Spatial Deconfounder

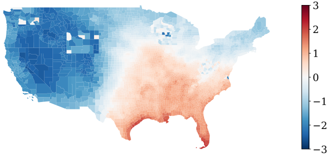
Interference-Aware Deconfounding for Spatial Causal Inference

Ayush Khot*[1] Miruna Oprescu*[2] Maresa Schröder[3] Ai Kagawa[4] Xihaier Luo[4]

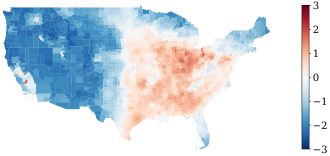
[1]University of Illinois at Urbana-Champaign [2]Cornell University, Cornell Tech
[3]LMU Munich, Munich Center for Machine Learning [4]Brookhaven National Laboratory

Spatial Causal Inference

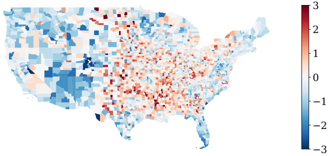
- Spatial index $s = (i, j)$
- Neighborhood $\mathcal{N}_s$ is a $(2r+1) \times (2r+1)$ square centered at s, excluding s itself
- **Features (Covariates)** $\mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}$
- **Unobserved confounder** $U(s)$
- **Interventions (Treatments)** $A_s, A_{\mathcal{N}_s}, \in \{0, 1\}$
- **Outcomes** Y_s



Humidity ($U(s)$)



$PM_{2.5}$ (A_s)



Mortality (Y_s)

1. Interference (SUTVA violation)

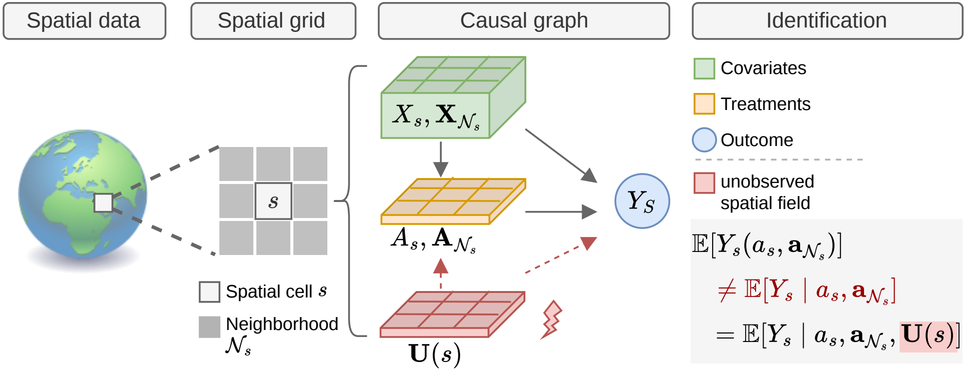
$Y_s(\mathbf{a}) = Y_s(a_s, \mathbf{a}_{\mathcal{N}_s})$

Pollution / disease spill across regions

2. Unobserved spatial confounding

$U(s)$ (humidity, topography, ...)

Jointly drives A_s and Y_s



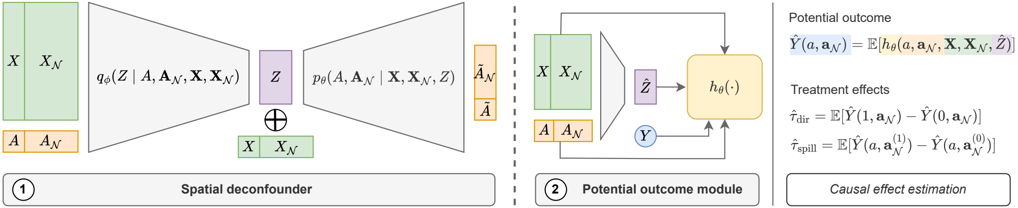
Interference is a Feature, Not a Bug

Existing methods handle one challenge by *assuming away* the other.

We show that they are deeply connected:
interference reveals structure in the latent confounder.

Each unit observes $(A_s, \mathbf{A}_{\mathcal{N}_s})$, all shaped by the *same* $U(s)$

- The deconfounder methods (Wang & Blei, 2019) assume i.i.d units with multiple simultaneous causes.
- Localized interference creates the multi-cause structure needed for deconfounding
- This yields an interference-driven deconfounding strategy that reconstructs latent spatial confounding *nonparametrically* and *without multiple treatment types*

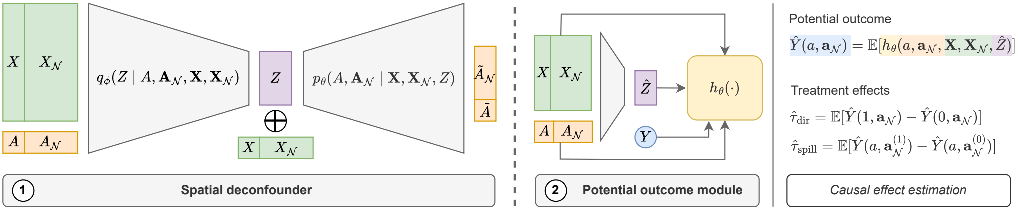


Stage 1: Confounder Reconstruction with C-VAE

- **Encoder** maps treatments and covariates into latent embedding Z_s:

$$q_\phi(Z_s \mid A_s, \mathbf{A}_{\mathcal{N}_s}, X_s, \mathbf{X}_{\mathcal{N}_s}) = \mathcal{N}(\mu_\phi(\cdot), \operatorname{diag} \sigma_\phi^2(\cdot))$$

- **Decoder** predicts A_s given covariates and latent: $p_\psi(A_s \mid X_s, \mathbf{X}_{\mathcal{N}_s}, Z_s)$
- **GMRF prior** enforces spatial smoothness: $p_\theta(Z) = \mathcal{N}\big(\mathbf{0},\ \tau^{-1}(L + \epsilon I)^{-1}\big)$



Stage 2: Potential outcome module

- After Stage 1, $\hat{Z}_s = \mathbb{E}_{q_\phi}[Z_s]$ and using a **flexible spatial model** h (e.g. U-Net):

$$\hat{Y}_s = h(A_s,\ \mathbf{A}_{\mathcal{N}_s},\ X_s,\ \mathbf{X}_{\mathcal{N}_s},\ \hat{Z}_s)$$

- **Direct Effect**: $\hat{\tau}_{\text{dir}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\Big[h(1, A_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, \hat{Z}_s) - h(0, A_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, \hat{Z}_s)\Big]$
- and analogously for the spillover effect $\hat{\tau}_{\text{spill}}$

Assumptions

Under **localized interference** and typical **deconfounding assumptions** (e.g. no single-cause confounding), we can reconstruct a confounder Z_s such that the direct and spillover effects are identifiable as

$$
\begin{aligned}
\tau_{\text{dir}} = \mathbb{E}_{\mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z}\Big[& \mathbb{E}_Y\big[Y_s \mid A_s = 1, \mathbf{A}_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z_s\big] \\
& - \mathbb{E}_Y\big[Y_s \mid A_s = 0, \mathbf{A}_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z_s\big]\Big], \qquad (1)
\end{aligned}
$$

$$
\begin{aligned}
\tau_{\text{spill}} = \mathbb{E}_{\mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z}\Big[& \mathbb{E}_Y\big[Y_s \mid a, \mathbf{A}_{\mathcal{N}_s} = \mathbf{a}^{(1)}_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z_s\big] \\
& - \mathbb{E}_Y\big[Y_s \mid a, \mathbf{A}_{\mathcal{N}_s} = \mathbf{a}^{(0)}_{\mathcal{N}_s}, \mathbf{X}_s, \mathbf{X}_{\mathcal{N}_s}, Z_s\big]\Big]. \qquad (2)
\end{aligned}
$$

Extending the SpaCE Benchmark

SpaCE (Tec et al., 2024): semi-synthetic spatial causal datasets from real datasets

Our extensions:

- Generate outcomes under two confounding regimes:

$$\text{(Local confounding)} \quad \hat{Y}_s = f(A_s, A_{\mathcal{N}_s}, X_s) + R_s, \tag{3}$$

$$\text{(Spatial confounding)} \quad \hat{Y}_s = f(A_s, A_{\mathcal{N}_s}, X_s, X_{\mathcal{N}_s}) + R_s, \tag{4}$$

  where f is a predictive function learned from the observed data
- Mask key covariates to simulate hidden $U(s)$

Datasets: $PM_{2.5} \rightarrow$ Mortality, (socioeconomic) $SO_4 \rightarrow PM_{2.5}$ (atmospheric)

Baselines:

- Spatial autoregressive models ($\textsc{s2sls-lag1}$, $\textsc{durbin}$, $\textsc{gmerror}$)
- Spline-based $\textsc{spatial}$ and $\textsc{spatial+}$ with exposure-mapped variants *E-MAP*
- Other models like $\textsc{gcnn}$, $\textsc{dapsm}$, and $\textsc{unet}$

Results: Local Confounding

Table 1: Standardized absolute bias, $\sigma_y^{-1}|\hat{\tau} - \tau| \pm 95\%$ CI. $p \approx 0.5$: good model fit; R: deconfounder neighborhood radius. Best and second-best values are bolded and underlined, respectively. **Environment:** $PM_{2.5} \rightarrow m$ $(r_d = 1)$.

hidden confounder: ρ_{pop}

Method	DIR	SPILL	p
C-VAE-SPATIAL+ ($R=1$)	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.02 \pm 0.00}$	0.51 ± 0.07
C-VAE-SPATIAL+ ($R=2$)	0.04 ± 0.01	$\underline{0.04 \pm 0.01}$	0.50 ± 0.05
DAPSM	0.25 ± 0.01	n/a	n/a
GCNN	0.36 ± 0.03	n/a	n/a
GMERROR	$\underline{0.03 \pm 0.00}$	n/a	n/a
DURBIN	$\underline{0.03 \pm 0.00}$	0.07 ± 0.00	n/a
S2SLS-LAG1	$\underline{0.03 \pm 0.00}$	0.07 ± 0.00	n/a
SPATIAL+	0.05 ± 0.02	n/a	n/a
SPATIAL	$\mathbf{0.02 \pm 0.00}$	n/a	n/a
E-MAP-SPATIAL+ ($R=1$)	0.04 ± 0.02	0.07 ± 0.07	n/a
E-MAP-SPATIAL ($R=1$)	$\mathbf{0.02 \pm 0.00}$	0.07 ± 0.07	n/a

hidden confounder: q_{summer}

Method	DIR	SPILL	p
C-VAE-SPATIAL+ ($R=1$)	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.02 \pm 0.00}$	0.53 ± 0.05
C-VAE-SPATIAL+ ($R=2$)	$\underline{0.04 \pm 0.01}$	$\underline{0.04 \pm 0.01}$	0.51 ± 0.05
DAPSM	0.30 ± 0.03	n/a	n/a
GCNN	0.41 ± 0.03	n/a	n/a
GMERROR	0.20 ± 0.00	n/a	n/a
DURBIN	0.20 ± 0.00	0.06 ± 0.00	n/a
S2SLS-LAG1	0.20 ± 0.00	0.06 ± 0.00	n/a
SPATIAL+	0.05 ± 0.02	n/a	n/a
SPATIAL	$\mathbf{0.02 \pm 0.00}$	n/a	n/a
E-MAP-SPATIAL+ ($R=1$)	0.05 ± 0.02	0.07 ± 0.07	n/a
E-MAP-SPATIAL ($R=1$)	$\mathbf{0.02 \pm 0.00}$	0.07 ± 0.07	n/a

Results: Spatial Confounding

Table 2: Standardized absolute bias, $\sigma_y^{-1}|\hat{\tau} - \tau| \pm$ 95% CI. $p \approx 0.5$: good model fit; R: deconfounder neighborhood radius. Best and second-best values are bolded and underlined, respectively.

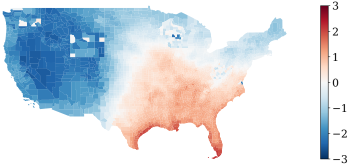
$PM_{2.5} \rightarrow m$, $(r_d = 1)$, conf: ρ_{pop}

Method	DIR	SPILL	p
C-VAE-UNET ($R = 1$)	0.06 ± 0.02	0.16 ± 0.08	0.53 ± 0.03
C-VAE-UNET ($R = 2$)	$\underline{0.04 \pm 0.01}$	0.07 ± 0.02	0.51 ± 0.04
DAPSM	0.20 ± 0.01	n/a	n/a
GCNN	0.17 ± 0.06	n/a	n/a
GMERROR	0.05 ± 0.00	n/a	n/a
DURBIN	0.05 ± 0.00	$\underline{0.05 \pm 0.00}$	n/a
S2SLS-LAG1	0.05 ± 0.00	$\mathbf{0.04 \pm 0.00}$	n/a
SPATIAL+	0.12 ± 0.09	n/a	n/a
SPATIAL	$\mathbf{0.03 \pm 0.00}$	n/a	n/a
UNET	0.06 ± 0.01	0.17 ± 0.04	n/a
E-MAP-SPATIAL+ ($R = 1$)	0.11 ± 0.09	0.06 ± 0.06	n/a
E-MAP-SPATIAL ($R = 1$)	$\mathbf{0.03 \pm 0.00}$	0.07 ± 0.06	n/a

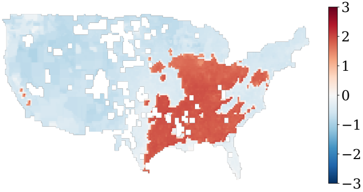
$SO_4 \rightarrow PM_{2.5}$, $(r_d = 1)$, conf: OC

Method	DIR	SPILL	p
C-VAE-UNET ($R = 1$)	0.09 ± 0.00	0.12 ± 0.02	0.48 ± 0.05
C-VAE-UNET ($R = 2$)	$\mathbf{0.06 \pm 0.01}$	0.07 ± 0.03	0.54 ± 0.03
DAPSM	1.57 ± 0.00	n/a	n/a
GCNN	0.42 ± 0.15	n/a	n/a
GMERROR	0.13 ± 0.00	n/a	n/a
DURBIN	0.13 ± 0.00	$\mathbf{0.01 \pm 0.00}$	n/a
S2SLS-LAG1	0.13 ± 0.00	0.08 ± 0.00	n/a
SPATIAL+	0.12 ± 0.08	n/a	n/a
SPATIAL	$\underline{0.07 \pm 0.00}$	n/a	n/a
UNET	$\underline{0.07 \pm 0.02}$	$\underline{0.05 \pm 0.02}$	n/a
E-MAP-SPATIAL+ ($R = 1$)	0.10 ± 0.07	$\underline{0.05 \pm 0.01}$	n/a
E-MAP-SPATIAL ($R = 1$)	0.07 ± 0.00	$\underline{0.05 \pm 0.01}$	n/a

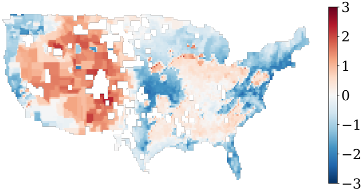
Recovering the Hidden Spatial Field



True $U(s)$
(summer humidity q_{summer})



PC1 of $\hat{Z}_s$
(captures treatment)



PC2 of $\hat{Z}_s$
(slightly recovers $U(s)$)

The second principal component of the latent space mirrors the large-scale spatial structure of the true (unobserved) confounder

Conclusions

Key Contributions:

- **Spatial Deconfounder:** A framework that **jointly addresses interference and unobserved spatial confounding** by treating neighborhood treatments as a multi-cause signal.
- We establish **identification from a C-VAE** with a spatial prior under assumptions on the **latent spatial field** and **outcome structure**.
- **Empirics:** In semi-synthetic datasets with both interference and unobserved spatial confounding, the **Spatial Deconfounder shows consistent bias reductions** relative to strong spatial and deep-learning baselines.

Broader Impact:

- Enables **more accurate causal estimates** that can inform policy decisions in applications with both interference and unobserved spatial confounding such as **environmental health, climate science, and social sciences.**

`github.com/moprescu/Spatial-Deconfounder`