Topics in spatial causal inference

Brian Reich, NC State

SRCOS, 2022, Jekyll Island

Collaborators: Can Cui, Andrew Giffin, Yawen Guan, Garritt Page, Ana Rappold, Massimo Ventrucci, Lili Wu, Shu Yang
Adjusting for unmeasured spatial confounders

- Environmental and epidemiological data are often observational and spatially correlated.

- A fundamental task is to estimate the effect of a treatment variable on a response variable.

- The key assumption that there are no missing confounders is generally impossible to verify.

- However, it may be possible to remove the effects of unmeasured confounders that are smooth spatial functions.

- We examine several methods to adjust for missing spatial confounding variables.
Motivating Example

Wu et al (2020)\(^1\) found that an increase of 1 $\mu g/m^3$ in ambient fine particulate matter (PM\(_{2.5}\)) is associated with a 15% increase in the COVID-19 mortality rate.

The response variable $Y$ is the cumulative COVID-19 mortality counts through May 12, 2020 for US counties.

The exposure variable $X$ is county-level average exposure to PM\(_{2.5}\) for 2000-2016.

---

Introduction

Exposure variable, PM$_{2.5}$ concentration

![Map showing mean PM$_{2.5}$ concentrations across the United States.](image)

- Mean PM$_{2.5}$ values range from 4 to 12.
Outcome variable, log mortality rate

Log(deaths/population)

Log(deaths/population)
Other applications

- Wildland fires $\rightarrow$ increased air pollution?

- Using an app $\rightarrow$ reduced air pollution exposure?

- Fishing regulations $\rightarrow$ ecological diversity?
Standard spatial linear regression model

- The response at location $s \in \mathbb{R}^2$ is $Y(s)$
- The treatment variable of interest is $A(s)$
- The model is $Y(s) = \beta_0 + A(s)\beta_A + \theta(s) + \varepsilon(s)$
- Our goal is to estimate $\beta_A$
- The residuals have two components: $\theta(s)$ is spatially correlated and $\varepsilon(s) \sim iid \text{Normal}(0, \tau^2)$
Standard spatial linear regression model

- The spatial term $\theta(s)$ is a Gaussian process with $E\{\theta(s)\} = 0$ and $\text{Var}\{\theta(s)\} = \sigma^2$

- The correlation decays with distance

  $\text{Cor}\{\theta(s_i), \theta(s_j)\} = \rho(d_{ij})$

  where $d_{ij}$ is the distance between $s_1$ and $s_j$

- For example, the exponential correlation is

  $\rho(d) = \exp(-d/\phi)$
Potential confounding variables

▶ Variable $X(s)$ is a **confounder** if it is correlated with $A(s)$ and $Y(s)$

▶ It is a **spatial confounder** if has spatial correlation

▶ It is an **missing spatial confounder** it is unknown or unobserved
Missing confounders and residual correlation

- Say the truth is \( Y(s) = \beta_0 + A(s)\beta_A + X(s)\beta_X + \epsilon(s) \)

- If \( X \) is observed, we can use non-spatial regression

- This would be fine even if \( X \) and \( A \) are correlated

- If \( X \) is unobserved, it would contribute to spatial component of the error term \( X(s) + \epsilon(s) \)

- We might fit the spatial model

\[
Y(s) = \beta_0 + A(s)\beta_A + \theta(s) + \epsilon(s)
\]

and hope that \( \theta \) accounts for \( X \)
Simulation with a missing spatial confounder

- We generate $X(s)$ and $A(s)$ as spatial processes

- The correlation between $X$ and $A$ ranges from 0 to 1

- The data are generated as

\[ Y(s) = A(s)\beta_A + X(s)\beta_X + \varepsilon(s) \]

where $\beta_A = \beta_X = 1$ and $\varepsilon(s) \overset{iid}{\sim} \text{Normal}(0, 0.1^2)$

- We fit non-spatial and spatial regression models that exclude $X$
**Coverage and root mean squared error for $\beta_A$**

- Spatial regression works well when $\text{Cor}\{X(s), A(s)\} = 0$
- However, even a small correlation causes huge problems
Possible remedies for a missing spatial confounder

1. Scientific knowledge

2. Propensity scores

3. Matching

4. Mean adjustments

---

Adding known confounders

- Wu et al include several potential confounder variables:
  - Climate variables such as temperature
  - SES variables including income and race
  - Access to ICU beds

- Call the known confounders \( W_1(s), \ldots, W_p(s) \)

- The fitted model is
  \[
  Y(s) = \beta_0 + \sum_{j=1}^{p} W_j(s)\alpha_j + A(s)\beta_A + \theta(s) + \epsilon(s)
  \]

- A logical extension to environmental application is to replace the \( W_j(s) \) with numerical model output
EXAMPLE: Adding mathematical/process models

- Fine particular matter (PM$_{2.5}$) is a criteria pollutant monitored by the EPA to protect human health.

- In the US, emissions of PM$_{2.5}$ from most sources are steadily declining except for forest fire smoke.

- How much PM$_{2.5}$ and health burden is causally-attributed to wildfires?

- This is difficult to answer directly because only total PM$_{2.5}$ (background + fire) can be measured.

- We$^2$ combine numerical models (CMAQ), observational data and causal inference.

---

CMAQ: The Community Multiscale Air Quality Modeling System

* plume rise, biogenic, lightning generated NO, sea salt, windblown dust, bidirectional exchange of ammonia
Annual average CMAQ (12km × 12km)

- [1.85, 3.07]
- (3.07, 4.4)
- (4.4, 6.79)
- (6.79, 30.4)
CMAQ run without fires
Difference between the runs (as % of total)

PM$_{2.5}$ (%)

- (23.5,91.8)
- (12,23.5)
- (7.43,12)
- [0.84,7.43]
PM$_{2.5}$ is measured every 3-6 days; this is the 2008-2012 average.
Time series plot for one site in Northern CA

PM$_{2.5}$ (µg/m$^3$)

- Observed, Total
- CMAQ, Total
- CMAQ, Background

Graph shows PM$_{2.5}$ levels from May to November 2008.
Data sources and potential outcomes notation

- Monitor data at location $s$ and day $t$: $Y_t(s)$
- CMAQ run without fires: $\hat{\theta}_t(s)$
- CMAQ run with fires: $\hat{\theta}_t(s) + \hat{\delta}_t(s)$
- The “treatment” is the regime
  - $A = 0$: world without wildland fires
  - $A = 1$: current world with wildland fires
- $Y_t(s, 0)$ and $Y_t(s, 1)$ are the potential PM$_{2.5}$ outcomes
- Model: $Y_t(s, 0) = \theta_t(s)$ and $Y_t(s, 1) = \theta_t(s) + \delta_t(s)$
- $\theta_t$ and $\delta_t$ are stochastic processes
Potential outcomes framework

▶ The causal effect is

$$\Delta(s) = E[Y_t(s, 1) - Y_t(s, 0)]$$

where the average is over the distribution of fires over the entire spatial domain

▶ **Challenge**: we never observe data under $A = 0$

▶ To address this we use

▶ CMAQ output
  ▶ Causal assumptions
Assumptions

We assume there exist

- $C_t(s) \in \{0, 1\}$ where $s$ is affected by smoke iff $C_t(s) = 1$
- Bias-correction functions $B_0$ and $B_1$

so that the following assumptions hold:

(A1) Consistency: $Y_t(s) = Y_t[s, C_t(s)]$

(A2) No unmeasured confounders given model output:

$$\theta_t(s) = B_0[\hat{\theta}_t(s)] + e_{1t}(s)$$
$$\delta_t(s) = B_1[\hat{\delta}_t(s)] + e_{2t}(s),$$

where $e_t(s) = [e_{1t}, e_{2t}]$ is independent of $A$ and $C$
Assumptions

Are these assumptions reasonable?

- (A1) assumes that we have some observations we are sure are not affected by fire smoke ... probably OK?

- (A2) assumes that the CMAQ modelers have included the important drivers of fine particulate matter .... maybe OK? Have we accounted for all feedback loops?
Causal interpretation

- Theory: We prove that under these assumptions the estimate effect has a causal interpretation.
- We specify parametric models for the bias correction functions $B_0$ and $B_1$ and the spatial process $e_t(s)$.
- Theory: We show that all parameters, including the correlation between counterfactuals, are identifiable.
- This serves as a basis for using a Bayesian approach to estimating $\Delta(s)$ and quantify its uncertainty.
Data and estimates for one site in CA

Graph showing PM$_{2.5}$ (µg/m$^3$) from May to November 2008 with the following line types:
- Observed, Total
- CMAQ, Total
- Estimated, Total
- CMAQ, Background
- Estimated, Background
Causal estimate, $\Delta(s)$, posterior mean

PM$_{2.5}$ ($\mu$g/m$^3$)

- [-0.072, 0.006]
- (0.006, 0.073]
- (0.073, 0.164]
- (0.164, 0.38]
- (0.38, 1.8]
- (1.8, 9.29]
Cumulative health burden by county
EXAMPLE: Propensity scores

- Smoke Sense is a cell phone app design by the US EPA to educate and engage citizens.

- The app provides users current and forecasted air quality and alerts of fires.

- Users record their smoke observations, health symptoms and preventative actions.

- Users also play educations games, earn badges, etc.
The smoke Sense app
Smoke Sense data and notation

- **Response**: $Y_{it}$ is the number of self-reported symptoms (asthma attack, chest pain, etc.) by user $i$ in week $t$

- **Treatment**: $A_{it}$ is a binary indicator that user $i$ took preventive measures (stay indoors, use mask, etc.)

- **Covariates**: baseline variables (age, gender, etc.) and time-varying variables (smoke exposure, visibility, etc.)

- We use $n = 1076$ users from 2019 and the number of observations ranges from 1-15

- Our objective is to determine if using the app leads to a reduction in symptoms
App-based platforms provide unprecedented opportunities to reach users, but pose challenges:

- **Self-selection**: Protective behaviors were left to the user and may depend on their characteristics

- **Informative missingness**: Participants are more likely to self-report when they experience smoke or symptoms

- **Spatial variation**: The causal effect of treatment may vary over the study’s large and diverse domain
Potential outcomes (PO)

- For simplicity, we\(^3\) drop the user subscript \((i)\) and assume regular observation times \(t \in \{1, 2, \ldots\}\)

- History is denoted by overbar, so \(\bar{A}_m\) and \(\bar{X}_m\) are the set of values before time \(m\)

- POs are denoted by \(\bar{a}_m\) is the set of treatments before time \(m\) and \(a_m\) is the set of treatment taken after time \(m < t\)

- The PO at time \(t\) is \(Y_t(\bar{a}_m, a_m)\)

- Our goal is to estimate the causal effect which is the mean difference between \(Y_t(\bar{a}_m, a_m)\) and \(Y_t(\bar{a}_m, 0)\)

- We do this within the structured nested mean model of Robins (1992)

Structured nested mean model

The causal effect $\gamma$ as a function of parameters $\theta$ is

$$E \{ Y_t(\bar{a}_m, a_m) - Y_t(\bar{a}_m, 0)|\bar{a}_m, \bar{x}_m \} = \gamma_{tm}(\bar{a}_m, \bar{x}_m; \theta) = \gamma_{tm}(\theta)$$

An example with covariate and lagged treatment effects is

$$\gamma_{tm}(\theta) = x_m \beta + \psi \exp \left\{ -\frac{(t - m - \mu)^2}{2\sigma^2} \right\} a_m$$

and $\theta = \{\beta, \psi, \mu, \sigma\}$

The treatment at time $m = t - \mu$ has the most impact

The magnitude of treatment at lag $\mu$ is $\psi$
Estimation

- Sequential randomization assumption: treatment at time $m$ is a function only of the propensity score
  \[ e_m = \text{Prob}(A_m = 1| \bar{A}_{m-1}, \bar{X}_m, \bar{Y}_{m-1}) \]

- To estimate the lag $m$ effect, we must remove effects for time $l \in \{m, \ldots, t-1\}$

- Let $R_{tm}(\theta) = Y_t - \sum_{l=m}^{t-1} \gamma_{tl}(\theta)$ and $\mu_{tm}(\theta) = E(R_{tm})$

- The weighted (by $w_{tm}$) estimating equation is
  \[ G(\theta) = \sum_{t=1}^{n_t} \sum_{m=1}^{t-1} w_{tm} \{ R_{tm}(\theta) - \mu_{tm}(\theta) \} (A_t - e_t) \]
Spatially-varying effects

- We extend this structured nested mean model to allow for spatially-varying $\theta$
- This could lead to tailoring the app to individual environments
- We use geographically weighted ($w_{tm}$)local polynomial estimation
- The local effect at $s^*$ is taken to be polynomial in $s$ near $s^*$
- Observations are kernel weighted by distance $||s - s^*||$
- Double robustness: We prove consistency if either the potential-outcome or mean models are correct
Smoke Sense – Local estimate of $\psi(s)$

Wild CIs of $\psi$ with bias correction
EXAMPLE: Spatial matching

- Commercial fishing is highly regulated for conservation
- There is variation in the type and degree of regulation
- Using data from global Gill et al. [4] we compare two broad classes of regulations: no take (NT) and multi-use (MU)
- Our objective is to compare the relative effectiveness of these two policies

---

Data and notation

- **Response:** $Y_i$ is the log biomass at site $i = 1, \ldots, n = 9987$

- **Treatment:** $A_i = 1$ for NT and $A_i = 0$ for MU

- **Covariates:** $X_i$ are known covariates

- **Dependence structure:** the $n$ observations are nested in 215 Marine Protection Areas (MPAs)
Response \((Y)\) by treatment \((A)\)
## Covariates \((X)\)

<table>
<thead>
<tr>
<th>Continuous</th>
<th>Site-level</th>
<th>MPA-level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Latitude (degree)</td>
<td>MPA size (km(^2))</td>
</tr>
<tr>
<td></td>
<td>Longitude (degree)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depth (m)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wave exposure (kW/m)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance to shoreline (km)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance to market (km)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coastal population (million/100km(^2))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sample date (year)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum sea surface temperature (°C)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chlorophyll-A (mg/m(^3))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reef area with 15km (km(^2))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPA age (years)</td>
<td></td>
</tr>
<tr>
<td>Categorical</td>
<td>Habitat type (16)</td>
<td>Country (43)</td>
</tr>
<tr>
<td></td>
<td>Marine ecoregion (56)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sampling protocol (6)</td>
<td></td>
</tr>
</tbody>
</table>
(3) Matching

Dependence structure
Challenges

- The treatment variables are not randomly assigned

- The allocation of treatments is likely influenced by biomass

- Sites are nested within MPA, giving (spatial) dependence
Potential outcomes (POs) and assumptions

▶ $Y_i(a)$ is the PO under treatment $a \in \{0, 1\}$

▶ The PO model for a site in MPA $r$ is

$$Y_i(a) = \mu_a(X_i) + \alpha_r + \varepsilon_i$$

where $\alpha_r$ is an MPA random effect

▶ Assumption: the POs for each unit are not influenced by the treatment assigned to other units

▶ Assumption: conditional independence between treatment and the POs when accounting for the covariates
Matching estimator

- We propose a matching estimator to account for spatial dependence and missing confounders.
- Each site is matched by space and covariates with $M$ observations with the opposite treatment.
- The missing PO (i.e., $Y_i(0)$ if $A_i = 1$) are imputed using the mean of these $M$ observations.
- The naive matching estimator for the average treatment effect (ATE) is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

- The average treatment effect on the treated (ATT) is similar.
Extensions

- The naive estimator is biased if $\mu_a(X)$ varies by $a$

- For fit several regression models including random forests to remove this bias

- Deriving the standard error is challenging because of MPA effects and sites are members of multiple pairs

- We develop a wild bootstrap for this purpose

- **Theory**: We prove that our estimator is consistent and asymptotically normal
### Results

**Table:** Summary of the average treatment effect (ATE) and the average treatment effect on the treated (ATT) with estimated standard errors in parentheses when comparing the multi-use (MU) policy and no-taken (NT) policy in MPAs where MU is considered as treatment group; Response is \( \log(\text{Fish Biomass}) \).

<table>
<thead>
<tr>
<th>Method</th>
<th>ATE</th>
<th>ATT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point Estimate</td>
<td>95% CI</td>
</tr>
<tr>
<td>Matching on MPA-level covariates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sieve Method</td>
<td>-0.49 (0.31)</td>
<td>(-1.10, 0.13)</td>
</tr>
<tr>
<td>Smooth Spline</td>
<td>-0.27 (0.25)</td>
<td>(-0.77, 0.23)</td>
</tr>
<tr>
<td>Regression Forest</td>
<td>-0.57 (0.35)</td>
<td>(-1.26, 0.12)</td>
</tr>
<tr>
<td>Matching on all covariates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sieve Method</td>
<td>-0.41 (0.17)</td>
<td>(-0.76, -0.07)</td>
</tr>
<tr>
<td>Smooth Spline</td>
<td>-0.34 (0.30)</td>
<td>(-0.93, 0.26)</td>
</tr>
<tr>
<td>Regression Forest</td>
<td>-0.70 (0.32)</td>
<td>(-1.32, -0.07)</td>
</tr>
</tbody>
</table>
EXAMPLE: Adjusting for unmeasured spatial confounders

▶ The key assumption that there are no missing confounders is generally impossible to verify

▶ However, it may be possible to remove the effects of unmeasured confounders that are smooth spatial functions

▶ We propose a method to adjust for missing spatial confounding variables using spectral methods.

▶ Motivating example Wu et al (2020) found that an increase of 1 $\mu g/m^3$ in PM$_{2.5}$ is associated with a 15% increase in the COVID-19 mortality rate.

---

The exposure variable $X$ is county-level average exposure to PM$_{2.5}$ for 2000-2016.
The response variable $Y$ is the cumulative COVID-19 mortality counts through May 12, 2020 for US counties.
Our approach

▶ We assume that local information is more informative than global information

▶ If you could design the experiment you would assign different $A(\mathbf{s})$ to neighbors, right?

▶ We propose a joint model for $A$ and the unobserved spatial confounder, $X$\textsuperscript{6}

▶ We derive the optimal confounder adjustment under the joint model

▶ We then discuss assumptions needed to identify the parameters in the confounder adjustment

Unmeasured confounders

- $X(s)$ is unknown or not measured

- Obviously, additional assumptions are required to estimate its correlation with $A(s)$

- We assume $[A(s), X(s)]$ follows a bivariate spatial GP

- Then $X$ can be written $X(s) = \hat{X}(s) + \delta(s)$ where
  - $\hat{X}(s) = E\{X(s)|A\}$
  - $\delta(s)$ is a Gaussian process that is independent of $A$
Unmeasured confounders

The observation equation then becomes

\[ Y(s) = \alpha + A(s)\beta_A + \hat{X}(s)\beta_X + \delta(s) + \epsilon(s) \]

- \( \hat{X}(s) \) is the required confounder adjustment
- The spatial term \( \delta(s) \) is independent of \( A \)
- Therefore, if we know \( \hat{X} \) we eliminate spatial confounding
- But how to estimate \( \hat{X} \)?
We model $A$ and $X$ using spectral methods

- If $A$ and $X$ are stationary they have spectral representations

- The spectral representations are

$$X(s) = \int \exp(i\omega^T s)x(\omega)d\omega$$

$$A(s) = \int \exp(i\omega^T s)a(\omega)d\omega$$

- $\omega \in \mathbb{R}^2$ is a frequency

- $x(\omega)$ and $a(\omega)$ are mean zero and independent across $\omega$
Spectral decomposition in 1D

- Full signal (black): $X(s) = \int \exp(i\omega^T s)x(\omega)d\omega$
- Low-frequencies (red): $\tilde{X}(s) = \int_{\|\omega\| < T} \exp(i\omega^T s)x(\omega)d\omega$
Spectral decomposition in 1D

- **Full signal (black):** $X(s) = \int \exp(i\omega^T s)x(\omega)d\omega$
- **High-frequencies (red):** $\tilde{X}(s) = \int_{||\omega||>T} \exp(i\omega^T s)x(\omega)d\omega$
We model $X$ and $A$ using spectral methods

- Dependence is measured by the coherence\(^1\):

$$\alpha(\omega) = \text{Cor}\{x(\omega), a(\omega)\}$$

which we allow to vary by frequency

- However, we assume that

$$\lim_{||\omega|| \to \infty} \alpha(\omega) = 0$$

and prove that we can identify the treatment effect

- This allows for confounding of large-scale features (small $\omega$) but assumes independence at small scales (large $\omega$)

- This is a similar motivation for matching spatial neighbors

\(^1\)Kleiber (2017) has a good review
Model in the spatial domain

- The model in the spatial domain is

\[ Y(s) = \beta_0 + A(s)\beta_x + \hat{X}(s)\beta_x + \delta(s) + \epsilon(s) \]

- The confounder adjustment is

\[ \hat{X}(s) = \int K(u - s)A(u)du, \]

where \( K(u - s) \) is the inverse Fourier transform of \( \alpha(\omega) \)

- \( \hat{X} \) is a smoothed version of \( A \) that is included to adjust for spatial confounding

- The methods can be fit using standard MCMC/ML methods
Plot of the covariate $A$ (left) versus $\hat{\mathbf{X}}$ (right)

$v_{xz} = 1v_x$

$v_{xz} = 3v_x$
COVID mortality/PM

\[ \exp(\beta_k) \]

- Spectral Parametric
- Spectral Semiparametric
- Standard
Spatial causal inference is an important and rapidly-evolving area.

We have covered a wide variety of methods, and this is just the surface.


Thanks to the NIH (R01ES031651-01,R01ES027892-01) and NSF (DMS-2152887)