

## 3 DATASET AND FEATURE ENGINEERING

The acquired dataset of Berlin's environmental and meteorological characteristics for the period 2023 is self-constructed and is derived from the integration of eleven sub-datasets, as detailed in the Data Source Statement. In addition to the hourly emission measurements, all land use characteristics describing the surroundings of each monitoring station are obtained from the [Berlin Geoportal](#) and are temporally consistent. Meteorological data, featuring an hourly temporal resolution without spatial variation, are obtained from the [Deutscher Wetterdienst - DWD](#). To capture the seasonal variation in green volume, the grid-based Leaf Area Index dataset from Yan et al. (2024) is transformed and included in the constructed dataset. An overview of the constructed feature is given in Figure 1, and the following section details the data transformation for each variable group, the required missing data imputation, and the feature selection applied to address multicollinearity.
















Group	Name	Feature Name	Preprocessing Steps	Dependency
Target variable	Nitrogen dioxide	NO2	missing values imputation with RF from data of existing sites at same time step	 
average emission	weighted average of NO2 for each time step	'weighted_mean_pollution'	the average of simultaneous NO2 concentration is weighted by difference in distance toward the city center	 
Emitter: Traffic	traffic volume	TVI_{radius} (for radius = [25, 50, 75, 100, 200])	traffic volume index (TVI) as vehicle count times street length inside different Radius	
Emitter: Traffic	distance to street distance to intersection	nearest_street, nearest_intersect	combine coordinates with shapefile of road network	
Emitter: Traffic	proportion of heavy vehicles	prop_main_tvi_200		
Emitter: Population	population density within 200 and 500 meter	pop_{radius} (for radius = [200, 500])		
Time	boolean for peak traffic hours	weekend, rushhour	weekend : Saturday, Sunday, Holidays (lookup table) rush hour : 6 am - 20 pm	
Greenery: GVI	Green Volume Index	gvi_{radius} (for radius = [25, 50, 75, 100, 200])	weighted average of green volume for different radius	
Greenery: LAI	Seasonal variation in Leaf Area Index	lai_factor	rioxarray: using grid-based dataset of LAI for buffer area around Berlin with 8 day and 5 km resolution	
Meteorological Data	hourly data for humidity, temperature, solar radiation, air pressure, precipitation, wind speed and direction	humidity, temp, radiation, air_pressure, precipitation_mm, precipitation_bool, wind_speed, wind_degree	imputation of missing data through LSTM based univariate multi-step forecasting and linear interpolation	
Street Canyon: captivity	capture degree of architectural density	prop_intercept_{radius} (for radius = [50, 200])	"detection" of surrounding buildings through spatial echo analysis	
Street Canyon: free wind access	no building in the current wind direction	free_wind	is obstacle in +/- 5° of wind direction?	 

Figure 1: This table enumerates the processing steps undertaken for each feature, detailing their respective dependencies on temporal or spatial variations. The summary provides insights into the methodologies applied for feature engineering.

### 3.1 Target Variable: Nitrogen Dioxide ( $NO_2$ )

The sixteen official monitoring stations of the Berliner-Luftgütemessnetz (2023) monitor various air pollutants, including particulate matter, nitrogen oxides, or near-surface ozone. Throughout 2023, these stations provided 140,144 hourly observations of  $NO_2$ . The stationary sensors are strategically placed along major traffic routes, residential areas, and suburban regions to capture a wide range of land-use types and varying emitter densities. The locations of these stations are depicted in the map shown in Figure 2. Periodic maintenance and calibration of the monitoring sensors ensure data reliability (Schümann et al., 2021, p. 4). The pollution measurements are skewed towards high values, as evidenced by the distribution shown in Figure 20 and 21. A total of 722 missing observations were identified and subsequently removed from the dataset due to their marginal proportion and the lack of temporal or between-site correlation in the missing data. Further correlation tests with external features could provide insights into the characteristics of the missing data but are not pursued due to the small proportion of absent data.

### 3.2 Feature Engineering: Weighted Average of $NO_2$ Concentration from other sensing Sites

The weighted average of simultaneous  $NO_2$  measurement at each time step is incorporated as a spatial lag variable. For this purpose, a custom weighing scheme based on *distance similarity to a centroid point* is developed. This similarity estimation is critical as an additional feature for both the RF model and the initialization of the graph representation. The significance of spatial correlation is incorporated through the weighted average and has previously been implemented using *inverse distance* (Vu et al., 2024, p. 6; Iskandaryan et al., 2023, p. 13), based on the assumption that spatial proximity correlates with similarity in local characteristics. However, inverse-distance is unable to consider distant but similar monitoring sites. Therefore, we propose a weighted average based on *distance similarity to the city center*, defined as the distance to Berlin's TV Tower. The weight matrix is calculated as:

$$M_{ij} = \sigma\left(\begin{cases} \frac{1}{|d_i - d_j|} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}\right) \quad (1)$$

where  $d_j$  and  $d_i$  represent the distance of the stations to the city center. The conditional statement ensures the identity matrix is zeroed out, and the sigmoid function  $\sigma$  normalizes the similarity estimations, thus obtaining the weights. Figure 2 and Figure 3 illustrate the different approaches and their respective levels of "awareness".

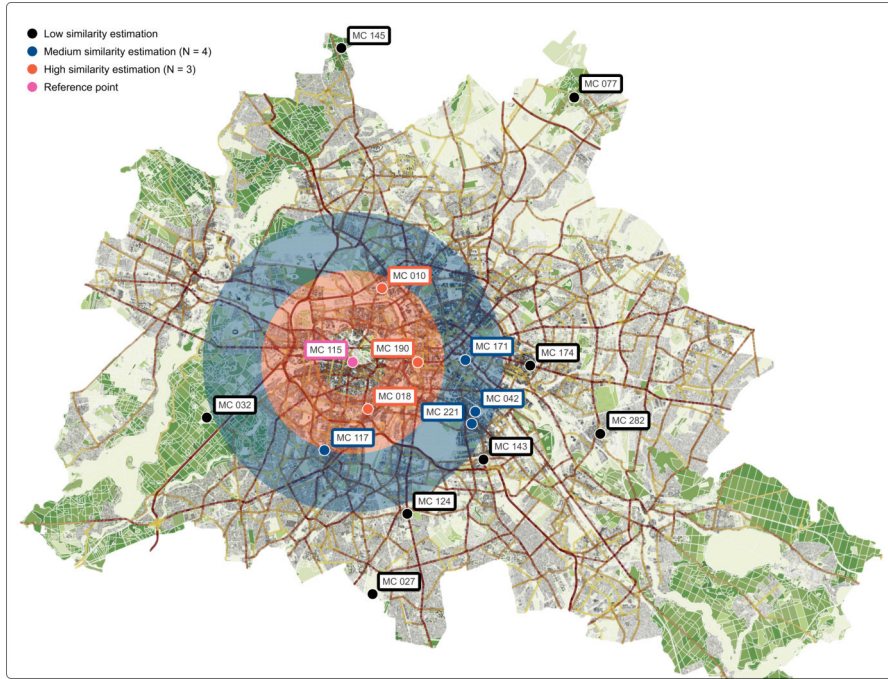


Figure 2: Similarity estimation based on inverse distance, which results in a low number of considered sites with high or medium similarity. The step-wise differentiation in “high” and “mediums” serves an illustrative purpose.

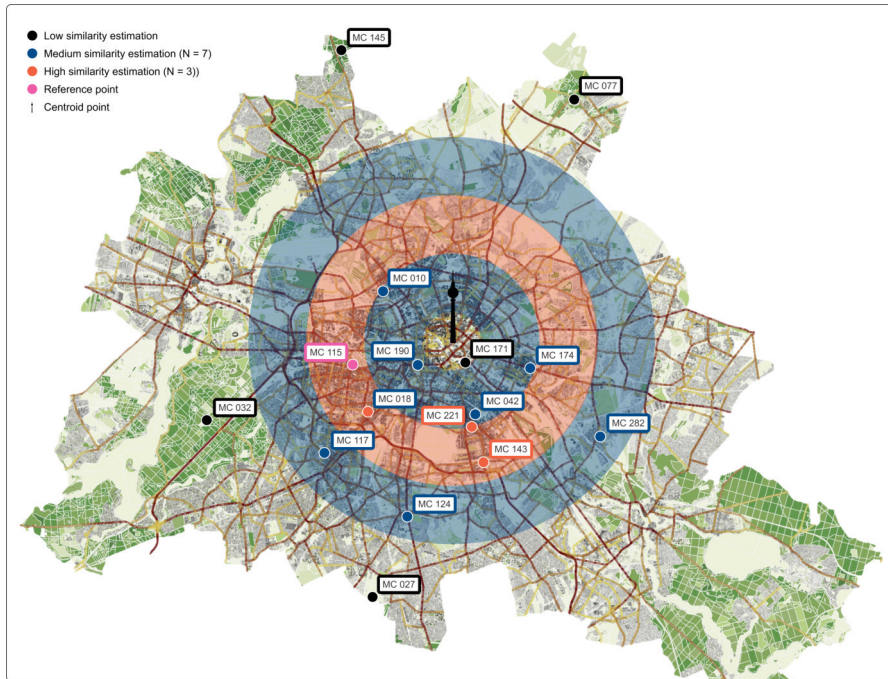


Figure 3: Similarity estimation based on the difference in distance to the centroid. Monitoring sites with a similar distance toward the city center are assigned a high similarity rating compared to those with a higher difference in distance.

This method captures the circular land-use transitions of urban areas, where surface sealing, traffic, and population density typically decrease with increasing distance from the centroid.

For empirical evaluation, a *Pearson correlation* test and a *mutual information* (MI) estimation are conducted to compare the true emission values with the weighted averages calculated by each method. The inverse distance method achieves a correlation of 0.648, surpassing the unweighted average. The proposed difference in distance approach achieves the best correlation score of 0.718. The entropy-based MI exceeds covariance (Kraskov et al., 2004, p. 1) and demonstrates the increased informative value of the *difference in distance* calculation, with a 40% increase compared to the inverse distance. Nevertheless, the simplified distance-based similarity assumption and small sample of stations limit this evaluation and the different spatial proximity estimations are further compared through model performance.

### 3.3 Feature Engineering: Traffic and Population Density

In 2019, the Berlin city administration captured traffic data by counting vehicles at 2,500 points, which was then extrapolated to represent vehicle counts for an average weekday across the main road infrastructure (SenMVKU, 2020, p. 12). To quantify site-specific traffic volume for different radii, the vehicle count is multiplied by its road length within each radius and normalized by the radius size, which is shown in Figure 4. Additionally, the proximity to the nearest intersection is considered, encoding the increase in emissions due to vehicle acceleration (Brokamp et al., 2017, p. 4). The specific traffic emission levels are further differentiated by vehicle type, under the consideration that trucks or buses above 3.5 tonnes exceed the  $NO_2$  emission by a factor of ten (Lighterink, 2017, p. 12). The proportion of these major emitters at the monitoring sites ranges from 2% to 5%, with two notable exceptions where they account for 11% and 18%. Emissions also vary based on the type of fossil fuel used and the traffic flow (Lighterink, 2017, p. 12), though these factors are assumed to be spatially consistent across the dataset. Given that time-dependent variations in traffic volume are not included, peak hours on weekdays are represented as boolean features, derived from traffic density measurements reported in a separate study on Berlin's traffic volume (SenMVKU, 2020, p. 12). As a general approximation for heating-related emissions, the population density in a 200- and 500-meter radius is included.



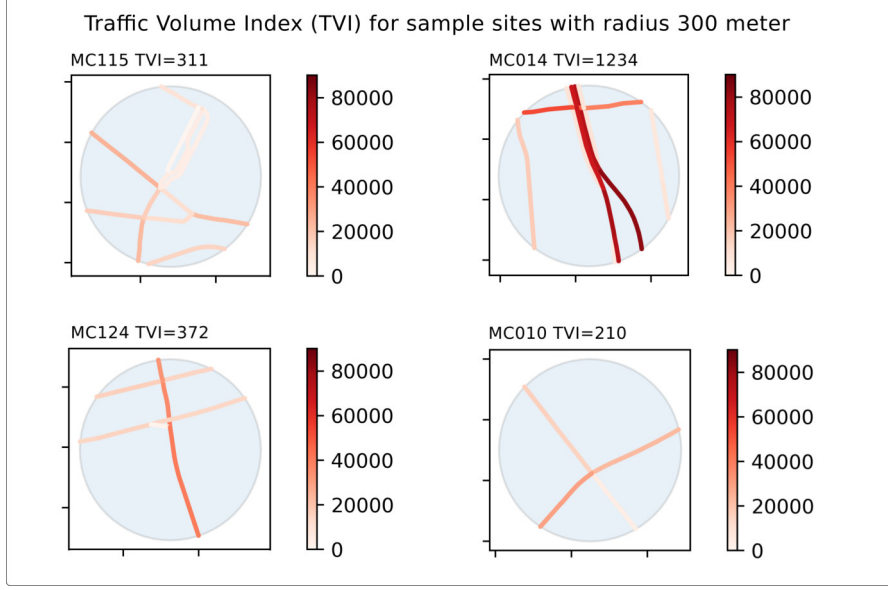


Figure 4: Visualization of different traffic densities surrounding different monitoring sites and their associated traffic volume index (TVI). The daily vehicle count is displayed as the color spectrum.

### 3.4 Feature Engineering: Urban Greenery

To identify the local greenery, the city documentation on urban vegetation from 2020, measured in green volume, is utilized. The dataset comprises 34,498 Berlin-wide units, each with a precise specification of green volume in  $m^3/m^2$ . The volume is calculated as the product of the absolute vegetation and average vegetation height, analyzed through *color infrared orthophotos* (LUP, 2021) as illustrated in Figure 5. These images were captured during an aerial summer flight when vegetation was at its maximum. Greenery fundamentally absorbs visible red light, distinguishing it from non-greenery, and enables the calculation of green volume through the *Normalized Difference Vegetation Index* (LUP, 2021, p. 5). Per-site greenery has been quantified for radii of 25, 50, 75, 100, 200, and 300 meters by intersecting the polygons of specific vegetation volume with an artificial buffer zone around each site. The calculation can be formalized as,

$$GVI_r = \frac{\sum_{i=1}^n (A_i GVI_i)}{\pi r^2}, \quad (2)$$

wherein the green volume index ( $GVI_r$ ) for each radius  $r$  is calculated as the sum of the vegetation volume  $GV_i$  for each polygon insight the radius, multiplied by their size  $A_i$  within the radius as a relative weight, before being normalized by the division of the total area size  $\pi r^2$  to obtain the original unit of  $m^3$  vegetation volume per  $m^2$ . The vegetation areas between the sites can be categorized into four categories with fluent borders. At two of the sixteen sites, the predominantly tree-free surroundings average less than one  $m^3/m^2$  of

vegetation. The second group of eight sites with frequent street-side trees have values below three. Five sites have a vegetation volume greater than three and are surrounded by nearby parks or a high number of trees, as shown in Figure 5. The last category is distinct from the others as it is located within the forest and scores above fifteen  $m^3/m^2$  of vegetation.

To integrate the seasonal variation in urban green, the *Leaf Area Index* (LAI) throughout the year is included as a predictor. To assess and quantify this variation, the grid-based dataset of Yan et al. (2024) on global LAI with an eight-day temporal and five-kilometer spatial resolution, is filtered for the target area with a buffer zone around Berlin to cover a total region of around 80 square kilometers. The area average for each time step is normalized over the months to create a vegetation factor that peaks at 1 during maximum vegetation and is then extrapolated across all hourly time steps.

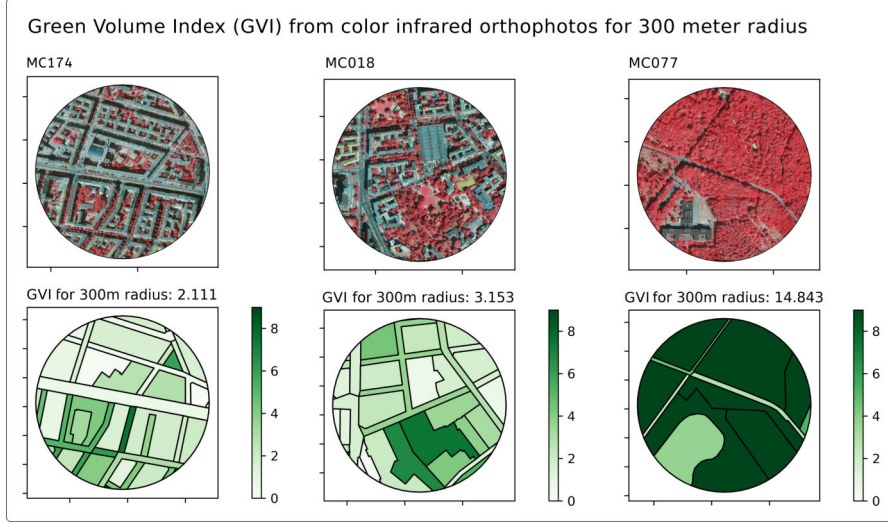


Figure 5: Green Volume Index as average green volume in  $m^3/m^2$  for different radii, built upon infrared imagery.

### 3.5 Feature Engineering: Street Canyon

The architectural features of each site that affect the flow of air and enable pollution trapping in dense architecture (Nazridoust & Ahmadi, 2006, p. 4) are estimated using a dataset on building heights. A spatial echo concept was developed to determine the surrounding area without any building interference within the 50- and 200-meter range, as visualized in Figure 6. This methodology generates 360 artificial lines of the radius length which are spread out evenly in all directions and analyzed for intersections with the polygons of surrounding buildings. This measures the distance from the monitoring site to the nearest building and allows for the feature engineering of total

proportion without interception. In addition, the exact degrees that are free of buildings are stored to construct an additional feature to assess whether the occurring wind from a specific direction is blocked by interfering buildings, which determines local air exchange (Nazridoust & Ahmadi, 2006, p. 2).

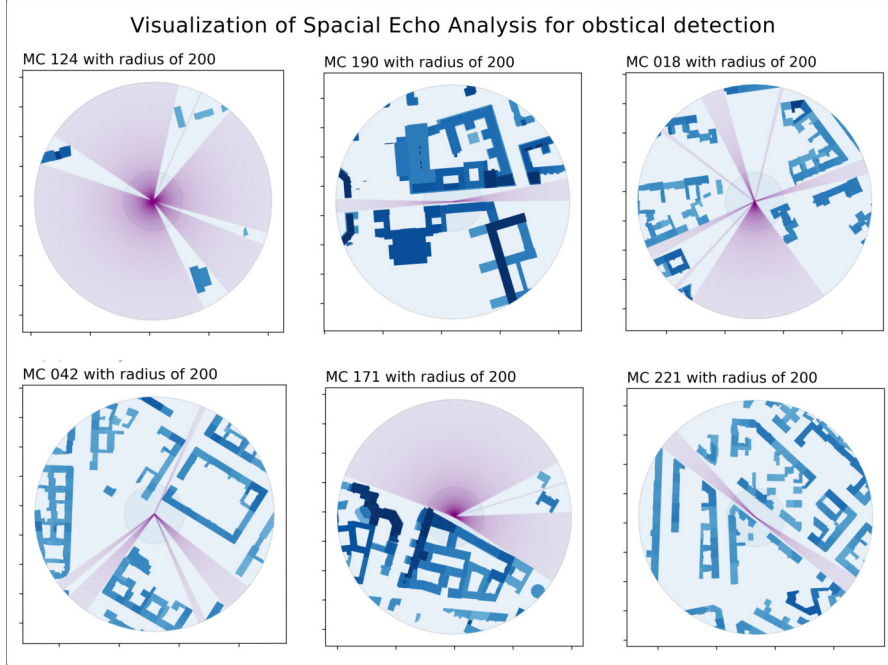


Figure 6: Spatial Echo Analysis for obstacle detection to quantify local wind flow. The obstacle-free degrees are displayed by purple lines, that indicate no intersection with surrounding polygons (buildings).

### 3.6 Missing Data Imputation: Meteorological Data

Throughout the literature on air pollution estimation, meteorological factors such as wind speed and air temperature are highly descriptive (Selmi et al., 2016, p. 196; Gocheva-Ilieva & Livieris, 2020, p. 54; Rybarczyk & Zalakeviciute, 2018, p. 10). In Berlin, the German Weather Service (*Deutscher Wetterdienst*) tracks and provides hourly data for temperature, precipitation, wind speed, and direction. These parameters are selected based on prior research implementations and with consideration for collinearity among the features, which is further discussed in the Section [Mitigation of Multicollinearity through Feature Selection](#).

The eight original weather parameters have a total of 344 missing instances over the year, which represent approximately two percent of time steps with one or more missing values. The imputation is preferred to avoid introducing bias through deletion when data are not missing at random. Plotting the absence of sensor data over time

reveals the existence of a time-dependent correlation in absence, as shown in Figure 7.

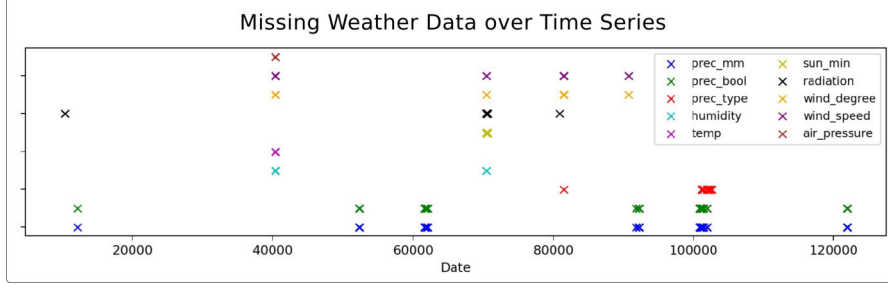


Figure 7: Missing meteorological data plotted over time.

This reliance on preceding and subsequent data points challenges the ability of traditional bidirectional imputation techniques, such as linear or spline interpolation, to accurately impute data for longer sequences of missing values. As the missing values are further clustered between features, as shown in Figure 7, traditional regression-based imputation using contemporaneous features is impractical. Given these complexities, linear and spline interpolation are evaluated against two distinct univariate Long-Short-Term-Memory (LSTM) based multi-step forecasting architectures for their per-feature prediction performance. The meteorological data from the same weather station for the preceding year is utilized to train and test these imputation methods with artificially generated missing values. For prediction, the 24 time steps preceding each missing instance are employed. The architecture of those two Recurrent Networks can be described as a single LSTM and a second hybrid model, which utilizes a convolutional layer before the LSTM layer. Both architectures and their specification and hyperparameter tuning, are further discussed in Appendix B.1 and motivated by the implementation of Brownlee (2020). The results are provided in the Section [Imputation Results of Missing Meteorological Data](#) and discussed in the Section [Assessment of Research Questions](#).

### 3.7 Mitigation of Multicollinearity through Feature Selection

The phenomenon of collinearity, which describes the shared informative value among two or more predictor variables (Chan et al., 2022, p. 2), can lead to inaccurate coefficient estimation in logistic regression, though its impact is considered marginal in random forest and neural network models (Veaux & Ungar, 1994, p. 5). The primary objective in reducing collinearity is the integrity of the post-hoc feature importance, since the contribution of individual features can otherwise be relativized and inconclusive (Molnar, 2023; Chan et al., 2022, p. 2). Furthermore, permutation-based model-agnostic methods can gener-



ate implausible feature value combinations, an issue that is discussed further in the Section [Feature Importance Analysis](#).

To mitigate the effects of collinearity, univariate feature selection is employed, selecting the most informative features based on their Mutual Information (MI) with the target, within clusters of multicollinear variables. This selection process uses only training data to prevent data leakage and is favored for its computational efficiency compared to incremental, wrapper-based methods (Chan et al., 2022, p. 4). While *Principal Component Analysis* (PCA) offers an effective alternative for preserving information that would otherwise be lost by excluding partly redundant features, its abstraction complicates the analysis of feature importance and prevents feature-wise residual analysis (Chan et al., 2022, p. 5). The collinearity of each predictor is estimated by their *Variance Inflation Factor* (VIF), with a threshold for exclusion set to ten (Chan et al., 2022, p. 5). VIF values are calculated by selecting each feature as the dependent variable in a linear regression model that uses the remaining features as predictors, with the results and their associated VIFs presented in Table 1.

Feature Selection Through MI and VIF		
Feature	MI	VIF
prec_mm	0.0016	1.136941
weekend	0.0055	1.034408
prec_bool	0.0061	1.571798
rushhour	0.0098	3.173482
wind_degree	0.0366	5.478492
temp	0.0411	5.813787
wind_speed	0.0679	6.148629
lai_factor	0.0701	16.142953
free_wind	0.0718	4.721348
pop_500	0.2417	6.090651
tvi_200	0.2446	5.697859
GVI_25	0.2447	2.219108
prop_main	0.2470	3.311182
prop_intercept_200	0.2470	11.466473
nearest_in	0.2477	3.185428
mean_pollution	0.3835	4.433558

Table 1: Feature Selection based on *mutual information* (MI) towards the target value  $NO_2$  and low multicollinearity estimated by *Variance Inflation Factor* (VIF) of the training data.