Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Masterthesis

Towards Fine Granular Particulate Matter Estimations From Satellite Data

Joshua Schimmelpfennig

joshua.schimmelpfennig@studium.uni-hamburg.de
MIN-Fakultät, Fachbereich Informatik
Informatik (M.Sc.)
Matr.-Nr. 6813643

Erstgutachter: Prof. Janick Edinger
Zweitgutachter: Dr. Philipp Kisters

Abgabe: 13.10.2025

Abstract

This study analyses the performance of satellite-based $PM_{2.5}$ estimation methods in urban, high-resolution environments. The methods based on AOD modified with direct satellite input, approximating the dark-target retrieval algorithm. PCA-GRNN, XGBoost, 2-layer and 5-layer MLP were tested. The case study was performed in the City of Hamburg with $N = 1460$. XGBoost ($R^2 = 0.855$) performed best, followed by PCA-GRNN ($R^2 = 0.723$).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

MAE	Digitial Number
PM	Particulate Matter
PM_{10}	Particulate Matter with an aerodynamic diameter of less than $10\mu m$
PM_1	Particulate Matter with an aerodynamic diameter of less than $1\mu m$
$PM_{2.5}$	Particulate Matter with an aerodynamic diameter of less than $2.5\mu m$
$RMSE$	Root Mean Square Error
ANN	Artificial Neural Networks
AOD	Aerosol Optical Depth
AOT	Aerosol Optical Thickness
CRS	Coordinate Reference System
CTM	Chemical Transport Model
DEM	Digital Elevation Map
DN	Digitial Number
DWD	Deutscher Wetterdienst
GIS	Geographic Information System
GRNN	General Regression Neural Network
HaLM	Hamburger Luftmessnetz
LaSRC	Land Surface Reflectance Code
MLP	Multilayer Perceptron

MODIS . . . . . Moderate Resolution Imaging Spectroradiometer

NDBI . . . . . Normalized Difference Built-Up Index

NDVI . . . . . Normalized Difference Vegetation Index

OLI . . . . . Operational Land Imager

PBLH . . . . . Planetary Boundary Layer Height

PCA . . . . . Principle Component Analysis

PCA-GRNN . . . . . Principle Component Analysis - General Regression Neural Network

PRS . . . . . atmospherirc pressure

RH . . . . . relative humidity

TEMP . . . . . temperature

TOA . . . . . Top of Atmosphere

WD . . . . . wind direction

WS . . . . . wind speed

XGBoost . . . . . Extreme Gradient Boosting

1 Introduction

1.1 Motivation

Air Pollution is an ongoing concern in the EU. In the 2023 EU air quality status report [19], the EU Commission states that air pollution remains a primary health concern and the most significant environmental health hazard for EU citizens, with pollution exposure levels exceeding EU and WHO standards.

Air pollution exposure has known negative health effects, including respiratory and cardiovascular diseases, such as asthma and tachycardia, leading to decreases in life quality and an increase in mortality.[7] Fine particulate matter, aerosols with an aerodynamic diameter of less than 2.5μm ($PM_{2.5}$)[62], are linked to increased hospitalization and mortality [7].

The EU Commission passed stricter standards under the Green Deal to reduce fine particulate matter ($PM_{2.5}$) exposure until 2030, aligning them more closely with the WHO standard [18]. In the standard, the reduction target for $PM_{2.5}$ is 18$\mu g/m^3$.

In 2015, Karagulian et al. [33] analyzed PM source studies and summarized that globally, urban $PM_{2.5}$ is caused by traffic, but also has contributions from industrial activities and domestic fuel burning. However, 22% of $PM_{2.5}$ contribution remains unexplained, and the authors additionally note strong heterogeneity between sources and data incompleteness.

Methods mitigating exposure require information about exposure rates and concentrations within urban environments. Good, high-quality data coverage is essential for lawmakers, local authorities, and activists to advocate for effective strategies and policies aimed at reducing fine particulate matter affecting their citizens. The best method to measure near-ground $PM_{2.5}$ aerosols is through *in situ* measurements, registering the near-ground concentrations directly at location [63].

In cities like Hamburg, Germany, official measurement networks operated by different agencies, like the DWD [57] and the HaLM [24]. While they provide high-quality, validated data, they are sparse and only provide a dot measurement of that specific location. The measured values are used to estimate the overall climate of a location, but offer little insight into local microclimates within cities or even streets. Furthermore, *in-situ* measurements require the installation and maintenance of said equipment, which may not be possible everywhere due to cost, location, sparse population, or other factors. Nevertheless, without proper spatial coverage, citizens can be exposed to local pollution hot spots.

While it is possible to increase coverage through specific projects, such as sensors on city buses [32], along bike routes [50] or bikes themselves [16], these methods only contribute point measurements.

This is also true for public citizen science projects like Sensor.Community, which allows citizens to contribute to a global database of self-build measurement stations [48].

In 2010, van Donkelaar et al. [52] released a milestone study, estimating global $PM_{2.5}$ from Aerosol Optical Depth (AOD), introducing remote-sensed satellite imagery as a solution to both sparse geo-spatial data resolution and poor *in situ* measurement infrastructure. The benefits of satellite-derived $PM_{2.5}$ are global data availability, even in regions without the necessary infrastructure. A single satellite also has no inter-sensor errors, because a single sensor applies biases like measurement drifts equally to all pixels, whereas this has to be controlled in distributed sensor networks. This method comes at the cost of long revisit times until the satellite completes its orbit and measures another sample.

The study found a good conformity with ground-based measurement stations and showed that $PM_{2.5}$ can be estimated using satellites. While the study produced a globally covered product, the raw pixels covered (at nadir) an area of $10km \times 10km$. This is a result of the available resolution of input data. In short, AOD is a science product measuring the amount of aerosols in a given vertical column above the surface, retrieved from satellite data. The source data used in this study is itself at $10km \times 10km$ resolution and thus limits the global model.

This normally ground-based measurement, measured for example by AERONET [29], requires extensive algorithmic work to be estimated from space. As such, this science product is only available at higher data levels and not every satellite platform or provider necessarily distributes this product.

Zang et al. used a combination of AOD data from the geostationary satellite Himawari-8 and PM1 measurements from ground stations to estimate hourly, daily, and monthly PM1 concentrations in China [59]. The estimation is tied to the data frequency of the Himawari-8 satellite.

In a Land Surface Temperature (LST) case study in Hamburg, Bechtel et al. noted the poor spatial resolution of available LST data [5]. The authors highlight that for an urban analysis, a resolution of 100m is similar to a housing block size and thus more reasonable for analysis. Further, they showed a tradeoff between image frequency and resolution, caused by the orbit of the instrument. In short, due to Hamburg's northern location, geostationary satellites only have a poor spatial resolution of the city. Contrasting this, polar or sun-synchronous orbits can give instruments an at-nadir perspective, at the cost of a much higher revisit time. The authors analyzed a down-scaling scheme, correlating predictors in low and high resolution, capable of estimating high-resolution imagery from low-resolution footage.

In contrast, newer approaches circumvent this issue by estimating $PM_{2.5}$ directly from satellite data, as seen in [61] and [20], both creating machine learning models generating

$PM_{2.5}$ from high-resolution (30m) Landsat data.

1.2 Problem Statement

As shown in the introduction, $PM_{2.5}$ is a significant health concern, affecting citizens in dense, urban environments strongly. Zhang et al. [63] link immune system disorders, cancers, and diseases in respiratory, cardiovascular, and neurological systems to $PM_{2.5}$ exposure.

While the adverse effects in urban environments are known, the same cannot be said for the distribution of $PM_{2.5}$. The missing geo-spatial resolution can be improved with satellite estimations, since they record data over a large spatial expanse. AOD based methods have shown great promise in the estimation of $PM_{2.5}$ from satellites, but are limited by the size constraints of the AOD products. High-resolution $PM_{2.5}$ estimations from satellites are a recent topic, and approaches from [61] and [20] are validated on a macroscopic national scale. They are not validated in microscopic, high-resolution urban environments, which are known to be difficult for AOD retrievals [8][63].

To find $PM_{2.5}$ hot-spots and improve citizen safety, $PM_{2.5}$ estimation methods have to be evaluated in high-resolution urban environments.

1.3 Scope

This work aims to analyze $PM_{2.5}$ estimation models in high-resolution urban environments, using, for example, the 30m satellite data seen in [20]. The goal is to find a model capable of estimating $PM_{2.5}$ directly from satellite footage to retain the high source-resolution. Towards this goal, AOD and satellite-based methods will be compared. The best performing methods will be used in an analysis specifically targeting such a challenging environment. To perform the analysis, a study is selected and a dataset is built for it. In it, the methods are compared against ground stations to evaluate the performance of the methods. The goal of this work is to answer the following research questions (RQ):

1. How accurately can fine Particulate Matter be estimated in urban environments directly from satellite images?

2. How strongly does the temporal effect of seasons influence estimations from satellite images?

3. With the employed methods, is there a difference in estimator performance when comparing rural and urban ground stations?

Worse performances are expected, since high-resolution urban areas contain a lot of infrastructure, interfering with satellite retrievals.

1.4 Thesis Structure

In this thesis, existing estimators and their performance will be analyzed in high-resolution urban environments. Chapter 2 introduces foundational knowledge necessary to understand estimations in remote sensing and specifically of particulate matter. This understanding will be necessary for chapter 3, where existing estimators and models will be discussed. The best performing models will be used for this thesis and inspected in more detail in chapter 4. The presented estimators utilize different independent variables, the data requirements for this thesis are formulated in chapter 4 as well. Chapter 5 illustrates the evaluation strategy and how the best performing method is determined. Further, a study area is selected, and the design of the evaluation system is detailed. The collected results are collected, analyzed, and interpreted in chapter 6.

Finally, chapter 7 summarizes the results of this thesis and provides an outlook for future work.

2 Fundamentals

This section briefly summarizes the background knowledge required for this thesis. First, Particulate Matter is introduced and defined. The difference between primary and secondary PM is shown. Second, relevant indices are introduced, and the calculation formula is shown. Lastly, the concept of aerosol inversion is presented.

2.1 Particulate Matter

Zhang et al. [62] provide an overview of the formation of Particulate Matter (PM).

PM is classified by its atmospheric diameter, because the particle shapes can be odd. The following classes are used to refer to specific PM sizes[62][59].

- Particulate Matter with an aerodynamic diameter of less than $10\mu m$ (PM_{10})
- Particulate Matter with an aerodynamic diameter of less than $2.5\mu m$ ($PM_{2.5}$)
- Particulate Matter with an aerodynamic diameter of less than $1\mu m$ (PM_1)

PM is distinguished into primary and secondary PM, depending on its formation. According to Zhang[62], primary PM is emitted directly into the atmosphere, generally through anthropogenic processes. Secondary PM forms through chemical processes, reactions of gases present in the atmosphere.

2.2 Satellite Indices

Indices are early image processing methods that resulted from the mapping of vegetation density [31]. They are quotients of different satellite bands, bringing out specific features in the raster, For example, vegetation density. A popular index is the Normalized Difference Vegetation Index (NDVI). It is used to assess vegetation health, indicated by greenness intensity of the vegetation's chlorophyll [37].

Equation 2.1 shows the Landsat 8 formula for this index.

$$\text{NDVI} = \frac{\text{Band 5} - \text{Band 4}}{\text{Band 5} + \text{Band 4}} \tag{2.1}$$

Another useful index is the Normalized Difference Built-Up Index (NDBI)[60]. According to Jensen, the index returns positive values for built-up or barren areas and was reported as 92% accurate[31].

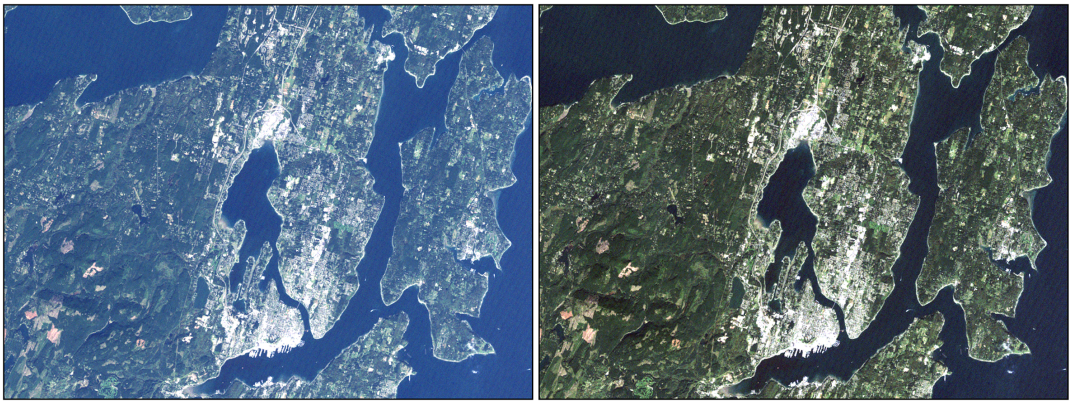

Figure 2.1: Inversion example showing TOA (left) and surface reflectance (right). Image taken from USGS [6].

Formula 2.2 is used to calculate the NDBI for Landsat 8:

$$\text{NDBI} = \left(\frac{\text{Band } 6 - \text{Band } 5}{\text{Band } 6 + \text{Band } 5}\right) - \text{NDVI} \tag{2.2}$$

2.3 Aerosol Inversion

When satellites acquire images outside the atmosphere, aerosols attenuate the solar radiation through scattering, reflection, and absorption [63]. Satellite data with attenuation is referred to as Top of Atmosphere (TOA) reflectance[53]. When analyzing satellite data, the surface is often the point of interest. In this case, aerosol attenuation is seen as noise, and aerosol inversion is the process to remove said noise from the image. The result is referred to as "surface" reflectance.

Figure 2.1 is an image in the public domain, taken from the USGS website. It shows the difference between TOA reflectance on the left and surface reflectance on the right. In physical models, Aerosol Optical Depth(AOD) describes the noise as an integrated extinction coefficient through the atmosphere. AOD and AOT describe the same property.

The theory behind aerosol inversion is as follows: aerosols attenuate to specific frequencies, at wavelengths between 0.25-0.7 microns. A negative example can then be procured by selecting images outside this frequency. An example is shown in the following section.

2.3.1 Retrieving AOD from MODIS

Koelemeijer et al. explain the retrieval of AOD from MODIS [34]. The basis of the algorithms is the reflective property of aerosols at specific frequencies. Fine aerosols have a higher attenuation at the 0.47 and 0.66μm and a negligible attenuation in the 2.1μm band. The 2.1μm band is used to determine the true reflectivity of the surface, assuming a constant relationship with the 0.47 and 0.66μm bands. Using the calibration, the 0.47 and

$0.66\mu m$ bands are then used to determine the proportion of reflectance attributed to fine aerosols, expressed as AOD. In other words, this is a de-noising of the lower frequencies using a calibration obtained at a higher frequency.

This process assumes that the majority of aerosols in the target area are fine and therefore reflective at a lower frequency. The authors note that concentrations of Saharan dust, observed for example over southern Europe, are reflective at $2.1\mu m$ and introduce a bias into the process. Secondly, the calibration correlates the surface reflectivity in $0.47\mu m$, $0.66\mu m$, and $2.1\mu m$, but surface reflectivity depends on the surface type and humidity content and is not a constant.

2.3.2 Inversion Algorithms

Dark Target

The dark target algorithm[38] has been described above. It is based on the assumption of negligible noise at 2.1 microns and retrieves AOD over dense vegetation. The algorithm is less reliable over bright surfaces, such as concrete [38].

Deep Blue

The deep blue[30] algorithm is based on high aerosol noise in red\near-ir bands over bright surfaces and low noise in the dark blue band (0.41 microns). To determine AOD, a reference image is used from long-term TOA observations.

3 Related Work

When researching the estimation of near-ground particulate matter (PM) from satellite imagery, Aerosol Optical Depth/Aerosol Optical Thickness (AOD/AOT) [2] has been found to have a strong correlation with near-surface PM [63], it is thus often used as a parameter when estimating PM.

Due to the importance of AOD, when analyzing research in this field, great care has to be taken to differentiate between research focusing on the estimation of AOD and that of estimating PM.

3.1 Correlation of Aerosol Optical Depth and Particulate Matter

Koelemeijer et al.[34] analyzed the correlation of satellite-derived AOD over Europe and PM concentrations near the ground. They collected data for 2003 from MODIS aboard both Terra and Aqua and performed their analysis against nationally reported $PM_{2.5}$ and PM_{10} recordings from the AIRBASE platform. The authors did not correct for biases introduced by testing methodology and national testing differences, but did further subdivide AOD into AODF, AOD from fine aerosols. The analysis is split between spatial and temporal correlation. In the spatial correlation, the authors found a good correlation between AOD (AODF) and PM: $PM_{2.5}$ 0.63 (0.77), PM_{10} 0.58 (0.53). The authors found a 15% difference when examining yearly averages of PM when comparing cloud-free and cloud-covered conditions. Increased PM averages during cloud-free conditions are relevant for AOD, because AOD can only be determined during these conditions. In their temporal correlation analysis, the authors found a strong negative correlation with precipitation, indicating the importance of aerosol extinction during rain. Contrary to this, this effect is less pronounced during winter seasons. Indeed, the lower boundary height during winter months appeared to counter the lower concentration during higher precipitation events, by concentrating aerosols in the atmosphere and, by extension, close to the ground. When correcting the temporal correlation of AOD and PM with the boundary layer height and a function of relative humidity (AOD*), the authors found a better correlation than without: hourly correlation of $PM_{2.5}$ is 0.77, PM_{10} is 0.68.

3.2 Estimation of Aerosol Optical Depth

Li et al. [39] evaluated the retrieval of AOD using the Land Surface Reflectance Code (LaSRC) on Landsat-8 30m and Sentinel-2A 10m imagery. Specifically, Top of the At-

mosphere (TOA) and surface reflectance were used to retrieve AOD over 20 Chinese cities. The retrieved AOD was compared to coincidental AERONET readings within a 10-minute retrieval time frame. The authors conclude that LaSRC is a good medium AOD retrieval algorithm for urban environments. Additionally, the accuracy for AOD derived from Sentinel-2A was higher compared to AOD retrieved from Landsat-8. Important in this research is the fact that the authors managed to retrieve AOD in a high spatial resolution, sufficient for urban analysis, directly from remotely sensed imagery.

3.3 Estimating Particulate Matter from Aerosol Products

Liu et al. performed a quantitative analysis of the relationship between AOT and $PM_{2.5}$ in St. Louis, Missouri[41]. Using two multiple linear regression models, they correlated MODIS or MISR AOT products together with meteorological parameters to near-ground $PM_{2.5}$. To match satellite pixels with meteorological ground stations, they averaged MISR and MODIS pixels that fell within a 30km search radius around each ground station. The fitted models achieved $R^2 = 0.62$ for the MISR AOT and $R^2 = 0.51$ with MODIS AOT with regards to daily average $PM_{2.5}$. After analyzing their predictors, the authors discovered that AOT predicted lower $PM_{2.5}$ during spring, when compared to the rest of the year.

A major milestone in the PM from AOD research is the research by van Donkelaar et al. [52]. In their 2010 study, the authors acknowledged a lack of global models on the presence of fine particulate matter with a diameter less than $2.5\mu m$ ($PM_{2.5}$), hindering the analysis of health impacts on populations without *in situ* measurements. Additionally, the authors concluded that the density of ground-based measuring stations does not improve the limited geo-spatial coverage and lack of regional representation of point-based measurements.

Without a global model, studies on the global impact of $PM_{2.5}$ cannot form meaningful conclusions. The authors developed a global estimator capable of predicting AOD from a fusion of MODIS and MISR instruments from NASA's Terra satellite, and the GEOS-Chem chemical transport model, on a $0.1 \circ \times 0.1 \circ$ or $\approx 10km \times 10km$ grid. The use of satellite-derived AOD, fused from MODIS and MISR sensors, in combination with the GEOS CTM, improved the performance of $PM_{2.5}$ estimation to an R of 0.77 for North America and 0.83 for the rest of the world. This model is one of the first global $PM_{2.5}$ concentration climatologies, allowing studies of the impact of PM in areas without a sufficient network of ground-based measurement stations.

Although the model is a breakthrough for long-term studies of $PM_{2.5}$ concentrations, there are several ways in which the $PM_{2.5}$ estimator can be improved to facilitate a better in-depth analysis. For example, the precision of $PM_{2.5}$ daily averages is not sufficient to analyze diurnal effects. While using a geostationary satellite will overfit a model to the image of the satellite, such a system could improve the temporal resolution of the model

and allow for analysis of diurnal effects. Additionally, improving the algorithms used to extract AOD could increase the accuracy of the following $PM_{2.5}$ estimation. Increasing spatial resolution could improve estimations in areas with denser geographic features and local pollution extrema, such as cities and metro areas.

Increasing and varying the source satellites used for AOD can reduce the sampling bias introduced by the reliance on a singular system. In this study, an excellent alternative would be the partner satellite Aqua. Aqua follows the same orbit as Terra, but crosses the Equator in the afternoon instead of the morning [44].

Zang et al. used a combination of AOD data from the geostationary satellite Himawari-8 and PM1 measurements from ground stations to estimate hourly, daily, and monthly PM1 concentrations in China [59].

Zamani Joharestani et al. [58] evaluate $PM_{2.5}$ estimations from AOD in the city of Teheran, an urban environment. The data availability was sparse due to missing MODIS-AOD retrievals. According to the authors, the desert environment interferes with the satellite retrieval, and the ground stations experience frequent outages. The authors evaluated RandomForest ($R^2 = 0.78$), XGBoost ($R^2 = 0.80$), and a deep MLP model ($R^2 = 0.77$), concluding that while XGBoost performed best, the performance of all models is similar.

3.3.1 Estimating Particulate Matter from Satellite Imagery

In 2018, Fernández-Pacheco et al. [20] correlate path radiance, the noise caused by aerosol scattering, to PM_{10} from ground stations using a random forest algorithm. They show that particulate matter can be estimated from satellite radiance directly. In 2019, Zhang et al. [61] employed a neural network to estimate $PM_{2.5}$ and PM_{10} from Landsat 8 OLI band reflectance data. They showed great conformity with ground stations and low RMSE.

3.4 Comparing Model Complexity

Nguyen et al. [45] compare traditional and deep learning algorithms in the task of estimating $PM_{2.5}$ with downscaling.

The previous AOD-based 3x3km $PM_{2.5}$ maps were deemed insufficient for urban and rural analysis. The previous maps were downscaled to 1x1km by resampling with the nearest neighbor algorithm, together with other PM precursors like humidity, temperature, windspeed, and other meteorological attributes. The parameters were resampled to a standard GeoTiff grid.

Traditional models include random forest for tree-based algorithms and catboost for gradient boosting methods. Deep models include LSTM and CNN. These models also introduce temporal and spatial attributes, with daily $PM_{2.5}$ maps as the temporal parameter with a lag of 1, 2, and 3 days. The spatial parameters for the CNN were varied by

Authors	**Model**	R^2	**Basis**
Fernández-Pacheco et al.	Random Forest	0.94	direct
van Donkelaar et al.	CTM	0.83	AOD
Nguyen et al.	Catboost (2-day lag)	0.81	$PM_{2.5}$
Nguyen et al.	ConvLSTM	0.81	$PM_{2.5}$
Zhang et al.	MLP	0.36\0.80	direct
Zamani Joharestani et al.	XGBoost	0.80	AOD
Zamani Joharestani et al.	MLP	0.77	AOD
Zang et al.	PCA-GRNN	0.65	AOD
Liu et al.	Multiple Linear Regression	0.62	AOD

Table 3.1: Models used sorted by the R^2 score.

the pixel size of the initial kernel. As a result, a 15x15 kernel was better for this task, but an 11x11 kernel could generalize better.

The goal of the algorithms was to estimate daily $PM_{2.5}$ maps in 1x1km resolution. They were trained using an 80/20 split of the dataset and evaluated in two conditions: the complete dataset and iterative dropping of a single station, testing the model's ability to generalize. The authors conclude that Catboost with a 2-day lag and ConvLSTM with an 11x11 kernel performed the best.

Summary

This chapter summarizes the current method for the estimation of AOD and $PM_{2.5}$. The history of AOD research is longer, due to the importance of AOD on aerosol inversion procedures from chapter 2.

Table 3.1 gives an overview of the models used in the related work. While the Random Forest approach from Fernández-Pacheco et al. reports the best R^2, the data and code cannot be verified.

4 Methods

The previous chapter 3 introduced the current methods of estimating near-ground particulate matter from satellite products.

The approaches can be broadly categorized into utilizing the science product AOD or the measured radiation directly from satellite images.

Common among both groups is estimating and analyzing Particulate Matter on a macro level, working with national [61] or even global[52] perspectives.

The available data at the time is a limiting factor on fine-granular estimations. AOD from MODIS is one of the earlier systems used for the estimation of PM, but it is only available at $1 \times 1km$ resolution [1][41][63]. Further, AOD products utilize surrounding grid-boxes and reduce the resolution of the final image [8].

New satellite systems, for example, Landsat 8/9 OLI, can record with a resolution of 30m at nadir. According to Bechtel et al., this resolution is appropriate for urban analysis [5]. Nevertheless, while there are studies estimating AOD from Landsat 8/9 OLI [39][21][40] and achieve acceptable performance, even in source resolution, there is no official AOD product.

Since methods to estimate AOD from Landsat 8/9 OLI produce good-fitting, high-resolution results and methods to estimate PM from AOD are similarly promising, we expect the combination of both steps to be possible. The combination of both steps to estimate PM from satellites directly, retaining the high resolution of the OLI system, has already been shown by Zhang et al. on a national scale in China [61]. The performance of this approach on a micro scale has to be evaluated as well, specifically in urban environments.

Urban environments are densely populated and, as discussed in chapter 2 are expected to have higher mass concentrations of anthropogenic, primary Particulate Matter, when compared to rural areas. MODIS-AOD products are, at least in part, derived via the dark target algorithm[1]. The dark target algorithm is known to be less accurate in areas with bright surfaces [8][63]. Urban environments are defined by dense infrastructure. Its reflective properties may interfere with direct estimations when the method is based on AOD.

For these reasons, urban environments differ strongly from the areas usually analyzed. Due to the difference in environments, it is also possible that this approach cannot adequately estimate PM in these areas.

In this thesis, we perform a quantitative analysis of current methods, estimating PM in urban environments directly from satellites. In addition, temporal and spatial effects

on the results will be considered as well.

4.1 Analysis Requirements

To perform a qualitative Analysis of existing methods, they have to be selected first. Once the list of methods is determined, a unified approach for the analysis has to be established. To achieve this, the models, preprocessing, and datasets of prior methods will be discussed. The sum of the discussion formulated the requirements for a unified method.

Since prior methods work on real-life data and not a shared, comparable dataset, a shared dataset needs to be established for the unified method. Especially the list of input parameters requires great care. If AOD based methods are selected, a method to adapt them for direct satellite radiation will be required as well.

Analysis of seasonal effects on prediction accuracy requires a function to assign seasons to the dataset. It also necessitates the presence of seasons with observable effects in the study area as well.

Spatial analysis will require assignment of ground stations to urban and more rural environments. This will also serve as a verification for the full qualitative analysis.

Lastly, it has to be noted that satellite estimations are constrained by their nature to clear weather scenarios. If the ground is not visible, Particulate Matter estimations become impossible.

4.2 Method Selection

Nguyen et al. [45] have shown that machine learning algorithms perform adequately when estimating PM. Further, they have shown that deep learning algorithms do not perform better at that task.

In this thesis, we aim to select methods with varying complexity in order to better understand the predictive performance in urban environments, given a fixed set of input parameters.

The first model is the shallow MLP by Zhang et al. [61]. It is selected because it directly estimates PM from satellite data.

The PCA-GRNN model by Zang et al. [59] performed well at high temporal resolution.

The XGBoost by Zamani Joharestani et al. [58] performed similarly to other machine learning algorithms and servers at its representative.

The deep MLP by Zamani Joharestani et al., as well, is selected as a direct deep learning comparison to the shallow MLP by Zhang et al.

The following tabular input parameter overviews are adapted from [61], since we found it very clear and easy to read.

Category	Parameter	Unit	Temporal Resolution
Particulate Matter	$PM_{2.5}$ PM_{10}	$\mu g/m^3$	1 h
Landsat 8 OLI	Band 1 Band 3 Band 7	Spectral Radiance	16 days
	NDVI	Spectral Index	
	Acquisition Month	unitless	
Meteorological Parameters	RH	%	3 h
	PRS	Pa	
	TEMP	K	
	WD	$^\circ$	
	WS	m/s	

Table 4.1: Dataset employed in [61] to train the MLP.

4.2.1 Multilayer Perceptron Model

Zhang et al. [61] use an Multilayer Perceptron in their direct estimation of Particulate Matter in China. The authors acknowledge previous research into the AOD-PM relationship and success in specific regions, but they note the need for further validation. Listed limitations include mismatching spatial-temporal resolutions and the link between AOD and the spatial distribution of $PM_{2.5}$. While research had focused on linear/non-linear regression models, artificial neural networks had proven better regression results. To address these limitations, Zhang et al. implemented a MLP model with back-propagation, combining meteorological and temporal factors in their estimation.

The model inputs can be seen in table 4.1. As the estimation target, $PM_{2.5}$ and PM_{10} data from ground stations are used. While previous research used AOD products as model inputs, spectral reflectance from Landsat 8 OLI Bands 1, 3, and 7, and the index NDVI is used instead. With this, the authors use 5 Landsat bands in total for their estimation.

In addition to Landsat data, meteorological parameters were gathered from independent ground stations.

To create features from the satellite images, the three datasets need to be integrated. The PM ground stations were focused, and coincident Landsat images gathered. Then, a circle with a radius of 15m was drawn around the ground stations, and the image pixels falling within were averaged. This method is similar to [41]. The meteorological parameters were assigned to the ground stations by measure of shortest distance. Finally, the

authors performed a dataset correction by means of a longitude-latitude zone algorithm, assigning each study area a station subset with the highest resulting R^2 due to the size of the study area.

The model consists of two hidden layers, with 15 neurons each. Before the input layer, the parameters were normalized because MLP's are sensitive to large features. The activation of the hidden layers is reported as tan-sigmoid, with purelin as the activation for the output layer. These are the names of the activation functions in MATLAB, tan-sigmoid is identical to tanh, and pureline is the identity function if used in the output layer.

The final R^2 for $PM_{2.5}$ reached 0.80 with the LLZ method, but only 0.37 without. While China is a geographically diverse region, the second result might reflect the actual model performance more accurately. In addition, the hidden layers of an MLP may contain more than one local minimum, leading to different validation results depending on weight initialization. Reporting a single three-way validation result can be inconclusive.

4.2.2 PCA-GRNN Model

Zang et al. estimated high-resolution hourly PM_1 using a hybrid PCA-GRNN model [59]. Since ANNs gained in popularity, especially General Regression Neural Networks, the authors focus on extending the previous methods by applying PCA before passing the components to the GRNN. The purpose of this was to reduce the effect of co-linearity between predictors, which had not been controlled in the past. PCA merges features while retaining their variability, producing a new set of indices in feature-space, organized by decreasing variability. In other use cases, PCA can be employed to reduce the dimensionality of the input by retaining only the first n components, assuming that they account for the most variability in the given features.

In addition to adding PCA, the authors also focused on estimating Particulate Matter with an aerodynamic diameter of less than $1\mu m$, a subset of $PM_{2.5}$ and PM_{10}.

Table 4.2 shows the dataset used in this study. Because Himawari-8 is a geo-stationary satellite and can provide hourly AOD products, the researchers included the hour of recording as an input parameter. In contrast to the model discussed in section 4.2.1, this model uses Digital Elevation Map for geographic reference and also Planetary Boundary Layer Height. While this study uses AOD as an input parameter instead of direct band reflectance, we are nonetheless interested in a direct comparison, since the datasets are otherwise so similar.

In preprocessing, meteorological parameters and satellite products were resampled to the AOD grid. Samples were collected within a 5km radius around the PM stations and within double the parameters temporal resolution, centered on the Himawari AOD measurements.

The topology of a GRNN is determined by the method itself. General Regression Neural Network is a single-pass function approximator that is capable of learning on sparse

Category	Parameter	Unit	Temporal Resolution
Ground Stations	PM_1	$\mu g/m^3$	5 min
	month	unitless	–
	hour	unitless	–
	longitude	$^\circ$	–
	latitude	$^\circ$	–
Meteorological Parameters	RH	%	6 h
	PRS	Pa	
	TEMP	K	
	WS	m/s	
	PBLH	m	3 h
Satellite Products	AOD	unitless	1 h
	NDVI	Spectral Index	16 days
	DEM	m	–

Table 4.2: Dataset employed in [59] to train the PCA-GRNN.

data [49]. To achieve this, the number of neurons in the single hidden layer is equal to the number of learning samples. The size of the input layer is equal to the feature dimension. In the hidden layer, the samples are aggregated using Gaussian kernels, similar to SVMs. The results are summed in the weighted and simple summation nodes before the sums are divided in the output layer.

GRNNs can learn from sparse data and can be of use in this problem, since the revisit time of satellites can be poor. The drawback of this method is the scaling with feature size, increasing computational intensity with each sample.

4.2.3 XGBoost

XGBoost (eXtreme gradient boosting) in Zamani Joharestani et al. [58] is applied to the urban environment of Tehran, achieving a comparable performance to Random Forest and deep learning models. The authors selected this method for its resilience against overfitting. Table 4.3 shows the data input used in the study. Unfortunately, the authors reported a large volume of missing data, both in ground station $PM_{2.5}$ ($\approx$ 54%) and satellite AOD 3km ($\approx$ 97%). Similar to previous methods, ground station $PM_{2.5}$ were extended with meteorological and satellite products.

The authors selected 200 estimators, a depth of 8, a gamma of 0.7, and a child weight of 8 as the hyperparameters for XGBoost with cross-validation grid search.

Nguyen et al. have reported good performance with machine learning approaches as

Category	Parameter	Unit	Temporal Resolution
Ground Stations	$PM_{2.5}$	$\mu g/m^3$	daily
	weekday	unitless	–
	day of year	unitless	–
	season	unitless	–
	longitude	$^\circ$	–
	latitude	$^\circ$	–
	altitude	m	–
Meteorological Parameters	RH	%	daily
	TEMP max	$^\circ C$	
	TEMP min	$^\circ C$	
	TEMP	$^\circ C$	
	Visibility	km	
	WS	m/s	
	WS (sustained)	m/s	
	PRS	Pa	
	dew point	$^\circ C$	
Satellite Products	AOD 3km	unitless	5 min
	AOD 10km	unitless	

Table 4.3: Dataset employed in [58] to train the PCA-GRNN.

Layer	Neurons	Regularization
1	270	–
2	120	L2
3	70	L2
4	50	L2
5	20	L2, L1
6	1	–

Table 4.4: Summary of Table 4 in [58], topology of the deep model.

well [45], compared to deep learning models. While the dataset was sparse and contained a lot of missing values, it still performed well in an urban environment. Since this study focuses on urban environments as well, we expect a similar or better performance, assuming a study area with better data availability is selected.

4.2.4 Deep Learning

As noted in the previous section 4.2.3, gradient boosting is performed similarly to deep learning in [58].

While [45] discovered that the estimation task of $PM_{2.5}$ is well fitting for machine learning algorithms and too sparse for deep learning, it was not focused on an urban environment.

A shallow MLP is included in section 4.2.1 already, the increasing complexity of the urban environment, as well as high resolution, should be more suitable for a deep model.

The data and processing methodology are the same as seen in section 4.2.3.

Table 4.4 was taken from Table 4 in [58]. It shows the topology of the deep model, all layers use the ReLU as their activation function. Regularization was added due to the missing values with $\alpha = 0.002$ and $\alpha = 0.001$ in the fifth layer.

4.3 Unified Method

4.3.1 Estimating PM from Satellite Images

This section discusses how PM may be estimated directly from satellite images and how we can adapt AOD based methods. Of the selected methods, only the shallow MLP model by Zhang et al.[61], presented in subsection 4.2.1, utilizes raw spectral radiance from Landsat 8. The bands used are 1, 3, and 7, with 4 and 5 included via the NDVI. Table 4.5 is the band reference table from the Landsat 8/9 OLI/TIRS Collection 2 Level 1 Data Format Control Book[35]. As can be referenced from the table, the bands cover the frequency ranges $0.435 - 0.451$, $0.533 - 0.590$, and $2.107 - 2.294 \mu m$ directly, $0.636 - 0.637$

Band Number	Band Description	Band Range (μm)
1	Coastal Aerosol	0.435-0.451
2	Blue	0.452-0.512
3	Green	0.533-0.590
4	Red	0.636-0.637
5	Near-Infrared (NIR)	0.851-0.879
6	Short Wavelength Infrared (SWIR)	1.566-1.651
7	SWIR 2	2.107-2.294
8	Panchromatic	0.503-0.676
9	Cirrus	1.363-1.384
10	Thermal Infrared Sensor (TIRS) 1	10.600-11.190
11	TIRS 2	11.500-12.510

Table 4.5: Landsat 8/9 Band Reference Table in μm, Table 2-1 from [35]

and $0.851 - 0.879\mu m$ indirectly through NDVI. As discussed in section 2.3 aerosol inversion, aerosols attenuate solar radiation at frequencies in the range $0.25 - 0.7\mu m$, which can be covered by Landsat bands 1-4 and 8. Further, the discussed dark target algorithm uses frequencies around $2.1\mu m$, due to negligent attunement at these frequencies, which is covered by Landsat band 7.

It appears that Zhang et al., with respect to their findings, successfully approximated the dark target algorithm. As a consequence, the bright pixel weakness of dark target could apply to this method, highlighting the relevance of analyzing its performance in urban environments.

Because the dark target approximation is successful, the dark target algorithm estimating AOD, we will utilize this method in this study to adapt and evaluate the AOD-based methods with direct satellite radiance. This enables these methods in search for the best fitting method, answering research question 1.

Additionally, Landsat 8 and 9 will be used for this evaluation, since this method is derived from their bands, and data is available globally due to their polar orbit.

4.3.2 Auxiliary Data

All data that is not either ground station PM or satellite reflectance is auxiliary data. It is used to increase the prediction accuracy due to PM being influenced by a variety of factors. [63] Aerosols suspended in the atmosphere are inherently influenced by changes in the atmosphere. For example, the local mass concentration of $PM_{2.5}$ may change drastically depending on wind speed. Physical factors also play a role. The hygroscopic growth effect, for example, increases the size of aerosols depending on relative humidity.

Other works, especially land-use regressions, include human density indicators as a link to primary PM emission, for example, through traffic.

A major category of auxiliary parameters used in the methods here is meteorological values. All methods employ relative humidity (RH), atmospherirc pressure (PRS), temperature (TEMP), and wind speed (WS), so they will be included in the analysis as well.

Wind direction (WD) is used in the MLP method and can give insight into time-dependent analysis and the movement of aerosols.

Planetary Boundary Layer Height (PBLH) is an important indicator, because it limits the vertical column in which PM moves. If the PBLH decreases, it concentrates the aerosols in a smaller vertical column, and we expect to increase the concentration near the ground as well. However, since this study analyses urban environments and the requirement for specialized monitoring equipment, a locally differentiated PBLH map may not be available.

Station altitude and Digital Elevation Map (DEM) are, in this case, the same. Digital Elevation Map are maps of the topology height of the surface, ignoring infrastructure. If it is only extracted over ground stations, it offers the same contribution as the stations' altitude.

Normalized Difference Vegetation Index (NDVI) is used by both the MLP and PCA-GRNN methods and will be used as well, since the dark target algorithm estimates AOD over vegetated areas, making NDVI a sensible indicator for these areas.

Visibility, TEMP max, TEMP min, WS (sustained), and dew point from the XGBoost method will not be included, due to redundancy or poor correlation with $PM_{2.5}$ in the Spearman correlation [58].

4.3.3 Preprocessing

During preprocessing, satellite data needs to be integrated with ground stations to form a coherent dataset. The methods employed have their root in prior research. Data integration can be traced back to the original MODIS/MISR AOD validation with AERONET ground station AOD [41][8]. For example, in the data integration for the multiple linear regression models by Liu et al. in 2006 [41], MISR/MODIS AOD and meteorological data were available in a comparatively coarse 17.6km, 10km and 40km grids. Meteorological data is integrated with ground stations by assigning the pixel values to ground stations contained within the pixels. The integration of AOD followed the pattern of previous AOD validation studies, averaging all satellite AOD pixels in a 30km radius around the ground station.

The approach of averaging satellite data in a radius around ground stations has been utilized by Zhang et al. with a radius of 5km [61] and 15m by Zang et al.[59], while Zamani Joharestani et al. did not mention the integration process [58]. The temporal resolution is related to the satellite revisit time, with the exception of the geo-stationary satellite used in the PCA-GRNN method. Station and satellite measurements taken within an

interval of ± 30 minutes are considered coincident and grouped together.

To establish a common integration method, the difference in search radii has to be addressed. The size of the radius trades between the specificity and generalizability of the model. A larger radius increases the chance of finding at least a single viable pixel in the radius if clouds are present. However, since this thesis focuses on high-resolution urban environments, a high search radius would invalidate the representativeness of the high resolution.

Therefore, a fitting radius needs to be established during the analysis and prior to the main quantitative analysis. Once a fitting radius is found, data integration can be used for the preprocessing of the main analysis.

4.3.4 Assigning Seasons

Season assignment depends on the study locality. Due to the limited samples, it is desirable to have seasons evenly distributed in time. This is a soft requirement for the study area, since the seasons themselves determine the atmospheric effects we want to examine. It does, however, require the study area to have measurable seasonal changes.

4.3.5 Standardization

Since MLPs are vulnerable against feature scaling, these methods will require standardization. We will adapt the standard scaling normalization from Zamani Joharestani et al. [58]. The formula can be seen in equation 4.1, where z_i is the i-th sample of scaled feature, x_i the i-th unscaled feature and $\overline{x}$, δ_x the mean and deviation of feature x

$$z_i = \frac{x_i - \overline{x}}{\delta_x} \tag{4.1}$$

Summary

This section highlights the importance of verifying the performance of prominent methods to estimate $PM_{2.5}$ in high-resolution and urban environments. Four methods are selected for deeper analysis, and their method are explored in detail. A unified method based on the dark target algorithm is derived, and a requirement for an analysis of both AOD- and satellite images-based methods by replacing AOD with satellite data. The following section implements these requirements and transforms them into a framework for an analysis, including data sources, transformations, and model estimations.

5 Evaluation Design

In the previous chapter, a unified method was derived to estimate PM from satellite images directly. An approach to replace AOD with raw satellite data was introduced, and the link to the dark target algorithm was established. Further, the parameters used in the different methods were discussed and condensed to a shared dataset requirement. Preprocessing steps and data integration were compared, and the question was posted at which radius pixel around a station should be averaged to a measurement at that station.

This chapter introduces the evaluation methodology of the selected models using the unified method. First, a study area for the analysis is selected, with respect to the seasonal and environmental requirements highlighted in chapter 4. Depending on the location, available data sources are introduced, and the parameters for the analysis are listed. Omitted parameters are discussed as well.

Secondly, the system implementation for data extraction, processing, and storage is presented. Methods of extraction are listed in broad detail and reference the code in the supplemental material when appropriate. Data storage and integration of satellite and ground stations is described in detail.

Thirdly, the analysis itself is introduced. Quantitative metrics for the evaluation are presented. Seasonal and spatial assignment methods are introduced here as well.

Lastly, the model implementation is discussed.

5.1 Study Area

The city of Hamburg is Germany's third largest city and contains a population of around 1.9 million citizen[23]. Housing structure varies from dense multi-stories closer to the city center to single-family homes towards the outside. Laying at a natural inland harbor, the port of Hamburg is Germany's largest and third largest port in Europe. The river Elbe splits the city center to the north and the main industrial areas, including the port, to the south. The highest natural elevation is the hill Hasselbrack[28], located at 116m to the southwest, in one of many parks and nature reserves in the city.

With the Hamburger Luftmessnetz (HaLM), or air-measuring network of Hamburg, it possesses a collection of 15 measurement stations capable of detecting PM_{10} and $PM_{2.5}$.

We selected this study area due to its size, diverse environment, relatively flat geography, and data collection availability.

5.1.1 Bounding Box

Hamburg's territory includes the islands Neuwerk, Schahörn, and Nigehörn in the Wadden Sea, part of the North Sea. Due to their negligible population and distance of $\approx 100km$ from the city center, they are excluded from this study. For this reason, we base the bounding box for this analysis on the 1988 geological map[22] and select it as $(53.393 - 53.747°\text{E}, 9.662 - 10.365°\text{N})$.

5.1.2 Time Frame

Since satellite measurements coincident with ground stations can be sparse if the satellites are not geostationary, the time frame for this study will be set from 2010-01-01 to 2025-01-01, to ensure enough data can be gathered.

5.1.3 Data Sources

Satellite Images

Since Zhang et al. [61], with their MLP method, utilized the Landsat 8 OLI[51] system for their study, it will be used here as well.

Landsat 8 and 9 both carry the OLI instrument in sun-synchronous orbit. This means that the satellite passes over the equator at the same local time. The spatial resolution of 30 meters is in contrast to the revisit time of 16 days. Landsat 8 and 9 have an 8-day offset, the revisit time of *any* OLI measurement is much shorter. A benefit of using Landsat in this analysis is that no approximation of band frequency is required, since the bands from the MLP method can be used directly. Furthermore, the Landsat website published a formula to calculate Landsat NDVI directly from the bands[37].

Landsat data is available through the Earth Explorer website[1] or the M2M machine interface[2] .

Generally, the data is separated into scene tiles, identified by partially overlapping paths and row indices. The scenes contain the band rasters and auxiliary data, such as qa pixel rasters containing information about cloud cover. Before release, the raw satellite footage is processed by the USGS. In this analysis, Collection 2, Level 1, Tier 1 tiles from Landsat 8 and 9 are used. Collection 2 describes the major release of the processing suite, fixing issues present in Collection 1, which was used by Zhang et al. The tile level describes the processing applied to it. At Level 1, the tiles are geometrically corrected and radiometrically calibrated. Importantly, level 1 products still contain the top-of-atmosphere reflectance, which contains the atmospheric noise introduced by aerosols. Level 2 products contain surface reflectance bands, which have already been corrected. Lastly, the tier system describes the data quality and is separated into real-time, Tier 1, and Tier 2. Real-time data is available for Landsat 8, before the scenes are processed to

[1] `https://earthexplorer.usgs.gov`
[2] `https://m2m.cr.usgs.gov/`

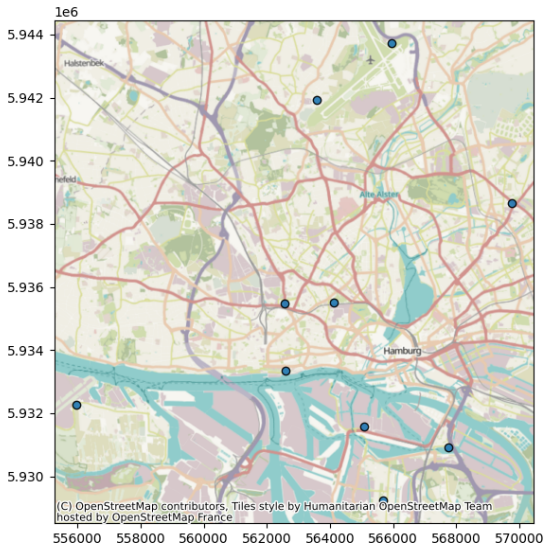

Figure 5.1: Spatial distribution of $PM_{2.5}$ stations in Hamburg.

either Tier 1 or Tier 2. Tier 1 data has to fulfill quality conditions in order to be suitable for analysis. Tier 2 contains data that does not meet the criteria.

Particulate Matter Ground Stations

As mentioned in section 5.1, the city of Hamburg operates its own atmospheric measurement network.

The HaLM network[24] consists of 16 stations, monitoring atmospheric pollution and meteorological parameters throughout the city. Pollutants include PM_{10} and $PM_{2.5}$, but also sulfur dioxide (SO_2), nitric oxide (NO), nitrogen dioxide (NO_2), Ozone (O_3), carbon monoxide (CO) and lead (Pb).

This thesis will focus on $PM_{2.5}$ pollution due to its effects on human health and anthropogenic sources. Not all stations measure all atmospheric components. Out of the 16 stations, 12 stations measure $PM_{2.5}$.

For the public, the $PM_{2.5}$ data is available as hourly means through the website[24] and an OGC API[26]. The sampling method can be found on the project website[25]. Figure 5.1 shows the distribution of ground stations in relation to the city. The dots mark the location of $PM_{2.5}$ capable HaLM ground stations.

Meteorological Ground Stations

Meteorological parameters are available from HaLM as well. Measured meteorological parameters are temperature, relative humidity, wind speed, wind direction, air pressure, precipitation, and solar irradiation [43]. These are available in hourly means as well.

Out of the 16 total stations, only 5 collected meteorological parameters. For this reason, the meteorological stations are supplemented with meteorological stations from DWD.

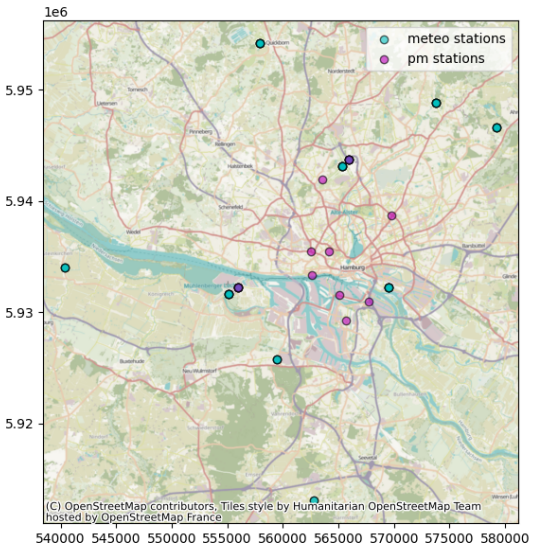

Figure 5.2: Spatial distribution of ground stations in Hamburg.

The Deutscher Wetterdienst (DWD) is a public institution under the German Federal Ministry of Transport[55]. It provides meteorological and climatological services for the entire Federal Republic of Germany, not just the city of Hamburg. As such, they operate multiple measurement stations throughout the nation, with six stations falling within the study area. For this thesis, we use this datasource to supplement the sparse meteorological measurements of the HaLM network. Data is requested through the climate data center portal at `https://cdc.dwd.de/portal/`, providing easy station selection and filtering. Similar to the HaLM network, temperature[13], relative humidity[14], wind speed[11], wind direction[15] and air pressure[12] datasets are requested from the DWD in hourly means. Purple stations are HaLM stations, that measure both.

Figure 5.2 shows the spatial distribution of ground stations used in the analysis, after processing has concluded. Meteorological ground stations have a cyan fill color, *PM* stations are magenta.

The implementation of supplementation is described in section 5.2.2 as part of the data integration.

Planetary Boundary Layer Height and Digital Elevation Maps

In the unified methods, PBLH and DEM are included in the auxiliary data. With the selected study area, these parameters will be omitted. At the location, PBLH is only available from a single measurement device and would thus be the same for every station. Digital elevation maps represent the topology of the ground, not any infrastructure built on top of it. Due to the relatively flat terrain, with the highest hill at just 116m, DEMs were omitted as well.

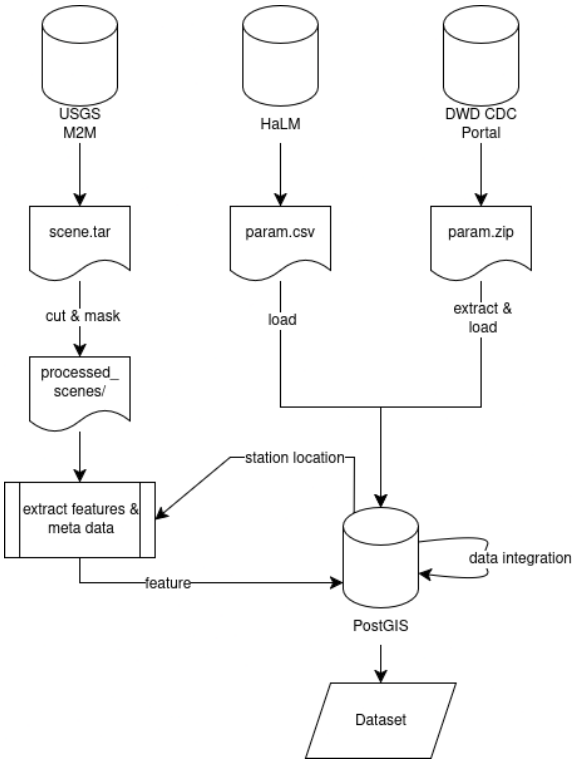

Figure 5.3: Ingestion System Overview.

5.2 Data Ingestion Design

This section provides an overview of the implementation of the ingestion pipeline used to extract the satellite and ground station features, store and preprocess them. Implementation decisions and utilized libraries are discussed at a high level. For the code, please refer to the supplemental materials.

Figure 5.3 shows an overview of the ingestion process.

First, data for $PM_{2.5}$ ground stations is loaded, metadata extracted, and ingested into the database. The same is done for meteorological ground stations. Landsat scenes are pulled as tar balls, containing all bands for a given scene. After the scenes have been reduced to the study and clouds removed, the $PM_{2.5}$ ground station positions from the database are used to extract the features from the raster. The extracted features are stored in the database, where they are integrated with the station recordings and presented as a ready-to-use dataset for estimations.

The following subsections provide more details on each step in the ingestion chain. The following section presents the database model and how the data is integrated into a dataset.

5.2.1 Extraction

To use data, it has to be extracted first. For this thesis, data ingestion was performed mostly with Python scripts, due to the simple application and quick iteration of the lan-

guage. If no Python script was used, the data was downloaded by hand. To download data, the third-party library `requests` has been used synchronously. The instances of the `session` class allow storing of credentials and managing a TCP connection pool. This is helpful in the respectful use of server resources when requesting large amounts of data. The download functions also allow the streaming of files, helping with large downloads like the Landsat rasters.

Additionally, the `dotenv` library was used to load secrets.

Wherever applicable, data was requested for the date range 2010-01-01 until 2025-01-01, i.e., 15 years. This ensures all historic data that may coincide with Landsat measurements can be retrieved, even though Landsat 8 was launched in 2013.

HaLM

While there is an OGC API provided by the Urban Data Platform Hamburg[26] for the HaLM data, it can also be downloaded via the HaLM homepage. The download URL `https://hamburg.luftmessnetz.de/core/pm25.csv?` did not link to a static file stored on a server but is instead a query containing the parameters set in the website's form element. This allows querying for data and downloading it as a CSV file directly. The full URL with query parameters can be found in the file `pull_halm.py` in the supplemental materials.

The endpoint is used to query for $PM_{2.5}$ hourly averages of each year. The meteorological parameters were requested in a similar, but required shorthands found on `https://luft.hamburg.de/meteorologie`.

The metadata was downloaded on 2025-07-13 from the Metaver collection[27], linking to the url `https://geodienste.hamburg.de/download?url=https://geodienste.hamburg.de/HH_WFS_Luftmessnetz&f=json`. It contains a `GeoJSON` feature collection of all stations, which is not available through the OGC endpoint. The meta `JSON` can be found in the supplemental materials at `jsons/app_luftmessnetz_messwerte_EPSG_4326.j`

Loading of measurements is performed in `load_halm.py`.

The station Marckmannstraße (41MM) was not available in the meta file, but available on the HaLM website. It was added manually later on. Stations Billbrook and Bramfeld were not available in the metadata file or website and were excluded from the dataset.

DWD

The DWD data has been downloaded by hand, after specifying the target area bounding box, time frame, and selecting the available datasets on `https://cdc.dwd.de/portal/`. After the request is specified, a link to the resource is received via email after a short while. The downloaded archive contains licensing agreements, request metadata, and the requested data as CSV files. The CSV files are separated by measurement data and station metadata. Similar to the HaLM data, this data is ingested via `load_dwd.py` and `load_dwd_meta.py`, which can be found in the supplemental data.

Landsat

The code for pulling Landsat data can be found in the `pull_landsat.py` in the supplemental materials. It is based on the band download scripting example provided by USGS[17]. Landsat data is pulled via the M2M proxy service[3] interface, which allows pulling large amounts of data. The M2M service functions as a proxy service, providing functionality to query dataset metadata, to search and build lists of scenes of interest, and finally submitting the scene list for a download request. The requested download is then passed to systems in the back, and the M2M service provides a download URL as soon as it is ready.

For this thesis, a scene list is built for the `landsat_ot_c2_l1` dataset with the temporal and spatial filters listed above. Instead of singular bands, all bands for a given scene are requested scene, which are provided as a tar ball. In total, 2475 scenes were downloaded with a total storage volume of 2.6TB.

Each scene in the tar file contains bands 1-11 with the frequencies listed in 4.5, metadata, and QA files. The band files are GeoTiffs, containing the measured radiation as digital numbers (DN). Digital numbers are used for a more precise data transfer and can be scaled back to the actual radiation. This was not done in the thesis, because the variables will be standardized before the models, regardless, so the original representation is not required. Meta files have information about the acquisition itself, such as the date, product name, and correction parameters. QA files encode quality information about the retrieved reflectance, such as cloud presence. The complete list of files can be found in the data format control book[35].

Due to its size, the tars are stored in the local filesystem before they are processed further and later loaded into the database. The Coordinate Reference System (CRS) for Landsat scenes in this analysis is `ESPG:32632`, where the units of the coordinates are in meters.

Difficulties in this extraction arise due to the unclear naming of the waiting list status, making the implementation of continues after program crashes more difficult.

5.2.2 Preprocessing

This subsection explains the preprocessing applied to the Landsat tars retrieved in the previous section. In this processing step, the features will be extracted from the rasters, a step in data integration from the unified method in section 4.3.3.

The preprocessing of Landsat data is done with Python, since the data is not yet in the database. This extraction method works on both rasters (Landsat GeoTiffs) and vector data (ground station buffers), which is not a well-supported operation in common Python GIS libraries. For this reason, different libraries with overlapping capabilities are used in this step.

[3] `https://m2m.cr.usgs.gov/`

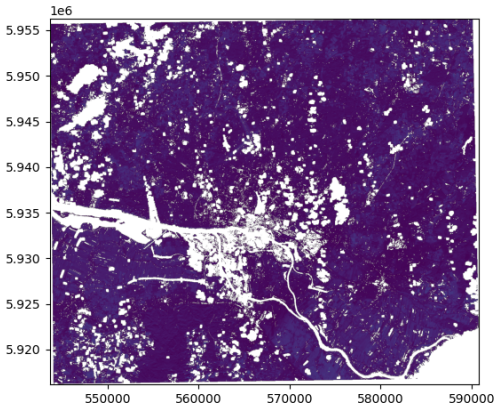

Figure 5.4: Raster after preprocessing being applied.

Rasterio is a library designed for large rasters. As such, its operations load rasters in chunks and not at once. This is necessary, because rasters can overwhelm a system's memory due to their size. Rasterio can still load in the entire raster for manipulation through numpy. Numpy is a popular library of mathematical functions, especially arrays. Pandas is a popular data processing library. Geopandas is an extension of Pandas, adding geospatial support to Pandas. Shapely is a vector library adding support for vector geometries to Python. Finally, rasterstats is a library used to calculate quantitative measures on rasters. The code can be found in `preprocess.py` in the supplemental materials.

The preprocessing consists of two major steps: Mask and crop with rasterio, and feature extraction with rasterstats.

Mask and Crop

Mask and crop prepares the scenes for in-memory processing. It extracts the scenes from the tar file and stores them in a temporary directory.

The QA_PIXEL file encodes the results of the cloud retrieval algorithms as a binary array, stored as an integer. Level 2 uses the same QA_PIXEL as Level 1 and the Level 2 data format control book gives examples for the integer values[36].

The example table states that the integer 21824 represents a clear sky with low cloud probability. This value is used as a binary filter on the QA_PIXEL and then applied as a mask to the other bands. Furthermore, the bands are cropped with the study area bounding box from section 5.1.1, drastically reducing the file size. The cropped bands and metadata files are stored in a directory for processed bands, and the temporary directory is cleared.

Figure 5.4 shows a raster after the mask and crop have been applied. The river Elbe is visible, running from west to south-east. The bright spots are caused by the cloud mask application.

Feature Extraction

In the feature extraction step, station coincident raster pixels are extracted from each band. First, indices are calculated so that they may be included as bands in the feature retrieval. The NDVI calculation follows equation 2.1. Additionally, the NDBI is calculated using equation 2.2:

As described in the preprocessing of the unified method in section 4.3.3, satellite data is extracted by averaging the retrieved pixels around a ground station and assigning the mean to that station. A concern was raised with regard to the size of the buffer, since the used radius varies greatly. Before the main analysis, this study will compare the model performance with different search radii.

To perform the extraction, the location of ground stations is loaded from the database. Next, Geopandas is used to create buffer polygons with a given radius around the ground station's coordinates. For each band, Rasterstats is used to calculate the means of the pixel values under each polygon. The resulting dataframe is unpivoted with Pandas. The stats calculated include the mean, pixel count, and count of missing values. The acceptance threshold for a feature is a mean with at least 50% of retrieved pixels. Finally, the features are pushed to the database.

5.2.3 Database

To store geo-referenced information such as station location or satellite rasters, this study uses the GIS extension PostGIS for the popular, open-source database PostgreSQL. A Geographic Information System (GIS) like PostGIS enables the storage and processing of geographic information, like satellite rasters and ground station locations. In PostgreSQL, PostGIS adds geographic information as a column type, which is stored in a separate table in the background. This way, the geographic location of a ground station can be stored in addition to regular PostgreSQL operations.

While PostGIS supports both raster and vector data and PostGIS was selected for these reasons, the raster data processor was later changed to Python, due to the interaction between raster and vector data. A major drawback of PostGIS was also solved through this change. The only official method of loading raster data into PostGIS is through the `raster2pgsql` command line interface. By processing in Python, only vector-referenced data has to be stored in the database. No raster loading is required.

Figure 5.5 shows the major transformation steps performed within the database. It is the continuation of processing where Figure 5.3 ended. The processing is organized into three distinct layers, implemented as schemas. The landing layer serves as a receptacle for the new data. Measurement tables in this layer have a `ldts` (load date timestamp) so only the most recent data is selected.

The code for this layer can be found in `sql_scripts\1_landing.sql` of the supplemental materials.

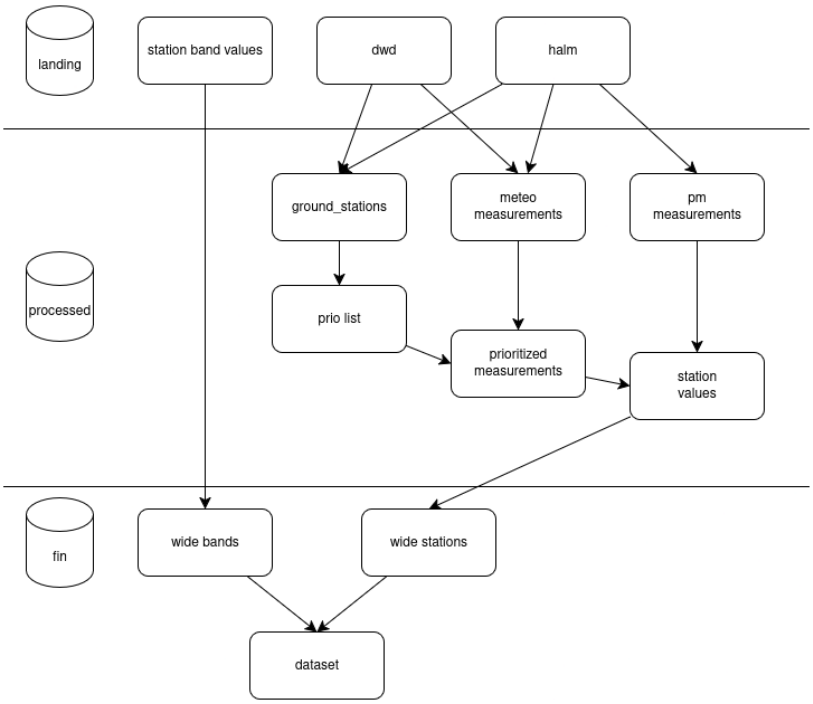

Figure 5.5: Database processing sketch, showing main transformation steps in PostGIS.

The second layer, `processed` is responsible for processing. Data from the landing schema is selected and refined into station and measurement data. While present in the DWD dataset, the atmospheric parameters for the HaLM ground stations have to be extracted first. The ground station and measured parameter, in essence, represent the measurement instrument, through which the ground stations are separated into *PM* and meteorological instruments.

As mentioned in section 5.1.3, only a few HaLM ground stations also record meteorological data and must thus be supplemented with DWD ground station. In addition, the data availability of HaLM ground stations can be low as well. This might be due to equipment maintenance or station retasking to a different area.

Since meteorological data must be included in a feature for it to be accepted, the risk of missing a retrieved feature to missing auxiliary data at specific station, must be reduced. This is achieved by creating a priority list for each *PM* ground station, weighting the distance to auxiliary instruments for each auxiliary parameter. In the best case, the distance for a given instrument is 0, because it was measured at the same ground station. If stations are missing, the first non-null value can be selected from the list of priority.

Satellite data is not processed in this layer because that processing is performed in Python, as described in section 5.2.2.

The code for this layer can be found in `sql_scripts\2_processed.sql` of the supplemental materials.

The final layer, `fin`, creates the dataset and prepares it for analysis. Band features and

station data are pivoted to wide-form and joined with the coincident method derived in section 4.3.3. Only non-null samples are considered in the join. Ground station features within ± 30 minutes of band feature acquisition are selected and averaged. The final dataset is now in the sample $\times$ parameter matrix form expected by the models.

The code for this layer can be found in `sql_scripts\3_final.sql` of the supplemental materials.

5.2.4 Dataset

The previous section laid out the transformations within the database in detail and the derivation of the final dataset was described. This section will summarize the final dataset and show which parameters will be used for the evaluation. The table `fin.dataset`, can be found as a pandas DataFrame parquet in `notebooks\dataset.parquet` in the supplemental materials 7.1.

Table 5.1 provides a summary of the dataset built in section 5.2.3, following the pattern of data overviews in chapter 4. The metadata for the ground station is static and has no temporal resolution. Historization was not implemented for metadata. The station name is used as the station identifier in the HaLM data and was kept. The Scene ID is similarly the product identifier for Landsat scenes, but encodes information about the satellite, processing tier, acquisition date, and retrieval location reference. It was added as a filtering shorthand for data analysis. Acquisition (Acq.) times represent the month and year in which the scene was retrieved from the satellite. The temporal resolution is 8\16 days due to Landsat 9 launching six years after Landsat 8, reducing the revisit time from 16 days to 8.

Table 5.2 shows the usage of the variables in the analysis. Variables marked as `input` are used as model inputs. While all bands were extracted, the selected bands for model input follow the established method in section **??**. It is important to note that while the Bands 4, 5 and 6 are used for the calculation of NDVI and NDBI indices, they are not themselves used in the analysis and thus not marked as inputs.

Variables marked as `info` provide or encode additional information that can be used for further analysis, but are not given to the models. Instead, they can be applied to filter or group the dataset in the analysis. The variable Buffer Radius has been marked as `condition`, because its effects need to be analyzed before the main analysis and then filtered to a final value, depending on the results. Finally, the $PM_{2.5}$ measurements are the dependent variable in this study and used as ground truth in the analysis.

In total, 1460 samples were extracted with the best performing radius and using the methods described above.

5.2.5 Season Assignment

As discussed in chapter 4, a seasonal assignment is required. When looking at the cities climate[23], specifically mean temperatures, months in this city can be grouped into two

Category	Variable	Unit	Temporal Resolution
Particulate Matter Ground Station Metadata	Station Name	-	-
	Buffer Radius	meters	-
	Latitude	meters	-
	Longitude	meters	-
PM measurement	$PM_{2.5}$	$\mu g/m^3$	24h moving hourly mean
Meteorological Parameters	RH	%	hourly mean
	PRS	Pa	
	TEMP	$^{\circ}C$	
	WS	m/s	
	PBLH	m	
Satellite Data	Scene ID	-	8/16 days
	Acq. Year	years	
	Acq. Month	months	
	Acq. Timestamp	milliseconds	
	Band 1	Radiance (DN)	
	Band 2		
	Band 3		
	Band 4		
	Band 5		
	Band 6		
	Band 7		
	Band 8		
	Band 9		
	Band 10		
	Band 11		
	NDVI	Index (DN)	
	NDBI		

Table 5.1: Variable overview of `fin.dataset` table in the database.

Variable	**Usage**
Station Name	info
Acquisition Timestamp	info
Scene ID	info
Buffer Radius	condition
$PM_{2.5}$	target
Latitude	input
Longitude	input
Acquisition Year	input
Acquisition Month	input
RH	input
PRS	input
TEMP	input
WS	input
PBLH	input
Band 1	input
Band 3	input
Band 7	input
NDVI	input
NDBI	input
Band 2	-
Band 4	-
Band 5	-
Band 6	-
Band 8	-
Band 9	-
Band 10	-
Band 11	-

Table 5.2: Usage of variables from the dataset.

Month	Meteorological Season	Assigned Value
January	Winter	0
February	Winter	0
March	Spring	1
April	Spring	1
May	Spring	1
June	Summer	2
July	Summer	2
August	Summer	2
September	Fall	3
October	Fall	3
November	Fall	3
December	Winter	0

Table 5.3: Meteorological Seasons in Germany[56] and their assigned value.

categories. The months June, July, and August have the highest mean temperature, whereas the months December, January, and February have the lowest. The two categories are separated by three months on both sides of a year, fulfilling the requirement for distinct seasons sufficiently. When looking at the meteorological seasons for Germany, this is the exact assignment[56]. Measurements are assigned to seasons by the month they were recorded in. The mapping can be seen in table 5.3.

5.3 Model Evaluation

This study analyses the model performance using the Python library `scikit-learn`. `scikit-learn` is a machine learning library providing methods and support for an entire estimation chain, from loading a dataset to visualizing the results. Due to the popularity of the library, many machine-learning packages provide interface functionality to integrate into `scikit-learn`'s validating procedures. The code for the evaluation itself and the model implementations can be found in the Python notebook under `notebooks/evaluation.ipynb` in appendix 7.1, supplemental materials. Plots created in this notebook or shown prior are created either with `matplotlib` directly or via another library's interface to `matplotlib`.

With regards to the small sample size, the models presented in chapter 4 will be evaluated using two-way 10-fold cross-validation [46]. In a two-way evaluation, the dataset is split into training and test splits. This way, the models are tested on unseen data and their generalizability can be evaluated. In k-fold cross-validation, the dataset is split into

k-folds, where one fold is withheld for the testing split and the rest is used for training. With this method, every fold is used as the testing split once, and the model is tested on every part of the dataset. This is important if there is a configuration of samples that the model performs well at, that might otherwise fall into the simple two-way test set randomly. For each fold, selected metrics report the performance on the given split. All metrics are then collected and expressed as the mean and standard deviation of the metric.

Cross-validation can be used for model evaluation or hyperparameter tuning. When used with hyperparameter tuning, a third validation split is taken to prevent statistical leakage into the hyperparameters. In this study, the hyperparameters are taken from the selected method, because they were already optimized for the estimation of $PM_{2.5}$ and the sample size of the extracted dataset is limited. This helps to establish the mean performance of the models on the dataset and eliminates randomly selecting good-performing samples in the test split.

In this analysis, the R^2-score, $RMSE$ and MAE are collected, following previous studies[59]. The R^2 is the coefficient of determination [9]. It shows the amount of variance the model can reproduce in the dependent variable. Values of R^2 are in the fixed range of $[0, 1]$ and are interpretable as a percentage of goodness of fit. Thus, the generalizability of different models on the same dataset can be compared. The R^2 can be negative in specific situations, such as predicting values non-linearly and worse than a least-squares method.

The MAE (Digitial Number)[42] is the mean difference between estimation and prediction using the Manhatten Distance. It measures the distance between the prediction and measured values in the same space as the dependent variable, allowing for interpretations in the same space. The $RMSE$ (Root Mean Square Error) [47] works similarly, but uses the Euclidean distance instead. This makes large errors contribute more strongly towards the resulting error, while still in the same space as the dependent variable. Both MAE and $RMSE$ allow for comparison between models, but not between variables using different scales.

Cross-validation, R^2, $RMSE$ and MAE are implemented using the `scikit-learn` suite. The implementation of $RMSE$ and MAE inverts them in the library, following the pattern of higher values representing better performance. In the code, the results for $RMSE$ and MAE are inverted back for this reason, to retrieve the actual distances.

5.4 Model implementation

As mentioned in the previous section 5.3, the evaluation is implemented using `scikit-learn`. The code for the evaluation itself and the model implementations can be found in the Python notebook under `notebooks/evaluation.ipynb` in appendix 7.1, supplemental materials. `scikit-learn` provides a pipeline object, that can chain multiple models and transformers together.

Hyperparameter	Value
n_estimators	200
max_depth	8
gamma	0.7
min_child_weight	8

Table 5.4: Hyperparameter used for XGBoost.

Following section 4.1, the input for all models will be standardized, since they are vulnerable to feature scaling and the scales between satellite data and meteorological parameters can be very different.

This is implemented using the pipeline feature and the `StandardScaler` preprocessor class from `scikit-learn`.

5.4.1 PCA-GRNN

The PCA-GRNN method described in section 4.2.2 is not a model implemented in the `scikit-learn` library. Instead, the analysis uses the compatible library pyGRNN[3], used to analyze isotropic and anisotropic GRNN implementations[4]. For this analysis, the isotropic implementation is used. The library is compatible with `scikit-learn` in that it implements the training and evaluation functions used by `scikit-learn`. This allows using the GRNN implementation like a regular `scikit-learn` model.

To implement a PCA-GRNN, a `scikit-learn` PCA model is added before the GRNN model in a `scikit-learn` pipeline.

5.4.2 XGBoost

For the implementation of XGBoost method from section 4.2.3, the open-source library XGBoost is used in Python[10]. The `scikit-learn` compatible interface of this library is the class `xgb.XGBRegressor`. The hyperparameters used are taken from XGBoost, described in section 4.2.3.

5.4.3 MLP

The shallow and deep MLP methods from sections 4.2.1 and 4.2.4 are implemented with Keras. Keras is a high-level interface for deep learning, capable of working with multiple backends like Tensorflow or PyTorch. This thesis uses torch as a backend. The functional API provides an easy method to implement deep neural networks. Keras provides an interface for `scikit-learn` as well, here it is the `keras.wrappers.SKLearnRegressor` wrapper class. This class can pass hyperparameters for the fit function, which are not

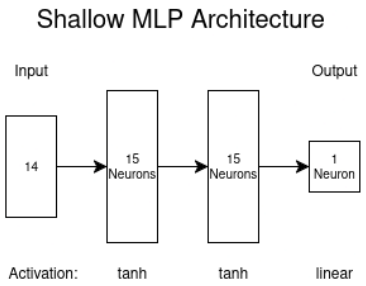

Figure 5.6: Architecture of the shallow multi-layer perceptron network.

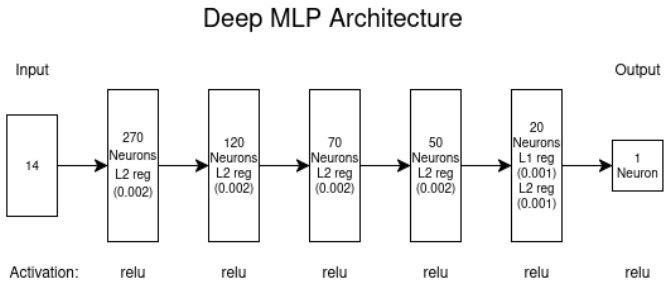

Figure 5.7: Architecture of the deep multi-layer perceptron network.

specified in the model itself. To use this wrapper, the models have to be compiled. Both the deep and shallow models use the Adam optimizer and mean squared error as the loss function. They are trained with a validation split of 10%, a batch size of 64, 1000 epochs and early stopping.

Shallow-MLP

The architecture for the shallow MLP has been taken from the shallow MLP method introduction in section 4.2.1, it can be seen in figure 5.6. The model consists of two dense hidden layers with 15 neurons each. They are activated with the tanh function, which has a sigmoidal curve. The activation function for the output neuron is linear.

Deep-MLP

The architecture for the deep MLP has been taken from the deep MLP method introduction in section 4.2.4. Figure 5.7 provides an overview of the model, which consists of 5 hidden dense layers with decreasing numbers of neurons. All dense layers use ReLU as their activation function and L2 regularization until the last hidden layer, which also uses L1 regularization.

Summary

This section shows the implementation and analysis of the unified method derived in chapter 4. The study area, data sources and their processing are discussed. The data integration is shown in the Python scene preprocessing and database station transformation. `Scikit-learn` and its validation methods and metrics are introduced and explained. Finally, the model implementation is shown, together with external libraries used. The dataset and analysis metrics are used in the following chapter to answer the research questions posed in chapter 1 and discuss the results.

6 Results

This chapter presents and discusses the results of the evaluation, using the methods designed in chapter 5. First, the station radius is analyzed and a distance is selected. Selecting the retrieval radius determines the sample size of the dataset, since it influences the retrieval performance. All experiments are evaluated using 10-fold cross-validation and reporting the mean value of the metrics achieved in each fold. After the radius selection, the models are analyzed using the dataset in three experiments. First, the overall performance and estimation ability are established on the entire dataset. Next, the dataset is grouped by the seasons, established using the seasonal assignment method from section 5.2.5. Lastly, the influence of the direct environment is established by separating the dataset into urban and rural ground stations.

6.1 Station Radii Analysis

The station radius determines the area around a ground station within which the pixels from satellite retrievals are averaged to a single value. This value is considered the satellite retrieval at that station for the used band. Changing the radius is a trade-off between the specificity and generalizability of the retrieval. Considering spotty cloud covers, it can also change the number of samples retrieved. Because the feature must be complete in order to be added to the dataset, missed retrievals must also be minimized.

To evaluate the retrieval amount, the sample counts for each radius can be compared. To evaluate retrieval quality, the retrievals will be used to train a simple multiple regression model.

The regressors will be analyzed in the same manner as the main models, cross-validation with 10 folds and an evaluation of metrics mean and standard deviations.

Table 6.1 shows the result of the radii pre-experiment. In the radius column, the selected radii are listed. They were selected as multiples of the base resolution of 30m. It is evident that the different radius sizes produce very similar results. Surprisingly, the 3000m radius produces the best R^2 with $R^2 = 0.15$, whereas the 15m radius performed the worst $R^2 = 0.11$. However, the standard deviation of the R^2 scores seen in table 6.2 from the cross-validation is significantly higher for the 3000m radius (0.19) than for all other conditions ($0.05 - 0.09$). Interestingly, the RMSE and MAE show an improvement of overall generalizability of the models with increasing radius (RMSE 15m: 6.99 $\rightarrow$ RMSE 3000m: 6.52 MAE 15m: 4.95 $\rightarrow$ MAE 3000m: 4.71). Finally, the retrieved samples show that the 15m and 3000m radii performed the worst ($N = 1217$ and $N = 1311$),

Radius in meters	Mean R^2	Mean RMSE	Mean MAE	N
15	0.116613	6.985351	4.952503	1217
30	0.140762	6.831198	4.846269	1451
60	0.137924	6.844605	4.858167	1460
90	0.130854	6.861759	4.866506	1460
120	0.132851	6.805180	4.840888	1449
150	0.128615	6.791880	4.832206	1448
300	0.130095	6.719610	4.772716	1420
3000	0.152431	6.519197	4.712831	1311

Table 6.1: Mean metrics from multiple regression cross-validation for different station radii.

Radius in meters	Std. R^2	Std. RMSE	Std. MAE
15	0.060064	0.636484	0.501074
30	0.050075	0.550093	0.448556
60	0.053795	0.590171	0.431194
90	0.067588	0.459214	0.323805
120	0.065045	0.448270	0.232827
150	0.074252	0.559189	0.228419
300	0.089616	0.630677	0.211228
3000	0.185508	0.520504	0.376303

Table 6.2: Standard deviation of metrics from multiple regression cross-validation for different station radii.

Model	R^2	$RMSE$	MAE
PCA-GRNN	0.723	3.698	2.237
XGBoost	**0.855**	**2.720**	**1.868**
Shallow MLP	-2.177	13.104	10.878
Deep MLP	0.096	6.988	5.109

Table 6.3: Results of the overall performance on the dataset with N=1460 with metric means.

when compared to 60m and 90m radii (both $N = 1460$). Except for the 15m and 3000m radius, all other radii result in very similar results. The radius of 60m (double the station radius) is selected, due to the highest sample retrieval and better model approximation when compared to 90m.

6.2 Overall Performance

Table 6.2 shows the results of testing the models on the entire dataset. The best performing scores are highlighted with bold text. On all metrics, XGBoost performed the best. This model can explain 85.5% of the variance in the dependent variable, while the errors are comparatively small. Closely following is the PCA-GRNN model with an R^2 of 0.723 and still minor errors.

Surprisingly, both MLP models performed worse, with a negative R^2 for the shallow MLP and 0.096 for the deep model. In all metrics, the shallow MLP is worse than the deep MLP. It appears that there has been a learning effect, but with an R^2 approaching 0, the deep model is just slightly better than static guessing.

It is possible that the models are overfitting due to the small size of the dataset. It might be possible to improve the performance of the MLP models by reducing the auxiliary features or increasing the sample size.

6.3 Temporal Analysis

This section analyses the effects of seasonal atmospheric changes by applying the models on each season separately. The seasonal split method from section 5.2.5 is used to assign a season to the samples. Table 6.4 shows the results of this experiment. The left side of the table shows the conditions, with the metrics' mean of the 10-fold cross-validation to the right. Best performing methods per category are highlighted in bold.

In line with the results of the overall analysis in section 6.2, the metrics returned for the XGBoost model are the best. In addition, the MLP models are performing worse in comparison. Of note is the performance of the PCA-GRNN method. In Winter, the R^2

Season	N	**Model**	R^2	$RMSE$	MAE
Winter	168	PCA-GRNN	0.204	8.109	5.242
		XGBoost	**0.878**	**3.329**	**2.225**
		Shallow MLP	-1.762	18.224	14.403
		Deep MLP	-1.575	17.660	13.704
Spring	436	PCA-GRNN	0.313	5.517	3.715
		XGBoost	**0.813**	**3.130**	**2.122**
		Shallow MLP	-2.500	15.093	12.721
		Deep MLP	-1.558	12.800	9.829
Summer	464	PCA-GRNN	-0.384	4.633	3.591
		XGBoost	**0.755**	**2.222**	**1.558**
		Shallow MLP	-4.322	10.367	9.301
		Deep MLP	-1.255	6.752	5.308
Fall	392	PCA-GRNN	0.257	4.511	3.038
		XGBoost	**0.835**	**2.373**	**1.711**
		Shallow MLP	-2.946	11.911	10.289
		Deep MLP	-1.532	9.547	7.600

Table 6.4: Model Performance by season with metric means.

is only 0.204, and even -0.384 in Summer. This is surprising, because the sample size in Summer ($N = 464$) is higher than in Winter ($N = 168$), yet the R^2 score is much worse. One would generally expect a better performance with more samples. However, the same appears to be true for the XGBoost model as well. In Winter, the $R^2 = 0.878$ is better than the overall performance $R^2 = 0.855$ and much better than in summer $R^2 = 0.755$. The reason for this difference in summer is perhaps the result of more noise being present in the data. In summer, the expected $PM_{2.5}$ content is much lower than in winter months, when heating is likely causing higher concentrations of $PM_{2.5}$, as seen in section **??** above. With lower concentrations in summer, it is possible that noise is influencing the errors more than in other seasons.

6.4 Spatial Analysis

The goal of this section is to analyze the influence of the direct station environment on the models. As noted in chapters 1 and 4, urban environments are likely to influence the retrieval of the dark target algorithm due to high reflectance, causing bright pixels. Furthermore, the topology of infrastructure surrounding ground stations can influence

Group	N	**Model**	R^2	$RMSE$	MAE
Urban	509	PCA-GRNN	0.052	5.779	3.763
		XGBoost	**0.646**	**3.735**	**2.399**
		Shallow MLP	-2.238	11.97	9.877
		Deep MLP	-1.192	9.534	6.940
Rural	951	PCA-GRNN	0.690	3.996	2.463
		XGBoost	**0.822**	**2.993**	**2.045**
		Shallow MLP	-2.406	13.891	11.649
		Deep MLP	-0.378	8.652	6.893

Table 6.5: Model Performance by environment with metric means.

atmospheric movement and thus the distribution of $PM_{2.5}$.

While all stations are in an urban environment, this analysis will separate $PM_{2.5}$ ground stations by the built-up area around them. The NDBI has been developed to detect land-cover changes through urbanization with an accuracy of 92.6%[60]. The NDBI is used as an abstraction of the environment, since the retrieval is performed in a radius around the ground stations. Averaging NDBI measurements on each station and locating the median of the means provides an even separator by NDBI concentrations. Stations with an NDBI below the median are considered urban, and rural if their mean is above the median.

Table 6.5 shows the means of 10-fold cross-validation when performed on the dataset separated by the urban environment detection method mentioned above. Best results are marked in bold. The trend of performance order from previous experiments repeats here as well. Results from the rural category align well with the performance on the entire dataset in section 6.2.

In the urban condition, only the XGBoost method achieved a mean R^2 of at least 0.5 with $R^2 = 0.646$ The PCA-GRNN method achieved $R^2 = 0.052$. While the R^2 scores are lower in comparison with station samples in the rural group, the errors of the XGBoost and PCA-GRNN methods do not worsen at the same rate. The XGBoost MAE is 2.045 in the rural group, but only rises to 2.399 in the urban group. Similarly, the PCA-GRNN MAE rises from 2.463 to 3.763. This indicates more noise in the data, which the model may not be able to model.

7 Conclusion

In this work, the high-resolution satellite-based PM estimation introduced by Zhang et al. [61] was related to the dark target AOD algorithm[1] and extrapolated into a unified dataset, evaluating AOD based algorithms with satellite data instead. The analysis was performed in the city of Hamburg, due to the known issues of the dark target algorithm in urban environments.

The accuracy of the models has been shown in the overall evaluation 6.2, showing that both XGBoost and PCA-GRNN perform well, with an explained variance ($R^2 = 0.855$, $R^2 = 0.723$) and errors $< 4\mu g/m^3$ aerosol mass ($MAE = 1.868$, $MAE = 2.237$). The MLP models were not able to adequately estimate $PM_{2.5}$. This result is in line with the findings by Nguyen et al. [45], where machine learning algorithms are capable of $PM_{2.5}$ estimation even on small sample sizes, where deep learning methods struggle.

The evaluation in section 6.3 shows a difference in retrieval availability based on the season, with significantly lower retrievals in Winter (XGBoost: $R^2_{\text{Summer}} = 0.755$, $R^2_{\text{Winter}} = 0.878$). This contributes to the fair-weather requirement of satellite estimations. However, while the sample size is much lower in Winter, the uncertainty, as expressed by the explained variance, is also lower when compared to summer, while the errors are higher. This is due to higher $PM_{2.5}$ concentrations during Winter, likely to higher anthropogenic emissions and the boundary layer height effect described by Koelemeijer et al. [34].

Finally, the effect of the urban environment was explored in the spatial analysis in section 6.4. The ground stations were separated into Urban and Rural ground stations using the median of station mean NDBI. In the urban condition, only the XGBoost method achieved a positive result ($R^2_{\text{Urban}} = 0.646$, $R^2_{\text{Rural}} = 0.822$). This suggests that more research is required into the features needed to estimate $PM_{2.5}$ in built-up high-resolution urban environments.

In summary, the change from AOD to satellite data based estimation in urban environments is successful for XGBoost and PCA-GRNN, but not for the MLP methods.

7.1 Future Work

Due to the data availability in the study area, the sample size remains small even with the supplementary actions taken in chapter 5 and the large temporal window of requested satellite data. A path for future work is increasing the ground station availability by selecting a better study area with better official measurement stations, or including public ground stations from citizen science projects like Sensor.Community.

Another approach is to increase the availability of satellite data instead, for example, by applying the downscaling method by Bechtel et al.[5] to geostationary satellites and estimating more high-resolution satellite images. This way, there would be a higher chance of a cloud-free retrieval and more satellite data available for coincident data integration.

Lastly, this approach is based on the dark target AOD retrieval method. It is surprising that XGBoost performed well. However, dark target is a very early algorithm and future work can instead attempt to model more modern algorithms, such as deep blue, time sequence, or even LaSRC itself [63][54].

Finally, XGBoost and PCA-GRNN can be used to calculate $PM_{2.5}$ on the entire satellite scene, given a mapping of ground stations for auxiliary data. Such a mapping can, for example, be determined through a Voronoi map.

Bibliography

[1] *Aerosol - LAADS DAAC*. URL: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/science-domain/aerosol/#modis (visited on 09/28/2025).

[2] *Aerosol Optical Depth*. URL: https://earthobservatory.nasa.gov/global-maps/MODAL2_M_AER_OD (visited on 01/15/2025).

[3] Federico Amato. *Federhub/pyGRNN*. Aug. 1, 2025. URL: https://github.com/federhub/pyGRNN (visited on 09/14/2025).

[4] Federico Amato et al. *On Feature Selection Using Anisotropic General Regression Neural Network*. Oct. 12, 2020. DOI: 10.48550/arXiv.2010.05744. arXiv: 2010.05744 [stat]. Pre-published.

[5] Benjamin Bechtel, Klemen Zakšek, and Gholamali Hoshyaripour. "Downscaling Land Surface Temperature in an Urban Area: A Case Study for Hamburg, Germany". In: *Remote Sensing 2012, Vol. 4, Pages 3184-3200* 4.10 (Oct. 19, 2012), pp. 3184–3200. ISSN: 2072-4292. DOI: 10.3390/RS4103184.

[6] Michelle A. Bouchard. *Example of Landsat Collection 2 Surface Reflectance | U.S. Geological Survey*. Oct. 6, 2010. URL: https://www.usgs.gov/media/images/example-landsat-collection-2-surface-reflectance-0 (visited on 09/11/2025).

[7] Bert Brunekreef and Stephen T. Holgate. "Air Pollution and Health". In: *The Lancet* 360.9341 (Oct. 19, 2002), pp. 1233–1242. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(02)11274-8. PMID: 12401268.

[8] D. A. Chu et al. "Validation of MODIS Aerosol Optical Depth Retrieval over Land". In: *Geophysical Research Letters* 29.12 (2002), MOD2-1-MOD2–4. ISSN: 1944-8007. DOI: 10.1029/2001GL013205.

[9] *Coefficient of Determination*. In: *Wikipedia*. Oct. 6, 2025. URL: https://en.wikipedia.org/w/index.php?title=Coefficient_of_determination&oldid=1315466760 (visited on 10/12/2025).

[10] *Dmlc/Xgboost*. Distributed (Deep) Machine Learning Community, Oct. 12, 2025. URL: https://github.com/dmlc/xgboost (visited on 10/12/2025).

[11] *DWD Climate Data Center (CDC): Hourly Mean of Station Observations of Wind Speed ca. 10 m above Ground in m/s for Germany, Version V21.3, Last Accessed: 2025-09-14.*

[12] *DWD Climate Data Center (CDC): Hourly Station Observations of Air Pressure at Station Level in hPa for Germany, Version V21.3, Last Accessed: 2025-09-14.*

[13] *DWD Climate Data Center (CDC): Hourly Station Observations of Air Temperature at 2 m above Ground in °C for Germany, Version V21.3, Last Accessed: 2025-09-14.*

[14] *DWD Climate Data Center (CDC): Hourly Station Observations of Relative Humidity in % for Germany, Version V21.3, Last Accessed: 2025-09-14.*

[15] *DWD Climate Data Center (CDC): Hourly Station Observations of Wind Direction ca. 10 m above Ground in Degree for Germany, Version V21.3, Last Accessed: 2025-09-14.*

[16] Shane B Eisenman et al. "BikeNet: A Mobile Sensing System for Cyclist Experience Mapping". In: *ACM Transactions on Sensor Networks (TOSN)* 6.1 (2007), p. 39. ISSN: 15504859. DOI: 10.1145/1653760.1653766.

[17] *EROS User Services / Machine to Machine (M2M) / M2M Landsat Bands Bundle and Band Groups Download · GitLab*. GitLab. Sept. 9, 2025. URL: https://code.usgs.gov/eros-user-services/machine_to_machine/m2m_landsat_bands_bundle_download (visited on 10/05/2025).

[18] *EU Air Quality Standards - European Commission*. URL: https://environment.ec.europa.eu/topics/air/air-quality/eu-air-quality-standards_en (visited on 03/19/2025).

[19] *Europe's Air Quality Status 2023*. [Publications Office of the European Union], 2023. DOI: 10.2800/59526.

[20] V. M. Fernández-Pacheco et al. "Estimation of PM10 Distribution Using Landsat5 and Landsat8 Remote Sensing". In: *Proceedings 2018, Vol. 2, Page 1430* 2.23 (Oct. 31, 2018), p. 1430. ISSN: 2504-3900. DOI: 10.3390/PROCEEDINGS2231430.

[21] Bijoy Krishna Gayen et al. "Estimation of High-Resolution Aerosol Optical Depth (AOD) from Landsat and Sentinel Images Using SEMARA Model over Selected Locations in South Asia". In: *Atmospheric Research* 298 (Mar. 1, 2024), p. 107141. ISSN: 0169-8095. DOI: 10.1016/J.ATMOSRES.2023.107141.

[22] *Geologische Karte 1:50 000 Hamburg - MetaVer*. URL: https://metaver.de/trefferanzeige?docuuid=A02341C6-1C34-11D4-B517-0060086B14D3&f=x1%3A9.6295%3By1%3A53.3161%3Bx2%3A10.4205%3By2%3A53.8379%3Boptions%3Ainside%3B (visited on 06/26/2025).

[23] *Hamburg*. In: *Wikipedia*. Sept. 4, 2025. URL: https://en.wikipedia.org/w/index.php?title=Hamburg&oldid=1309549370#Geography (visited on 09/10/2025).

[24] *Hamburger Luftmessnetz*. URL: https://luft.hamburg.de/ (visited on 03/19/2025).

[25] *Hamburger Luftmessnetz – FHH*. URL: https://luft.hamburg.de/allgemeine-informationen/messverfahren-775712 (visited on 10/05/2025).

[26] *Hamburger Luftmessnetz (HaLm)*. URL: https://api.hamburg.de/datasets/v1/luftmessnetz/collections (visited on 10/03/2025).

[27] *Hamburger Luftmessnetz (HaLm) | MetaVer*. URL: https://metaver.de/trefferanzeige?docuuid=EFCA6E2A-1D5B-408A-BBB1-C8AC41A3E5F7#detail_links (visited on 10/05/2025).

[28] *Hasselbrack*. In: *Wikipedia*. July 8, 2024. URL: https://de.wikipedia.org/w/index.php?title=Hasselbrack&oldid=246565026 (visited on 10/02/2025).

[29] Brent N. Holben et al. "AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization". In: *Remote Sensing of Environment* 66.1 (Oct. 1, 1998), pp. 1–16. ISSN: 0034-4257. DOI: 10.1016/S0034-4257(98)00031-5.

[30] N.C. Hsu et al. "Deep Blue Retrievals of Asian Aerosol Properties During ACE-Asia". In: *IEEE Transactions on Geoscience and Remote Sensing* 44.11 (Nov. 2006), pp. 3180–3195. ISSN: 1558-0644. DOI: 10.1109/TGRS.2006.879540.

[31] John R. Jensen. *Introductory Digital Image Processing: A Remote Sensing Perspective*. 4th ed. Pearson Series in Geographic Information scienceAlways Learning. Glenview, Ill.: Pearson Education, 2016 [publ. 2015]. xxxi+623. ISBN: 978-0-13-405816-0.

[32] Sami Kaivonen and Edith C.H. Ngai. "Real-Time Air Pollution Monitoring with Sensors on City Bus". In: *Digital Communications and Networks* 6.1 (Feb. 1, 2020), pp. 23–30. ISSN: 2352-8648. DOI: 10.1016/J.DCAN.2019.03.003.

[33] Federico Karagulian et al. "Contributions to Cities' Ambient Particulate Matter (PM): A Systematic Review of Local Source Contributions at Global Level". In: *Atmospheric Environment* 120 (Nov. 1, 2015), pp. 475–483. ISSN: 1352-2310. DOI: 10.1016/J.ATMOSENV.2015.08.087.

[34] R. B. A. Koelemeijer, C. D. Homan, and J. Matthijsen. "Comparison of Spatial and Temporal Variations of Aerosol Optical Thickness and Particulate Matter over Europe". In: *Atmospheric Environment* 40.27 (Sept. 1, 2006), pp. 5304–5315. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2006.04.044.

[35] *Landsat 8-9 OLI/TIRS Collection 2 Level 1 Data Format Control Book | U.S. Geological Survey*. Mar. 25, 2025. URL: https://www.usgs.gov/media/files/landsat-8-9-olitirs-collection-2-level-1-data-format-control-book (visited on 09/25/2025).

[36] *Landsat 8-9 OLI/TIRS Collection 2 Level 2 Data Format Control Book | U.S. Geological Survey*. July 21, 2023. URL: https://www.usgs.gov/media/files/landsat-8-9-olitirs-collection-2-level-2-data-format-control-book (visited on 10/05/2025).

[37] *Landsat Normalized Difference Vegetation Index | U.S. Geological Survey*. URL: https://www.usgs.gov/landsat-missions/landsat-normalized-difference-vegetation-index (visited on 09/14/2025).

[38] R. C. Levy et al. “Global Evaluation of the Collection 5 MODIS Dark-Target Aerosol Products over Land”. In: *Atmospheric Chemistry and Physics* 10.21 (Nov. 5, 2010), pp. 10399–10420. ISSN: 1680-7316. DOI: 10.5194/acp-10-10399-2010.

[39] Zhongbin Li et al. “Evaluation of Landsat-8 and Sentinel-2A Aerosol Optical Depth Retrievals across Chinese Cities and Implications for Medium Spatial Resolution Urban Aerosol Monitoring”. In: *Remote sensing* 11.2 (Jan. 1, 2019), 10.3390/rs11020122. ISSN: 20724292. DOI: 10.3390/RS11020122. PMID: 32021701.

[40] Tianchen Liang et al. “Estimation of Aerosol Optical Depth at 30 m Resolution Using Landsat Imagery and Machine Learning”. In: *Remote Sensing* 14.5 (Mar. 1, 2022), p. 1053. ISSN: 20724292. DOI: 10.3390/RS14051053/S1.

[41] Yang Liu et al. “Using Aerosol Optical Thickness to Predict Ground-Level PM2.5 Concentrations in the St. Louis Area: A Comparison between MISR and MODIS”. In: *Remote Sensing of Environment*. Multi-Angle Imaging SpectroRadiometer (MISR) Special Issue 107.1 (Mar. 15, 2007), pp. 33–44. ISSN: 0034-4257. DOI: 10.1016/j.rse.2006.05.022.

[42] *Mean Absolute Error*. In: *Wikipedia*. Feb. 16, 2025. URL: https://en.wikipedia.org/w/index.php?title=Mean_absolute_error&oldid=1276071917 (visited on 10/12/2025).

[43] *Messstationen*. URL: https://luft.hamburg.de/luftmessstationen-hamburg (visited on 10/03/2025).

[44] *MODIS Web*. URL: https://modis.gsfc.nasa.gov/about/ (visited on 02/14/2025).

[45] Phuong D.M. Nguyen et al. “Mapping of High-Resolution Daily Particulate Matter (PM2.5) Concentration at the City Level through a Machine Learning-Based Downscaling Approach”. In: *Environmental Monitoring and Assessment* 197.1 (Jan. 1, 2025), pp. 1–22. ISSN: 15732959. DOI: 10.1007/S10661-024-13562-6/FIGURES/10.

[46] Sebastian Raschka. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. Nov. 11, 2020. DOI: 10.48550/arXiv.1811.12808. arXiv: 1811.12808 [cs]. Pre-published.

[47] *Root Mean Square Deviation*. In: *Wikipedia*. Oct. 9, 2025. URL: https://en.wikipedia.org/w/index.php?title=Root_mean_square_deviation&oldid=1315956734 (visited on 10/12/2025).

[48] *Sensor.Community: Build Your DIY Sensor and Become Part of the Worldwide Citizen Science, Open Data, Civic Tech Network. Supported by a Lot of Contributors*. URL: https://sensor.community/en/ (visited on 03/19/2025).

[49] Donald F. Specht. “A General Regression Neural Network”. In: *IEEE transactions on neural networks* 2.6 (1991), pp. 568–576. URL: http://www.inf.ufrgs.br/~engel/data/media/file/cmp121/GRNN.pdf (visited on 09/20/2025).

[50] Amy Thai, Ian McKendry, and Michael Brauer. "Particulate Matter Exposure along Designated Bicycle Routes in Vancouver, British Columbia". In: *Science of The Total Environment* 405.1–3 (Nov. 1, 2008), pp. 26–35. ISSN: 0048-9697. DOI: 10.1016/J.SCITOTENV.2008.06.035. PMID: 18701140.

[51] *USGS EROS Archive - Landsat Archives - Landsat 8-9 Operational Land Imager and Thermal Infrared Sensor Collection 2 Level-1 Data | U.S. Geological Survey*. Nov. 27, 2020. URL: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-8-9-operational-land-imager-and (visited on 10/05/2025).

[52] Aaron van Donkelaar et al. "Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application". In: *Environmental Health Perspectives* 118.6 (June 2010), pp. 847–855. ISSN: 00916765. DOI: 10.1289/EHP.0901623/SUPPL_FILE/0901623ART_SUPPL.PDF. PMID: 20519161.

[53] E. Vermote et al. "LaSRC (Land Surface Reflectance Code): Overview, Application and Validation Using MODIS, VIIRS, LANDSAT and Sentinel 2 Data's". In: Institute of Electrical and Electronics Engineers. New York, NY, United States, July 27, 2018. URL: https://ntrs.nasa.gov/citations/20190001670 (visited on 09/11/2025).

[54] E. Vermote et al. "LaSRC (Land Surface Reflectance Code): Overview, Application and Validation Using MODIS, VIIRS, LANDSAT and Sentinel 2 Data's". In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia: IEEE, July 2018, pp. 8173–8176. ISBN: 978-1-5386-7150-4. DOI: 10.1109/IGARSS.2018.8517622.

[55] *Wetter Und Klima - Deutscher Wetterdienst - About Us*. URL: https://www.dwd.de/EN/aboutus/aboutus_node.html (visited on 10/04/2025).

[56] *Wetter Und Klima - Deutscher Wetterdienst - Glossar - J - Jahreszeiten*. URL: https://www.dwd.de/DE/service/lexikon/Functions/glossar.html;jsessionid=BCFC2DA08872F2CDD49F1E501886AE89.live11053?lv2=101304&lv3=101324 (visited on 09/25/2025).

[57] *Wetter Und Klima - Deutscher Wetterdienst - Messung Des Stadtklimas*. URL: https://www.dwd.de/DE/klimaumwelt/klimaforschung/klimawirk/stadtpl/messung_stadtklima/messung_stadtklima_node.html (visited on 03/19/2025).

[58] Mehdi Zamani Joharestani et al. "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data". In: *Atmosphere* 10.7 (July 2019), p. 373. ISSN: 2073-4433. DOI: 10.3390/atmos10070373.

[59] Lin Zang et al. “Estimating Hourly PM1 Concentrations from Himawari-8 Aerosol Optical Depth in China”. In: *Environmental Pollution* 241 (Oct. 1, 2018), pp. 654–663. ISSN: 0269-7491. DOI: 10.1016/J.ENVPOL.2018.05.100. PMID: 29902748.

[60] Y. Zha, J. Gao, and S. Ni. “Use of Normalized Difference Built-up Index in Automatically Mapping Urban Areas from TM Imagery”. In: *International Journal of Remote Sensing* 24.3 (Jan. 1, 2003), pp. 583–594. ISSN: 0143-1161. DOI: 10.1080/01431160304987.

[61] Bo Zhang et al. “Estimation of PMx Concentrations from Landsat 8 OLI Images Based on a Multilayer Perceptron Neural Network”. In: *Remote Sensing 2019, Vol. 11, Page 646* 11.6 (Mar. 16, 2019), p. 646. ISSN: 2072-4292. DOI: 10.3390/RS11060646.

[62] Renyi Zhang et al. “Formation of Urban Fine Particulate Matter”. In: *Chemical Reviews* 115.10 (May 27, 2015), pp. 3803–3855. ISSN: 15206890. DOI: 10.1021/ACS.CHEMREV.5B00067/ASSET/IMAGES/MEDIUM/CR-2015-00067K_0032.GIF.

[63] Ying Zhang et al. “Satellite Remote Sensing of Atmospheric Particulate Matter Mass Concentration: Advances, Challenges, and Perspectives”. In: *Fundamental Research* 1.3 (May 1, 2021), pp. 240–258. ISSN: 2667-3258. DOI: 10.1016/J.FMRE.2021.04.007.

Supplemental Materials

The referenced code and files can be found in this GitHub project archive: `https://github.com/joshuaschimmel/spacedust`.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe. Sofern im Zuge der Erstellung der vorliegenden Abschlussarbeit generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten Wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, den 11.10.2025 Unterschrift [redacted]

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den 11.10.2025 Unterschrift [redacted]