Australasian Applied Statistics Conference 2024

Cokrig-and-Regress for Spatially Misaligned Environmental Data

---

Zhi Yang Tho

Joint work with Dr. Francis Hui, Prof. Alan Welsh, Dr. Tao Zou

Australian National University

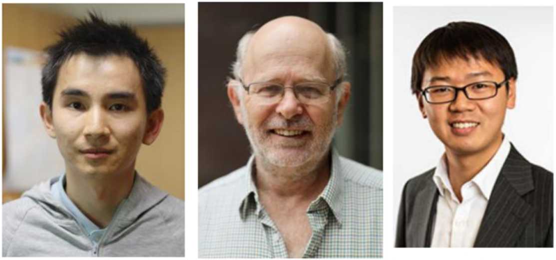

Figure: Left to right: Dr. Francis Hui, Prof. Alan Welsh, Dr. Tao Zou

Overview

Motivation

Motivation

Spatial misalignment problem between $PM_{2.5}$ concentration (response) and meteorological variables (covariates) such as temperature and precipitation.

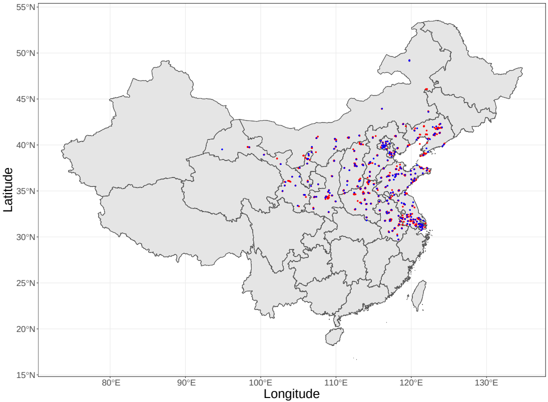

Figure: Map of China with geographic locations of pollution monitoring stations and meteorological stations.

Motivation (Cont'd)

Existing methods to construct spatially aligned datasets:

- **Nearest-neighbor interpolation** (Jhun et al., 2015; Greenstone et al., 2022);
- **Kriging** for each meteorological covariate separately and treating predicted covariates as **fixed** (Reich et al., 2011; Liu et al., 2020);
- **Krige-and-regress (KNR)** method that accounts for additional variability of the predicted covariate (Madsen et al., 2008; Szpiro et al., 2011; Pouliot, 2023)

However, KNR method only allows for

- A single misaligned meteorological covariate;
- Simple linear pollution-meteorological relationship.

Model Set-Up

Model Set-Up

Denote the pollution stations and meteorological stations as $S = \{s_1, \cdots, s_N\}$ and $\tilde{S} = \{\tilde{s}_1, \cdots, \tilde{s}_M\}$, respectively.

Let $\boldsymbol{y} = (y_1, \cdots, y_N)^\top = (y(s_1), \cdots, y(s_N))^\top$,
$\boldsymbol{x}_k = (x_{1k}, \cdots, x_{Nk})^\top = (x_k(s_1), \cdots, x_k(s_N))^\top$ and
$\tilde{\boldsymbol{x}}_k = (\tilde{x}_{1k}, \cdots, \tilde{x}_{Mk})^\top = (x_k(\tilde{s}_1), \cdots, x_k(\tilde{s}_M))^\top$,

$$y_i = \beta_0 + \sum_{k=1}^{K} \boldsymbol{f}_k(x_{ik})^\top \boldsymbol{\beta}_k + \rho_i + \epsilon_i, \text{ for } i = 1, \cdots, N, \tag{1}$$

where **only $\boldsymbol{y}$ and $\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_K$ are observed but not $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K$.**

Model Set-Up (Cont'd)

Let $\boldsymbol{x} = (\tilde{\boldsymbol{x}}_1^\top, \boldsymbol{x}_1^\top, \cdots, \tilde{\boldsymbol{x}}_K^\top, \boldsymbol{x}_K^\top)^\top$ denote the stacked $K(M+N)$-vector of K covariates at locations in both $\tilde{S}$ and S. Assume

$$\begin{aligned} &\boldsymbol{x} \sim N(\boldsymbol{\mu} \otimes 1_{M+N}, \boldsymbol{\Sigma}), \\ &\boldsymbol{\Sigma} = \mathrm{Bdiag}(\boldsymbol{L}_1, \cdots, \boldsymbol{L}_K)(\boldsymbol{R} \otimes \boldsymbol{I}_{M+N})\mathrm{Bdiag}(\boldsymbol{L}_1, \cdots, \boldsymbol{L}_K)^\top, \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)^\top$, $\boldsymbol{R}$ is a $K \times K$ cross-correlation matrix for the K meteorological covariates, $\boldsymbol{L}_k$ is the lower Cholesky factor of $\boldsymbol{\Sigma}_k$ which is the Matern spatial covariance matrix for the k-th covariate.

Cokrig-and-Regress (CNR)

CNR Step 1

Estimate the parameters of the **joint** distribution of the meteorological covariates, based on the **observed misaligned meteorological data** $\boldsymbol{x}_{\tilde{S}} = (\tilde{\boldsymbol{x}}_1^\top, \cdots, \tilde{\boldsymbol{x}}_K^\top)^\top$.

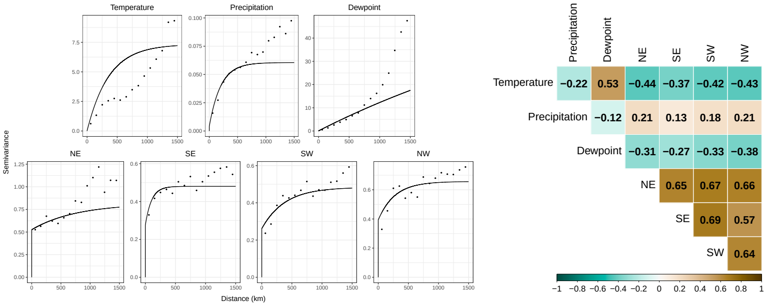

Figure: Estimated marginal Matérn covariance models (left), estimated cross-correlation matrix (right) from CNR Step 1.

CNR Step 2

Predict the unobserved $\boldsymbol{x}_S = (\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_K^\top)^\top$ by **cokriging**, based on the observed $\boldsymbol{x}_{\tilde{S}}$ and estimated parameters for the joint distribution of the meteorological covariates from CNR Step 1.

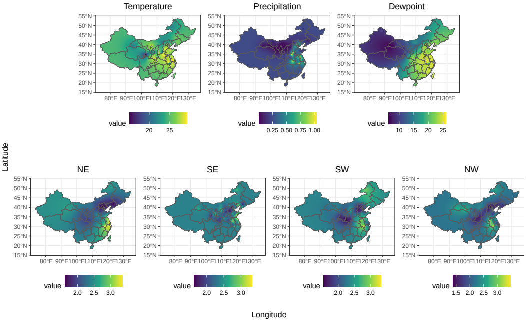

Figure: Spatial maps for the cokriging predictor of each meteorological covariate in CNR Step 2.

CNR Step 3

Replace the unobserved meteorological covariates with their cokriging prediction and **fit the spatial linear mixed model** (1) i.e., $y_i = \beta_0 + \sum_{k=1}^{K} \boldsymbol{f}_k(\hat{x}_{ik})^\top \boldsymbol{\beta}_k + \rho_i + \epsilon_i$.

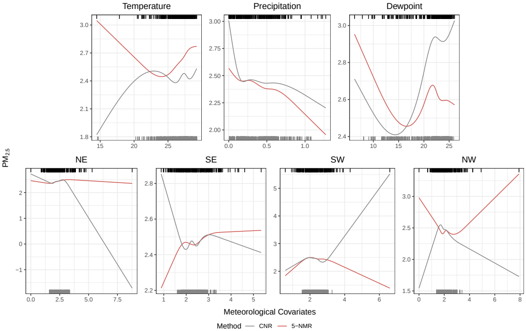

Figure: Estimated conditional smoothers for the seven meteorological covariates based on CNR and 5-NMR using natural cubic splines.

Uncertainty Quantification

Bias-Correction and Uncertainty Quantification

To estimate variance of the CNR estimates or construct confidence intervals for $\boldsymbol{\beta}_k$:

(i) Perform **preliminary parametric bootstrap** to bias-correct CNR spatial covariance parameter estimates (which was shown to be biased).

(ii) Based on the bias-corrected CNR spatial covariance parameter estimates, perform **secondary parametric bootstrap** to obtain bootstrap samples of CNR estimates for $\boldsymbol{\beta}_k$.

Bootstrap CI compared to Naive Variance Estimator

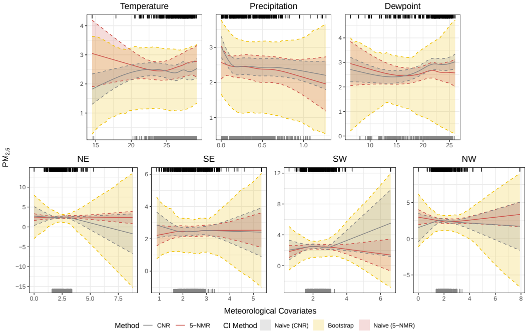

Figure: Estimated conditional smoothers for the seven meteorological covariates based on CNR (grey solid lines) and 5-NMR (red solid lines) using natural cubic splines. Also shown are 95% confidence bands based on the naive variance estimator for CNR (shaded regions between dashed lines in grey), bootstrap percentile confidence bands for CNR (shaded regions between dashed lines in yellow) and the naive variance estimator for 5-NMR (shaded regions between dashed lines in red).

Simulation Studies

Simulation Settings

- $\tilde{S}$ and S are locations of $M = 243$ meteorological stations and $N = 796$ pollution monitoring stations, respectively.
- $K = 5$ covariates with $\boldsymbol{f}_k(x_{ik})^\top \boldsymbol{\beta}_k = x_{ik}\beta_k$ for $k = 1, \cdots, 5$, and $\boldsymbol{\beta} = (2, 1, 0.5, 1, 0.5, 1)^\top$.
- A total of 400 simulated datasets.

Simulation Settings

- Bias and RMSE for β_k:
  - CNR;
  - 5-nearest-matching-and-regress (5-NMR).

- Ratio of average estimated standard error to empirical standard deviation (ASE/ESD), empirical coverage probability of 95% CIs for β_k:
  - Naive variance estimator (Naive) by ignoring the cokriging prediction uncertainty;
  - Proposed bootstrap approach.

Simulation Results

Table: Bias, RMSE, ASE/ESD and empirical coverage of β_k for $k = 1, \cdots, 5$.

		Point Estimation				
	Method	$\beta_1 = 1$	$\beta_2 = 0.5$	$\beta_3 = 1$	$\beta_4 = 0.5$	$\beta_5 = 1$
Bias	CNR	**−0.0068**	**0.0021**	**−0.0070**	**0.0046**	**−0.0111**
	5-NMR	−0.1529	−0.0752	−0.1591	−0.0665	−0.1580
RMSE	CNR	**0.1733**	**0.1670**	**0.2068**	**0.2137**	**0.1948**
	5-NMR	0.2516	0.2113	0.2705	0.2535	0.2872
		ASE/ESD				
Method	Inference Method	$\beta_1 = 1$	$\beta_2 = 0.5$	$\beta_3 = 1$	$\beta_4 = 0.5$	$\beta_5 = 1$
CNR	Naive	0.7495	0.7749	0.7236	0.7811	0.7750
	Bootstrap	**1.0896**	**1.0789**	**1.0300**	**1.0628**	**1.0928**
5-NMR	Naive	0.8096	0.8159	0.8571	0.8541	0.7827
		Empirical Coverage				
Method	Inference Method	$\beta_1 = 1$	$\beta_2 = 0.5$	$\beta_3 = 1$	$\beta_4 = 0.5$	$\beta_5 = 1$
CNR	Naive	0.8400	0.8650	0.8275	0.8875	0.8750
	Bootstrap	**0.9525**	**0.9625**	**0.9600**	**0.9600**	**0.9550**
5-NMR	Naive	0.7950	0.8500	0.8175	0.8750	0.8125

References I

Greenstone, M., He, G., Jia, R., and Liu, T. (2022). Can technology solve the principal-agent problem? Evidence from China's war on air pollution. *American Economic Review: Insights*, 4:54–70.

Jhun, I., Coull, B. A., Schwartz, J., Hubbell, B., and Koutrakis, P. (2015). The impact of weather changes on air quality and health in the United States in 1994–2012. *Environmental Research Letters*, 10:084009.

Liu, Y., Zhou, Y., and Lu, J. (2020). Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. *Scientific Reports*, 10:14518.

Madsen, L., Ruppert, D., and Altman, N. S. (2008). Regression with spatially misaligned data. *Environmetrics*, 19:453–467.

References II

Pouliot, G. A. (2023). Spatial econometrics for misaligned data. *Journal of Econometrics*, 232:168–190.

Reich, B. J., Eidsvik, J., Guindani, M., Nail, A. J., and Schmidt, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *The Annals of Applied Statistics*, 5:2425–2447.

Szpiro, A. A., Sheppard, L., and Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12:610–623.

R Codes

R codes available at https://github.com/Zy1225/CNR.

THE
END

THANKS!