Space-time modelling: A case study

Johan Lindström[1,2]

[1]Centre for Mathematical Sciences
Lund University

[2]Department of Statistics
University of Washington

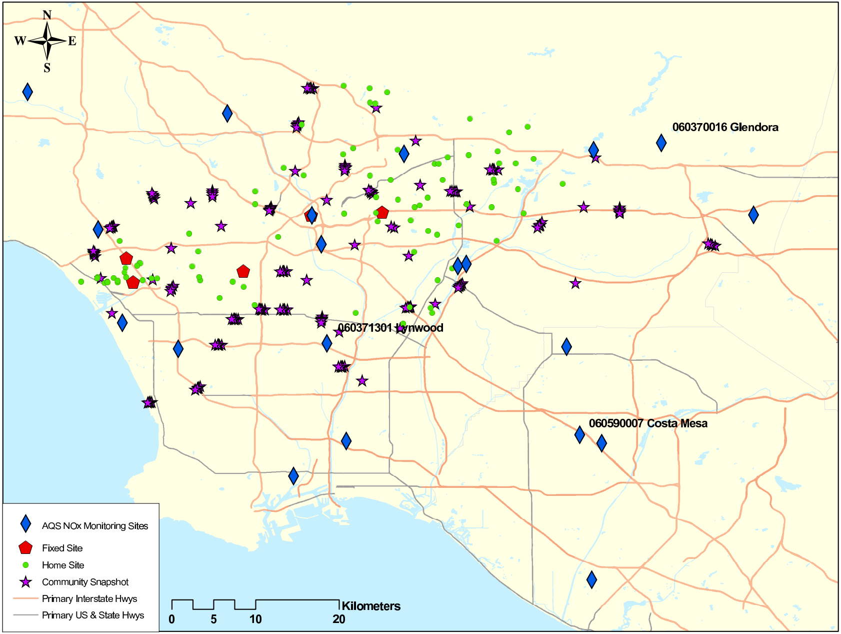
Pan-American Advanced Study Institute
Búzios
June 19, 2014

The MESA Air study

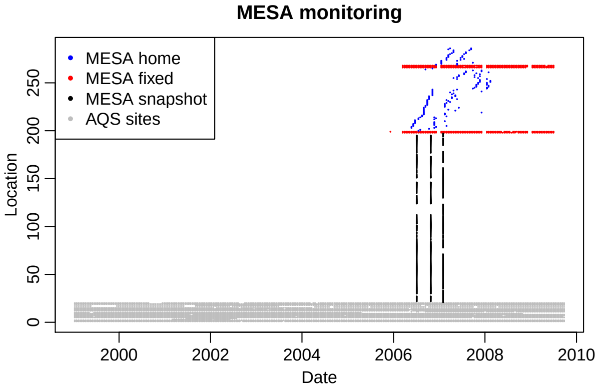
- The Multi-Ethnic Study of Atherosclerosis (MESA) is a large study of cardiovascular diseases.
- It follows more than 6 000 people from six communities.
  - Baltimore
  - Chicago
  - **Los Angeles**
  - Minneapolis – Saint Paul
  - New York
  - Winston–Salem

Basic problem: People don't live where we monitor.

(Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2013)


N
W
E
S
060370016 Glendora
060371301 Lynwood
060590007 Costa Mesa
AQS NOx Monitoring Sites
Fixed Site
Home Site
Community Snapshot
Primary Interstate Hwys
Primary US & State Hwys
0
5
10
20
Kilometers

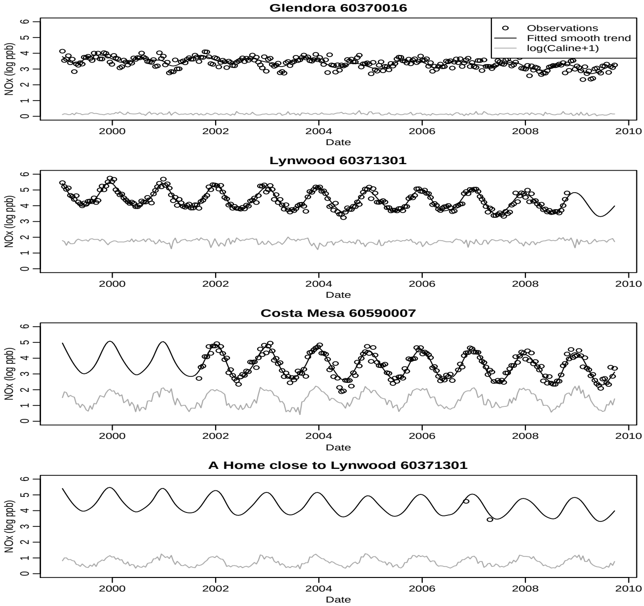
Available data — Los Angeles



Type of site	sites	Start date	End date	Nbr. of obs.
AQS	20	1999–01–27	2009–10–07	4180
MESA fixed	5	2005–12–07	2009-07-01	399
MESA home	84	2006–05–24	2008–02–13	155
MESA snapshot[1]	177	2006–07–05	2007–01–31	449

[1]Snapshot dates: 2006–07–05, 2006–10–25, and 2007–01–31

Available data — Los Angeles



Spatio-temporal model

We model the logarithm of each 2-week average as

$$y(s,t) = \sum_{l=1}^{L} \gamma_l \mathcal{M}_l(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t) + \nu(s,t).$$

$\mathcal{M}_l(s,t)$ Spatio-temporal covariates with coefficients γ_l.

$f_i(t)$ Smooth temporal trends with $f_1(t) \equiv 1$ and $f_2(t), \ldots, f_m(t)$ mean zero.

$\beta_i(s)$ Spatially varying coefficients for the temporal trends.

$\nu(s,t)$ Residuals, modelled as a mean zero Gaussian field that is independent in time but has spatial structure.

The smooth temporal trends, $f_i(t)$ are compute using a singular value decomposition of the data matrix, Y (see Fuentes et al., 2006, and the computer exercise).

Spatio-temporal model (cont.)

$$\beta_i(s) \in \mathsf{N}\left(X_i\alpha_i, \Sigma_{\beta_i}(\theta_B)\right)$$

- X_i Design matrices, that includes geographical covariates (different for each i).
- α_i Regression coefficients.
- Σ_{β_i} Covariance matrix describing additional spatial dependence not captured by the geographical covariates.
- θ_B Parameters of the covariance matrices.

$$\nu(s,t) \in \mathsf{N}\left(0, \Sigma_\nu(\theta_\nu)\right)$$

- Σ_ν Block diagonal covariance matrix for the residuals.
- θ_ν Parameters of the residual covariance matrix.

Combined model

Introducing

$$B = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{bmatrix} \qquad \Sigma_B(\theta_B) = \begin{bmatrix} \Sigma_{\beta_1}(\theta_B) & 0 & 0 \\ 0 & \Sigma_{\beta_2}(\theta_B) & 0 \\ 0 & 0 & \Sigma_{\beta_3}(\theta_B) \end{bmatrix}$$

our model becomes

$$[Y|\theta_B, \theta_\nu, \alpha] \in \mathsf{N}\left(FX\alpha, \Sigma_\nu(\theta_\nu) + F\Sigma_B(\theta_B)F^\top\right),$$

and parameters can be estimated by maximising the log-likelihood

$$l(\theta_B, \theta_\nu, \alpha|Y).$$

Likelihood simplifications

- Matrix algebra can be used to "simplify" the likelihood (Harville, 1997; Petersen and Pedersen, 2012).
- As an example we study the determinant of the log-likelihood

$$\log \left|\Sigma_\nu + F\Sigma_B F^\top\right| = \log|\Sigma_\nu| + \log|\Sigma_B| + \log\left|\Sigma_B^{-1} + F^\top \Sigma_\nu^{-1} F\right|$$

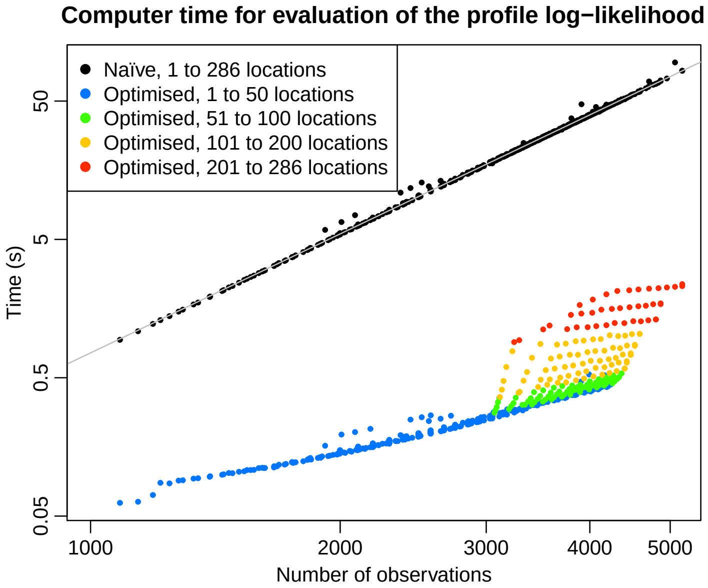
This may not seem simpler but:

1. $\Sigma_\nu + F\Sigma_B F^\top$ is dense $N \times N$-matrix, and computing the determinant requires $\mathcal{O}\left(N^3\right)$ operations.
2. Σ_ν and Σ_B are both block diagonal, with "small" blocks.
3. $\Sigma_B^{-1} + F^\top \Sigma_\nu^{-1} F$ is a dense $mn \times nm$-matrix. Computing the determinant requires $\mathcal{O}\left(m^3 n^3\right)$ operations, with $mn \ll N$.

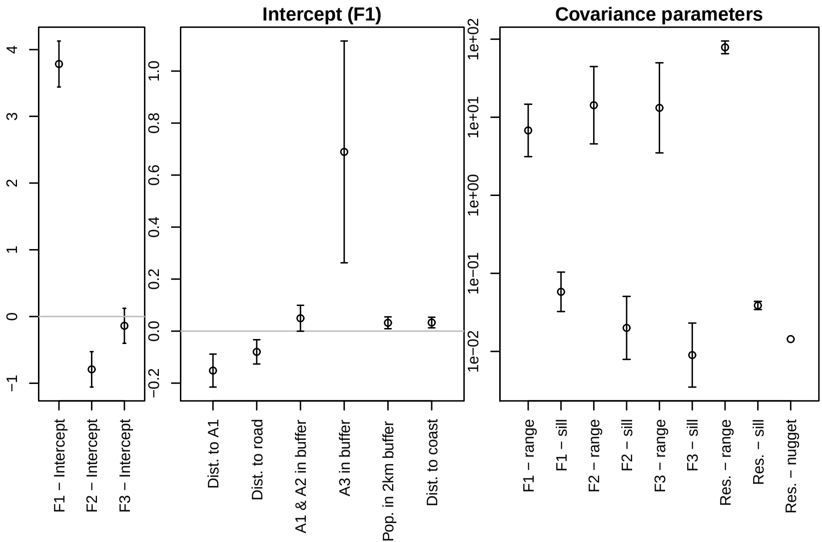
Where:

- N Total number of observations.
- n Total number of observed sites.
- m Number of temporal basis functions (incl. intercept).

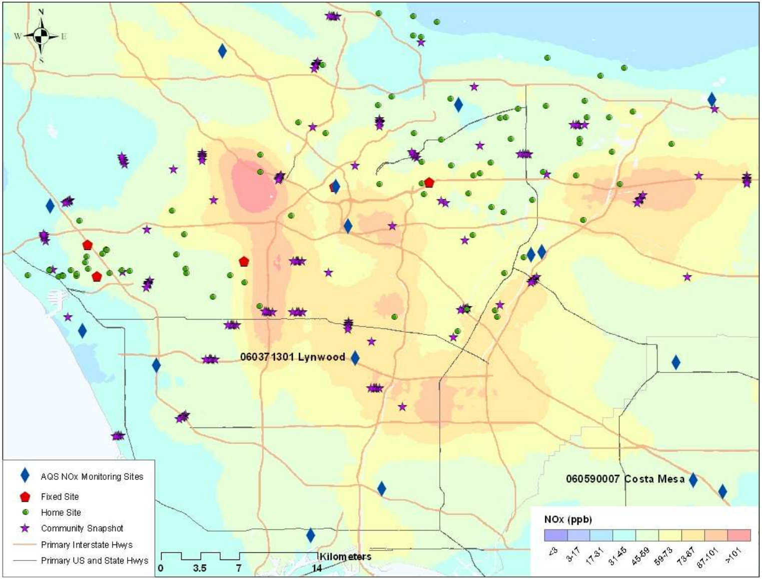
Computational issues



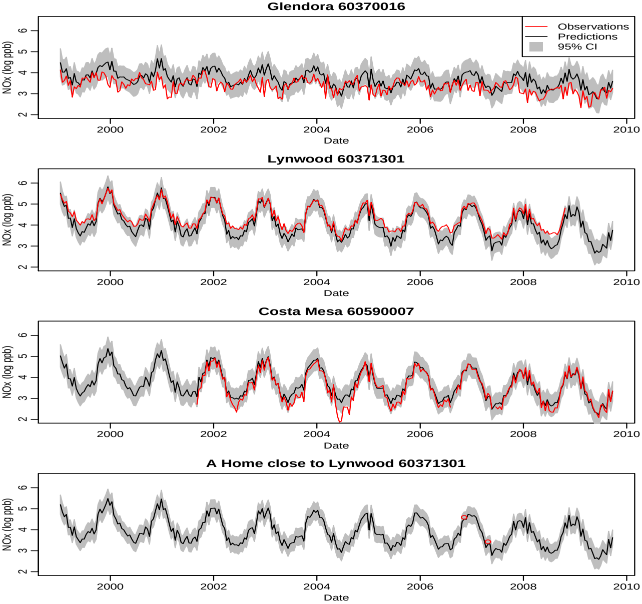
Estimated parameters



Predicted average NO_x concentration — Los Angeles



Model validation — NO_x in Los Angeles



Bibliography I

Fuentes, M., Guttorp, P., and Sampson, P. D. (2006), "Using transforms to analyze space-time processes," in *Statistical Methods for Spatio-Temporal Systems*, eds. Finkenstädt, B., Held, L., and Isham, V., Chapman & Hall/CRC, pp. 77–150.

Harville, D. A. (1997), *Matrix Algebra From a Statistician's Perspective*, Springer, 1st ed.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A., Richards, M., Larson, T., and Sheppard, L. (2013), "A Flexible Spatio-Temporal Model for Air Pollution with Spatial and Spatio-Temporal Covariates," *Environ. Ecol. Stat.*, 1352-8505, 1–23.

Petersen, K. B. and Pedersen, M. S. (2012), "The Matrix Cookbook," http://matrixcookbook.com.

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., and Kaufman, J. D. (2011), "Pragmatic Estimation of a Spatio-temporal Air Quality Model with Irregular Monitoring Data," *Atmos. Environ.*, 45, 6593–6606.

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S., and Kaufman, J. (2010), "Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies," *Environmetrics*, 21, 606–631.