Nonparametric regression of spatio-temporal data using infinite-dimensional covariates (IMS-APRM Conference 2026)

Soudeep Deb

Indian Institute of Management Bangalore
Bannerghatta Road, Bengaluru 560069, India.
Email: soudeep@iimb.ac.in.

June 15, 2026

Outline

Acknowledgments

Subhrajyoty Roy

Postdoc fellow, Wash Univ St.Louis

Sayar Karmakar

Asst Prof, U FLorida.

Rishideep Roy

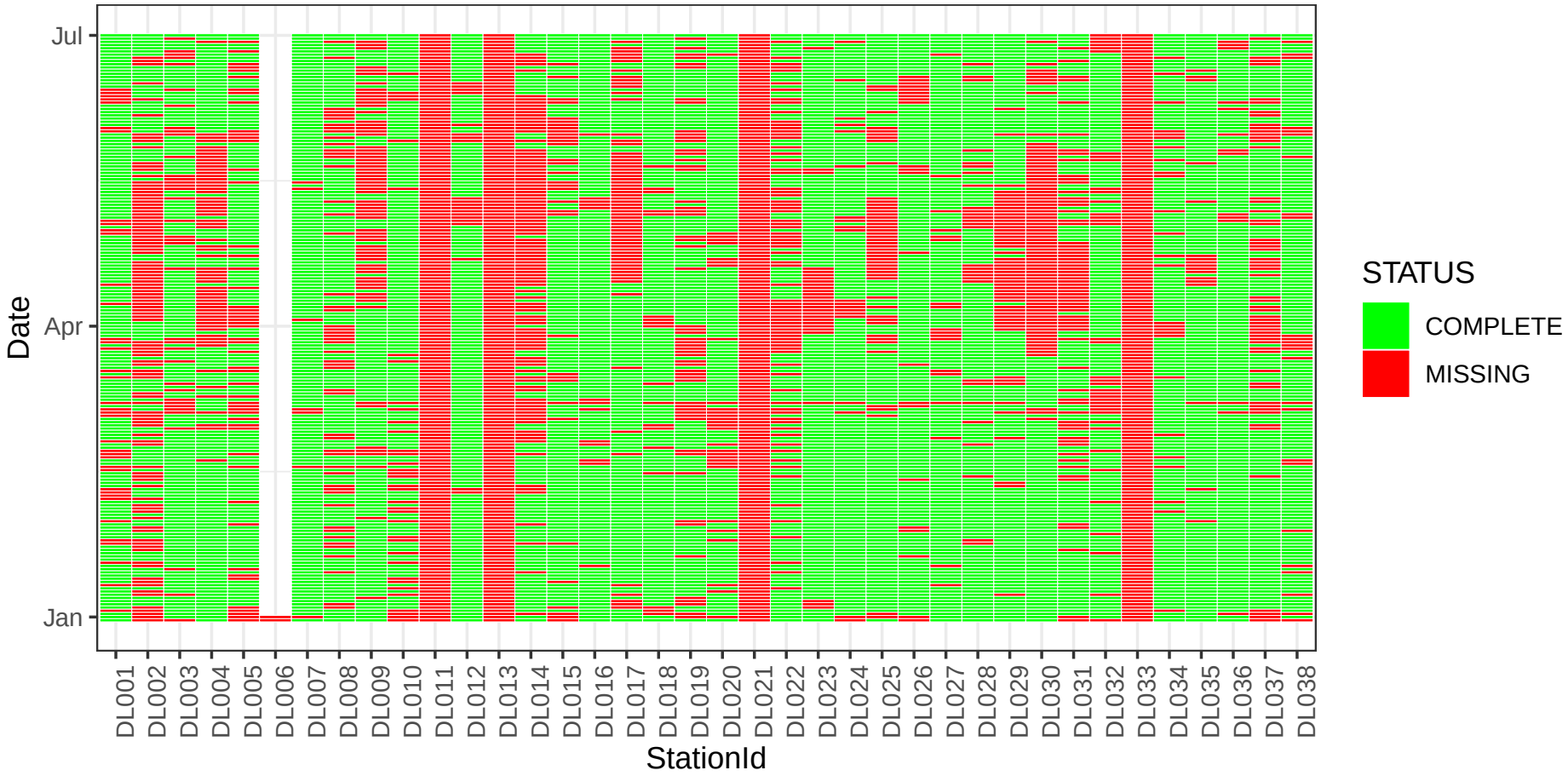
Lecturer, Univ of Essex.

Outline

Motivation: Spatio-temporal Data

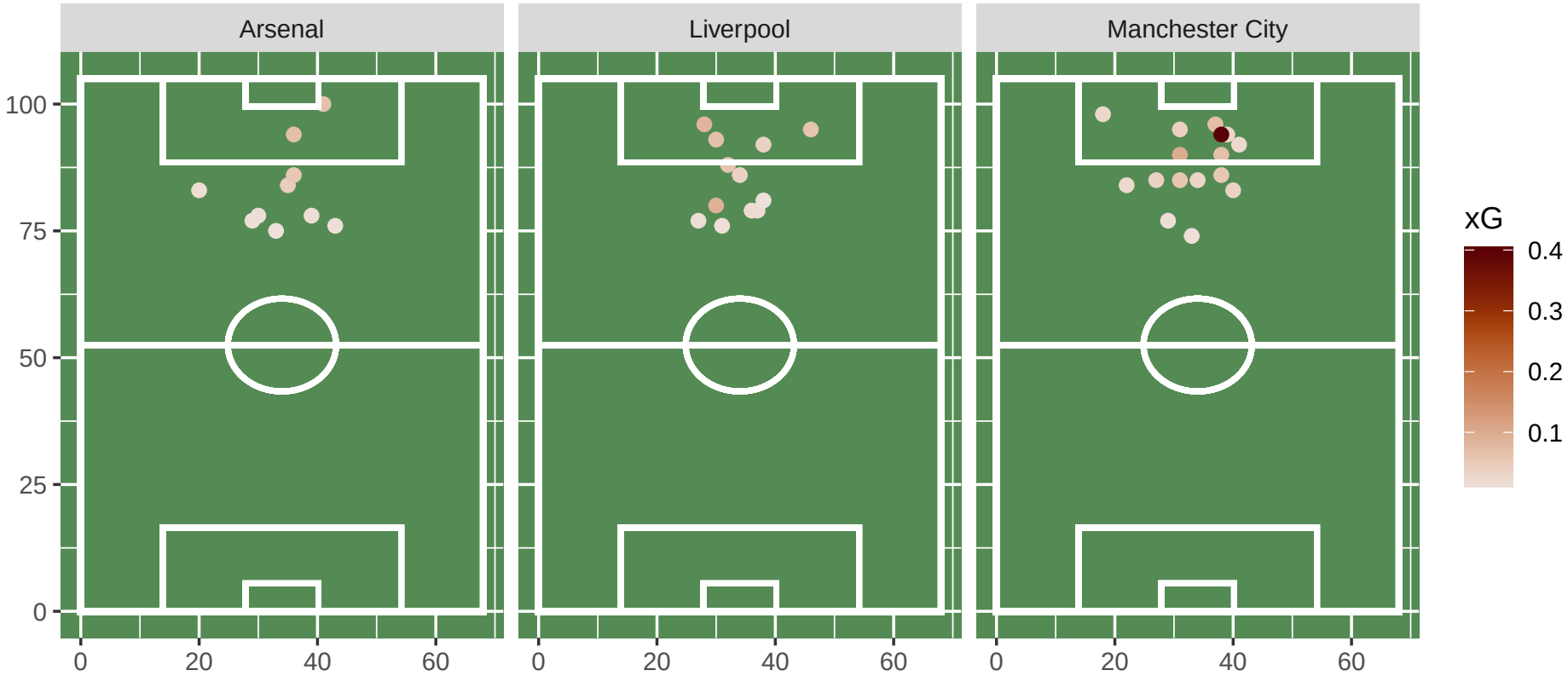
- Many modern data sets are indexed by
  - time $t \in T$,
  - space $s \in S$.
- Observations often come from
  - environmental monitoring,
  - seismology,
  - air pollution and climate studies,
  - epidemiology,
  - sports.
- Covariates can be high-dimensional time series or functional objects.

Motivation: Practical Example I



A heatplot indicating the missingness of the $\mathrm{PM}2.5$ observations across locations (columns) and timepoints (rows).

Motivation: Practical Example II



Position and expected goals (xG) of shots taken against Chelsea during the league 2014-15 by the players of Arsenal, Liverpool, and Manchester City.

Challenges

- **Irregular sampling**
  - Time points need not be equally spaced.
  - Spatial locations differ from time to time.
- **High or infinite-dimensional covariates**
  - Functional covariates or long historical vectors.
- **Complex dependence**
  - Temporal dependence in covariates.
  - Spatial correlation in errors.
- **Nonlinearity**
  - Relationship between response and covariates may be nonlinear.

Existing literature: what is covered and what is missing

- In nonparametric statistics, most existing methods (e.g., Deb et al., 2025) mainly focus on finite-dimensional covariates and regular grids.
- Functional regression with infinite-dimensional covariates in iid setup (Ferraty & View, 2004; Xiang et al., 2013; Omar & Wang, 2019).
- Recent work allows temporal dependence, but assumes regular designs (e.g., Li & Yang, 2023).
- **Closest related work:** Hong & Linton (2020) considers infinite-dimensional covariates, but in a time series setup rather than a fully spatiotemporal irregular-design setting.

Our contribution

We develop nonparametric inference for

$$Y_t(s) = \mu(\mathbf{X}_t, s) + \sigma(\mathbf{X}_t, s)\epsilon_t(s).$$

Main contributions

- **Model:** A spatiotemporal regression framework with irregular timepoints, time-varying spatial locations, and infinite-dimensional covariates.
- **Theory:** Kernel-based estimation of μ and σ using a spatial basis expansion, with relevant asymptotic theory.
- **Application:** Two domains – air pollution, and soccer analytics.

Outline

Data structure

- Response: $Y_t(s)$ observed at time t and location s.
- Covariate: $\boldsymbol{X}_t$ is a second-order stationary process in an affine subspace of a weighted L^2 Banach space:

$$\chi \subset \boldsymbol{D}\mathcal{L}^2 = \left\{ (z_1, z_2, \dots) \in \mathbb{R}^\infty : \sum_{i=1}^{\infty} \zeta_i^{-2} z_i^2 < \infty \right\}.$$

- Time points: $0 = t_0 < t_1 < \cdots < t_n = 1$.
- At each time t_i, observe locations

$$s_{t_i,1}, \dots, s_{t_i,n_i} \in \mathcal{S}.$$

Model framework

$$Y_t(s) = \mu(\mathbf{X}_t, s) + \sigma(\mathbf{X}_t, s)\epsilon_t(s), \qquad s \in \mathcal{S},\ t \in \mathcal{T}.$$

- $\mathbf{X}_t \in \chi \subseteq \mathbb{R}^\infty$: temporal-level, possibly infinite-dimensional covariate.
- $\mu(\mathbf{X}_t, s)$: conditional mean surface.
- $\sigma(\mathbf{X}_t, s)$: conditional volatility surface.
- Observed data:

$$\mathcal{D} = \{(t_i, s_{t_i j}, \mathbf{X}_{t_i}, Y_{t_i}(s_{t_i j}))\}.$$

Key feature

Timepoints are irregular, and spatial locations may change across time.

Spatial basis reduction

Assume

$$\mu(\mathbf{x}, s) = \sum_{k=1}^{\infty} \mu_k(\mathbf{x}) b_k(s), \qquad \sigma(\mathbf{x}, s) = \sum_{k=1}^{\infty} \sigma_k(\mathbf{x}) b_k(s).$$

Projecting the response onto basis function b_k gives

$$Y_{tk}^* = \int_{\mathcal{S}} Y_t(s) b_k(s)\, ds = \mu_k(\mathbf{X}_t) + \eta_{tk}.$$

Main idea

The spatiotemporal problem reduces to estimating scalar regression components $\mu_k(\mathbf{x})$ with infinite-dimensional covariates.

Estimator

For each basis coefficient,

$$\widehat{\mu}_k(\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left(\|\mathbf{H}_n^{-1}(\mathbf{x} - \mathbf{X}_{t_i})\|\right) \widehat{Y}^*_{t_i k}}{\sum_{i=1}^{n} K\left(\|\mathbf{H}_n^{-1}(\mathbf{x} - \mathbf{X}_{t_i})\|\right)}.$$

- Kernel smoothing is done in the covariate space χ.
- $\widehat{Y}^*_{t_i k}$ approximates $\int_{\mathcal{S}} Y_{t_i}(s) b_k(s)\, ds$.
- Because locations are irregular, $\widehat{Y}^*_{t_i k}$ is computed using a modified Monte Carlo grid procedure.

Modified Monte Carlo aggregation

Goal

Approximate the spatial integral using irregular locations $\{s_{t_i j}\}_{j=1}^{n_i}$ without over-weighting dense spatial clusters.

1. Compute the effective spatial resolution

$$\epsilon_{t_i}^* = \inf\left\{\epsilon > 0 : \mathcal{S} \subseteq \bigcup_{j=1}^{n_i} B(s_{t_i j}, \epsilon)\right\}.$$

2. Partition $\mathcal{S}$ into grid cells $H_1, \ldots, H_{r_i}$ with diameter $\epsilon_{t_i}^*$.
3. Pick one representative observation per cell:

$$s_{t_i, j_l} = \arg\min_j \|s_l^c - s_{t_i j}\|, \qquad l = 1, \ldots, r_i.$$

4. Average over representatives:

$$\widehat{Y}_{t_i k}^* = \frac{1}{r_i} \sum_{l=1}^{r_i} Y_{t_i}(s_{t_i, j_l}) b_k(s_{t_i, j_l}).$$

Key result on spatial aggregation

Proposition 2

For fixed t_i and k,

$$\mathbb{E}\left|\widehat{Y}^*_{t_i k} - Y^*_{t_i k}\right| = O\left(\delta_k(\epsilon^*_{t_i})\right).$$

Hence, if the effective spatial resolution satisfies

$$\epsilon^*_{t_i} \to 0,$$

then the modified Monte Carlo estimate is consistent.

Outline

Assumptions: high-level view

- χ is equipped with a scaled L^2 metric, allowing infinite-dimensional $\mathbf{X}_t$, which satisfies polynomial moment contraction (PMC):

$$\theta_2(m) = O(m^{-\tau}), \qquad \tau > 2.$$

- Errors have spatial covariance

$$\mathrm{Cov}(\epsilon_t(s), \epsilon_t(s')) = \rho(s, s').$$

- Type-I bounded-support kernel and bandwidth satisfying small-ball probability conditions.

PMC vs Mixing Conditions

- PMC is weaker than classic strong mixing assumptions.
- Allows for:
  - some non-mixing processes,
  - more flexible temporal structures.
- Also weaker than geometric moment contraction (GMC).
- Still strong enough to obtain
  - laws of large numbers,
  - central limit theorems for sums.

Consistency

Theorem 1

For fixed k and $\mathbf{x} \in \chi$,

$$\widehat{\mu}_k(\mathbf{x}) - \mu_k(\mathbf{x}) = o_{\mathbb{P}}(1).$$

Theorem 2

For large enough K_ϵ,

$$\sum_{k=1}^{K_\epsilon} \widehat{\mu}_k(\mathbf{x}) b_k(s) - \mu(\mathbf{x}, s) = O_{\mathbb{P}}(\epsilon).$$

If $b_\infty(s)$ is continuous, the result holds uniformly over $s \in \mathcal{S}$.

Asymptotic normality

Theorem 3

For fixed k and $\mathbf{x}$,

$$\frac{\sqrt{n\phi_{\mathbf{x}}(h_n\lambda)}\,\xi_1}{\sqrt{\xi_2\sigma_{kk}(\mathbf{x})}}\left[\widehat{\mu}_k(\mathbf{x}) - \mu_k(\mathbf{x}) - \widetilde{b}_{nk}(\mathbf{x})\right] \Rightarrow N(0,1).$$

- Effective sample size: $n\phi_{\mathbf{x}}(h_n\lambda)$.
- Bias term: $\widetilde{b}_{nk}(\mathbf{x}) = O(h_n)$.
- Bias correction:

$$\widehat{\mu}_k^* = 2\widehat{\mu}_k^{(h_n)} - \widehat{\mu}_k^{(2h_n)}.$$

Construction of confidence intervals

Corollary

For fixed k and $\mathbf{x}$,

$$\widehat{\mu}_k(\mathbf{x}) - \widetilde{b}_{nk}(\mathbf{x}) \pm \frac{z_{1-\alpha/2}\sqrt{\xi_2 \widehat{\sigma}_{kk}(\mathbf{x})}}{\xi_1 \sqrt{n\phi_{\mathbf{x}}(h_n\lambda)}}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for $\mu_k(\mathbf{x})$.

The same conclusion holds with the bias-corrected estimator $\widehat{\mu}_k^*(\mathbf{x})$.

Outline

Why simultaneous confidence bands?

Pointwise intervals answer:

"Is $\mu(\mathbf{x}, s)$ well-estimated at one fixed $\mathbf{x}$?"

Simultaneous bands answer:

"Is $\mu(\mathbf{x}, s)$ well-estimated over many covariate values $\mathbf{x} \in \chi_n$?"

Goal

Construct bands such that, for fixed $s \in \mathcal{S}$, with high probability

$$\mu(\mathbf{x}, s) \in \text{band}(\mathbf{x}, s) \quad \text{for all } \mathbf{x} \in \chi_n.$$

Band construction

Use the truncated, bias-corrected estimator

$$\widehat{\mu}^*_{1:K}(\mathbf{x}, s) = \sum_{k=1}^{K} \widehat{\mu}^*_k(\mathbf{x}) b_k(s).$$

Estimate its local variance by

$$Q_K(\mathbf{x}, s) = \mathbf{b}_{1:K}(s)^\top \widehat{\mathbf{\Sigma}}_{1:K,1:K}(\mathbf{x}) \mathbf{b}_{1:K}(s).$$

A simultaneous band has the form

$$\widehat{\mu}^*_{1:K}(\mathbf{x}, s) \pm \frac{\sqrt{\xi_2 Q_K(\mathbf{x}, s)}}{\xi_1 \sqrt{n \phi_{\mathbf{x}}(h_n \lambda)}} B_{m_n}(z), \qquad \mathbf{x} \in \chi_n.$$

Theorem 5: simultaneous validity

Let $m_n = |\chi_n|$. Under the assumptions of the paper,

$$\mathbb{P}\left(\sup_{\mathbf{x} \in \chi_n} \frac{\xi_1 \sqrt{n \phi_{\mathbf{x}}(h_n \lambda)}}{\sqrt{\xi_2 Q_{K_\epsilon}(\mathbf{x}, s)}} \left| \widehat{\mu}^*_{1:K_\epsilon}(\mathbf{x}, s) - \mu(\mathbf{x}, s) \right| < B_{m_n}(z) \right) \gtrsim e^{-2e^{-z}}.$$

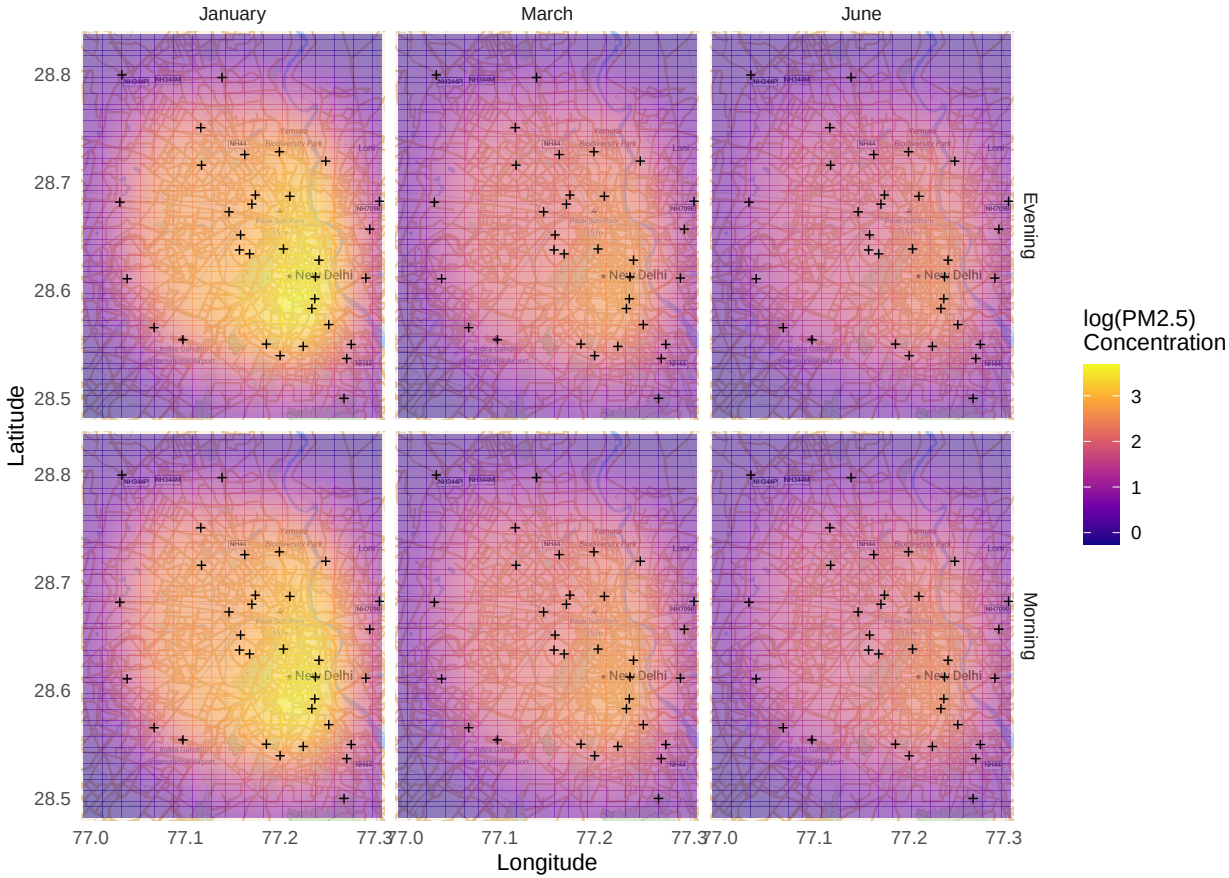
- $B_{m_n}(z)$ is the Gumbel threshold for the maximum over χ_n.
- Larger m_n gives wider bands.
- K_ϵ controls truncation error from the basis expansion.

Outline

Air pollution in Delhi

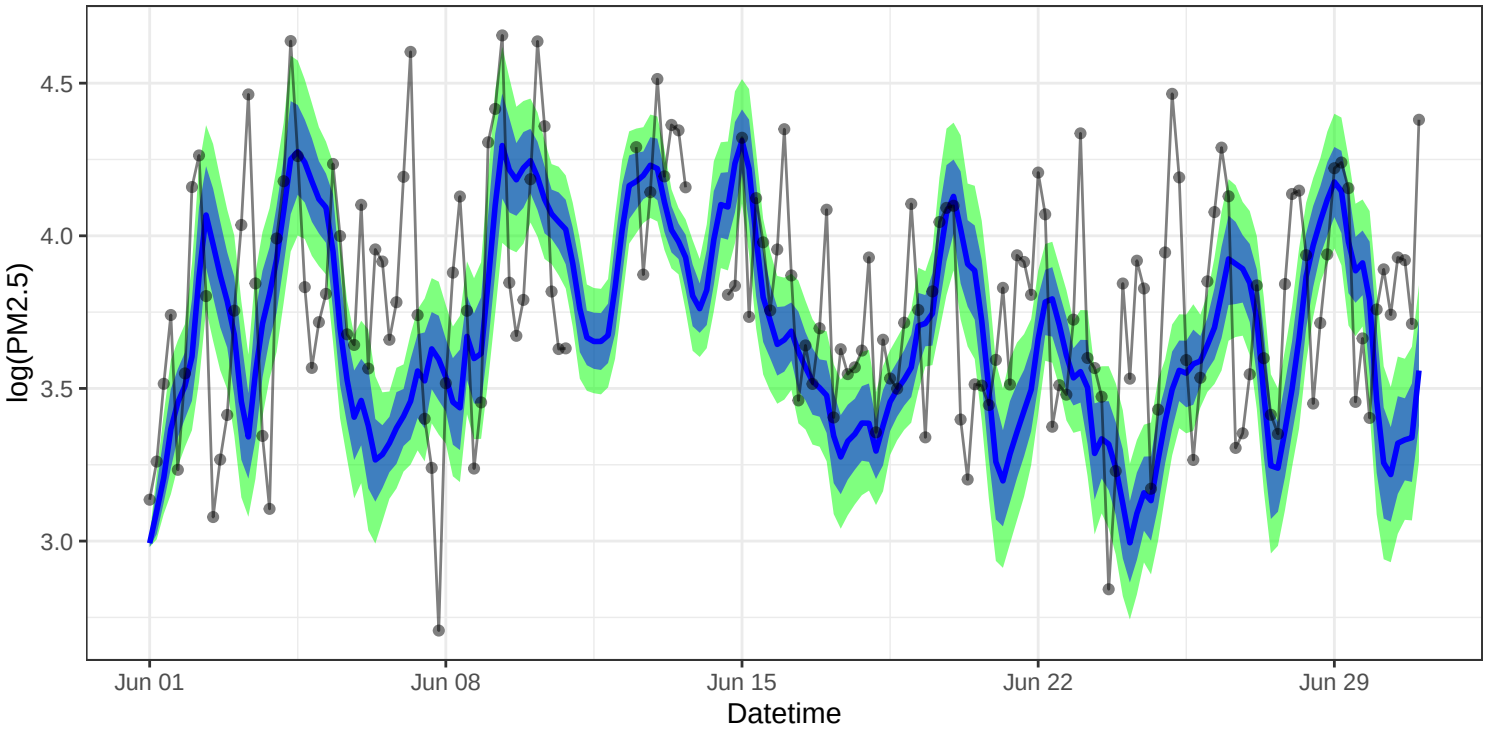
- Real data example: air pollution in the Delhi region.
- Features:
  - irregular spatial locations of 38 monitoring stations,
  - missing observations due to sensor downtime,
  - multiple covariates such as other pollutants and weather variables.
- Fits naturally into the proposed spatio-temporal regression framework.

Estimated mean surfaces



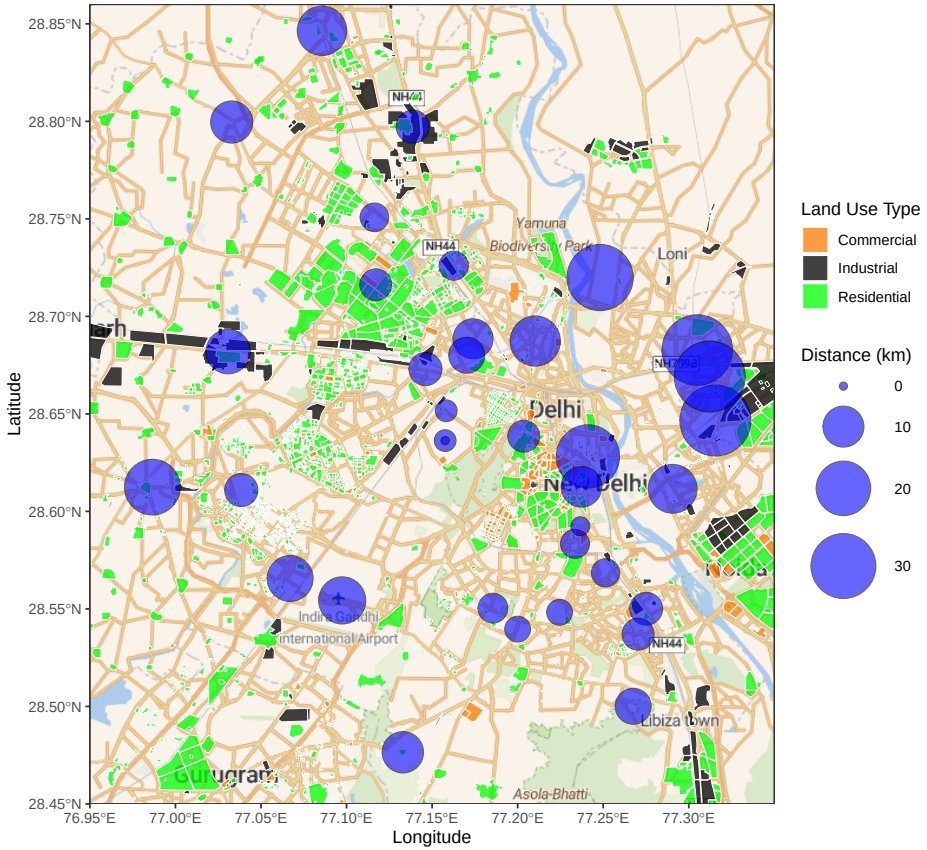
Estimated mean levels of PM2.5-concentration across the entire region for three different months and during two specific times of the day (10 AM and 8 PM).

Prediction of PM2.5



The predicted mean response (log of PM2.5 measurements) for one station. The blue band depicts the pointwise confidence interval, and the green band depicts the simultaneous confidence band across all timepoints in June 2020.

Best choices of spatial covariates



The best choice of the number of spatial covariates (in km) in predicting the PM2.5 measurements for June 2020 in every station at Delhi, along with landuse data.

Outline

Summary

- We study nonparametric regression for spatiotemporal data with irregular design and infinite-dimensional covariates.
- Focus is on conditional mean and volatility surfaces:

$$Y_t(s) = \mu(\mathbf{X}_t, s) + \sigma(\mathbf{X}_t, s)\epsilon_t(s).$$

- Spatial basis projection reduces the problem to componentwise functional regression.
- Theory provides consistency, CLT, variance estimation, and simultaneous confidence bands under PMC dependence.
- Applications to $\mathrm{PM}2.5$ prediction and soccer xG surfaces demonstrate practical relevance.

Practical choices to keep in mind

- Basis $\{b_k\}$:
  - choose according to geometry of S and data.
- Bandwidth h_n:
  - cross-validation or plug-in rules,
  - possibly different for different k.
- Truncation level K_n:
  - based on cumulative variance explained,
  - or information criteria.

Computational considerations

- Projection step:
  - requires evaluating $b_k(s_{t,j})$ for many (t, j, k).
- Kernel regression:
  - potentially expensive in high dimension,
  - use dimension reduction or fast nearest-neighbour search when needed.
- Parallelization:
  - estimation for different k can be parallelized.

Limitations

- Requires sufficient spatial coverage over time.
- Choice of basis and tuning parameters can influence results.
- Infinite-dimensional regression is inherently data-hungry.
- Theoretical results rely on technical conditions (PMC, smoothness, small-ball probabilities) that may be hard to verify in practice.

References

Roy, S., **Deb, S.**, Karmakar, S., & Roy, R. (2026). Nonparametric regression of spatio-temporal data using infinite-dimensional covariates. arXiv preprint arXiv:2604.01593.

- **Deb, S.**, Neves, C., Roy, S. (2025+) Nonparametric quantile regression for spatio-temporal processes. `https://arxiv.org/abs/2405.13783`.
- Ferraty, F., & Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. Nonparametric Statistics, 16(1-2), 111-125.
- Hong, S. Y., & Linton, O. (2020). Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff. Journal of Econometrics, 219(2), 389-424.
- Li, J., & Yang, L. (2023). Statistical inference for functional time series. Statistica Sinica, 33(1), 519-549.
- Omar, K. M. T., & Wang, B. (2019). Nonparametric regression method with functional covariates and multivariate response. Communications in Statistics-Theory and Methods, 48(2), 368-380.
- Xiang, D., Qiu, P., & Pu, X. (2013). Nonparametric regression analysis of multivariate longitudinal data. Statistica Sinica, 769-789.

IIMB
तेजस्वि नावधीतमस्तु

Thank you for listening.

Contact: soudeep@iimb.ac.in

Personal Webpage

Pre-print of the paper