A Geospatial Approach to Measuring Economic Activity*

Anton Yang[1], Jianwei Ai[2,3], and Costas Arkolakis[1]

[1]Yale University
[2]Renmin University of China
[3]Cornell University

March 20, 2025

Abstract

We introduce a new methodology to detect and measure economic activity using geospatial data and apply it to steel production, a major industrial pollution source worldwide. Combining plant output data with geospatial data, such as ambient air pollutants, nighttime lights, and temperature, we train machine learning models to predict plant locations and output. We identify about 40% (70%) of plants missing from the training sample within a 1 km (5 km) radius and achieve R^2 above 0.8 for output prediction at a 1 km grid and at the plant level, as well as for both regional and time series validations. Our approach can be adapted to other industries and regions, and used by policymakers and researchers to track and measure industrial activity in near real time.

1 Introduction

Economic measurement is the cornerstone of the economic discipline, yet traditional methods, particularly for industrial output, often fall short of providing up-to-date and granular information. Census-based surveys or government reports, the mainstay of industrial output measurement, are typically infrequent, incomplete, or prohibitively expensive to collect. For instance, according to the United Nations, only a small portion of developing countries collect industrial statistics on an annual basis, while many conduct surveys in 5 or 10-year intervals, and others have not done so for over 15 years (Upadhyaya and Todorov, 2009), and in Sub-Saharan Africa,

*Yang: Yale University (e-mail: anton.yang@yale.edu); Ai: Renmin University of China and Cornell University (e-mail: aijianwei@ruc.edu.cn); Arkolakis: Yale University and NBER (e-mail: costas.arkolakis@yale.edu). We thank Adel Daoud, Amit Khandelwal, Jennifer Marlon, Joseph Shapiro, Kei Irizawa, Lance Pangilinan, Sam Kortum, Tillmann von Carnap, Tianyu Fan, Yuansen Li, Mushfiq Mobarak, Hyunjoo Yang, and the seminar participants at Yale University and Yale Center for Geospatial Solutions for their helpful feedback and suggestions. Part of this work was carried out while Ai was visiting the Department of Economics at Yale University. He gratefully acknowledges their hospitality and the financial support provided by the China Scholarship Council. Ziyang Long and Meng Xia provided excellent research assistance. This is a short version of our work that excludes various sections. All errors are our own.

fewer than half of countries have published reliable industrial production metrics for many years (United Nations Industrial Development Organization, 2016).

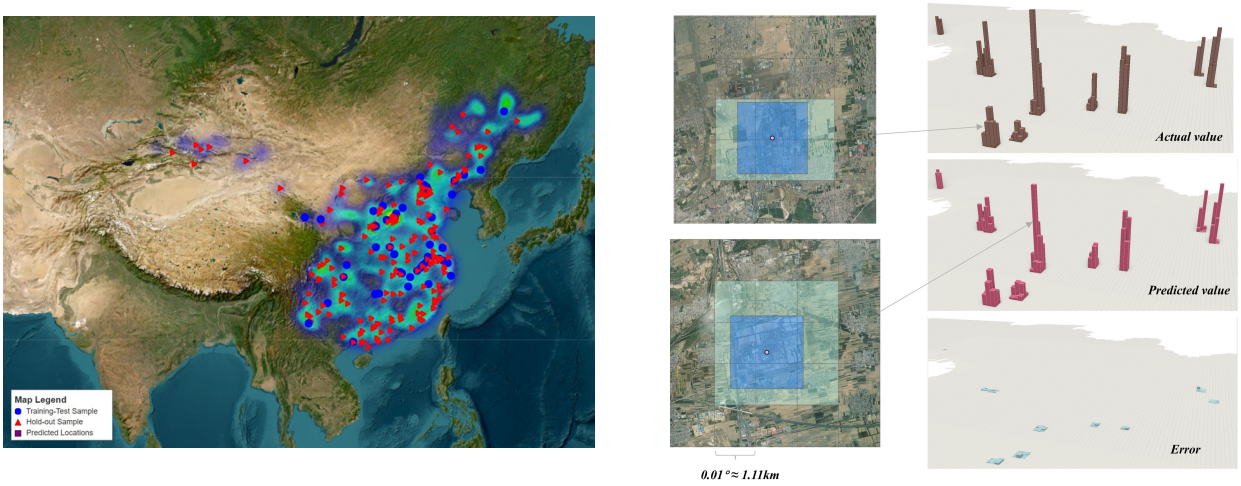
Geospatial data—e.g., satellite data—have emerged as a new tool for addressing these limitations, offering a cost-effective and comprehensive alternative (Burke et al., 2021). For example, researchers developed methods to estimate GDP using nighttime light (NTL) data from satellites (Henderson, Storeygard and Weil, 2012; Nordhaus and Chen, 2015; Martinez, 2022), but these studies focus on aggregate-level predictions and often remain inaccurate at times (Chen and Nordhaus, 2019). Recent advancements have focused on integrating satellite data with machine learning (Khachiyan et al., 2022; Sherman et al., 2023; Vogel et al., 2024), but these procedures more accurately reflect variation in population than in GDP per capita (Khachiyan et al., 2022; Ahn et al., 2023). Rossi-Hansberg and Zhang (2025) adapts such methodologies to measure regional GDP with high-resolution data across the globe and shows the importance of introducing geospatial indicators such as CO_2.

We develop a methodology to detect economic activity and measure its intensity with high spatial and temporal precision. Unlike alternative GDP estimates derived from satellite data, our approach directly links industrial output and location to real-time environmental signals, providing a more accurate and scalable economic monitoring tool. The premise of the methodology is that industrial activity has a distinct environmental imprint, so that both its location and intensity can be remotely measured by a variety of geospatial indicators. We thus choose to use it to detect and predict economic activity in the steel industry, known for its major economic role yet also stands as a global major polluter. The methodology involves two steps: (1) predicting plant locations within grid cells by matching known coordinates of steel plants with data of various geospatial datasets and applying a standard neural network model; (2) using data to predict crude steel output based on these identified plant locations.

The first main contribution of our paper is to illustrate how to accurately predict plant locations not reported in training and testing datasets (Figure 1a). In this step, we employ a neural network model to predict the presence of steel production in a dataset containing approximately 1.4 million grids. We identify approximately 40% of plants missing from the training sample within a 1 km radius and 70% within a 5 km radius. Using a local interpretation method based on the Shapley value concept from cooperative game theory, we find that particulate matter (PM), including PM_{10} and $PM_{2.5}$, is a key predictor of plant locations.

Our second main contribution is to show it is possible to achieve high accuracy in predicting steel output (Figure 1b) by associating various environmental indicators with ground-truth data and training our model using standard machine learning techniques. In this step, we apply tree-based machine learning models to predict steel output at both the 1 km grid and at the plant levels. Our findings show that ozone (O_3), NTL, and heat are key predictors of steel output at the grid level, while O_3 and NTL data are the key indicators at the plant level. We perform K-fold cross-validation on the full sample (Stone, 1974); and complement it with additional validations leaving a selected region and time period from the training sample. We find that our predictions

Figure 1: Predicted Locations and Output of Steel Plants



(a) Plant Locations (b) Plant Output

Note: The left panel (a) shows our predicted plant locations across China using Esri Satellite Imagery. The blue circle points represent plants with complete monthly output data from 2019 to 2022 used in the training, while the red triangle points correspond to the plants observed from the holdout sample. In the northwest region, the predicted locations (purple 'foggy') are derived from areas not included in the training sample. The right panel (b) shows a 3D map of crude steel output predictions. Figure A.6 shows the distribution of steel output across China.

of steel output using geospatial statistics fit very well with an R^2 above 0.8 at the plant level.

To provide further validation of our methodology we show that it can capture output fluctuations from two major events. The first is the Spring Festival, which typically occurs from late January to February, during which steel production drops significantly due to temporary shutdowns for the holiday, with operations quickly resuming afterward. Our model accurately predicts this decrease and the quick recovery afterward. The second event is the unparalleled COVID-19 pandemic, which lasted from early 2020 to the end of 2022. In Wuhan, the pandemic's epicenter, strict lockdowns led to an unprecedented and steep drop in steel production. While steel plants outside Wuhan also experienced declines, these were notably less severe and exhibited less volatility compared to the extraordinary disruption observed in Wuhan. Our predictions fit well with these broad patterns and show consistent trends with the reported output.

2 Steel Production and Geospatial Environmental Factors

As one of the most intensive industrial processes, steel production releases large amounts of air pollutants and heat, which can be observed remotely. The production process involves three main stages: mining iron ore, turning the ore into iron, and smelting the iron into steel. In traditional blast furnaces, iron ore is heated with metallurgical coke at high temperatures to produce molten iron or pig iron. This process uses a hot air blast containing oxygen and releases substantial carbon dioxide (CO_2) and carbon monoxide (CO), the latter of which can further oxidize into CO_2.

Electric arc furnaces may still produce pollution if powered by electricity from nonrenewable sources. Other air pollutants from steel production include PM, sulfur dioxide (SO_2), and nitrogen oxides (NO_x). Steel industries are major emitters of volatile organic compounds (VOC*s*) and NO_x, which under certain atmospheric conditions may form ozone (O_3) as a by-product. Meanwhile, steel plants have high costs associated with shutting down and restarting operations, and thus typically operate continuously, both day and night. This operational characteristic results in consistent heat emissions and potentially NTL, making temperature and NTL data instrumental in detecting industrial activity (Liu et al., 2018; Zhang et al., 2019; Xie et al., 2024). We detail the data sources and dataset construction below.

2.1 Geospatial Environmental Indicators Used as Predictors

We use the constructed dataset of geospatial indicators to predict steel locations and output. We harmonize each of the geospatial indicators at 1 km $\times$ 1 km grid cell, which provides approximately 1.4 million data points for the entire China. This allows us to associate emissions with industrial activity using high-quality, high-frequency pollution data obtained at fine-grained resolutions (Wei et al., 2021*a*,*b*, 2022*a*,*b*; Cooper et al., 2022; Halder et al., 2023; Wei et al., 2023).

Remote sensing data collected by satellites includes ambient air pollutants, NTL, and land surface temperature (LST). Ambient air pollution data are sourced from two main datasets. The first is ChinaHighAirPollutants (CHAP), a high-resolution dataset specific to China that provides high-quality geospatial data, including satellite remote sensing observations and ground-based measurements (Wei et al., 2023). The second is Sentinel-5P, a fully open-access dataset with a lower spatial resolution that does not include particulate matter measurements. We use CHAP to obtain our primary results and Sentinel-5P for robustness checks.

The CHAP dataset provides key indicators, including concentrations of NO_2, SO_2, O_3, CO, $PM_{2.5}$, PM_{10}, at 1 km resolution. LST and NTL intensity come from NASA and NOAA satellite observations, as well as ground-based monitoring stations. Missing LST data are filled using ground measurements (Tang et al., 2024).

2.2 Steel Plants Output and Location

For ground-truth data on steel production, we use information from 146 steel plants, but only 70 are included in our machine learning model because of missing values in production or environmental indicators. These data are provided by the Chinese Iron and Steel Association (CISA), which collects monthly production volumes for crude steel and pig iron and provides plant coordinates (see also Brandt et al., 2022). The 146 plants represent over 70% of China's steel production capacity. The dataset includes details on production technology (blast furnace, electric arc furnace, or integrated processes). To verify the accuracy of these output data, we compare CISA's data with aggregate-level data from the National Bureau of Statistics of China and find the two data sources to be highly consistent, except that CISA's data are available at a much finer

Figure 2: Steel Plants from Satellite Images



(a) Example of Steel Plant 1 (b) Example of Steel Plant 2 (c) Squared Grid Construction

Note: These figures show the locations and shapes of steel plants identified from satellite imagery. Panels (a) and (b) show two examples of steel plants in China. We know each plant's POI, but the coordinates can sometimes be inaccurate, as shown in Panel (b) (i.e., the pink dot outside the red-lined polygon). We verify each POI and its location individually. Panel (c) shows how we construct the shape of plants in our machine learning model. To facilitate replicability, we use a 2 km square to represent each plant instead of creating detailed polygons, as shown in Panel (c). Appendix A.4 shows the typical size distribution of steel plant areas in China. We replicate the main exercise introduced in this paper using detailed polygons and find that the results are similar.

level of granularity (Figure A.3).

We then verify the location of each steel plant by cross-checking its point of interest (POI) with satellite images. Since POI coordinates can occasionally be imprecise, we manually adjust them to ensure alignment with the actual plant locations. To make the data easier to replicate, we represent each plant using a 2 km $\times$ 2 km square instead of detailed shapes, based on the typical size of steel plants in China. Figure 2 provides three illustrative examples.

We match the ground-truth output data with the input data using spatial information from ambient air pollutant datasets. The 1 km grid cells are defined by their latitude and longitude coordinates based on geolocation information provided in the CISA dataset for each of the 70 observed plants and the associated grid cells. Steel output is calculated at the grid level using an area-weighted approach that aggregates grid-level data to the plant level based on the proportion of each grid covered by the plant. The calculation is as follows:

$$Y_g = \sum_{p \in \mathcal{P}_g} \frac{S_{gp}}{S_p} P_p, \tag{1}$$

where $\mathcal{P}_g$ is the set of plants overlapping grid g, S_{gp} is the area of grid g that overlaps with plant p, S_p is the total area of plant p, P_p is the total production of plant p, and Y_g is the total production output assigned to grid g.

Equation (1) calculates grid-level production by summing contributions from all plants that intersect with a grid. Note that a single grid cell can partially contain multiple plants, and a single plant may span several grid cells.

3 Method

We predict two primary outcomes: the locations of steel plants and their production output. Because these predictions rely on labeled data, our approach falls under supervised learning (Athey and Imbens, 2019). Specifically, we use (i) a neural network model to identify plant locations and (ii) a regression model to estimate plant output at both the grid and plant levels.

3.1 Model framework

First, we predict the locations of steel plants at the grid-cell level using a deep learning model to estimate the probability of a plant being located in a specific grid cell. As expected, our dataset contains far fewer grid cells with steel plants (minority class) than those without (majority class), potentially leading to class imbalance (Leevy et al., 2018). This imbalance can bias the model, causing it to perform well on the majority class but poorly on the minority class since the model effectively assigns more weight to the majority class during training. To address this issue, we apply the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which generates synthetic examples of grid cells with steel plants to balance the dataset, allowing the model to learn effectively from both classes (Chawla et al., 2002).

Next, we apply a neural network classifier to the SMOTE-balanced dataset to classify grid cells associated with steel plants. Neural networks are well-suited for this task as they handle large, sparse datasets and capture complex non-linear relationships in the data (LeCun, Bengio and Hinton, 2015). We use the `ReLU` activation function (Nair and Hinton, 2010), Dropout layers to prevent overfitting (Srivastava et al., 2014), and binary cross-entropy as the loss function (Murphy, 2012). The model is trained for 100 epochs, with accuracy as the performance metric. The dataset is split into 80% for training and 20% for testing (Hastie, Tibshirani and Friedman, 2009).

We use tree-based models to predict crude steel production.[1] While deep learning excels with large text and image data, tree-based methods remain the state-of-the-art for medium-sized datasets (Grinsztajn, Oyallon and Varoquaux, 2022). Specifically, we apply XGBoost and other gradient-boosted regression tree models, supervised learning models that build an ensemble of shallow trees sequentially, where each tree corrects errors made by the previous ones, to map input features to the production outputs.

To fix ideas, we model the conditional expectation of Y based on a set of p predictors, $\mathbf{X}$, using the following specification:

$$Y = f\left(X_1, X_2, \ldots, X_p\right) + \epsilon, \tag{2}$$

where Y represents the log of crude steel measured at both the grid and plant levels, $X_1, X_2, \ldots, X_p$

[1] We note that pig iron is a crucial intermediate in crude steel production, but including both pig iron data and environmental indicators in our predictive model is likely not a robust approach due to strong multicollinearity. To validate this point, we replicate the same procedure to predict pig iron location and output. Our R^2 for predicting pig iron exceeds 0.8, and the fact that the resulting feature importance largely overlaps with those in predicting steel confirms a high correlation between pig iron and the environmental imprints used to predict steel.

represent the predictor variables defining the feature vector $\mathbf{X}$, and ϵ is the prediction error, with $\mathbb{E}[Y \mid X_1, X_2, \ldots, X_p] = f(X_1, X_2, \ldots, X_p)$ and zero conditional expectation. To evaluate and compare model performance, we apply several models, including linear regression, Lasso, kernel ridge regression, ElasticNet, random forest, gradient boosting, LightGBM, and XGBoost. We also use an ensemble model, which combines multiple machine learning algorithms to improve predictive accuracy and robustness. It also reduces overfitting and bias while enhancing generalization (Wolpert, 1992). We show model details in Appendix B.

3.2 Input Features

We link the data from Section 2 to the model's input features used as predictors. To optimize the model's ability to extract information from pollution data, we create three distinct feature groups. The first group consists of ambient air pollutants. The second group consists of the centroid latitude and longitude of each grid and each plant, as well as the steel plant's production mode: electric, blast furnace, and mixed (which combines blast furnace and electric). The third group includes additional environmental factors LST, NTL intensity, along with other features: year, month, and cyclical characteristics derived from trigonometric functions, $\sin(2\pi \cdot \text{Month}_t/12)$ and $\cos(2\pi \cdot \text{Month}_t/12)$, to account for seasonal variations.

In the first step, when predicting whether a grid contains a steel plant, we exclude the second group of features. In the second step, when predicting output levels, we use all the input features discussed above. We additionally incorporate lagged geospatial features to capture temporal trends in production.

3.3 Training and Validation

We follow standard practice in machine learning by splitting the data into training and testing sets. In the first step, we focus on grid cells throughout China but restrict the data to November of each year from 2019 to 2022.

In the second step, we train two output simulation models: the grid-level model using grids of cities with steel plants and the plant-level model using steel plants in our sample with non-zero output. Due to the large size of our dataset, we evaluate our model using K-fold cross-validation, where each fold produces a train-test split with 80% training and 20% testing. The labeled data are pooled and randomly partitioned into five equal subsets for each grid or plant. A model is trained on four subsets and evaluated on the remaining subset. This process is repeated for all five-folds, and the final performance metrics are averaged across the iterations. This technique allows us to assess model performance on different subsets of the data, reducing the risk of overfitting and ensuring that the model generalizes well to out-of-sample data. Ideally, the process is repeated multiple times to estimate the distribution of out-of-sample errors.

We employ hyper-parameter tuning to strengthen the model's focus on learning from pollutant features while improving its overall performance and interpretability (Hazan, Klivans and

Yuan, 2017). We prioritize environmental features because they are direct indicators of steel production. To mitigate overfitting, we apply standard L1 and L2 regularization, which are commonly referred to as Lasso and Ridge penalties (Tibshirani, 1996; Hoerl and Kennard, 1970).

4 Main Results

We assess our model performance using several common metrics. For the neural network model, we use accuracy, precision, recall, and area under the receiver operating characteristic curve (AUROC). The AUROC measures the probability that the classifier ranks a randomly chosen positive example higher than a randomly chosen negative example, with scores ranging from 0.5 (random classifier) to 1.0 (perfect predictor). For the output prediction we simply report the R^2.

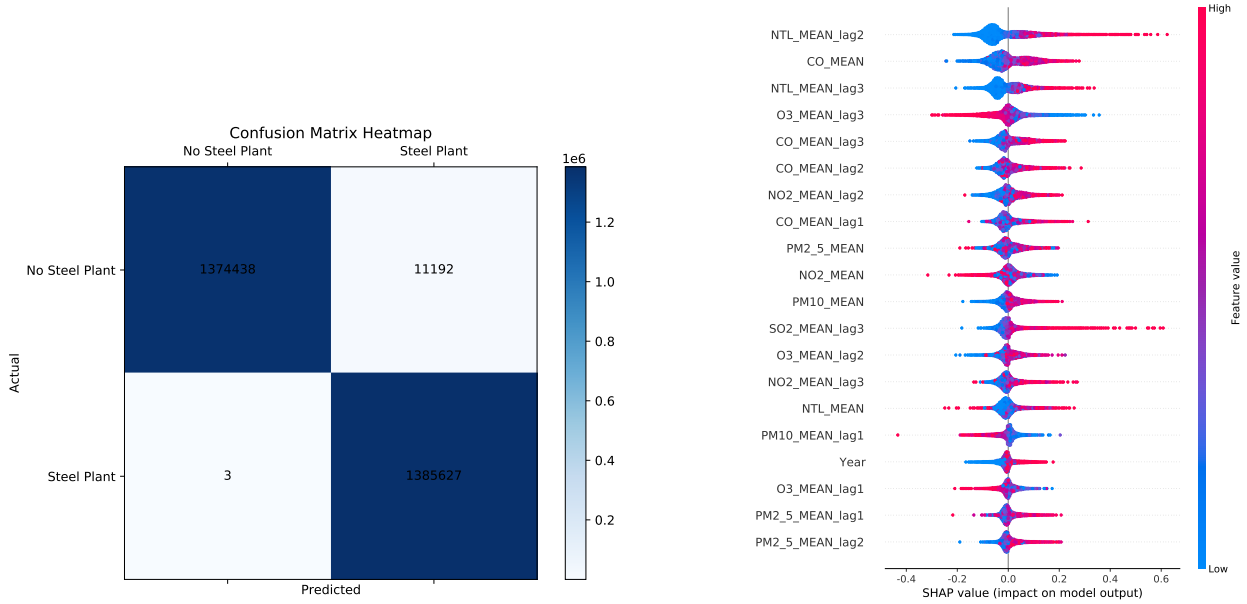
4.1 Location Prediction

We use our model to predict steel plant locations and enhance interpretability with SHapley Additive exPlanations (SHAP), a model-agnostic post hoc method. Our goal is to train a classification model that uses input features, including geospatial data and spatial attributes, to predict whether a given grid cell contains a steel plant. For training and testing, we include plants that have complete monthly output data from 2019 to 2022. To test out-of-sample performance, we use plants from the CISA list that are excluded from the training sample as the holdout set.

Using the open-source Global Steel Plant Tracker as a holdout sample, we identify nearly 40% of the plants missing from the CISA dataset within a 1 km radius and 70% within a 5 km radius. Increasing the radius to 10 km and applying the lowest probability threshold (>0.5) increases the coverage to nearly 90% (Figure A.7). Our model successfully predicts plant locations in northwestern China that are not reported by the CISA (Figure 1a). We then use the confusion matrix in Figure 3a to evaluate the model's performance in predicting plant locations across over 1.3 million grid cells after applying SMOTE. With a standard threshold of 0.5, the matrix shows a high count of true negatives (1,374,438), where the model correctly identifies grids without steel plants. False negatives and false positives—where the model either misses or incorrectly predicts the presence of steel plants—are relatively low, at 3 and 11,192 instances, respectively. Although there are some false positives, the spatial clustering shown previously (Figure 1a) implies that these misclassifications typically occur in close proximity to actual steel plant locations, which suggests that the model's errors are often geographically near true positives. Meanwhile, true positives are notably high, at 1,385,627 instances. The model achieves near-perfect discrimination between steel and non-steel grids (AUROC close to 1.0) (Figures A.10) and maintains a strong balance between precision and recall, as reflected by a high harmonic mean (99.2%). Precision is exceptionally high (98.5%), suggesting that when the model flags a grid as containing a steel plant, it is almost always correct (Figure 3a).

To evaluate model performance and interpret the results, we examine feature importance to understand the relationship between the outcome (i.e., dependent variable) and input features

Figure 3: Confusion Matrix and Shapley Values in the Classification Model



(a) Confusion Matrix

(b) Shapley Value

Note: The left panel (a) shows the confusion matrix. SMOTE effectively addresses class imbalance by synthesizing new data for minority classes (steel-producing grids). A standard threshold of 0.5 is used for the confusion matrix. The left panel (b) shows the Shapley values of each feature in our classification model. The Shapley value shows the contribution of each feature to the model's prediction. Each dot represents a sample in the model, and the density reflects the distribution of feature values.

(i.e., independent variables). To address the common criticism of machine learning models as "black boxes" and to clarify how inputs affect outputs, we use SHAP to quantify the contribution of each feature to the model's predictions. SHAP is a widely used approach for interpreting the impact of individual features on model outcomes (Lundberg, 2017). In contrast to Partial Dependence Plots (PDP), which focus on average effects, SHAP's game-theoretic foundation provides feature contributions at the observation level. The SHAP decomposition can be equivalently written as follows:

$$\widehat{Y}_i = \beta_0 + \text{shap}\left(X_{1,i}\right) + \text{shap}\left(X_{2,i}\right) + \cdots + \text{shap}\left(X_{p,i}\right), \tag{3}$$

where $\widehat{Y}_i$ is the model prediction for the observation of i, β_0 is the mean prediction of the model across all observations (referred to as the base value of the model output), i.e., the prediction without any inputs, and $\text{shap}\left(X_{p,i}\right)$ is the marginal contribution of feature p for observation i. Thus, the sum of Shapley terms in Equation (3) equals the difference between the actual prediction and the average prediction.[2]

[2]We use a specialized SHAP method tailored to geo-referenced data. Specifically, we use the GeoShapley package, a game theory-based approach for measuring spatial effects in machine learning models (Li, 2024).

Our model accurately identifies known steel plants with few false positives. Figure 3b shows the SHAP feature importance plot, which captures the average contribution of each feature to the model's predictions. Each dot represents the impact of a single feature value. SHAP values on the x-axis denote the magnitude and direction of a feature's impact on the model's output. We find that our location prediction heavily relies on NTL, NO_2, and CO. NTL data is particularly useful, as continuously operating steel plants emit significant light at night, making it a strong indicator of industrial activity. NO_2 and CO are also critical, as they are direct byproducts of the steel-making process.

4.2 Output Prediction

Given the predicted locations, we perform an intensive margin analysis to predict steel output. We estimate steel output at a finer grid-level resolution and predict output for each plant.

Grid-level Prediction. To estimate steel output across grid cells, we apply several machine learning models trained on remote sensing data. Our results show that the model is highly accurate in predicting steel production. We provide a detailed analysis of model performance: predicted versus actual values, and feature importance.

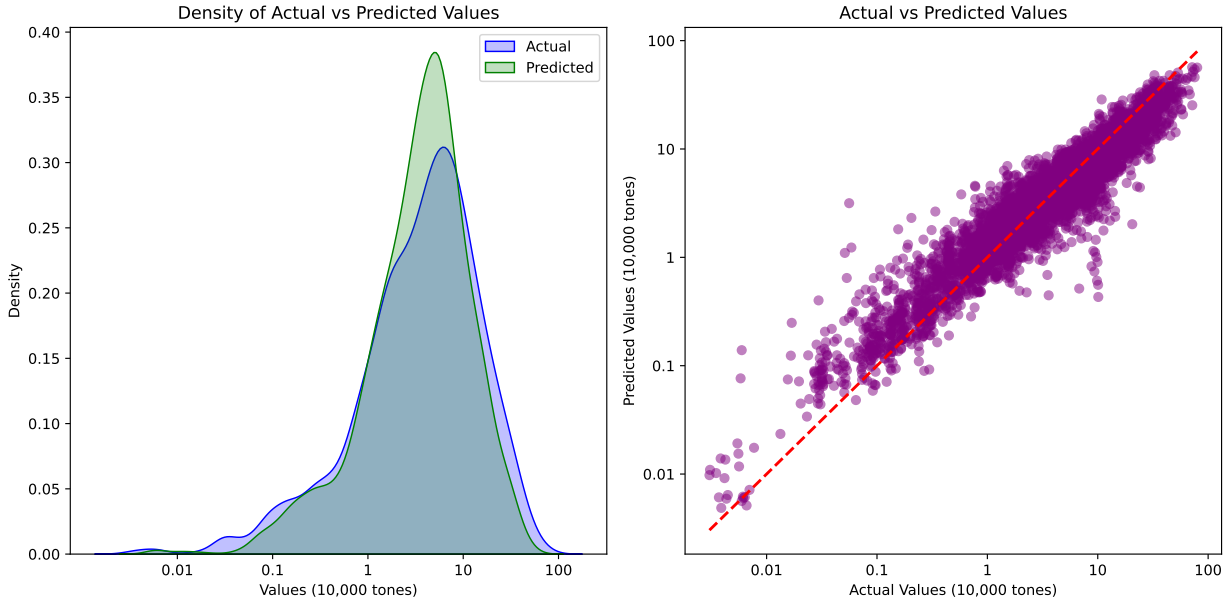
We evaluate several models to determine the most effective approach for predicting steel output. Table A.2 compares their performance. Gradient-boosting models perform well, but the ensemble learning model, which combines multiple algorithms, achieves the highest R^2 value of 0.934. Figure 4 compares actual and predicted steel output at the grid level. The scatter plot shows a strong correlation, with points clustering around the diagonal red line representing perfect predictions. This indicates that the model reliably captures the spatial distribution of outputs across regions. The distribution plots show that the predicted values closely mirror the actual distribution, further validating the accuracy of the model.

To understand the key drivers behind steel output predictions, we analyze feature importance using SHAP values, which quantify the contribution of each feature. The SHAP values are efficiently estimated by the tree-based SHAP algorithm (Lundberg, 2017; Lundberg, Erion and Lee, 2018). Finally, we use a non-parametric bootstrap to estimate the uncertainties in SHAP value estimates. Figure 5a and 5b show the lagged values of pollutants, O_3, and LST, which suggest that past environmental conditions are also crucial in predicting output.

Plant-level Prediction. We estimate steel production using identified plant locations and various machine learning models trained on features such as pollutant levels, temperature data, and NTL intensity. Figure A.12 shows that predicted outputs closely align with actual values, indicating high accuracy; the density plot also demonstrates a tight match between predicted and observed distributions.

To understand what drives these predictions, we examine feature importance scores. Our results show that pollutant concentrations, including O_3, $PM_{2.5}$, PM_{10}, and SO_2, are the most

Figure 4: Actual and Predicted Value at the Grid Level



Note: This figure shows the model fit at the grid level in our sample. The left-hand side shows the distribution fit between actual and predicted values, while the right-hand side shows the value of crude steel output fit between actual and predicted values expressed in units of 10,000 tons.

significant contributors, followed by NTL and land surface temperature. The XGBoost model achieves a strong R^2 of 0.88 at the plant level. While prior studies typically produce grid-level estimates (Ahn et al., 2023), our approach allows for granular plant-level predictions.

4.3 Robustness Check Using Open-Source Sentinel Data

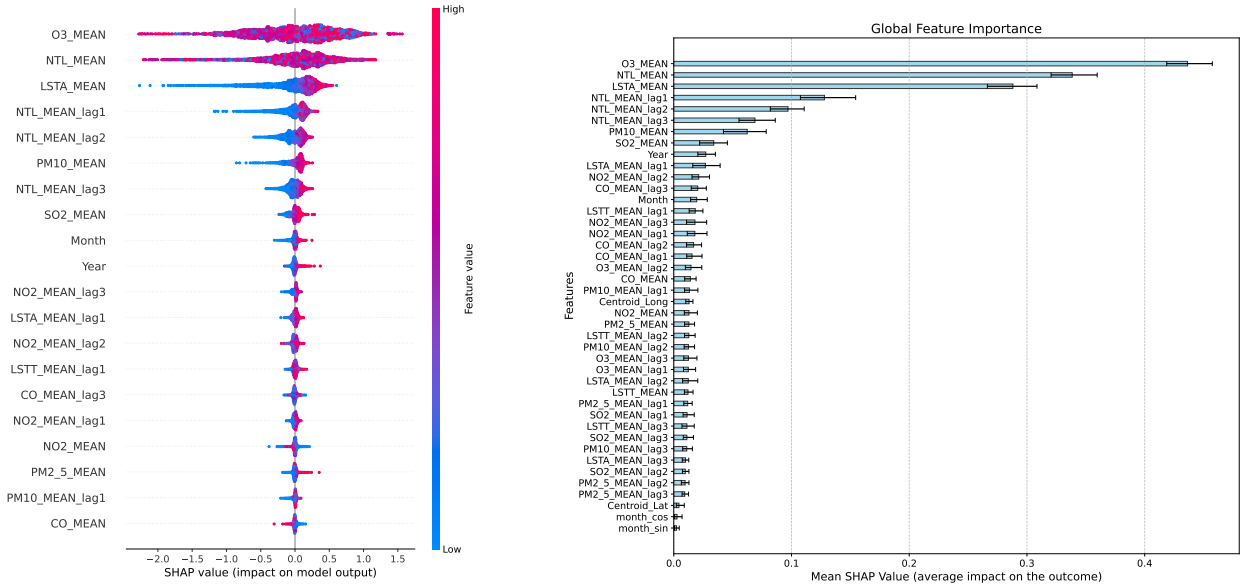
We use air pollutant data from Sentinel-5P to predict output at both grid and plant levels, following the same procedure. Sentinel-5P measures air quality at a resolution of 3.5 km $\times$ 7 km, while the CHAP dataset provides finer 1 km resolution data for China. To make the two datasets compatible, we resample the Sentinel-5P data down to the 1 km grid of the CHAP dataset using the Google Earth Engine platform. Our results show R^2 values above 0.8 for both grid- and plant-level predictions, which suggests that our method remains robust even when using open-source satellite data with lower quality but greater accessibility (Figures A.13 and A.14).

4.4 Event Analysis

We assess the impact of two major events—the Spring Festival and the COVID-19 pandemic—and evaluate whether our methodology can capture fluctuations in these periods. We view this as a basic validation of predictions of the model for output aggregated across regions.

During the Spring Festival, which typically occurs from late January to February, steel plants

Figure 5: Shapley Values for Features in Regression Model at the Grid Level



(a) Shapley Values

(b) Simulation-Based SHAP Values

Note: The left panel (a) shows the Shapley values estimated using a non-parametric bootstrap method based on residual resampling. The right panel (b) shows the simulation-based SHAP values recalculated using the same approach. In both panels, model residuals were resampled with replacement and added to the predicted values to generate new dependent variables. The XGBoost model was retrained using the original hyperparameters, and SHAP values were recalculated and stored. We repeat the process 5,000 times and present 95% confidence intervals.

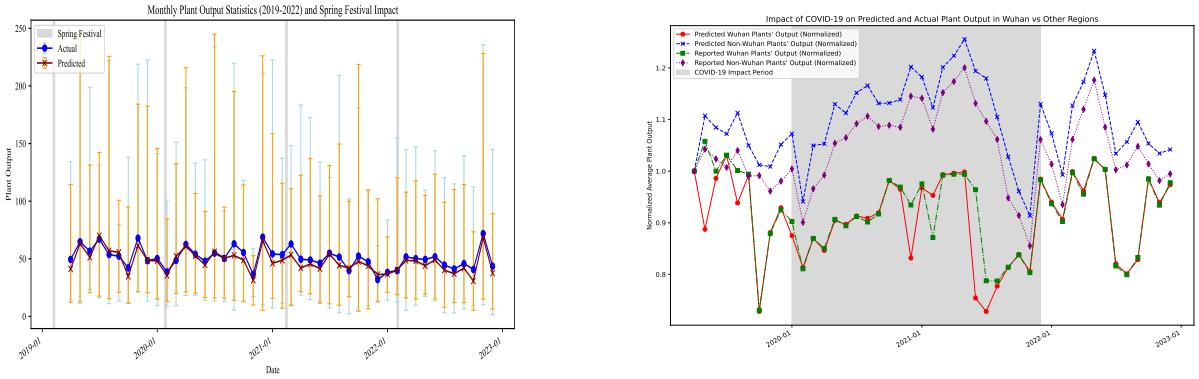
temporarily reduce output due to holiday shutdowns. Figure 6a shows these consistent declines, which are short-lived and followed by rapid rebounds as operations resume. The model accurately predicts these fluctuations and demonstrates its capability to forecast aggregate production trends driven by seasonal factors.

The COVID-19 pandemic, which began in early 2020 and extended toward the end of 2022, caused a sharp and sustained decline in steel production, particularly in Wuhan, the outbreak's epicenter, as shown in Figure 6b. While non-Wuhan plants also experienced declines due to lockdowns and shutdowns, the impact was less severe and there was an immediate rebound, suggesting that the pandemic affected the entire country but had a disproportionately intense impact in Wuhan. Our predictions fit closely with these trends, with only minor underestimation of output during the middle and end of the pandemic period.

5 Generalization Capabilities: Holdout Validation

We now validate the predictions of our model across different regions and time periods using holdout data, excluded from the training and testing sets.

Figure 6: Event Analysis on Spring Festivals and COVID-19



(a) Plant Output During the Spring Festival

(b) Plant Output During COVID-19

Note: Figure 6a shows the crude steel's actual and predicted values during the Spring Festival. The gray bars are the periods of the Spring Festival each year. The yellow vertical line indicates the monthly range of steel output. Figure 6b shows our actual and predicted values during the COVID-19 period. We normalized the values to those of the first month, January 2019. The data source for this figure is the reported plant output from the CISA. We repeat the process 5,000 times and calculate the arithmetic mean.

5.1 Time Series Validation

For the time series validation, we examine steel output trends over a specific period. Specifically, we train the model on historical data from 2019 to 2021 and generate predictions for 2022. That is, we exclude 2022 from the training and testing sets, treating it as an external sample.

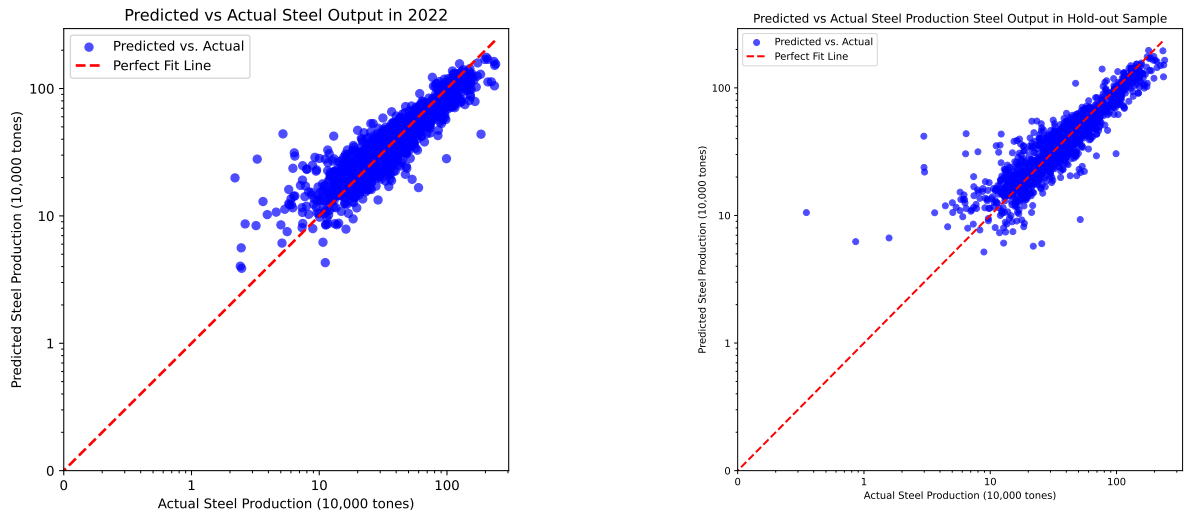
The model effectively captures the temporal variability of steel output. As shown in Figure 6a, predicted values closely align with actual production in 2022, particularly during significant shifts. For instance, the model accurately predicts lower output during winter months due to environmental regulations and higher levels during peak industrial activity.[3] Figure 7a demonstrates strong performance when we validate the model with 2022 data, with an R^2 of 0.84. These results suggest the model reliably tracks steel production fluctuations over time and can potentially predict future years.

5.2 Regional Validation

The regional validation examines the model's ability to predict output in regions without ground-truth data. To do so, we train the model on steel production data from one region and use another region as an external validation sample. Specifically, we train the model on data from northern China and validate it using steel output data from southern China. As shown in Figure 7b, the model achieves an R^2 of 0.81 in predicting output in the southern region. Although some outliers fall outside the training sample, the model successfully captures overall trends across many grids. Our result shows that the model does not exclusively depend on localized

[3]For instance, China's winter coal substitution policies, which aim to reduce air pollution by replacing coal with cleaner energy sources for heating in northern regions during winter.

Figure 7: Using External Sample as a Holdout Validation



(a) 2022 as External Sample

(b) Southern China as External Sample

Note: Figure 7a shows the model forecasting capability using 2022 as an external validation. We use the years from 2019 to 2021 to train and test the model, and then use 2022 as the holdout sample to validate the model (70 plants $\times$ 12 months $=$ 840 observations). Figure 7b shows the model generalization capability using Southern China as an external validation. We train and test our model on data from Northern China, and use Southern China as the holdout sample to validate the model (31 plants $\times$ 48 months $=$ 1,488 observations).

patterns in China. These findings suggest that our procedure has the potential to generalize across different spatial and temporal contexts.

6 Conclusions

The use of geospatial data to measure economic statistics has been gradually increasing in economics but has also faced criticism for its various limitations (Chen and Nordhaus, 2019; Ahn et al., 2023). We develop a new methodology to measure economic activity that relies on the premise that certain industries have distinct environmental imprints detectable remotely. This methodology involves two steps, each of its own interest: detecting the location of production and measuring its intensity at each location.

In addition to measurement, we believe our approach provides two additional insights. First, it demonstrates the possibility of using the trained model to predict plant locations not included in our ground-truth sample. Second, it shows that environmental footprints can be applied beyond the time periods used to train our model, suggesting predictive power that extends to future time horizons. Our methodology can be adapted to industries beyond steel, regions beyond China, and used to measure economic activity across a broader spectrum.

References

Ahn, Donghyun, Minhyuk Song, Seungeon Lee, Yubin Choi, Jihee Kim, Sangyoon Park, Hyunjoo Yang, and Meeyoung Cha. 2023. "Fine-Grained Socioeconomic Prediction from Satellite Images with Distributional Adjustment." 3717–3721.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics*, 11(1): 685–725.

Brandt, Loren, Feitao Jiang, Yao Luo, and Yingjun Su. 2022. "Ownership and productivity in vertically integrated firms: evidence from the Chinese steel industry." *Review of Economics and Statistics*, 104(1): 101–115.

Burke, Marshall, Anne Driscoll, David B Lobell, and Stefano Ermon. 2021. "Using satellite imagery to understand and promote sustainable development." *Science*, 371(6535): eabe8628.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A scalable tree boosting system." 785–794.

Chen, Xi, and William D Nordhaus. 2019. "VIIRS nighttime lights in the estimation of cross-sectional and time-series GDP." *Remote Sensing*, 11(9): 1057.

Cooper, Matthew J, Randall V Martin, Melanie S Hammer, Pieternel F Levelt, Pepijn Veefkind, Lok N Lamsal, Nickolay A Krotkov, Jeffrey R Brook, and Chris A McLinden. 2022. "Global fine-scale changes in ambient NO2 during COVID-19 lockdowns." *Nature*, 601(7893): 380–387.

Datta, Anupam, Shayak Sen, and Yair Zick. 2016. "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems." 598–617, IEEE.

Efron, Bradley. 1992. "Bootstrap methods: another look at the jackknife." In *Breakthroughs in Statistics: Methodology and Distribution*. 569–593. Springer.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in Neural Information Processing Systems*, 35: 507–520.

Halder, Bijay, Iman Ahmadianfar, Salim Heddam, Zainab Haider Mussa, Leonardo Goliatt, Mou Leong Tan, Zulfaqar Sa'adi, Zainab Al-Khafaji, Nadhir Al-Ansari, Ali H Jawad, et al. 2023. "Machine learning-based country-level annual air pollutants exploration using Sentinel-5P and Google Earth Engine." *Scientific Reports*, 13(1): 7968.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Vol. 2, Springer.

Hazan, Elad, Adam Klivans, and Yang Yuan. 2017. "Hyperparameter optimization: A spectral approach." *arXiv preprint arXiv:1706.00764*.

Henderson, J Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring economic growth from outer space." *American Economic Review*, 102(2): 994–1028.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. "The ERA5 global reanalysis." *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049.

Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12(1): 55–67.

Khachiyan, Arman, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, and Amit K Khandelwal. 2022. "Using neural networks to predict microspatial economic growth." *American Economic Review: Insights*, 4(4): 491–506.

Kossen, Jannik, Neil Band, Clare Lyle, Aidan N Gomez, Thomas Rainforth, and Yarin Gal. 2021. "Self-attention between datapoints: Going beyond individual input-output pairs in deep learning." *Advances in Neural Information Processing Systems*, 34: 28742–28756.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature*, 521(7553): 436–444.

Leevy, Joffrey L, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. 2018. "A survey on addressing high-class imbalance in big data." *Journal of Big Data*, 5(1): 1–30.

Lipovetsky, Stan, and Michael Conklin. 2001. "Analysis of regression in game theory approach." *Applied Stochastic Models in Business and Industry*, 17(4): 319–330.

Liu, Yongxue, Chuanmin Hu, Wenfeng Zhan, Chao Sun, Brock Murch, and Lei Ma. 2018. "Identifying industrial heat sources using time-series of the VIIRS Nightfire product with an object-oriented approach." *Remote Sensing of Environment*, 204: 347–365.

Li, Ziqi. 2024. "GeoShapley: A Game Theory Approach to Measuring Spatial Effects in Machine Learning Models." *Annals of the American Association of Geographers*, 1–21.

Lundberg, Scott. 2017. "A unified approach to interpreting model predictions." *arXiv preprint arXiv:1705.07874*.

Lundberg, Scott M, Gabriel G Erion, and Su-In Lee. 2018. "Consistent individualized feature attribution for tree ensembles." *arXiv preprint arXiv:1802.03888*.

Martinez, Luis R. 2022. "How much should we trust the dictator's GDP growth estimates?" *Journal of Political Economy*, 130(10): 2731–2769.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.

Nair, Vinod, and Geoffrey E Hinton. 2010. "Rectified linear units improve restricted Boltzmann machines." 807–814.

Nordhaus, William, and Xi Chen. 2015. "A sharper image? Estimates of the precision of nighttime lights as a proxy for economic statistics." *Journal of Economic Geography*, 15(1): 217–246.

Rossi-Hansberg, Esteban, and Jialing Zhang. 2025. "Local GDP Estimates Around the World." National Bureau of Economic Research NBER Working Paper 33458.

Sherman, Luke, Jonathan Proctor, Hannah Druckenmiller, Heriberto Tapia, and Solomon M Hsiang. 2023. "Global high-resolution estimates of the United Nations Human Development Index using satellite imagery and machine-learning." National Bureau of Economic Research.

Shetty, Shobitha, Philipp Schneider, Kerstin Stebel, Paul David Hamer, Arve Kylling, and Terje Koren Berntsen. 2024. "Estimating surface NO2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning." *Remote Sensing of Environment*, 312: 114321.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, 15(1): 1929–1958.

Stone, Mervyn. 1974. "Cross-validatory choice and assessment of statistical predictions." *Journal of the royal statistical society: Series B (Methodological)*, 36(2): 111–133.

Tang, Wenbin, Ji Zhou, Jin Ma, Ziwei Wang, Lirong Ding, Xiaodong Zhang, and Xu Zhang. 2024. "TRIMS LST: a daily 1 km all-weather land surface temperature dataset for China's landmass and surrounding areas (2000–2022)." *Earth System Science Data*, 16(1): 387–419.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

United Nations Industrial Development Organization. 2016. "Report on Industrial Statistics to the Forty-seventh Session of the United Nations Statistical Commission." *Working document, United Nations Statistical Commission, 47th Session*, Accessed: 2025-03-15.

Upadhyaya, S, and V Todorov. 2009. "UNIDO Data Quality: A quality assurance framework for UNIDO statistical activities." *Vienna: UNIDO*.

Vogel, Kathryn Baragwanath, Gordon H Hanson, Amit Khandelwal, Chen Liu, and Hogeun Park. 2024. "Using Satellite Imagery to Detect the Impacts of New Highways: An Application to India." National Bureau of Economic Research.

Wei, Jing, Song Liu, Zhanqing Li, Cheng Liu, Kai Qin, Xiong Liu, Rachel T Pinker, Russell R Dickerson, Jintai Lin, KF Boersma, et al. 2022*a*. "Ground-level NO2 surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence." *Environmental Science & Technology*, 56(14): 9988–9998.

Wei, Jing, Zhanqing Li, Alexei Lyapustin, Lin Sun, Yiran Peng, Wenhao Xue, Tianning Su, and Maureen Cribb. 2021*a*. "Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications." *Remote Sensing of Environment*, 252: 112136.

Wei, Jing, Zhanqing Li, Jun Wang, Can Li, Pawan Gupta, and Maureen Cribb. 2023. "Ground-level gaseous pollutants (NO 2, SO 2, and CO) in China: Daily seamless mapping and spatiotemporal variations." *Atmospheric Chemistry and Physics*, 23(2): 1511–1532.

Wei, Jing, Zhanqing Li, Ke Li, Russell R Dickerson, Rachel T Pinker, Jun Wang, Xiong Liu, Lin Sun, Wenhao Xue, and Maureen Cribb. 2022*b*. "Full-coverage mapping and spatiotemporal variations of ground-level ozone (O3) pollution from 2013 to 2020 across China." *Remote Sensing of Environment*, 270: 112775.

Wei, Jing, Zhanqing Li, Wenhao Xue, Lin Sun, Tianyi Fan, Lei Liu, Tianning Su, and Maureen Cribb. 2021*b*. "The ChinaHighPM10 dataset: generation, validation, and spatiotemporal variations from 2015 to 2019 across China." *Environment International*, 146: 106290.

Wolpert, David H. 1992. "Stacked generalization." *Neural Networks*, 5(2): 241–259.

Xie, Yanmei, Caihong Ma, Yindi Zhao, Dongmei Yan, Bo Cheng, Xiaolin Hou, Hongyu Chen, Bihong Fu, and Guangtong Wan. 2024. "The Potential of Using SDGSAT-1 TIS Data to Identify Industrial Heat Sources in the Beijing–Tianjin–Hebei Region." *Remote Sensing*, 16(5): 768.

Zhang, Ping, Chencheng Yuan, Qiangqiang Sun, Aixia Liu, Shucheng You, Xianwen Li, Yaping Zhang, Xin Jiao, Danfeng Sun, Minxuan Sun, et al. 2019. "Satellite-based detection and characterization of industrial heat sources in China." *Environmental Science & Technology*, 53(18): 11031–11042.

Appendix: A Geospatial Approach to Measuring Economic Activity

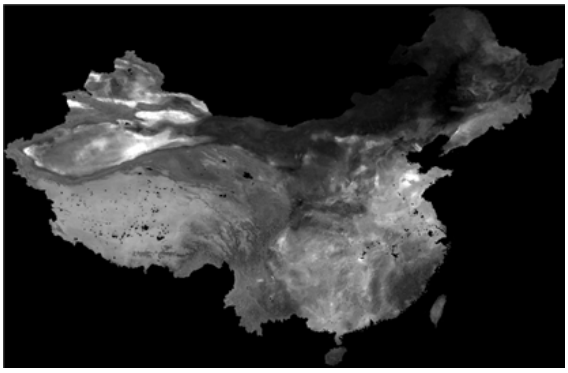
Anton Yang[1], Jianwei Ai[2,3], and Costas Arkolakis[1]

[1]Yale University
[2]Renmin University of China
[3]Cornell University

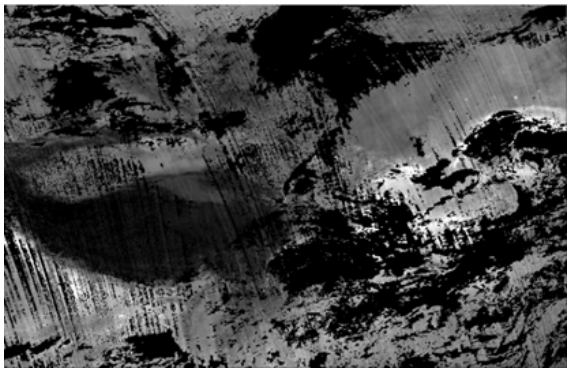
A Additional Figures and Tables

In this section, we provide supplementary figures and tables in our paper. Figure A.1 shows a comparison of the two databases, CHAP and Sentinel 5-P, using raw NO_2 data as an illustrative example in TIFF format. Figure A.2 shows the relationship between these two data sources, using NO_2 as the variable of interest. Figure A.3 shows the steel output monthly in China across different data sources. Figure A.4 shows the spatial distribution of Chinese steel plants in China. Figure A.5 shows the correlation matrix of the input features used in the machine learning models. Figure A.6 shows the location and output distribution across space of China's steel plants. Figure A.7 shows the steel plant location prediction accuracy when using different thresholds and tolerance distances. Figure A.8 shows the spatial heterogeneity of key input features at the grid level, both with and without steel plants. Figure A.9 shows relative humidity, air temperature, and solar radiation in the western, southern, and northern regions for April, July, and December. Figure A.10 shows the classification model performance. Figure A.11 provides a partial dependence plot (PDP) generated using the SHAP method and bootstrap analysis at the grid level. Figure A.12 compares the actual and predicted values of the output variable at the plant level. Figure A.13 shows the output prediction accuracy using Sentinel 5-P data at the grid level. Figure A.14 shows the output prediction accuracy at the plant level. Figure A.15 shows the Shapley values for each feature in the regression model at the plant level. Figure A.16 shows the Shapley value in the regression model at the plant level using Sentinel 5-P. Figure A.17 shows the Shapley values derived using a non-parametric bootstrap method. Figure A.18 shows the Shapley value of features using Sentinel 5-P data. Figure A.19 shows the PDP, again using the SHAP method and bootstrap analysis, at the plant level. Figure A.20 shows the box plots of out-of-sample results based on a 5-fold cross-validation. Table A.1 summarizes the model setting in the first location prediction task. Table A.2 shows the metrics comparison across models in output prediction at both the grid and plant level. Table A.3 shows the pollutant exposure index and population above pollution threshold in the western and southern regions.

Figure A.1: Ambient Air Pollutants NO_2 from Different Sources (Raw Data)



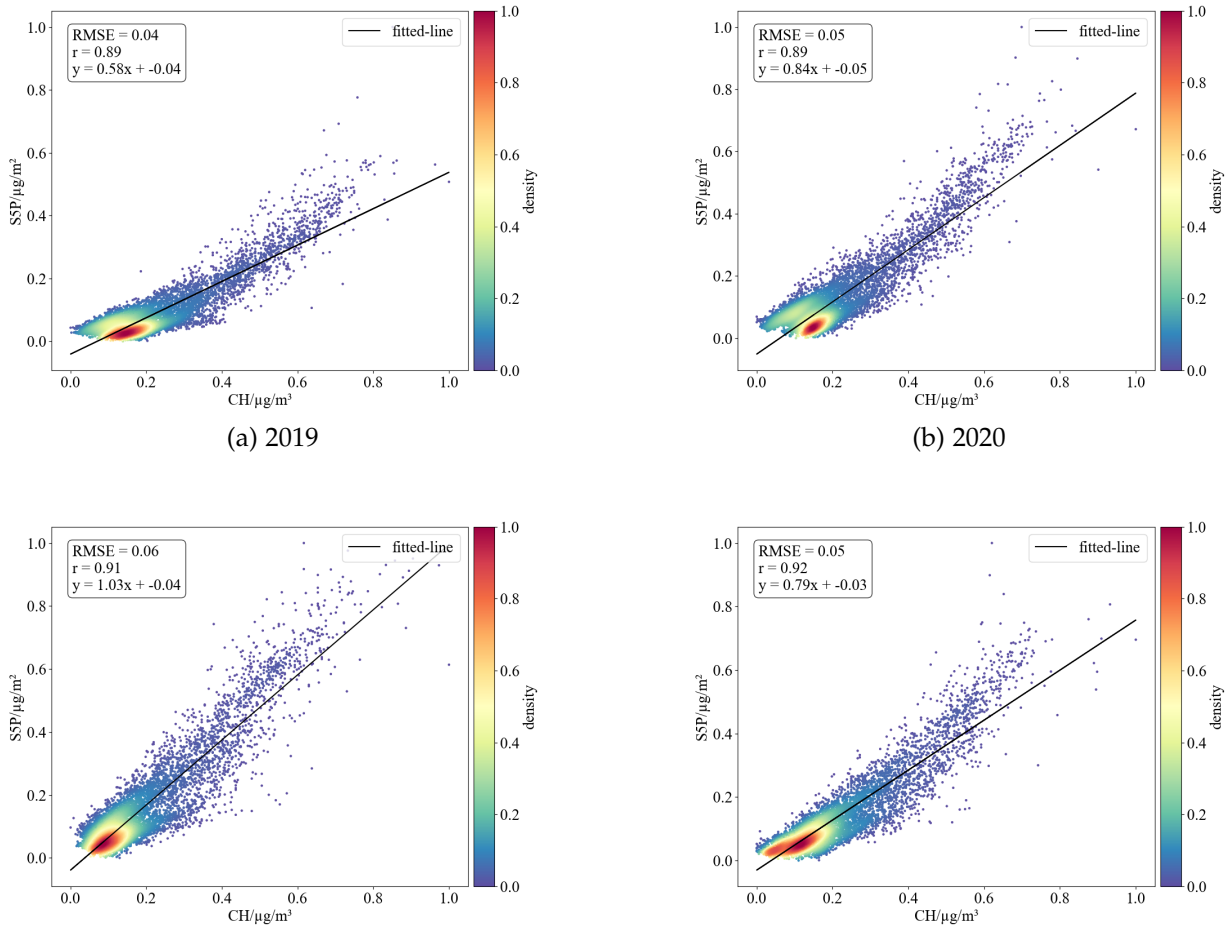
(a) NO_2 from CHAP



(b) NO_2 from Sentinel 5-P

Note: This figure compares two sources of NO_2 data: (a) CHAP and (b) Sentinel-5P. CHAP provides high-quality air pollution data specifically for China and is used in our main analysis. We also include results from our model trained on Sentinel-5P, as it is more accessible for researchers to replicate our exercises.

Figure A.2: Comparative Density Plots for Different Years: CHAP and Sentinel-5P

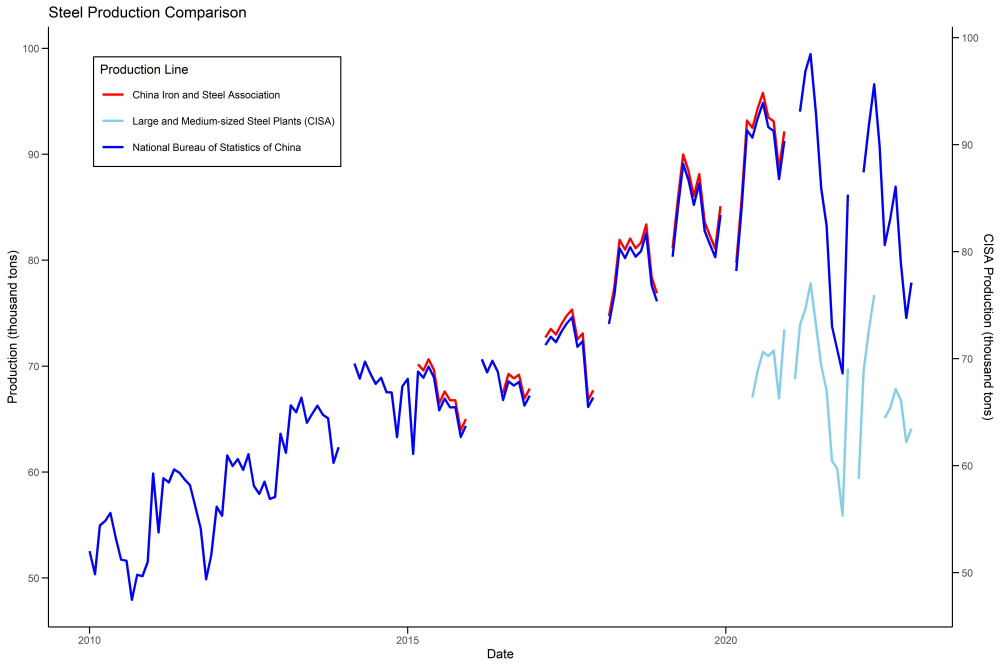


(a) 2019 (b) 2020

(c) 2021 (d) 2022

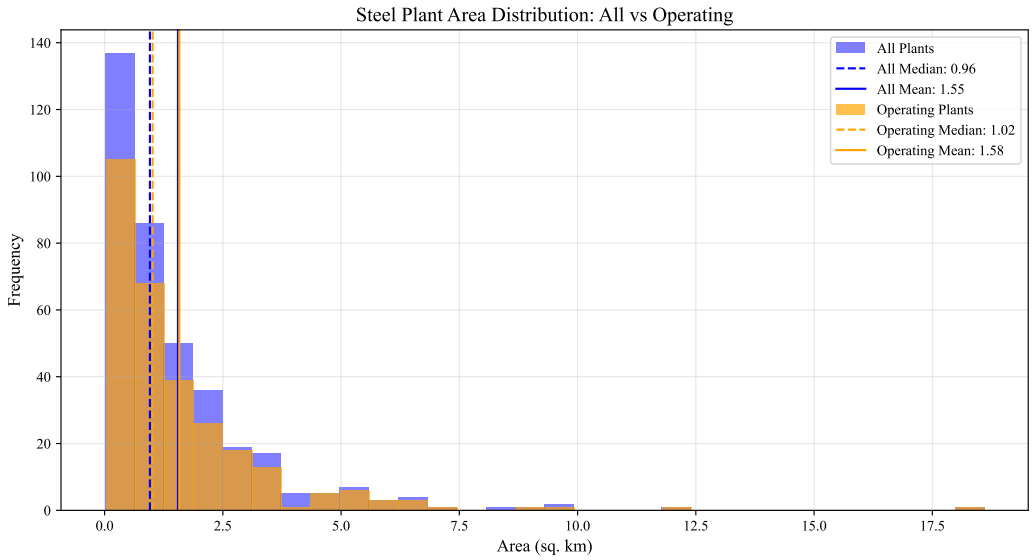
Note: This figure shows the fitted lines for two different sources of NO_2: (a) CHAP and (b) Sentinel-5P. The processing of Sentinel-5P NO_2 data followed the methods outlined in Shetty et al. (2024), which included filtering TROPOMI data based on a quality assurance flag and an uncertainty threshold. Specifically, only data with a quality assurance flag above 0.75 were retained to ensure high-quality retrievals. Additionally, data affected by clouds (cloud radiance fraction > 0.5) or influenced by snow, ice, or other errors were excluded. Since the Level 3 product from Google Earth Engine (GEE) already incorporates quality filtering, only cloud masking was applied.

Figure A.3: Steel Output Monthly in China across Sources



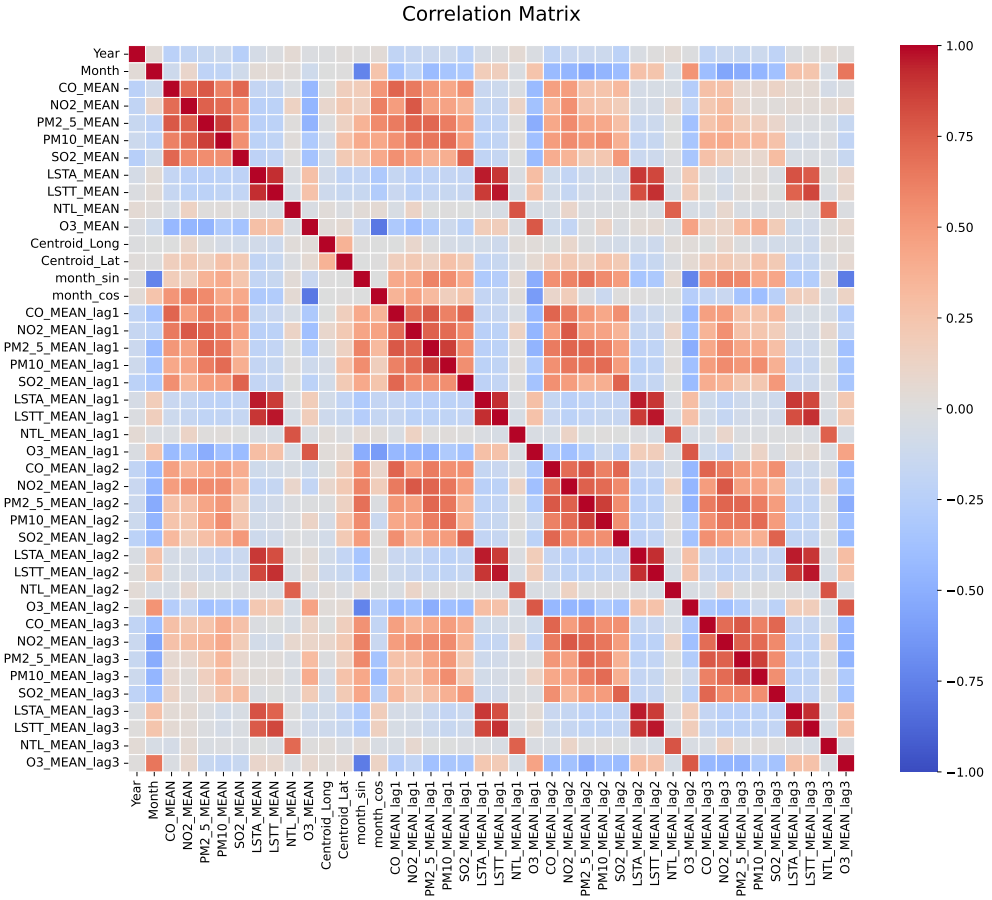
Note: This figure shows monthly steel output trends in China. The blue line represents data from the National Bureau of Statistics of China (2010-2024), while the red line represents data from the CISA. The lighter blue line represents the output of major steel plants reported by the CISA, available from 2020 onward. These data sources show consistent trends, especially during our research period (2020-2022). Our sample focuses on major steel plants reported by the CISA, which account for a significant share of China's total steel output. Note that both the government and the CISA report output data only for large and medium-sized steel plants, and individual plant-level data are not available for all plants. Gaps in the lines indicate missing data.

Figure A.4: Spatial Distribution of Chinese Steel Plants



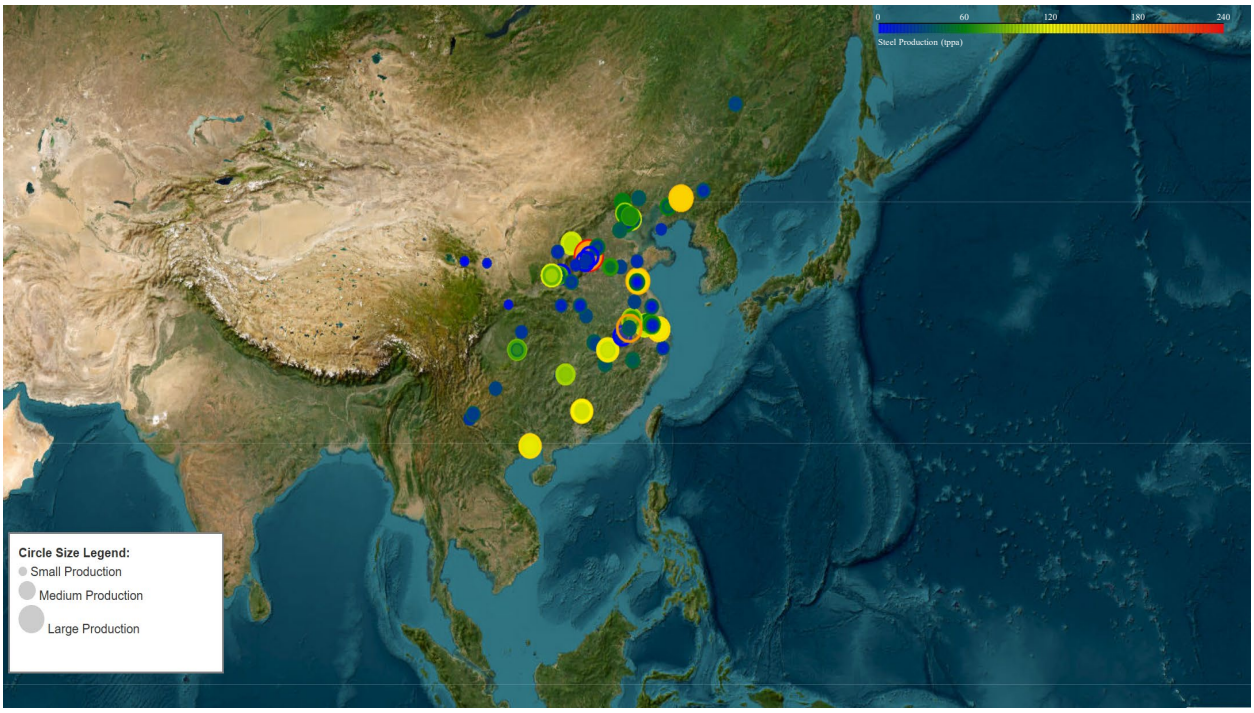
Note: This figure shows the spatial distribution of Chinese steel plants. The steel plant areas are calculated based on polygons using the coordinates provided by the Global Energy Monitor and satellite imagery.

Figure A.5: Correlation Matrix of the Input Features



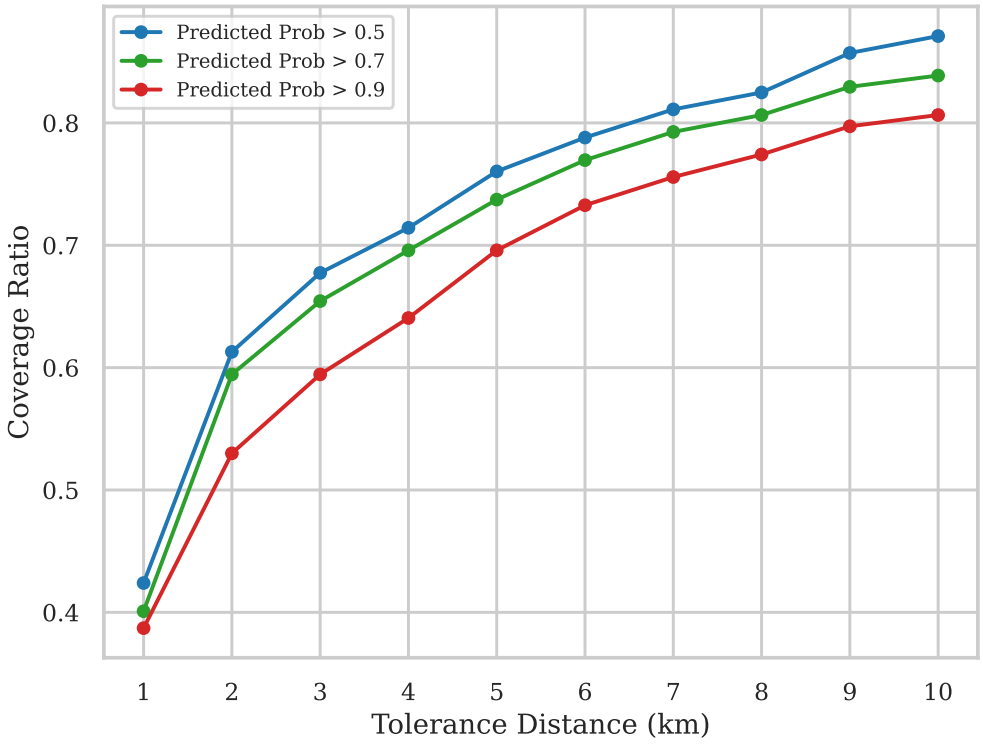
Note: This figure shows the correlation matrix of the input features in our machine-learning model. The results show that ambient air pollutants exhibit high intercorrelation among their lagged terms. In contrast, land surface temperatures (LSTA and LSTT), derived from Aqua (afternoon) and Terra (morning) satellites respectively, show a slight negative correlation with air pollutants, with Terra recording warmer temperatures. NTL appears to have a low correlation with other features in the model.

Figure A.6: Location and Steel Output Distribution of China's Steel Plants



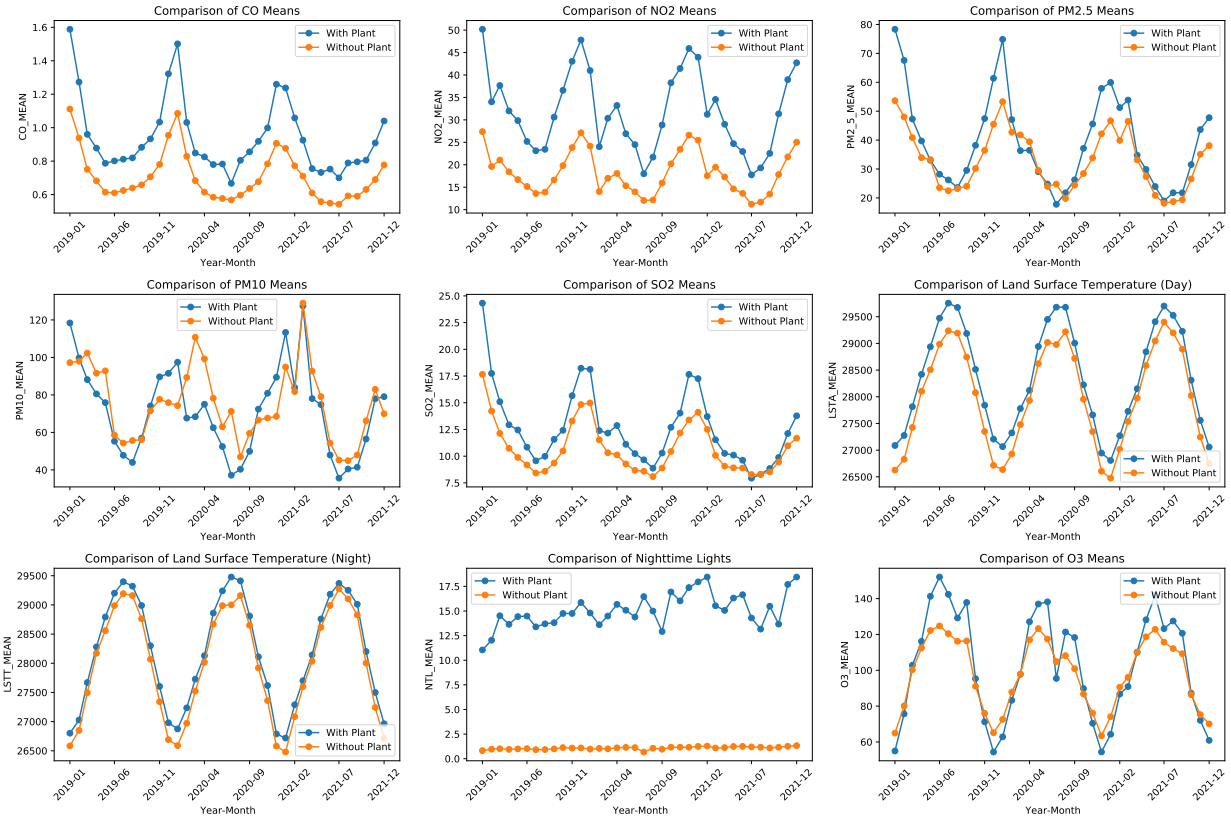
Note: This figure shows the location and output distribution of China's steel output in our training-test sample.

Figure A.7: Percentage of Identified Steel Plants from the Holdout Sample



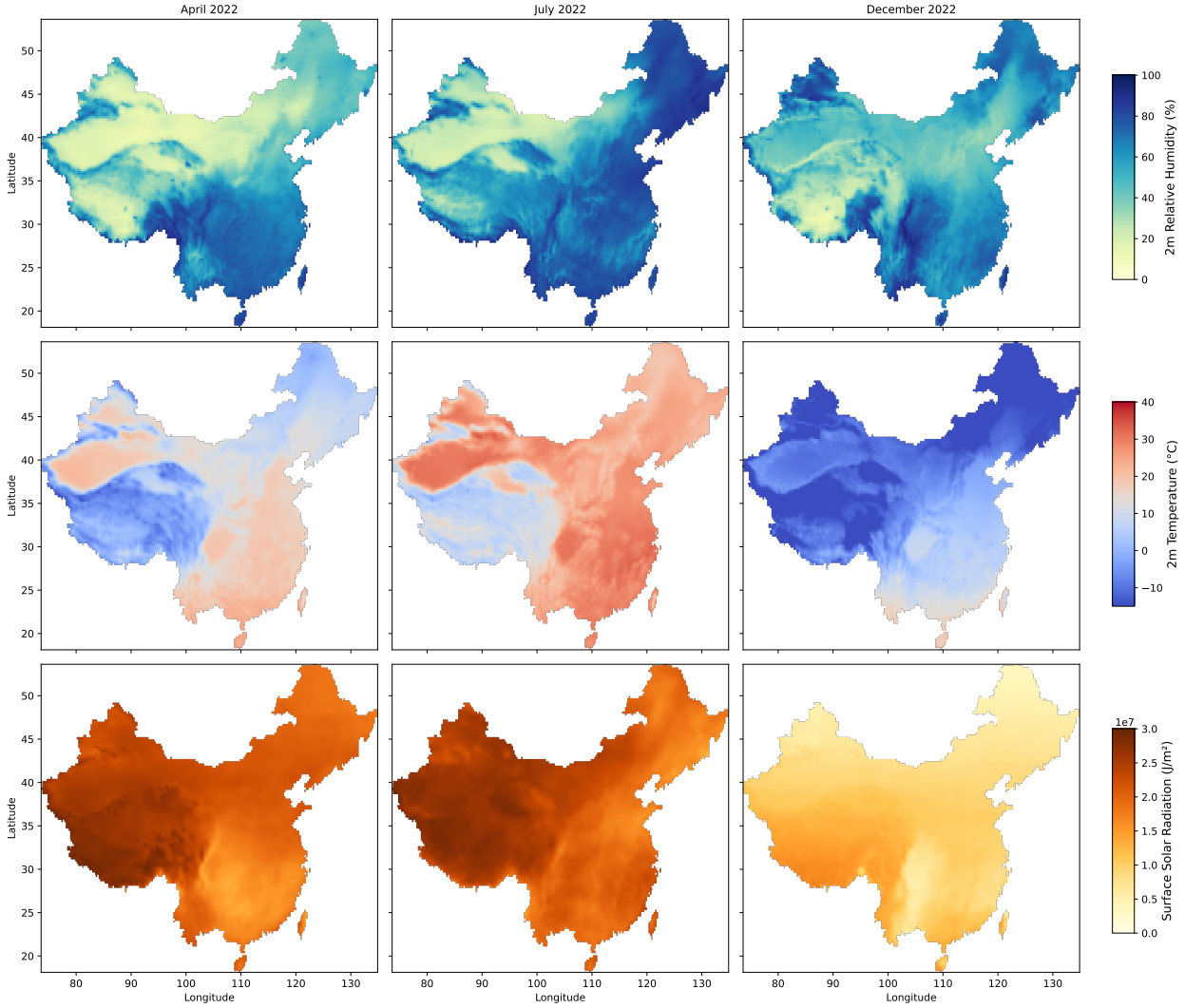
Note: The coverage ratio shows the fraction of holdout sample points (red markers shown in Figure 1a) that lie within a specified distance from high-probability prediction points (i.e., points with a predicted probability above the given threshold). The tolerance distance varies from 1 km to 10 km.

Figure A.8: Cross-Grid Spatial Heterogeneity With and Without Steel Plants



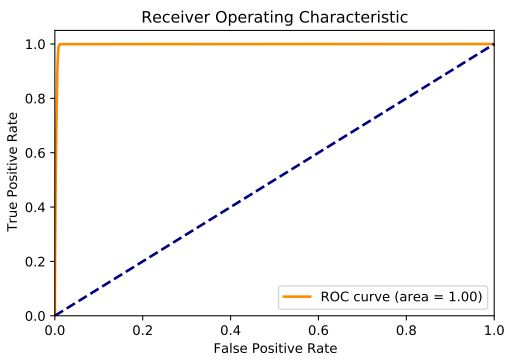
Note: This figure compares spatial characteristics based on average measurements in grids with and without steel plants. One may argue that many geospatial statistics are influenced by various industrial activities, and our predictions may capture general industrial activity rather than being specific to steel production. The evaluation metrics along with this figure support this point. At the grid level, analyzing feature importance shows that morning temperature (LSTA) significantly enhances predictive power. However, at the plant level, our results suggest that LSTA, which exhibits less spatial heterogeneity across different industries, becomes an irrelevant indicator. Moreover, the higher predictive power of O_3 can be explained by its seasonal variations. Typically, during winter, O_3 levels are very low due to insufficient sunlight, so the difference between grids with and without steel plants is minimal. In contrast, during summer, when sunlight is abundant, even the same VOC*s* and NO_x emissions from steel plants are subjected to more intense photochemical reactions, leading to higher O_3 levels and, consequently, creating a sharper contrast between grids with and without steel plants. This stark variation compared to January–when there is practically no sunlight–makes it easier for the machine learning model to learn which grids have steel plants (also see Figure A.9 for seasonal variations across environmental factors). As a result, O_3 stands out with a greater predictive power.

Figure A.9: Relative Humidity, Air Temperature, and Solar Radiation

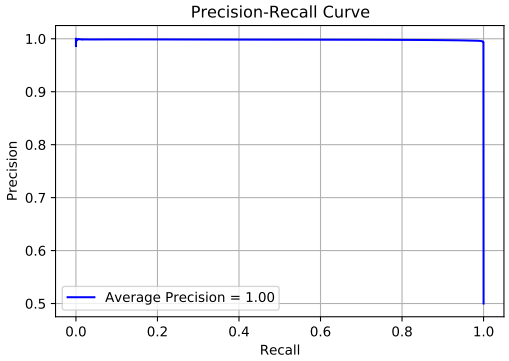


Note: ERA5 monthly mean 2 m air temperature (i.e., air temperature measured at two meters above ground level) and 2 m dewpoint (i.e., the temperature to which air at the height would have to be cooled for water vapor to condense or to reach saturation) were obtained from the Copernicus Climate Data Store (CDS) "reanalysis-era5-single-levels-monthly-means" dataset for April, July, and December 2022 (Hersbach et al., 2020). We calculate 2 m relative humidity using the Magnus formula from the 2 m dewpoint and temperature fields. The surface solar radiation downwards (SSRD) for these same months was also taken from the ERA5 single-level monthly means dataset, which represents the total incoming solar radiation (both direct and diffuse) reaching the Earth's surface (in $J\,m^{-2}$).

Figure A.10: Classification Model Performance



(a) AUROC



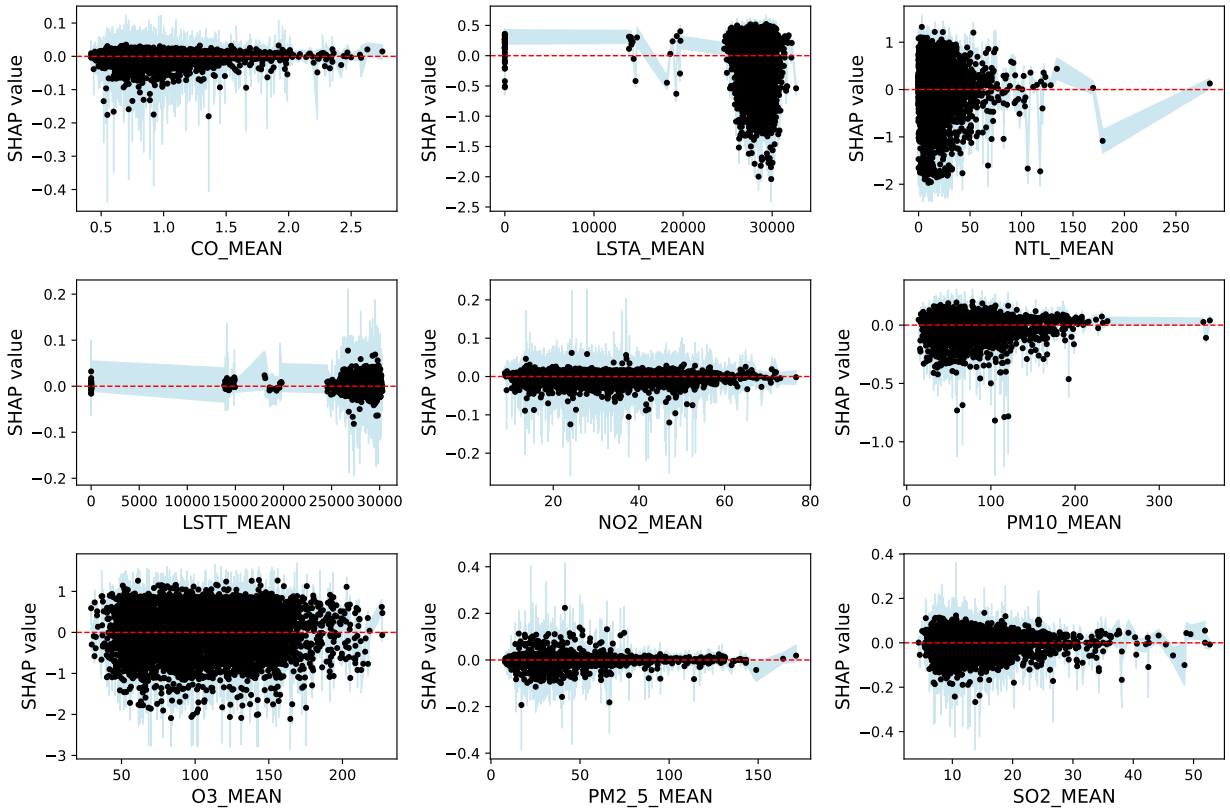
(b) Precision-Recall Curve

(c) Model Metrics Report

Class	Precision	Recall	F1-Score	Support
0	1.00	0.99	0.99	1,385,630
1	0.99	1.00	0.99	1,385,630

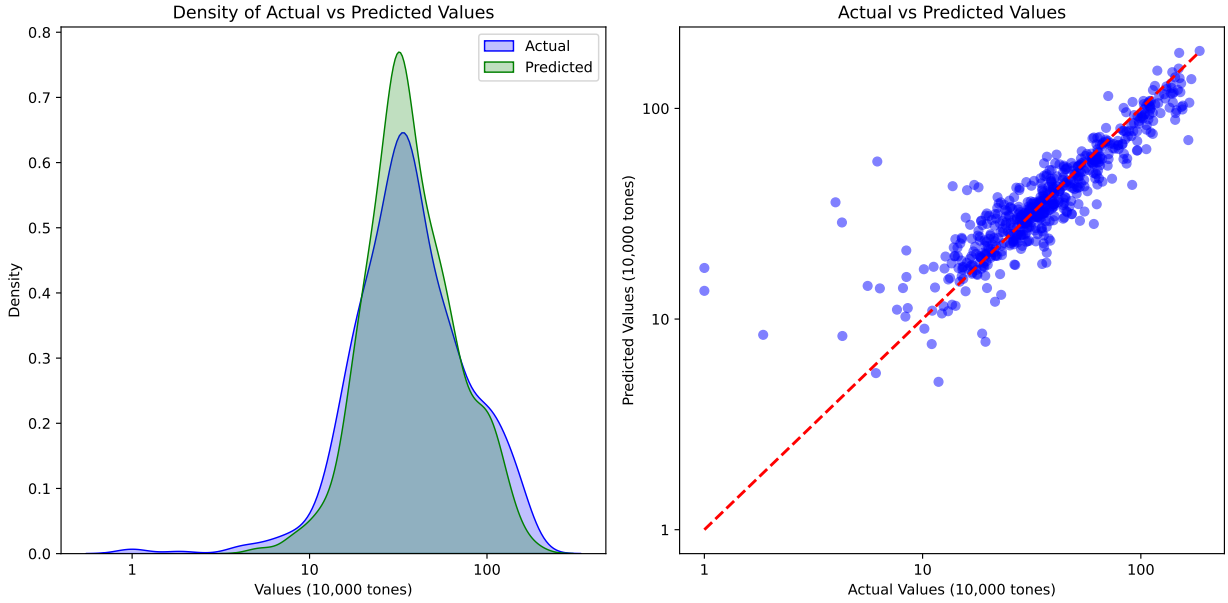
Note: These figures and table show the classification model performance using AUROC and precision-recall as evaluation metrics. Figure A.10a shows the Receiver Operating Characteristic (ROC) curve, while Figure A.10b shows the precision-recall curve. Table A.10c shows the classification report for model performance after applying the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset. Class 1 corresponds to grid cells containing steel plants, while Class 0 corresponds to the minority class of grid cells without steel plants.

Figure A.11: Partial Dependence Plot (PDP) at the Grid Level



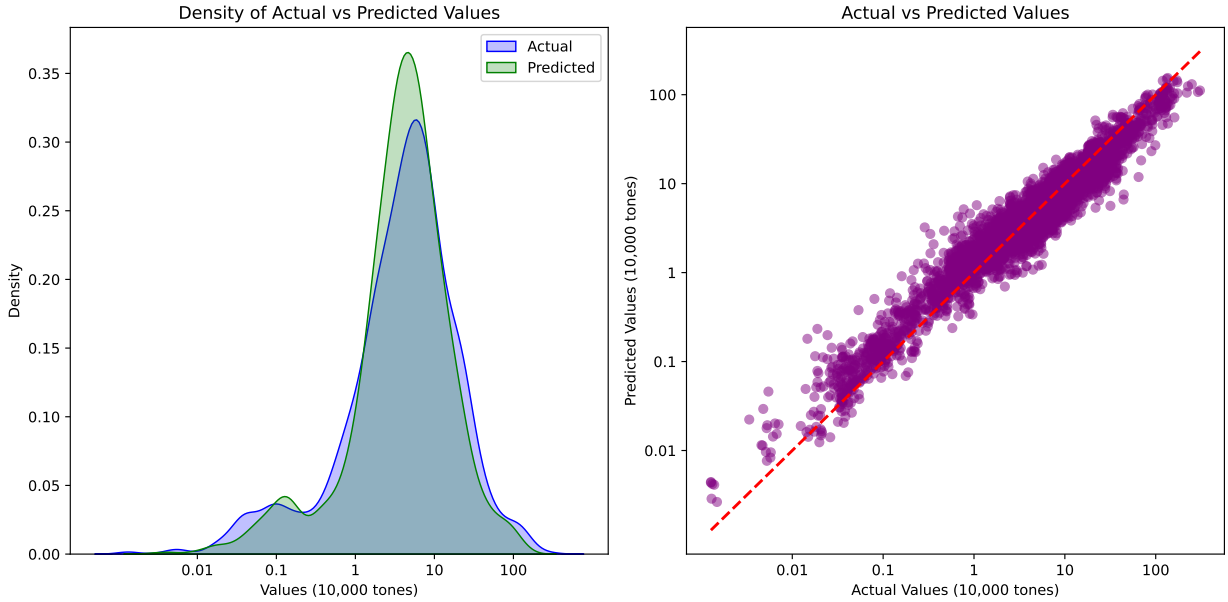
Note: This figure shows the PDP for all non-location features at the grid level. The blue shading represents the 95% bootstrap confidence interval for the SHAP values.

Figure A.12: Actual and Predicted Value at the Plant Level Using CHAP



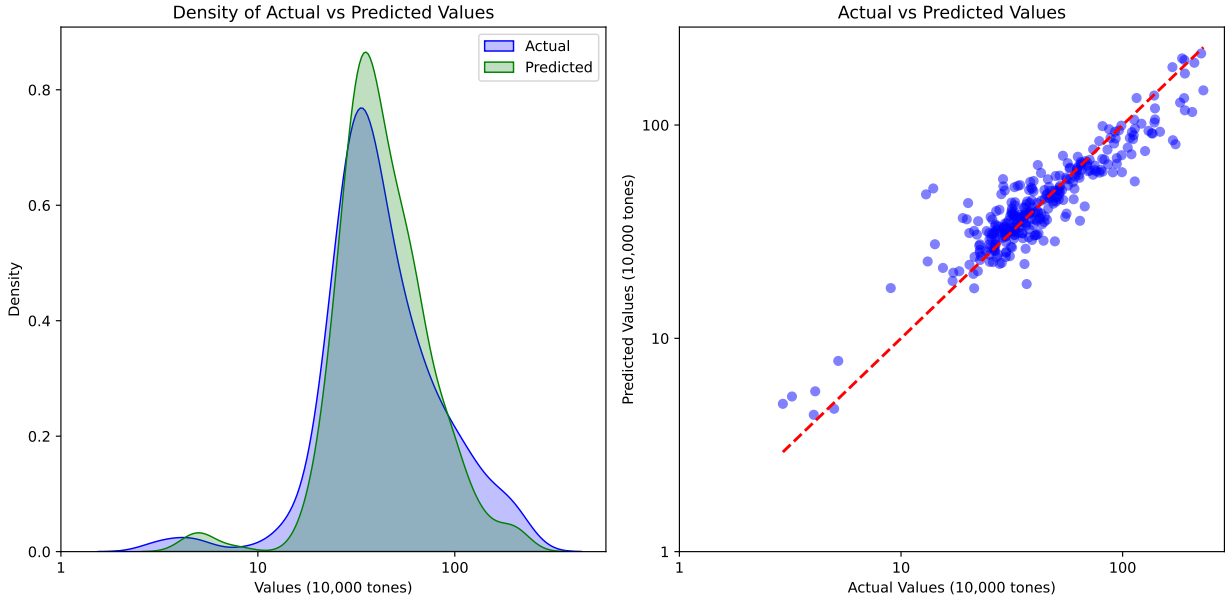
Note: This figure shows the model fit at the plant level using CHAP data. The left panel shows the distribution of actual and predicted values, while the right panel compares crude steel output between actual and predicted values (measured in units of 10,000 tons).

Figure A.13: Actual and Predicted Value at the Grid Level Using Sentinel 5-P



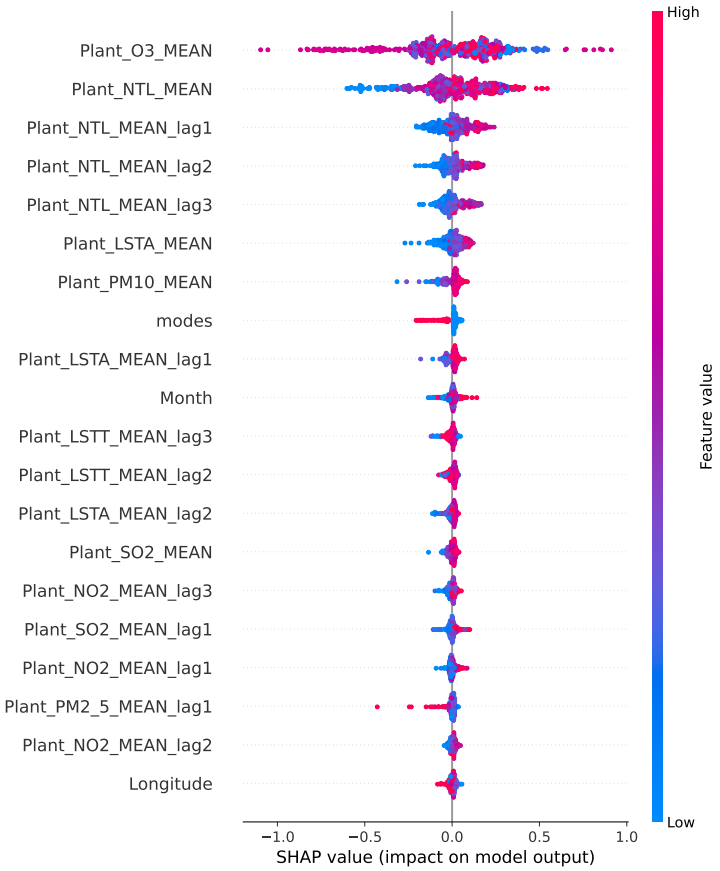
Note: This figure shows the model fit at the grid level using Sentinel 5-P satellite data. The left panel shows the distribution of actual and predicted values, while the right panel compares crude steel output between actual and predicted values (measured in units of 10,000 tons).

Figure A.14: Actual and Predicted Value at the Plant Level Using Sentinel 5-P



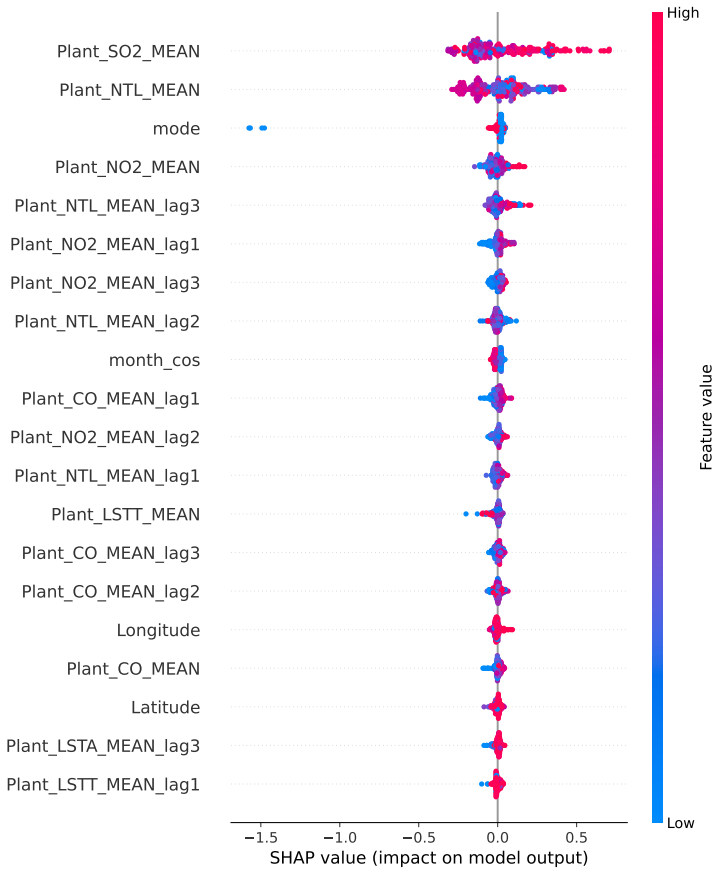
Note: This figure shows the model fit at the plant level using Sentinel 5-P satellite data. The left panel shows the distribution of actual and predicted values, while the right panel compares crude steel output between actual and predicted values (measured in units of 10,000 tons).

Figure A.15: Shapley Value for Each Feature in the Regression Model at the Plant Level



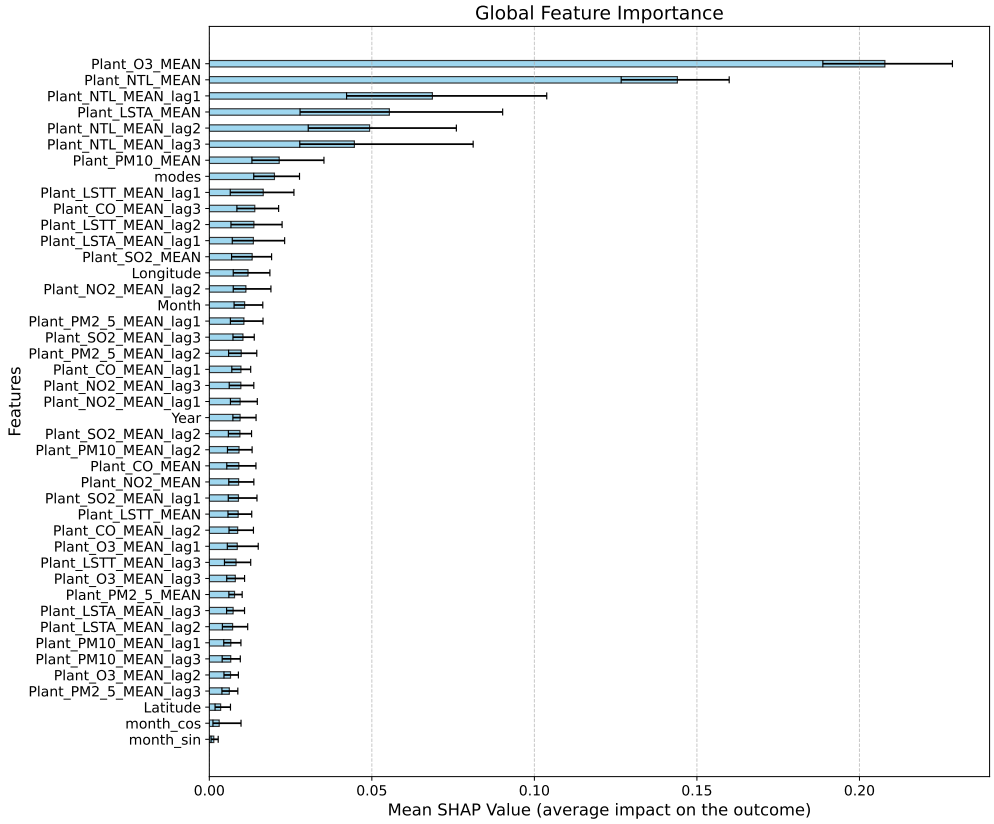
Note: This figure shows the Shapley values of each feature in our classification model. The Shapley value shows the contribution of each feature to the model's prediction. Each dot represents a sample in the model, and the density reflects the distribution of feature values. The blue color on the left tail shows that smaller values of the feature are associated with a lower-than-average model output, meaning that lower feature values reduce steel output (negative SHAP value). Alternatively, if red appears on the left tail, it suggests that higher feature values reduce steel output. On the other hand, if blue appears on the right tail, it means that lower feature values contribute to an increase in steel output, while red on the right tail indicates that higher feature values drive steel output upward. For example, LSTA (temperature) shows a clear blue tail on the left, meaning that lower temperatures (blue) reduce steel output.

Figure A.16: Shapley Value in the Regression Model at the Plant Level Using Sentinel 5-P



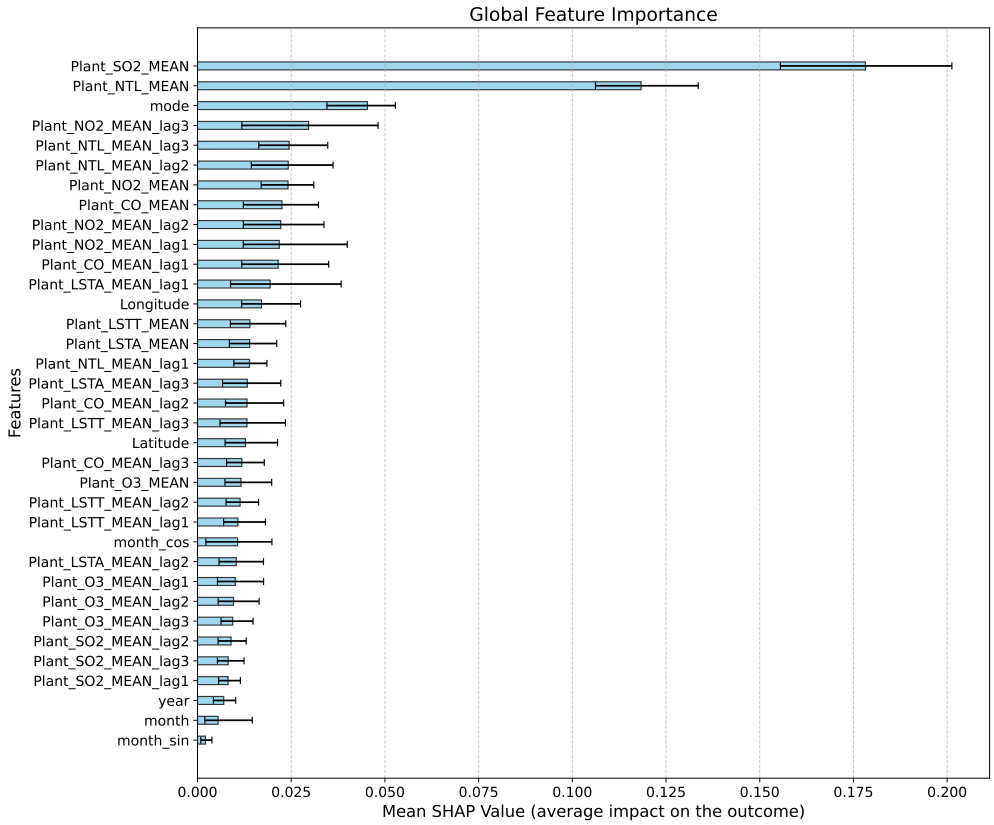
Note: This figure shows the Shapley values of each feature in our classification model using Sentinel 5-p. The Shapley value shows the contribution of each feature to the model's prediction. Each dot represents a sample in the model, and the density reflects the distribution of feature values.

Figure A.17: Shapley Value for Each Feature in Regression Model at the Plant Level



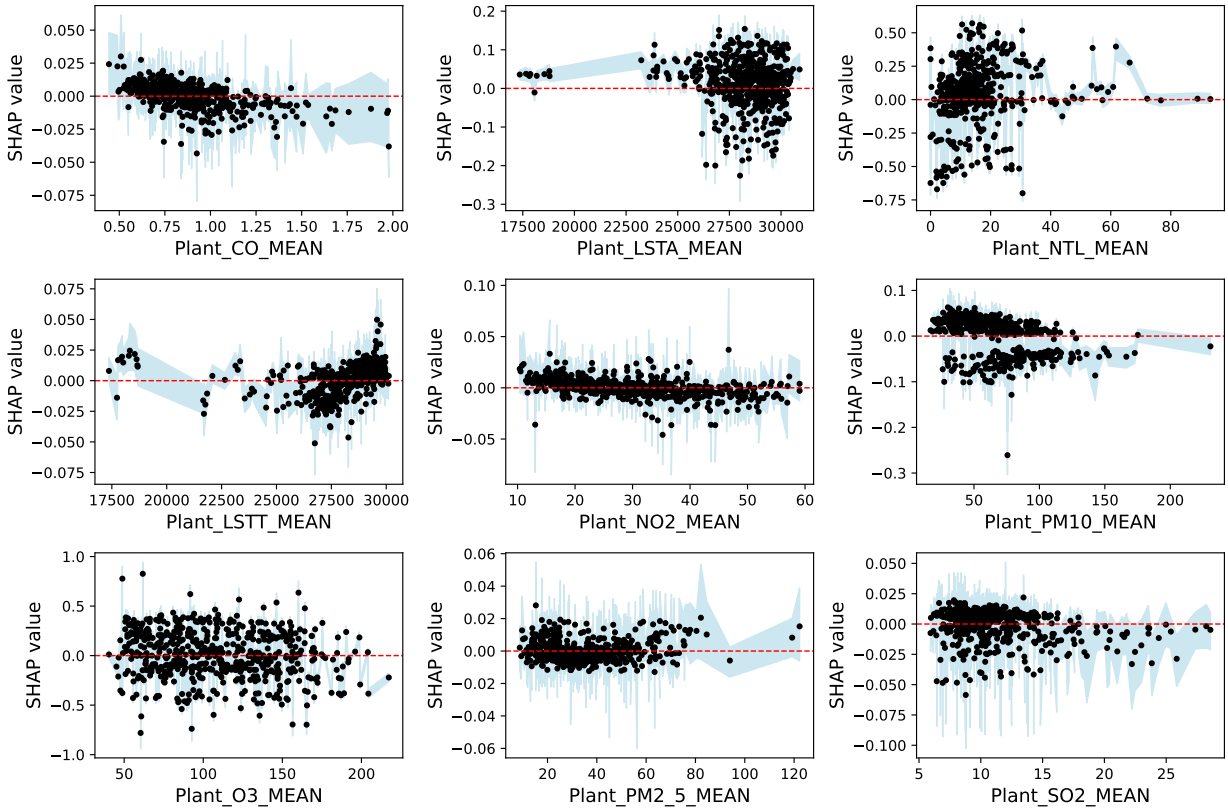
Note: This figure shows the Shapley values for each feature in our regression model, which were estimated using a non-parametric bootstrap based on residual resampling (Efron, 1992). Specifically, model residuals are resampled with replacement and added to the predicted values to generate new dependent variables. We retrained the XGBoost model using the original hyperparameters, and the SHAP values are recalculated and stored. We repeat the process 5,000 times and present 95% confidence intervals.

Figure A.18: Shapley Value in Regression Model at the Plant Level using Sentinel 5-P



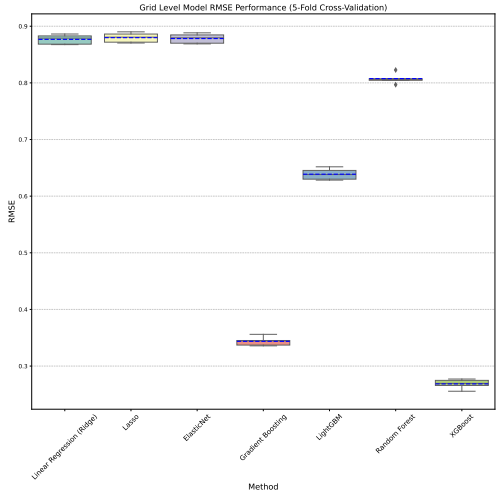
Note: This figure shows the Shapley values for each feature in our regression model, which were estimated using a non-parametric bootstrap based on residual resampling (Efron, 1992). Specifically, model residuals are resampled with replacement and added to the predicted values to generate new dependent variables. We retrained the XGBoost model using the original hyperparameters, and the SHAP values are recalculated and stored. We repeat the process 5,000 times and present 95% confidence intervals. The satellite data is from Sentinel 5-P.

Figure A.19: Partial Dependence Plot (PDP) at the Plant Level

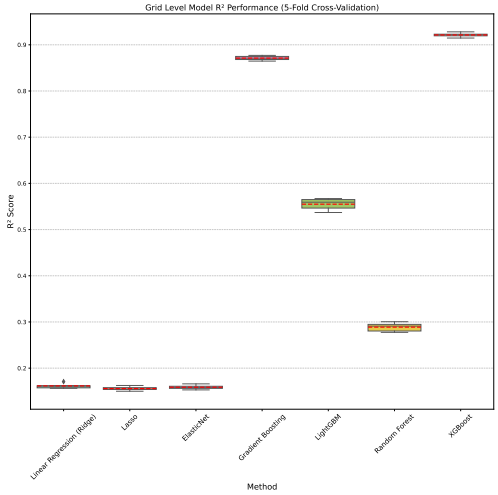


Note: This figure shows the PDP for all non-location features at the plant level. The blue shading represents the 95% bootstrap confidence interval for the SHAP values.

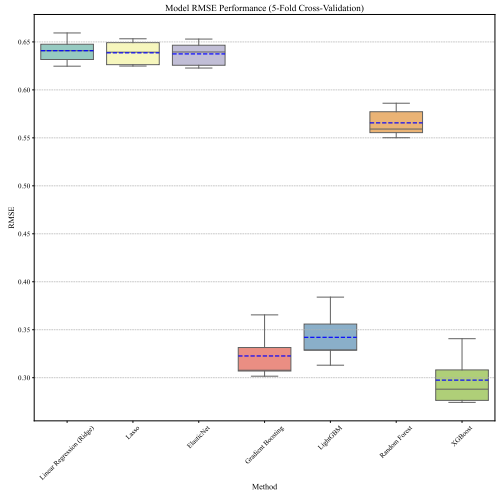
Figure A.20: Box Plots of Out-of-Sample Results Based on a 5-Fold Cross-Validation



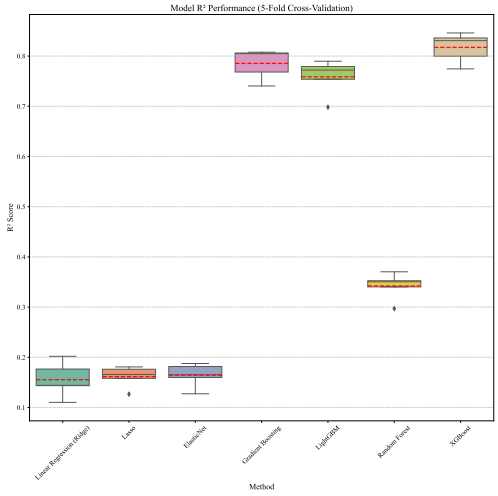
(a) RMSE at the Grid Level



(b) R^2 at the Grid Level



(c) RMSE at the Plant Level



(d) R^2 at the Plant Level

Note: Figure A.20a and A.20c show model performance using RMSE as the metrics at both the grid and plant levels. Fiture A.20b and A.20d use R^2 as the metric at both grid and plant levels. These box plots suggest that the XGBoost performs better than other machine learning models.

Table A.1: Model Summary of the Location Prediction

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	2624
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2080
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 1)	33

Note: The neural network model we build is a fully connected framework containing two hidden layers. Each hidden layer has 64 and 32 neurons, respectively. The third column shows the number of parameters we use in each layer.

Table A.2: Metrics Comparison Across Models

Panel (a): Grid Level	**MSE**	**RMSE**	**WMAPE**	**R2 Score**
Linear Regression	0.770281	0.877621	0.437954	0.160030
Lasso	0.770067	0.877500	0.437944	0.160263
ElasticNet	0.770075	0.877505	0.437943	0.160254
Gradient Boosting	0.184679	0.429669	0.202237	0.794800
LightGBM	0.511671	0.715285	0.349945	0.442062
Random Forest	0.662404	0.813841	0.403364	0.277673
XGBoost	0.102480	0.319965	0.150515	0.888198
Ensemble Learning	0.060117	0.245188	11.144664	0.933764
Panel (b): Plant Level				
Linear Regression	0.410790	0.640814	0.139372	0.155307
Lasso	0.407932	0.638591	0.137763	0.161500
ElasticNet	0.406521	0.637483	0.137732	0.164365
Gradient Boosting	0.104663	0.322463	0.060021	0.785451
LightGBM	0.117655	0.342091	0.063710	0.758615
Random Forest	0.320097	0.565605	0.119762	0.342104
XGBoost	0.089104	0.297479	0.053238	0.817400
Ensemble Learning	0.011405	0.332277	5.480388	0.785455

Note: The ensemble learning model combines the following base models, Lasso, LightGBM, Random Forest, and XGBoost, with a Linear Regression meta-model.

Table A.3: Pollutant Exposure Index and Population above Pollution Threshold

Pollutant	Pollutant Exposure Index		Population in Cells above Threshold		
	West	South	West	South	Threshold Value
CO	0.79	0.78	1 415 098	73 661 267	0.8
SO_2	10.86	9.86	7 957 498	129 801 983	10.0
NO_2	27.30	26.26	174 560	10 625 705	30.0
O_3	103.52	108.42	442 478	162 742 874	110.0
$PM_{2.5}$	28.86	31.12	171 406	27 095 035	35.0
PM_{10}	54.27	55.19	366 991	54 936 831	60.0

Note: The value of threshold is either in $\mu g/m^3$ or parts per billion (ppb). These values are pollutant levels exceeding moderate thresholds based on estimated emissions from industrial plants. However, since direct pipe emissions are not observed and pollutant levels are derived from satellite-based data, the reported values may be lower than plant-specific regulatory range.

B Model Description

In this section, we outline the main models used in our methodology. We focus on tree-based algorithms and the application of Shapley values for interpretability.

For completeness, we also specify the additional parameters used in our tree-based model. The `Learning Rate` determines how much each new tree corrects the current model, while `Max Depth` sets the maximum depth of each tree, which controls model complexity. We also tune parameters related to the number of trees and the data sampling strategy based on our selected evaluation metrics.

B.1 Extreme Gradient Boost Method

XGBoost, or extreme gradient boosting, is a widely used tree-based model and remains a preferred tool for many practitioners (Chen and Guestrin, 2016; Kossen et al., 2021). It supports column subsampling (random feature selection) and employs hyperparameters, such as regularization terms, maximum tree depth, and the number of trees, to mitigate overfitting. We show how these parameters contribute to minimizing the regularized objective function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell\left(Y_i, \widehat{Y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{B.1}$$

where $\ell\left(Y_i, \widehat{Y}_i\right)$ is the loss for the i^{th} observation, and $\Omega\left(f_k\right)$ is a regularization term that helps reduce overfitting. The regularization term is defined as follows:

$$\Omega\left(f_k\right) = \gamma T_k + \frac{1}{2}\lambda \sum_j w_{kj}^2, \tag{B.2}$$

where T_k is the total number of leaves in the k^{th} tree, w_{kj} is the prediction for the j^{th} leaf in that tree, and γ and λ are hyperparameters controlling regularization. XGBoost uses an additive learning approach, where predictions are updated iteratively:

$$\hat{Y}_i^{(t)} = \hat{Y}_i^{(t-1)} + f_t\left(X_i\right), \tag{B.3}$$

with $f_t(X_i)$ representing the output of the t^{th} tree. Each tree is constructed to minimize the overall objective function. For a given split, XGBoost calculates the following gain to decide whether the split improves the model:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma, \tag{B.4}$$

where G_L and G_R are the sums of first-order gradients for the left and right splits, respectively, H_L and H_R are the sums of second-order gradients (Hessians), and λ is the L2 regularization term. To avoid overfitting and highlight pollutant features in our model, we adopt several strategies. First, we use column subsampling (`colsample_bytree`) to select a subset of features for each tree, thus increasing the likelihood of pollutant features appearing across the ensemble. Specifically, we set `colsample_bytree` to 0.6 to favor pollutant features during training. Second, we assign an L1 regularization coefficient $\alpha = 0.2$ to shrink the weights of less relevant features toward zero. Finally, we use an L2 regularization coefficient $\lambda = 0.5$ to penalize large coefficients for features with high impact, such as satellite-labeled pollutants.

B.2 Shapley Value in Machine Learning

SHAP builds on concepts from cooperative game theory. The "game" is the machine learning prediction task. The "players" are the input features (e.g., SO_2, O_3), and the "payout" is the model's prediction for a specific observation (e.g., grid-level steel output). The objective of SHAP is to distribute this payout among all features based on their respective contributions.

The contribution of a feature X_p in a model can be written as:

$$X_p = \sum_{S \subseteq N \setminus \{p\}} \frac{|S|!\,(n-|S|-1)!}{n!}\Big[f(S\cup\{p\}) - f(S)\Big], \tag{B.5}$$

where N is the set of all features, n is the total number of features, S is any subset of N excluding p, and $f(S)$ is the model's prediction using only the features in S. Shapley values satisfy desirable properties such as efficiency, symmetry, and additivity, making them well-suited for explaining model predictions (Lipovetsky and Conklin, 2001; Datta, Sen and Zick, 2016). In practice, $f(S)$ is commonly approximated via marginal expectations, that is:

$$f(S) = \mathbb{E}\big[f(X) \mid X_S\big], \tag{B.6}$$

where X_S denotes the features in subset S. For tree-based models (e.g., XGBoost), SHAP values

can be computed efficiently using specialized algorithms such as TreeExplainer.

A positive SHAP value ($\phi_i > 0$) means that a feature increases the predicted outcome relative to the baseline, while a negative value ($\phi_i < 0$) means that it decreases the outcome. The absolute magnitude of ϕ_i reflects the feature's impact on the prediction.

PDPs complement SHAP by visualizing the average relationship between a single feature and the predicted outcome while marginalizing over all other features. For a feature X_j, the PDP is:

$$\widetilde{f}_{\text{PDP}}(X_j) = \frac{1}{n}\sum_{i=1}^{n} f(X_j, \mathbf{X}_{i,-j}), \tag{B.7}$$

where X_j is the value of the feature of interest, $\mathbf{X}_{i,-j}$ are all other features of observation i, and n is the total number of observations. A flat PDP suggests minimal influence, while an increasing or nonlinear PDP indicates positive or more complex relationships.

In our application, PDPs for pollutants like NO_2 show thresholds where rising emissions no longer substantially increase steel output. For NTL, the PDP shows diminishing returns, which suggests that once an area is sufficiently illuminated, additional brightness adds little predictive value. Meanwhile, SHAP provides local interpretations by isolating each feature's effect on individual predictions, complementing the global perspective of PDP.