Bipartite interference: Estimating health effects of power plant interventions

Fabrizia Mealli

European University Institute
fabrizia.mealli@eui.eu

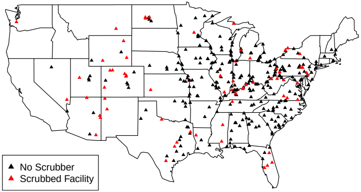
with Cory Zigler (UT Austin) and Laura Forastiere (Yale)

ESEM - August 28, 2023

Major Pollution Source: Power Plants

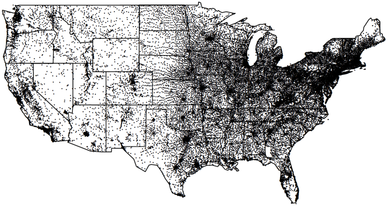
⇒ Many regulations to reduce emissions

Interventions at Power Plants:
"Scrubber" Installation
⇝ Reduce SO_2 emissions



Affect?
⟶

Health at Population Locations
(Zip Codes)



Question: Do scrubbers on coal-fired power plants causally affect hospitalizations for Ischemic Heart Disease (IHD) among Medicare beneficiaries?

Formalization of Bipartite Structure

Two types of observational units:

1. *Intervention Units*
   - $\mathcal{P} = \{p_1, p_2, \ldots, p_j, \ldots, p_J\}$: a set of J power plants, where interventions occur (or not)
   - $S_j = 1/0$ denotes presence/absence of intervention
   - Treatment allocation: $\mathbf{S} = (S_1, \ldots, S_j, \ldots, S_J)$
   - Space of possible vectors of treatment allocations: $\mathcal{S}(J)$
   - Covariates $\mathbf{X}^{int} = (\mathbf{X}_1^{int}, \ldots, \mathbf{X}_j^{int}, \ldots, \mathbf{X}_J^{int})$
2. *Outcome Units*
   - $\mathcal{M} = \{m_1, m_2, \ldots, m_i, \ldots, m_n\}$: a set of n locations, where outcomes are measured
   - Y_i denotes the outcome of interest
   - Covariates: $\mathbf{X}^{out} = (\mathbf{X}_1^{out}, \ldots, \mathbf{X}_i^{out}, \ldots, \mathbf{X}_n^{out})$

Potential Outcomes

- $Y_i(\mathbf{s})$ potential outcome that would be observed at outcome unit i under treatment allocation $\mathbf{s}$, $\mathbf{s} \in \mathcal{S}(J)$
- No multiple versions of treatment: $Y_i(\mathbf{s}) = Y_i(\mathbf{s}')$ $\forall i$ when $\mathbf{s} = \mathbf{s}'$

(*Zigler and Papadogeorgou, 2021; Papadogeorgou, Mealli, Zigler, 2019*)

Examples and Challenges

Examples:

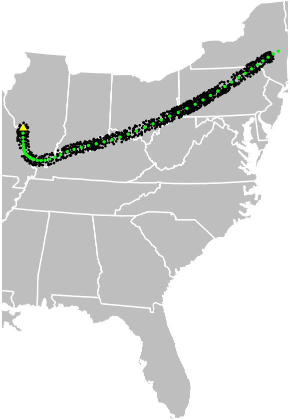
1. *Economics of housing (Stock, 1989)*
    - Intervention units: Hazardous-waste disposal sites where treatments (cleaning up) can be applied
    - Outcome units: Locations where outcomes (e.g., housing values) are measured
2. *Education economics (Crema, 2022)*
    - Intervention units: Neighborhoods which may be exposed to openings of charter schools (treatment)
    - Outcome units: Traditional public schools (TPS) where outcome (e.g., racial segregation) are measured

Non trivial differences in:

- the formulation of the estimands
- the assignment mechanism
- different types and sources of confounding

Interference Mapping and Causal Quantities

Using Long-range Pollution Transport Models (HyADS)
Henneman et al (2019)



- Emissions originating at a power plant (**yellow triangle**) *move*

  ⇝ Long distances towards conversion to harmful pollution

- Potential outcomes for zip code i:

  ⇝ Y_i(power plants that pollute over i)

- Estimands that compare the potential outcomes

  ⇝ if the *most influential (or closest)* power plant were treated

  ⇝ if the 10 most influential power plants were treated

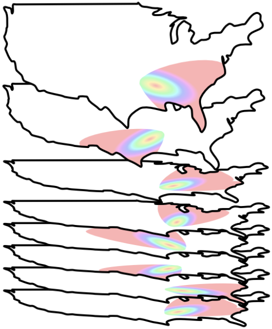
  ⇝ any other intervention allocation

Interference Mapping and Causal Quantities

Using Long-range Pollution Transport Models (HyADS)

Henneman et al (2019)

Use atmospheric transport models to simulate emissions events starting at individual stack

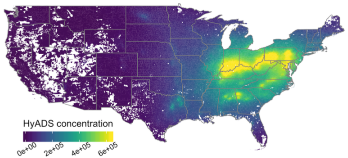


- Emissions originating at a power plant (**yellow triangle**) *move*
  - ⇝ Long distances towards conversion to harmful pollution
- Potential outcomes for zip code i:
  - ⇝ Y_i(power plants that pollute over i)
- Estimands that compare the potential outcomes
  - ⇝ if the *most influential (or closest)* power plant were treated
  - ⇝ if the 10 most influential power plants were treated
  - ⇝ any other intervention allocation

Interference Mapping

"Source-Receptor Matrix" or "Bipartite Weighted Directed Network"

- $t_{ij} \equiv$ influence of j^{th} power plant on i^{th} location
  - i.e., # of times dispersed parcels from p_j pass over location m_i (re-scaled to a max of 1)
  - directly output by HyADS atmospheric model



- Interference mapping: $n \times J$ matrix, T:

$$
\begin{array}{c} \\ ZIP_1 \\ \vdots \\ ZIP_i \\ \vdots \\ ZIP_n \end{array}
\begin{array}{c} \begin{array}{ccccc} PP_1 & \dots & PP_j & \dots & PP_J \end{array} \\
\begin{pmatrix}
t_{11} & \dots & t_{1j} & \dots & t_{1J} \\
\vdots & \dots & \vdots & \dots & \vdots \\
t_{i1} & \dots & t_{ij} = (\text{influence of } PP_j \to ZIP_i) & \dots & t_{iJ} \\
\vdots & \dots & \vdots & \dots & \vdots \\
t_{n1} & \dots & t_{nj} & \dots & t_{nJ}
\end{pmatrix} \end{array}
$$

Bivariate Treatment

1. Outcome unit i 's "key associated" intervention unit: $j^*_{(i)}$, i.e., the power plant that is most influential for ZIP code i
   - $j^*_{(i)} \equiv j : t_{ij} = \max_j \{t_{i1}, \ldots, t_{ij}, \ldots t_{iJ}\}$

2. "Individual" Treatment: $Z_i = S_{j^*_{(i)}}$, i.e., scrubber status of the *key associated* power plant $j^*_{(i)}$
   - $Z_i = 0, 1$ according to whether the key power plant has a scrubber or not

3. "Upwind" Treatment: Function of the scrubber statuses of all other "linked" power plants
   - "Exposure Mapping:" $g_i(\cdot; T) : \{0, 1\}^{J-1} \rightarrow \mathcal{G}_i$
   - $g_i(\mathbf{S}, T) = G_i = \sum_{j \neq j^*_{(i)}} t_{ij} S_j$
   - HyADS-weighted upwind treatment rate

(*Aronow, Samii, 2017; Forastiere et al., 2021*)

"Upwind Interference" Assumption

(Adaptation of SUTVA)

$$Y_i(\mathbf{S}) = Y_i(Z_i, G_i) = Y_i(Z_i', G_i') = Y_i(\mathbf{S}')$$

- Reduces interference to depend only on Z_i and G_i (not on the entire $\mathbf{S}$)
- Hospitalizations are the same under different $\mathbf{S}$ if the treatment of the key power plant and the "upwind" treatment rate are the same
- **The key assumption** about interference

Assignment mechanism governing the joint treatment:

$$P\left(\mathbf{Z}, \mathbf{G} \mid \mathbf{X}^{out}, \mathbf{X}^{int}, \{Y_i(z, g), z \in \{0, 1\}, g \in \mathcal{G}\}\right)$$

“Direct” Effects in the Bipartite Setting

Average dose-response under key-associated treatment z and upwind treatment g

$$\mu(z,g) = \mathbb{E}_{X^{int},X^{out}}\left[\mathbb{E}_{Y(\cdot)|X^{int},X^{out}}\left[Y_i(z,g) \mid X_i^{int}, X_i^{out}\right]\right]$$

The “direct” effect of treating the key plant while holding fixed the “upwind” treatments:

$$\tau(g) = \mu(1,g) - \mu(0,g)$$

“Overall” average over the distribution of “upwind” treatments:

$$\tau = \sum_{g \in \mathcal{G}} \tau(g) P(G_i = g)$$

“The effect on IHD hospitalizations that we are expecting to have in a zipcode from installing a scrubber on the *key associated* power plant”

"Indirect" or "Spillover" Effects in the Bipartite Setting

The "indirect" or "upwind" effect of installing scrubbers on upwind power plants without changing the key plant:

$$\delta(g; z) = \mu(z, g) - \mu(z, g^{min})$$

"Overall" average over the distribution of "upwind" treatments:

$$\Delta(z) = \sum_{g \in \mathcal{G}} \delta(g; z) P(G_i = g)$$

"The effect on IHD hospitalizations that we are expecting to have in a zipcode from installing more scrubbers on upwind plants"

Confounding Specification

Ignorability of the Joint Treatment:

$$Y_i(z, g) \perp\!\!\!\perp Z_i, G_i \mid \{\mathbf{X}_j^{int}\}_{j:t_{ij}>0}, \mathbf{X}_i^{out}, \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}_i, \forall i$$

$\mathbf{X}_i^{out}$ includes: population, urbanicity, race/ethnicity, education, HH income, poverty, occupied housing, migration, smoking, region, temperature, humidity, Medicare age, sex.

$$\{\mathbf{X}_j^{int}\}_{j:t_{ij}>0} = \left\{\mathbf{X}_{j^*_{(i)}}^{int}, \{\mathbf{X}_j^{int}\}_{j \neq j^*_{(i)}:t_{ij}>0}\right\}$$

$\mathbf{X}_{j^*_{(i)}}^{int}$ includes: operating time, heat input, %operating capacity, ARP phase II participation, NO_x controls sulfur content of coal *of power plant* $j^*_{(i)}$

$\{\mathbf{X}_j^{int}\}_{j \neq j^*_{(i)}:t_{ij}>0}$ includes: "upwind" versions of power plant characteristics (e.g., the HyADS-weighted operating time of upwind plants)

Individual and Neighborhood Propensity Scores

Extension of Forastiere et al. (2021)

View "individual" and "upwind" treatments as bivariate treatment and estimate a **joint propensity score**:

$$\psi_i(z, g; x^{int}, x^{out}) = P\left(Z_i = z, G_i = g \mid \{\mathbf{X}_j^{int}\}_{j:t_{ij}>0} = x^{int}, \mathbf{X}_i^{out} = x^{out}\right)$$

Further decompose into:

$$\begin{aligned}
&\psi_i(z, g; x^{int}, x^{out}) \\
= \quad & P\left(Z_i = z \mid \{\mathbf{X}_j^{int,z}\}_{j:t_{ij}>0} = x^{int,z}, \mathbf{X}_i^{out,z} = x^{out,z}\right) \qquad (1) \\
& \times P\left(G_i = g \mid Z_i = z, \{\mathbf{X}_j^{int,g}\}_{j:t_{ij}>0} = x^{int,g}, \mathbf{X}_i^{out,g} = x^{out,g}\right) \qquad (2)
\end{aligned}$$

$(1) \equiv \phi_i(z; x^{int,z}, x^{out,z}) \equiv$ "individual propensity sore"

$(2) \equiv \lambda_i(g; z, x^{int,g}, x^{out,g}) \equiv$ "upwind propensity score"

Estimation strategy

Extension of Forastiere et al. (2021)

1. Estimate ϕ_i and λ_i: $\hat{\phi}_i$, $\hat{\lambda}_i$
2. Stratify zip codes into $K = 5$ strata based on based on $\hat{\phi}_i$
3. Specify and estimate $Y_i(z, g)|\hat{\lambda}_i \sim f^y(z, g, \hat{\lambda}; \theta_k)$; derive predicted values $\hat{Y}_i(z, g)$
4. Estimate within-stratum dose-response function $\mu_k(z, g)$:

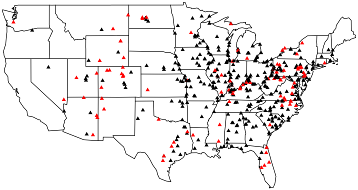
$$\widehat{\mu}_k(z, g) = \frac{\sum_{i \in n_k} \widehat{Y}_i(z, g)}{n^k}$$

5. Obtain an overall estimate: $\widehat{\mu}(z, g) = \sum_{k=1}^{K} \widehat{\mu}_k(z, g)\pi^k$
6. Calculate Direct and Upwind effects
7. Intervention-Unit Bootstrap - not exactly right, but shown via simulations to be conservative as a measure of model and design variability

Power Plant and Zip Code Data

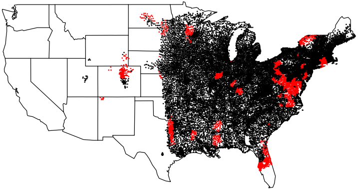
Integrated with HyADS atmospheric model

472 Coal-fired power plants operating in 2005



Affect →

Medicare IHD Hospitalizations in 2005 at 25,553 zip codes (overlap criteria)

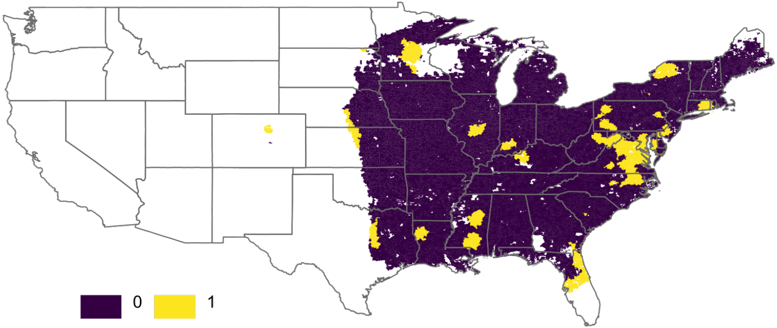


- 106 with scrubbers installed
- Information on emissions, plant size, operating capacity, other controls, etc. (EPA)

- Information on population demographics (Census, Medicare), weather (NOAA), smoking rates (CDC)

Individual Treatment

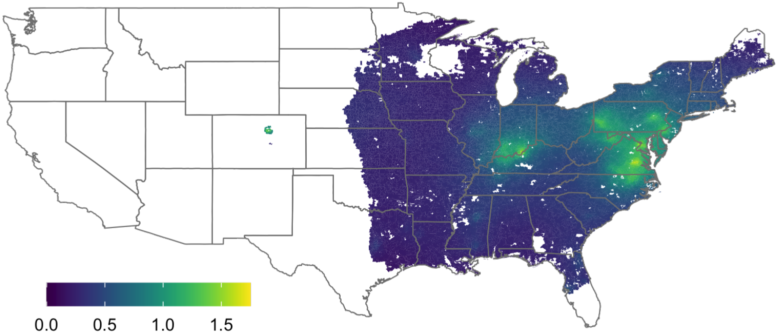
Whether the **most influential** plant has a scrubber



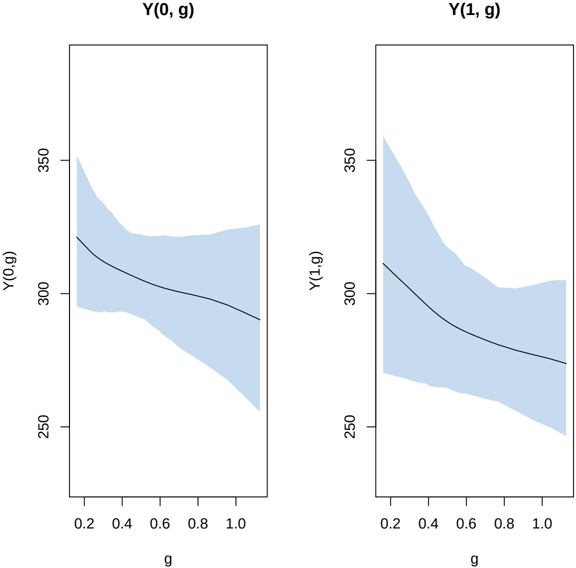
$Z_i = 0$ if j^* has no scrubber
$Z_i = 1$ if j^* has scrubber (2,753 zip codes)
278 power plants are key-associated and 35 have scrubber

"Upwind" Treatment

HyADS-weighted treatment rate of upwind plants



Average Dose-Response for $Y(z, g)$



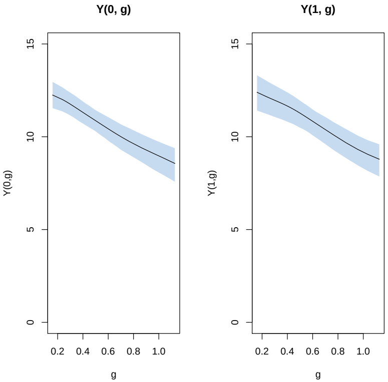
Estimated Effects of Scrubbers on IHD Hospitalization

- Average “direct” effect of installing a scrubber on the key power plant (τ)
  - **Most influential plant**: -23 (-38, 14)

  ... hospitalizations per 10K person-years

- Average “upwind” effect of installing more scrubbers on “upwind” plants ($\Delta(z)$):
  - **Most influential** plant:
    - If $z = 1$: -18 (-39, 2)
    - If $z = 0$: -14 (-37, 2)

Average Dose-Response for $PM_{2.5}(z, g)$



$\hat{\tau}$=-0.37 (-0.32, 0.84);

$\Delta(0)$=-1.34 (-1.75, -0.99); $\Delta(1)$=-1.16 (-1.63, -0.80)

Summary

- Causal inference in bipartite observational settings
  - ▷ Introduce new types of causal questions and estimands
  - ▷ Complex exposure patterns generate different types of interference, going beyond unit-to-unit or spatial interference
  - ▷ New notions of confounding and homophily
  - ▷ New methods for **bipartite causal inference**
- Relevance for many other interventions with a clear distinction between intervention and outcome units

References

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912-1947.

Crema A. (2022) School competition and classroom segregation. Unpublished Manuscript.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534), 901-918.

Henneman LRF, Christine Choirat C., Ivey C., Cummiskey K., and Zigler CM (2019). Characterizing population exposure to coal emissions sources in the United States using the HyADS model. *Atmospheric Environment*, 203:271-280.

Papadogeorgou G., Mealli F., Zigler CM (2019) Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics* 75 (3), 778-787.

Stock JH (1989) Nonparametric policy analysis. *Journal of the American Statistical Association*, 84 (406):567-575.

Zigler, C., Forastiere, L., and Mealli, F. (2020). Bipartite interference and air pollution transport: Estimating health effects of power plant interventions. *arXiv preprint* arXiv:2012.04831.

Zigler CM and Papadogeorgou G. (2021) Bipartite Causal Inference with Interference. *Statistical Science*, 36, 109-123.