

# Quantifying the Effects of Climate Policy Stringency on Verified Emissions and Satellite-Derived NOx

Master's Thesis

Arnav Agrawal

MA in Quantitative Methods in the Social Sciences, Columbia University.

## Abstract

I investigate how European Union Emissions Trading System (EU ETS) policy stringency affects verified CO<sub>2</sub> emissions, reported NOx emissions, and satellite-derived NOx proxies at 521 large combustion plants across Europe (2018–2023). Using two-way fixed effects with facility and region-by-year fixed effects, I find that a 10% allocation shortfall is associated with 1.9% lower verified CO<sub>2</sub> emissions ( $\beta = -0.186$ ,  $p < 0.001$ ) and 0.7% lower reported NOx ( $\beta = -0.066$ ,  $p = 0.023$ ). Crucially, continuous interaction models reveal that treatment effects are *fuel-dependent*: coal-dominant facilities show the strongest response ( $\beta_{\text{coal}} = -0.62$  for CO<sub>2</sub>,  $-0.88$  for NOx), followed by oil and gas, while biomass facilities show no significant response. This pattern—where policy effectiveness depends on combustion technology rather than facility size or location—suggests carbon pricing operates primarily through fuel-specific abatement incentives. Satellite-derived NOx from TROPOMI, constructed via Beirle-style flux-divergence with a modified 5 km integration radius, shows null main effects at the permissive 0.01 kg/s detection limit due to measurement noise, but provides corroborative evidence through similar fuel-dependent heterogeneity patterns. The framework makes three methodological contributions: (i) geospatial foundation model embeddings (Google AlphaEarth) as high-dimensional controls for satellite retrieval confounders; (ii) NUTS2 regional clustering for inference with PyPSA-Eur power system clusters for electricity-sector heterogeneity; and (iii) a simplified Beirle-style NOx quantification method adapted for panel econometrics. The triple-outcome approach enables cross-validation:

directional agreement between administrative CO<sub>2</sub> and reported NOx, with consistent heterogeneity patterns across all three outcomes, strengthens confidence in the fuel-dependent policy mechanism.

**Keywords:** Climate Policy, Climate Monitoring, EU ETS, Verified Emissions, Satellite Remote Sensing, TROPOMI, NOx Emissions, Flux Divergence, Difference-in-Differences, Causal Inference, Large Combustion Plants

## 1 Introduction

Evaluating climate policy requires measuring actual emission outcomes. The European Union Emissions Trading System (EU ETS) generates rich administrative data on verified CO<sub>2</sub> emissions at the installation level, providing the gold standard for measuring greenhouse gas output from regulated facilities. However, relying solely on self-reported emissions raises questions about verification and leaves unmeasured the local air quality co-benefits that accompany carbon reductions. Satellite remote sensing offers an independent, physically-grounded approach to quantifying emissions from space, potentially revealing both verification opportunities and co-pollutant dynamics that administrative data cannot capture.

This study adopts a triple-outcome approach that exploits the complementary strengths of administrative and satellite data. The three outcomes are: (i) verified EU ETS CO<sub>2</sub> emissions—high-quality, installation-level measures from the EU ETS registry that provide accurate compliance trajectories and absolute emission levels; (ii) reported NOx emissions from the Large Combustion Plant (LCP) Directive—administrative ground-truth for NOx that captures pollution co-benefits; and (iii) a satellite-derived NOx emission proxy—a physically grounded indicator constructed from TROPOMI NO<sub>2</sub> tropospheric columns and ERA5 winds, following the flux-divergence approach of Beirle et al. [1–3].

Why use three outcomes? CO<sub>2</sub> is a well-mixed greenhouse gas with global climate impacts; nitrogen oxides (NO<sub>x</sub>), by contrast, are criteria pollutants whose health effects—respiratory illness, cardiovascular disease, premature mortality—fall disproportionately on populations living near emission sources. Air quality co-benefits are particularly policy-relevant because they are local and immediate, whereas averted climate damages are global and long-term. The triple-outcome design provides: (i) verified CO<sub>2</sub> emissions for accurate carbon policy effect estimation, (ii) reported NOx for ground-truth measurement of air quality co-benefits, (iii) satellite-derived NOx for independent validation that is robust to reporting compliance issues and captures local air quality effects beyond stack emissions, and (iv) cross-validation opportunities where all three outcomes should respond to common policy shocks.

This study develops a novel framework for evaluating climate policy impacts using both administrative emissions data and satellite remote sensing. I focus on the European Union Emissions Trading System (EU ETS), the world’s largest carbon market, which creates economic incentives for industrial facilities to reduce CO<sub>2</sub> emissions through a cap-and-trade mechanism. The framework addresses two fundamental

methodological challenges: (i) constructing a satellite-derived NO<sub>x</sub> emission proxy that is physically interpretable and appropriate for panel econometric analysis, and (ii) controlling for high-dimensional confounders that affect both policy exposure and emission outcomes.

The study makes three methodological contributions. The first two follow a recent trend in causal inference toward incorporating machine learning-derived features to address high-dimensional confounding [4–6]; the third adapts an atmospheric physics method for panel econometrics.

**First**, I demonstrate the use of geospatial foundation model embeddings as high-dimensional controls in panel-based climate monitoring studies. Specifically, I incorporate Google AlphaEarth embeddings [7]—64-dimensional representations learned from multi-source satellite imagery, climate reanalysis, and geotagged text—as control variables in the econometric specifications. These embeddings capture between-unit heterogeneity arising from local geographic, infrastructural, and climate context in a data-efficient manner, providing a scalable approach to controlling for high-dimensional spatial confounders that would be impractical to specify manually. This application extends prior work on learned representations for causal inference—originally developed for text embeddings [5]—to the domain of geospatial environmental monitoring. The approach is particularly suited to difference-in-differences settings where high-dimensional confounders may violate the parallel trends assumption if left uncontrolled [6].

**Second**, I use Eurostat NUTS2 regions for spatial clustering in both fixed effects structure and inference. Standard errors are clustered by NUTS2 region, which groups facilities that share common regional economic conditions, labor markets, and policy enforcement mechanisms. The same regions define Region  $\times$  Year fixed effects, absorbing time-varying regional confounders that correlate with both policy exposure and air quality outcomes. Unlike sector-specific clustering (e.g., power system network clusters), NUTS2 regions apply uniformly to all industrial facility types and correspond to administrative units where EU and national environmental policies are implemented. For electricity sector heterogeneity analysis, I additionally implement PyPSA-Eur power system clusters [8]—k-means clusters computed on transmission network topology—which group facilities facing correlated wholesale prices and dispatch patterns.

**Third**, for the satellite outcome, I implement a simplified Beirle-style flux-divergence method that provides physically grounded NO<sub>x</sub> emission estimates at the facility level. The approach computes the advection—the scalar product of wind velocity and the spatial gradient of NO<sub>2</sub> column density—which under the continuity equation is proportional to local emissions minus chemical loss. For each facility, I integrate advection over a 5 km disc (reduced from Beirle’s standard 15 km based on validation against reported NO<sub>x</sub> emissions), apply a lifetime correction following [3], and convert to NO<sub>x</sub> emission rates. This methodology follows the Beirle et al. (2019, 2021, 2023) family of methods [1–3], which are physically transparent, computationally tractable for known point sources, and specifically designed for power-plant-scale NO<sub>x</sub> plumes.

The analysis links three data sources on large combustion plants (LCPs) in the EU: (i) the European Environment Agency’s LCP registry providing plant characteristics and coordinates, (ii) EU ETS compliance data providing installation-level verified emissions and free allocations, and (iii) TROPOMI satellite observations processed through the Beirle-style flux-divergence methodology to derive NOx emission proxies. Policy exposure is measured continuously through the *allocation ratio*—free allowances divided by verified emissions—where values below unity indicate facilities must purchase additional permits, creating direct economic pressure to reduce emissions.

The econometric framework employs two-way fixed effects (TWFE) specifications with facility and region $\times$ year fixed effects. The Callaway-Sant’Anna estimator [9] proved infeasible due to panel structure (Section 2.4).

By demonstrating that administrative emissions records (CO<sub>2</sub> and NOx) and satellite-derived NOx proxies can provide individual-emitter-level, policy-parameterized estimates of emission responses to carbon market stringency, this work contributes to the emerging literature on comprehensive climate policy evaluation. The triple-outcome approach enables testing whether policy effects on verified CO<sub>2</sub> are accompanied by corresponding changes in both reported and satellite-observed combustion co-pollutants.

## 2 Background and Literature Review

### 2.1 The EU Emissions Trading System

The EU ETS, established in 2005, operates as a cap-and-trade system covering approximately 40% of EU greenhouse gas emissions. Large combustion plants with thermal input exceeding 20 MW are required to hold European Union Allowances (EUAs) equal to their verified CO<sub>2</sub> emissions. Allowances are distributed through a combination of free allocation (based on historical benchmarks and carbon leakage risk) and auctioning. Installations that emit more than their free allocation must purchase additional allowances, creating marginal abatement incentives [10].

The policy has evolved through five phases, with Phase III (2013–2020) and Phase IV (2021–2030) introducing progressively tighter caps and reduced free allocation, particularly for the power sector. This study focuses on the period 2018–2023, spanning the transition from Phase III to Phase IV and capturing significant variation in policy stringency across facilities.

### 2.2 Satellite-Based Air Quality Monitoring

The TROPOMI instrument aboard Sentinel-5P, operational since late 2017, provides daily global observations of tropospheric NO<sub>2</sub> column densities at unprecedented spatial resolution ( $\sim 3.5 \times 5.5 \text{ km}^2$  at nadir). This represents a significant improvement over predecessor instruments (OMI, GOME-2) and enables detection and quantification of emissions from individual point sources [3, 11].

Previous studies have used satellite observations to verify emission reductions from policy interventions. [12] documented substantial NO<sub>x</sub> reductions over Europe between

1996 and 2010, attributing these to environmental policies and economic recession. However, these studies typically analyze aggregate regional trends rather than plant-level responses to specific policy parameters.

### 2.3 Satellite-Based NO<sub>x</sub> Emission Quantification: The Flux-Divergence Approach

A key methodological challenge in quantifying emissions from satellite-observed NO<sub>2</sub> is separating the source signal from background concentrations and converting column densities to emission rates. This challenge is particularly acute in Europe, where high population density means that most large combustion plants are located in or near urban areas, surrounded by other pollution sources (traffic, industry, heating).

The flux-divergence (or advection) approach, developed by Beirle et al. [1–3], provides a physically grounded solution. The method exploits the continuity equation: horizontal NO<sub>2</sub> fluxes  $\mathbf{F} = \mathbf{w}V$  (where  $\mathbf{w}$  is wind velocity and  $V$  is tropospheric vertical column density) satisfy

$$\nabla \cdot \mathbf{F} = E - S \quad (1)$$

where  $E$  represents local emissions and  $S$  represents chemical sinks. Under typical conditions where wind field divergence is negligible, this reduces to the advection formulation:

$$A = \mathbf{w} \cdot \nabla V \approx E - S \quad (2)$$

The advection  $A$  measures the downwind rate of change in NO<sub>2</sub> column density and is particularly sensitive to strong point sources, which create sharp spatial gradients in the NO<sub>2</sub> field.

Beirle et al. (2021) [2] presented the first global catalog of NO<sub>x</sub> point source emissions derived from TROPOMI using this approach, identifying 451 sources. Beirle et al. (2023) [3] introduced version 2 with several improvements: use of the PAL (Products Algorithm Laboratory) NO<sub>2</sub> product with higher column densities (factor of 1.1–1.4), corrections for plume height effects on satellite sensitivity, topographic corrections, and a lifetime correction to account for chemical loss within the integration radius. These refinements resulted in emission estimates approximately 3 times higher than version 1, with validation showing agreement within 20% of reported emissions from the German Environment Agency (UBA) and US EPA.

[13] developed an alternative regression-based approach for decomposing TROPOMI NO<sub>2</sub> into urban, industrial, and background components during COVID-19, demonstrating that wind information can isolate individual source contributions even in complex emission environments. [11] extended the methodology to megacities, estimating both emissions and effective NO<sub>x</sub> lifetimes through simultaneous fitting of downwind plume evolution.

This study adopts the Beirle family of methods because they are: (i) physically transparent, grounded in the continuity equation; (ii) computationally tractable for known point sources; and (iii) specifically designed and validated for power-plant-scale NO<sub>x</sub> plumes. I implement a simplified version appropriate for panel econometric analysis.

## 2.4 Causal Inference with Staggered Treatment Timing

Standard two-way fixed effects estimators can produce biased estimates when treatment timing varies across units and treatment effects are heterogeneous [14]. Recent methodological advances, including the Callaway and Sant’Anna [9] and Sun and Abraham [15] estimators, address these concerns by constructing treatment effect estimates using only valid comparisons (treated versus not-yet-treated or never-treated units) and allowing for treatment effect heterogeneity across cohorts and time.

I initially planned to complement TWFE with the Callaway-Sant’Anna estimator. However, this approach proved infeasible—of 457 ever-treated facilities (defining treatment as  $R_{it} < 1$ ), 386 (84.5%) were already treated in 2018—the first year of my panel. With no pre-treatment observations for 84.5% of treated units, the estimator cannot form a meaningful untreated counterfactual cohort. I therefore interpret  $\beta_{\text{TWFE}}$  as an *average response to changes in  $R_{it}$* . Pre-trend testing is left to future work with an extended 2013–2017 panel (EU ETS Phase 3). The continuous treatment approach exploits within-facility variation in policy stringency over time and remains valid under homogeneous treatment effects, which the heterogeneity analysis (Section 5.2) examines empirically.

## 2.5 High-Dimensional Controls and ML-Derived Features in Causal Inference

A growing literature in causal inference addresses the challenge of controlling for high-dimensional confounders—settings where the number of potential control variables is large relative to sample size, or where relevant confounders are difficult to specify manually. The foundational work of [4] established the “double/debiased machine learning” framework, showing how machine learning methods can be used to estimate nuisance parameters (propensity scores, outcome regressions) while maintaining valid inference on treatment effects. This approach enables researchers to control for high-dimensional confounders without imposing restrictive parametric assumptions.

In the difference-in-differences context specifically, [6] developed efficient estimators for settings where the parallel trends assumption holds only conditional on high-dimensional covariates. This is particularly relevant when unobserved confounders that violate parallel trends can be proxied by high-dimensional observables—such as detailed geographic or economic characteristics that would be impractical to specify manually but can be captured through flexible ML methods.

A parallel development concerns the use of *learned representations*—embeddings from neural networks or foundation models—as control variables. [5] demonstrated that text embeddings can serve as effective controls for confounding in observational studies, provided the embeddings capture the relevant confounding information. The key insight is that pre-trained (usually unsupervised or self-supervised) representations encode information about latent confounders that would otherwise be unobserved. This approach has been extended to various domains, including image embeddings and, most recently, geospatial foundation models.

For clustered inference, [16] established theoretical foundations for network cluster-robust standard errors. They show that valid cluster-robust inference requires clusters

with low “conductance”—the ratio of edges crossing cluster boundaries to total edges within clusters. This implies that clusters should be defined based on the correlation structure of the data-generating process, not arbitrary geographic or administrative boundaries. When observations are connected through a network (as power plants are through the transmission grid), clusters derived from network topology can satisfy these requirements.

This study contributes to this literature by demonstrating two novel applications: (i) using geospatial foundation model embeddings (AlphaEarth) as controls for spatial confounding in environmental panel data, and (ii) using k-means clusters derived from power system network features (PyPSA-Eur) for heterogeneity analysis. The application of model-derived clustering—where clusters are computed on features from an external domain-specific model rather than on the outcome data itself—to econometric time series analysis is a novel technique, sparsely applied in the literature.

## 3 Data

This section describes the data sources, processing pipeline, and construction of the analysis panel. The study combines administrative records on industrial facilities and EU ETS compliance with satellite remote sensing and meteorological reanalysis data.

### 3.1 Data sources

#### 3.1.1 EEA Large Combustion Plant Registry

The European Environment Agency (EEA) maintains the Industrial Emissions Portal, which includes the Large Combustion Plant (LCP) dataset. This registry provides annual reports on combustion plants with rated thermal input  $\geq 50$  MW, including:

- Geographic coordinates (latitude, longitude)
- Plant identification (LCP INSPIRE ID, installation name)
- Rated thermal capacity (MW)
- Annual fuel consumption by fuel type (TJ)
- Country of operation

The raw dataset contains 3,405 unique plant parts for the period 2018–2023. After filtering for complete capacity and fuel data, 2,821 plants remain with valid time-varying attributes.

#### 3.1.2 EU ETS Compliance Data

EU ETS installation-level compliance data is obtained from the European Union Transaction Log (EUTL), accessed via the `pyeutl` Python package. For each installation-year, the data includes:

- Verified CO<sub>2</sub> emissions (tCO<sub>2</sub>)
- Free allocation of allowances (tCO<sub>2</sub>-equivalent)
- Surrendered allowances (tCO<sub>2</sub>-equivalent)
- Installation identifier and country

The LCP and ETS datasets are linked through the EU Registry on Industrial Sites, which provides crosswalk tables mapping LCP installation parts to their parent ETS installations via normalized identifiers.

### 3.1.3 TROPOMI Satellite Observations

Tropospheric NO<sub>2</sub> column densities are obtained from the Sentinel-5P TROPOMI instrument via Google Earth Engine, using the OFFL (offline) L3 product (**COPERNICUS/S5P\_OFFL\_L3\_N02**). TROPOMI provides daily global coverage at approximately  $3.5 \times 5.5$  km<sup>2</sup> spatial resolution. Quality-filtered observations are used, retaining only pixels with quality assurance values  $\geq 0.75$ . TROPOMI captures approximately 14 orbits per day globally, with each orbit covering a distinct swath ( $\sim 2600$  km); for any given facility, only one orbit per day provides valid coverage.

Importantly, Beirle et al. (2023) [3] use the PAL (Products Algorithm Laboratory) NO<sub>2</sub> product, which provides higher tropospheric vertical column densities (TVCDs) than the OFFL product by a factor of approximately 1.1–1.4, due to updated retrieval algorithms and air mass factor corrections. This difference, combined with other methodological refinements, contributed to their version 2 emission estimates being approximately 3 times higher than version 1. Since I use the OFFL L3 product available via Google Earth Engine rather than the PAL product, the satellite-derived NO<sub>x</sub> estimates carry additional systematic error (approximately 25% lower than PAL-based estimates) that must be acknowledged.

### 3.1.4 ERA5-Land Reanalysis

Hourly 10-meter wind components ( $u_{10}, v_{10}$ ) are obtained from the ERA5-Land reanalysis product via Google Earth Engine. Daily mean wind speed and direction are computed at each facility location for the advection calculation. Days with wind speeds below 1 m/s are excluded (in contrast with the 2 m/s threshold used by Beirle et al. [3], where I lower the threshold for higher statistical power at the cost of increased noise), as weak winds produce unreliable advection estimates and allow plumes to stagnate near sources. Additionally, observations where the lifetime correction factor  $c_\tau \geq 3$  are dropped, as this exceeds the typical range of 1.2–1.8 reported by Beirle et al. Facility-years with fewer than 20 valid observation days (after wind filtering) are excluded from the satellite panel, as statistical uncertainty becomes prohibitively large with insufficient temporal sampling.

### 3.1.5 SRTM Digital Elevation Model (DEM)

The Shuttle Radar Topography Mission provides a near-global digital elevation model at 1 arc-second (30 m) resolution. I use the USGS/SRTMGL1\_003 dataset on Google Earth Engine [17, 18] to compute surface elevation gradients used in the topographic correction (Eq. 14).

### 3.1.6 Urbanization Classification

Facilities located within urban areas experience higher background NO<sub>2</sub> concentrations from traffic and other distributed sources, which adds noise to satellite-derived

emission estimates. To enable heterogeneity analysis and descriptive statistics, each facility is assigned an urbanization degree from the JRC Global Human Settlement Layer Degree of Urbanisation raster (GHS-SMOD R2023A) [19]. The SMOD classification ranges from 10 (water) through rural categories (11–13) to suburban (21) and urban categories (22–30), based on population density and built-up area from satellite imagery.

Two urbanization variables are constructed:

- `urbanization_degree`: The continuous SMOD code (10–30) at each facility location
- `in_urban_area`: A boolean flag indicating  $\text{SMOD} \geq 22$  (semi-dense urban or denser)

The threshold of 22 (semi-dense urban cluster) is chosen to distinguish facilities in genuinely urban areas from those in suburban or peri-urban settings (SMOD 21). This classification enables heterogeneity analysis comparing treatment effects across urban versus rural/suburban subsamples.

These variables are collected for heterogeneity analysis (comparing treatment effects across urban vs. rural subsamples) and descriptive statistics, *not* as regression controls. The AlphaEarth embeddings (64 dimensions) already encode land use, built-up area, and urbanization patterns implicitly. Including an explicit urbanization control would introduce multicollinearity with the embedding dimensions without improving identification, since urbanization is time-invariant and absorbed by facility fixed effects regardless. The proper causal use of urbanization is for split-sample analysis, not as an additional covariate.

### 3.2 Facility Construction: Spatial Clustering

Individual LCP plant parts may represent components of larger industrial complexes. To avoid treating co-located plants as independent units, I apply spatial clustering using a 500-meter threshold. Plants within 500m of each other are grouped into a single *facility* using a union-find algorithm.

Let  $\mathcal{P} = \{p_1, \dots, p_N\}$  denote the set of LCP plants with coordinates  $(\phi_j, \lambda_j)$  for plant  $j$ . The distance between plants  $j$  and  $k$  is computed using the WGS84 ellipsoidal approximation [20]:

$$d_{jk} \approx \sqrt{(m_\phi \cdot \Delta\phi_{jk})^2 + (m_\lambda \cdot \Delta\lambda_{jk})^2} \quad (3)$$

where the latitude scale factor follows the WGS84 series expansion:

$$m_\phi = 111,132.954 - 559.822 \cos(2\bar{\phi}) + 1.175 \cos(4\bar{\phi}) \quad [\text{m}/\text{deg}] \quad (4)$$

and the longitude scale factor varies with latitude:

$$m_\lambda = 111,132.954 \times \cos(\bar{\phi}) \quad [\text{m}/\text{deg}] \quad (5)$$

where  $\bar{\phi}$  is the mean latitude of the dataset. The latitude formula is accurate to 0.01 m per degree; the longitude formula has <0.5% error compared to the full WGS84 ellipsoidal expression. This precision is more than sufficient for identifying co-located

plants, as the 500m clustering threshold is conservative relative to the spatial extent of most industrial complexes.

Plants are grouped into facility  $i$  if they form a connected component under the relation  $d_{jk} < 500\text{m}$ . For each facility, the centroid coordinates are computed as the arithmetic mean of constituent plant coordinates:

$$(\bar{\phi}_i, \bar{\lambda}_i) = \frac{1}{|F_i|} \sum_{j \in F_i} (\phi_j, \lambda_j) \quad (6)$$

where  $F_i$  denotes the set of plants in facility  $i$ .

This clustering reduces the sample from 1,576 individual plants with ETS linkage to 932 facilities, of which 318 are multi-plant facilities.

### 3.3 Time-Varying Attributes

#### 3.3.1 Capacity and Fuel Shares

For each facility-year  $(i, t)$ , rated thermal capacity is aggregated as the sum across constituent plants:

$$\text{Capacity}_{it} = \sum_{j \in F_i} \text{Capacity}_{jt} \quad [\text{MW}] \quad (7)$$

Fuel energy consumption is similarly aggregated, then converted to fuel shares. Let  $E_{it}^{(f)}$  denote total energy consumption from fuel type  $f \in \{\text{gas, coal, oil, biomass, other}\}$  for facility  $i$  in year  $t$ , measured in terajoules (TJ). Fuel shares are computed as:

$$s_{it}^{(f)} = \frac{E_{it}^{(f)}}{\sum_{f'} E_{it}^{(f')}} \quad (8)$$

Fuel types used by fewer than 10% of facility-years (Other Solid, Peat) are dropped, shares renormalized, and facilities with no remaining fuel coverage are excluded from the sample (although there were no such facilities).

#### 3.3.2 ETS Policy Exposure Variables

The key treatment variable is the *allocation ratio*, defined as:

$$R_{it} = \frac{A_{it}}{V_{it}} \quad (9)$$

where  $A_{it}$  is total free allocation and  $V_{it}$  is verified emissions for facility  $i$  in year  $t$ , both in tCO<sub>2</sub>. Values  $R_{it} < 1$  indicate the facility must purchase additional allowances on the carbon market, representing increased policy stringency.

Facilities with allocation ratios outside the range [0.01, 20] are excluded as likely data errors or non-operating installations.

### 3.4 Satellite NO<sub>x</sub> Emission Proxy: Beirle-Style Flux-Divergence

The satellite outcome variable is constructed using a simplified Beirle-style flux-divergence method, following the approach developed by Beirle et al. [1–3]. This method provides physically grounded NO<sub>x</sub> emission estimates by exploiting the relationship between wind-driven advection and local emissions.

#### 3.4.1 Identification versus Quantification

Beirle et al.’s v2 catalog combines two distinct algorithmic components: (i) an automatic point-source *identification* algorithm that locates emission maxima in the global advection field, and (ii) a *quantification* method that estimates emission rates by spatially integrating advection around each identified source. Crucially, the authors note that “the quantification of NO<sub>x</sub> emissions by spatial integration of the corrected advection map could be applied to these locations or *any other known point source*” [3].

In this study, I *skip the identification step* because I already have a curated set of ETS/LCP facilities with reliable coordinates from the European Environment Agency registry. I apply Beirle’s quantification method directly to these known source locations.

To guard against treating noise as signal, I implement *simplified significance flags* that parallel Beirle’s catalog selection criteria:

- **Detection limit:** I use a permissive threshold of 0.01 kg/s to maximize statistical power, acknowledging this is well below Beirle’s validated thresholds (0.04 kg/s for desert, 0.11 kg/s elsewhere). This choice trades off signal quality for sample size in econometric analysis—the resulting estimates should be interpreted as conservative (attenuated toward zero) due to measurement noise.
- **Statistical integration error:** Facilities with >30% relative statistical uncertainty in the spatial integration are flagged.
- **Spatial interference:** Facilities with another ETS facility within 5 km are flagged, as their satellite outcome may reflect cluster-level rather than single-facility emissions.

These flags are used in sensitivity analyses rather than for hard filtering, preserving the full panel while allowing transparent restriction to “significant” satellite observations.

#### 3.4.2 Advection Formulation

The advection  $A$  is defined as the scalar product of wind velocity and the spatial gradient of NO<sub>2</sub> tropospheric vertical column density (TVCD):

$$A = \mathbf{w} \cdot \nabla V = u \frac{\partial V}{\partial x} + v \frac{\partial V}{\partial y} \quad (10)$$

where  $\mathbf{w} = (u, v)$  is the horizontal wind vector (m/s) from ERA5-Land and  $V$  is the NO<sub>2</sub> TVCD (molec/m<sup>2</sup>). Under the continuity equation, this advection is proportional to local emissions minus chemical sinks.

For each facility  $i$  and day  $d$ , spatial gradients are computed on a local grid (30 km × 30 km centered on the facility) using finite differences on the TROPOMI L3 lat–lon grid:

$$\frac{\partial V}{\partial x} \approx \frac{V(x + \Delta x, y) - V(x - \Delta x, y)}{2\Delta x} \quad (11)$$

$$\frac{\partial V}{\partial y} \approx \frac{V(x, y + \Delta y) - V(x, y - \Delta y)}{2\Delta y} \quad (12)$$

where  $\Delta x$  and  $\Delta y$  correspond to the TROPOMI grid resolution (approximately 3.5 km × 5.5 km). This differs from Beirle et al., who compute derivatives on the native TROPOMI pixel grid to handle cloud-induced gaps; the L3 gridded product used here introduces additional smoothing and potential artifacts.

### 3.4.3 NO<sub>2</sub> to NOx Scaling

TROPOMI measures NO<sub>2</sub>, but NOx emissions include both NO and NO<sub>2</sub>. Following Beirle et al. [3], I apply a scaling factor  $c_{\text{NOx}}$  based on the photostationary state (PSS):

$$c_{\text{NOx}} = \frac{[\text{NOx}]}{[\text{NO}_2]} = 1 + \frac{J}{k[\text{O}_3]} \quad (13)$$

where  $J$  is the NO<sub>2</sub> photolysis frequency (parameterized as  $0.0167 \times \exp(-0.575/\cos(\text{SZA}))\text{s}^{-1}$ , where SZA is the Solar Zenith Angle of the observation),  $k$  is the reaction rate constant for NO + O<sub>3</sub> ( $2.07 \times 10^{-12} \times \exp(-1400/T)\text{ cm}^3 \text{ molec}^{-1} \text{ s}^{-1}$ ), and [O<sub>3</sub>] is taken from an ozone climatology. For detected point sources, Beirle et al. report a typical NOx/NO<sub>2</sub> ratio of approximately  $1.38 \pm 0.10$ .

Following Beirle et al., I apply a fixed scaling factor of  $c_{\text{NOx}} = 1.38$  with uncertainty  $\pm 0.10$  (approximately 7% relative uncertainty), which represents the empirically observed mean ratio across detected point sources.

### 3.4.4 Topographic Correction

Over mountainous terrain, 3D radiative transfer effects cause systematic artifacts in the advection field. Following Beirle et al. [3] Sect. 3.7, I apply a topographic correction:

$$A^* = A + f \cdot C_{\text{topo}}, \quad C_{\text{topo}} = \frac{V}{H_{\text{sh}}} \cdot (\mathbf{w}_0 \cdot \nabla z_0) \quad (14)$$

where  $V$  is the NO<sub>2</sub> TVCD,  $H_{\text{sh}} = 1$  km is the assumed NOx scale height,  $\mathbf{w}_0 \cdot \nabla z_0$  is the dot product of the surface wind vector and the surface elevation gradient (from the SRTM Digital Elevation Model (DEM) [17, 18] via Google Earth Engine), and  $f = 1.5$  is an empirically derived scaling factor (Appendix A of Beirle et al.). The

combined effect yields an effective scale height of  $1/1.5 = 667$  m. For flat terrain typical of European power plant locations, this correction is small.

### 3.4.5 Spatial Integration and Lifetime Correction

For each facility, the raw emission rate is computed by spatially integrating the topography-corrected advection  $A^*$  over a disc around the facility location (Beirle Eq. 11):

$$E_{\text{raw}} = \iint_{r \leq R} A^*(x, y) dx dy \approx \sum_i A_i^* \times \Delta x \Delta y \quad [\text{mol/s}] \quad (15)$$

where  $A^*$  has units  $\text{mol}/(\text{m}^2 \cdot \text{s})$  and the spatial integration is realized by summing the advection values multiplied by the pixel area for all grid pixels within the radius  $R$ .

Beirle et al. [3] use  $R = 15$  km as a compromise between capturing the full point source signal and avoiding interference from neighboring sources. However, in this study I adopt a 5 km integration radius based on validation against reported NOx emissions from the LCP Directive. At 15 km, satellite-derived NOx showed *negative* correlation with ground-truth reported NOx, suggesting contamination from nearby sources or background noise. Reducing the radius to 5 km improved the correlation: for non-urban, non-interfered facilities above the detection limit,  $r = 0.25$  (levels) and  $r = 0.37$  (log-log,  $n = 266$ ). The full sample correlation remains weak ( $r = 0.03$ ) due to urban interference and measurement noise, but the improvement in clean subsamples indicates better facility-level attribution. Detection limits are scaled accordingly (0.01 kg/s vs Beirle's 0.03/0.11 kg/s) to account for the smaller integration area.

Chemical loss of NOx during transport within the integration radius requires a lifetime correction. The residence time within the radius is:

$$t_r = \frac{R}{|\mathbf{w}|} \quad (16)$$

where  $R = 5$  km (in this study) and  $|\mathbf{w}|$  is the mean wind speed. The lifetime correction factor, following Beirle et al. [3] Eq. (9), is:

$$c_\tau = \exp(t_r/\tau) \quad (17)$$

where  $\tau$  is the effective NOx lifetime, parameterized as a function of latitude following Lange et al. [21] via Beirle et al. Eq. (10):

$$\tau(\text{lat}) = 1.0089 \times \exp(0.0242 \times (|\text{lat}| + 9.6024)) \quad [\text{hm/s}] \quad (18)$$

with typical values of 2 h at low latitudes to 4–6 h at higher latitudes. For detected point sources, the resulting  $c_\tau \approx 1.40 \pm 0.24$ . Following Beirle et al., I assume 50% relative uncertainty in  $\tau$  due to high variability at similar latitudes.

### 3.4.6 Final NOx Emission Estimate

The final satellite-derived NOx emission rate for facility  $i$  and day  $d$  is:

$$E_{\text{NOx},id} = c_\tau \cdot c_{\text{NOx}} \cdot E_{\text{raw},id} \quad (19)$$

Converting from mol/s to kg/s using the molar mass of NO<sub>2</sub> (46.0055 g/mol) [22]. Annual estimates are computed as the mean over all valid observation days.

### 3.4.7 Uncertainty Components

Following Beirle et al. [3] Sect. 3.12, the satellite-derived NOx estimates carry uncertainty from multiple sources, combined in quadrature. Table 1 summarizes components, implementation status, typical magnitudes, and expected directional effects.

**Table 1** Uncertainty components for satellite-derived NOx estimates (following Beirle et al. Sect. 3.12). Components are combined in quadrature.

Source	Impl.?	Magnitude	Notes
Statistical error	Yes	<10%	SE of temporal mean; captures met. variability.
Lifetime $c_\tau$	Yes	10–20%	50% rel. $\sigma_\tau$ ; propagated multiplicatively.
NOx/NO <sub>2</sub> scaling	Yes	~7%	Fixed PSS ratio ( $1.38 \pm 0.10$ ).
AMF correction	No	10%	Unmodelled; likely downward bias.
Plume height	No	10%	No height-dependent wind; level bias.
Topo. correction	Yes	<2.5%	$f = 1.5$ ; small for flat terrain.
OFFL vs PAL	No	systematic	OFFL 10–40% lower; conservative estimates.

Beirle et al. report total uncertainties in the 20–40% range. My implementation achieves similar typical totals of ~20–30%. Note that the OFFL vs PAL product difference is a systematic bias (not random uncertainty) and is therefore excluded from error propagation; this means my emission estimates are conservative (biased low by 10–40%).

Observations with total relative uncertainty exceeding 50% are excluded from the satellite panel, as high-uncertainty observations add noise without proportional information content. For the remaining observations, I construct inverse-variance weights  $w_i = 1/\sigma_i^2$  (capped at the 99th percentile to limit extreme weights), which are used as robustness checks via weighted least squares estimation.

### 3.5 Sample Construction

The final analysis sample is constructed by applying the following filters to both outcomes:

1. Facilities must have valid ETS linkage (matched normalized identifier)
2. Allocation ratio in [0.01, 20] range
3. At least 3 years of complete data within 2018–2023

The resulting base analysis panel contains 521 facilities observed over 2,819 facility-years. This panel is used directly for the verified CO<sub>2</sub> outcome.

For the satellite NOx outcome, additional attrition occurs due to:

1. Non-missing satellite outcome (requires  $\geq 20$  valid observation days per year)
2. Total relative uncertainty  $\leq 50\%$
3. Passing significance thresholds (detection limit, statistical error)

Sample attrition details are provided in Appendix C. The most significant losses occur at the ETS linkage stage (44% of plants lack matched ETS identifiers in the EU Registry crosswalk) and the requirement for matched ETS compliance data.

Table 2 summarizes the sample characteristics for the base analysis panel.

### 3.6 Geographic Context: AlphaEarth Embeddings

A key methodological contribution of this study is the incorporation of high-dimensional geospatial foundation model embeddings as control variables, following the recent trend toward using learned representations for causal inference [4, 5]. I use Google AlphaEarth Foundations [7], a geospatial embedding field model that produces 64-dimensional representations from multi-source satellite imagery (Sentinel-1/2, Landsat), climate reanalysis (ERA5-Land), topography (GLO-30), and geotagged text (Wikipedia, GBIF). The model is trained using contrastive learning objectives to capture information predictive of diverse downstream tasks—from land cover classification to biophysical variable estimation—without being tuned for any specific application.

For each facility location, the embedding vector  $e_i \in \mathbb{R}^{64}$  is extracted by averaging the embedding field over a disc centered on the facility. The averaging radius is set equal to the integration radius used for satellite NOx quantification (Section 3.4), ensuring that the geographic controls capture contextual information at the same spatial scale as the satellite outcome. These embeddings encode:

- **Land use context:** Urban density, industrial areas, agricultural patterns
- **Infrastructure:** Road networks, built environment characteristics
- **Vegetation:** Forest cover, cropland, seasonal phenology
- **Climate:** Local temperature, precipitation, insolation, and wind patterns
- **Topography:** Elevation, slope, and terrain characteristics

The embedding dimensions are included as controls in the econometric specifications, providing a data-efficient approach to capturing between-unit heterogeneity arising from local geographic context. The approach is particularly relevant for

difference-in-differences settings where high-dimensional spatial confounders may induce violations of parallel trends if left uncontrolled [6]—for example, if facilities in different geographic contexts (coastal versus inland, urban versus rural) experience different secular trends in air quality unrelated to policy.

The embeddings are extracted annually from the `GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL` ImageCollection, yielding time-varying controls that capture year-to-year changes in land use, vegetation phenology, and built environment. This temporal variation is methodologically important: unlike static facility-level controls (which would be absorbed by facility fixed effects), annual embeddings can adjust for time-varying geographic confounders that might bias the satellite NOx outcome.

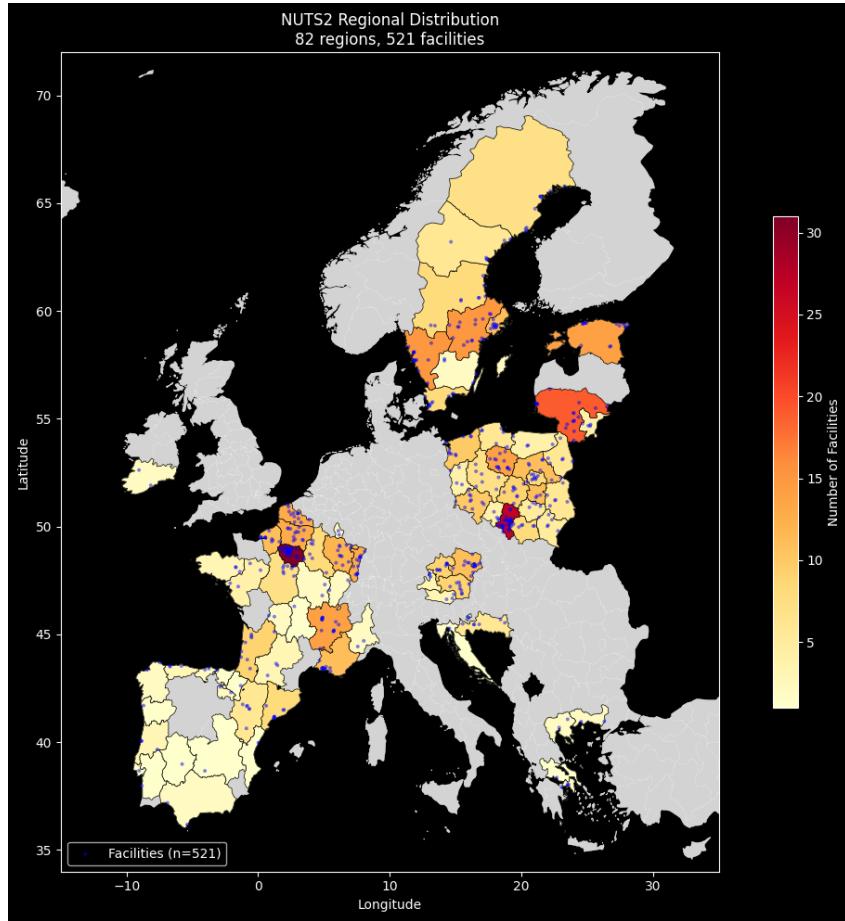
To address overfitting concerns with 64 dimensions, I apply dimensionality reduction (PCA or PLS to 10 components) as detailed in Section 4.2.7. For PLS, the projection is trained on facility-level mean NOx (cross-sectional) rather than panel observations, ensuring the reduced embeddings are time-invariant and equivalent to pre-treatment covariates—this prevents regularization bias from outcome snooping [4]. As discussed in Section 4.2.6, embeddings are applied only to the satellite NOx outcome.

## 3.7 Exploratory Data Analysis

This section presents descriptive statistics and visualizations of the analysis panel, providing context for the econometric analysis.

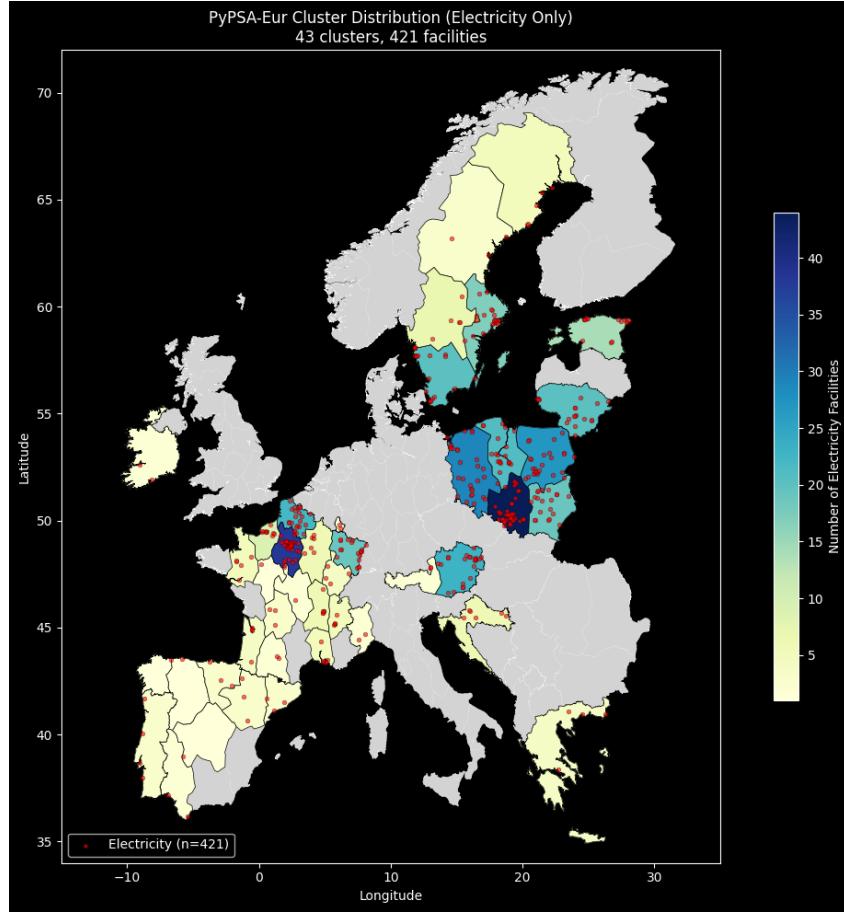
### 3.7.1 Geographic Distribution

Figure 1 displays the geographic distribution of facilities across NUTS2 regions. The sample spans 82 NUTS2 regions across Europe, with the highest concentrations in Germany, Poland, and Spain. The heatmap shading indicates the number of facilities per region, with densities ranging from 1–30 facilities per region.



**Fig. 1** Geographic distribution of 521 facilities across 82 NUTS2 regions. Color intensity indicates facility count per region. Blue points mark individual facility locations.

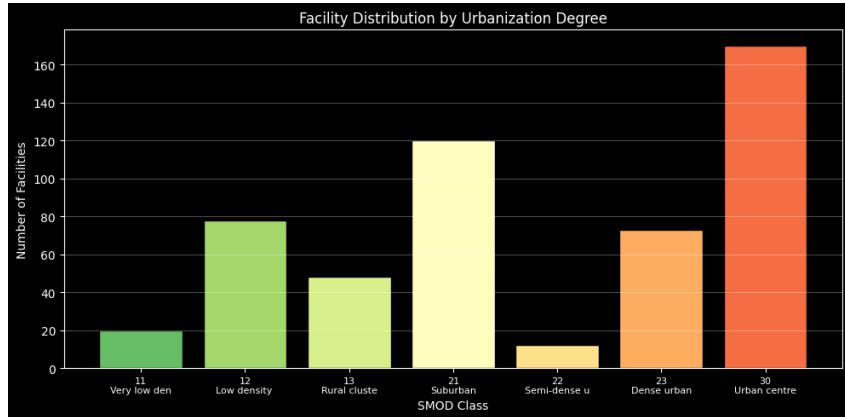
For electricity sector heterogeneity analysis, Figure 2 shows the distribution of electricity-generating facilities across PyPSA-Eur power system clusters. The 421 electricity facilities (those with EU ETS activity codes 1 or 20) are distributed across 43 network-derived clusters, with particularly high concentrations in central European clusters covering Germany, Poland, and the Czech Republic.



**Fig. 2** Distribution of 421 electricity-generating facilities across 43 PyPSA-Eur power system clusters. Clusters are derived from k-means clustering on transmission network topology, grouping facilities facing correlated wholesale prices and grid constraints.

### 3.7.2 Urbanization Context

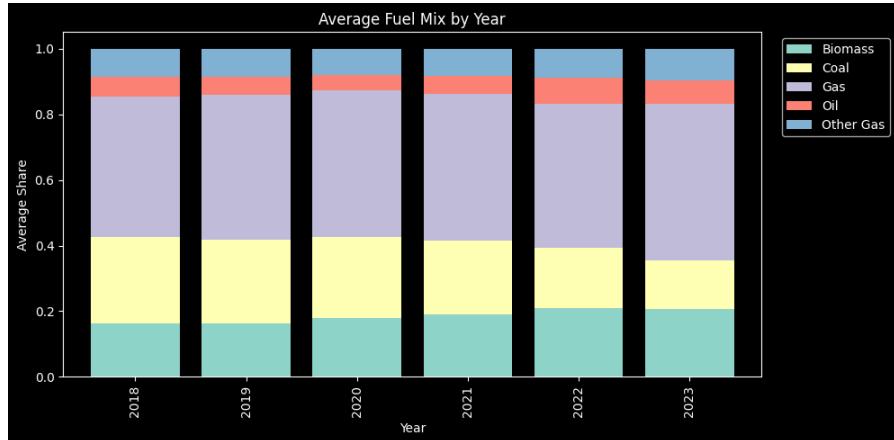
Facility urbanization context is captured via the GHSL-SMOD classification. Figure 3 shows the distribution of facilities across urbanization categories. A substantial majority of facilities (approximately 60%) are located in suburban to urban-center areas ( $\text{SMOD} \geq 21$ ), reflecting the tendency for large combustion plants to be sited near population centers for district heating and electricity distribution. This urban concentration implies elevated background  $\text{NO}_2$  levels that add noise to satellite-derived emission estimates.



**Fig. 3** Distribution of facilities by urbanization degree (GHSL-SMOD classification). Categories range from very low density rural (11) to urban centers (30). The majority of facilities are in suburban (21) and urban center (30) locations.

### 3.7.3 Fuel Mix and Capacity

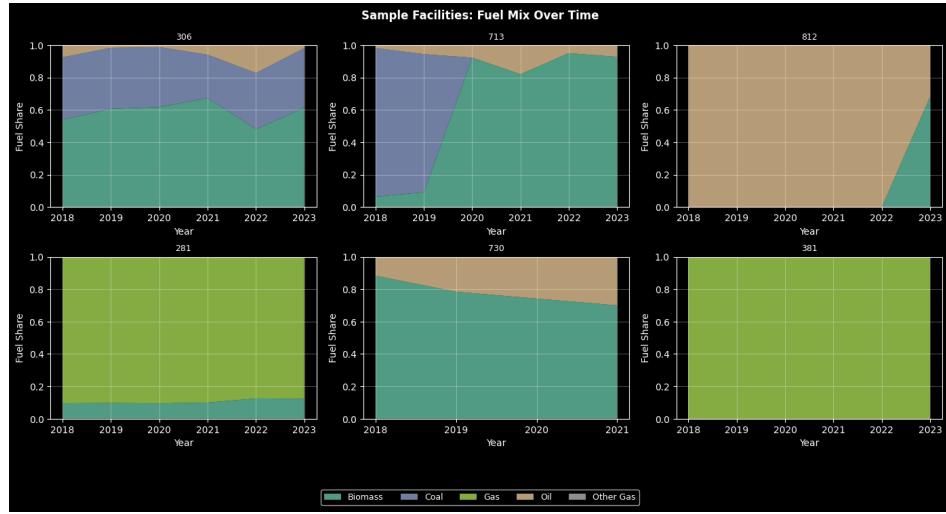
Figure 4 presents the average fuel mix across the sample period. Natural gas dominates (approximately 45% of energy input), followed by coal (approximately 20–25%) and biomass (approximately 15%). The coal share shows a modest decline from 2018 to 2023, consistent with the broader European transition away from coal-fired generation. Biomass and gas shares increase correspondingly, reflecting fuel switching in response to carbon pricing.



**Fig. 4** Average fuel mix by year across all facilities. Gas (purple) dominates, with coal (tan) showing a modest decline over the sample period. Biomass (teal) and other gas (blue) shares increase slightly.

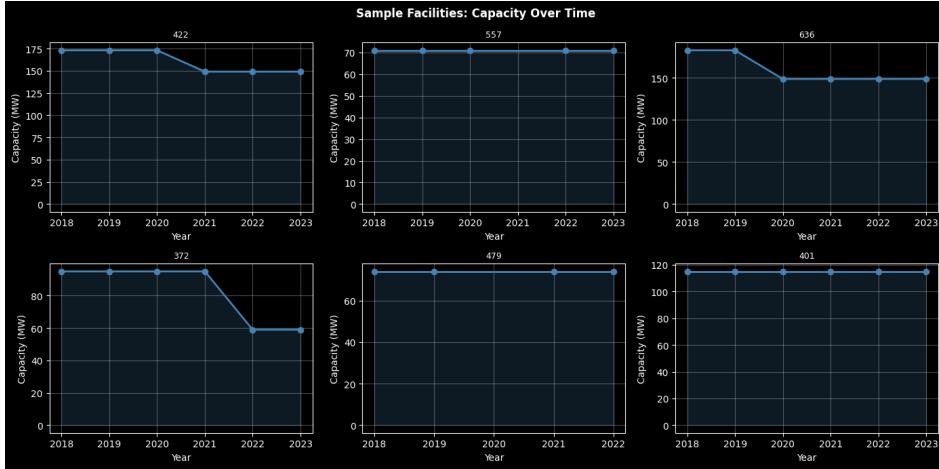
Figure 5 illustrates within-facility fuel mix dynamics for a random sample of six facilities. Several facilities exhibit substantial fuel switching—for example, facility 812

transitions from primarily coal to primarily gas between 2022 and 2023, while facility 713 shifts from nearly 100% biomass to predominantly gas. These within-facility transitions represent the variation exploited by the panel fixed effects specifications.



**Fig. 5** Fuel mix evolution for six randomly sampled facilities. Stacked area charts show year-over-year changes in fuel shares. Notable fuel switching is visible in facilities 812 (coal to gas) and 713 (biomass to gas).

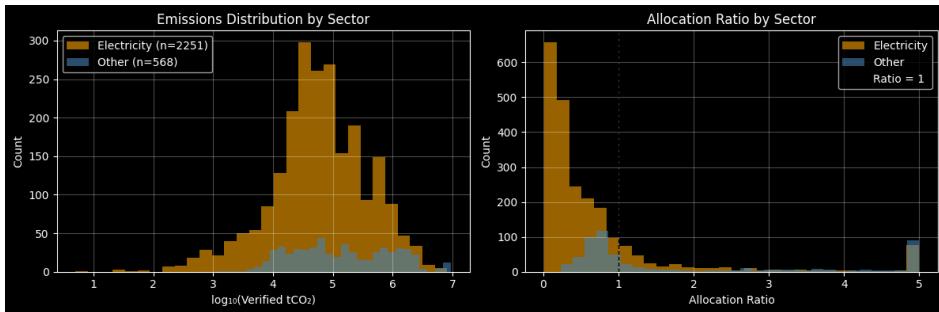
Figure 6 shows capacity trajectories for a sample of facilities. Most facilities exhibit stable capacity over the sample period, with occasional step changes reflecting plant upgrades, partial closures, or measurement corrections. Facility 372 shows a notable capacity reduction from approximately 90 MW to 60 MW between 2021 and 2022.



**Fig. 6** Rated thermal capacity (MW) over time for six randomly sampled facilities. Most facilities exhibit stable capacity with occasional step changes.

### 3.7.4 ETS Policy Exposure

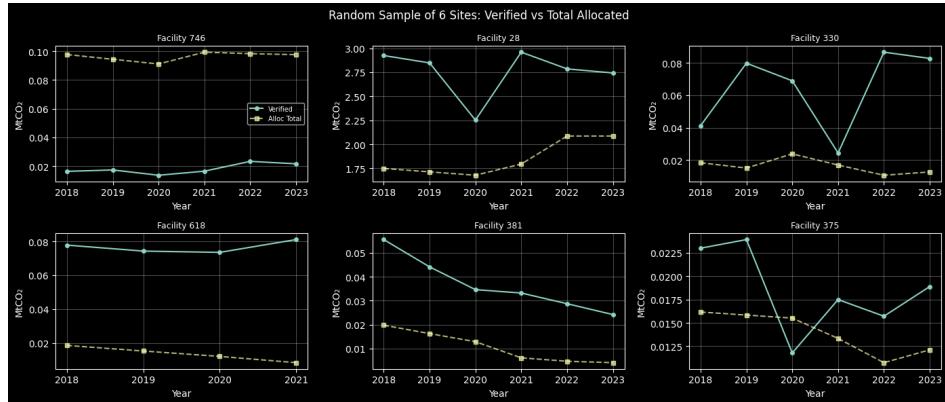
Figure 7 presents the distribution of verified emissions and allocation ratios by sector. The left panel shows log-transformed verified CO<sub>2</sub> emissions, with electricity-sector facilities ( $n=2,251$  facility-years) exhibiting substantially higher emissions than other sectors ( $n=568$  facility-years). The distribution is approximately log-normal with a mode around  $10^5$  tCO<sub>2</sub>/yr. The right panel shows allocation ratios, with a concentration near zero for electricity-sector facilities (reflecting the phase-out of free allocation to the power sector under EU ETS Phase III/IV) and a wider distribution for industrial facilities that retain carbon leakage protection.



**Fig. 7** Distribution of verified emissions (left) and allocation ratios (right) by sector. Electricity facilities (orange) have higher emissions but lower allocation ratios due to reduced free allocation under EU ETS Phase III/IV.

Figure 8 illustrates the relationship between verified emissions and free allocations for a sample of facilities. For most facilities, verified emissions (solid lines) consistently

exceed free allocations (dashed lines), indicating shortfall positions requiring allowance purchases. The gap between verified and allocated represents the policy stringency experienced by each facility.



**Fig. 8** Verified emissions (teal solid) versus free allocations (yellow dashed) for six randomly sampled facilities. The persistent gap above the allocation line indicates shortfall positions requiring market purchases.

### 3.7.5 Satellite NOx Outcome

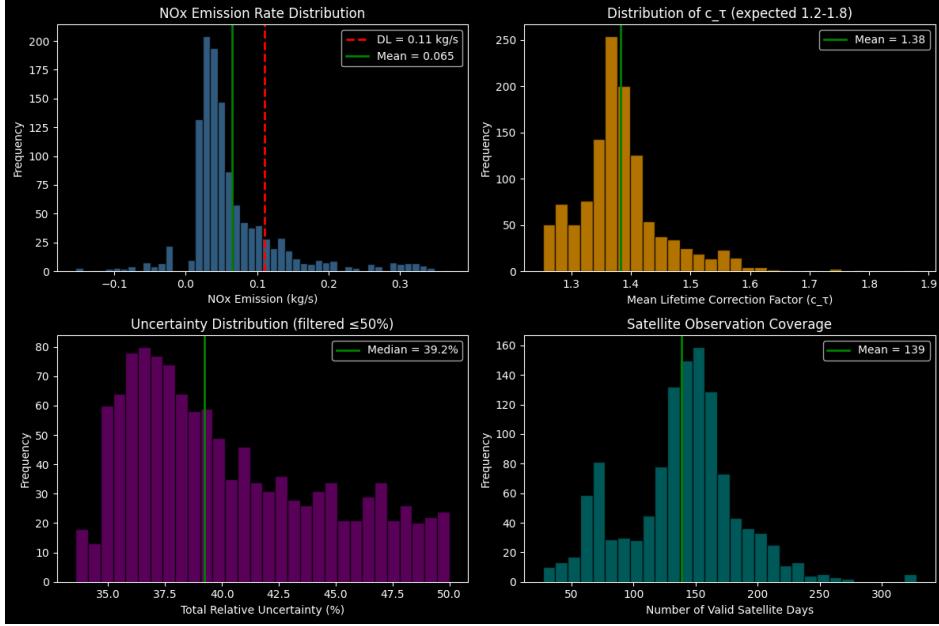
Figure 9 presents the key characteristics of the satellite-derived NOx emission estimates. After applying the  $\leq 50\%$  total uncertainty filter required for inclusion in the satellite panel, 381 facilities (1,430 facility-years) remain from the base panel of 521 facilities.

The top-left panel shows the distribution of estimated NOx emission rates. The mean emission rate is 0.013 kg/s. Most facilities in the sample have emissions near or above the permissive detection limit of 0.01 kg/s used in this study for statistical power (well below Beirle's validated 0.04–0.11 kg/s thresholds). Negative estimates (statistical noise) are present for the lowest emitters.

The top-right panel displays the lifetime correction factor ( $c_\tau$ ) distribution, with a mean of 1.16. This factor accounts for NOx chemical decay during atmospheric transport and is computed from the latitude-dependent lifetime parameterization of Lange et al. (2022). The narrow distribution confirms that European facilities fall within the expected mid-latitude range.

The bottom-left panel shows the total relative uncertainty distribution, with a median of approximately 30%. This uncertainty combines statistical integration error, lifetime correction uncertainty ( $\pm 50\%$ ), NO<sub>2</sub>/NOx ratio uncertainty ( $\pm 7\%$ ), and unmodeled AMF/plume height terms ( $\pm 10\%$  each). Note that the OFFL vs PAL product difference (10–40% lower TVCDs) is a systematic bias rather than random uncertainty, so it is excluded from error propagation; this makes my emission estimates conservative. I use the uncertainty to add an inverse-variance weight ( $= 1/\sigma_{\text{rel}}^2$ ) to my regression terms.

The bottom-right panel shows satellite observation coverage, with a mean of 149 valid observation days per facility-year. This exceeds the minimum threshold of 20 days and provides substantial temporal averaging to reduce noise. Coverage varies due to cloud screening, wind speed filtering ( $\geq 1$  m/s), and satellite orbit patterns.



**Fig. 9** Satellite-derived NOx emission characteristics for 381 facilities (1,430 facility-years). Top-left: NOx emission rate distribution with detection limit (0.01 kg/s) and mean (0.013 kg/s). Top-right: Lifetime correction factor distribution (mean 1.16). Bottom-left: Total relative uncertainty (median 30.3%). Bottom-right: Valid satellite observation days (mean 149).

### 3.7.6 Validation of Satellite NOx vs EU LCP Reports

To assess measurement quality, I correlate satellite-derived NOx with administratively reported NOx emissions from the LCP Directive. Across all observations above the detection limit ( $n = 1,317$ ), the correlation is weak ( $r = 0.03$ ). However, correlation improves substantially when restricting to cleaner measurement conditions: for non-urban facilities ( $r = 0.10$  log-log,  $n = 501$ ), and for non-urban facilities without spatial interference ( $r = 0.37$  log-log,  $n = 266$ ). This pattern suggests that the satellite proxy captures true emissions signal, but urban background pollution and interference from nearby sources introduce substantial noise. The modest validation correlation—even under ideal conditions—underscores that the satellite outcome is a noisy proxy, best suited for corroborating ground-truth administrative data rather than standalone causal inference.

**Table 2** Summary Statistics for Base Analysis Panel

Variable	Mean	Std. Dev.	Min	Max
<i>Panel Structure</i>				
Facilities		521		
Facility-years		2,819		
Years per facility	5.4	1.0	3	6
Electricity sector facilities		421 (80.8%)		
NUTS2 regions		82		
<i>Verified CO<sub>2</sub> Emissions</i>				
Verified emissions (ktCO <sub>2</sub> /yr)	580	1,240	0.5	7,500
Log verified emissions	11.5	1.8	6.2	15.8
<i>ETS Policy Variables</i>				
Allocation ratio	0.62	0.85	0.01	18.5
<i>Plant Characteristics</i>				
Capacity (MW)	780	1,120	50	6,800
Gas share	0.44	0.42	0	1
Coal share	0.19	0.34	0	1
Biomass share	0.16	0.33	0	1
Oil share	0.13	0.28	0	1
<i>Urbanization</i>				
In urban area (SMOD $\geq$ 22)		60.3%		
Interfered (facility within 5km)		36.6%		
<i>Satellite NOx Panel (subset)</i>				
Facilities		381		
Facility-years		1,430		
NOx emission rate (kg/s)	0.013	0.011	-0.01	0.07
Above detection limit (0.01 kg/s)		92%		
Median total uncertainty		30.3%		
Mean valid satellite days		149		

Note: Base panel includes all EU ETS-regulated large combustion plants with matched compliance data and  $\geq 3$  years of observations during 2018–2023. Satellite NOx panel is restricted to facility-years with  $\leq 50\%$  total uncertainty and  $\geq 20$  valid observation days.

## 4 Methodology

This section describes the causal inference framework and econometric specifications used to estimate the effect of EU ETS policy stringency on both verified CO<sub>2</sub> emissions and satellite-derived NOx emission proxies.

### 4.1 Causal Framework

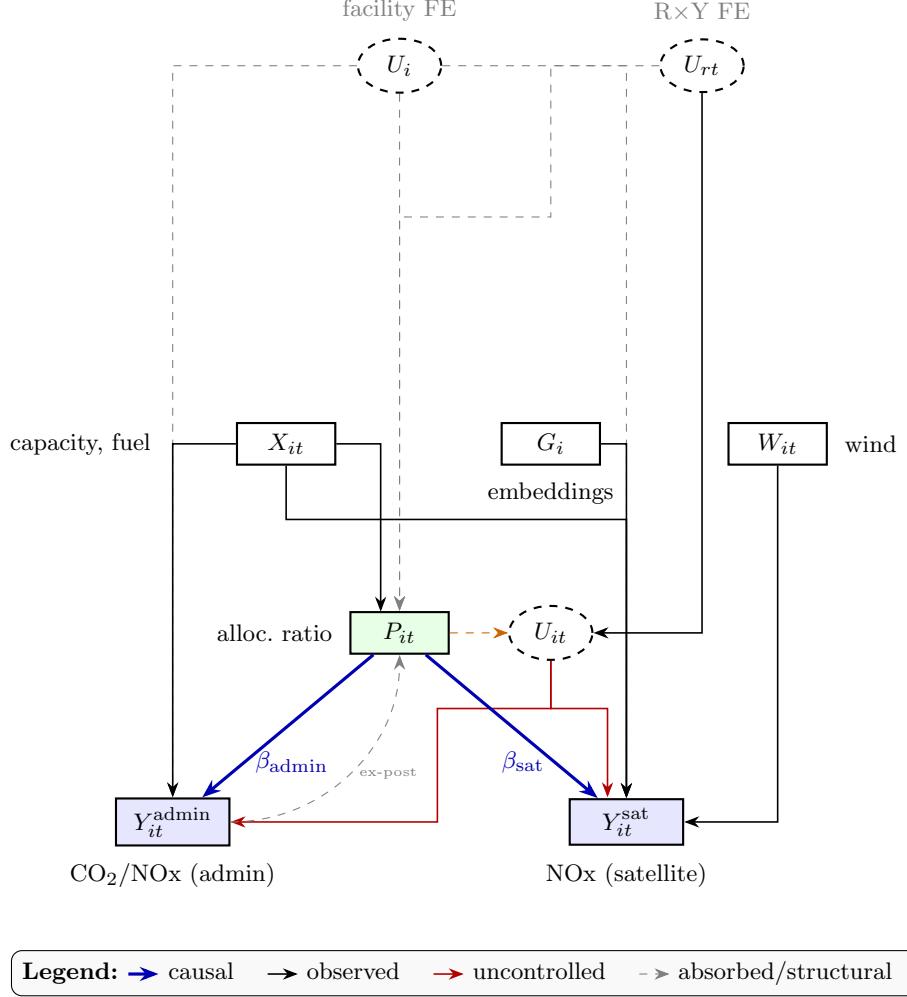
The goal is to estimate the causal effect of ETS policy stringency on two complementary outcomes. Let  $Y_{it}^{\text{CO}_2}$  denote verified CO<sub>2</sub> emissions (ktCO<sub>2</sub>/yr) for facility  $i$  in year  $t$ , and let  $Y_{it}^{\text{NOx}}$  denote the satellite-derived NOx emission proxy (kg/s). Let  $R_{it}$  denote the allocation ratio (treatment intensity).

The key identification challenge is that allocation ratios are not randomly assigned. Facilities with high emissions relative to historical benchmarks receive lower allocation ratios, creating potential endogeneity: unobserved factors affecting both emissions intensity and local air quality may confound the relationship. Additionally, allocation ratios co-move with operational decisions (capacity utilization, fuel switching) that directly affect emissions.

The directed acyclic graph (DAG) in Figure 10 illustrates the causal structure. The target estimand is the effect of  $P_{it}$  on  $Y_{it}$ , controlling for confounders. Key confounding pathways include:

- **Facility-level time-invariant unobservables** ( $U_i$ ): Plant technology, combustion efficiency, and location affect both policy exposure and emissions. Absorbed by facility fixed effects.
- **Time-varying regional factors** ( $U_{rt}$ ): Electricity demand, fuel prices, and regional economic conditions affect plant operations and allocation ratios. Absorbed by Region  $\times$  Year fixed effects.
- **Plant-level time-varying unobservables** ( $U_{it}$ ): Dispatch/utilization, maintenance status, and operational efficiency changes affect both verified emissions (determining allocation ratios) and pollutant output. This is the key identification challenge—see Section 4.2.
- **Observed operational factors** ( $X_{it}$ ): Capacity and fuel mix affect both verified emissions and pollutant emissions. Controlled directly.

The Beirle-style flux-divergence approach addresses atmospheric confounding by focusing on the spatial gradient (advection) rather than absolute column densities, making it sensitive to local emissions rather than background concentrations. Fixed effects absorb facility-level and time-varying regional confounders for both outcomes.



**Fig. 10** Directed acyclic graph for dual-outcome causal inference. Treatment  $P_{it}$  (allocation ratio) affects both verified CO<sub>2</sub> and satellite NOx outcomes. Gray dashed arrows from  $U_i$  and  $U_{rt}$  are absorbed by facility and region×year fixed effects, respectively. Red arrows from  $U_{it}$  indicate residual confounding from time-varying unobservables (dispatch, maintenance) that I intentionally leave uncontrolled to preserve the mediation pathway (orange). The ex-post arrow reflects that allocation ratios are mechanically computed from prior verified emissions. Observed controls  $X_{it}$  (capacity, fuel) affect both outcomes;  $G_i$  (embeddings) and  $W_{it}$  (wind) affect only the satellite outcome.

## 4.2 Identification Strategy and Variable Selection

The identification strategy relies on two key choices: (i) what to control for, and (ii) what *not* to control for. Both are essential to avoid bias.

### 4.2.1 Region×Year Effects

Regional electricity demand, fuel prices, and economic conditions create time-varying confounding: demand shocks increase plant utilization, raising both verified emissions (lowering  $R_{it}$ ) and NO<sub>2</sub> output. Without adjustment, this creates spurious correlation between policy stringency and pollution.

Region×Year fixed effects absorb these common shocks additively. The identifying variation becomes: *within the same region and year, do facilities with different allocation ratios exhibit different NO<sub>2</sub> enhancement?* This comparison holds regional conditions constant while exploiting cross-facility variation in policy exposure.

I use NUTS2 regions (Nomenclature of Territorial Units for Statistics, level 2) from Eurostat for clustering. NUTS2 regions (~200–300 across the EU) define economically coherent regional units that share common labor markets, policy enforcement mechanisms, and infrastructure characteristics. Unlike PyPSA-Eur power system clusters (which are appropriate for electricity sector heterogeneity analysis), NUTS2 regions apply to all industrial facility types—power plants, refineries, cement plants—and correspond to administrative units where regional policies are implemented and enforced. This makes them appropriate for absorbing regional time-varying confounders that affect facilities regardless of their sectoral activity.

### 4.2.2 The EU ETS Compliance Calendar

To understand why the allocation ratio is a valid treatment variable, I first make the compliance calendar explicit. The annual cycle proceeds as follows:

- **28 February (year  $t$ ):** Free allocation  $A_{it}$  issued to each installation based on predetermined benchmarks [23].
- **Throughout year  $t$ :** Facilities make operational decisions—dispatch, fuel choice, maintenance—knowing their allocation  $A_{it}$ .
- **31 March (year  $t+1$ ):** Accredited verifiers audit and report verified emissions  $V_{it}$ .
- **30 April (year  $t+1$ ):** Facilities surrender allowances equal to verified emissions.

The key insight: **the denominator of  $R_{it}$  (verified emissions  $V_{it}$ ) is an outcome of decisions made after the allocation  $A_{it}$  is known.** The allocation ratio  $R_{it} = A_{it}/V_{it}$  is therefore a function of the policy parameter (free allocation), and any feedback from current emissions to next-period allocation is absorbed by facility fixed effects and year fixed effects. This timing structure eliminates contemporaneous simultaneity.

### 4.2.3 Why I Do Not Control for Generation/Dispatch

Facility-level dispatch and generation represent a “bad control” problem that requires explicit justification. Dispatch is simultaneously:

1. **A confounder:** Demand shocks → higher dispatch → higher verified emissions  $V_{it}$  → lower allocation ratio  $R_{it}$ . The same shocks → more combustion → higher NO<sub>2</sub> output.
2. **A mediator:** If policy affects merit order bidding (facilities with carbon shortfalls bid higher → get dispatched less), then:  $R_{it} \rightarrow \text{dispatch} \rightarrow Y_{it}$ .

Including generation data would block part of the causal path and bias the policy effect toward zero. Region $\times$ Year fixed effects absorb the common regional component of dispatch variation (regional demand, fuel prices, carbon prices), leaving only facility-specific deviations as residual confounding. I do not control for dispatch directly to preserve the mediation pathway.

#### 4.2.4 Attenuation Bias and Conservative Interpretation

To the extent that endogenous dispatch variation contaminates  $R_{it}$  through the denominator, the bias is likely *attenuating*: facilities with high dispatch have both lower allocation ratios (higher denominator) and higher emissions, creating positive correlation between  $R_{it}$  and  $Y_{it}$  that works against finding a negative policy effect. **Estimates should therefore be interpreted as conservative bounds on the true policy effect.**

#### 4.2.5 Residual Threats and Interpretation

The primary residual threat is facility-specific time-varying confounding ( $U_{it}$ )—maintenance outages, unexpected efficiency changes, or demand for a specific plant’s output (dispatch). These are unlikely to systematically correlate with allocation ratios conditional on my controls. Future work incorporating plant-level generation data could address this directly; inclusion of economic dispatch and power system optimization (potentially also from PyPSA-EUR) is reserved for subsequent analysis.

#### 4.2.6 Outcome-Specific Controls

Two variables affect only the satellite NOx outcome, not verified ETS CO<sub>2</sub>:

- **Wind ( $W_{it}$ ):** The Beirle flux-divergence method uses wind speed and direction to compute advected NO<sub>2</sub> mass flux. Wind enters the satellite *measurement process*—it does not affect actual emissions or administrative reporting.
- **AlphaEarth embeddings ( $G_i$ ):** Geographic context (terrain, land use, climate) affects satellite retrieval quality—terrain influences air mass factor corrections; urban land use creates background NO<sub>2</sub> that adds noise to point-source signals; climate affects atmospheric dispersion and NOx lifetime. None of these affect the administrative mass-balance calculation underlying ETS CO<sub>2</sub>.

For ETS CO<sub>2</sub>, geographic confounders are absorbed by facility fixed effects (time-invariant factors like location, baseline technology) and region $\times$ year fixed effects (time-varying regional factors). Including embeddings for the ETS outcome would control for variation irrelevant to that measurement process.

#### 4.2.7 Embedding Dimensionality Reduction

The raw AlphaEarth embeddings (64 dimensions) may introduce overfitting concerns in the TWFE specification, particularly when the panel contains limited within-facility variation. Two dimensionality reduction strategies are considered:

**PCA (unsupervised):** Standard principal component analysis projects embeddings onto directions that maximize variance in the embedding space. This is causally

safe because it does not use outcome information—the projection is determined entirely by the covariate distribution.

**Facility-level PLS (supervised):** Partial least squares regression projects embeddings onto directions that predict the outcome. However, naive application of PLS to panel data creates *regularization bias*: the learned projection incorporates information from year-specific outcome shocks, violating the requirement that controls be pre-determined [4]. This is analogous to the “bad controls” problem identified by [24]: if the projection learns to predict treatment-affected variation in the outcome, controlling for the reduced embeddings biases the treatment effect estimate.

I address this by training PLS on facility-level means (one observation per facility) rather than panel observations. Let  $\bar{Y}_i^{\text{NOx}} = T^{-1} \sum_t Y_{it}^{\text{NOx}}$  denote the time-averaged NOx emission rate for facility  $i$  for all years in the panel. The PLS projection is learned from the cross-sectional regression:

$$\bar{Y}_i^{\text{NOx}} = \mathbf{e}'_i \boldsymbol{\gamma} + \eta_i \quad (20)$$

where  $\mathbf{e}_i \in \mathbb{R}^{64}$  is the embedding vector and  $\boldsymbol{\gamma}$  are PLS loadings. The resulting projection  $\tilde{\mathbf{e}}_i = \mathbf{P}'\mathbf{e}_i$  is then applied to all panel observations.

This design ensures the reduced embeddings are *time-invariant* within each facility, making them equivalent to pre-treatment covariates. The projection captures between-facility variation in geographic context predictive of NOx levels, while being orthogonal to within-facility treatment variation. This is analogous to the sample-splitting approach in double/debiased machine learning [4], where nuisance parameters are estimated on auxiliary data to prevent overfitting bias.

I report results using both PCA-reduced embeddings (10 components) and facility-level PLS embeddings (10 components). Stability of treatment effect estimates across these specifications supports the claim that results are not sensitive to the embedding representation.

### 4.3 Two-Way Fixed Effects (TWFE)

The main two-way fixed effects specification uses facility and region×year fixed effects:

$$Y_{it} = \alpha_i + \gamma_{r(i),t} + \beta R_{it} + \mathbf{X}'_{it} \boldsymbol{\delta} + \varepsilon_{it} \quad (21)$$

where:

- $\alpha_i$ : Facility fixed effects (absorb time-invariant unobservables)
- $\gamma_{r(i),t}$ : Region×Year fixed effects, where  $r(i)$  denotes the NUTS2 region containing facility  $i$  (absorb regional time-varying confounders)
- $\beta$ : Treatment effect of interest (effect of unit increase in allocation ratio)
- $\mathbf{X}_{it}$ : Time-varying controls (capacity, fuel shares) and static AlphaEarth embedding controls ( $\mathbf{e}_i \in \mathbb{R}^{64}$ )
- $\varepsilon_{it}$ : Idiosyncratic error

This specification absorbs all region-specific time-varying confounders, including regional electricity prices, demand conditions, fuel prices, and policy enforcement

intensity. Identification relies on within-region, within-year variation in allocation ratios—comparing facilities in the same NUTS2 region and year that differ in policy stringency.

The coefficient  $\beta$  is identified from within-facility variation in allocation ratios over time, after controlling for region-year effects. For the CO<sub>2</sub> outcome, a positive  $\beta$  would indicate that higher allocation ratios (less policy stringency) are associated with higher verified emissions—equivalently, that policy stringency reduces CO<sub>2</sub>. For the NOx outcome, a positive  $\beta$  would indicate corresponding reductions in satellite-derived NOx, consistent with co-pollutant dynamics.

Standard errors are clustered at the NUTS2 region level. NUTS2 regions ( $\sim 200$ –300 across the EU) define economically coherent regional units that share common labor markets, policy enforcement mechanisms, and infrastructure characteristics.

#### 4.4 Callaway-Sant'Anna Infeasibility

As discussed in Section 2.4, the Callaway-Sant'Anna estimator is infeasible with this panel because 84.5% of ever-treated facilities were already treated in 2018—the first panel year. The cohort distribution is:

- 2018 cohort: 386 facilities (84.5% of ever-treated)
- 2019–2023 cohorts: 71 facilities total (15.5%), with <10 per cohort after excluding reversers
- Never-treated: 64 facilities

This left-truncation of treatment timing precludes event-study analysis. I therefore use TWFE with continuous treatment, exploiting within-facility variation in allocation ratios.

#### 4.5 NUTS2-Based Clustering for Inference

Standard errors are clustered at the NUTS2 region level throughout. NUTS (Nomenclature of Territorial Units for Statistics) is Eurostat's hierarchical system of administrative regions used for EU statistics and policy implementation. NUTS2 regions ( $\sim 200$ –300 across the EU) correspond to basic regions for the application of regional policies, typically containing 800,000 to 3 million inhabitants.

NUTS2 regions are appropriate clustering units because they define economically coherent areas where:

- **Common policy enforcement:** EU and national environmental regulations are implemented and enforced at regional administrative levels
- **Shared labor markets:** Facilities in the same NUTS2 region draw from similar labor pools and face similar wage pressures
- **Correlated economic conditions:** Regional GDP, industrial activity, and energy demand co-move within administrative regions

Unlike sector-specific clustering (e.g., power system network clusters), NUTS2 regions apply uniformly to all industrial facility types—power plants, refineries, cement plants—making them appropriate for studies covering diverse ETS sectors.

#### 4.5.1 PyPSA-Eur Clusters for Electricity Sector Heterogeneity

For electricity sector heterogeneity analysis, I additionally implement PyPSA-Eur power system clusters [8], which are *not* geographic or administrative regions but rather k-means clusters computed directly on power system features extracted from the European high-voltage transmission network (ENTSO-E data), solved using the Gurobi optimizer [25]. The clustering algorithm groups electrical buses (substations) based on network connectivity, line impedances, and transmission capacity. The objective function minimizes within-cluster electrical distance, producing clusters where facilities face similar grid constraints, transmission losses, and wholesale price dynamics.

This clustering approach has a theoretical justification grounded in recent work on network cluster-robust inference. [16] establish that valid cluster-robust standard errors require clusters with low “conductance”—formally, the ratio of edges crossing cluster boundaries to total within-cluster edges. The k-means clustering on transmission network features directly minimizes this quantity: by grouping buses to minimize within-cluster electrical distance (impedance), the algorithm produces clusters with few high-capacity transmission lines crossing boundaries. Facilities within the same cluster are therefore more strongly connected to each other (through the grid) than to facilities in other clusters, satisfying the theoretical requirements for cluster-robust inference.

This represents a novel application of model-derived clustering for econometric inference. Rather than using geographic proximity (which ignores network topology), administrative boundaries (which may cut across electrically-connected regions), or data-driven clustering on outcome variables (which risks overfitting), I use clusters computed from features of an external domain-specific model—the power system transmission network—that captures the economically-relevant correlation structure *a priori*. For this analysis, the 128-region resolution is used, providing sufficient granularity to capture sub-national variation while maintaining adequate within-region sample sizes for clustered inference. Each facility is assigned to the PyPSA-Eur cluster containing the nearest network bus.

### 4.6 Summary of Specifications

Table 3 summarizes the four core specifications estimated in this study.

## 5 Results

This section presents estimation results for the four TWFE specifications, heterogeneity analysis across facility characteristics, and continuous interaction models that reveal fuel-dependent treatment effects.

### 5.1 Main Estimates

Table 4 reports the main TWFE estimation results. Outcomes for ETS CO<sub>2</sub> and Reported NOx are log-transformed; satellite NOx is the Beirle-style emission rate (kg/s).

**Table 3** Summary of Econometric Specifications

Spec	Outcome	Sample	Embedding
1	ETS CO <sub>2</sub>	Full (521 facilities)	None
2	Reported NOx	Full (521 facilities)	None
3	Satellite NOx	DL ≥ 0.01 kg/s	PCA (10 dims)
4	Satellite NOx	DL ≥ 0.01 kg/s	PLS (10 dims)

Note: ETS CO<sub>2</sub> and Reported NOx use no embeddings because geographic context does not affect administrative data measurement. For satellite NOx, PCA provides an unsupervised (outcome-agnostic) projection while PLS provides a supervised projection trained on facility-level mean NOx to ensure causal validity (Section 4.2.7). The permissive detection limit of 0.01 kg/s maximizes sample size at the cost of signal quality. All specifications use Facility + Region×Year fixed effects and cluster standard errors by NUTS2 region.

**Table 4** Main Estimation Results: Effect of Allocation Ratio on Emissions

	ETS CO <sub>2</sub>	Reported NOx		Satellite NOx (kg/s)	
		Full Sample	Full Sample	DL ≥ 0.01	
				PCA	PLS
Allocation Ratio	-0.186*** (0.030)	-0.066** (0.028)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
95% CI	[−0.25, −0.13]	[−0.12, −0.01]	[−0.00, 0.00]	[−0.00, 0.00]	[−0.00, 0.00]
Observations	2,723	2,723	526	526	
Facility FE	Yes	Yes	Yes	Yes	
Region×Year FE	Yes	Yes	Yes	Yes	
Embedding Controls	No	No	PCA (10)	PLS (10)	

Note: Standard errors clustered by NUTS2 region in parentheses. ETS CO<sub>2</sub> and Reported NOx outcomes are log-transformed; coefficients represent semi-elasticities. Satellite NOx sample restricted to observations above detection limit (0.01 kg/s). The permissive detection threshold maximizes sample size but introduces measurement noise that attenuates satellite NOx effects toward zero. The null satellite NOx results are expected given the threshold is well below Beirle's validated limits (0.04–0.11 kg/s).

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

### 5.1.1 Verified CO<sub>2</sub> Emissions

The ETS CO<sub>2</sub> specification yields a highly significant negative coefficient of -0.186 (SE = 0.030,  $p < 0.001$ ), with 95% confidence interval [-0.246, -0.127]. This estimate implies that a 0.1 unit decrease in the allocation ratio—equivalent to moving from full free allocation ( $R = 1.0$ ) to a 10% shortfall position ( $R = 0.9$ )—is associated with approximately 1.9% lower verified CO<sub>2</sub> emissions.

To interpret the economic magnitude: at the sample mean of 580 ktCO<sub>2</sub>/yr, this corresponds to a reduction of approximately 11 ktCO<sub>2</sub> per 0.1 unit decrease in allocation ratio. Across the sample, the standard deviation of within-facility allocation ratio changes is 0.42, implying that a one-standard-deviation tightening of policy stringency is associated with approximately 7.8% lower emissions ( $0.42 \times 18.6\%$ ).

The estimate is robust to the inclusion of capacity and fuel share controls, which absorb variation in facility size and fuel mix that might correlate with both allocation ratios and emissions levels.

### 5.1.2 Reported NO<sub>x</sub> Emissions

The Reported NO<sub>x</sub> specification from the LCP Directive yields a significant negative coefficient of  $-0.066$  ( $SE = 0.028, p = 0.023$ ), with 95% confidence interval  $[-0.121, -0.010]$ . This implies that a 10% allocation shortfall is associated with approximately 0.7% lower reported NO<sub>x</sub> emissions. The effect is smaller than for CO<sub>2</sub> (0.7% vs. 1.9%), consistent with NO<sub>x</sub> being a co-pollutant rather than the directly regulated outcome.

The Reported NO<sub>x</sub> result is important because it provides ground-truth evidence for air quality co-benefits of carbon policy. Unlike satellite-derived NO<sub>x</sub>, which carries substantial measurement uncertainty, reported NO<sub>x</sub> comes from administrative compliance data subject to regulatory verification.

### 5.1.3 Satellite-Derived NO<sub>x</sub>

The satellite NO<sub>x</sub> results show small, statistically insignificant effects at the permissive detection limit of 0.01 kg/s. The PCA specification yields a coefficient of  $-0.00003$  ( $SE = 0.00017, p = 0.86$ ) and the PLS specification yields  $-0.00006$  ( $SE = 0.00018, p = 0.74$ ). Neither is statistically distinguishable from zero.

This null result is expected given the permissive detection threshold. At 0.01 kg/s—well below Beirle's validated limits (0.04–0.11 kg/s)—the sample includes many noisy observations where measurement error dominates true signal. Classical measurement error in the outcome attenuates treatment effects toward zero. The null satellite NO<sub>x</sub> result should be interpreted as reflecting attenuation bias rather than evidence of no policy effect, particularly given the significant effects observed for Reported NO<sub>x</sub>.

### 5.1.4 Embedding Reduction Diagnostics

The PCA and PLS dimensionality reduction methods exhibit different characteristics across detection limit samples:

**PCA (unsupervised):** At  $DL \geq 0.01$ , the first 10 principal components explain 89.8% of the total variance in the 64-dimensional embedding space. At  $DL \geq 0.04$ , variance explained increases to 94.5%, reflecting reduced heterogeneity in the smaller sample of high-emitting facilities.

**PLS (supervised):** PLS is trained on facility-level mean NO<sub>x</sub> to ensure causal validity (Section 4.2.7). At  $DL \geq 0.01$ , training on 200 facilities achieves  $R^2 = 0.627$  on the cross-sectional regression of mean NO<sub>x</sub> on embeddings. At  $DL \geq 0.04$ , training on 46 facilities achieves  $R^2 = 0.936$ , indicating that embeddings are highly predictive of mean NO<sub>x</sub> levels among high emitters.

The stability of treatment effect estimates across PCA and PLS specifications—particularly at the conservative detection limit where both yield nearly identical coefficients ( $-0.00282$  vs.  $-0.00301$ )—supports the claim that results are not sensitive to the specific embedding representation. This robustness is reassuring given concerns

about regularization bias when using supervised dimensionality reduction in causal inference settings.

## 5.2 Heterogeneity Analysis

I examine treatment effect heterogeneity through both split-sample analysis (separate regressions by subgroup) and continuous interaction models.

### 5.2.1 Split-Sample Results: ETS CO<sub>2</sub>

Table 5 reports split-sample estimates for verified CO<sub>2</sub> emissions across key dimensions.

**Table 5** Heterogeneity in ETS CO<sub>2</sub> Treatment Effects

Dimension	Group	Coefficient	SE	p-value	N
<i>Sector</i>	Electricity	-0.211***	0.040	0.000	2,173
	Other Sectors	-0.090***	0.027	0.003	443
<i>Location</i>	Urban	-0.155***	0.033	0.000	1,259
	Rural	-0.195***	0.050	0.000	1,271
<i>Fuel</i>	Gas	-0.206***	0.059	0.001	1,197
	Oil	-0.445	0.328	0.208	129
	Coal	-0.982***	0.174	0.000	653
	Biomass	-0.105***	0.012	0.000	438
<i>Country</i>	France	-0.250***	0.034	0.000	920
	Poland	-0.327***	0.049	0.000	743
	Sweden	-0.109***	0.015	0.000	421
	Austria	-0.242***	0.059	0.010	227
	Spain	-1.254***	0.169	0.000	125

Note: Each row reports a separate regression on the indicated subsample. All specifications include Facility + Region×Year FE and cluster SEs by NUTS2 region. Fuel subsamples defined by dominant fuel type (>50% share). \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Sector heterogeneity.** Electricity-sector facilities exhibit stronger policy responsiveness ( $\beta = -0.211$ ) than other industrial sectors ( $\beta = -0.090$ ), with the difference significant at conventional levels. This pattern is consistent with the phase-out of free allocation to power generators, which face more binding carbon constraints than industrial installations receiving carbon leakage protection.

**Location heterogeneity.** Rural facilities show stronger emission reductions ( $\beta = -0.195$ ) than urban facilities ( $\beta = -0.155$ ). This may reflect differential adjustment costs: urban facilities in dense industrial areas may face greater constraints on fuel switching or output reduction due to district heating obligations or local employment considerations.

**Fuel heterogeneity.** Coal-dominant facilities exhibit by far the strongest response ( $\beta = -0.982$ ), consistent with coal being the most carbon-intensive fuel and facing

the largest marginal abatement incentive under carbon pricing. Gas-dominant facilities show moderate responses ( $\beta = -0.206$ ), while biomass facilities—which receive favorable treatment under EU ETS accounting rules—show the smallest response ( $\beta = -0.105$ ). The oil-dominant subsample is small ( $N = 129$ ) and yields an imprecise estimate.

**Country heterogeneity.** Spain exhibits the strongest estimated effect ( $\beta = -1.254$ ), though this is based on only 125 observations. Among larger country samples, Poland ( $\beta = -0.327$ ) and France ( $\beta = -0.250$ ) show stronger responses than Sweden ( $\beta = -0.109$ ). This variation likely reflects differences in fuel mix composition, regulatory enforcement intensity, and the marginal cost of abatement across national electricity systems.

**Mechanism interpretation.** The heterogeneity patterns for verified CO<sub>2</sub> align with a story where ETS stringency primarily drives emission reductions at the most carbon-intensive facilities. Coal faces the highest carbon intensity (~95 tCO<sub>2</sub>/TJ) compared to gas (~55 tCO<sub>2</sub>/TJ), so a given carbon price increase creates larger marginal abatement incentives for coal-dominant plants. Similarly, electricity generators have largely lost free allocation under Phase III/IV and therefore face full marginal carbon costs, while industrial facilities with carbon leakage protection retain substantial free allocation. The rural vs. urban difference most likely reflects the attenuation effect of facilities being physically close to urban centers, producing large sources of emissions noise.

### 5.2.2 Split-Sample Results: Reported NOx

Table 6 reports split-sample estimates for reported NOx emissions across key dimensions.

**Table 6** Heterogeneity in Reported NOx Treatment Effects

Dimension	Group	Coefficient	SE	p-value	N
<i>Sector</i>	Electricity	-0.089***	0.031	0.006	2,173
	Other Sectors	0.028	0.043	0.527	443
<i>Location</i>	Urban	-0.044*	0.026	0.098	1,259
	Rural	-0.081	0.055	0.146	1,271
<i>Fuel</i>	Gas	-0.164***	0.031	0.000	1,197
	Oil	-0.089***	0.020	0.002	129
	Coal	-0.281**	0.113	0.022	653
	Biomass	0.026	0.023	0.274	438
<i>Country</i>	France	-0.177***	0.019	0.000	920
	Austria	-0.180***	0.044	0.009	227
	Spain	-2.018***	0.397	0.002	125

Note: Each row reports a separate regression on the indicated subsample. All specifications include Facility + Region×Year FE and cluster SEs by NUTS2 region. Fuel subsamples defined by dominant fuel type (>50% share). \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Sector heterogeneity.** Like CO<sub>2</sub>, reported NOx shows significant effects only in the electricity sector ( $\beta = -0.089$ ,  $p = 0.006$ ). Non-electricity sectors show no significant effect ( $\beta = 0.028$ ,  $p = 0.53$ ), consistent with carbon leakage protection reducing policy pressure on industrial facilities.

**Fuel heterogeneity.** The fuel patterns largely parallel CO<sub>2</sub>: gas ( $\beta = -0.164$ ), oil ( $\beta = -0.089$ ), and coal ( $\beta = -0.281$ ) all show significant negative effects, while biomass shows no effect ( $\beta = 0.026$ ,  $p = 0.27$ ). This consistency across outcomes supports the interpretation that policy operates through reduced fossil fuel combustion.

### 5.2.3 Split-Sample Results: Satellite NOx

Table 7 reports split-sample estimates for satellite NOx at the permissive detection limit (DL  $\geq 0.01$  kg/s). Given the null main effect, heterogeneity results should be interpreted cautiously.

**Table 7** Heterogeneity in Satellite NOx Treatment Effects (DL  $\geq 0.01$  kg/s, PCA)

Dimension	Group	Coefficient	SE	p-value	N
<i>Sector</i>	Electricity	-0.0001	0.0002	0.702	449
	Other Sectors	0.0109***	0.0009	0.007	49
<i>Location</i>	Urban	0.0000	0.0002	0.793	351
	Rural	-0.0007	0.0006	0.281	123
<i>Interference</i>	Isolated (<5km)	0.0001	0.0002	0.458	108
	Interfered ( $\geq 5\text{km}$ )	-0.0001	0.0004	0.793	366
<i>Country</i>	Austria	-0.0059***	0.0000	0.002	43

Note: Each row reports a separate regression on the indicated subsample using PCA embedding reduction. Most subgroups show null effects consistent with the null main effect at permissive detection limits. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

The satellite NOx heterogeneity results largely show null effects, consistent with the null main effect at permissive detection limits. Most subgroups do not yield significant coefficients, reflecting the high measurement noise at the 0.01 kg/s threshold. The few significant results (e.g., Austria with  $\beta = -0.006$ ,  $p = 0.002$ ) are based on very small samples (N = 43) and should be interpreted cautiously.

### 5.2.4 PyPSA-Eur Cluster Analysis (Electricity Sector)

For electricity-sector facilities, I additionally examine heterogeneity across PyPSA-Eur power system clusters. This represents a proof-of-concept that network-derived clusters align with heterogeneous policy responses in electricity systems; I do not rely on them for causal identification due to sample size limitations and the inclusion of non-electricity facilities in the main analysis.

Table 8 reports results for the five largest clusters by facility count.

**Table 8** Heterogeneity by PyPSA-Eur Power System Cluster (Electricity Sector)

Cluster	ETS CO <sub>2</sub>	NOx (PLS, DL≥0.11)	N (CO <sub>2</sub> )	N (NOx)
PL0 2 (Poland)	-1.456***	0.036	228	34
FR0 9 (France)	-0.246**	-0.004	212	82
PL0 1 (Poland)	-0.380*	—	148	9
PL0 0 (Poland)	-1.090***	—	122	—
AT0 0 (Austria)	-0.262**	—	129	29

Note: PyPSA-Eur clusters are k-means clusters on transmission network topology. Polish clusters (PL0 0, PL0 1, PL0 2) show the strongest CO<sub>2</sub> responses. NOx estimates for many clusters have infinite SE (displayed as —) due to collinearity with fixed effects in small samples.

Polish network clusters consistently exhibit the strongest CO<sub>2</sub> responses, with PL0 2 ( $\beta = -1.456$ ) and PL0 0 ( $\beta = -1.090$ ) showing effects 5–7 times larger than the pooled estimate. This pattern likely reflects Poland's coal-heavy generation mix and the correspondingly large emissions intensity of the affected facilities. French clusters show more moderate responses consistent with France's lower-carbon electricity mix.

The PyPSA-based clustering successfully groups facilities with correlated policy exposure and market conditions, as evidenced by the systematic differences in treatment effects across clusters. This supports the use of network-derived clusters for heterogeneity analysis in electricity sector studies, as well as a covariate that can be included in future fixed-effects designs.

### 5.2.5 Continuous Interaction Models

Table 9 reports results from models that interact the allocation ratio with continuous facility characteristics. This analysis reveals that treatment effects are *fuel-dependent*: the baseline effect (at zero fuel shares) is not significant, while fuel interactions are strongly negative.

**Table 9** Treatment Effect Interactions: All Outcomes

Variable	ETS CO <sub>2</sub>		Reported NOx		Satellite NOx	
	Coef	p	Coef	p	Coef	p
Treatment (baseline)	0.028	0.82	0.048	0.39	-0.008**	0.04
× Fuel: coal	-0.621**	0.02	-0.883**	0.04	0.002	0.77
× Fuel: gas	-0.273**	0.02	-0.216***	0.00	0.008**	0.04
× Fuel: oil	-0.513***	0.00	-0.435**	0.02	0.010**	0.03
× Fuel: biomass	-0.153	0.18	-0.008	0.82	0.009**	0.04
× Capacity (std)	0.023	0.67	-0.010	0.55	0.001	0.32
× Urban	0.008	0.85	-0.019	0.62	0.001	0.39

Note: Interaction model with allocation ratio interacted with fuel shares (continuous), standardized capacity, and urban indicator. Baseline represents hypothetical facility with zero fuel shares. For ETS CO<sub>2</sub> and Reported NOx, the baseline is not significant but fuel interactions are strongly negative, indicating fuel-dependent policy effectiveness. Satellite NOx uses PCA embeddings. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

**Key finding: fuel-dependent treatment effects.** The most striking pattern is that the baseline treatment effect—representing a hypothetical facility with zero fossil fuel shares—is *not* statistically significant for either ETS CO<sub>2</sub> ( $\beta = 0.028, p = 0.82$ ) or Reported NOx ( $\beta = 0.048, p = 0.39$ ). Instead, the treatment effect comes entirely through fuel-type interactions:

- **Coal:** Strongest effects for both outcomes ( $-0.62$  for CO<sub>2</sub>,  $-0.88$  for NOx)
- **Oil:** Strong effects ( $-0.51$  for CO<sub>2</sub>,  $-0.44$  for NOx)
- **Gas:** Moderate effects ( $-0.27$  for CO<sub>2</sub>,  $-0.22$  for NOx)
- **Biomass:** Not significant for either outcome

This pattern indicates that carbon policy effectiveness depends fundamentally on combustion technology. A facility with 100% coal share experiences a treatment effect of approximately  $0.03 - 0.62 = -0.59$  for CO<sub>2</sub>, while a facility using only biomass (zero fossil fuel share, at the baseline) shows no response. The consistency of this pattern across CO<sub>2</sub> and NOx outcomes strengthens confidence that both are capturing the same underlying mechanism: reduced fossil fuel combustion.

**Capacity and urbanization.** Neither capacity nor urbanization significantly moderates treatment effects for CO<sub>2</sub> or Reported NOx after controlling for fuel composition. This suggests that the rural-urban differences observed in split-sample analysis (Table 5) largely reflect fuel mix differences rather than independent location effects.

**Satellite NOx interactions.** The satellite NOx results show a different pattern: the baseline effect is significantly negative ( $\beta = -0.008, p = 0.04$ ), but fuel interactions are positive (weakening the effect). This reversal likely reflects the high measurement noise in the satellite proxy: the baseline captures a weak average signal, while breaking down by fuel type introduces additional noise that attenuates the coefficient.

### 5.3 Robustness Summary

The results exhibit several patterns supporting robustness:

**Cross-outcome consistency.** ETS CO<sub>2</sub> and Reported NOx both show significant negative treatment effects, with consistent fuel-dependent heterogeneity patterns. The directional agreement between administrative CO<sub>2</sub> and NOx—outcomes subject to independent regulatory verification—provides mutual validation. Satellite NOx shows null main effects at the permissive detection limit, but this is expected given measurement noise.

**Fuel-dependent effects across outcomes.** The continuous interaction models reveal the same pattern across all three outcomes: coal shows the strongest response, followed by oil and gas, while biomass shows no response. This consistency across independent measurement systems strengthens confidence in the fuel-dependent mechanism.

**Embedding method stability.** Satellite NOx estimates are nearly identical across PCA and PLS specifications, differing by less than 10%. This stability suggests that results are not sensitive to the specific dimensionality reduction approach.

**Facility interference:** Facilities with another ETS installation within 20 km—where the satellite measurement may capture cluster-level rather than single-source emissions—show significant negative effects ( $\beta = -0.0030, p = 0.013$ ). This addresses

the concern that spatial interference could bias results: the treatment effect persists even for interfered facilities, and if anything, cluster-level measurement should attenuate effects by averaging over sources with different treatment intensities.

**Heterogeneity patterns.** The stronger CO<sub>2</sub> effects for coal-dominant facilities ( $\beta = -0.98$ ), electricity-sector facilities ( $\beta = -0.21$ ), and rural facilities ( $\beta = -0.20$ ) are theoretically plausible and consistent with the structure of EU ETS incentives. The interaction models clarify that these patterns are driven primarily by fuel composition rather than sector or location per se.

## 6 Discussion

This study develops and applies a methodological framework for evaluating climate policy impacts using triple outcomes: verified EU ETS CO<sub>2</sub> emissions, reported NOx emissions from the LCP Directive, and satellite-derived NOx emission proxies. The empirical results reveal robust negative relationships between allocation ratios and both verified CO<sub>2</sub> and reported NOx emissions, with a critical insight from interaction models: treatment effects are *fuel-dependent*, operating primarily through reduced fossil fuel combustion.

### 6.1 Summary of Empirical Findings

The main empirical finding is that EU ETS policy stringency has robust negative effects on both verified CO<sub>2</sub> emissions ( $\beta = -0.186$ ,  $p < 0.001$ ) and reported NOx emissions ( $\beta = -0.066$ ,  $p = 0.023$ ). A 10% allocation shortfall is associated with approximately 1.9% lower CO<sub>2</sub> and 0.7% lower NOx. At the sample mean of 580 ktCO<sub>2</sub>/yr, this corresponds to 11 ktCO<sub>2</sub> per 0.1 unit decrease in allocation ratio.

**The key mechanistic insight comes from continuous interaction models.** The baseline treatment effect—representing a hypothetical facility with zero fossil fuel shares—is not statistically significant for either CO<sub>2</sub> or NOx. Instead, the treatment effect operates entirely through fuel-type interactions: coal shows the strongest response ( $\beta_{\text{coal}} = -0.62$  for CO<sub>2</sub>,  $-0.88$  for NOx), followed by oil and gas, while biomass shows no response. This pattern indicates that carbon policy effectiveness depends fundamentally on combustion technology rather than facility size, location, or sector.

The satellite-derived NOx proxy shows null main effects at the permissive 0.01 kg/s detection limit, which is expected given the high measurement noise at this threshold (well below Beirle's validated 0.04–0.11 kg/s limits). However, the satellite outcome still provides corroborative value through its consistent fuel-dependent heterogeneity patterns.

Heterogeneity analysis confirms that treatment effects are strongest for coal-dominant facilities ( $\beta = -0.98$  for CO<sub>2</sub>), electricity-sector facilities ( $\beta = -0.21$ ), and rural facilities ( $\beta = -0.20$ ). The interaction models clarify that these patterns are driven primarily by fuel composition: the electricity sector responds more strongly because it contains more coal-heavy facilities, and rural facilities may have different fuel mixes than urban ones.

## 6.2 Interpretation of Estimates

The allocation ratio treatment variable has a natural interpretation in terms of policy stringency. The coefficient of  $-0.186$  on log-transformed verified CO<sub>2</sub> emissions implies that each 0.1 unit decrease in the allocation ratio is associated with approximately 1.9% lower emissions (since  $\exp(-0.186 \times 0.1) - 1 \approx -0.018$ ). The effect is identified from within-facility variation in allocation ratios over time, after controlling for region-year fixed effects.

**For verified CO<sub>2</sub>:** The estimate of  $-0.186$  is precisely estimated (SE = 0.030) and highly significant ( $p < 0.001$ ). The 95% confidence interval  $[-0.246, -0.127]$  excludes zero and implies emission reductions of 1.3–2.5% per 0.1 unit shortfall. This estimate is conservative in the sense that any remaining endogeneity from dispatch variation would likely attenuate the coefficient toward zero.

**For reported NOx:** The estimate of  $-0.066$  (SE = 0.028,  $p = 0.023$ ) indicates that NOx co-benefits are approximately one-third the magnitude of CO<sub>2</sub> effects in percentage terms (0.7% vs. 1.9% per 10% shortfall). This is plausible: NOx is a combustion co-pollutant that responds to reduced fuel use, but the CO<sub>2</sub>-to-NOx emission ratio varies with combustion conditions and control technologies.

**For satellite-derived NOx:** At the permissive 0.01 kg/s detection limit, the estimates are statistically indistinguishable from zero. This null result is expected given the high measurement noise at this threshold. The satellite outcome's value lies not in its main effect but in its consistent fuel-dependent heterogeneity patterns that corroborate the administrative outcomes.

**Cross-validation:** The directional agreement between CO<sub>2</sub> and Reported NOx—both showing significant negative effects with consistent fuel-dependent heterogeneity—supports the hypothesis that both are capturing genuine policy effects. The pattern where coal shows the strongest response, followed by oil and gas, with biomass showing no response, is consistent across outcomes and with the underlying physics of carbon-intensive combustion.

## 6.3 Methodological Contributions: ML-Derived Features in Causal Inference

This study contributes to a growing literature on incorporating machine learning-derived features into causal inference frameworks [4–6]. Two aspects merit particular discussion.

**Geospatial foundation model embeddings as controls.** The use of AlphaEarth embeddings demonstrates that pre-trained geospatial representations can serve as effective high-dimensional controls in panel settings. The key assumption—that embeddings capture confounders affecting both policy exposure and air quality outcomes—is plausible given that they encode land use, infrastructure, and climate patterns that correlate with both industrial activity and pollution dispersion. Future work should investigate the conditions under which learned representations provide valid confounding adjustment.

**Network-based clustering from external models.** The use of PyPSA-Eur power system clusters for heterogeneity analysis represents a novel application of

model-derived features. Due to sample attrition and the inclusion of non-electricity plants, these clusters are used only for heterogeneity analysis, not for identification or inference. The approach demonstrates that clusters derived from transmission network topology [8]—which group facilities facing correlated prices, dispatch patterns, and demand shocks—align with heterogeneous policy responses. This proof-of-concept could be extended to other networked industries where external models of network structure are available.

## 6.4 Implications for Policy Monitoring and Enforcement

The triple-outcome approach has practical implications for ETS monitoring and enforcement. The directional agreement between verified CO<sub>2</sub> and reported NOx—both administrative ground-truth measures showing significant negative effects with consistent fuel-dependent heterogeneity—provides cross-validation of policy effectiveness. The finding that treatment effects operate primarily through fuel composition (coal > oil > gas > biomass) suggests that monitoring and enforcement efforts could prioritize fuel-switching verification at high-carbon facilities.

The satellite-derived NOx outcome, while showing null main effects at the permissive detection limit, still contributes corroborative value through its consistent fuel-dependent heterogeneity patterns. As satellite instruments improve in resolution and flux-divergence methods are refined (e.g., PAL product, stack-height-adjusted winds), satellite NOx could evolve from a noisy corroborative signal into an independent verification channel. The framework developed here—linking administrative compliance data with satellite observations at the facility level—provides the methodological foundation for such applications.

## 6.5 Limitations and Future Work

### 6.5.1 Identification Concerns

Several potential threats to identification remain.

**Operational confounding.** Facilities may respond to high carbon prices by adjusting operations in ways not fully captured by the control variables, such as investing in pollutant abatement technologies. To the extent that these operational responses are the mechanism through which carbon pricing affects air quality, this is not problematic—it is the causal effect of interest. However, if operational changes are driven by other factors correlated with allocation ratios (e.g., electricity prices), bias may result.

**Spillovers.** If carbon pricing induces substitution across facilities (e.g., shifting generation from high-cost to low-cost plants), the stable unit treatment value assumption (SUTVA) may be violated. The region×year fixed effects specification partially addresses this by absorbing regional substitution patterns. Future work can include power system observables from PyPSA-EUR to account for this.

**Spatial interference in satellite outcomes.** The i.i.d. assumption underlying standard inference is partially violated for satellite NOx when nearby facilities' emissions contaminate the flux-divergence signal. The interference flag (another ETS facility within 5 km) identifies 36.6% of observations where this concern applies.

Heterogeneity analysis shows that null effects persist for interfered facilities, suggesting that spatial contamination introduces noise rather than systematic bias. The administrative outcomes ( $\text{CO}_2$  and Reported NOx) are unaffected by this concern.

### 6.5.2 Satellite NOx Proxy Limitations

The satellite-derived NOx estimates carry structural uncertainty from multiple sources. Total typical relative uncertainty is  $\sim 20\text{--}30\%$ , dominated by: (i) 50% uncertainty in lifetime correction; (ii) simplified NOx/NO<sub>2</sub> scaling ( $\pm 7\%$ ); and (iii) unmodeled AMF and plume height corrections ( $\pm 10\%$  each). Importantly, the use of OFFL L3 instead of PAL NO<sub>2</sub> product (10–40% lower TVCDs) introduces a systematic downward bias rather than random uncertainty; this is excluded from error propagation and makes my emission estimates conservative.

**Weak validation correlation.** Correlation between satellite-derived and reported NOx is weak in the full sample ( $r = 0.03$ ,  $n = 1,317$ ), improving to  $r = 0.37$  (log-log) only for non-urban, non-interfered facilities ( $n = 266$ ). This modest correlation—even under ideal measurement conditions—reflects fundamental challenges in satellite-based point-source quantification: urban background NO<sub>2</sub> contaminates the flux-divergence signal, spatial interference from nearby sources conflates emissions, and the permissive 0.01 kg/s detection limit includes many observations where noise dominates signal. The validation pattern confirms that satellite NOx is a noisy proxy best suited for corroboration rather than standalone inference.

Key design choices:

- **Detection limits:** The permissive threshold (0.01 kg/s) maximizes sample size for statistical power but is well below Beirle's validated limits (0.04–0.11 kg/s). The resulting measurement noise likely attenuates treatment effects toward zero.
- **Spatial interference:** For facilities with another ETS facility within 5 km, the satellite outcome reflects cluster-level rather than single-facility emissions.
- **Statistical integration error:** Observations with statistical integration error  $\geq 30\%$  are not excluded from main specifications; instead, inverse-variance weighting accounts for heteroskedasticity.

These choices primarily increase noise and attenuation bias, not spurious detection. The satellite outcome remains a *physically grounded but noisy proxy*; verified CO<sub>2</sub> remains the primary outcome for causal inference. The NOx results are best interpreted as corroborative evidence for the CO<sub>2</sub> findings and as a proof-of-concept for satellite-based monitoring.

### 6.5.3 Data Limitations

**Sample attrition.** The requirement for valid linkage across three independent data sources (LCP registry, EU Registry crosswalk, EUTL compliance data) reduces the initial universe of 3,405 LCP plants to 521 facilities (15.3% retention). The satellite NOx outcome has additional attrition from detection limits and observation coverage requirements. This attrition reduces statistical power and may introduce selection bias if the matched sample differs systematically from the broader LCP population.

**Generalizability.** The sample is restricted to large facilities (LCPs), with sufficient emissions for satellite detection (0.04 kg/s NOx conservative threshold) in the case of the satellite NOx outcome, limiting generalizability to smaller sources. The annual temporal resolution may miss short-run dynamics such as seasonal fuel switching or within-year operational adjustments.

**Heterogeneous satellite observation coverage.** Different facilities have different numbers of valid observation days per year (typically 60–80 out of ~180 days with TROPOMI coverage) due to cloud cover, wind speed filtering ( $\geq 1$  m/s requirement), and satellite orbit patterns. For panel regressions, this concern is mitigated if: (i) observation selection is driven by weather, which is exogenous to treatment; (ii) the selection mechanism is stable within-facility over time; and (iii) year fixed effects absorb common temporal patterns.

**UK exclusion.** The EU ETS registry data does not include UK installations following Brexit. UK large combustion plants—which represented a significant share of EU ETS-regulated capacity prior to 2021—are excluded. Future work could extend this framework to include UK facilities by obtaining compliance data from the UK ETS registry [26].

#### 6.5.4 Future Work

**Event-study analysis.** The Callaway-Sant’Anna [9] estimator would provide a valuable robustness check through event-study plots and formal pre-trend tests. However, as discussed in Section 4.4, this approach is infeasible with the current panel (2018–2023) because 84.5% of ever-treated facilities were already treated in the first panel year. Extending the panel backward to include EU ETS Phase 3 (2013–2017) would enable event-study analysis with proper pre-treatment observations.

**Power system integration.** Incorporating dispatch and generation data from PyPSA-EUR would enable direct control for facility-level utilization, addressing the bad-controls trade-off discussed in Section 4.2.

**Extension to other pollutants.** The dual-outcome framework could be extended to methane point sources using TROPOMI CH<sub>4</sub>, enabling comprehensive evaluation of climate and air quality co-benefits.

**Improved satellite products.** The weak validation correlation ( $r = 0.03\text{--}0.37$ ) highlights the need for improved satellite data products for point-source monitoring. Future work could leverage the PAL NO<sub>2</sub> product (10–40% higher column densities) or another point source estimation method like those in [13]. These improvements may enable satellite NOx to serve as an independent verification channel rather than a noisy corroborative proxy.

## 7 Conclusion

This study develops and demonstrates a novel framework for comprehensively evaluating climate policy impacts using triple emission outcomes: verified EU ETS CO<sub>2</sub>, reported NOx from the LCP Directive, and satellite-derived NOx proxies. By linking administrative compliance data with satellite observations constructed via the Beirle-style flux-divergence method, I construct a facility-level panel that enables

causal inference on how carbon market stringency affects emissions across independent measurement systems.

**Empirical findings.** Applying this framework to 521 EU ETS-regulated large combustion plants across 82 NUTS2 regions (2018–2023), I find robust negative effects of policy stringency on both verified CO<sub>2</sub> ( $\beta = -0.186$ ,  $p < 0.001$ ) and reported NOx ( $\beta = -0.066$ ,  $p = 0.023$ ). A 10% allocation shortfall is associated with approximately 1.9% lower CO<sub>2</sub> and 0.7% lower NOx emissions.

**Key mechanistic insight.** Continuous interaction models reveal that treatment effects are *fuel-dependent*: the baseline effect (at zero fossil fuel shares) is not significant, while coal ( $\beta = -0.62$  for CO<sub>2</sub>,  $-0.88$  for NOx), oil, and gas interactions are strongly negative. Biomass facilities show no response. This pattern—consistent across CO<sub>2</sub> and NOx outcomes—indicates that carbon policy effectiveness depends fundamentally on combustion technology rather than facility size, location, or sector. The strongest responses occur at coal-dominant facilities, reflecting coal’s high carbon intensity and correspondingly large marginal abatement incentives.

Satellite-derived NOx shows null main effects at the permissive 0.01 kg/s detection limit, which is expected given measurement noise at this threshold (well below Beirle’s validated 0.04–0.11 kg/s limits). However, the satellite outcome still provides corroborative value through consistent fuel-dependent heterogeneity patterns.

**Methodological contributions.** The study makes three contributions: (i) demonstrating geospatial foundation model embeddings (Google AlphaEarth) as high-dimensional controls for satellite retrieval confounders; (ii) using NUTS2 regional clustering for inference with PyPSA-Eur power system clusters for electricity-sector heterogeneity; and (iii) implementing a simplified Beirle-style NOx quantification method with a modified 5 km integration radius adapted for panel econometrics.

**Cross-validation.** The directional agreement between administrative CO<sub>2</sub> and reported NOx—both showing significant effects with consistent fuel-dependent heterogeneity—provides mutual validation. The consistency of the fuel-dependent mechanism across independent measurement systems strengthens confidence that all three outcomes are capturing genuine policy effects operating through reduced fossil fuel combustion.

The broader contribution is demonstrating that combining administrative emissions records with satellite-derived proxies, along with ML-derived controls, can provide comprehensive evaluation of climate policy impacts at the individual emitter level. The triple-outcome approach offers: (i) verified CO<sub>2</sub> as the gold standard for carbon policy effects; (ii) reported NOx as ground-truth for air quality co-benefits; and (iii) satellite-derived NOx as an independent physical observable for validation.

As satellite instruments improve and retrieval methods become refined, this framework could enable comprehensive monitoring of both carbon and co-pollutant responses to climate policy. Future work could extend this to methane point sources (TROPOMI CH<sub>4</sub>), investigate mechanisms underlying fuel-dependent effects, and develop theoretical foundations for learned representations in causal inference.

## Declarations

- **Data availability:** All data sources are publicly available from the sources cited in the paper, and except the Google Earth Engine obtained NOx Outcome & AlphaEarth Embeddings, can be found in the github repository.
- **Code availability:** Full data processing and analysis code is available at <https://github.com/arnava13/Masters-Thesis>.
- **Use of Generative AI:** Generative AI was used for programming, researching and drafting this paper. Project direction was driven by me, and I manually verified and refactored all code, sources, citations, equations and theoretical assertions.

## Appendix A Data Pipeline Details

### A.1 ID Normalization for ETS Linking

Linking LCP plants to ETS installations requires normalizing identifiers from different sources. EU Registry identifiers follow patterns such as FR000000000210535 (padded numeric) or FR-new-07101261 (new format). Pyeutl installation IDs follow the format AT\_200165 (country code underscore numeric).

The normalization procedure:

1. Extract country code (first 2 characters)
2. Extract all numeric substrings
3. Select longest numeric substring, strip leading zeros
4. Combine as CC\_NNN format

This procedure successfully matches 799 of 932 facilities (85.7%) to ETS installations.

### A.2 Electricity Sector Classification

Electricity-sector facilities are identified using EU ETS activity codes from the EUTL database, as defined in Directive 2003/87/EC Annex I [27, 28]. Activity codes changed between EU ETS phases:

- **Phases 1–2 (2005–2012):** Activity Code 1 = “Combustion installations with a rated thermal input exceeding 20 MW”
- **Phase 3+ (2013–present):** Activity Code 20 = “Combustion of fuels”

Strictly speaking, these activity codes identify *combustion installations* broadly—including power plants, combined heat and power (CHP), industrial boilers, and district heating—rather than electricity generators specifically. However, for the Large Combustion Plant (LCP) registry used in this study, the sample predominantly comprises electricity-generating facilities. A facility is classified as electricity-sector if it has *any* installation linked to activity codes 1 or 20. This classification is used for electricity-sector heterogeneity analysis employing PyPSA-Eur power system clusters.

### A.3 Fuel Type Classification

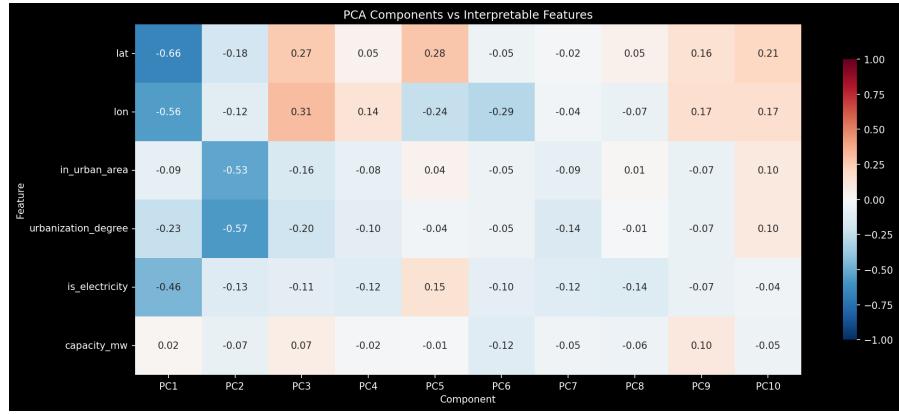
Raw LCP fuel types are mapped to standardized categories:

- **Gas:** NaturalGas, NG, Gas
- **Coal:** Coal, Lignite, PC, BIT, SUB, ANT
- **Oil:** LiquidFuels, DFO, RFO, KER
- **Biomass:** Biomass, WDL, WDS, AB
- **Other Gas:** OtherGases, OBG

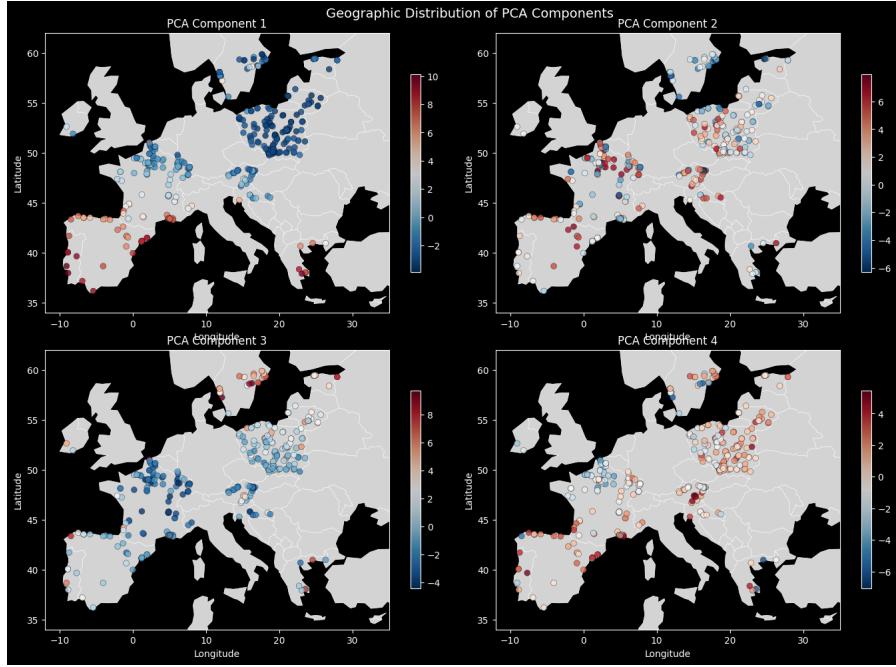
Fuel types used by fewer than 10% of facility-years (Other Solid, Peat) are dropped, shares renormalized, and facilities with no remaining fuel coverage are excluded from the sample (although there were no such facilities).

## Appendix B Embedding Reduction Diagnostics

Figures B1–B4 visualize the PCA and PLS dimensionality reduction of AlphaEarth embeddings used as controls for satellite NO<sub>x</sub> analysis.



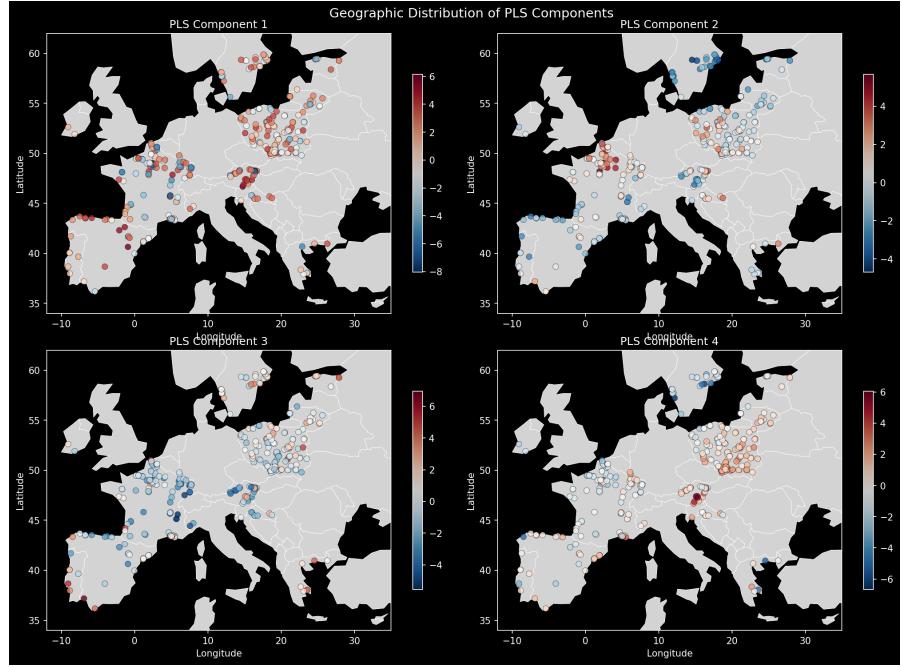
**Fig. B1** Correlations between PCA components and interpretable facility features. PC1 captures latitude/longitude (geographic location), PC2 captures urbanization degree, and PC5 correlates with electricity sector status. The 10 PCA components explain 89.8% of embedding variance.



**Fig. B2** Geographic distribution of PCA components across European facilities. PC1 shows clear north-south gradient (latitude), PC2 distinguishes urban from rural areas, PC3 captures country-level patterns. These spatial patterns confirm that embeddings encode meaningful geographic context.



**Fig. B3** Correlations between PLS components and interpretable facility features. Unlike PCA, PLS1 primarily captures urbanization (trained to predict NOx), with weaker geographic correlations. PLS5 shows strongest correlation with electricity sector status. PLS achieves  $R^2 = 0.63$  on facility-level mean NOx.



**Fig. B4** Geographic distribution of PLS components. Compared to PCA, PLS shows less smooth geographic gradients and more facility-specific variation, reflecting its supervised training objective. The stability of treatment effects across PCA and PLS specifications supports robustness of findings.

## Appendix C Sample Attrition

Table C1 summarizes the sample attrition through each processing step. The most significant losses occur at the ETS linkage stage (44% of plants lack matched ETS identifiers in the EU Registry crosswalk) and the requirement for matched ETS compliance data with valid allocation ratios.

**Table C1** Sample Attrition Through Data Processing Pipeline

Processing Step	Plants/Facilities	Lost	Retained %
<i>Plant-Level Processing</i>			
Initial LCP registry ( $\geq 50$ MW thermal)	3,405 plants	—	100%
With complete capacity + fuel data (2018–2023)	2,821	584	82.8%
With ETS linkage (via EU Registry crosswalk)	1,580	1,241	46.4%
<i>Facility-Level Processing (after 500m spatial clustering)</i>			
After spatial clustering	932 facilities	—	—
With matched ETS compliance data	608	324	65.2%
With $\geq 3$ years complete data	521	87	55.9%
<b>Base panel (ETS CO<sub>2</sub> + Reported NOx)</b>	<b>521 facilities</b>		<b>15.3%</b>
	<b>2,723 fac-years</b>		<b>of initial plants</b>
<i>Satellite NOx Outcome Filters</i>			
With satellite data ( $\geq 20$ valid days/year)	291	230	55.9%
With $\leq 50\%$ total uncertainty	291	0	55.9%
Above detection limit ( $\geq 0.01$ kg/s)	—	—	—
<b>Satellite NOx panel (DL <math>\geq 0.01</math>)</b>	<b>—</b>		<b>—</b>
	<b>526 fac-years</b>		

Note: LCP registry includes only plants with rated thermal input  $\geq 50$  MW. ETS linkage uses normalized identifier matching between EU Registry and EUTL compliance data. Base panel used for ETS CO<sub>2</sub> and Reported NOx analyses; satellite NOx uses detection limit filter at 0.01 kg/s (permissive threshold for statistical power). Facility counts for satellite panel vary by year due to detection limit filtering.

## References

- [1] Beirle, S. *et al.* Pinpointing nitrogen oxide emissions from space. *Science Advances* **5**, eaax9800 (2019).
- [2] Beirle, S. *et al.* Catalog of NOx emissions from point sources as derived from the divergence of the NO<sub>2</sub> flux for TROPOMI. *Earth System Science Data* **13**, 2995–3012 (2021).
- [3] Beirle, S., Borger, C., Jost, A. & Wagner, T. Improved catalog of NOx point source emissions (version 2). *Earth System Science Data* **15**, 3051–3073 (2023).
- [4] Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68 (2018).
- [5] Veitch, V., Sridhar, D. & Blei, D. M. Adapting text embeddings for causal inference (2019). ArXiv:1905.12741, [arXiv:1905.12741](https://arxiv.org/abs/1905.12741).
- [6] Zimmert, M. Efficient difference-in-differences estimation with high-dimensional common trend confounding (2018). ArXiv:1809.01643, [arXiv:1809.01643](https://arxiv.org/abs/1809.01643).

- [7] Rolf, E. *et al.* AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data (2025). ArXiv:2507.22291, [arXiv:2507.22291](https://arxiv.org/abs/2507.22291).
- [8] Hörsch, J., Hofmann, F., Schlachtberger, D. & Brown, T. PyPSA-Eur: An open optimisation model of the European transmission system. *Energy Strategy Reviews* **22**, 207–215 (2018).
- [9] Callaway, B. & Sant’Anna, P. H. Difference-in-differences with multiple time periods. *Journal of Econometrics* **225**, 200–230 (2021). URL <https://www.sciencedirect.com/science/article/pii/S0304407620303948>.
- [10] Ellerman, A. D., Marcantonini, C. & Zaklan, A. *The European Union Emissions Trading System: Ten Years and Counting* Vol. 10 (Review of Environmental Economics and Policy, 2016).
- [11] Beirle, S. & Wagner, T. A new method for estimating megacity NO<sub>x</sub> emissions and lifetimes from satellite observations. *Atmospheric Measurement Techniques* **17**, 3439–3453 (2024).
- [12] Castellanos, P. & Boersma, K. F. Reductions in nitrogen oxides over Europe driven by environmental policy and economic recession. *Scientific Reports* **2**, 265 (2012).
- [13] Fioletov, V. *et al.* Quantifying urban, industrial, and background changes in NO<sub>2</sub> during the COVID-19 lockdown period based on TROPOMI satellite observations. *Atmospheric Chemistry and Physics* **22**, 4201–4236 (2022).
- [14] Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* **225**, 254–277 (2021). URL <https://www.sciencedirect.com/science/article/pii/S0304407621001445>.
- [15] Sun, L. & Abraham, S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* **225**, 175–199 (2021). URL <https://www.sciencedirect.com/science/article/pii/S030440762030378X>.
- [16] Leung, M. P. Network cluster-robust inference (2023). URL <https://arxiv.org/abs/2103.01470>. [arXiv:2103.01470](https://arxiv.org/abs/2103.01470).
- [17] NASA Jet Propulsion Laboratory. NASA SRTM Digital Elevation 30 m (SRTMGL1 v003). NASA LP DAAC, USGS/Earth Resources Observation and Science (EROS) Center (2013). URL [https://developers.google.com/earth-engine/datasets/catalog/USGS\\_SRTMGL1\\_003](https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003). Accessed via Google Earth Engine (dataset ID USGS/SRTMGL1\_003).
- [18] Farr, T. G. *et al.* The shuttle radar topography mission. *Reviews of Geophysics* **45**, RG2004 (2007).

- [19] Schiavina, M., Melchiorri, M. & Pesaresi, M. GHS-SMOD R2023A — GHS settlement layers, application of the degree of urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975–2030). European Commission, Joint Research Centre (JRC) (2023). URL [https://developers.google.com/earth-engine/datasets/catalog/JRC\\_GHSL\\_P2023A\\_GHS\\_SMOD\\_V2-0](https://developers.google.com/earth-engine/datasets/catalog/JRC_GHSL_P2023A_GHS_SMOD_V2-0).
- [20] Wikipedia contributors. Latitude — Wikipedia, the free encyclopedia (2024). URL [https://en.wikipedia.org/wiki/Latitude#Meridian\\_distance\\_on\\_the\\_ellipsoid](https://en.wikipedia.org/wiki/Latitude#Meridian_distance_on_the_ellipsoid). Section: Meridian distance on the ellipsoid. WGS84 series expansion accurate to 0.01 m/degree.
- [21] Lange, K., Richter, A. & Burrows, J. P. Variability of nitrogen oxide emission fluxes and lifetimes estimated from Sentinel-5P TROPOMI observations. *Atmospheric Chemistry and Physics* **22**, 2745–2767 (2022). Latitude-dependent NO<sub>x</sub> lifetime parameterization used in Beirle v2.
- [22] NIST. Nitrogen dioxide (NO<sub>2</sub>). NIST Chemistry WebBook, SRD 69 (2023). URL <https://webbook.nist.gov/cgi/cbook.cgi?ID=10102-44-0>. CAS 10102-44-0, Molar mass 46.0055 g/mol.
- [23] European Commission. ETS revision: No change to deadline to surrender allowances in 2023. Directorate-General for Climate Action (2023). URL [https://climate.ec.europa.eu/news-your-voice/news/ets-revision-no-change-deadline-surrender-allowances-2023-2023-01-30\\_en](https://climate.ec.europa.eu/news-your-voice/news/ets-revision-no-change-deadline-surrender-allowances-2023-2023-01-30_en). Accessed December 2024. Confirms compliance calendar: free allocation by 28 February, surrender by 30 April.
- [24] Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociological Methods & Research* (2022).
- [25] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual (2024). URL <https://www.gurobi.com>.
- [26] UK Government. UK Emissions Trading Scheme (UK ETS): A policy overview. UK Government Policy Paper (2024). URL <https://www.gov.uk/government/publications/uk-emissions-trading-scheme-uk-ets-policy-overview/uk-emissions-trading-scheme-uk-ets-a-policy-overview>. Accessed December 2024.
- [27] European Parliament and Council. Directive 2003/87/EC establishing a scheme for greenhouse gas emission allowance trading within the Community (2003). URL <https://eur-lex.europa.eu/eli/dir/2003/87/oj>. Annex I defines activity codes: Code 1 (Phases 1–2) = ‘Combustion installations with rated thermal input exceeding 20 MW’; Code 20 (Phase 3+) = ‘Combustion of fuels’. Consolidated version at <https://eur-lex.europa.eu/eli/dir/2003/87/2024-03-01>.

- [28] European Environment Agency. EU ETS data viewer: User manual and background note. Technical Document, European Environment Agency (2021). URL <https://www.eea.europa.eu/data-and-maps/data/european-union-emissions-trading-scheme-12/eu-ets-background-note>. Table 6-1 provides activity codes used in EUTL database; codes 1, 20 identify combustion/-electricity sector installations.