A Design-Based Approach to Spatial Correlation

Ruonan Xu[1] Jeffrey M. Wooldridge[2]

[1]Rutgers University

[2]Michigan State University

ESEM, August 2022

Research Questions

- Spatial correlation
  - A merger between two gasoline companies (Houde, 2012)
  - Closure and demolition of public housing (Aliprantis and Hartley, 2015)
  - Spillover effects to adjacent entities
- Q: Whether standard errors should be adjusted for spatial correlation? If so, when?
  - Sampling scheme
  - Assignment design
  - Model specification

Finite Population Framework

- Population size $\to \infty$
- Sampling-based v.s. design-based uncertainty (Abadie et al, 2020)
- "With spatial data, it is common for the sample and the population to be the same," (Pinkse et al, 2007)
- Can explicitly introduce different sampling schemes

Contribution

- Derive new laws of large numbers and a central limit theorem for near-epoch dependent (NED) processes
  - Nonstationary processes, unbounded moments, irregularly spaced lattices
  - Cluster sampling, cluster correlation on top of spatial correlation
  - Accommodate sampling from superpopulations as a special case
- Examine the necessity of SHAC standard errors for a general class of estimators
  - Finite population asymptotic properties of M-estimators
  - Functions of M-estimators: e.g., average partial effect

Literature

- Limit theorems for random fields: Jenish and Prucha (2009), Jenish and Prucha (2012), Bradley and Tone (2017)
- Finite population inference: Abadie et al. (2020), Abadie et al. (2022), Xu (2021a, 2021b), Bojinov et al. (2021), Savje et al. (2021), Savje (2021), Leung (2022)
- Spatial econometrics: e.g., Xu and Lee (2019)

Notation

- $D \subseteq \mathbb{R}^d$, $d \geq 1$: lattice of (possibly) unevenly placed locations in $\mathbb{R}^d$
- $\{D_M\}$, $|D_M| \to \infty$
- Relax SUTVA: potential outcome function $y_{iM}(\boldsymbol{x}_M)$, $\boldsymbol{x}_M = \{x_{iM}, i \in D_M, M \geq 1\}$
- X_{iM}: assignment variables, z_{iM}: attributes, $Y_{iM} = y_{iM}(\boldsymbol{X}_M)$: realized outcome
- Denote $W_{iM} = (\boldsymbol{X}_M, Y_{iM})$ for brevity
- Conditioned on the potential outcomes and attributes in the population
- Assume that the finite population parameters are identified

Estimand

$$\theta_M^* = \arg\min_\theta \frac{1}{|D_M|} \sum_{g=1}^{G_M} \sum_{j \in D_{gM}} \mathbb{E}\big[q_{jM}(W_{jM}, \theta)\big]$$

$$= \arg\min_\theta \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big[q_{iM}(W_{iM}, \theta)\big]$$

Spatial M-estimator

$$\hat{\theta}_N = \arg\min_{\theta} \frac{1}{|D_N|} \sum_{g=1}^{G_M} \sum_{j \in D_{gM}} R_{jM} q_{jM}(W_{jM}, \theta)$$
$$= \arg\min_{\theta} \frac{1}{|D_N|} \sum_{i \in D_M} R_{iM} q_{iM}(W_{iM}, \theta)$$

- R_{iM}: binary sampling indicator
- $|D_N| = \sum_{i \in D_M} R_{iM}$

Sampling Scheme

Assumption 2

(i) The sampling scheme consists of two steps. In the first step, a random group of clusters is drawn according to Bernoulli sampling with probability $\rho_{cM} > 0$; in the second step, units are independently sampled, according to a Bernoulli trial with probability $\rho_{uM} > 0$, from the subpopulation consisting of all the sampled clusters. (ii) The sequence of sampling probabilities ρ_{cM} and ρ_{uM} satisfies $\rho_{cM} \to \rho_c \in [0, 1]$, $\rho_{uM} \to \rho_u \in [0, 1]$, and $|D_M|\rho_{uM}\rho_{cM} \to \infty$ as $M \to \infty$.

Sampling Scheme

- $\rho_{cM} = \rho_{uM} = 1$: observe the entire population
- $\rho_{cM} = 1$, $\rho_{uM} < 1$: random sampling
- $\rho_{cM} < 1$, $\rho_{uM} \leq 1$: cluster sampling
- $\rho_c = 0$: a negligible fraction of clusters are sampled from a population of a large number of clusters
- $\rho_c = 1$, $\rho_u = 0$: a negligible portion of units are randomly drawn from a large population

Selected Assumptions

Assumption 1

Increasing domain asymptotics

Assumption 3

$\max_{1 \leq g \leq G_M} |D_{gM}| \leq C < \infty$ as $M \to \infty$.

Assumption 4

The sampling indicators, $R = \{R_{iM}, i \in D_M, M \geq 1\}$, are independent of the assignment variables, $X = \{X_{iM}, i \in D_M, M \geq 1\}$, and the underlying mixing random fields, $U = \{U_{iM}, i \in T_M, M \geq 1\}$, where $D_M \subseteq T_M \subseteq D$.

Spatial Dependence

Assumption 5 (Mixing condition)

For the input random field U: (i) $\bar{\alpha}(r) \to 0$ as $r \to \infty$; (ii) $\lim_{r \to \infty} \bar{\rho}(r) < 1$.

Definition

Assumption 6 (NED condition)

The random field $g = \{g_{iM}(W_{iM}, \theta), i \in D_M, M \geq 1\}$ is L_2-NED on $U = \{U_{iM}, i \in T_M, M \geq 1\}$ with the scaling factors d_{iM} and the NED coefficients $\psi(s)$ of size $-2d(r-1)/(r-2)$ for some $r > 2$. The $g(\cdot)$ function includes $q_{iM}(W_{iM}, \theta)$, $m_{iM}(W_{iM}, \theta)$, $\nabla_\theta m_{iM}(W_{iM}, \theta)$, $f_{iM}(W_{iM}, \theta)$, and $\nabla_\theta f_{iM}(W_{iM}, \theta)$ defined in Appendix A.

Asymptotic Distribution

Theorem 1

Under Assumptions 1-6, and Assumption A.1 in Appendix A, $V_M^{-1/2}|D_N|^{1/2}(\hat{\theta}_N - \theta_M^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_k)$.

$$V_M = H_M(\theta_M^*)^{-1} S_M H_M(\theta_M^*)^{-1}$$

$$\begin{aligned} S_M = &\Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) + \rho_{uM}\rho_{cM}\Delta_{spatial,M}(\theta_M^*) \\ &- \rho_{uM}\rho_{cM}\Delta_{E,M} - \rho_{uM}\rho_{cM}\Delta_{EC,M} - \rho_{uM}\rho_{cM}\Delta_{ES,M} \end{aligned}$$

▸ Notation

The SHAC standard errors?

$$\begin{aligned} S_M =& \Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) + \rho_{uM}\rho_{cM}\Delta_{spatial,M}(\theta_M^*) \\ & - \rho_{uM}\rho_{cM}\Delta_{E,M} - \rho_{uM}\rho_{cM}\Delta_{EC,M} - \rho_{uM}\rho_{cM}\Delta_{ES,M} \end{aligned}$$

Report the SHAC standard errors when

- Assignment variables are spatially correlated
  - Holds for spatial assignments both at the individual level or the cluster level
- Spillover effects are specified in the model
  - Even in the absence of spillover effects in the potential outcome function

Sampling Probability Matters

$$
\begin{aligned}
S_M = & \Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) + \rho_{uM}\rho_{cM}\Delta_{spatial,M}(\theta_M^*) \\
& - \rho_{uM}\rho_{cM}\Delta_{E,M} - \rho_{uM}\rho_{cM}\Delta_{EC,M} - \rho_{uM}\rho_{cM}\Delta_{ES,M}
\end{aligned}
$$

- $\rho_u = 0$: reporting the EHW standard error would suffice
- $\rho_c = 0$: reporting the cluster-robust standard error would suffice

Usual SHAC Variance Estimator is Conservative

$$\hat{V}_{SN} = \hat{H}_N(\hat{\theta}_N)^{-1}\hat{S}_N(\hat{\theta}_N)\hat{H}_N(\hat{\theta}_N)^{-1}$$

$$\hat{S}_N(\theta) = \frac{1}{|D_N|}\sum_{i\in D_M}\sum_{j\in D_M} R_{iM}R_{jM}\cdot\omega\left(\frac{\nu(i,j)}{b_M}\right)m_{iM}(W_{iM},\theta)m_{jM}(W_{jM},\theta)'$$

Theorem 2

Under Assumptions 1-7, and Assumptions A.1-A.2 in Appendix A, $\hat{V}_{SN} - (V_M + \rho_{uM}\rho_{cM}V_E) \xrightarrow{p} \mathbf{0}$, where $V_E = H_M(\theta_M^*)^{-1}S_E H_M(\theta_M^*)^{-1}$ and $S_E = \frac{1}{|D_M|}\sum_{i\in D_M}\sum_{j\in D_M}\omega\left(\frac{\nu(i,j)}{b_M}\right)\mathbb{E}\left[m_{iM}(W_{iM},\theta_M^*)\right]\mathbb{E}\left[m_{jM}(W_{jM},\theta_M^*)\right]'$.

Simulation Designs

- Spatial assignments at the individual level
- Spatial assignments at the cluster level
- Spatial assignments allowing for spillover effects
- Spatial assignments in nonlinear models

Spatial Correlation at the Individual Level

- Uneven lattice, $\sqrt{M} \times \sqrt{M}$
- $y_{ig}(x_{ig}) = a \cdot \beta_{ig} x_{ig} + c_g + u_{ig}$
- $u_M = p_u W_u u_M + \epsilon_M$
- $\epsilon_M \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- W_u: contiguity matrix, units i and j are neighbors if $\nu(i,j) \leq \sqrt{2}$
- Regress Y_i on 1 and X_i
- Expected size of each dimension is 18 $\Rightarrow$ expected sample size of 324
- Clusters in the sampling scheme: group the consecutive three units by order $\Rightarrow$ expected number of 108 clusters in the sample

Five Sampling Schemes

- Observe the entire population
- Independently sample clusters with a probability of 0.25
- Independently draw units with a probability of 0.25
- Independently sample clusters with a probability of 0.01
- Independently draw units with a probability of 0.01

Independent Assignment

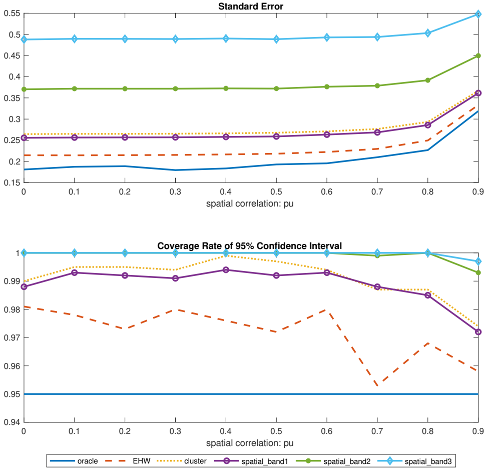

Figure: Independent Assignments Observing Entire Population

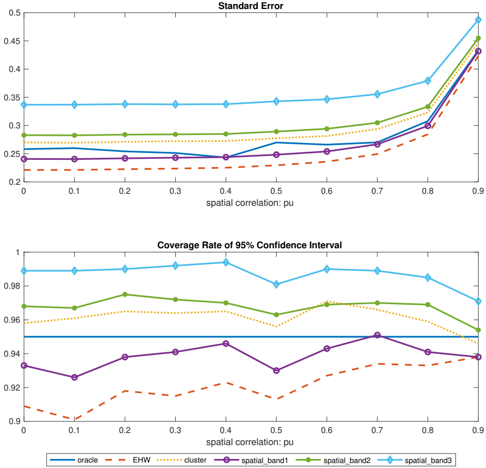

Figure: Independent Assignments with Cluster Sampling 0.25

Spatial Assignments

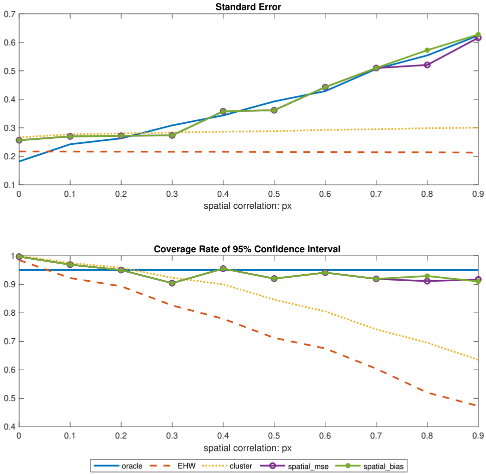

Figure: Spatial Assignments Observing Entire Population

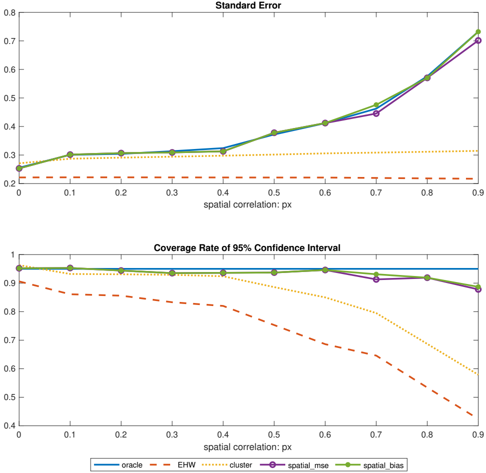

Figure: Spatial Assignments with Cluster Sampling 0.25

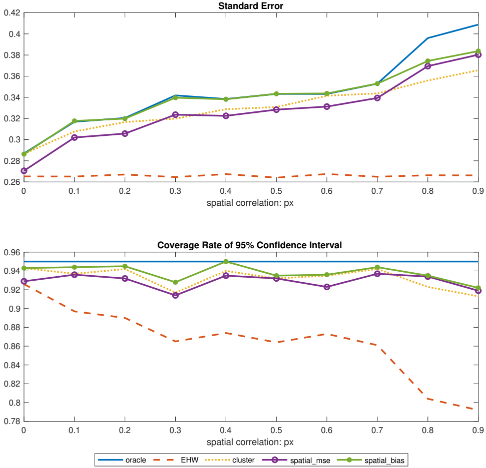

Figure: Spatial Assignments with Cluster Sampling 0.01

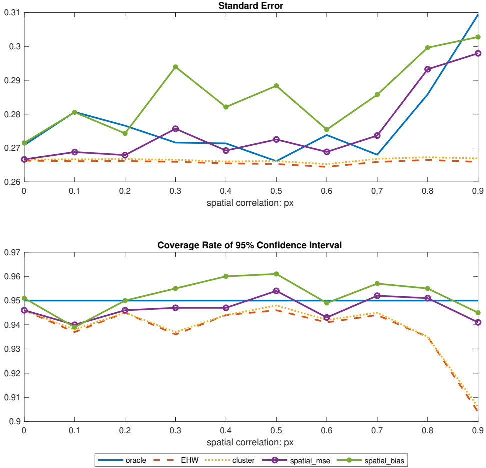

Figure: Spatial Assignments with Independent Sampling 0.01

Spillover Effects

- $y_{ig}(\boldsymbol{x}_M) = 2\beta_{ig}x_{ig} + \gamma W_x \boldsymbol{x}_M + \epsilon_{ig}$
- Three assignment and spillover combinations
  - No spatial assignments and no spillover effects
  - No spatial assignments with spillover effects
  - Spatial assignments combined with spillover effects
- Regress Y_i on 1, X_i, and $W_s\boldsymbol{X}_s$

Table: Specifying Spillover Effects

	entire population			cluster sampling			independent sampling		
	$p_x = 0$, $\gamma = 0$	$p_x = 0$, $\gamma = 1$	$p_x = 0.1$, $\gamma = 1$	$p_x = 0$, $\gamma = 0$	$p_x = 0$, $\gamma = 1$	$p_x = 0.1$, $\gamma = 1$	$p_x = 0$, $\gamma = 0$	$p_x = 0$, $\gamma = 1$	$p_x = 0.1$, $\gamma = 1$
coeff	0.002	1.002	1.061	0.004	0.636	0.546	-0.001	0.500	0.472
std	0.133	0.110	0.136	0.113	0.128	0.169	0.137	0.149	0.160
EHW	0.101	0.083	0.100	0.087	0.094	0.123	0.103	0.112	0.121
EHW_CI	(0.863)	(0.860)	(0.855)	(0.860)	(0.843)	(0.836)	(0.854)	(0.853)	(0.848)
cluster	0.107	0.096	0.109	0.109	0.115	0.150	0.111	0.121	0.129
cluster_CI	(0.887)	(0.910)	(0.887)	(0.938)	(0.924)	(0.911)	(0.883)	(0.883)	(0.875)
SHAC1	0.129	0.107	0.135	0.112	0.122	0.167	0.134	0.145	0.159
SHAC1_CI	(0.934)	(0.938)	(0.950)	(0.943)	(0.941)	(0.947)	(0.936)	(0.935)	(0.935)
SHAC2	0.129	0.108	0.136	0.113	0.123	0.169	0.135	0.146	0.160
SHAC2_CI	(0.936)	(0.941)	(0.936)	(0.941)	(0.943)	(0.941)	(0.939)	(0.933)	(0.937)

Conclusion

- Using a design-based approach, we identify the sources of uncertainty underlying spatial data
- Whenever there are spatial assignments or when spillover effects are estimated, the SHAC standard errors must be used, unless the sampling probability is negligible

Definition

- α-mixing and maximal correlation coefficient in Bradley and Tone (2017) ▸ Definition 1
- NED random fields in Jenish and Prucha (2012) ▸ Definition 2
- *m*-dependent random fields in Moricz, Stadtmuller, and Thalmaier (2008) ▸ Definition 3

◂ return

Definition

Definition 1

Let $\mathcal{A}$ and $\mathcal{B}$ be two sub-σ-algebras of $\mathcal{F}$, and let

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup(|P(AB) - P(A)P(B)|, A \in \mathcal{A}, B \in \mathcal{B})$$

and

$$\rho(\mathcal{A}, \mathcal{B}) = \sup |corr(f, g)|, f \in L^2_{real}(\mathcal{A}), g \in L^2_{real}(\mathcal{B}).$$

For $K \subseteq D_M$ and $V \subseteq D_M$, let $\sigma_M(K) = \sigma(U_{iM}, i \in K)$ and $\alpha_M(K, V) = \alpha(\sigma_M(K), \sigma_M(V))$. Then, the α-mixing coefficient for the random field U is defined as:

$$\bar{\alpha}(r) = \sup_M \sup_{K,V}(\alpha_M(K, V), \nu(K, V) \geq r).$$

The maximal correlation coefficient is defined as:

$$\bar{\rho}(r) = \sup_M \sup_{K,V}(\rho_M(K, V), \nu(K, V) \geq r).$$

return

Definition

Definition 2

Let $W = \{W_{iM}, i \in D_M, M \geq 1\}$ be a random field , let $U = \{U_{iM}, i \in T_M, M \geq 1\}$ be another random field, where $|T_M| \to \infty$ as $M \to \infty$, and let $d = \{d_{iM}, i \in D_M, M \geq 1\}$ be an array of finite positive constants. Then the random field W is said to be $L_p(d)$-near-epoch dependent on the random field U if

$$\|W_{iM} - E(W_{iM}|\mathcal{F}_{iM}(s))\|_p \leq d_{iM}\psi(s)$$

for some sequence $\psi(s) \geq 0$ with $\lim_{s\to\infty} \psi(s) = 0$. The $\psi(s)$ are called the NED coefficients, and the d_{iM} are called the NED scaling factors. W is said to be L_p-NED on U of size $-\lambda$ if $\psi(s) = O(s^{-\mu})$ for some $\mu > \lambda > 0$.

return

Definition

Definition 3

A random field $U = \{U_{iM}, i \in D_M, M \geq 1\}$ is called *m*-dependent if for all finite subsets $K, V \subset D$ with $\nu(K, V) > m$ the σ-algebras $\sigma(U_{iM}, i \in K)$ and $\sigma(U_{iM}, i \in V)$ are independent.

◂ return

Matrix Notation

$$\Delta_{ehw,M}(\theta) = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\left[m_{iM}(W_{iM}, \theta) m_{iM}(W_{iM}, \theta)'\right]$$

$$\Delta_{E,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\left[m_{iM}(W_{iM}, \theta_M^*)\right] \mathbb{E}\left[m_{iM}(W_{iM}, \theta_M^*)\right]'$$

$$\Delta_{cluster,M}(\theta) = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{1}(C_{iM} = C_{jM}) \mathbb{E}\left[m_{iM}(W_{iM}, \theta) m_{jM}(W_{jM}, \theta)'\right],$$

$$\Delta_{EC,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{1}(C_{iM} = C_{jM}) \mathbb{E}\left[m_{iM}(W_{iM}, \theta_M^*)\right] \mathbb{E}\left[m_{jM}(W_{jM}, \theta_M^*)\right]'$$

$$\Delta_{spatial,M}(\theta) = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{1}(C_{iM} \neq C_{jM}) \mathbb{E}\left[m_{iM}(W_{iM}, \theta) m_{jM}(W_{jM}, \theta)'\right]$$

$$\Delta_{ES,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{1}(C_{iM} \neq C_{jM}) \mathbb{E}\left[m_{iM}(W_{iM}, \theta_M^*)\right] \mathbb{E}\left[m_{jM}(W_{jM}, \theta_M^*)\right]'$$

$$H_M(\theta) = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\left[\nabla_\theta m_{iM}(W_{iM}, \theta)\right]$$

return

APE Estimator

$$\gamma_M^* = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\left[f_{iM}(W_{iM}, \theta_M^*)\right]$$

$$\hat{\gamma}_N = \frac{1}{|D_N|} \sum_{i \in D_M} R_{iM} f_{iM}(W_{iM}, \hat{\theta}_N)$$

$$\begin{aligned} V_{f,M} = & \Delta^f_{ehw,M} + \rho_{uM} \Delta^f_{cluster,M} + \rho_{uM}\rho_{cM} \Delta^f_{spatial,M} \\ & - \rho_{uM}\rho_{cM}\Delta^f_{E,M} - \rho_{uM}\rho_{cM}\Delta^f_{EC,M} - \rho_{uM}\rho_{cM}\Delta^f_{ES,M} \end{aligned}$$

Results Carry Over

Theorem 3

Under Assumptions 1-7, and Assumptions A.1-A.3 in Appendix A,

(1) $V_{f,M}^{-1/2}|D_N|^{1/2}(\hat{\gamma}_N - \gamma_M^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_q)$;

(2) $\hat{V}_{f,SN} - (V_{f,M} + \rho_{uM}\rho_{cM}V_{f,E}) \xrightarrow{p} \mathbf{0}$.