

Policy Effects on Site-Level Emissions: Master's Thesis Initial Proposal

Arnav Agarwal

October 2025

QMSS, Columbia University

Motivation

- Environmental policies are designed to cut pollution — but do they make a clear, local difference around major industrial sites?
- A site-level view keeps the focus on the places where emissions actually occur and where communities are affected.
- New satellite data now lets us track air quality consistently across countries and over time.
- High quality AI-derived embeddings also allow for encoding of high-dimensional geospatial controls.
- We follow plants before and after policy rules take effect to see how local air quality changes.
- Clear evidence helps governments and industry target efforts, improve enforcement, and prioritise investments where they matter most.

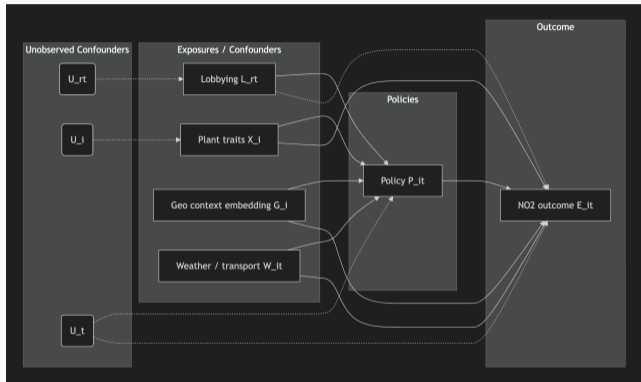
Research Questions

- **RQ1: Average effect.** What is the effect of policy exposure on tropospheric NO₂ within buffers around treated facilities?
- **RQ2: Heterogeneity.** How do effects vary by primary fuel, capacity, and controls (SCR/SNCR/FGD), and by local context?
- **RQ3: Moderation.** How do seasonal and meteorological conditions (winds, PBLH) moderate short-run impacts?
- **RQ4: Dynamics.** Event-time profile (anticipation, persistence).

Causal Diagram & Variables

Key variables

- P_{it} : policy exposure (US EPA EASEY Service; EU ETS) at facility i , month t .
- E_{it} : NO₂ around facility (e.g., 10 km buffer from TROPOMI).
- X_i : plant traits (capacity, primary fuel); abatement
- G_{it} : geographic/land-use embeddings around the site.
- W_{it} : weather/transport (winds, PBLH, temperature, precip.).



Data Sources

Plants & geometry

- EIA-860 (US, annual): capacity, fuel, coords.
- EIA-860M (US, monthly): operability mask.
- EPA FRS (US, static): backup geocodes.
- EEA LCP (EU, annual): plant attrs & coords.

Policy / exposure

- US EPA EASEY Service: policy x time x location flags under various US government programmes; ozone season.
- EU ETS (Energy Trading System): credits exposure.

Outcomes & controls

- TROPOMI NO₂ (D→M): buffered columns.
- ERA5 (monthly): winds, PBLH, temp, precip.
- AlphaEarth (annual): geographic/context embeddings.

Methodology (Post-Data Processing)

1) Feature engineering

- Outcome: satellite NO₂ around each facility (monthly).
- Exposure: monthly policy indicators; operability status.
- Covariates: core plant traits; weather/climate *embeddings* capturing spatiotemporal variation (as in prior MLC work).

2) Identification

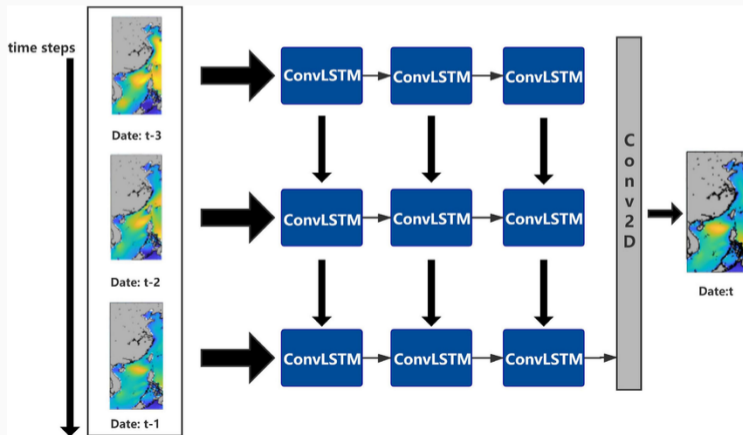
- Staggered policy timing \Rightarrow event-study/DiD with unit and time effects.

3) Estimation

- Perform estimation with simpler (e.g. linear) to complex (e.g. deep learning) models.
- Probe ATE and CATEs (fuel, size, controls, context) and run robustness checks.

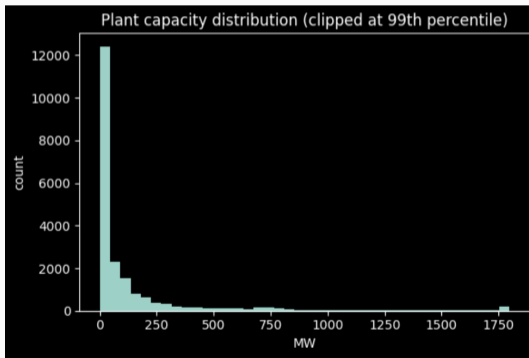
Example Spatiotemporal Encoder: Conv-LSTM

- For capturing spatiotemporal variation in weather/climate variables we can use encoders like Conv-LSTMs (or Conv3D, Temporal Convolutional Networks, etc.) and get vectorised embeddings.



Initial EDA: Facility Universe (US & EU)

- Construct one row per plant with {lat, lon, capacity_MW, fuel_primary, abatement_flags}.
- Merge US EIA-860 and EU LCP harmonised fields into a common schema.



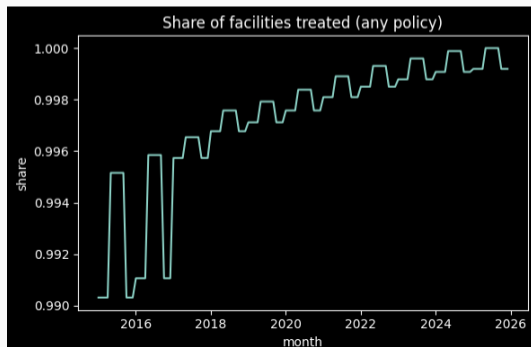
Initial EDA: Facility Universe (US & EU)

country		EU	US
fuel_group_top			
Gas		0.482	0.122
Solar		0.000	0.392
Other/Unknown		0.000	0.208
Other		0.128	0.033
Oil		0.106	0.053
Biomass		0.150	0.007
Coal		0.134	0.013
Hydro		0.000	0.088
Wind		0.000	0.083

	country	fuel_group		n
16	US	Solar		6331
15	US	Other/Unknown		3350
2	EU	Gas		2160
9	US	Gas		1963
11	US	Hydro		1425
18	US	Wind		1344
13	US	Oil		859
0	EU	Biomass		674
1	EU	Coal		600
4	EU	Other gas		514
3	EU	Oil		475
17	US	Waste		341

Initial EDA: Policy Panel & Timing

- Monthly flags: `treat_any`, `treat_cap`; ozone–season logic for CSAPR NO_x OS (May–Sept).
- Mask US exposures using EIA–860M operability where available (leave EU unchanged).
- Shares are computed over operable rows when coverage exists.



Key Challenges (Probably Scope) & Design Choices

Linkage & scope

- EU ETS installations \neq EEA-LCP IDs \Rightarrow fuzzy + geo join ($\pm 1\text{--}2$ km).
- Various data sources, fused using multiple methods (combining deep learning and standard estimation methods), can be data-hungry and can be prone to instability

Measurement

- Satellite QA/cloud filtering; buffer-size sensitivity (5–20 km); urban background vs rural.

Key Challenges (Probably Scope) & Design Choices

Identification

- Time-Dependent Treatment Effects (TDTTE): Determining timescales for when we expect the full effect of our treatments to have manifested. Similarly, anticipation windows / placebo leads.
- Although capturing policy attributes alleviates, issue with encoding varying treatment features.
- Spatial spillovers (e.g. plants within close geographical proximity resulting in independence violation).