Spatiotemporal Exposure Prediction with Penalized Regression

Nathan A. Ryder and Joshua P. Keller (2022)

Scott Cleven and Annika Helverson

BIOS 7240

May 8, 2023

Epidemiological Motivation

Particulate matter (PM) is a mixture of tiny particles that are suspended in the air.

- Includes nitrates, sulfates, organic matter, metals, and soil dust.
- Contributes to 4.7% of disability-adjusted life years (Murray et al. 2020).
- Both long and short term exposures have a causal relationship with adverse respiratory and cardiovascular health.
- EPA identifies sulfate (SO_4^{2-}) as associated with respiratory and cardiovascular health effects, but health effects vary by component.

Methodological Motivation

The epidemiological question at hand requires spatio-temporal predictions of ambient PM concentrations.

- Many viable spatio-temporal modeling approaches already exist with universal kriging (UK) performing the best for PM data (Berrocal et al. 2020).
- UK fits a new model for each day, ignoring temporal information.
- Spatiotemporal extension of UK was introduced by Lindström et al. (2014) and implemented in `SpatioTemporal` R package but iterative optimization is used to estimate covariance parameters so it is computationally burdensome.

Proposal

This paper introduces a penalized regression model for spatiotemporal data that:

- Penalizes overfitting.
- Smooths over adjacent time points.
- Is computationally efficient.
- Yields accurate predictions over both time and space.
- Can be used for spatiotemporal applications beyond PM data as well.

Penalized Regression Model

Objective Function:

$$Q(\beta|X,y) = \sum_{i=1}^{n}\sum_{t=1}^{T} I_{it}(x_{it} - r_{it}^T\beta_t)^2 + \sum_{t=1}^{T} g_1(\lambda_1, \beta_t) + \lambda_2 \sum_{i=1}^{n}\sum_{t=2}^{T}(r_{it}^T\beta_t - r_{i(t-1)}^T\beta_{t-1}) \quad (1)$$

Terms:

1. Quadratic loss
2. Discourages overfitting (Lasso)
3. Smooths over adjacent predictions

Spatial And Temporal Covariates

- Spatiotemporal Predictors r_{it}.
- Thin Plate Regression Splines (TPRS) at each site location.
  - Provide covariate values at any site we want to predict.
- Allows for predictive flexibility but with added computational costs.

Estimation

$\hat{\beta}$ has a closed form solution:

$$\hat{\beta} = (R_{obs}^T R_{obs} + \Lambda_1 + \lambda_2 R^T D^T D R)^{-1} R_{obs}^T x \tag{2}$$

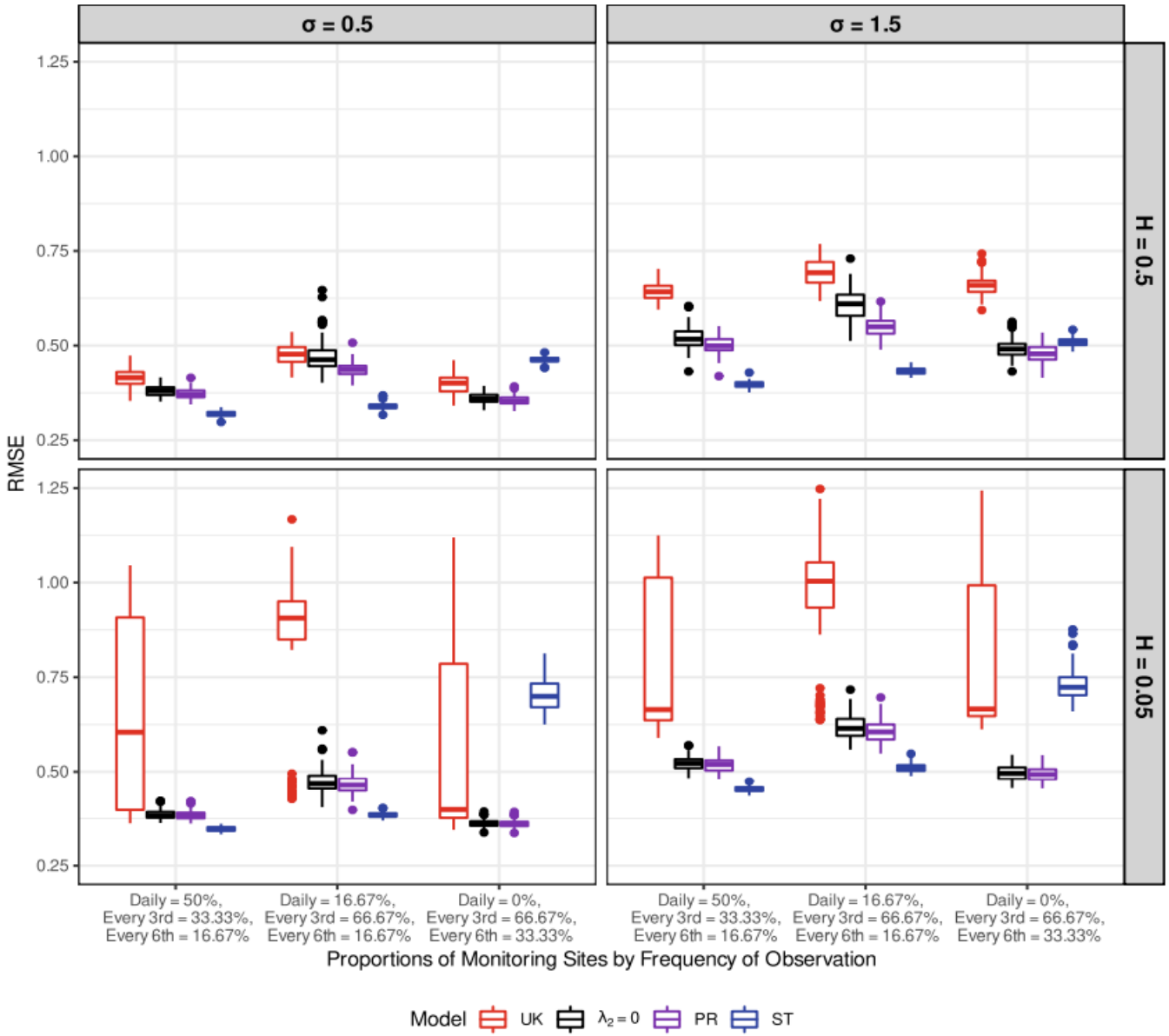
- R_{obs} is a matrix of stacked block diagonal matrices of observed covariate values at for each location and time.
- R also includes blocks for unobserved times.
- Λ_1 is the corresponding block diagonal matrix of λ_1 values.
- D is the distance matrix with potential entries {0, 1, and -1}

And each λ is estimated using cross validation.

Setup

- A 64x64 spatial grid over 60 time points.
- Of the 4096 grid points, 500 training and 1000 testing locations were randomly selected.
- error standard deviation $\sigma = 0.5$ or 1.5
- H representing the random walk over time of H = 0.5 and 0.05 (larger H means greater variation across time).
- Compared their model (PR) to the universal kiging (UK), ridge regression ($\lambda_2 = 0$), and to the `SpatioTemporal` (ST) models.
- Model comparison was done using root mean square error (RSME).
- Daily missingness was also accounted for where observations are observed daily, every three days, and every six days.

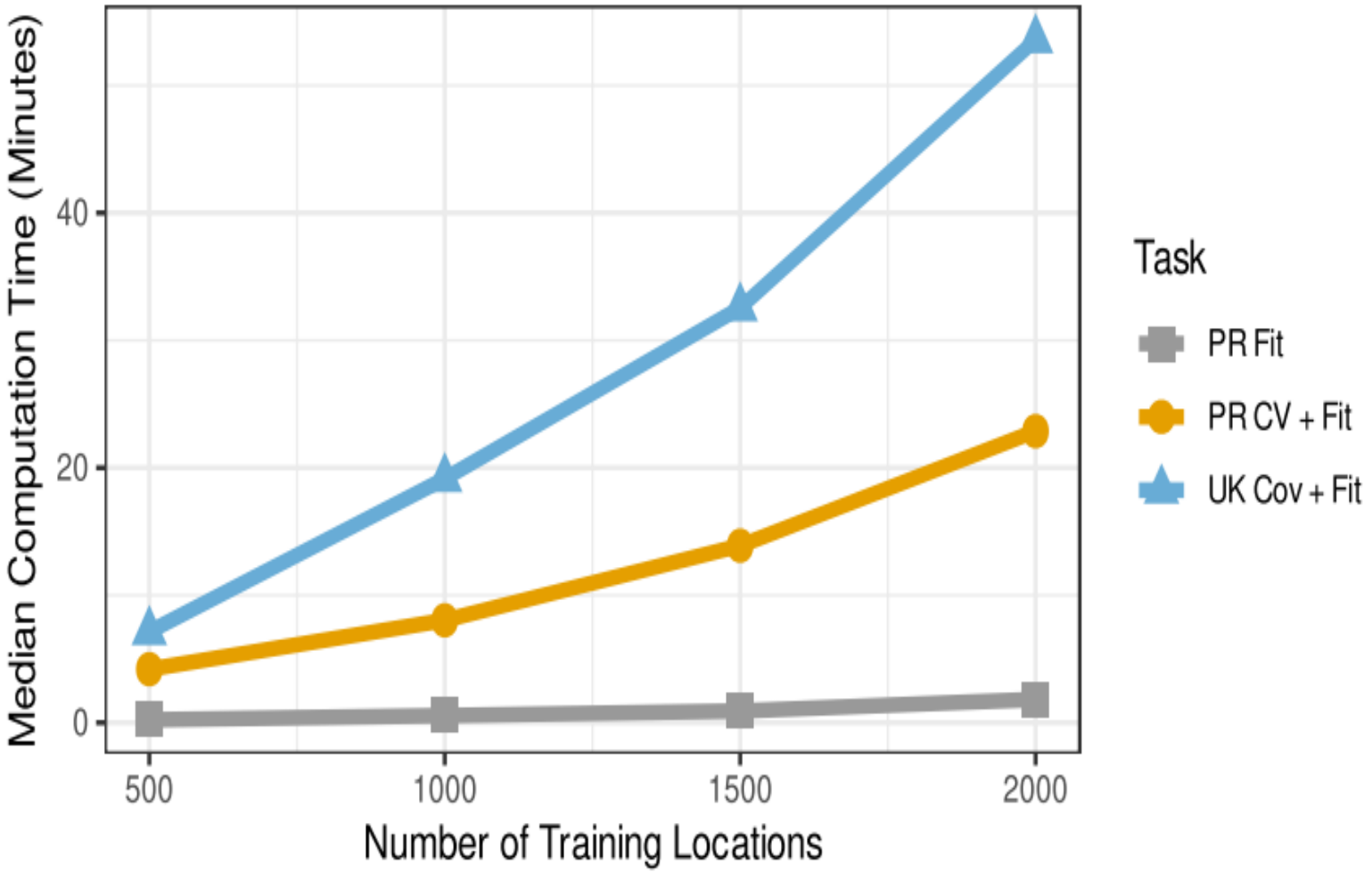
Simulation Results



Simulation Results Cont.

- PR has equal or lower RMSE to UK and ridge regression across all scenarios.
- ST has lowest RMSE except when all locations are only observed every 3rd or 6th day.
- When no monitors have daily data, PR fits every third day as if it were daily and outperforms ST.
- If all monitors have some data with each frequency (daily, every 3rd, every 6th) and similar number of observations made on each date, PR does not beat ST.
- As non-spatial error (σ) increases, all model fits worsen.
- Increased daily fluctuations (H = 0.05) decrease predictive accuracy in all models and affect ridge more that PR.

Computation Time Results



Computation Time Results Cont.

- PR requires inversion of a sparse $pT \times pT$ matrix to estimate β (Eq. 2).
- UK requires inversion of an $nT \times nT$ matrix to estimate covariance parameters (sill and range).
- ST uses an optimization algorithm to get ML estimates of its parameters in space and time. It takes 181.2 min. for $N_{train} = 500$ and 5,486.5 min. for $N_{train} = 2000$ and thus is infeasible for many locations or with limited resources.

Setup

- PM was measured for concentrations of at least 2.5 and 10, as well as Sulfate (SO_4^{2-}) and Silicon (Si) over the eastern United States.
- The goal is to predict pollutant particle density over time.
- PM tended to be measured daily while SO_4^{2-} and Si tended to be measured every three or six days.
- Spatiotemporal predictors:
  - 3-hour average surface temperatures from the North American Regional Reanalysis (NARR).
  - The Community Multiscale Air Quality Modeling Systems's (CMAQ) model results for air quality based on atmospheric dispersion.
  - The EPA's daily $PM_{2.5}$ Predictions.

Data Filtering and Transformation

- Only use federal reference method (FRM) monitors which yield 24-hour data.
- Remove concentrations equal to 0 (likely invalid).
- If monitors are collocated, only keep the earliest registered monitor at that site.
- PM concentrations are skewed so log transform observed concentrations and CMAQ values.
- Similarly filter sulfate and silicon concentrations and additionally replace negative values with the lowest observed positive value.

Model Fitting

- Predict each pollutant using the objective function from Eq. 1 with daily average temperatures and TPRS basis functions as predictors.
- Similarly predict $PM_{2.5}$ with CMAQ as an additional predictor.
- CMAQ is a deterministic predictor of $PM_{2.5}$ and excluded from the overfitting penalty Γ_1.
- Fix $\lambda_2 = 0$ to select λ_1, then select λ_2.
- Also fit UK and ST with 10-fold cross validation for comparison.

RMSE Results (Annual Average)

Table 2. Cross-validated RMSE and R^2 values across all dates and sites ("Overall") and from by-site annual average predictions and observations ("Annual Average")

		Overall		Annual Average	
Pollutant	Model	RMSE	R^2	RMSE	R^2
$PM_{2.5}$	PR	2.184	0.731	0.917	0.629
	ST	1.973	0.780	0.899	0.634
	UK	2.389	0.677	0.875	0.651
PM_{10}	PR	9.569	0.311	4.375	0.260
	ST	9.275	0.350	4.446	0.235
	UK	13.503	0.009	4.746	0.130
SO_4^{2-}	PR	0.456	0.530	0.193	0.627
	ST	0.535	0.373	0.223	0.592
	UK	0.771	0.057	0.279	0.434
Si	PR	0.119	0.455	0.038	0.508
	ST	0.137	0.308	0.047	0.309
	UK	0.138	0.308	0.041	0.472

We fit penalized regression (PR) per Eq. (3) using daily average temperature values and some number of TPRS basis functions (see Sect. 4.5) as predictors. For $PM_{2.5}$, we also added logged CMAQ values as predictors

Discussion

The proposed penalized regression model for spatiotemporal prediction penalizes overfitting and smooths over adjacent time points. TPRS basis functions are used as predictors and we can get daily average values at any spatial location in the domain.

- Can outperform UK and ST depending on the data conditions (infrequent measurement times).
- Scales well computationally to increasing spatial locations.
- Predictive accuracy similar to UK but worse than ST for $PM_{2.5}$ and PM_{10} but best (or tied best) performance for the lesser studied, sulfate and silicon concentrations.
- Less sensitive to spatial variability and missingness than UK.
- Only predicts dates with at least one observation but can interpolate to get estimates at missing or unobservable dates.
- Flexible definition of temporal adjacency: can set any amount of time as adjacent (1 day, 1 week, 11 hours, etc.).

Discussion cont.

Extensions to the model may be possible through clever application of new penalty terms.

- May be possible to to predict multiple pollutants at once.
- Extensions to other fields may require slight manipulation.
- However, the penalty must remain convex or else the solution to $\hat{\beta}$ will no longer be closed-form.
  - Generalized optimization techniques can be used to get around this problem but the computation time will increase.

Questions?

Matrix Forms

$$\mathbf{x} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1T} \\ x_{21} \\ \vdots \\ x_{nT} \end{bmatrix}, \quad \boldsymbol{\beta}_t = \begin{bmatrix} \beta_{0t} \\ \vdots \\ \beta_{(p-1)t} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_T \end{bmatrix}.$$

Matrix Forms cont.

$$\underset{p\times p}{\Gamma_1} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & \lambda_1 & 0 & \cdots \\ 0 & 0 & \lambda_1 & \\ \vdots & & \ddots & \ddots \end{bmatrix}, \text{ so that } \underset{pT\times pT}{\Lambda_1} = \begin{bmatrix} \underset{p\times p}{\Gamma_1} & 0 & 0 & \cdots \\ 0 & \underset{p\times p}{\Gamma_1} & 0 & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}.$$

Matrix Forms cont.

$$\mathbf{r}_{it} = \begin{bmatrix} r_{0it} \\ \vdots \\ r_{(p-1)it} \end{bmatrix} \quad \text{and} \quad \underset{nT \times pT}{\mathbf{R}} = \begin{bmatrix} \mathbf{r}_{11}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{r}_{12}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{0}^\top & \ddots & & \\ \vdots & \vdots & & & \\ & & & & \mathbf{r}_{1T}^\top \\ \mathbf{r}_{21}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{r}_{22}^\top & \mathbf{0}^\top & & \\ \mathbf{0}^\top & \mathbf{0}^\top & \ddots & & \\ \vdots & \vdots & & & \\ & & & & \mathbf{r}_{nT}^\top \end{bmatrix}.$$

Matrix Forms cont.

$$\underset{(T-1)\times T}{\mathbf{d}} = \begin{bmatrix} -1 & 1 & 0 & \cdots \\ 0 & -1 & 1 & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}, \quad \text{so that} \quad \underset{n(T-1)\times nT}{\mathbf{D}} = \begin{bmatrix} \underset{(T-1)\times T}{\mathbf{d}} & 0 & 0 & \cdots \\ 0 & \underset{(T-1)\times T}{\mathbf{d}} & 0 & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}.$$

Matrix Forms cont.

$$\mathrm{R}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{r}_{11}^{\top}\boldsymbol{\beta}_1 \\ \mathbf{r}_{12}^{\top}\boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{r}_{1T}^{\top}\boldsymbol{\beta}_T \\ \mathbf{r}_{21}^{\top}\boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{r}_{nT}^{\top}\boldsymbol{\beta}_T \end{bmatrix} \quad \text{and} \quad \mathrm{DR}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{r}_{12}^{\top}\boldsymbol{\beta}_2 - \mathbf{r}_{11}^{\top}\boldsymbol{\beta}_1 \\ \mathbf{r}_{13}^{\top}\boldsymbol{\beta}_3 - \mathbf{r}_{12}^{\top}\boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{r}_{1T}^{\top}\boldsymbol{\beta}_T - \mathbf{r}_{1(T-1)}^{\top}\boldsymbol{\beta}_{T-1} \\ \mathbf{r}_{22}^{\top}\boldsymbol{\beta}_2 - \mathbf{r}_{21}^{\top}\boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{r}_{nT}^{\top}\boldsymbol{\beta}_T - \mathbf{r}_{n(T-1)}^{\top}\boldsymbol{\beta}_{T-1} \end{bmatrix}.$$

Summary Results

Table 1. Summary statistics for ambient pollutant concentrations in 2017 after log transformation, along with the distribution of monitoring sites by frequency of observation

									Number of monitoring sites by days observed		
Pollutant	N	Mean	SD	Min	Q_1	Median	Q_3	Max	0–61 days	62–122 days	123–365 days
$\log(PM_{2.5})$	69,693	1.88	0.568	−2.78	1.55	1.93	2.28	4.47	85 (15.3%)	402 (72.2%)	70 (12.6%)
$\log(PM_{10})$	11,383	2.60	0.633	0	2.20	2.64	3.00	6.16	137 (81.5%)	26 (15.5%)	5 (3.0%)
$\log(SO_4^{2-})$	9,160	−0.30	0.860	−10.82	−0.68	−0.24	0.19	2.61	130 (80.7%)	29 (18.0%)	2 (1.2%)
log(Si)	8,868	−3.48	1.351	−10.82	−4.07	−3.35	−2.73	1.31	132 (82.0%)	29 (18.0%)	0 (0%)

The monitoring sites are split into groups that were observed for 61 days (one-sixth of a year) or less, observed between 61 and 122 days (one-third of a year), or observed between 123 and 365 days

Summary Stats of the Results

Table 3. Summary statistics for daily cross-validated RMSE values across all monitoring sites

Pollutant	Model	T	Mean	SD	Min	Q_1	Median	Q_3	Max
$PM_{2.5}$	PR	365	2.339	0.912	1.076	1.719	2.145	2.740	7.166
	ST	365	2.117	0.838	0.951	1.572	1.951	2.43	6.2
	UK	365	2.329	0.844	1.021	1.776	2.187	2.588	6.249
PM_{10}	PR	325	11.768	8.463	1.641	6.380	9.234	15.029	59.477
	ST	325	11.211	8.162	1.072	6.354	8.974	13.088	57.135
	UK	325	15.380	8.248	2.215	10.239	13.127	18.169	61.267
SO_4^{2-}	PR	173	0.420	0.279	0.069	0.264	0.374	0.519	3.007
	ST	173	0.464	0.269	0.102	0.316	0.420	0.523	2.760
	UK	124	0.712	0.287	0.163	0.533	0.676	0.845	1.942
Si	PR	122	0.079	0.085	0.012	0.030	0.049	0.087	0.441
	ST	122	0.084	0.106	0.010	0.026	0.046	0.087	0.502
	UK	122	0.076	0.088	0.009	0.026	0.044	0.080	0.464

We fit penalized regression (PR) per Eq. (3) using daily average temperature values and some number of TPRS basis functions (see Sect. 4.5) as predictors. For $PM_{2.5}$ we also added logged CMAQ values as predictors