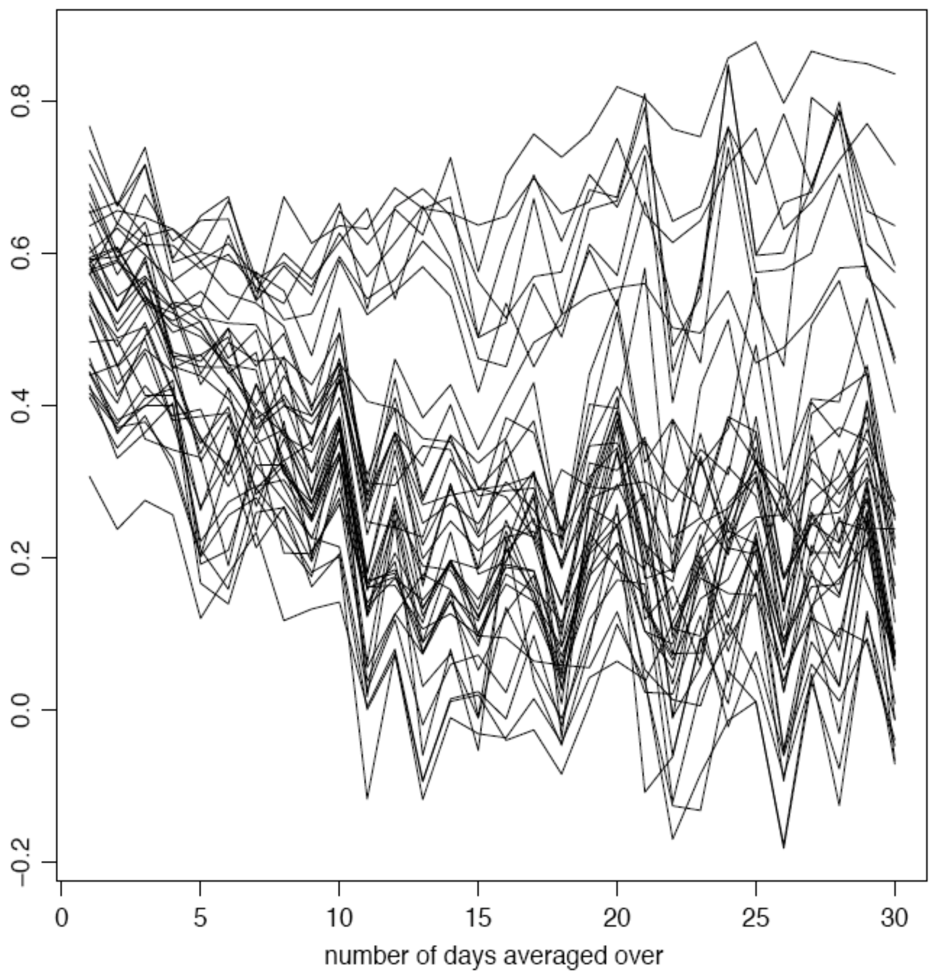
Seasonal trends on approximately monthly time scale:

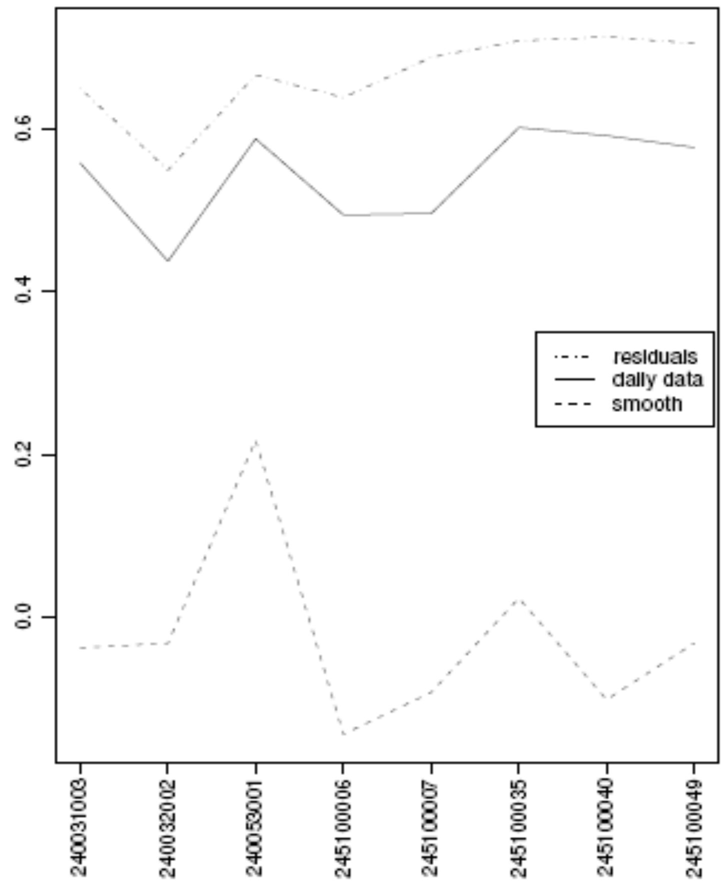
AQS

CMAQ



Correlations Between CMAQ and AQS: Effect of Temporal Averaging



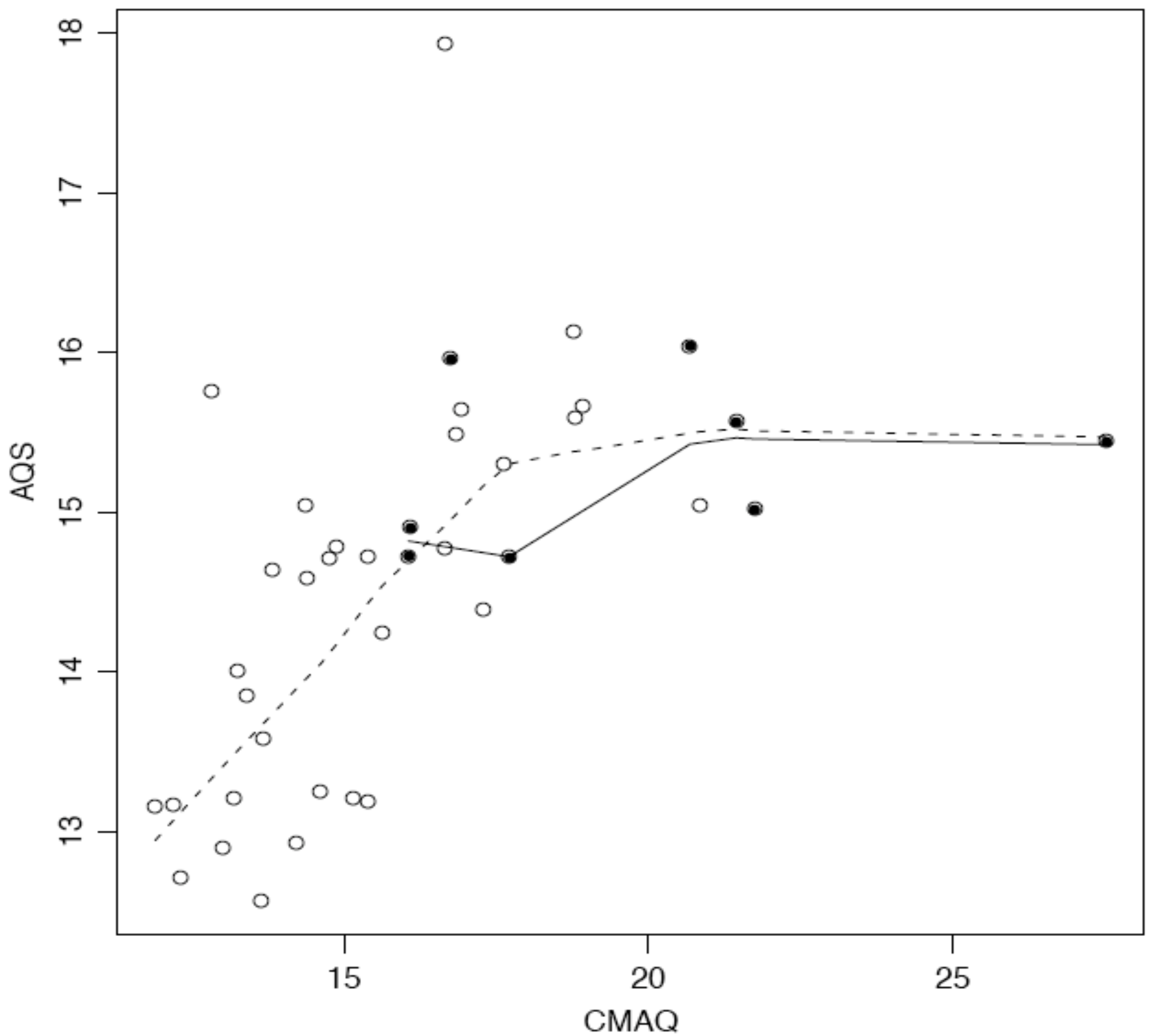


Correlations by site:
Effect of number of days averaged over

Correlations by model component:
Impact at each AQS site in Baltimore

Association of Annual Averages Across Sites: CMAQ vs. AQS

Solid points:
8 sites in
Baltimore City



Comments on CMAQ for Application to the MESA Air Spatio-Temporal Model

- Preliminary conclusion: Unlikely that CMAQ will improve the MESA Air spatio-temporal model
  - Weaker correlation of AQS and CMAQ at longer time scales
  - Seasonal structures are different
  - However
    - To date we have only evaluated one year of CMAQ predictions
    - There is some spatial correlation between CMAQ and AQS annual averages at larger spatial scales
    - There might be a benefit to including seasonally detrended CMAQ predictions
- Logistical issue: The MESA Air model needs air quality model predictions for ten years and many spatial locations

Summary and Discussion

- Evaluation of air quality model output for health studies should be done in the context of the exposure of interest in the health analysis
  - Cohort studies: Long-term average exposure
- Multiple options are available for exposure prediction. Method selection should consider:
  - Data at hand
  - Prediction goal
- All exposure models require validation
  - Validation should focus on the end use of the predictions
- Air quality model predictions have not improved the MESA Air spatio-temporal model
  - Results should be viewed in the context of the MESA Air study design and data
- Use of air quality model output and exposure predictions in health studies must also consider the health study design and data