FROM LEAVES TO BREEZES:

PREDICTING NITROGEN DIOXIDE CONCENTRATION FROM SURROUNDING URBAN GREENERY AND METEOROLOGICAL, SPATIAL, AND TRAFFIC CHARACTERISTICS

RICHARD SCHMIDT

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE SOCIETY AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES OF TILBURG UNIVERSITY

STUDENT NUMBER

458423

COMMITTEE

dr. Sharon Ong
Kyana van Eijndhoven

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 22th, 2024

WORD COUNT

8799

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr Sharon Ong, for her guidance throughout this process, which has often exceeded both my expectations and the duration set for our weekly meetings. Secondly, I would like to take this opportunity to acknowledge and thank my girlfriend Johanna for her ever-forward-thinking support throughout the iterative creation of this thesis. Equally, I would like to thank my parents for their unconditional support in getting me to where I am today. I would like to further express my gratitude to the Berlin Geoportal and the Deutscher Wetter Dienst – DWD for their efforts in promoting and establishing an open data culture, which is the foundation of this and many preceding research. Finally, I would like to express my appreciation to SciHub, without which I would have been left without access to diverse scientific research and thus, in many places, the professional depth of this study would have been diminished.

FROM LEAVES TO BREEZES:

PREDICTING NITROGEN DIOXIDE CONCENTRATION FROM SURROUNDING URBAN GREENERY AND METEOROLOGICAL, SPATIAL, AND TRAFFIC CHARACTERISTICS

RICHARD SCHMIDT

Abstract

This thesis addresses the challenge of predicting nitrogen dioxide (NO_2) levels using urban greenery and meteorological, spatial, and traffic data. Despite the variety of previous machine learning approaches to air pollution estimation, the consideration of the interaction effects of urban greenery and its seasonal variations is underrepresented. Previous studies integrated local vegetation through land-use regression and land-use random forest models without the consideration of seasonal variation. This thesis distinguishes itself by adapting dynamic land-use characteristics to a state-of-the-art graph representation, to examine the seasonal variation in the mitigation effect of vegetation. In order to achieve this, a random forest regressor (RF) is compared to multistep graph embedding for spatial representation learning, focusing on predictive performance, robustness, and interpretability of feature contribution and interaction. Utilizing data from sixteen monitoring stations in Berlin throughout 2023, this thesis integrates extensive meteorological and geological datasets provided by the Berlin Geoportaland the Deutscher Wetterdienst – DWD. The results show that the graph neural network outperforms the RF, with a notable reduction in the between-test-site variation of predictive accuracy, with a standard deviation of 0.06 compared to 0.16 for the RF. Nevertheless, due to the complexity and homophily of the graph structure, only the RF allows for local interpretability through Shapley value, which indicates that urban greenery mitigates NO2 levels in dependence on seasonal variation. However, further research is required to control for confounding attributes. This research contributes to the fields of urban planning and environmental policy by quantifying the pollution mitigation potential of urban greenery and contributing to the awareness of seasonal variation.

DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The constructed data frame merges geographical and meteorological data under open source regulation, which is attributed by the Berlin Geoportaland the Deutscher Wetterdienst(German Weather Service) respectively. The following enumeration of attribution lists the original title of the datasets and their legal ground for use and reproducibility.

The usage of the meteorological data is regulated by the "Creative Commons BY 4.0" (CC BY 4.0) and covers the following data sets:

- Hourly station observations of 2 m air temperature and humidity for Germany, Version v24.03
- Hourly station observations of precipitation for Germany, Version v24.03
- Hourly station observations of pressure for Germany, Version v24.03
- Hourly station observations of solar incoming (total/diffuse) and longwave downward radiation for Germany, Version v24.03
- Hourly mean value from station observations of wind speed and wind direction for Germany, Version v24.03

The usage of the geographical and land-use data is regulated by the "Data licence Germany – attribution – version 2.0" (DL-DE->BY-2.0) and covers the following data sets:

- Geoportal Berlin / Digitale Color-Infrarot-Orthophotos 2020 (DOP20CIR) – Sommerbefliegung
- Geoportal Berlin / Gebäudehöhen (Umweltatlas)
- Geoportal Berlin / Einwohnerdichte 2022 (Umweltatlas)
- Geoportal Berlin / Verkehrsmengen DTV 2019 (Umweltatlas)
- Geoportal Berlin / Messdaten des Berliner Luftgütemessnetzes

Thirdly, the spatial-temporal leaf area index dataset by Yan et al. (2024), licensed under "Creative Commons Attribution 4.0", is integrated.

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.

All the figures belong to the author and have been created through the Python libraries Matplotlib and Seaborn or Figma. The thesis code can be accessed on GitHub. The used packages and libraries are listed in the Appendix A. Part of the code for graph embedding has been adapted by the authors from Vu et al. (2024) and Zhan et al. (2018). The architecture for multi-step forecast, applied for missing data imputation, has been altered from the tutorial by Brownlee (2020). The adapted code fragments are indicated in the notebook. Furthermore, GPT4 was used for simplification, expansion, specific adjustments, and error handling of the written code. In terms of writing, the author used assistance with the language of the paper. Both *Deepl* and *Grammarly* were applied to improve the author's original content, for paraphrasing, spell-checking, and grammar. Nevertheless, no generated text has been inserted in the following thesis. No other typesetting tools or services were used.

CONTENTS

1 INTRODUCTION

1.1 *Project Definition*

This research aims to predict the air pollution concentration of nitrogen dioxide (NO_2) at spatially fixed monitoring sites, based on the simultaneous concentration at neighboring stations as well as meteorological and land-use data. Therefore, the NO_2 measurements of sixteen monitoring sites in Berlin are used. The integration of urban greenery, its interaction with other predictors, and the dynamic seasonal variation of greenery have been neglected in the prediction of emission concentrations in previous work. This study seeks to address these gaps. To capture the variation between the sixteen stations over an hourly time series of the year 2023, the site-specific traffic and population characteristics, the variation in green volume, and the surrounding building structure are included as predictor variables. Berlin is favored as a research area due to the lack of previous applications of machine learning in air pollution prediction and the availability of open data, which provides detailed ecological land-use data.

This study compares a random forest regressor (RF) and a graph neural network (GNN) in terms of their predictive performance, robustness, and interpretability of feature contribution and interaction. Recent applications of graph-based representational learning in the context of air pollution estimation have argued in favor of its ability to extract spatial relationships (Iskandaryan et al., 2023, p. 3; Vu et al., 2024, p. 2), but fall short in evaluating its capabilities towards explainability.

1.2 *Societal Motivation*

Despite a significant reduction in air pollution-induced fatalities within the *European Union*, air pollution concentrations continued to exceed the World Health Organization guidelines in 96% of urban areas in 2020 (EEA, 2023). The urgent need for mitigation is further emphasized by the perspective of a potential increase due to ongoing changes in atmospheric composition (Silva et al., 2017). This study aims to explore the local potential of urban greenery for air pollution mitigation. Insights into the dynamics of these interactions could inform urban planning and prompt a reconsideration of land-use policies. Estimating the mitigation capacity of local greenery and the impact of seasonal vegetation changes could support discussions and the development of countermeasures against air pollution.

1.3 *Scientific Motivation*

Despite the extensive history of air pollution prediction research (Rybarczyk & Zalakeviciute, 2018), this thesis is motivated by the insufficient integration of land use characteristics within graph representations. Consequently, it aims to address the significant underrepresentation of scientific studies on vegetation's capacity to mitigate air pollution through this novel perspective of graph representation. Unlike previous research, this thesis analyzes green volume as a dynamic predictor, recognizing its variability over time. Additionally, it develops a comprehensive land use dataset to characterize the local environment at each monitoring site and is the first machine learning application on Berlin air pollution data.

1.4 *Research Questions*

RQ1: *To what extent can the emission level of NO2 at monitoring sites in Berlin be predicted by urban greenery, meteorological, spatial, and traffic factors?*

As a stepping stone towards the main research question, the missing meteorological data is imputed at 344 time steps. The missing sensor data is clustered both over time and between features, which poses a challenge to traditional imputation and interpolation techniques. Therefore, the capability of recurrent-based forecasting is explored in the first sub-question:

SQ1: *To what extent can univariate multistep forecasting be used as an imputation technique for meteorological features, in contrast to linear and spline interpolation?*

The subsequent questions investigate the ability and interaction effects of green volume for air pollution decomposition and thereby contribute to existing academic decent by establishing an innovative interpretation through data science techniques.

The distance between the emitter and vegetation, under which the latter contributes to emissions reduction, is under academic debate and varies extensively (Eeftens et al., 2012, p. 11199; Brokamp et al., 2017, p. 4). This research aims to contribute to this discussion and provide an orientation for future research by evaluating a significant buffer radius.

SQ2: *To what extent can the optimal buffer radius from a sensing site to surrounding vegetation be determined through entropy- and correlation-based similarity estimations?*

Previous studies have yielded conflicting results regarding the mitigation effects of greenery in interaction with seasonal changes in vegetation volume. While Setälä et al. (2013, p. 107) find no significant seasonal variations, Escobedo and Nowak (2009, p. 109) contends that total leaf area is pivotal for reducing air pollution. This study addresses these discrepancies by quantifying seasonal variations in green volume through the Leaf Area Index (LAI), an established measure derived from satellite imagery in the target area. Unlike prior machine learning analyses that treated vegetation as a static predictor, this thesis thereby acknowledges the dynamic nature of vegetation's seasonal variations.

SQ3: *To what extent can Shapley values, as local, model agnostic interpretation, be utilized to analyze variations in the mitigation of NO_2 concentrations relative to seasonal changes in the Leaf Area Index?*

1.5 *Summary of Findings and Contributions*

Both models exhibit comparable predictive performance, with R^2 scores of approximately 0.6. Nonetheless, the graph neural network (GNN) demonstrates superiority over the random forest (RF) model by significantly reducing the variation in predictive accuracy across test sites, evidenced by a standard deviation of 0.06, compared to 0.16 for the RF.

The conducted residual and feature analysis aligns with prior findings, underscoring the importance of vehicle type and the interaction between seasonal variations and the mitigation capabilities of vegetation. Further, post-hoc interpretation of individual feature contributions, analyzed using Shapley values, reveals that the local NO_2 concentration estimates at stations with higher green volume exhibit greater variability across seasons. Notably, at stations with high Leaf Area Index (LAI), the negative contribution to NO_2 estimation during the summer months—when green volume peaks—is more than double that observed at stations with lower levels of greenery. However, these findings are constrained by a small sample size of test stations, confounding land-use characteristics, and the complex interactions involved in aerosol decomposition. The observed marginal decrease in the contribution of urban vegetation to NO_2 mitigation under peak temperatures highlights areas for further research and integration with existing domain knowledge.

Additionally, the methodologies developed for the *spatial similarity estimation* (3.2) and the *stratified, spatial cross-validation* (4.3) prove to be superior to the approaches they were compared against.

2 LITERATURE REVIEW

The estimation of air pollution levels through machine learning represents a broad and continuously evolving field that bridges environmental science and urban planning (Jain et al., 2022, p.1). The advancement of data-driven approaches for predicting air pollution is a key focus of research, driven by the need to model the non-linear relationships inherent in heterogeneous data sets (Rybarczyk & Zalakeviciute, 2018, p.2). These techniques also account for spatial and temporal dependencies. A comparative study from 2018 found neural networks, ensemble models, support vector machines, and multiple linear regression among the best-performing models (Rybarczyk & Zalakeviciute, 2018, p. 22; Turek & Kamińska, 2022, p. 1). However, no single model consistently demonstrated the highest performance and the R^2 scores varied considerably across prediction, interpolation, or forecasting tasks (Rybarczyk & Zalakeviciute, 2018, p. 22; Turek & Kamińska, 2022, p. 1). Despite these advancements in model architecture, the specific impact of vegetation on NO_2 mitigation has not been fully explored, representing a gap in the interception of environmental factors with technological innovations.

2.1 *Vegetation for NO_2 Mitigation*

Historically, the mitigating effect of vegetation on air pollution has been primarily investigated through hypothesis testing and the use of rule-based environmental model software. Among these, the I-Tree software, which models the impact of urban green spaces on environmental interactions, is a widely recognized tool in this field of research (Selmi et al., 2016, p. 193). However, these predefined systems are limited by their inability to adapt to changing environmental conditions or to learn from new data. While various approaches agree on the positive mitigating effects of vegetation, they significantly diverge in their quantification of this effect (Selmi et al., 2016, p. 198; Srbinovska et al., 2021, p. 11).

2.2 *Vegetation in Machine Learning: land-use Models*

Shams et al. (2021, p. 2) highlighted the general lack of local vegetation for air pollution estimation but found them to be highly predictive (Shams et al., 2021, p. 5). Land-use regression or land-use random forest models incorporating their surrounding conditions of monitoring sites, but represent vegetation on a superficial and static level, with mixed or no explicit results as to their feature importance (Larkin et al., 2023, p. 7; Brokamp et al., 2017). In addition to vegetational characteristics, land-use models incorporate environmental and combustion-related predictors.

2.3 *Temporal dependency*

Throughout the research, time dependency has been recognized as a critical descriptive predictor, extending beyond pollution forecasting. Traditional methods have leveraged temporal auto-correlation through time-series analysis techniques such as ARIMA or STRK (Gocheva-Ilieva & Livieris, 2020, p. 58; Zhan et al., 2018, p. 468). In addition, the discrete encoding of features with time delay was found to be predictive to capture temporal dependencies (Shams et al., 2021, p. 5). In the domain of recurrent-based sequence-to-sequence modeling, various architectures incorporating Long-Short-Term-Memory (LSTM) or Gated Recurrent Units (GRUs) have been utilized (Vu et al., 2024; C. J. Huang & Kuo, 2018)and are continuously integrated into novel deep-learning architectures.

2.4 *Spatial Dependency*

Recent advancements in state-of-the-art architectures for spatial air quality estimation have sought to leverage spatial relationships through graph-based data representation. Traditional geostatistical techniques, such as *Kriging* require the specification of spatial correlation estimations that depend on domain-specific knowledge of underlying spatial dependencies (Vu et al., 2024, p. 2). In contrast, previous data-driven approaches like grid-based analysis using convolutional neural networks (CNNs) often rely on the simplistic assumption of uniform regional distribution, which is limited in its ability to generalize to varied geographies (S. Wang et al., 2020, p. 2).

Overcoming these limitations, graph-based representations effectively incorporate inter-station dependencies by aggregating attributes from neighboring stations as spatial lag. This neighborhood configuration may be initialized based on predefined proximity estimations or dynamically trained through attention mechanisms (S. Wang et al., 2020, p. 3). Different methodologies for graph embeddings can therefore be conceptualized as a geographical layer, characterizing the relationships between individual monitoring stations (Vu et al., 2024, p. 3; Iskandaryan et al., 2023, p. 2736). These methodologies are predominantly implemented as undirected and temporally static graph structures (Iskandaryan et al., 2023, p. 2731).

In contrast, the implementation of knowledge-enhanced dynamic or directional graph structures facilitates the consideration of variable spatial dependencies, such as current wind direction and strength, as evidenced by studies like those of Xiao et al. (2022) or S. Wang et al. (2020). Particularly, the promising results of the sequential encoding of meteorological characteristics led to the incorporation of recurrent architectures to graph neural networks. However, despite these innovations, different novel spatio-temporal graph network approaches,

which utilize comprehensive spatial and temporal data, have barely surpassed previous achievements (Iskandaryan et al., 2023, p. 2740).

Currently, there is no established benchmark dataset for graph-based spatial air pollution estimation. A dataset covering six air pollutants in Beijing for the year 2017 (H. Wang, 2019) is frequently used in various graph-based approaches (Vu et al., 2024; Y. Huang et al., 2021; Seng et al., 2021). The Berlin dataset significantly differs from the Beijing dataset, which features a more densely distributed network of 35 monitoring stations and higher overall emission distribution.

2.5 *Research Strategy*

The existing research gap in machine learning towards the integration of vegetation lies in the exclusion of seasonal variations and reliance on land-use regression and random forest models. Additionally, there is a notable lack of clarity concerning the magnitude of vegetation's impact on air pollution prediction, which will be addressed through feature importance analysis for both models. This thesis addresses these gaps by first implementing a land-use random forest model that leverages Berlin's monitoring and geological data. This model is distinct from its predecessors as it examines the effects of seasonal variation in leaf area. Consequently, post-hoc local interpretability through Shapley values is employed to investigate the temporal variation in the negative contribution of NO_2 estimation under changing vegetation volumes.

Secondly, a novel integration of land-use characteristics to graph neural networks is implemented as state-of the-art architecture to assess whether incorporating spatial dependencies enhances explanatory power. The architecture, adapted from Vu et al. (2024), is characterized as a static, bidirectional, and non-temporal graph representation. Unlike the majority of machine learning research in air pollution prediction, which predominantly focuses on temporal correlations, this study emphasizes the exploration of spatial dependencies.

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 Geoportaland 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 engeneering.

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 NO2 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:

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

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 σ normalizes the similarity estimations, thus obtaining the weights. Figure 2 and Figure 3 illustrate the different approaches and their respective levels of "awareness".

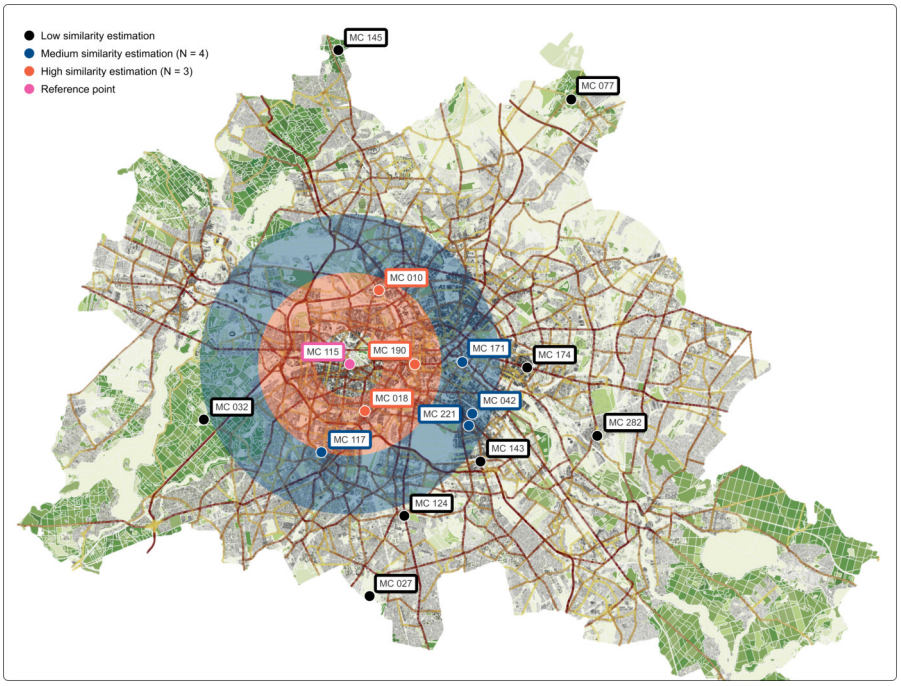

Figure 2: Similarity estimation based on inverse distance, which results in a low number of considered sites with high or medium similarity. The stepwise differentiation in "high" and "mediums" serves an illustrative purpose.

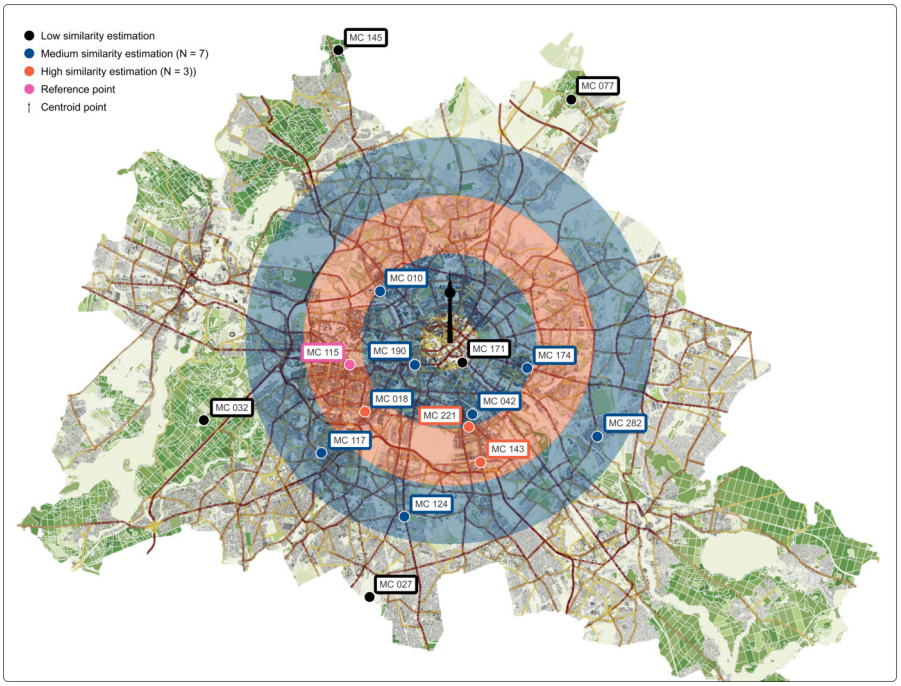

Figure 3: Similarity estimation based on the difference in distance to the centroid. Monitoring sites with a similar distance toward the city center are assigned a high similarity rating compared to those with a higher difference in distance.

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.

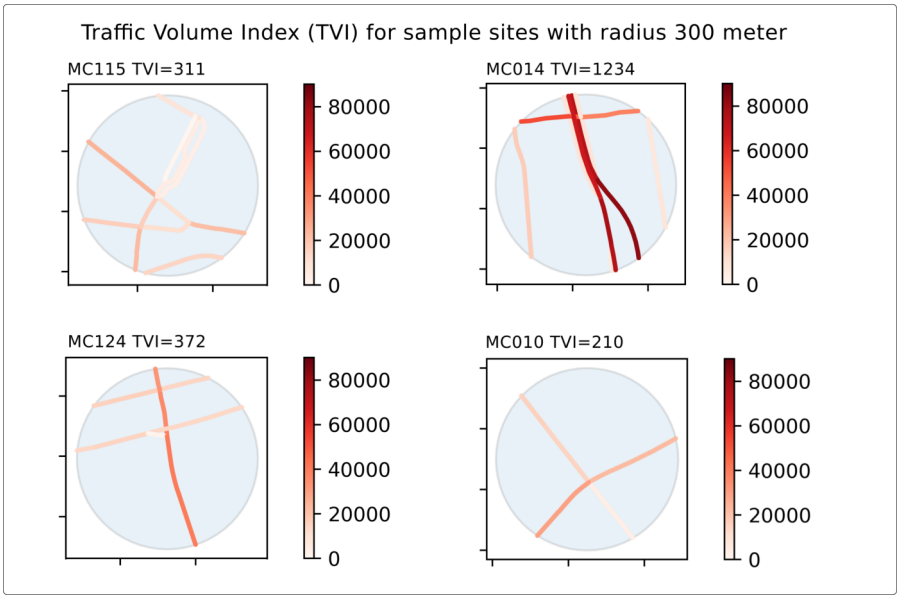

Figure 4: Visualization of different traffic densities surrounding different monitoring sites and their associated traffic volume index (TVI). The daily vehicle count is displayed as the color spectrum.

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,

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

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 πr^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.

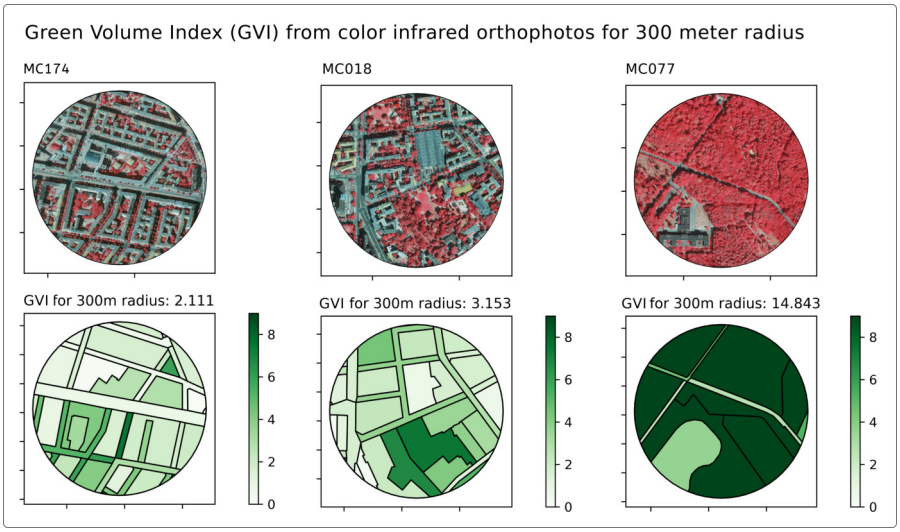

Figure 5: Green Volume Index as average green volume in m^3/m^2 for different radii, built upon infrared imagery.

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).

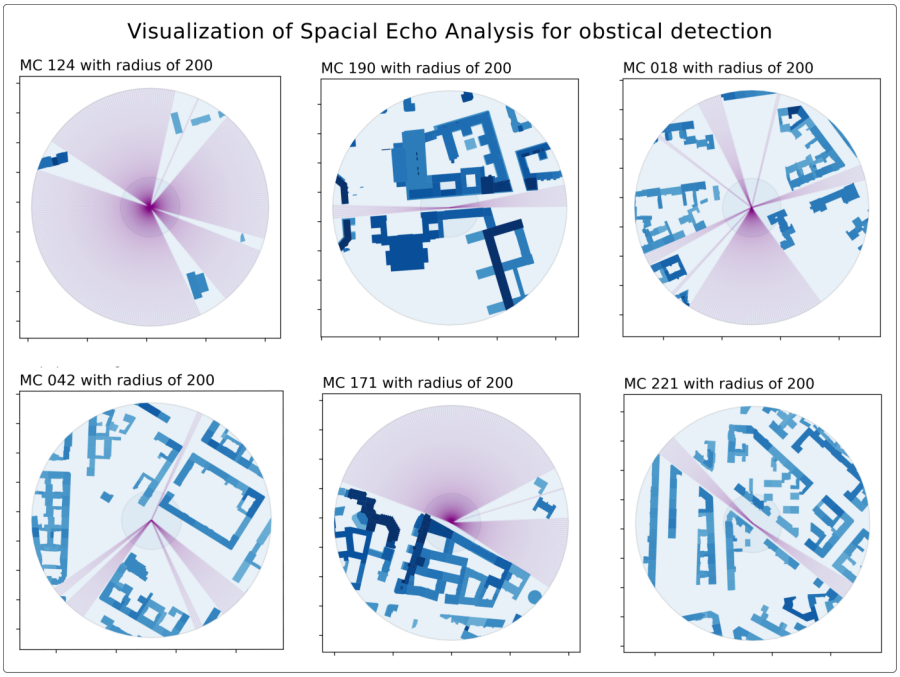

Figure 6: Spatial Echo Analysis for obstacle detection to quantify local wind flow. The obstacle-free degrees are displayed by purple lines, that indicate no intersection with surrounding polygons (buildings).

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.

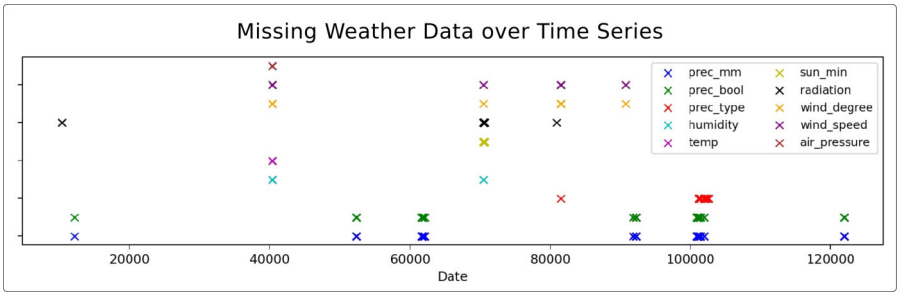

Figure 7: Missing meteorological data plotted over time.

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 *Variation 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.

4 METHODOLOGY

Both architectures are adapted using comparable training schemes and evaluated against the same held-out data set to maintain the integrity of their comparison. The development of stratified, spatial cross-validation and its implications for the generalizability of the test results are detailed in the Section Generalization through stratified, spatial Cross-Validation, following a brief introduction of both architectures. The implementation details further entail the introduction of the Bayesian-based Hyperparameter Tuning and conclude with a methodological introduction of the proposed comparison scheme through selected Evaluation Metrics and Feature Importance Analysis. The data flow is illustrated in Figure 8.

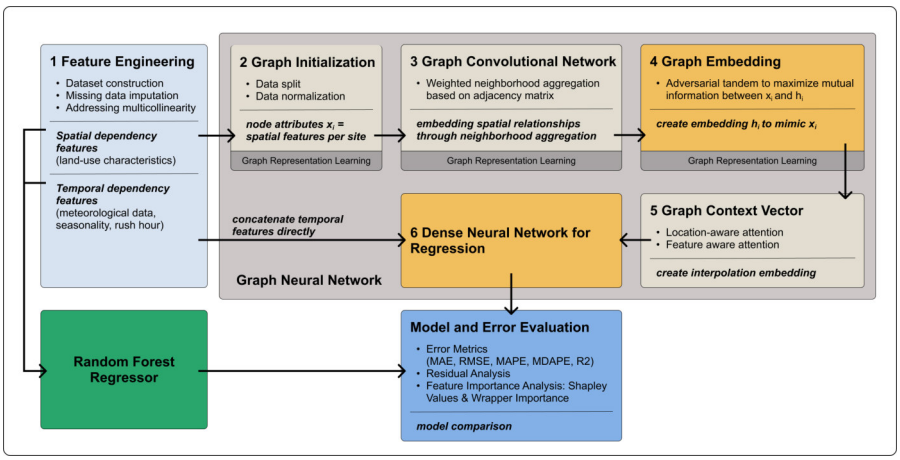

Figure 8: Illustration of data flow through both models after initial data preparation.

4.1 *Random Forest*

A land-use random forest regression has been demonstrated to be a robust approach in pollution estimation, effectively modeling non-linear relationships within heterogeneous data sources (Rybarczyk & Zalakeviciute, 2018, p. 8; Gocheva-Ilieva & Livieris, 2020, p. 54). This algorithm combines the transparency and simplicity of decision trees with the ability to structure non-linear relationships. The robustness and generalizability of the regression are enhanced by the ensemble structure, which aggregates multiple trees alternated through instance bootstrapping and feature bagging (Brokamp et al., 2017, p. 2). Spatial cross-validation is applied to adjust hyperparameters, details of which are provided in Table 2.

Hyperparameter Random Forest		
Parameter	**Parameter Range**	**Best Performaning**
criterion	default=squared error, friedman mse, absolute error	squared error
max depth	default=None, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100	None
min samples leaf	default=1, 2, 4	4
min samples split	default=2, 5, 10	5
n estimators	default=100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 2000	100

Table 2: Hyperparameter for random forest model.

4.2 *Graph Neural Network*

A contrastive, multi-step graph embedding is implemented to address the spatial dependencies among the monitoring sites. This architecture and its implementation in PyTorch are inspired by the work of Vu et al. (2024) and a detailed description of the applied alterations to the architecture can be found in Table 3.

4.2.1 *Graph Structure and Initialization*

The graph embedding process is segmented into three consecutive steps of neighborhood aggregation, wherein the feature representation of each monitoring station is enhanced by integrating attribute vectors from geographically proximate stations. A comprehensive, literature- and formula-based description of step-wise feature aggregation is available in Appendix C. Unlike traditional tabular data structures, graph representations benefit from an additional layer of information encoded in the relationships between instances, which facilitates the exploitation of spatial relations (Vu et al., 2024, p. 7). The Graph representations, denoted as $G = (N, A)$,comprise nodes N, that represent different monitoring sites with their specific land-use attributes.Although meteorological features are not embedded directly into the graph, they are combined with the graph embeddings and used as input for a subsequent dense neural network. The interconnectivity between nodes, or edge relations, is defined by reciprocal similarity scores among the stations. This dense and undirected network is encapsulated in an adjacency matrix A, characterized by the *difference in distance to the centroid*.

Adaptation to T-GCN by Vu et al. (2024)	
Architectural component	**Description of applied alterations**
Considered Features	While Vu et al. (2024) incorporate meteorological and air pollution data, the applied feature set has been extended to consider the verity of land-use characteristics listed in Table 4.
Graph initialization	Vu et al. (2024) define the inter-station relationship through **inverse distance**, which builds the adjacency matrix. The similarity estimation has been adjusted to the proposed **difference in distance to centroid** approach. The implication of this difference cannot be understated, as both the convolutional neighborhood aggregation and the attention scoring rely on this initialization.
Encoding Architecture	Since we do not include spatio-temporal features, the recurrent unite is excluded from the graph representation learning.
Integration of Hyperopt	To utilize the Bayesian parameter Search, *hyperopt* has been integrated in the model pipeline.
Adjusted Hyperparameter	Parameter tuning has been utilized as detailed in Table 8

Table 3: Applied adaptations to the architecture and implementation by Vu et al. (2024).

4.2.2 *Graph Representation Learning*

The initial propagation of neighborhood information, weighted by the adjacency matrix, employs a *graph convolutional transformation*, initially proposed by Kipf and Welling (2016) and illustrated in Figure 9. Subsequently, a lower-dimensional attribute embedding is learned through a contrastive generator-discriminator architecture, inspired by the *Spatio-temporal Deep Graph Infomax* approach of Opolka et al. (2019). This embedding process aims to maximize the mutual information between the attribute vector and the embedding, facilitating the distinction between positive and corrupted attribute-embedding pairs. Lastly, a *location and a feature-aware attention* block is developed to selectively enhance the relevance of spatial similarities among the characteristics of the monitoring stations and the learned embeddings.

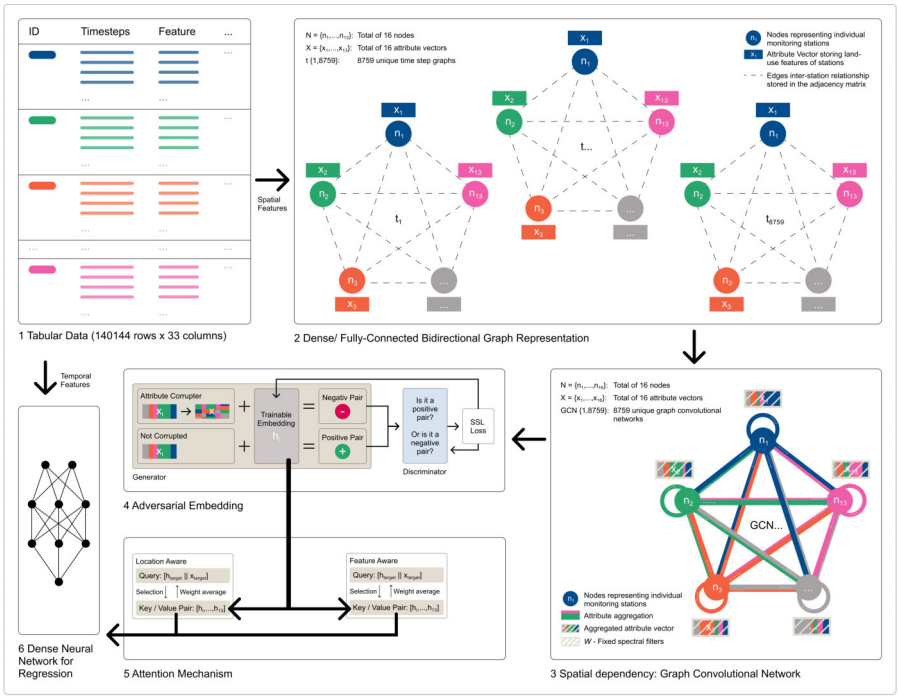

Figure 9: Detailed illustration of data processing steps for graph representation learning. The initial tabular structure is transformed into a graph structure, where each monitoring site is represented as an individual node. During the Graph Convolution, the neighboring information is propagated, wherein the site similarity is defined in the adjacency matrix. The aggregation is transformed by an adversarial embedding before being processed by two attention blocks. The output is concatenated with the meteorological data as an input layer for the neural network for regression.

4.2.3 *NO_2 prediction with Dense Neural Network*

For the regression of the NO_2 concentration at the target locations, the previously learned location and feature-aware context vectors are concatenated with the temporal features to construct the input layer for the following neural network. The meteorological data, listed in Table 4 enriches the context of the target value and is therefore directly included. The architecture of the dense network, along with the parameter search process, is described in the following section, with all selected, best-performing parameters detailed in Table 8.

Spatial and Temporal Features for GNN	
Spatial Features (temporal consistent)	'prop_intercept_200', 'GVI_25', 'tvi_200', 'prop_main_', 'nearest_in', 'pop_500', 'free_wind'
Temporal Features (spatial consistent)	'weekend', 'rushhour', 'lai_factor', 'prec_mm', 'prec_bool', 'temp', 'wind_speed', 'wind_degree'

Table 4: Considered spatial and temporal features.

4.3 *Implementation Details*

4.3.1 *Bayesian-based Hyperparameter Tuning*

The *Tree-structured Parzen Estimator (TPE)*, initially proposed by Bergstra et al. (2011), is employed as an adaptive method for hyperparameter optimization. This model, initialized with a predefined dictionary of hyperparameters and their respective assumed distributions, has demonstrated superior computational efficiency compared to traditional random or grid-based search methods. This efficiency stems from the TPE's ability to estimate the expected improvement (EI) in performance for a new set of parameters via surrogate models (Bergstra et al., 2011, p. 3). To refine the search, two surrogate models are progressively updated based on the current performance outcomes of the tested parameters, thereby iteratively approaching the optimal parameter set (Bergstra et al., 2011, p. 3). This process utilizes Bayesian statistics, an iterative method of likelihood estimation that leverages prior knowledge for posterior updating.

4.3.2 *Generalization through stratified, spatial Cross-Validation*

The proposed data split has been developed with consideration of spatial autocorrelation and the potential selection bias (N= 16), which determine the generalizability of the test results. Due to the computational cost associated with the hyperparameter tuning of graph neural network, a fully nested leave-one-station-out cross-validation is impractical. Therefore, three sites are designated as held-out test data to mitigate the overfitting bias on spatial characteristics. The selected stations are displayed in Figure 10. The selection of these sites is driven by two key considerations: first, to manage the risk of spatial autocorrelation, sites for the training set are chosen to be physically distant from those in the test set (Lovelance et al., 2024). Simultaneously, it is vital that these held-out stations reflect distinct and representative land-use characteristics, enhancing the generalizability of results across diverse locations. To achieve this conditioned selection, a k-mean algorithm was utilized to divide the stations into three clusters, while excluding the target value. One station from each cluster was then selected as a test site to ensure the inclusion of varied predictive scenarios. While the remaining risk of spatial autocorrelation cannot be ruled out entirely, the minimal distance between one training and one test site is 2.5 kilometers. This approach ensures the integrity of the test set, by compromising on the amount and diversity of training data. To maximize the value derived from the remaining thirteen training stations, a station-wise leave-one-out cross-validation is employed for hyperparameter tuning. For each iteration of adjustment, a total of thirteen folds are trained and evaluated on the i^{th} excluded station.

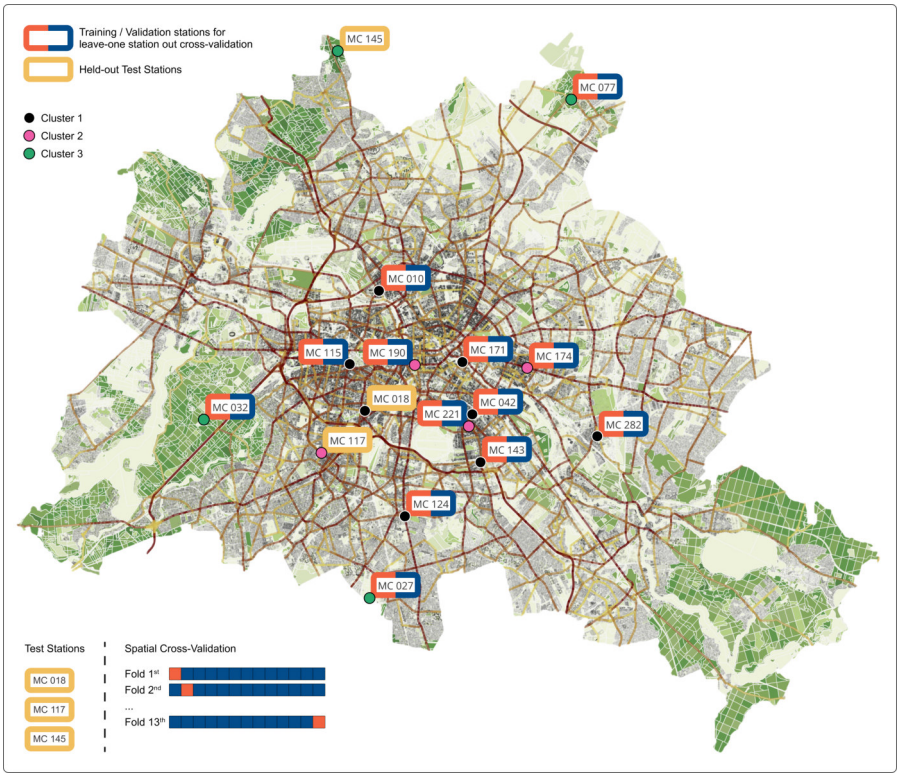

Figure 10: Spatial Leave-One-Station-Out Cross-Validation. The map depicts the held-out stratified test stations, which are selected by k-mean clustering. The remaining 13 stations are used for spatial cross-validation.

4.3.3 *Evaluation Metrics*

To contrast the predictive abilities and robustness of both models, their performances are compared using a variety of selected regression metrics, detailed in Table 5. This selection includes the *mean absolute error* (MAE) and the *root mean squared error* (RMSE), which offer a scale-dependent interpretation and highlight the variation in error distribution through the discrepancies between these two metrics. The squaring operation in RMSE particularly emphasizes the impact of larger residual values, thereby increasing the disparity between MAE and RMSE in the presence of extreme residuals. The *mean absolute percentage error* (MAPE) is a scale-independent error term that can be interpreted intuitively without domain knowledge. It calculates the average percentage deviation from the ground truth. However, due to its sensitivity to near-zero target values, where constant residuals lead to increased relative errors, MAPE can be misleading. This limitation is partly mitigated by its median-based variant, the *median absolute percentage error* (MDAPE). The R^2 score, a well-established metric, is used to evaluate a model's ability to explain the variance in the target value. The interpretability of the R^2 score is not dependent on the scale of the target values, nor is it biased by the presence of near-zero values. In-depth analysis of the performance and robustness of both models is

further enhanced through temporal and feature-wise residual analysis, along with a feature importance analysis, which will be introduced in the following section.

Evaluation Metrics for Regression	
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
Mean Absolute Percentage Error	$MAPE = 100 * \frac{1}{n}\sum_{i=1}^{n}\lvert\frac{y_i - \hat{y}_i}{y_i}\rvert$
R2	$R^2 = \frac{\sum_{i=1}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}(y_i - \overline{y}_i)^2}$

Table 5: Evaluation Metrics for regression performance.

4.3.4 *Feature Importance Analysis*

In order to compare the global reliance of each model on the given features, the feature contribution in the RF is examined through the post-hoc, local analysis of Shapley values, and the feature importance of the GNN is determined by a wrapper-based feature permutation. The construction of Shapley values as a model-agnostic interpretation is not applicable to the graph representation, as detailed in the subsequent section. This results in a compromise on the informative variation in local dependencies on different features between individual observations, provided by Shapley values. For each considered feature and time step, Shapley values are calculated as the difference in prediction between a subset of features, referred to as a coalition, and the same subset including the considered feature. This conditional feature contribution is repeated and averaged over all possible feature coalitions. The feature coalitions are achieved by permuting the not considered features, in order to preserve the shape of the training data. Due to this permutation process, collinear features had to be excluded, as otherwise conflicting combinations of feature values would be possible and could lead to an inflated risk of misprediction (Molnar, 2023).

The application of the Shapley value analysis to the graph structure is not feasible, given that the construction of $2 * 2^k$ graphs for each possible coalition, under k considered features, is required for each time step. Moreover, a sensitivity analysis of the pre-trained graph representation, such as the Shapley value or feature-wise permutation, does not accurately reflect the true performance in the absence of certain features due to the interdependence between nodes. The underlying homophily in graph representations implies that the node-level regression does not solely rely on the attributes of the test node. In contrast, it incorporates the feature-based embedded attributes

of similar nodes (Ma et al., 2023, p. 1). Consequently, the post-hoc sensitivity analysis only partially masks out the contribution of the considered attribute. To address this limitation, a wrapper-based feature-wise permutation is proposed, in which the model is iteratively retrained and then ranked based on the sensitivity of the predictive performance towards the changed feature. Although this methodology overcomes interdependence, required to isolate individual feature contributions, it is limited by the computational expense and the assumption of non-existing feature correlation.

5 RESULTS: PREPROCESSING, MODEL- AND ERROR-ANALYSIS

The Results section is organized into two main parts: data preparation findings and a comprehensive comparison of both models. The former addresses the outcomes of imputation techniques and the identification of the optimal buffer radius through similarity estimations. The Model and Error Analysis includes evaluation metrics, temporal and feature-wise residual analysis, and both global and local feature importance assessments.

5.1 *Imputation Results of Missing Meteorological Data*

Although the single LSTM architecture demonstrates a slight advantage over the hybrid model, both methods adequately capture the general sequential dependency, as depicted in Figure 11. Bidirectional linear interpolation outperforms the LSTM approach as the number of prediction steps increases. Application-oriented limitations and potential architectural enhancements are discussed in Section Assessment of Research Questions. Based on the per-feature evaluation, the standard LSTM was employed to impute the missing values for precipitation, humidity, and solar radiation, while the linear approach was favored for temperature, pressure, and wind speed. For the ordinary scaled wind degree, the *Last Observation Carried Forward* method was applied.

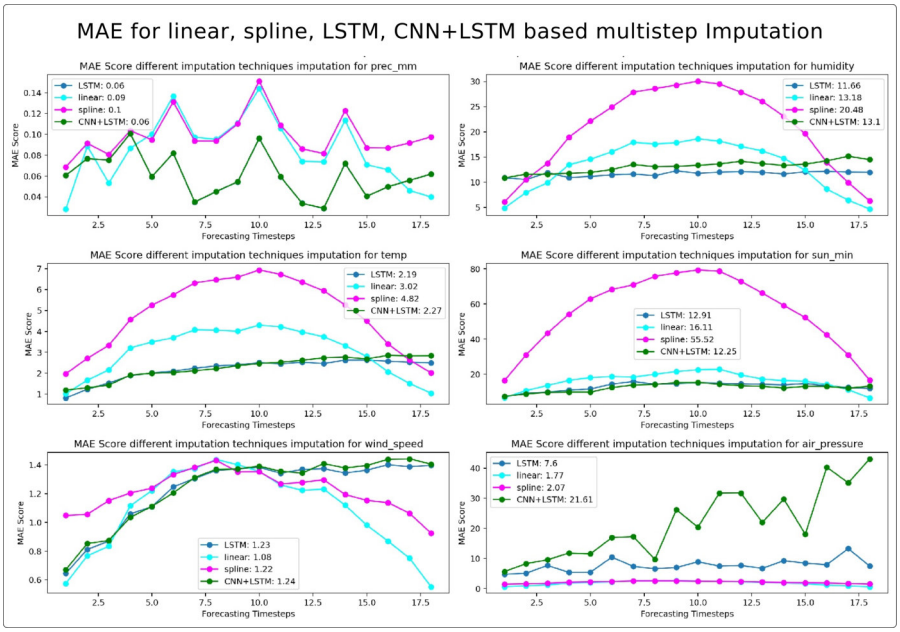

Figure 11: The between-feature comparison illustrates Linear, Spline, LSTM, and CNN+LSTM-based multi-step imputation, highlighting the predictive advantage of sequential modeling for short-range forecasting. In contrast, only the bidirectional linear and spline imputation methods achieve a decreasing error rate in the final imputation steps.

5.2 *Optimizing Buffer Radius: Insights from Mutual Information*

To determine the most informative buffer radius for surrounding vegetation, filter-based feature importance methods were used to assess the relationship between the radius-wise Green Volume Index (GVI) and station-wise average NO_2 concentrations. As indicated in Table 6, no significant differences were observed between the radii, substantiated by *Pearson correlation* and *mutual information* measures. This uniformity stems from the marginal variation in GVI across different radii, with the relative standard deviation ranging from 1.6% to 30%, and six stations recording variations below 10%. The results are further reflected in the Section Assessment of Research Questions.

Similarity estimation for different Buffer Radii					
	GVI 25	GVI 50	GVI 75	GVI 100	GVI 200
correlation	-0.6605	-0.6639	-0.6576	-0.6577	-0.6600
mutual. info.	0.1503	0.1012	0.1058	0.1128	0.1101

Table 6: Pearson correlation and mutual information between average NO_2 concentrations and the Green Volume Index (GVI) across different radii, highlighting minimal variability.

5.3 *Evaluation of NO_2 Prediction using Performance Metrics*

The deviation in ranking of the best-performing model per station, highlighted in bold in Table 7, elucidates the differing emphases of each performance metric and offers insights into the distribution of residuals. The discrepancy between the mean absolute error (MAE) and the root mean squared error (RMSE)—36% for the GNN and 33% for the RF—underscores the significant impact of large errors on overall performance. The trend of increasing residuals with higher target values is depicted in Figures 13, 12 as well as Figure 20 and 21 in Appendix D.1. This pattern highlights the limitation of both models in estimating peak values, an implication further explored in Section Predicting Peak Emission Values.

Performance Evaluation for GNN and RF per station					
	MAE	**RMSE**	**MAPE**	**MDAPE**	R^2
Gnn mc117	5.924	7.772	0.445	**0.282**	**0.654**
Gnn mco18	5.528	7.582	**0.367**	0.36	0.548
Gnn mc145	**2.392**	**3.486**	0.446	0.339	0.637
Gnn mean	4.615	6.28	0.419	0.327	0.613
RF mc117	7.379	9.226	0.596	0.362	0.512
RF mco18	3.768	5.475	**0.302**	**0.232**	**0.764**
RF mc145	**3.012**	**4.19**	0.667	0.418	0.475
RF mean	4.72	6.297	0.522	0.337	0.584
weighted mean	6.693	8.431	0.999	0.762	-0.274

Table 7: Performance metrics for GNN and RF models across test stations. Stations with the best performance per model and metric are highlighted. The variance in performance rankings between absolute (MAE, RMSE) and relative (MAPE, MDAPE, R^2) metrics reflects the scale and distribution of the target values. The weighted average of simultaneous NO_2 levels based on the *difference in distance similarity* serves as baseline performance.

The disparity in performance ranking between absolute and relative metrics underscores the sensitivity of MAPE and MDAPE to minor deviations from small target values, leading to substantial relative errors, as discussed in Section Evaluation Metrics. This sensitivity is particularly evident in the higher relative error of the RF model, caused by the increased overestimation of low values, as observed in Figure 12. The R^2 score comparison further demonstrates the GNN's superior predictive consistency across monitoring sites, which might be due to its utilization of spatial dependencies through the graph embedding. The discrepancy in performance rankings among different metrics underscores their limitations for a comprehensive comparison, influenced by the error distribution and scale of the

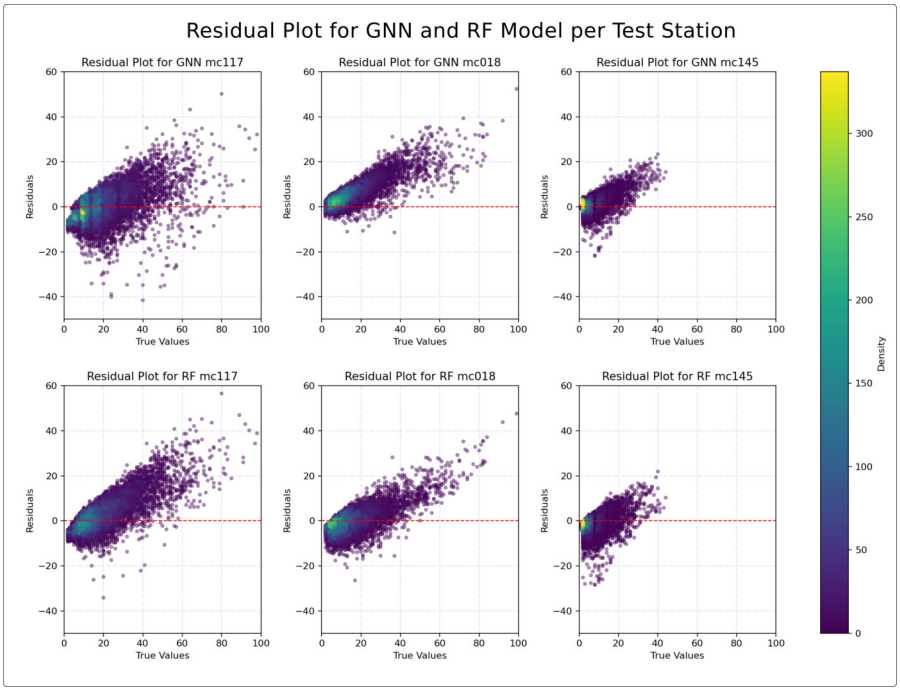

Figure 12: Residual Plot for GNN and RF models across test station. Increasing target values correlate with increasing residual, due to the underprediction of peak emission values. For both models, an increased scattering of the residuals for increasing target values, mainly present at the station mc117 underlines the shortcoming of both models to capture the complex variability related to high NO_2 concentrations. The RF model shows a general trend to overestimate low target values across all stations, which contributes to its higher percentage errors. The smaller range of the target value for station mc145 explains the deviation between performance ranking in the absolute and relative evaluation metrics.

data. Consequently, the predictive performance is further contrasted through detailed feature-wise and temporal residual analyses.

5.4 *Residual Analysis per meteorological Feature*

The comparison of predictive consistency in relation to variations in meteorological features highlights certain predictive limitations. The implications for model improvements are further explored in Section Methodological Shortcomings and further Research. Bivariate scatter plots, displayed in Figures 14 and 15, illustrate correlations but do not establish direct causation. However, they can be analyzed across different stations and models to ascertain whether the unexplained variance arises from the underrepresentation of specific characteristics or a confounding feature. Given that meteorological data is spatially independent, observed discrepancies in residual distribution among stations suggest unexplained spatial variability. This is particularly evident in the consistent overestimation by both models for westside

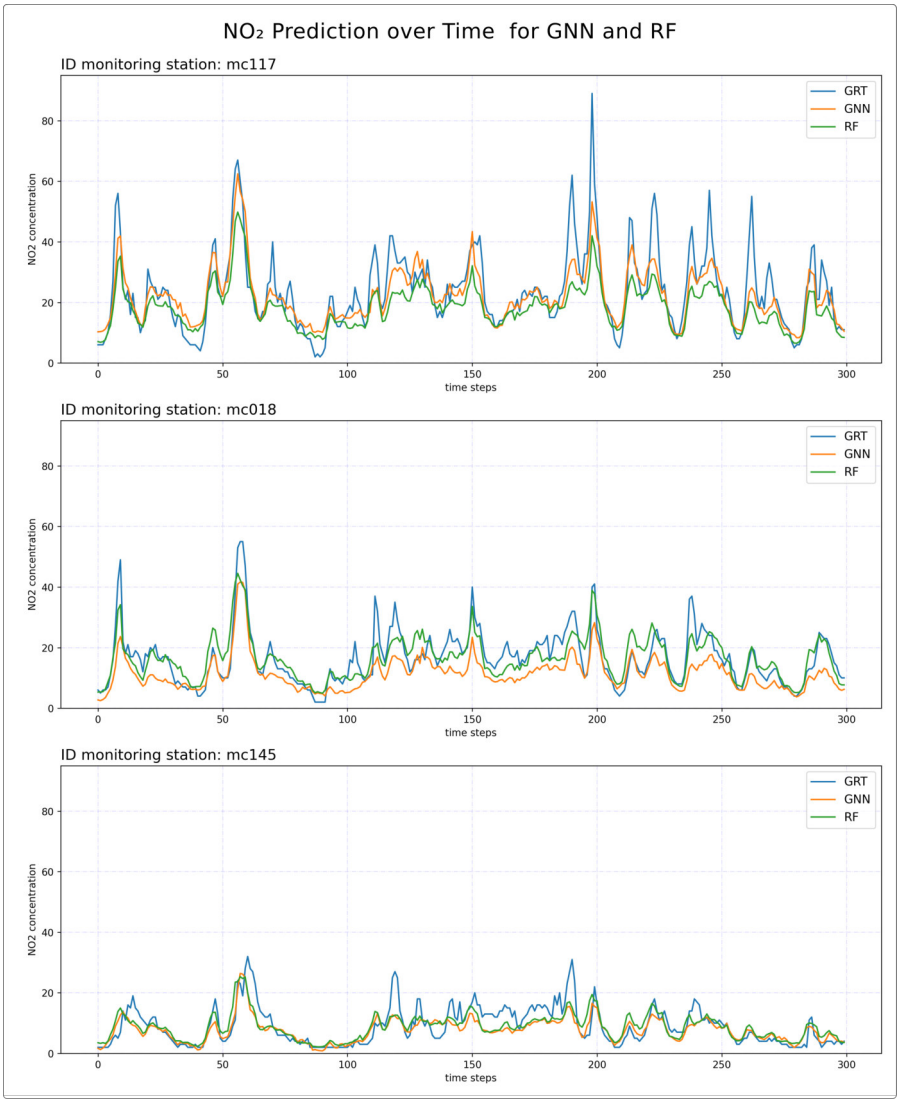

Figure 13: Ground truth (GRT) against predicted values over time period between 1^{st} and 12^{th} of January 2023 at target stations mc117, mc018, mc145.

winds at station mc117, as displayed in Figure 14. Potential factors such as spatial positioning or unaccounted physical barriers, and their methodological implications, are discussed in Section Methodological Shortcomings and further Research. Furthermore, residual analysis concerning wind speed reveals a correlation with the distribution of residuals, marked by a funnel-shaped error distribution as seen in Figure 14 and 15. The possibility of unaccounted NO_2 aerosol accumulation over time, which could explain these observations, is addressed in Section Integration of Temporal Dependency.

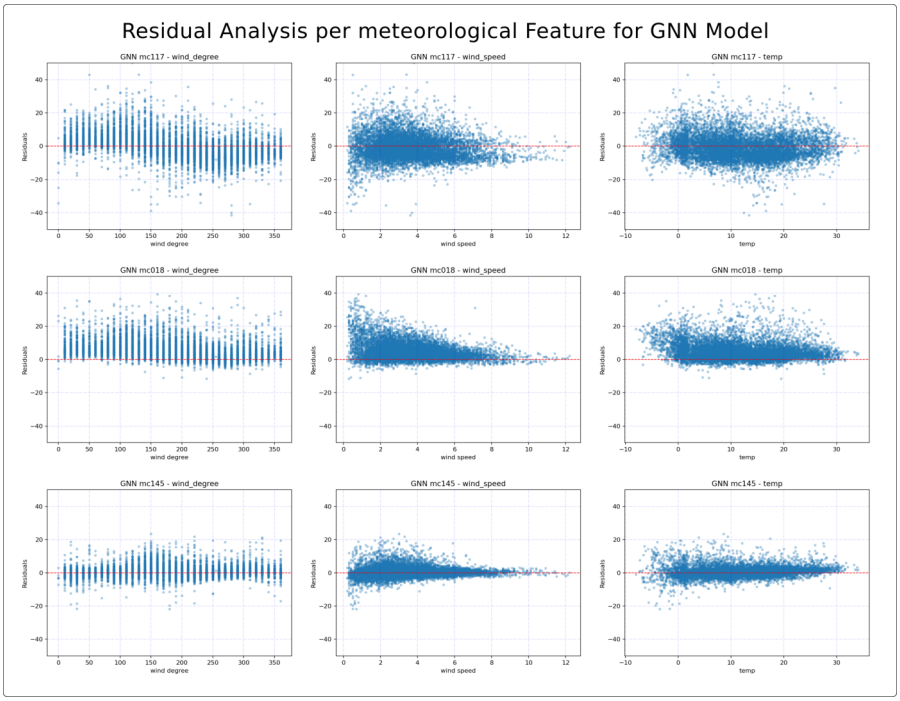

Figure 14: Scatter plot illustrating the correlation between residual variation and changes in meteorological features for the GNN model. While residuals are evenly distributed with temperature changes, variations in wind speed and direction highlight predictive inconsistencies for different conditions.

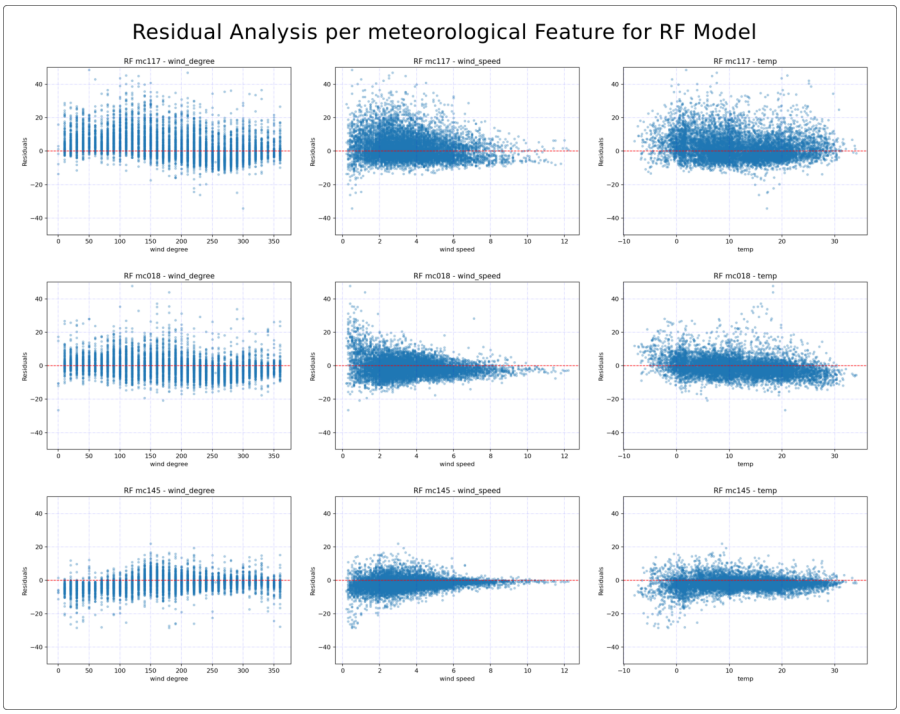

Figure 15: Scatter plot of correlation between variation in residual and changes in different meteorological features for the RF model.

5.5 *Temporal residual Analysis*

Analysis of the residuals across different temporal scales highlights notable seasonal and diurnal variations. Particularly evident is the diminished predictive performance during winter, as indicated by the broader inter-quarter ranges depicted in Figure 16. This decrease in performance during the winter months is not well-documented in existing literature, with only sparse support from previous findings, such as those by Zhan et al. (2018, p. 468), which shows comparable trends. In contrast, the predictive performance during the summer months is characterized by a lower spread and fewer peak values. The daytime variation is not further analyzed, due to the comparable alignment in residual and NO2 distribution.

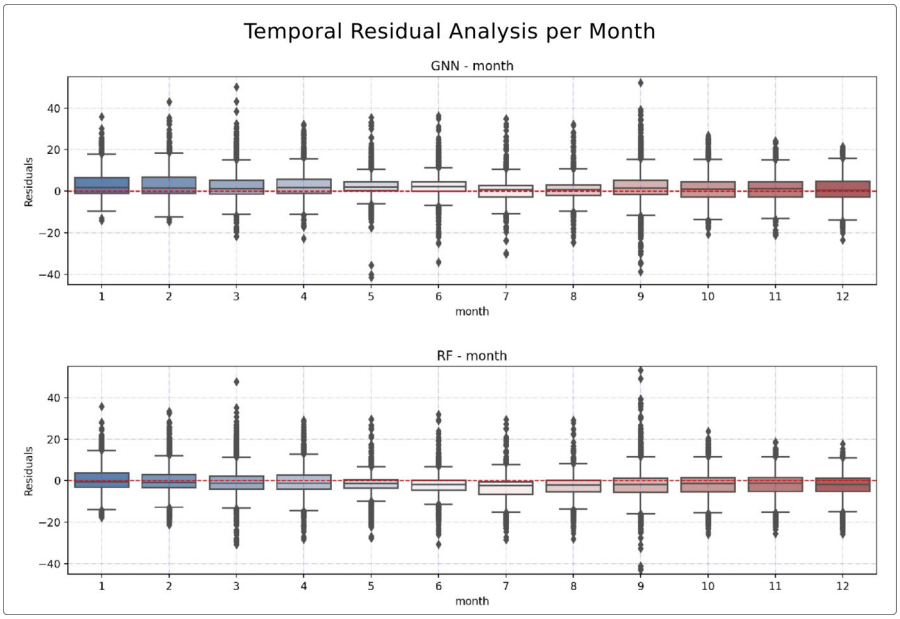

Figure 16: Distribution of residuals over all months for GNN and RF model, displayed as box plop, with a notable decrease in the spread of residuals in the summer months (May till August).

5.6 *Global Feature Importance*

The aggregated local feature contributions, quantified as Shapley values, provide a global interpretation of feature importance for the Random Forest model, as illustrated in Figure 17. The *weighted_mean_pollution*, representing the weighted average NO_2 concentration, emerges as the most influential feature. This metric encapsulates the temporal variation of all emission sources, including meteorological factors that affect decomposition. Consequently, meteorological variations are associated with a subordinate contribution, despite their significance for aerosol decomposition. Additionally, emitter-related predictors,

such as the average traffic volume within a 200-meter radius (*tvi_200*) and the proportion of heavy vehicles (*prop_main_*), are identified as having substantial contributions.

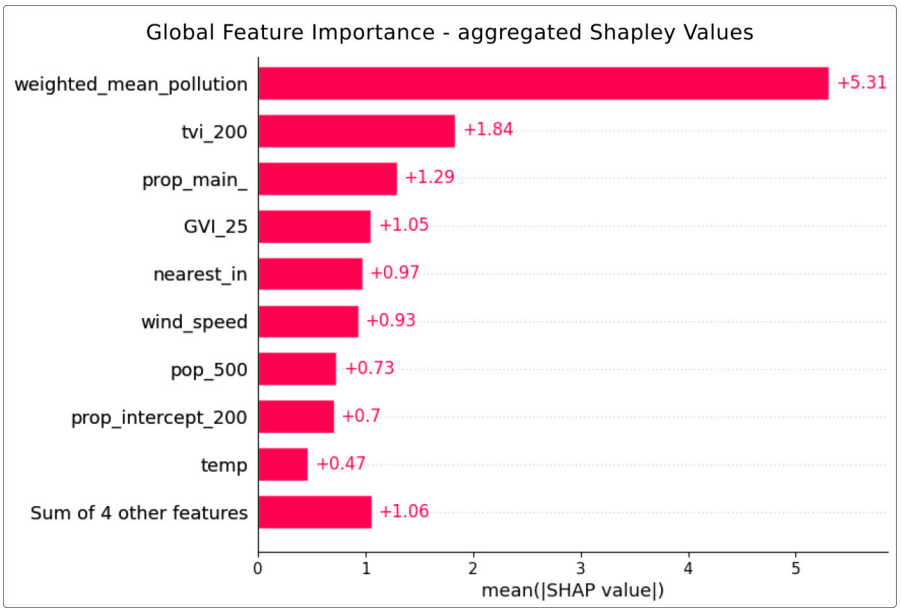

Figure 17: Global Feature Contribution based on local Shapley value for RF.

The graph embedding does not integrate the weighted average pollution as an additional variable; instead, it adjusts the attention coefficients in dependence on the land-use and meteorological features. Consequently, the dependence of the model on the target variable for the graph embedding is crucial and cannot be isolated through the chosen approach. Nevertheless, the observed deviation in the R^2-Score, as shown in Figure 18 following the permutation of the considered feature, underscores the graph representation's capability to extract informative value from these predictors, particularly towards meteorological features. The minimal contribution of spatial features suggests an alignment among these characteristics, wherein the absence of one feature is compensated by the presence of others.

5.7 *Localized Contribution Analysis via Shapley Values*

To assess the relationship between the changing Leaf Area Index (LAI) and NO_2 mitigation, this study analyzes the variation in local contributions to NO_2 concentration estimates in response to changing vegetation, as depicted in Figure 19. The left-side plot presents the variation under changing temperatures, used as a proxy for LAI due to their high correlation, resulting in a more gradient and uniformly distributed depiction. Both plots reveal a clear seasonal trend of increasing negative contribution in the NO_2 estimation, correlating with higher temperatures and greater LAI. Importantly, the use of

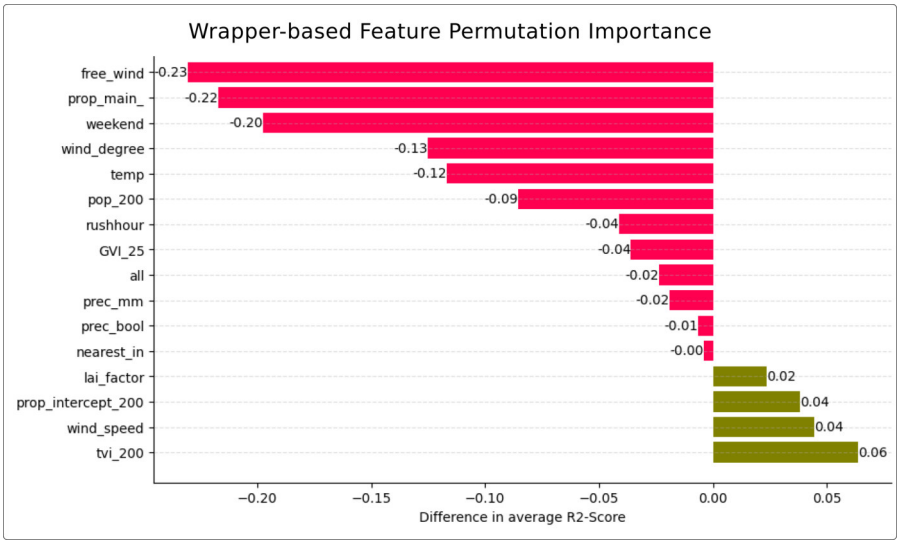

Figure 18: Global Feature Contribution from wrapper-based permutation importance for GNN. Divergence in R^2-Regression performance under the permutation of a considered feature outlines the feature contribution.

site-specific green volume as a color variable enables the isolation of greenery's contribution by highlighting differences across test stations. The Shapley values indicate that seasonal variations are more pronounced at the station with higher green volume, which exhibits more than twice the negative contribution compared to stations with less greenery during the summer months. Furthermore, the general U-shaped trend observed in both plots of Figure 19 indicate that the greatest negative contribution of green volume on NO2 estimation is associated with mid-range temperature and LAI. The subsequent Section Assessment of Research Questions will discuss these findings within the context of existing literature and reflect on the methodological limitations.

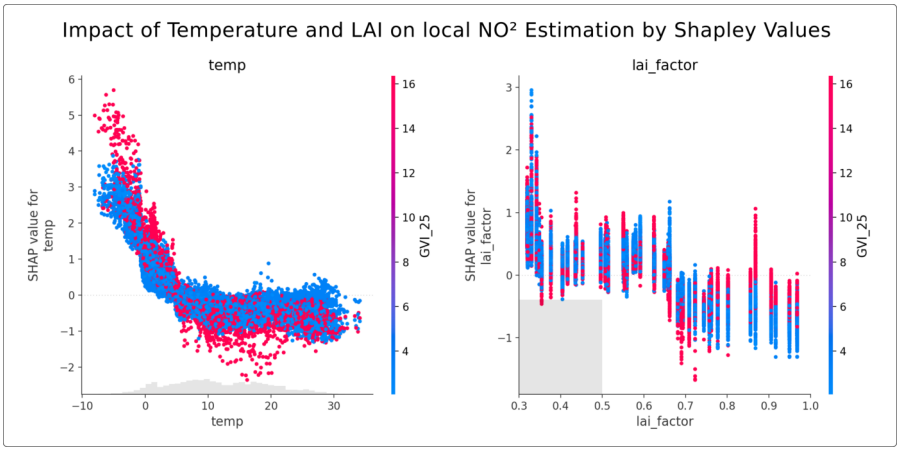

Figure 19: Variation in conditional Green Volume Index (GVI) contribution under changing temperature and leaf area index (LAI).

6 DISCUSSION

6.1 *Assessment of Research Questions*

RQ1: *To what extent can the emission level of NO2 at monitoring sites in Berlin be predicted by urban greenery, meteorological, spatial, and traffic factors?*

Both models leverage the majority of unexplained variation and achieve a similarly predictive performance with R2- Scores of approximately 0.6 as shown in Table 7. Nevertheless, the achieved results underperform many prior approaches to air pollution estimation (Rybarczyk & Zalakeviciute, 2018, p. 22; Turek & Kamińska, 2022, p. 1). This discrepancy can be partly attributed to the higher heterogeneity inherent in spatial estimation compared to forecasting-based approaches, which typically yield higher performances. Moreover, the reliance on a distance-based similarity weighting that assumes a linear gradient in land-use change radiating from a singular, universal centroid does not accurately reflect the complex spatial dynamics. Additionally, the sparse distribution of the sixteen stations contrasts with more densely sampled datasets from prior graph-based approaches (H. Wang, 2019).

In a direct comparison, the GNN model demonstrates superior performance over the RF model, evidenced by the reduction in the between-test-site variation of predictive accuracy, with standard deviations of 0.06 compared to 0.16 for the RF, as presented in Table 7. Nevertheless, due to the complexity and homophily of the graph structure, only the RF model allows for local interpretability through Shapley values. Analyzing the predictive error of both models reveals their limitation in predicting peak values, which is found to pose a significant challenge throughout the literature (Rybarczyk & Zalakeviciute, 2018, p. 23). These considerations are elaborated upon in Section Predicting Peak Emission Values.

The residual analysis underscores the influence of seasonal and meteorological variations on predictive performance, with increased errors observed during winter months and under conditions of low wind. The analyzed feature importance underscores the capacity of both models to extract informative value from the chosen predictors, with wind-related attributes and temperature emerging as the most descriptive meteorological features, aligning with prior findings (Selmi et al., 2016, p. 196; Gocheva-Ilieva & Livieris, 2020, p. 54; Rybarczyk & Zalakeviciute, 2018, p. 10). The identified contribution of the proportion of heavy vehicles (*prop_main_*) correlates with the significant increase in NO_2 emission for diesel-powered trucks or buses (Lighterink, 2017, p. 12). The proposed mitigation for multicollinearity using feature selection maintained the ability of model agnostics

but reduced the informative value of additional meteorological and land-use features, which could have been retained through principal component analysis. Further architectural improvement is discussed in the Section Methodological Shortcomings and further Research.

SQ1: *To what extent can univariate multistep forecasting be used as an imputation technique for meteorological features, in contrast to linear and spline interpolation?*

The results of the LSTM-based univariate, multi-step forecasting approach underscore the efficacy of recurrent-based imputation techniques, particularly for short-range forecasting. The accuracy of both models declines with increasing forecast steps, suggesting that a bidirectional modification might enhance performance. Additionally, considering the interconnectivity among various meteorological phenomena, a multivariate approach could yield more comprehensive insights. This methodology primarily utilizes training data from the previous year, assuming the generalizability of meteorological trends.

SQ2: *To what extent can the optimal buffer radius from a sensing site to surrounding vegetation be determined through entropy- and correlation-based similarity estimations?*

The data set, comprising local greenery surrounding monitoring sites in Berlin, shows no significant variation across different radii and suffers from a limited sample size of only sixteen stations. These factors compromise the reliability of the chosen similarity estimations and hinder conclusive answers to the sub-question based on the current data set.

SQ3: *To what extent can Shapley values, as local, model agnostic interpretation, be utilized to analyze variations in the mitigation of NO_2 concentrations relative to seasonal changes in the Leaf Area Index?*

The observed variation in the negative contribution of greenery to local NO_2 estimation supports the hypothesis of seasonal variation, aligning with previous findings (Escobedo & Nowak, 2009, p. 105). However, due to the limited number of stations with static land-use characteristics (N=3), these observations of high GVI are confounded by less populated, low-traffic areas. Consequently, this correlation complicates the attribution of NO_2 mitigation solely to green volume. To investigate the intricate relationship influenced by changing Leaf Area Index (LAI), future studies should consider controlled experimental designs, expand the sample size, or incorporate advanced statistical methods such as bootstrapping. The interpretation is further

complicated by the substantial total number of observations (N = 26,217), which diminishes the significance of outliers in both plots.

Despite data limitations and methodological constraints, it is concluded that temperature should be employed as a proxy for LAI to avoid inducing collinearity due to their high correlation, which can be seen in Table 1. Moreover, the unexpected correlation between high LAI and less effective NO2 mitigation in low GVI areas, suggests the presence of additional complexities.

6.2 *Methodological Shortcomings and further Research*

6.2.1 *Predicting Peak Emission Values*

The current limitation of both models to underestimate peak values is of particular concern in the context of application-oriented early warning systems targeting harmful concentrations. The predictive limitations arise partly from the underrepresentation of peak values within the training data, and align with previous approaches to air pollution estimation (Rybarczyk & Zalakeviciute, 2018, p. 23). Tamas et al. (2016) proposed an unsupervised pre-training clustering of the data, which creates three training sets for separate models to address the data imbalance. An alternative approach could involve the oversampling of peak instances through augmentation.

In the context of the current studies, architectural modifications such as adjustments to the loss function and scaling of the target value were explored during the hyperparameter optimization but did not demonstrate an increase in performance.

6.2.2 *Integration of Temporal Dependency*

The temporal dependency of the data has been neglected in the current study, since only the target and meteorological data hold temporal variation. Nevertheless, the increased predictive error for low wind conditions and peak values could be addressed through the encoded representation of preceding time steps. Integrating a recurrent unit that quantifies the sequential variation of meteorological data could indicate the persistence of low wind, facilitating the accumulation of concentrations (Nazridoust & Ahmadi, 2006, p. 3). Furthermore, including the sequential development of NO_2 concentration as a temporal lag feature could enhance predictions of peak values, assuming the presence of a preceding temporal trend.

6.2.3 *Revision of Spatial Dependency*

The absence of significant performance enhancement between the models may stem from the graph representation's constrained ability to effectively embed spatial relationships. In consideration of the correlation of westside wind and the underestimation for the test station located in the city west, displayed in Figure 14, the spatial positioning relative to Berlin's center could be integrated into the graph representation. The hypothesis—that emissions from the city center, when carried towards these sites by prevailing winds, increase NO_2 concentrations—could be empirically tested by examining changes in NO_2 levels in response to wind directions aligning with site positions relative to the city center. Architectural adaptations that consider wind conditions have been explored in directional graph representations, such as those proposed by Xiao et al. (2022) and S. Wang et al. (2020), which selectively incorporate sites based on current wind direction alignment.

Alternatively, integrating spatial positioning to include the locations of industrial plants, which contribute approximately one-third of Berlin's NO_2 emissions (Environmental-Atlas-Berlin, 2021), could enhance the model's accuracy. This approach aligns with strategies employed in other studies, such as the integration of localized *point-of-interest* (POI) features, which consider industrial sites (Xu et al., 2021).

6.2.4 *Adaptations of Feature Engineering*

The variation in residuals across different wind directions, combined with the minimal associated feature relevance of obstacle-related features, contrasts with existing literature on aerosol captivity, suggesting deficiencies in the proposed encoding methodology (Nazridoust & Ahmadi, 2006, p. 20). This feature engineering was limited to an abstraction that captures tendencies of surrounding architecture but that cannot account for the complex mechanics of fluid dynamics resulting, such as swirling effects (Nazridoust & Ahmadi, 2006, p. 20). Moreover, the simplification inherent in restricting the analysis to a 200-meter radius and excluding other significant obstacles like trees further reduces the complexity of the engineered features.

7 CONCLUSION

This research highlights the complexity of interactions between the environmental features that determine the local decomposition of NO_2. A quantification of the individual contribution of different land-use characteristics, which could guide urban planning in pollution mitigation, is further impeded by the typical collinearity among these features. While the incorporation of spatial dependencies may enhance model capabilities, it does not necessarily deepen the understanding of the underlying interactions. Despite these interpretable restrictions, the results allow for the support of existing findings such as the significance of the vehicle type for NO_2 concentration or the seasonal variation in the ability of vegetation to mitigate pollution. Additionally, the discovered marginal decrease in the contribution of urban vegetation to NO_2 mitigation under peak temperatures encourages further research and intersection with existing domain knowledge. As novel methodologies continue to evolve, the pursuit of understanding and managing these complex interactions remains a dynamic field, crucial for fostering healthier and more sustainable urban environments.

REFERENCES

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf (cit. on p. 25).

Berliner-Luftgütemessnetz. (2023). *Berliner luftgütemessnetz.* http://luftdaten.berlin.de (cit. on p. 13).

Brody, S., Alon, U., & Yahav, E. (2021). How attentive are graph attention networks? *CoRR*, *abs/2105.14491*. https://arxiv.org/abs/2105.14491 (cit. on p. 51).

Brokamp, C., Jandarov, R., Rao, M., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment*, *151*, 1–11. https://doi.org/https://doi.org/10.1016/j.atmosenv.2016.11.066 (cit. on pp. 7, 9, 15, 21).

Brownlee, J. (2020). *Multi-step time series forecasting with machine learning for electricity usage*. https://machinelearningmastery.com/multi-step-time-series-forecasting-with-machine-learning-models-for-household-electricity-consumption/ (cit. on pp. 3, 19).

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, *10*(8). https://doi.org/10.3390/math10081283 (cit. on pp. 19, 20).

EEA. (2023). *Air quality in europe 2022*. https://www.eea.europa.eu/publications/air-quality-in-europe-2022 (cit. on p. 6).

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., & Cirach, M. e. a. (2012). Development of land use regression models for pm2.5, pm2.5 absorbance, pm10 and pmcoarse in 20 european study areas. *Environmental Science Technology*, *46*(20), 11195–11205. https://doi.org/https://doi.org/10.1021/es301948k (cit. on p. 7).

Environmental-Atlas-Berlin. (2021). *Long-term development of air quality 2021*. https://www.berlin.de/umweltatlas/en/air/development-of-air-quality/since-1989/map-description (cit. on p. 41).

Escobedo, F. J., & Nowak, D. J. (2009). Spatial heterogeneity and air pollution removal by an urban forest. *Landscape and Urban Planning*, *90*(3), 102–110. https://doi.org/https://doi.org/10.1016/j.landurbplan.2008.10.021 (cit. on pp. 8, 39).

Gocheva-Ilieva, A. V., Snezhana G.and Ivanov, & Livieris, I. E. (2020). High performance machine learning models of large scale air pollution data in urban area. *Cybernetics and information technologies*, *20*(6). https://doi.org/https://doi.org/10.2478/cait-2020-0060 (cit. on pp. 10, 18, 21, 38).

Huang, C. J., & Kuo, P. H. (2018). A deep cnn-lstm model for particulate matter (pm2.5) forecasting in smart cities. *Sensors*, *18*(7). https://www.mdpi.com/1424-8220/18/7/2220 (cit. on p. 10).

Huang, Y., Ying, J. J.-C., & Tseng, V. S. (2021). Spatio-attention embedded recurrent neural network for air quality prediction. *Knowledge-Based Systems*, *233*, 107416. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107416 (cit. on p. 11).

Iskandaryan, D., Ramos, J. F., & Trilles Oliver, S. (2023). Graph neural network for air quality prediction: A case study in madrid. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2023.3234214 (cit. on pp. 6, 10, 11, 13).

Jain, S., Kaur, N., Verma, S. K., Hosen, A. S. M. S., & Sehgal, S. S. (2022). Use of machine learning in air pollution research: A bibliographic perspective. *Electronics*, *11*(21), 111–128. https://doi.org/https://doi.org/10.3390/electronics11213621 (cit. on p. 9).

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *CoRR*, *abs/1609.02907*. http://arxiv.org/abs/1609.02907 (cit. on pp. 23, 49).

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, *69*, 066138. https://doi.org/10.1103/PhysRevE.69.066138 (cit. on p. 15).

Larkin, A., Anenberg, S., Goldberg, D. L., Mohegh, A., Brauer, M., & Hystad, P. (2023). A global spatial-temporal land use regression model for nitrogen dioxide air pollution. *Frontiers in Environmental Science*, *11*. https://doi.org/https://doi.org/10.3389/fenvs.2023.1125979 (cit. on p. 9).

Lighterink, N. (2017). Real-word vehicle emissions. *Strategies for Mitigating Air Pollution in Mexico City, Discussion Paper No. 2017-06*. https://www.itf-oecd.org/sites/default/files/docs/real-word-vehicle-emisions.pdf (cit. on pp. 15, 38).

Lovelance, R., Nowosad, J., & Muenchow, J. (2024). – chapter 12: Statistical learning. introduction to (spatial) cross – validation. In *Geocomputation with r*. Curran Associates, Inc. https://r.geocompx.org/spatial-cv (cit. on p. 25).

LUP, L. U. P. (2021). Bestimmung von vegetationshöhen in berlin – aktualisierung 2020. https://www.berlin.de/umweltatlas/biotope/vegetationshoehen/2020/methode/ (cit. on p. 16).

Ma, Y., Liu, X., Shah, N., & Tang, J. (2023). Is homophily a necessity for graph neural networks? https://doi.org/https://doi.org/10.48550/arXiv.2106.06134 (cit. on p. 28).

Molnar, C. (2023). 8: Global model-agnostic methods. In *Interpretable machine learning- a guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/feature-importance.html (cit. on pp. 19, 27).

Nazridoust, K., & Ahmadi, G. (2006). Airflow and pollutant transport in street canyons. *Journal of Wind Engineering and Industrial Aerodynamics*, *94*, 491–522. https://doi.org/10.1016/j.jweia.2006.01.012 (cit. on pp. 17, 18, 40, 41).

Opolka, F. L., Solomon, A., Cangea, C., Velickovic, P., Liò, P., & Hjelm, R. D. (2019). Spatio-temporal deep graph infomax. *CoRR*, *abs/1904.06316*. http://arxiv.org/abs/1904.06316 (cit. on pp. 23, 50).

Pan, S., Hu, R., Fung, S., Long, G., Jiang, J., & Zhang, C. (2019). Learning graph embedding with adversarial training methods. *CoRR*, *abs/1901.01250*. http://arxiv.org/abs/1901.01250 (cit. on p. 50).

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12). https://www.mdpi.com/2076-3417/8/12/2570 (cit. on pp. 7, 9, 18, 21, 38, 40).

Schümann, L., Grunow, K., & Kaupp, H. (2021). Luftverunreinigungen in berlin monatsbericht januar 2023. https://www.berlin.de/sen/uvk/_assets/umwelt/luft/luftqualitaet/luftdaten-archiv/monats-und-jahresberichte/januar2021.pdf (cit. on p. 13).

Selmi, W., Weber, C., Rivière, E., Blond, N., Mehdi, L., & Nowak, D. (2016). Air pollution removal by trees in public green spaces in strasbourg city, france. *Urban Forestry Urban Greening*, *17*, 192–201. https://doi.org/https://doi.org/10.1016/j.ufug.2016.04.010 (cit. on pp. 9, 18, 38).

Seng, D., Zhang, Q., Zhang, X., Chen, G., & Chen, X. (2021). Spatiotemporal prediction of air quality based on lstm neural network. *Alexandria Engineering Journal*, *60*(2), 2021–2032. https://doi.org/https://doi.org/10.1016/j.aej.2020.12.009 (cit. on p. 11).

SenMVKU. (2020). Sraßenverkehrszählung berlin teil a – ergebnisbereicht – verkehrsmengenkarte dtv kfz/ lkw 2019. *Senate Department for Mobility, Transport, Climate Protection and the Environment*. https://www.berlin.de/sen/uvk/_assets/verkehr/verkehrsmanagement/verkehrserhebungen/ergebnisbericht-2019-teil-a.pdf (cit. on p. 15).

Setälä, H., Viippola, V., Rantalainen, A.-L., Pennanen, A., & Yli-Pelkonen, V. (2013). Does urban vegetation mitigate air pollution in northern conditions? [Selected Papers from Urban Environmental Pollution 2012]. *Environmental Pollution*, *183*, 104–112. https:

//doi.org/https://doi.org/10.1016/j.envpol.2012.11.010 (cit. on p. 8).

Shams, S. R., Jahani, A., Kalantary, S., Moeinaddini, M., & Khorasani, N. (2021). Artificial intelligence accuracy assessment in - no2 concentration forecasting of metropolises air. *Scientific Reports*, *11*. https://doi.org/https://doi.org/10.1038/s41598-021-81455-6 (cit. on pp. 9, 10).

Silva, R. A., West, J. J., Lamarque, J.-F., Shindell, D. T., & Collins, W. J. e. a. (2017). Future global mortality from changes in air pollution attributable to climate change. *Nature Climate Change*, *7*, 647–651. https://doi.org/https://doi.org/10.1038/nclimate3354 (cit. on p. 6).

Srbinovska, M., Andova, V., Mateska, A. K., & Krstevska, M. C. (2021). The effect of small green walls on reduction of particulate matter concentration in open areas. *Journal of Cleaner Production*, *279*, 123306. https://doi.org/https://doi.org/10.1016/j.jclepro.2020.123306 (cit. on p. 9).

Tamas, W., Notton, G., Paoli, C., Nivet, M.-L., & Voyant, C. (2016). Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. *Aerosol and Air Quality Research*, *16*(2), 405–416. https://doi.org/10.4209/aaqr.2015.03.0193 (cit. on p. 40).

Turek, T., & Kamińska, J. A. (2022). A comparative study of using random forests (rf), extreme learning machine (elm) and deep learning (dl) algorithms in modelling roadside particulate matter (pm10 pm2.5). (cit. on pp. 9, 38).

Veaux, R. D. D., & Ungar, L. H. (1994). Multicollinearity: A tale of two nonparametric regressions. https://api.semanticscholar.org/CorpusID:14684655 (cit. on p. 19).

Vu, V., Nguyen, D., Nguyen, T., Nguyen, Q., P.L., N., & Huynh, T. (2024). Self-supervised air quality estimation with graph neural network assistance and attention enhancement. *Neural Computing and Applications*. https://doi.org/https://doi.org/10.1007/s00521-024-09637-7 (cit. on pp. 3, 6, 10, 11, 13, 22, 23, 50, 51).

Wang, H. (2019). *Air pollution and meteorological data in Beijing 2017-2018*. https://doi.org/10.7910/DVN/USXCAK (cit. on pp. 11, 38).

Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., & Gao, F. (2020). Pm2.5-gnn: A domain knowledge enhanced graph neural network for pm2.5 forecasting. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/3397536.3422208 (cit. on pp. 10, 41).

Xiao, X., Zhiling, J., Wang, S., Xu, J., Peng, Z., Wang, R., Shao, W., & Hui, Y. (2022). A dual-path dynamic directed graph convolutional network for air quality prediction. *Science of The Total*

Environment, *827*, 154298. https://doi.org/10.1016/j.scitotenv.2022.154298 (cit. on pp. 10, 41).

Xu, J., Chen, L., Lv, M., Zhan, C., Chen, S., & Chang, J. (2021). Highair: A hierarchical graph neural network-based air quality forecasting method. *CoRR*, *abs/2101.04264*. https://arxiv.org/abs/2101.04264 (cit. on p. 41).

Yan, K., Wang, J., Peng, R., Yang, K., Chen, X., Yin, G., Dong, J., Weiss, M., Pu, J., & Myneni, R. B. (2024). Hiq-lai: A high-quality reprocessed modis leaf area index dataset with better spatiotemporal consistency from 2000 to 2022. *Earth System Science Data*, *16*(3), 1601–1622. https://doi.org/10.5194/essd-16-1601-2024 (cit. on pp. 2, 12, 17).

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., & Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across china using random forest for human exposure assessment. *Environmental Pollution*, *233*, 464–473. https://doi.org/https://doi.org/10.1016/j.envpol.2017.10.029 (cit. on pp. 3, 10, 35).

A APPENDIX: PACKAGES AND LIBRARIES

The following Packages have been used with pythonVersion: 3.11.3:

- pandas- Version: 1.5.3
- geopandas- Version: 0.14.1
- Shapely- Version: 2.0.2
- numpy- Version: 1.26.0
- matplotlib- Version: 3.7.1
- rioxarray- Version: 0.15.0
- sklearn- Version: 1.2.2
- seaborn- Version: 0.12.2
- keras- Version: 2.12.0
- requests- Version: 2.28.2
- shap- Version: 0.45.1
- statsmodels- Version: 0.14.1
- torch- Version: 2.0.1+cpu

B APPENDIX: PREPROCESSING

B.1 *Missing Data Imputation with LSTM-Model and CNN-LSTM-Model*

For the LSTM models, the previous year's training data is reshaped into 454 samples with 24 time steps. The first model consists of a 200-unit LSTM layer for learning representations of long- and short-term dependencies within the training sequences, before being processed from a 100-unit dense layer. The single output unit, with a linear activation function as well as a mean absolute error as loss, allows for regression. All previous layers are also constructed with a ReLu activation function. The second hybrid architecture combines additional convolutional layers for spatial feature extraction with the LSTM layer for temporal sequence analysis. The initial convolutional layer of 64 and 124 filters is transformed with three-by-three kernels and the default of single stride and no applied padding before applying max pooling and flattening for dimensionality reduction. A RepeatVector layer then transforms the flattened data into a sequential structure, which is then used by the LSTM layer. Both architectures undergo minor hyperparameter tuning in terms of model capacity, a chosen number of batches and epochs, and different activation functions.

C APPENDIX: METHODOLOGY

C.1 *Neighborhood Aggregation: Graph Convolutional Network*

In order to enhance the informative value of the graph representation, all 8759 unique graphs, one for each time step, are transformed through graph convolutions, initially proposed by Kipf et al. (2017). Through this process, an abstract and thereby more generalizable representation of the spatial relations can be aggregated. The term convolution reflects the similarity in both compositional assumptions and initial aspirations compared to kernel transformations in convolutional representation learning for image processing. In both instances, the transformation is based on the assumption that a local reception field is informative, that locally extracted patterns can be transferred to other locations, and that hierarchical abstractions can be extracted.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{3}$$

While a kernel transformation in the convolutional layer in image processing considers the natural neighborhood determined by surrounding pixels in the imagery, the nodes that are considered for the aggregation of neighboring node information in graph convolutions H are determined by the edge relations and their strength defined in the adjacency matrix A (Kipf & Welling, 2016, p. 2). Within a non-fully connected graph, only the connected nodes contribute to the aggregation of attributes for the considered node. In the applied dense graph representation for monitoring stations, the attributes from all other nodes are considered, but their influence is weighted in accordance with the adjacency matrix, building upon the difference in distance similarity (Kipf & Welling, 2016, p. 2). As in convolutional layers of image processing, the centroid pixel value or the centroid node for graph representation contributes to the aggregation, which is referred to as self-connections and denoted with $\tilde{A}$ (Kipf & Welling, 2016, p. 2). This extended adjacency matrix is normalized through the inverted square root of $\tilde{D}$, symbolizing the row-wise summation of $\tilde{A}$ as displayed in Formula 3. An illustration of the weighted aggregation is, graphically represented in Figure 9. Another similarity to convolutional layers is the shared and trainable matrix A used to transform each aggregated attribute vector denoted as *fixed local spectral filters* for the convolution layer l, which not only fosters the training of generalizable transformations but also reduces the computational cost (Kipf & Welling, 2016, p. 2). After the convolutional graph transformation 8759 new graphs, are further processed in the graph embedding.

C.2 *Graph Embedding: Spatio Deep Graph Infomax*

The temporal graph convolution representation is additionally embedded through a contrastive learning methodology, that utilizes an adversarial generator-discriminator architecture and is inspired by the *Spatio-temporal Deep Graph Infomax* of Opolka et al. (2019). Graph embedding not only increases the abstraction and thereby generalizability of learned latent space, but further "transfers [the] graph data into a low dimensional, compact, and continuous feature space" (Pan et al., 2019, p. 1), which can be further processed by the subsequent regression algorithms. Therefore, a self-supervised adversarial graph embedding approach is implemented, which is constituted of an encoder that trains the embedding as well as a discriminator challenging this embedding. The objective of adversarial learning is to maximize the mutual information between the embedding and the previously extracted Graph Convolutional Representation by utilizing a Deep Infomax (Vu et al., 2024, p. 9). As initialization each node-attribute matrix $x_1, ..., x_n$ of each graph, that entails the information gained by the GCN is paired with a trainable embedding of this matrix $h_1, ..., h_n$ and are referred to as *positive pair*. Additionally, each of the node embeddings $h_1, ..., h_n$ is part of a negative pair with a corrupted version of the node-attributes as counterpart denotes as $\tilde{x}_1, ..., \tilde{x}_n$ The introduced diversity of the manipulated graph encourages the training of generalizable and thereby robust embedding vectors (Vu et al., 2024, p. 9). The corruption of the original GCN data structure is introduced through a randomized interchange of feature attributes on the node dimension as well as an augmentation of this data array through uniformly distributed, but random rescaling (Vu et al., 2024, p. 9).

The discriminator D returns the probability of each pair being a *positive pair* and is reciprocally trained by self-supervised loss L_{ssl} calculated by the combined sum of the Discriminators prediction as log wherein N represents the number of two pairs (Vu et al., 2024, p. 9). Within this framework, the mutual information between embedding and attribute vector is maximized when the discriminator is best able to distinguish positive from corrupted pairs since this implies that the embeddings contain substantial information about the original data while being distinctive from others. The Discriminator is utilized through a shared weight matrix W and a sigmoid activation function $\sigma()$.

$$D(x_i, h_i) = \sigma(h_i W x_i) \tag{4}$$

$$D(x_i, h_i) = \sigma(h_i W \tilde{x}_i) \tag{5}$$

$$L_{ssl} = \frac{1}{2N}(\sum_{i=1}^{N}[\log(D(x_i, h_i)] + \sum_{i=1}^{N}[\log(D(\tilde{x}_i, h_i)]) \tag{6}$$

C.3 *Location and Feature aware attention Mechanism*

The encoder architecture is concluded with two separate graph attention mechanisms, which selectively attend to locations with similarities in the latent embedding space based on both the land-use and the temporal characteristics at the target location and moment. To initialize the location-aware attention, an interpolated embedding vector h for the target station is calculated and combined with the location-specific land-use features of this new station. This new interpolated embedding vector h for the target location is based on the weighted sum derived by the previously discussed similarity ranking. The location-aware attention creates a finite and thereby static scoring function that describes the relationship between the concatenation of the interpolated embedding h_{target} and the corresponding feature matrix as query and the other embeddings at the current time step as key-value pair (Brody et al., 2021, p. 4). Consequently, the self-attention coefficient $e_{i;target}$, which describes the inter-station relation towards the target embedding can be depicted as $a(Wh_i * U[h_{target}|x_{target}])$, with learnable weight matrix W and U, wherein a denotes the normalization of the attention coefficients through *SoftMax*. The feature-aware attention block generates a selective context vector of feature attributes and thereby "reduc[ing] the impact of irrelevant components and emphasiz[ing] those [...] highly correlated" (Vu et al., 2024, p. 10). Therefore, the query remains identical to that of the other attention block, while the key/ value pair consists of the single latent features stored in the previously learned embedding vectors h.

D APPENDIX: MODEL COMPARISON

D.1 *Station independent scatter Plot of ground Truth against predicted Values of both Models.*

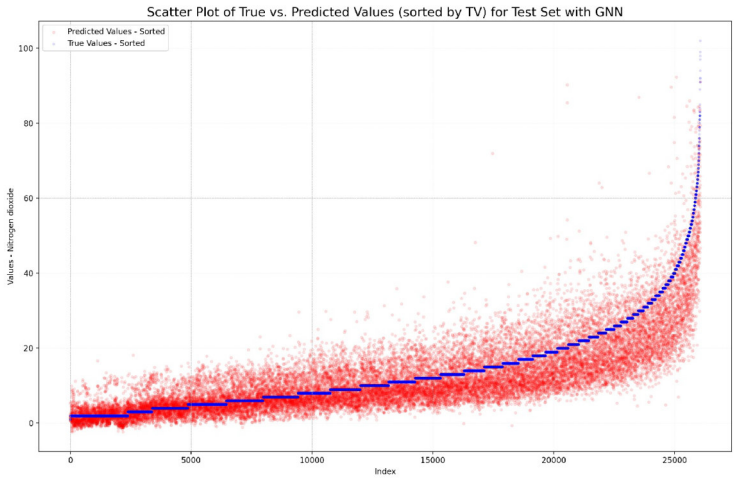

Figure 20: Scatter plot of true values, highlighted in blue against predicted values, highlighted in red for GNN. The Plot combines all stations per model and is sorted by increasing target value.

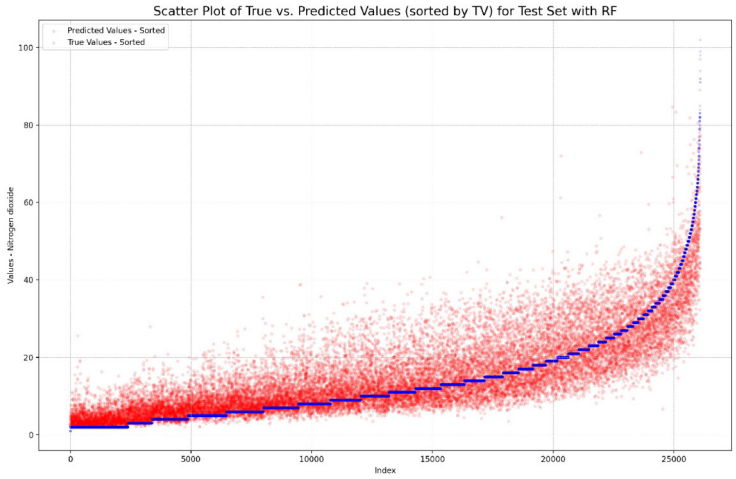

Figure 21: Scatter plot of true values, highlighted in blue against predicted values, highlighted in red for RF. The Plot combines all stations per model and is sorted by increasing target value.

E APPENDIX: GNN SPECIFICATION- HYPERPARAMETER

Hyperparameter Graph Neural Network		
Parameter	**Parameter Range**	**Best Performing**
Distance estimation	Inverse distance, the difference in distance	Difference in distance
Batch size	26, 34, 42, 50, 58	26
Encoder **(Generator)**		
units dense layer 1	48, 64, 80, 96, 112	48
units dense layer 2	48, 64, 80, 96, 112	64
Optimizer encoder	Adam, SGD, RMSprop	Adam
Learning Rate encoder	Log uniform distribution: 0.0005 – 0.003	0.001790
Momentum encoder	Uniform distribution: 0.5 – 0.99	-
L2 regularization encoder	Uniform 0 – 0.02	0.007949
Decoder **(Discriminator)**		
Optimizer decoder	Adam, SGD, RMSprop	RMSprop
Learning Rate decoder	Log uniform distribution: 0.0005 – 0.003	0.000612
Momentum decoder	Uniform distribution: 0.5 – 0.99	0.52711
L2 regularization decoder	Uniform 0 – 0.02	0.000203
(Dense Layer)		
scaled target	none, log transform	none
Loss function	MSE, Huber loss	MSE
Optimizer	Adam, SGD, RMSprop	SGD
Learning Rate	Log uniform distribution : 0.0005 – 0.003	0.00185
Momentum	Uniform distribution: 0.5 – 0.99	0.8177
L2 regularization	Uniform 0 – 0.02	6.642e-5
Batchsize	16, 24, 32, 40, 48, 56	56
Similarity measures	Inverse distance, the difference in distance (DID)	DID
Activation function decoder	Relu, Swish	Relu
n layer dense network	2, 3	3
Unit dense layer 1	32, 48, 62, 80, 96, 112, 128	112
Unit dense layer 2	32, 48, 62, 80, 96, 112, 128	112

Table 8: Hyperparameter tuning results for GNN: Parameter adversarial network for graph embedding and dense network for regression, evaluated on Tree-structured Parzen Estimator with 50 iterations.