Enhancing emission-oriented Land-Use-Regression data with Vision-Language-Models

– implementation for NO_x in The Netherlands

Master Thesis Business Analytics

Quirinus de Ruijter

Vrije Universiteit Amsterdam - Caeli B.V.

Enhancing emission-oriented Land-Use-Regression data with Vision-Language-Models

— implementation for NO_x in The Netherlands

by

Quirinus de Ruijter

to obtain the degree of Master of Science
at the Vrije Universiteit Amsterdam

Supervisor Caeli	Tim Spijkerman
Supervisor VU Amsterdam	prof. dr. Mark Hoogendoorn
Second reader VU Amsterdam	dr. Joost Berkhout
Additional reader Caeli	Martin Smit
Duration	February - August, 2025
Faculty	Faculty of Science, Vrije Universiteit Amsterdam
Student ID	2709397

Caeli B.V.	https://caeli.nl/
Cover	Copernicus Sentinel data (2017), processed by ESA, modified
Version	1.0 printed on 29th August 2025
Disclaimer	no form of Generative AI was used for writing this thesis report
Template	by Daan Zwaneveld, licensed under CC BY-NC 4.0

Abstract

Keywords : *Land Use Regression, Vision Language Model, Emission, Nitrogenoxides*

Air-polluting emissions negatively impact health and the environment, with NO_x being a significant contributor. This research explores a novel approach of enhancing Land Use Regression (LUR) data by using a Vision-Language-Model (VLM). For modelling NO_x emission data a set of 42.323 1 km^2 squares in the Netherlands is used, emissions (kg/km^2/year) are estimated by the RIVM. The VLM (Google's Gemini Flash 2.0) reclassifies OpenStreetMap obtained 'landuse:industrial' parcels — four times with different prompt settings — to more detailed landuse by analysing aerial images. Additionally, the VLM returns a yearly emission estimate for the parcel which is added as a new feature. An Elastic Net, Random Forest (RF), and Histogram-based Gradient Boosting Regression (HGBR) are fitted to the four adapted datasets and compared with performance on the benchmark dataset.

The results show improvements in two of the four enhanced datasets compared to the benchmark dataset evaluated with RMSE. On the full test-set, the benchmark-HGBR combination results in an RMSE of 18.318 kg, and the best scoring enhanced-HGBR combination results in an RMSE of 17.762 kg, achieving a 3,0% decrease. When only considering the 1.127 'enhanced' 1 km^2 squares of the test-set, an improvement is made from 48.138 kg RMSE to 48.094 kg RMSE, resulting in a 0,1% decrease in RMSE loss. On the other 'unenhanced' 1km^2 squares, an improvement is made from 4.949 kg RMSE to 4.684 kg RMSE, resulting in a 5,4% decrease in RMSE loss. Due to the wide confidence intervals of RMSE, statistical significances on these numbers could not be claimed.

The study highlights challenges in modelling high emission values but demonstrates that VLM-enhanced open-source data can introduce meaningful improvements in predictive performance. Although RMSE improvements are modest and not statistically conclusive, other metrics suggest that the approach has practical potential to strengthen LUR modelling of emissions.

Contents

Nomenclature

Recurring Abbreviations

Abbreviation	Definition
AQG	Air Quality Guideline
CBS	Centraal Bureau voor de Statistiek
E-PRTR	European - Pollutant Release and Transfer Register
HGBR	Histogram-based Gradient Boosting Regression
LUR	Land Use Regression
MMMU	Massive Multi-discipline Multimodal Understanding
NO	Nitrogenoxide
NO_2	Nitrogendioxide
NO_x	Nitrogen Oxides
O_3	Ozone
PM	Particulate Matter
RF	Random Forest
RIVM	Rijks Instituut voor Volksgezondheid en Milieu
VLM	Vision-Language Model

List of Figures

List of Tables

1 Introduction

It is known that release of NO_x (nitrogen-oxides) has negative impact on air quality. A group of researchers from the WHO [1] state that, after adjustment for particulate matter (PM), long-term exposures to NO_2 are associated with; respiratory and cardiovascular mortality, and with childrens respiratory symptoms and lung functioning. In addition to the impact on health, the impact on biodiversity through nitrogen deposition and acid rain is even more extensive according to the research by Clark [2]. The loss of biodiversity has immense and unseen effects on the circle of life according to consensus studies by Cardinale [3] and Tilman [4], justifying raise of alarms.

There have been investigations into the sources of NO_x emissions for a long time. Figure 1.1 displays the distribution of emissions per category in The Netherlands for the year 2022. The dependency on combustion engines remains a large source, with *traffic, transport* and *inland navigation* accounting for more than 65% of the yearly NO_x emissions. Additionally, the *energy* an *industry and waste* sectors account for another 24% of NO_x, again due to their use of fossil fuels required to create energy and products.

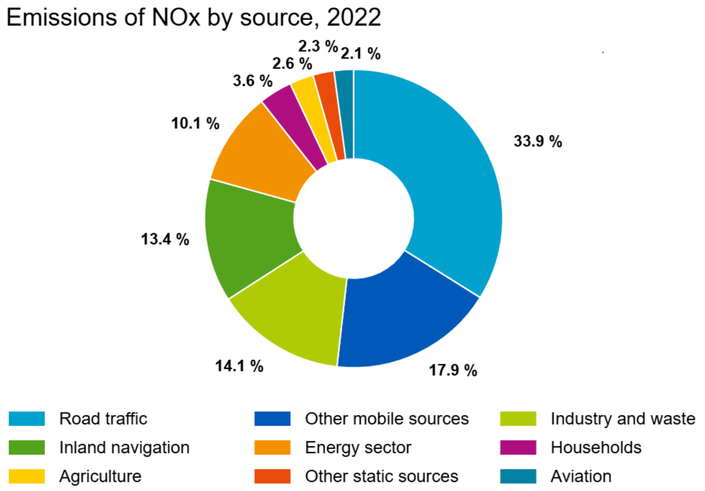


Figure 1.1: Emissions of NO_x in the Netherlands in 2022 by source. Road traffic and mobile sources account for 51.8% and are by far the largest contributor to NO_x emissions. The industry sector (industry, waste, and energy) accounts for 24.2% of emissions. Source: CBS [5]

RIVM[1] has been constructing insights in air quality since 1990 by measuring matters such as NO_x, SO_x, PM (sulphur-oxides, particulate-matter). Their models are based on detailed layers of estimated emissions and annual environmental reports of businesses. Estimations are based on the principle of Emission = Activity × Emission Factor, followed by allocation to sector and geographical location. For example, the amount of fuel types used in combination with scientifically derived emission factors is converted into an amount of emission per company sector[6]. Or emissions from manure management based on cattle livestock counts and emission factors by the National Emission Model for Agriculture (NEMA)[7]. The RIVM combines these layers

[1] Rijksinstituut voor Volksgezondheid & Milieu, National Institute for Public Health and the Environment

and many more by mathematical models which result in detailed insights to specific data-slices aggregated per year to; sectors, provinces, municipalities, and even square kilometres [8].

National institutions like the RIVM provide insights of high quality, however, their frequency of reporting is annually and lagged by two years. At the moment of writing, February 2025, the emissions available for inspection only include the years up to 2022. Governmental institutions responsible for installing regulations that result in emissions reduction will require insight into what actions to take against which costs for what benefits. Yearly reports are not enough to investigate effects of these measures taken, and therefore the frequency of reporting should be increased and the lag should be minimised.

It is thought that by incorporating near-live data from satellites, measurement stations, traffic movements, and energy use in machine learning models, contrary to the method presented by the RIVM, the frequency and lag of emission reports could be bettered. This is what Caeli accomplishes by moving to a data-driven approach. Additionally, more advanced methods could be used across broader areas, crossing national borders, and thus research areas of national institutes.

Hoek [9] conducted a review of Land-Use-Regression (LUR) models constructed between 1993 and 2008 assessing air pollution, of which 21 are about NO_2 concentrations. The typical model is built on only a few weeks of measurements from monitoring sites. Spatial variations are recorded over time and an estimation model is built to capture these patterns. Models achieve an R^2 of 0.60.8 on validation sets, depending on factors such as research size, location, and model complexity. Relevant predictors include *traffic* intensity, distances to *roads*, *population-* and *housing density*, *land-cover* factors, and *topographical* data. The sources of these features are often experimental based —traffic counts to estimate traffic density— or obtained from open source databases. Thus, while LUR models have proven effective, they reach a limit due to the availability and precision of their input, and are not scalable to other regions without performing experiments.

One of the largest open source databases about "*roads, trails, cafés, railway stations, and much more, all over the world*" is OpenStreetMap (OSM), where a community of 'mappers' is constantly improving a twin model of the world. One of the ways that OSM lacks precision is when an object is given the tag "*landuse:industry*" solely, as it does not differentiate between industry types. This omission is significant because emission levels vary drastically between sectors. For example, the largest energy industry facilities can emit up to 18 million kg of NO_x annually [10], whereas warehouses themselves emit almost none. When these facilities are only represented in the "*landuse:industry*" area feature, the difference in order of magnitude could not be captured by the models.

To be less dependent on others and allow for automation, a Vision Language Model (VLM) could be used to make adaptations to the LUR data. The hypothesis is that VLMs are capable of analysing aerial images and create output of several features linked to the area, and thereby improve quality of the dataset. Scientifically, this approach holds the potential to enhance the precision of LUR modelling by including diverse environmental, visual, and documental data, which can easily be adapted for other pollutants such as NH_3 or PM, and scaled to different geographic regions. For Caeli, the application of this methodology can markedly improve the accuracy of their emission and air quality models.

In short, the goal of this thesis is to contribute to the scientific frontier of LUR modelling and thereby help Caeli improve their activities. Therefore, this report investigates **to what extent Vision-Language-Model-derived features can complement traditional Land Use Regression datasets for modelling NO_x emissions in the Netherlands**. The research will begin with additional information in Chapter 2. The data sources resulting in a benchmark dataset are presented in Chapter 3. Followed by training several LUR models on this data in Chapter 4. An investigation will be held on making the best use of a VLM to enhance the benchmark dataset in Chapter 5. This will result in the fitting of new LUR models on the newly created datasets in Chapter 5, eventually comparing the results in Chapter 6.

2 Additional Insights

In this chapter, additional information is presented; what are nitrogen-oxides, current emission levels and targets, and the relevant literature on LUR modelling and the use of VLMs.

2.1. What is the 'Matter'?

Nitrogen-dioxide, with the chemical formula NO_2, is an inorganic compound of nitrogen and oxygen. The pure substance occurs as a poisonous reddish-brown gas, which is highly soluble in water. It forms nitric acids. The gas is a strong oxidiser, heavier than air, and reacts violently with other substances, such as metals.

Approximately 78% of the air we breathe is composed of nitrogen (N_2) while only 21% is made up of oxygen (O_2), both crucial for sustaining life as we know it. It is the burning of fossil fuels at high temperatures and pressures that allows NO to form from N_2. First, nitrogen reacts with oxygen through $N_2 + O \Leftrightarrow NO + N$ and $N + O_2 \Leftrightarrow NO + O$, known as the *Zeldovich* reactions. Since NO is an unstable molecule at room temperature and atmospheric pressure, it oxidises with ozone to form NO_2 by $NO + O_3 \Leftrightarrow NO_2 + O_2$, with a rate depending on factors such as sunlight, temperature, and humidity as stated by Walters [11] and Beychok [12]. When measuring a source that emits NO into fresh air, first around 10% of nitrogen is in the form of NO_2 which increases to 90% in a timespan of minutes or a few hours, measured by Wild [13]. Over more hours, this NO_2 reacts to other products such as O_3 (ozone), HNO_3 (nitric-acid) and others, disappearing from the scene. NO and NO_2 together are notated as NO_x and named *nitrogen oxide.*

It is crucial to emphasise the difference between *emission* and *concentration. Emission* is a measure of mass accumulated over time, whilst *concentration* is the presence of matter and thereby a measure of mass per volume at a certain moment in time. Emissions undergo dispersion because of meteorological effects and/or the movement of sources, which makes the analysis of the relation between concentration measurements and emission sources complex. This will require dispersion modelling with calibrations by measurements of ground stations, sound balloons, satellites, etc., which is outside the scope of this research, only emission is considered.

2.2. Guidelines

The raise of alarms has resulted in actions by means of laws and policies since the World Health Organisation (WHO) has set guidelines of limitations on particulate matter ($PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and carbon monoxide (CO) in 1987[14]. In 2005 and 2021 there have been global updates of these guidelines (see Table 2.1) which stimulated authorities to intensify efforts to study and restrict harmful emissions. The levels set are ought not to be harmful to human health and natural ecosystems.

The European Union has set up a Zero Pollution Action Plan for 2050, in which it demands their member states to comply with the WHO 2021 guidelines on air quality by 2030[15]. The research of Zara [16] has shown already massive reductions in the annual mean NO_x emissions of -35% from 2005 up to 2018 throughout Europe, with high reductions in high-emission zones the Po valley, Ruhr Area, and BeNeLux.

Figure 2.1 provides a closer understanding of the *concentration* numbers in the Netherlands. The daily average NO_2 concentration over 2023 is 9-11 mug/m^3 in regional areas, 16-18 mug/m^3 in urban areas, and 20-22

Pollutant	Averaging Time	2005 AQGs	2021 AQGs	Unit
$PM_{2.5}$	annual	10	5	µg/m^3
	24-hour^a	25	15	µg/m^3
PM_{10}	annual	20	15	µg/m^3
	24-hour^a	50	45	µg/m^3
O_3	peak season^b	-	60	µg/m^3
	8-hour^a	100	100	µg/m^3
NO_2	annual	40	10	µg/m^3
	24-hour^a	-	25	µg/m^3
SO_2	24-hour^a	20	40	µg/m^3
CO	24-hour^a	-	4	µg/m^3

Table 2.1: Air Quality Guidelines of the WHO per emission type of 2005 and 2021. ^{a}: 99th percentile (i.e. 34 exceedance-days per year), ^{b}: average of daily maximum 8-hour mean O_3 concentration in the six consecutive months with the highest six-month running-average O_3 concentration.

mug/m^3 in traffic dense areas. Since 2017 there has been no exceedance of the EU guidelines. The numbers are derived from the Luchtmeetnet initiative [17] where the measurements of validated ground stations are published near real-time.

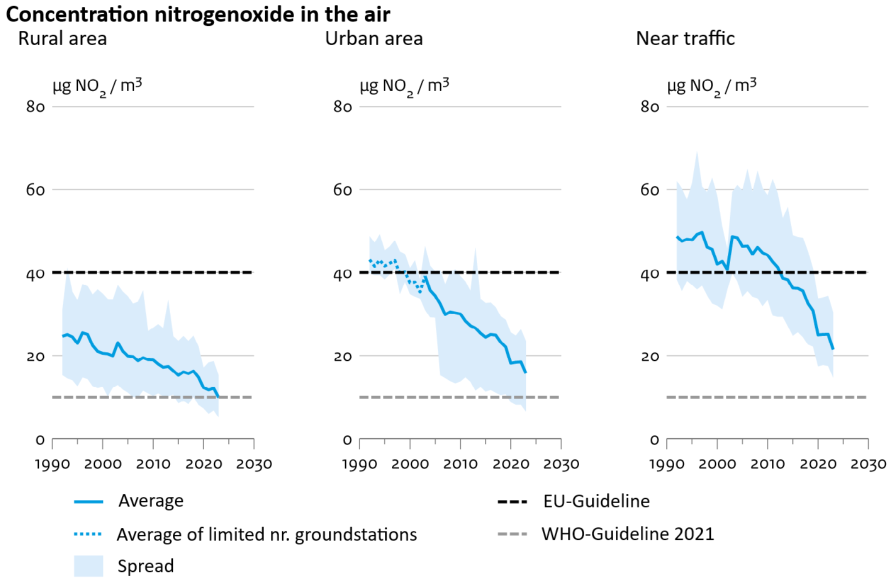


Figure 2.1: Yearly averages of daily NO_2 µg/m^3 in the Netherlands in regional, urban and traffic dense locations from 1993-2023. Since 2017 there has been no exceedance of the European Union guidelines of 40 µg/m^3. The trend reaches towards WHO AQGs of 10 µg/m^3 in 2030. Source: RIVM[18]

2.3. Relevant Literature

2.3.1. About Land Use Regression

There have been methods to estimate *emissions* for a certain area. These methods start with registering emissions by stations or samples for a certain period to obtain averages of emissions near highways, urban areas, industry sites and more. Then these measurements are predicted using Land Use Regression (LUR) models, which learn dependencies on geographical features such as proximity to highways, population density, and industrial areas, as shown in Figure 2.2. When model performance is deemed satisfactory, it can be used for the estimation of emissions in areas where no measurements are taken.

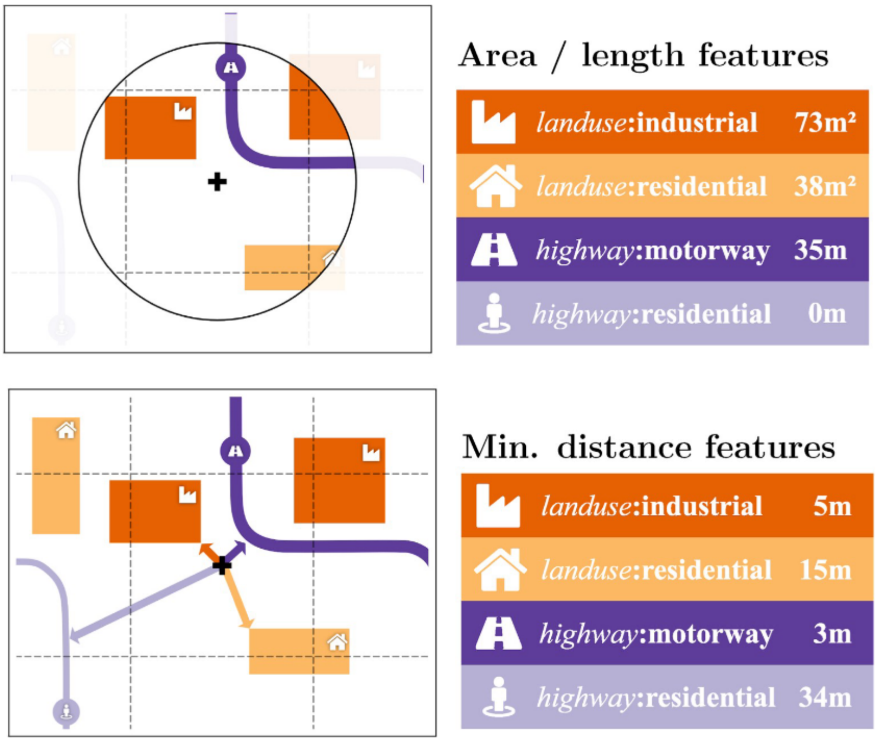


Figure 2.2: A schematic illustration of how eight different geographical features on a location, of which four areal and four distance-based features. The features are used to predict some emission or concentration in Land Use Regression models. From:[19]

2.3.2. Development and Usage of Vision-Language Models

The research of Radford [20] ’Contrastive Language-Image Pre-training’ (CLIP) combined text- and image-encoders with Large Language Models (LLMs) to bridge the gap between textual and visual data comprehension. Unlike traditional image classification datasets that only predict on a set of predefined labels, such as CIFAR-10/100[21], the model builds a conceptual understanding of the 400 million image-text pairs and thus also describes images on unseen topics, known as *zero-shot learning*.

Improvements in the area of interest— counting objects and analysing of satellite images— have been made by the research of Liu [22] with ’RemoteCLIP’. Their model outperforms the CLIP baseline by 10% on average on 9 out of 12 *downstream datasets* for remote sensing, datasets containing images obtained by satellites or unmanned aerial vehicles (UAVs). To further improve on their model, they state that larger models, larger datasets, and higher data quality are issues to address.

Ever since, there have been rapid developments in Vision Language Models (VMLs) and new usecases were unlocked. For example, Pan [23] uses VLMs to extract information from satellite and street-view images for housing-attributes in a real estate setting. Or the research of Roberts [24] classifies aerial images to be used for detection of deforestation or land-use planning. A different approach is presented by Steininger [25], which uses aerial images directly to build a deep-learning model for estimating air pollution on that scene.

For this research, the model used should be capable of analysing aerial clip-shots of industrial areas. Therefore, it is preferable that the model excels in object detection from real-world satellite-like images, counting of objects, reading text from POIs, whilst not losing its knowledge on gauging emissions.

One of the features relevant for the estimation of NO$_x$ emission estimation is population density [9]. The feature is, in combination with other LUR features, an expression of traffic intensity, traffic load, transportation of goods, energy production, commercial activity, and more.

3
Data Insights

Now that the context of LUR modelling and emissions has been set, a description of several datasets forming the *benchmark* dataset is presented.

3.1. Emission data by the RIVM

It is decided to train the LUR models on the RIVM emission data of 2022 spanning the entirety of the Netherlands with a $1km^2$ resolution. The data are considered of high quality due to its build-up from measurements, scientific research and environmental annual reports. An estimate on annual total NO_x emissions with Monte Carlo simulations results in a 95% confidence interval relative to the mean of 19% [26]. Thus, when total emissions are estimated at 100, the actual emissions are likely to be in the range of 90,5 and 109,5.

3.1.1. Creators of the data

The RIVM publishes registrations on 375 different emissions, such as NO_x, from Dutch sources of each year, which is commissioned by the ministries of I&W[1], KGG[2], and LVVN[3]. The collaboration involves CBS[4], PBL[5], WUR [6], and Deltares[7], led by the RIVM [8]. Besides these, different institutions provide other relevant information, such as Rijkswaterstaat[8] on traffic and the road network. The involvement of all these contractors and institutes highlights both the importance and complexity of this matter.

The construction of the result involves a total of seven task forces, each responsible for a specific sector: LULUCF[9], ENINA[10], VenV[11], MEWAT[12],WESP[13], and TRV[14].[27] The task-forces are each responsible for:

- Calculating emissions using the best available methodologies based on the results of scientific research.
- Determining necessary methodological changes based on new (inter)national scientific insights.
- Quality control of data in the relevant work fields.
- Jointly approving the data under the responsibility of the task-force chairperson.
- Annually updating methodology reports.
- Defining necessary research to maintain and/or improve the quality of emission calculations.

[1] Infrastructuur en Waterstaat, Infrastructure and Water Management
[2] Klimaat en Groene Groei, Climate Policy and Green Growth
[3] Landbouw, Visserij, Voedselzekerheid en Natuur; Agriculture, Fisheries, Food Security, and Nature
[4] Centraal Bureau voor de Statistiek, Statistics Netherlands
[5] Planbureau voor de Leefomgeving, Environmental Assessment Agency
[6] Wageningen University & Research
[7] Dutch knowledge institute for water and the subsurface
[8] Executive agency of I&W
[9] Land Use, Land Use Change, and Forestry
[10] Energie, Industrie en Afval; Energy, Industry, and Waste
[11] Verkeer en Vervoer, Traffic and Transport
[12] Methodeontwikkeling Emissies Water, Method Development for Water Emissions
[13] Werkgroep Servicebedrijven en Productgebruik, Working Group Service Companies and Product Use
[14] Taakgroep Ruimtelijke Verdeling, Spatial Distribution

3.1.2. Construction of the Data

Each year, the emissions of year 1990 up to year t-2 are presented based on the most up to date knowledge on models and data. One of the reasons of this delay is caused by the requirement of large companies to report their digital Environmental Annual Report (elektronisch Mulieujaarverlag, eMJV), which is part of the European-PRTR registration[10]. These statements are gathered and audited by the RIVM and used in the construction of annual emission data. The threshold for companies having to report their emissions has been set such that approximately 85%-90% of emissions from companies should be captured by the selected group. The remaining percentages for companies are estimated by scientific methods based on activity data and emission factors.

Besides the emissions from these companies, the task-forces estimate emissions from other sources. This is an accumulation of very detailed emission sources that range from energy usage of passenger and railway traffic, to airport movements, to cold engine starts, to waste incineration, to inland shipping and wood burning[28]. Here, not all aspects are covered, but one example will be given to provide an insight in the methodology that the task-forces apply. One of the determining factors of NO_x is combustion engines running on fossil fuels. The RIVM distinguishes between traffic of passenger cars, public transport, construction machinery and more[29]. The passenger cars traffic itself, is divided into six categories based on maximum speed and traffic type which is obtained by the NWB[15]. The traffic intensity is obtained from the DatMobility, which itself is based on the road network in combination with the research of ODiN by CBS taking into account the population density and locations of employment. This results in a *distribution grid* for the specific categories, for which the total NO_x emission estimations will be allocated to, of which two out of six categories are shown in Figure 3.1. All these distribution grids combined with emission factors of fuel consumption per car type and car usage (derived by TNO[16] [30]) result in a final emission layer of NO_x from the road network annually.

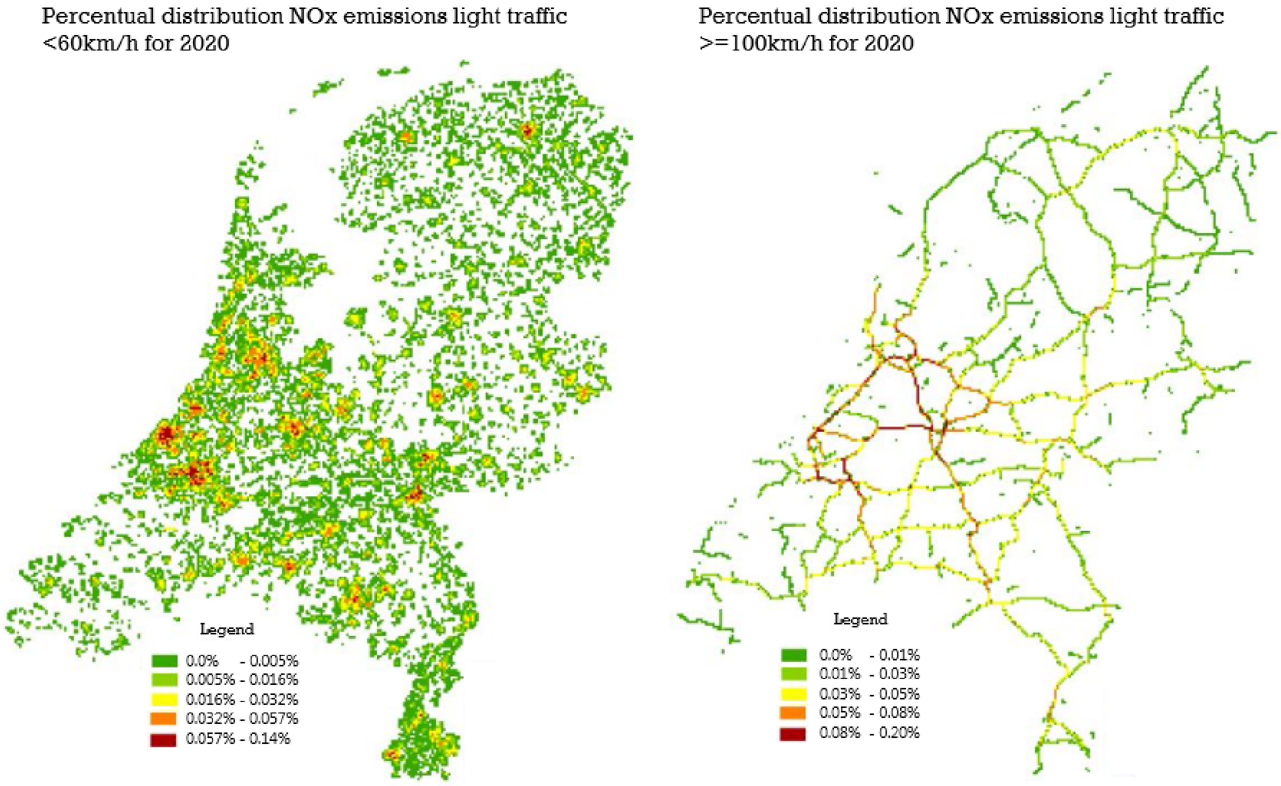


Figure 3.1: Two distributions keys on NO_x emissions per $1km^2$ for the **light traffic** category on roads with <60km/h and on roads for >=100km/h, constructed by the RIVM methodology. Source:[29]

[15] Nationaal Wegenbestand, National Road-network-reference

[16] Nederlandse organisatie voor Toegepast-Natuurwetenschappelijk Onderzoek, Dutch Organisation of Applied-Scientific Research

This example only provides a sketch on one of the layers constructed by a single task force, which again highlights the complexity and level of detail that is captured in the data. Eventually, all the results of these layers are gathered and reported in a document as well as an accessible dashboard[8].

The highest resolution of emission data is by square kilometre and fits to RD-coordinates[17], formalised by the NCG[18]. In this coordinate reference system (CRS, EPSG:28992) units are steps of 1 km, which makes the Netherlands range from 0 to 280 on x-coordinates, and from 300 to 625 in y-coordinates. Including the Exclusive Economic Zone (EEZ), the coordinates range from -40 to 280 and 300 to 860 for x- and y-coordinates respectively.

3.1.3. Analysis of the RIVM Emission Data

As of February 2025, the most recent data of emissions in The Netherlands with a resolution of 1km^2 is on the year 2022. Since the data are massively rightly skewed, a $\log_{10}$ transformation of the emission data provides better visualisations of the data in Figure 3.2 and Figure 3.3.

A summary of the data is shown in Table 3.1. Because later in the research the scope is narrowed down to industrial facilities, it decided to only train models on the land data-points and discard the sea subset.

Set		Count	Mean	Std.	Min	25%	50%	75%	Max
Total	NO_x (kg/km^2/y)	99.698	2.993	18.087	0	159	1.002	2.809	2.686.139
	$\log_{10}(NO_x)$	99.698	2,85	0,80	0	2,20	3,00	3,45	6,43
Land	NO_x (kg/km^2/y)	42.323	5.182	27.479	0	1.140	2.069	3.813	2.686.139
	$\log_{10}(NO_x)$	42.323	3,23	0,72	0	3,06	3,32	3,58	6,43
Sea	NO_x (kg/km^2/y)	57.375	1.379	2.303	0	105	249	1.600	20.155
	$\log_{10}(NO_x)$	57.375	2,57	0,74	0	2,03	2,40	3,20	4,30

Table 3.1: Summary statistics for NO_x and $\log_{10}(NO_x)$ emissions for total coverage and separated into *land* an *sea* surface subsets. Percentage columns denote quantile values.

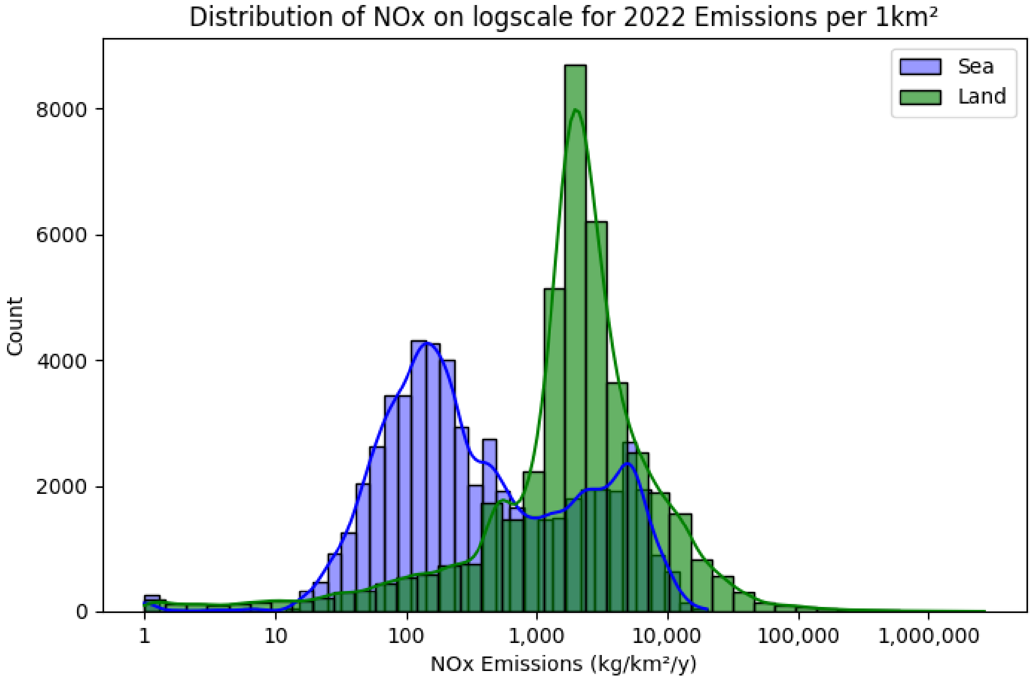


Figure 3.2: The distribution of logged NO_x emissions of 2022 reveals a difference between *land* and *sea* emissions per 1km^2 squares.

[17] Rijksdriehoekscoordinaten, National Triangulation Coordinates

[18] Nederlands Centrum voor Geodesie en Geo-Informatica, Netherlands Centre for Geodesy and Geo-Information

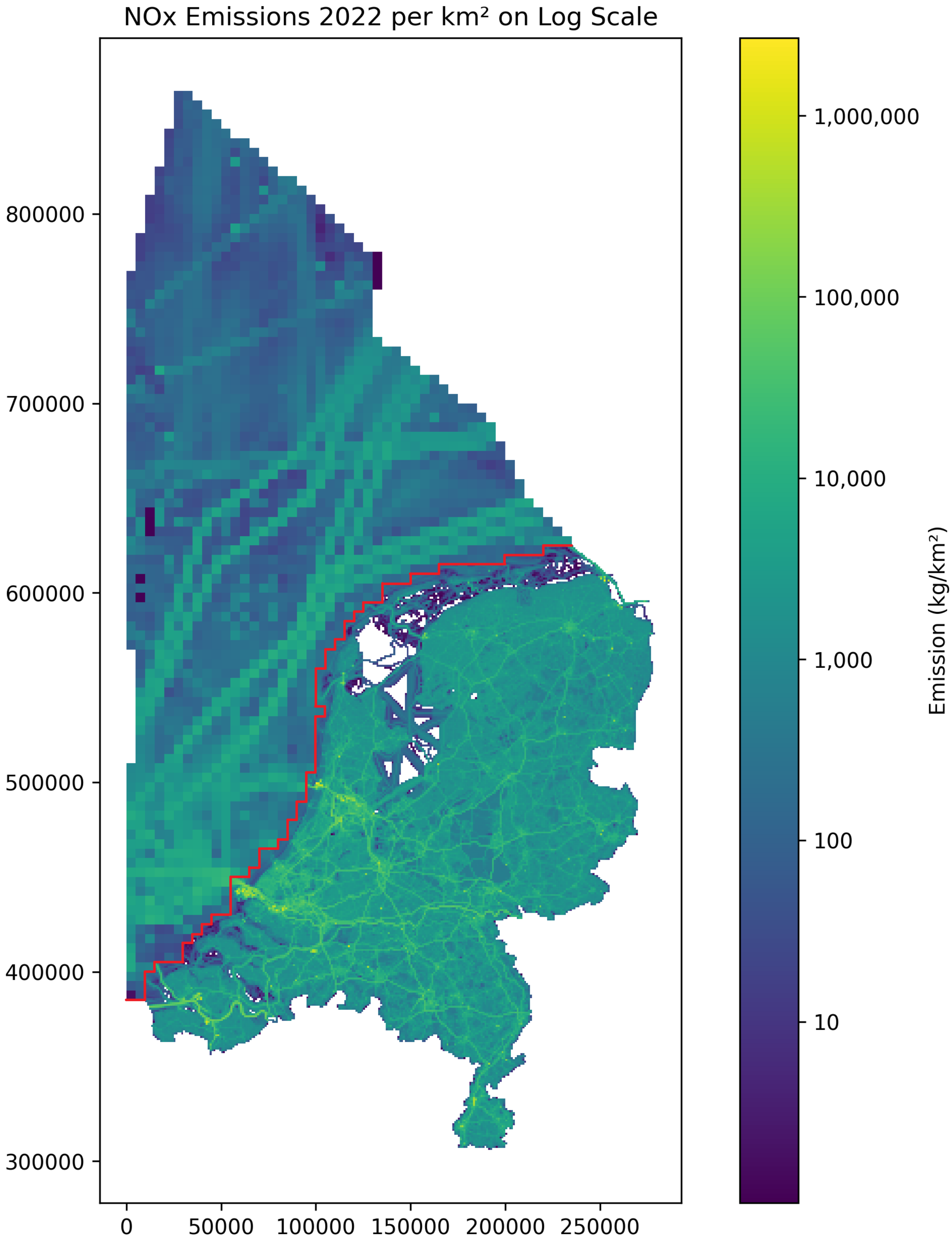


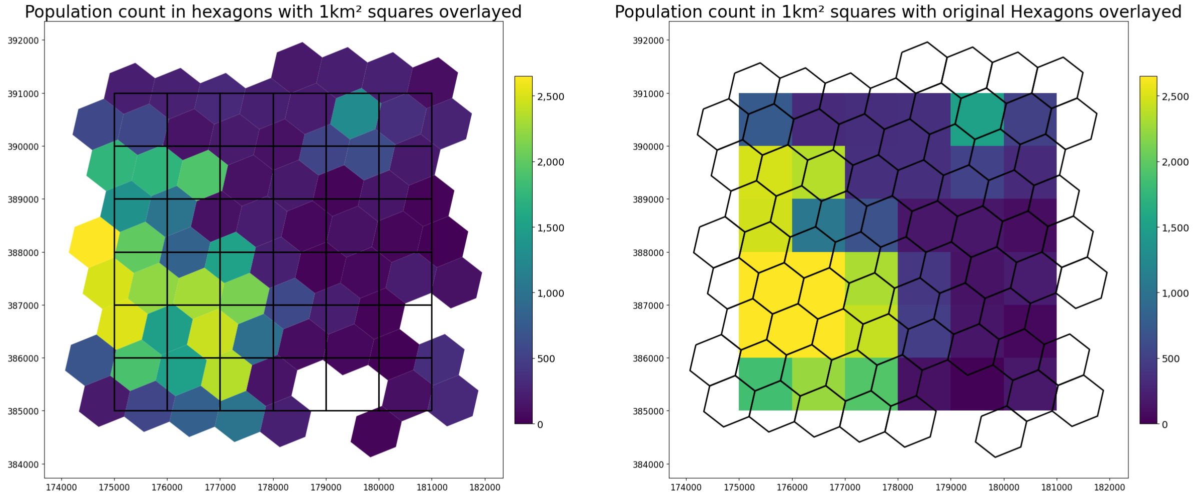
Figure 3.3: Logged NO_X emissions of 2022 in the Netherlands. Red border splits between *land* and *sea* subsets. The map reveals shipping routes of maritime activities with its connection points to the port of Amsterdam, Rotterdam, and Antwerp. Likewise, the road network is clearly visible. The axes are meters in the RD-coordinate system. Data source: RIVM[8]

3.2. Population Density

The population count per square kilometre was obtained from a dataset at a resolution of 400m stored in a Vector H3 hexagons format with global coverage[31]. The dataset was obtained by a fusion of the Global Human Settlement Layer (GHSL), Facebook, Microsoft Buildings, Copernicus Global Land Service Land Cover, Land Information New Zealand, and OpenStreetMap data. The hexagons have been mapped by equation (3.1) to RD-coordinates by intersection of geometric shapes between hexagons and squares, as illustrated in Figure 3.4.

$$P_s = \sum_{h \in H} \left(\frac{A_h \cap A_s}{A_h} \cdot P_h \right) \tag{3.1}$$

where P_s is the total population of square s, P_h is the population of hexagon h, $A_h \cap A_s$ is the intersection area between hexagon h and square s, A_h is the area of a full hexagon, and H is the set of intersecting hexagons with square s.


Population count in hexagons with 1km² squares overlayed

Population count in 1km² squares with original Hexagons overlayed

Figure 3.4: Subset of the Netherlands that illustrates the mapping of the population count from the original data in hexagons to the RD-coordinate 1km^2. The axes denote RD-coordinates which is used for the emission-registration.

3.3. Maritime Routing Data

Besides the emissions of vehicles on the road network, the Netherlands has a large network of maritime transportation and access entries of the ports of Amsterdam, Rotterdam and Antwerp. Shipping is an activity that still relies massively on the burning of fossil fuels. Two datasets from PDOK[19] have been used to obtain the maritime routes. Navigability is classified from 0 to VI and is determined by the CEMT[20] shown in Table 3.2. The shipping routes on sea have been mapped to the VIc class, one that allows for the largest ships to navigate. Figure 3.5 shows the shipping routes obtained from the datasets. Each meter of shipping lane class has been projected to the corresponding 1km^2 squares.

[19] Publieke Dienstverlening Op de Kaart , Public Service On Charts, "Vaarweg Netwerk Data Service-bevaarbaarheid" has been used for inland routes, and "Nationaal Wegen Bestand - Vaarwegen" has been used for routes at sea.

[20] Conférence Européenne des Ministres des Transport

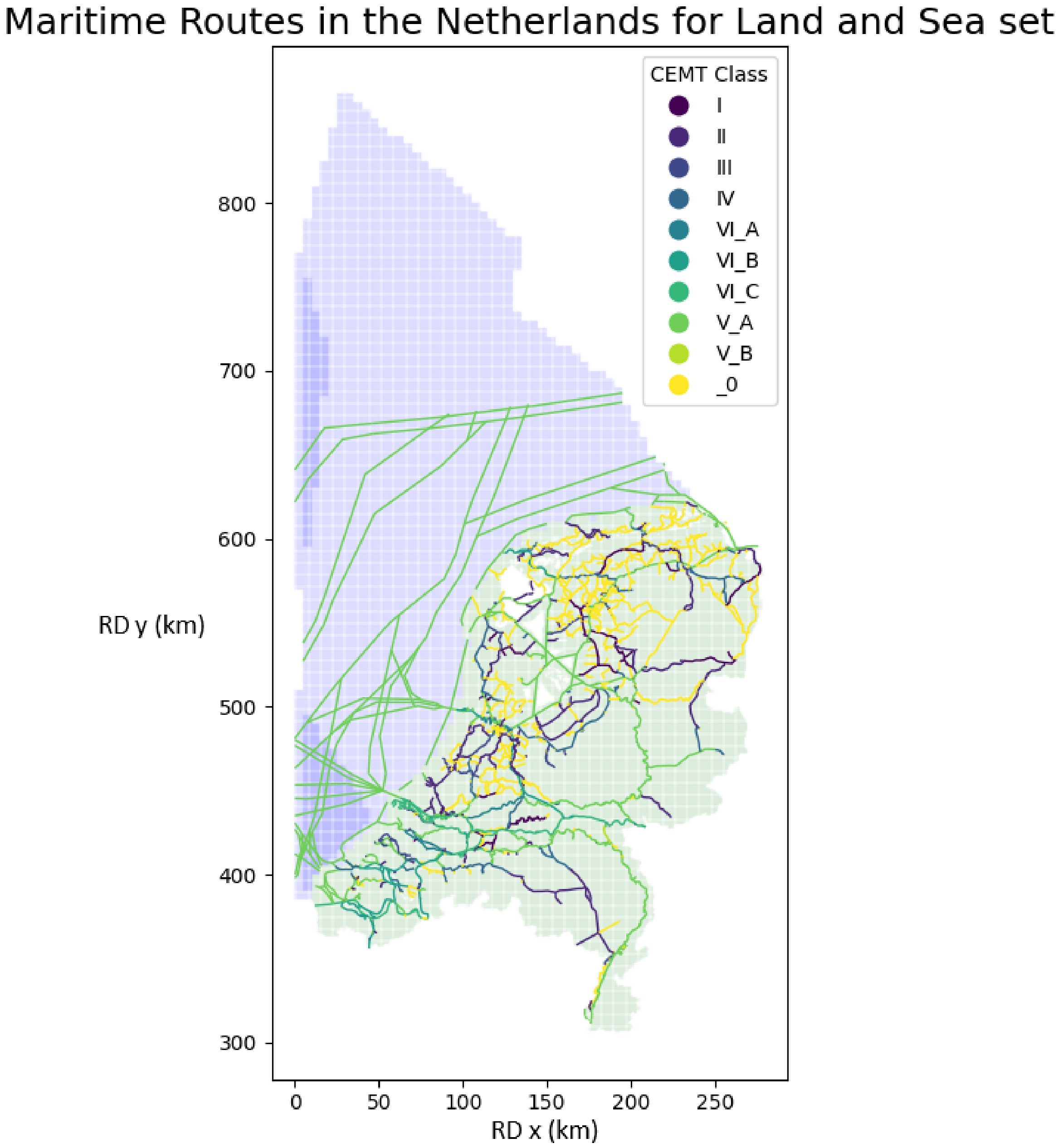


Figure 3.5: Display of the maritime routes in the Netherlands classified by CEMT categories. Data source: PDOK

CEMT Classification	Shiptype
0	Small vessels and recreational shipping
I	Spits
II	Kempenaar
III	Dortmund - Eemskanaal ship
IV	Rhine-Herne canal ship, Single push-tow unit
Va	Large Rhine ship, Single push-tow unit
Vb	Two-barge push-tow unit (long formation)
VIa	Two-barge push-tow unit (wide formation)
VIb	Four-barge push-tow unit
VIc	Six-barge push-tow unit

Table 3.2: CEMT classification standards for accessibility for inland European waterways. Source:[32]

3.4. Open Street Map Data

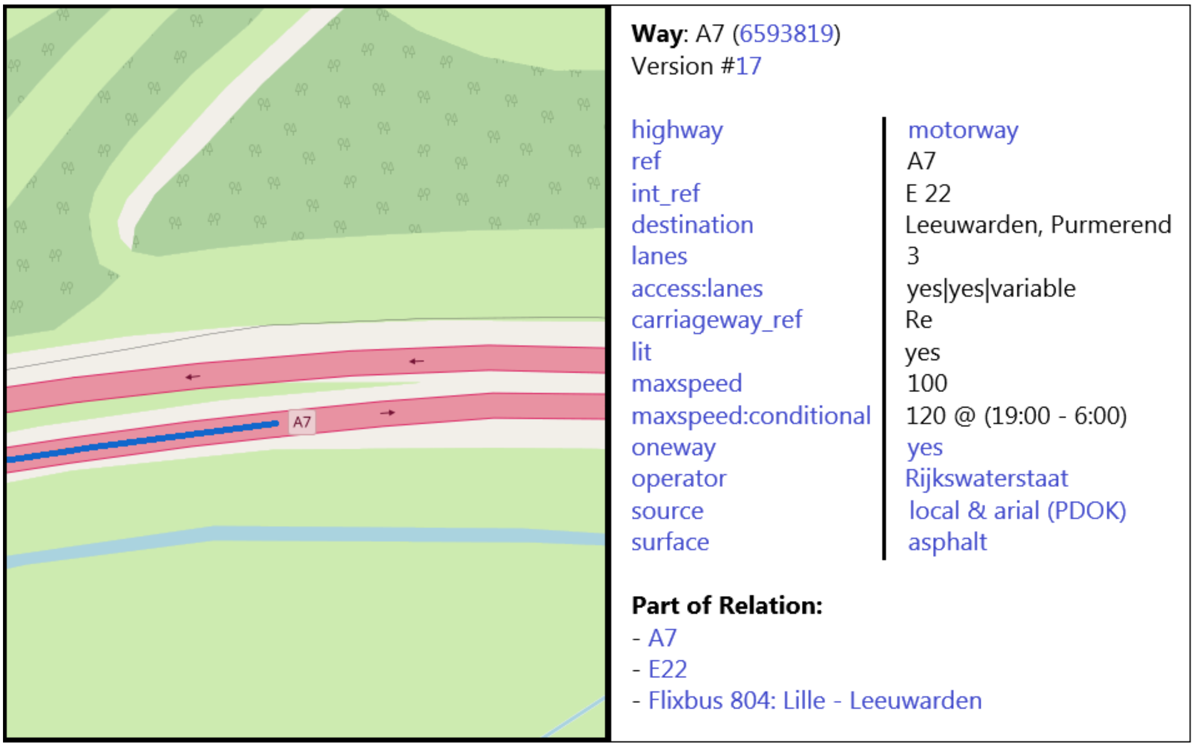
3.4.1. Mapping the World

The largest and most complex data-source used is obtained from Open Street Map [33]. This is an online database and map of detailed attributes which is built and kept up to date by a community of more than 40.000 monthly active users. The data contains features to be expected such as roads, land-uses, buildings, railways, but also more detailed amenities like waste-baskets, windmills, bus-stops and traffic-lights. The objects on the map are one of three geometrical types; *node*, representing a point in space, *way*, representing lines and areas, or *relation*, representing a set of nodes and/or ways. In total, the dataset contains almost $9 \cdot 10^9$ nodes, $1 \cdot 10^9$ ways and $1 \cdot 10^7$ relations. Nowadays around $2 \cdot 10^6$ nodes and $2 \cdot 10^5$ ways are added each day to the database, resulting in new objects or increased precision in existing objects [34]. Additionally, each object has one or more *tags* which is a combination of a *key* and a *value* and describe an object. As of February 2025 there are about $13 \cdot 10^3$ frequently used tags[35].

For illustration, Figure 3.6 shows a *way* object part of the A7 highway, listing detailed tags about the number of lanes, routing information, speed limits and surface type. Figure 3.7 shows another *way* object, the industrial complex of the Enecogen power plant in Europoort Rotterdam, listing address tags, but more important: energy output in Megawatts by gas combustion. Although this energy output might not be accurate and up-to-date, it is an indicator of the capacity of this plant. It must be noted that an energy output tag is provided for some energy plants only, and thus not different types of industrial activities.

Objects in OSM are created and updated through an *proposal process*, which allows different users to validate and update changes in objects or tags. It has to be noted that the database is continuously improving, and thus not always factual. Once changes are approved by consensus between some users (8 votes and at least 75% approval rate), they are added to a temporary *changeset* which will eventually be added to the database. This results in *active* and *inactive* versions of the database, allowing the tracking of changes in objects or locations over time. The complete active dataset stores around 2.000 GB of information.

In the Netherlands, the OSM data is of great quality due to its many users. When using the map elsewhere one has to keep in mind that the completeness of OSM map objects is highest in western countries, mostly European. However, according to a research by Zhou [36], countries with low coverage still have high quality in objects.

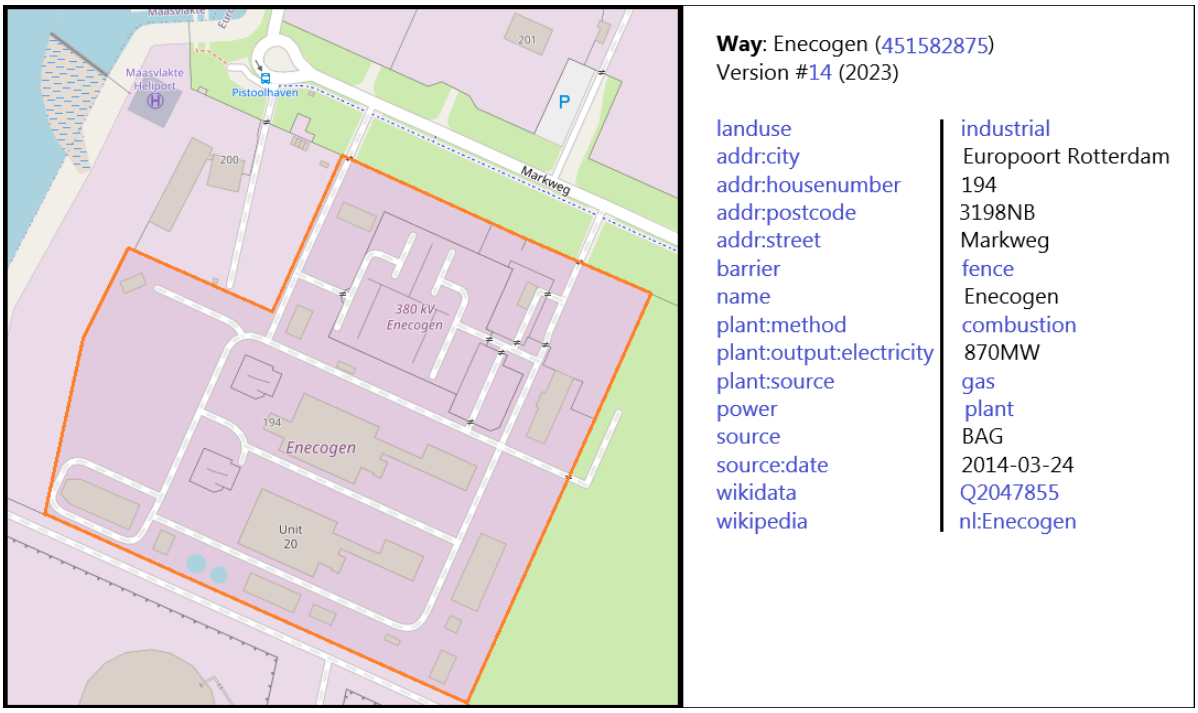


Way: A7 (6593819)
Version #17

highway	motorway
ref	A7
int_ref	E 22
destination	Leeuwarden, Purmerend
lanes	3
access:lanes	yes|yes|variable
carriageway_ref	Re
lit	yes
maxspeed	100
maxspeed:conditional	120 @ (19:00 - 6:00)
oneway	yes
operator	Rijkswaterstaat
source	local & arial (PDOK)
surface	asphalt

Part of Relation:
- A7
- E22
- Flixbus 804: Lille - Leeuwarden

Figure 3.6: Illustration of an OpenStreetMap *way* object, part of the A7 highway, with 14 tags providing information about names, lanes, speed limits etc. The tag '*highway:motorway*' stores the geometric information with a set of *nodes*. Additionally *relations* to which the object belongs to are shown. Hyperlinks point to different versions, tag usages, instances and relations.



Way: Enecogen (451582875)
Version #14 (2023)

landuse	industrial
addr:city	Europoort Rotterdam
addr:housenumber	194
addr:postcode	3198NB
addr:street	Markweg
barrier	fence
name	Enecogen
plant:method	combustion
plant:output:electricity	870MW
plant:source	gas
power	plant
source	BAG
source:date	2014-03-24
wikidata	Q2047855
wikipedia	nl:Enecogen

Figure 3.7: Illustration of an OpenStreetMap *way* object, the Enecogen power plant in Europoort Rotterdam, with 15 tags providing information about names, address, power plant capacity and energy source. The tag '*landuse:industrial*' stores the geometric information with a set of *nodes*.

3.4.2. Which data to collect?

For the estimation of NO_x emissions by LUR, one requires features of high quality that will likely correlate with emission sources. The collection of OSM objects in a certain square km^2 will be aggregated into features representing area, length, or count on a certain tag. For visualisation reasons only, they have been separated into six tag-topics. The descriptions of the topics **traffic**, **traffic-related**, **industry**, **buildings**, **port**, and **land-use** are given below. The complete list of tags used is presented in the Appendix A.

Since traffic is one of the most emitting sectors as seen in 1.1, the road network should be represented in the dataset. OSM provides twelve distinct types of roads for vehicles under the **tag key** '*highway*' with **tag values** such as '*motorway*' ,'*secondary*', '*residential*', '*motorway_link*', '*railway*' and '*unclassified*', see Figure 3.8. For each type, the length is multiplied by the number of lanes to obtain length features for 1 km^2 squares.

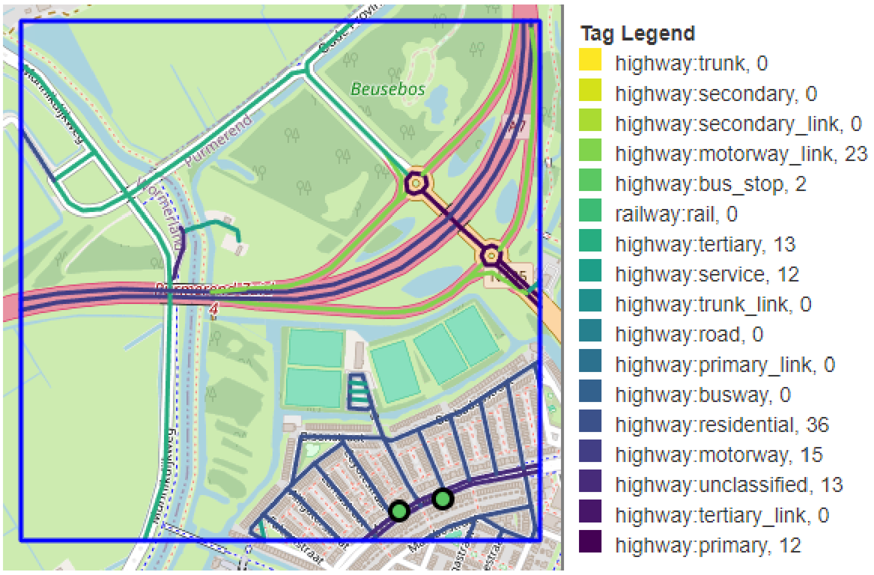


Figure 3.8: An example of OpenStreetMap feature retrieval on the tag-topic 'roads' for a 1km^2 square in Purmerend.

Besides the road network, extra traffic related features are interesting, since they reveal the locations of traffic congestions and traffic in acceleration and/or idle state. This is important since combustion engines emit more when running in non-ideal states, by Zhang and Johnson [37], [38]. Therefore objects like '*traffic_signals*', '*parking_space*', '*shop:car*', and '*bus_stop*' are gathered in their own tag-topic, see Figure 3.9.

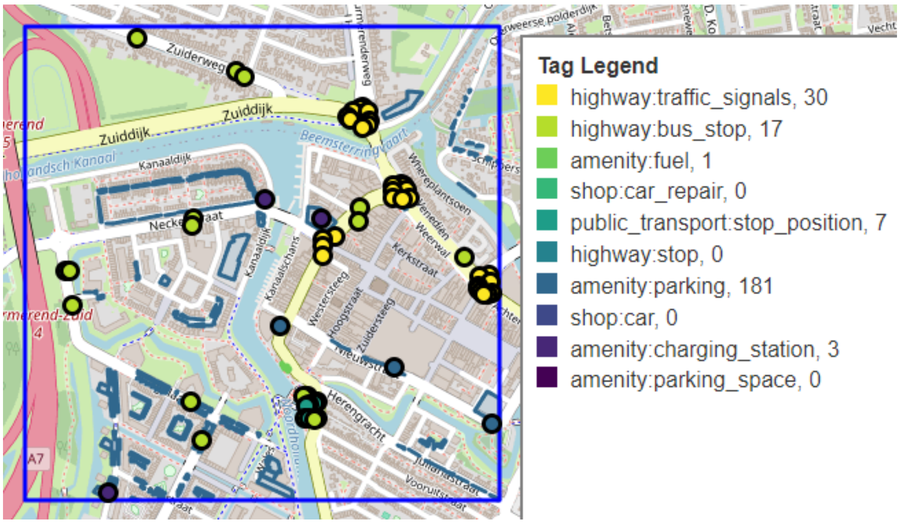


Figure 3.9: An example of OpenStreetMap feature retrieval on the tag-topic 'road related' for a 1km^2 square in Purmerend.

Another sector responsible for high emissions is industry, which generates power, either for production of goods or for consumption by households, often relying on fossil fuels. The tag-topic 'industry' contains features such as '*power:plant*', '*man_made:chimney*', and '*building:industrial*', see Figure 3.10

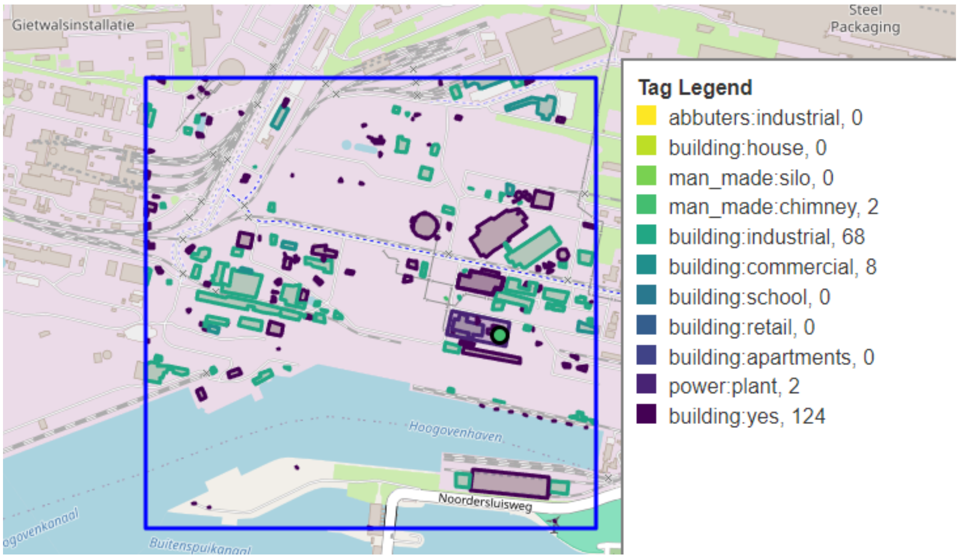


Figure 3.10: An example of OpenStreetMap feature retrieval on the tag-topic 'industrial' for a 1km^2 square located at the Tata Steel IJmuiden factory.

Features like '*building:commercial*' and '*building:retail*' might have an effect on traffic density aside from the '*population*' counts together with infrastructural features. See Figure 3.11 for an example of these features.

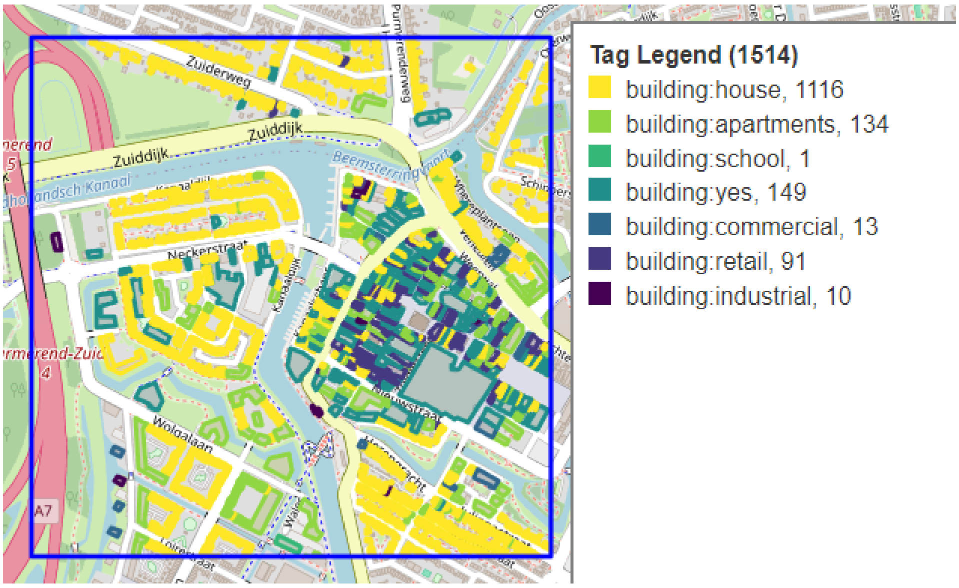


Figure 3.11: An example of OpenStreetMap feature retrieval on the tag-topic 'building' for a 1km^2 square located around the city centre of Purmerend.

Besides transportation on land there is transportation over water both inland and outland, together with activities in ports. This means that the **port** tag-topic contains features like '*man_made:pier*',' *man_made:goods_conveyor*', '*man_made:pipeline*', and '*amenity:ferry_terminal*', see Figure 3.12.

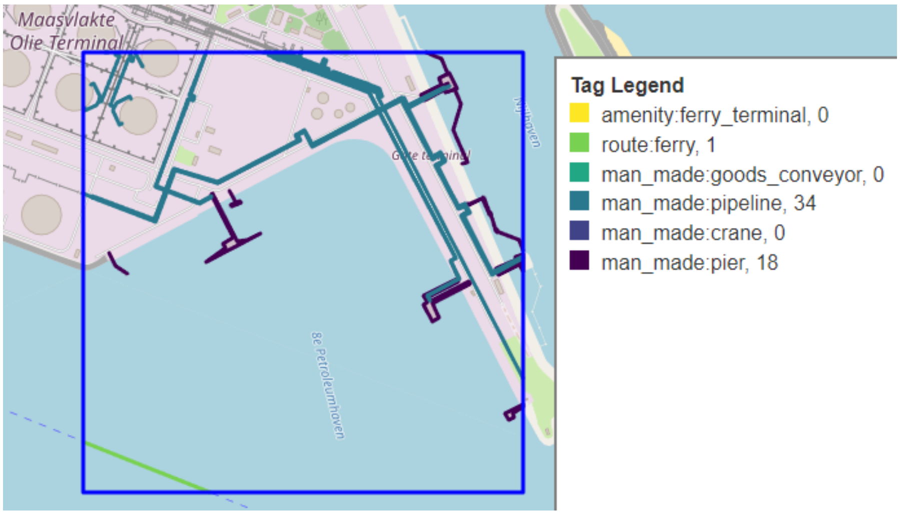


Figure 3.12: An example of OpenStreetMap feature retrieval on the tag-topic 'port' for a 1km^2 square located at the Maasvlakte, Port of Rotterdam.

Finally a **landuse** tag-topic is created to compute areas on tags such as '*landuse:construction*',' *landuse:residential*', '*landuse:farmland*' and '*landuse:forest*'.

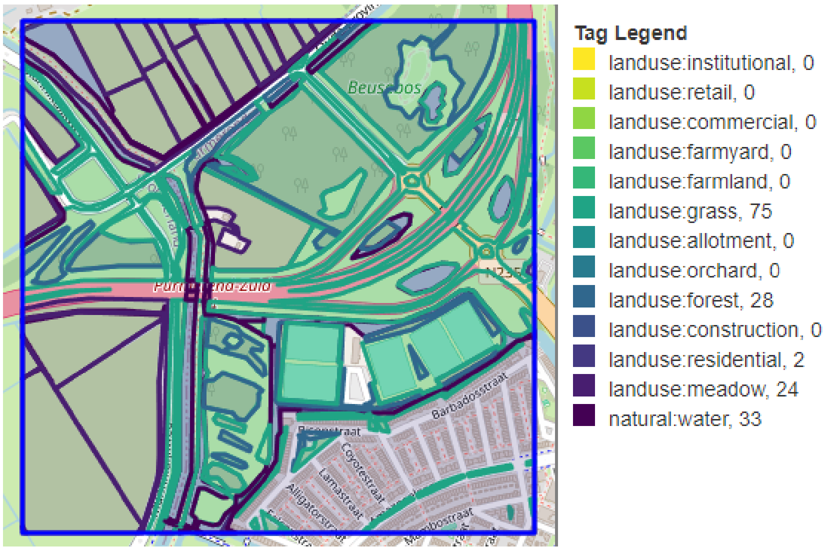


Figure 3.13: An example of OpenStreetMap feature retrieval on the tag-topic 'landuse' for a 1km^2 square in Purmerend.

3.4.3. OSM Feature Creation

Since not all objects that are returned by the API are required, or in the required format, some functionalities have been built in. For reproducibility on each of the functionalities, see the respective implementation settings in Appendix A

Filtering. Some features retrieve too many objects. For example '*building:yes*' is labelled to some industrial buildings, but also to garden sheds. The aggregated area does not distinguish this difference. To make the '*building:yes*' feature more suitable for industrial purposes, it is decided to filter out single objects below a certain threshold.

Overlap Resolving. Some objects pointing to certain locations overlap. For example the tag '*landuse:industrial*' is allocated to industrial areas, but in some cases also to industrial facilities belonging to this industrial area. This means that any smaller object is subtracted from larger objects such that the smaller area is used only once instead of twice.

Tag Setting. When retrieving an object based on a single tag, such as '*power:plant*', many more tags could belong to this object such as '*plant_source:gas*' as seen in Figure 3.7. First, the objects are given a *simple* tag based on the order of importance. This importance makes sure that when an object has more tags, the more detailed tag is selected (e.g. '*industrial:refinery*' is prioritised over '*landuse:industrial*'). Then, several objects are further specified into a *final* tag by a secondary tag of the object. For example, objects having the '*power:plant*' tag often come with the tag '*plant_source*', which makes the final tag '*power:plant_<source>*'. Resulting tags could become '*power:plant_gas*', '*power:plant_waste*' or '*power:plant_solar*' etc.

Tag Remapping. Since some tags differ slightly in name, they are remapped to existing tags.

Object Weighting. This is a functionality that reweights objects based on certain secondary tags belonging to objects, such as '*lanes=2*' of '*highway:motorway*'. For now, this is done by multiplying the object *length* by '*lanes*'.

Object Splitting. This functionality ensures that the energy output is distributed proportionally over different km^2 squares when an object spans more than one grid square. For example a '*power:plant*' object with '*plant:output:electricity=736MW*' that overlaps two grid squares will have its output divided proportionally to the area of the facility within each grid.

3.5. Final Benchmark Dataset Creation

The final dataset is a spatial join on all 1 km^2 squares with all datasets below:

- RD Coordinates, containing geographic information about 1 km^2 squares,
- RIVM emission data, containing emissions for each 1 km^2 square in 2022,
- Population data, containing population count on each 1 km^2 square,
- Maritime data, containing maritime routes on sea and land for each 1 km^2 square,
- OpenStreetMap features, containing objects aggregated on area, length or count for each 1 km^2 square.

This dataset is used as a *benchmark* dataset on which the VLM will indirectly make alterations. All 139 *area*, 19 *length*, 12 *count* features, 10 *maritime*, and the single *population* features have coverage on all the 42.323 grids. The 19 *MW* columns affect only a few grids, at most 25.

Figure 3.14 shows features correlated with NO_x emissions using the Kendall-correlation. The strongest features are related to energy, urban, and industrial features. At first glance, the scatter-plots appear almost exponentially decreasing, not suggesting positive relationships. However, the correlations are explained by patterns in the lower ranges of the data, shown in Figure 3.15, where the y-axes are limited to the 97th percentile of NO_x emissions. This illustrates why simple bivariate inspection can be misleading and why advanced modelling approaches are required to uncover meaningful relationships.

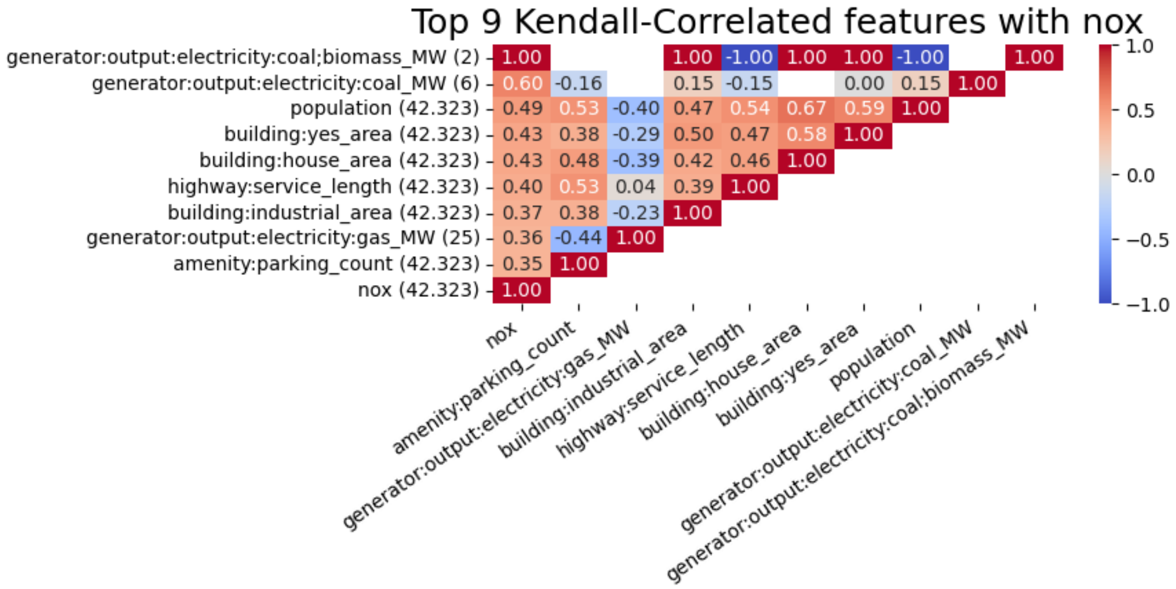


Figure 3.14: The correlation matrix for top 9 features on the *benchmark* dataset sorted on absolute values of correlation with target. The non-parametric Kendall is used since the features are highly rightly skewed. On the y-axis a features count is shown.

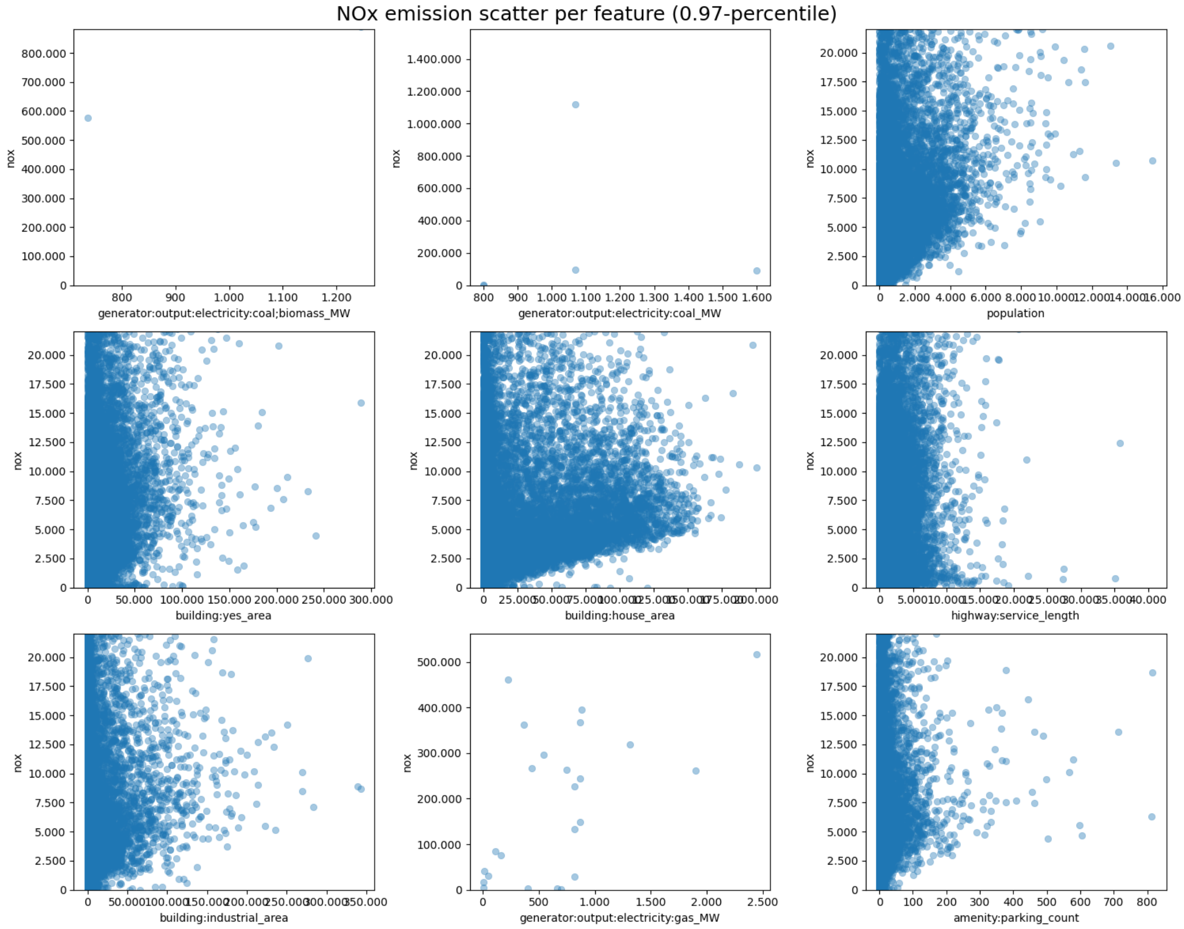


Figure 3.15: Scatter-plots on the Top 9 Kendall-correlated features on NO_X of the *benchmark* dataset.

4
Land Use Regression

The possible improvements by the VLM enhanced dataset will be assessed by several models; **Random Forest** since it is a model used in many LUR researches, **Histogram-XGBoost** since it is another sophisticated tree ensemble model, and **Elastic-Net** for displaying the gains of tree ensemble models. The selection of these different types of models should provide insight in what kind of models are better at alleviating the enhanced data that is created through the usage of VLMs. A brief explanation of their function follows to support intuitive understanding. Finally, the implementation of the training is outlined.

4.1. From Decision Trees to Gradient Boosting Trees

Elastic Net is a regularised Linear Regression model which combines L1 regularization (Lasso Regression) and L2 regularization (Ridge Regression). The first method adds a penalty for the sum of absolute coefficients, tending to filter out non-predicting features. The second method adds a penalty for the sum of squares of coefficients, reducing the size of large coefficients. Elastic Net is then a hybrid, adding L1 and L2 penalties with strength λ, and a L1-L2 proportion of α, resulting in the loss function described by Equation 4.1.

$$\mathrm{L}oss_{\text{ElasticNet}} = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda\left([\alpha]\sum_{j=1}^{n}|\beta_j| + [1-\alpha]\sum_{j=1}^{n}\beta_j^2\right), \text{ where } \lambda \in [0,\infty), \alpha \in [0,1] \tag{4.1}$$

Decision Tree is a model that partitions the input space into regions by binary splits based on any feature of choice. The model could be used for classification, as shown in Figure 4.1, and regression. An upside-down tree-like structure separates the dataset starting from the *root* node through *decision* nodes down to *leaf* nodes. A collection of data-points in a leaf is used for the classification or regression of unseen samples, for example by selecting the majority class or average value, respectively. Training the DT on a training set involves determining the decision variables at each binary split, the number of leaves allowed, the depth of the tree, etc. For a full description of DTs consult the work of Leo [39] and Bishop [40].

Random Forest is a so-called *tree ensemble* model. This is a collection of Decision Trees, in the order of hundreds or thousands, hence a *Forest*, that aggregate the predictions of all separate trees into a final prediction. The method prevents overfitting twofold. First, a subset of predictors is sampled for each tree, resulting in different estimations per tree. Secondly, a bootstrapping technique is used that samples a random subset of the data on which a tree is built. Typically 67% of the dataset is used to build the tree, and the remaining 33% is used to calculate the Out Of Bag (OOB) error. This results in a built-in method prevents overfitting, see Breiman [41] and Fife [42].

Gradient Boosting could be seen as a more sophisticated version of a Random Forest model, which does not outperform an RF by definition. It builds not just a collection of trees independently, but sequentially, where a new tree is based on errors of previous tree created and thereby complements previous predictions. The trees are built with respect to a loss function, allowing for gradient computations for faster convergence to an optimum. Consider Chen's work for a full explanation [43].

Histogram-based Gradient Boosting Regression Tree is an adaptation to the Gradient Boosting Regression Tree structure available in the scikit-learn library [44] which is recommended for larger datasets (n > 10.000). It uses a smaller memory by placing each feature instance in a bin, which additionally reduces the number of

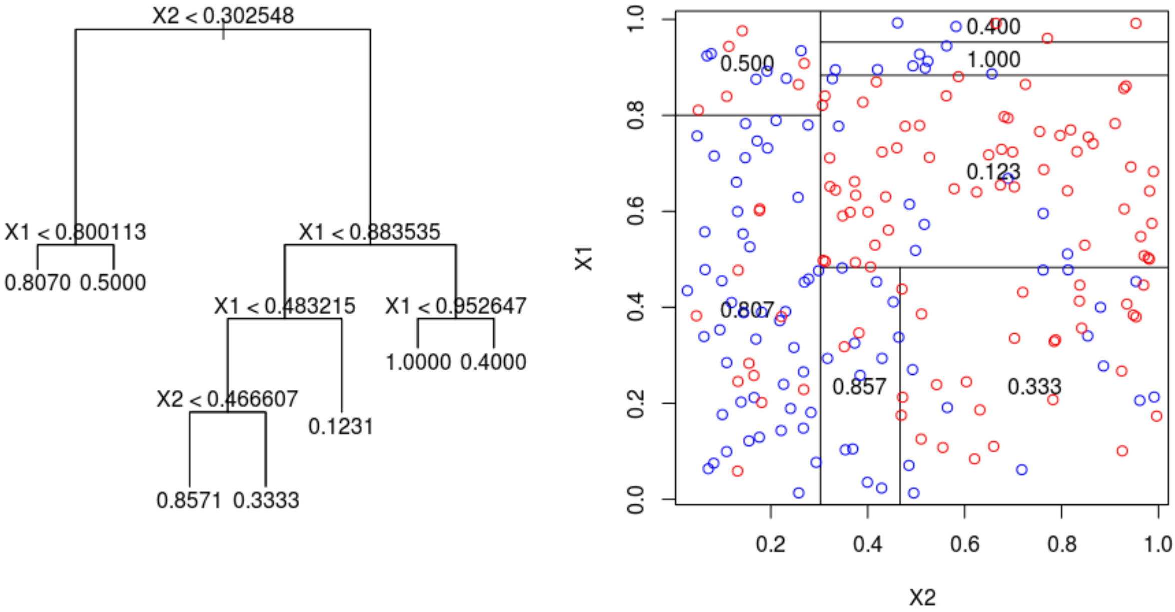


Figure 4.1: Example of a Decision Tree that partitions a 2D input space for binary classification: blue or red. (left): At each decision node a binary split is made based on the *X1* or *X2* value. The leaf nodes of the tree denote the probability of observing a blue class. (right): A visualisation of the splits by the DT with the data-points in the training set, and the probabilities of observing blue that coincide with the values in the leaf nodes. New nodes will be classified as blue or red based on these probabilities.

potential binary splitting-points. The model is significantly faster and at least as good as the base model, as stated by researches of Piotr [45] and Pedregosa [44].

4.1.1. Why Tree Models work for LUR

Tree ensemble methods have proven to be more effective than other traditional models when used for LUR modelling. Besides being able to capture complex nonlinear relations, the model optimisation algorithm contains functionalities for the prevention of overfitting. The latter is done by using boosting, bagging, feature sampling, tree pruning, and more. On top of that, the models allow for usage of different data-types, missing values, scalability to large datasets, insight in feature importance, and fast model inference.

4.2. Research Implementation

4.2.1. Target Scaling & Model Performance Metrics

Feature scaling is not required for tree ensemble models, thus RF and HGBR, since they bin features based on order of values, which is not affected by scaling. On the other hand Elastic Net requires feature scaling, thus a scaler object standardises the training data for this case only. The testdata are scaled using the same parameters of the scalar per feature obtained from scaling the training data.

On the other hand, a deliberate choice has to be made on whether to transform the target data or not. Since emission target values are highly positively skewed and sparse in the right tail of the distribution (Fig. 3.2), the optimisation metrics behave differently on the raw or log-transformed scale. It is decided to train models twice, once on the original target scale and once on the transformed target scale $\log_{10}$ and compare the model fits on the original target scale.

In this research, the Root Mean Squared Error loss function (RMSE) will be used for model optimisation (Eq. 4.2). Its selected to emphasise more on larger errors. By rooting, it becomes a more interpretable unit of $NO_x/y/km^2$.

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_{\text{true}} - y_{\text{predicted}}\right)^2} \in [0,\infty),\ n\text{: number of observations},\ y\text{: target value} \tag{4.2}$$

The RMSE measure will be used for the comparison of model performance within this research. Secondly, the

R^2 metric (Eq. 4.3) will be computed to compare with methods outside this of research, as it is often reported in other research on NO_x estimations. When $R^2 = 0$, the model performs equally well as a baseline model which always predicts the mean $\overline{y}$. When $R^2 = 1$, the model performs perfectly, predicting the target data without errors. When $R^2 < 0$, the model performs worse than the baseline model, subsequently rejecting the model. The R^2 metric is not comparable between models trained on different target scales. Therefore, the metrics are computed after inverse transformations of log scale predictions, making them comparable on the original scale.

$$R^2 = 1 - \frac{(\hat{y}_{\text{target}} - \hat{y}_{\text{model}})^2}{(\hat{y}_{\text{target}} - \overline{y})^2} \in (-\infty, 1], \text{ where } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ , } n\text{: number of observations, } y\text{: target value.} \tag{4.3}$$

4.2.2. Train-Test Split Creation

All models will be trained on the same 80% of the data and tested once on the remaining 20%, (33.807 train, 8.516 test). Since the data are about spatial objects, the data instances are not independent and identically distributed, i.e., high population numbers in the surrounding area, likely results in more traffic on roads. Therefore, the testset will be split based on 4 × 4 grid blocks, instead of randomly sampling directly. Due to the imbalanced target (Table 3.1), the grid blocks are sampled by *stratified sampling* based on the maximum target value within the grid blocks. This ensures that the sampling of grid blocks is not fully random, but results in similar distribution of target values between the training and test set. In this research, the square kilometres are divided into training and test set using *strategy='quantile'* and *'n-bins'=8* to preserve equal bin sizes, Figure 4.2 shows a part of the result.

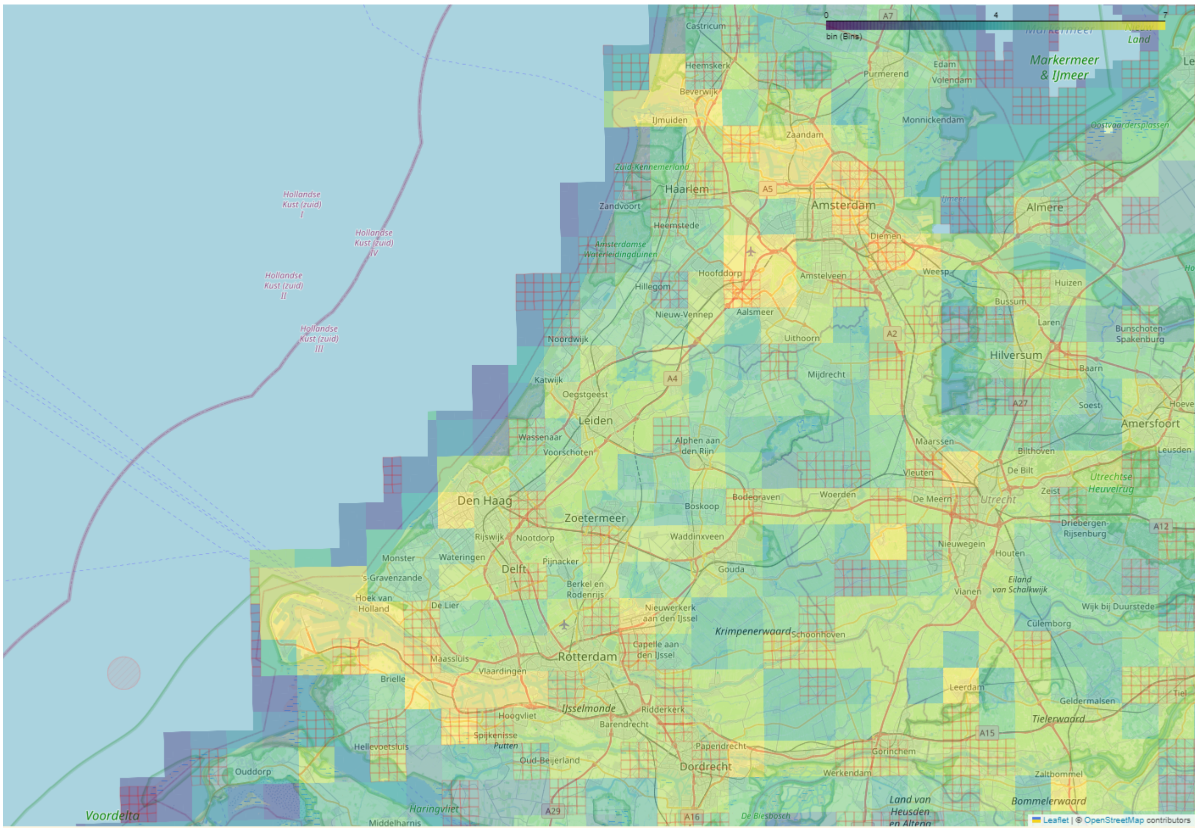


Figure 4.2: A section of the 80-20 train-test split created by stratified-sampling. All 1 km^2 squares have been grouped to 4 × 4 chunks, and binned to one of the 8 target bins, see colour scale. By random sampling of chunks per bin, a test set was created denoted in red.

Hyperparameter selection is performed by employing the Optuna framework, see Section 4.2.4. During training *k*-**Fold Stratified Cross Validation** is used, creating *k* subsets of the training data where during each fold one subset acts as validation set. This prevents overfitting the training data as a whole whilst using its entirety. Again, stratified sampling is used to create the $k = 5$ subsets of the training data with similar distributions (*strategy='quantile','n-bins'=4 due reduced number of data-points*). When the best hyperparameters for a model are found, a new model instance is trained on the entire training set of 80%.

4.2.3. Test Set Evaluation

The metrics RMSE and R^2 computed on the test set are only point estimates, since they are obtained from a finite sample of the underlying distribution of real emission values. Consequently, the composition of this sample has a strong influence on the reported metric. A procedure known as bootstrapping, where the test set is repeatedly resampled with replacement, allows for obtaining uncertainty estimates and enabling statistical testing. For each resample, the metrics are recalculated, producing a distribution of performance differences.

The differences between a *baseline* and an *alternative* model are given by Equation 4.4. Significance is then assessed by the proportion of bootstrap instances where the difference is in the hypothesized direction. For metric M, the probability that the *baseline* model performs better than the *alternative* is given by Equation 4.5. If $p > 0{,}975$, the *alternative* model is significantly worse based on M. If $p < 0{,}025$, the *alternative* model is significantly better. When $p = 0{,}5$, the *baseline* and *alternative* methods are equal based on M.

$$\Delta \mathrm{M}^b = \mathrm{M}^b_{baseline} - \mathrm{M}^b_{alternative} \quad \text{, where M : Metric , } b = 1, 2, \dots, \mathrm{B} \tag{4.4}$$

$$p = \Pr(\Delta \mathrm{M} < 0) \approx \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} (\Delta \mathrm{M}^b < 0) \quad \text{, for large B} \tag{4.5}$$

Parameter B must be chosen such that the collection of all M converges. Based on the graphs with different values of B in Figure 4.3, B = 5.000 was selected. This B was also sufficient for the other combinations of model and dataset.

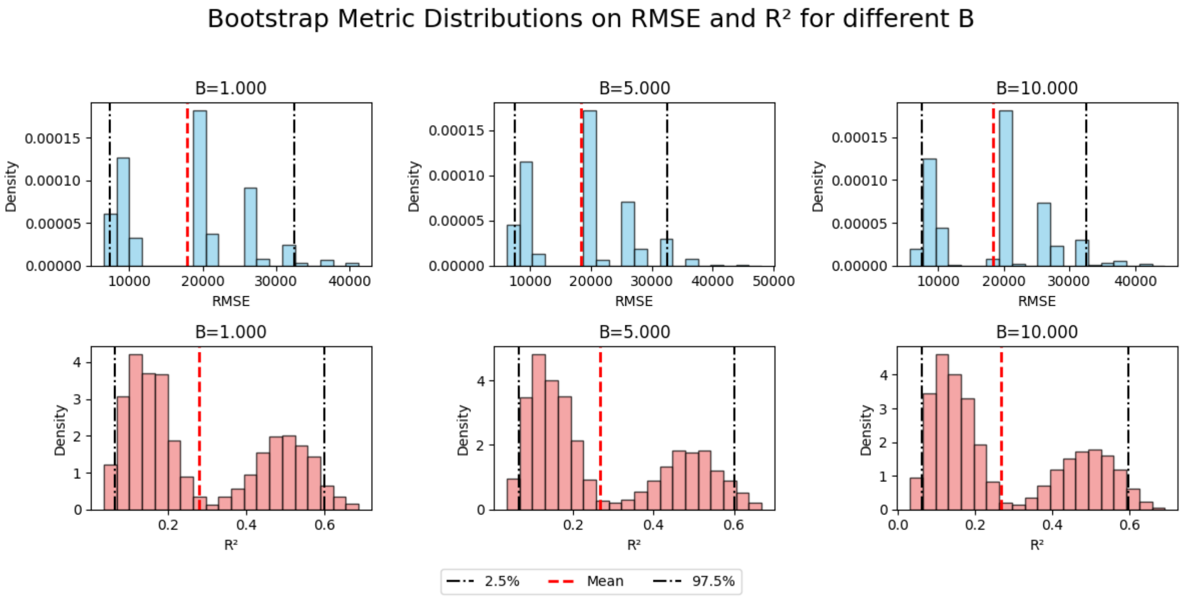


Figure 4.3: Histograms of computed metrics RMSE and R^2 obtained from bootstrapping on the test-set predictions of the HGBR-nox model trained on the *benchmark* dataset with different number of samples B.

Bootstrapping on the test set serves a role conceptually similar to cross-validation on the training set: both approximate variability through repeated resampling. In cross-validation, subsets of the training data are resampled to estimate how well a model generalises during development. In contrast, test-set bootstrapping is applied after training to quantify the uncertainty of final performance estimates and to enable significance testing between models.

4.2.4. Model Hyperparameter Tuning

All the hyperparameters for the three models are tuned by the Optuna framework [46]. It uses Bayesian optimisation to sample promising hyperparameters, resulting in a combination of the best hyperparameters and model parameters. The hyperparameter ranges were determined through an iterative process, where a few preliminary trials were used to identify unstable or irrelevant values, after which the search space was narrowed to focus on plausible regions. All models are trained using 40 trials, after which the best set of hyperparameters is saved.

Parameter	**Range / Values**	**Notes**
alpha	10^{-4} – 10.0 (log scale)	Regularization strength
L1_ratio	0.0 – 1.0	L1 vs L2 mix
fit_intercept	{True, False}	Fit intercept flag
selection	{cyclic, random}	Coordinate descent mode
random_state	5	For reproducibility)

Table 4.1: Hyperparameter search space for ElasticNet

Parameter	**Range / Values**	**Notes**
n_estimators	50 – 500	Number of trees
max_depth	3 – 20	Max tree depth
min_samples_split	2 – 20	Minimum samples to split
min_samples_leaf	1 – 20	Minimum samples per leaf
max_features	{sqrt, log2, None}	Feature subset strategy
max_leaf_nodes	200 – 1200	Tree complexity control
min_impurity_decrease	0.0 – 1.0	Minimum impurity decrease
bootstrap	{True, False}	Sampling with replacement
max_samples	0.5 – 1.0	If bootstrap=True
random_state	5	For reproducibility)

Table 4.2: Hyperparameter search space for RandomForestRegressor

Parameter	**Range / Values**	**Notes**
learning_rate	0.01 – 0.2	Step size, sample log scale
max_iter	50 – 500	Number of boosting trees
max_depth	6 – 15	Max depth single tree
max_leaf_nodes	200 – 1000	Max leaves single tree
min_samples_leaf	1 – 50	Minimum samples per leaf
L2_regularization	0.0 – 1.0	L2 penalty
max_features	0.1 – 1.0	Fraction of features
max_bins	128 – 255	Histogram bins
early_stopping	True	Fixed (enabled)
validation_fraction	0.1 – 0.3	Validation split fraction
tol	10^{-7} – 10^{-4} (log scale)	Tolerance for stopping
n_iter_no_change	10	Fixed
random_state	5	For reproducibility)

Table 4.3: Hyperparameter search space for HistGradientBoostingRegressor

5 Enhanced Land Use Regression

In the previous chapters, the *benchmark* dataset is built, followed by a brief explanation of the types of models and the implementation of research. The following chapter addresses how alterations to the *benchmark* dataset are generated using a VLM.

5.1. Vision Language Model Selection

Some models are *open source* meaning that the model parameters and structure are available for analysis. They allow for running the models locally and even fine-tune to specific tasks. Other models are *open data*, providing all the data the models were trained on. Finally, there exist *closed-source* models, for which the model structure, training data and parameters are disclosed. Oftentimes these are high performing models that are only accessible through applications or APIs and thereby trying to protect their advantages.

To compare all kinds of models — LLM, VLM, open-source, closed-source, multi-modal, etc.— benchmark datasets are created. Typically each benchmark dataset focusses on a specific task such as question answering, image captioning, 'Optical Character Recognition' (OCR), object detection, reasoning, relation finding, etc. Providing an example, the 'Massive Multi-discipline Multimodal Understanding' (MMMU)[47] is one of the most comprehensive benchmark datasets. It consists of 11.500 questions from college exams, quizzes, and textbooks, covering core disciplines such as Engineering, Science, and Medicine. The questions consist of an image-text pair with multiple-choice options featuring diverse image types such as charts, diagrams, maps, and chemical structures. The dataset is constructed to test models on skills such as perception, reasoning, and knowledge. In February 2025, the leaderboard on accuracy sets *Human Expert (High)* on top with an accuracy of 88,6% while the best performing model 'o1', by OpenAI, achieves an accuracy of 78,2%. By July 2025 the best performing model is Google's Gemini 2.5 Pro Deep-Think at an accuracy score of 84,0%, depicting the active improvements of this subject.

To get an overview of performance by model and benchmark dataset, there exist communities (most preeminent Hugging Face[48]) that test and evaluate models. One method is to test VLMs against a collection of benchmark datasets such as OpenVML Leaderboard[49]. Another method is based on user interaction with anonymous models and voting such as Arena(Vision)[50] which evaluate the models on unseen prompts, images, documents, and questions.

At the start of this research the two best performing models on the leaderboard of Arena(Vision) were versions of OpenAI's GPT-3 and Google's Gemini Flash 2.0 model. There was no insight in which documents or datasets have been used in training these models. It would at least be reasonable to assume for Google to incorporate trainingdata on its own services such as Google Maps, which contains aerial image understanding. For OpenAI this would have been illegal regarding Google's policies [51], although a different dataset could have been used, which, in return, would not have included Google's point-of-interests. Therefore, this research uses the **Gemini Flash 2.0** model with a knowledge cut-off in June 2024 [52]. The model will be used through an API and will not be fine-tuned.

5.2. Vision Language Model Deployment

5.2.1. 'Enhancement' Approach

To improve the quality of the OSM data constructed for the benchmark dataset model, several options could be investigated. Options that come to mind when inspecting Figure 1.1 are traffic and industry related, which account for 51,8% and 24,2% respectively. Considering the comprehensive coverage of the traffic network, consisting of more than 18 types including rails and waterways, in combination with population densities, one could argue that the low hanging fruit has been included for traffic.

While some facilities, such as refineries in Rotterdam, have been extensively tagged by OSM mappers, others remain underspecified. For instance, the steel factory of Tata-Steel and the ICL fertilizer facility are not clearly categorised in terms of their operational use. In general, the geometries are accurately mapped, but the attribution through tags lacks consistency. For example, both raw material storage sites and concrete factories may be labelled as ’*landuse:industrial*’, although the first does not directly emit NO_x, whereas the latter does. This inconsistency motivates why this research investigates the enhancement through the ’*landuse:industrial_area*’ feature.

In total, 4.629 OSM objects are tagged with ’*landuse:industrial*’. Many of the objects span an entire industrial region, see Figure 5.1. Therefore, these objects are partitioned into cadastral parcels with the [53] dataset. An algorithm is applied to filter only objects with ’reasonable’ sizes and shapes, see Section 5.3.1 for more details. This is done only for computational and cost reasons and therefore could be omitted. The result is 80.992 parcel objects and thus aerial images that will be processed by the VLM, see Section 5.3.2 for image creations.

In total, the feature ’*landuse:industrial_area*’ in the benchmark dataset represents an area of 482 km^2, which is 1,14% of the total dataset’s area spanning 42.323 km^2. The selected parcel objects for reclassification span an area of 296 km^2, which is 0,7% of the original dataset area. The parcel objects are located in 5.697 of the 1 km^2 squares, thus, 13,5% of the benchmark dataset rows will be altered in several columns.

Which columns will be altered is given by a list of tags displayed in Table 5.1. It is a selection made from tags available by OSM but also new ones, consisting of a wide range of operations whilst not being too extensive. The selection was guided by the types of companies listed on the site of RIVM displaying emissions [8]. Documentation on prompting techniques did not provide what number of tags would work best for structured outputs [54].

In addition to the modifications of the *area* columns, a new column, *NOx_emission_estimate*, obtained from the VLM, is introduced (see Section5.2.2). This feature is generated by the VLM through reasoning on context it produces. The purpose of adding this feature is to guide the model in distinguishing between facilities with vastly different emission levels, thereby aiming to improve its ability to predict NO_x emissions across several orders of magnitude. This process of indirectly altering the *benchmark* dataset will be done four times with different settings for the VLM, resulting in four ’*enhanced*’ datasets, more in Section 5.2.2.

amenity:parking
building:commercial
building:industrial
industrial:asphalt *
industrial:car-terminal *
industrial:cement
industrial:chemical
industrial:dry-bulk
industrial:fertilizer *
industrial:liquid-bulk
industrial:logistics
industrial:manufacturing
industrial:metal-construction
industrial:port
industrial:recycling
industrial:wastewater-plant
industrial:infrastructure *
landuse:construction
landuse:grass
landuse:other *
landuse:retail
leisure:marina
man-made:pier
natural:water
power:plant-biomass
power:plant-coal
power:plant-diesel
power:plant-gas
shop:car
shop:car-repair

Table 5.1: *landuse:industrial* tags are reclassified by the VLM to one of the following tags above. * non-existing tags in *benchmark* dataset.

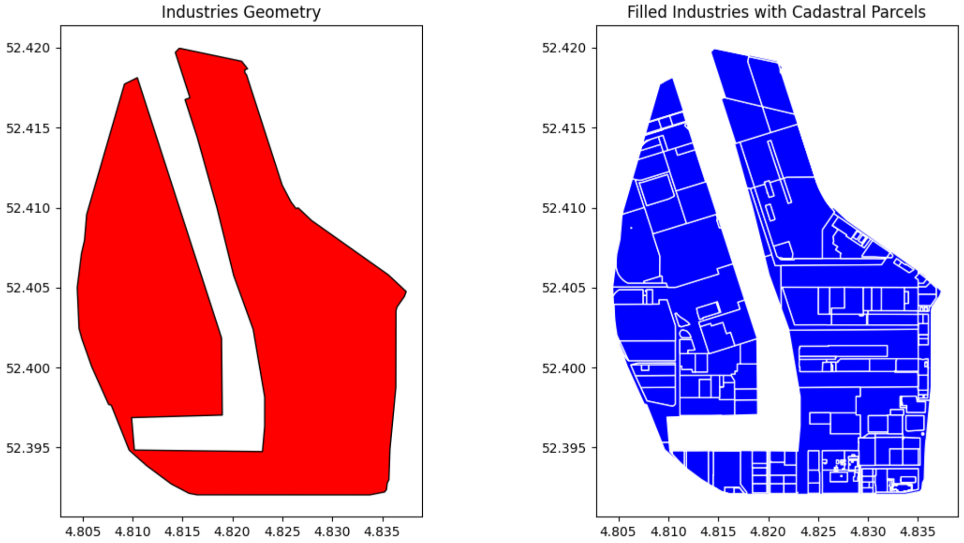


Figure 5.1: Red: geometric shape of the 'Westhaven' OSM object near Amsterdam which is only labelled with *landuse:industrial* (id: 188873529). The object spans a huge industrial area with many different industrial facilities of which no geometric shape exists in OSM yet, by inspection: car terminals, warehouses, dry-bulk transport and refineries. Blue: the partitioned OSM object into cadastral parcels, resulting in a complex collection of geometries that not only divide the facilities but also fragment the facility locations themselves.

5.2.2. About Model Inferencing

The Google Gemini documentation [54] provides prompt engineering guidelines that should improve model performances, suggesting to (paraphrased):

- provide context on the inputs and overall objective,
- narrow down the focus by step-by-step problem solving,
- provide examples, known as *few-shot-prompting*,
- add prefixes like "Answer is:", "Examples:", "English text:",
- add confirming patterns, instead of anti patterns,
- aggregate intermediate task answers to a final answer
- provide constraints on output format and size.

The *system instruction*, provided at the beginning of each model initialisation, will give the model context, an objective, and 'personality'. This system instruction is used with every prompt following in the conversation. The system instruction, presented below, declares a role, context, and desired output with constraints.

```
"""
Role: You are an aerial image analyst classifying industrial parcel landuses and estimating
    its direct (not indirect activities outside location) yearly nitrogen-oxide emissions in
    kilograms based on industrial attributes visible.
Constraints: You are allowed to come up with any label as long as it is fitting, concise, and
    conventional. Each step pairs with some *optional* examples.
About the Image: The image is an aerial view from Google Maps with a magenta highlighted
    cadastral parcel. There is space around the highlighted area that could serve as context.
Output: The output is a structured json.
"""
```

The creation of *four* alterations mentioned earlier originate from evaluating two different prompts with each two different model *temperature* settings. The first aim is to investigate whether providing the VLM with more *steps* guides the model to better output. Additionally, the *temperature* setting is lowered from the default value of 1,0 to 0,1. A lower temperature makes the VLM output more deterministic —which could be interpreted as less creative— which might result in different output behaviours. Hence, the four 'enhanced' datasets created are named 'steps_1.0', 'steps_0.1', 'direct_1.0', and 'direct_0.1'.

The first prompt was adjusted by trial and error on a small selection of examples, which resulted in the several steps declared. The prompts are linked with a different variable containing the tags listed in Table 5.1, providing the VLM with a set of options to choose from.

```
# Prompt with Steps
"""
**Instructions:**
 1.**name**: Identify a text label from the image that most probably fits the highlighted
     area if it is a facility. The name may be outside the magenta border if there’s a
     clear link (e.g., signage, adjacent facilities, surrounding fences). If no name
     suffices, set it to "NO_NAME".
 *Examples:* "Betoncentrale Van Kamp BV", "NO_NAME", "Tata Steel IJmuiden", etc.
 2.**context_features**: List distinct objects and features *outside* the magenta border
     for context. Focus on industrial, or infrastructural elements. Do NOT include names
     or text from the image.
 *Examples:* "cement truck", "road", "storage", "parking lots", "used water access",
     "railway", "ore piles", "industrial chimney", "red/brown deposit", etc.
 3.**parcel_features**: Now, list distinct objects with size or amount only *inside* the
     magenta border.
 *Examples:* "mixing drum", "loading bay", "few large storage silos", "warehouse", "low
     office building", "large facility", "many silos", etc.
 4.**activity_description**: Based on observations from the previous steps, infer and
     describe the primary activity occurring within the *highlighted* area.
 *Examples:* "production and distribution of concrete", "storage and administration",
     "steel smelting and relocation", "navigation pilots harbour", etc.
 5.**landuse_tag**: Based on observations from the previous steps, classify the
     *highlighted* area’s land use. Choose concise but conventional labels.
 6.**NOx_reasoning**: In one sentence, reason about the direct annual NOx emissions for
     the highlighted area in kilogram/year. Is it a large heavy energy industry facility
     or smaller? etc.
 *Examples:* "one of the largest steel facilities, exceptionally high on energy usage",
     "industrial but not energy demanding besides office buildings and hangars having low
     emissions", etc.
 7.**NOx_emission_estimate**: Now with all the above information, estimate the annual NOx
     emissions for the highlighted area in kilogram/year.
 *Examples:* office buildings: 1, material deposit: 1000, fertilizer production: 14000,
     energy plant: 654000, steel factory: 4138000, etc.
 """
```

The second prompt is created from reducing the first prompt, thus, containing fewer but exact similar steps.

```
# Prompt with more Direct approach
"""
**Instructions:**
 1.**name**: Identify a text label from the image that most probably fits the highlighted
     area if it is a facility. The name may be outside the magenta border if there’s a
     clear link (e.g., signage, adjacent facilities, surrounding fences). If no name
     suffices, set it to "NO_NAME".
 *Examples:* "Betoncentrale Van Kamp BV", "NO_NAME", "Tata Steel IJmuiden", etc.
 2.**parcel_features**: Now, list distinct objects with size or amount only *inside* the
     magenta border.
 *Examples:* "mixing drum", "loading bay", "few large storage silos", "warehouse", "low
     office building", "large facility", "many silos", etc.
 3.**landuse_tag**: Based on observations from the previous steps, classify the
     *highlighted* area’s land use. Choose concise but conventional labels.
 4.**NOx_emission_estimate**: Now with all the above information, estimate the annual NOx
     emissions for the highlighted area in kilogram/year.
 *Examples:* office buildings: 1, material deposit: 1000, fertilizer production: 14000,
     energy plant: 654000, steel factory: 4138000, etc.
 """
```

5.3. Technical details

5.3.1. Parcel Filter Algorithm

It is decided to filter out some peculiarly shaped cadastral parcels to not overload the VLM with too many images based on difficult geometries. These ’difficult’ parcels mainly consist of public infrastructural spaces being subdivided into parcels over time, resulting in complex combinations of shapes. These parcels remain classified as ’*landuse:industrial*’. Filtering involves the *convex-hull* of a geometric shape, which is the smallest periphery on a set of points, see Figure 5.2. Equation 5.1 is used to compute a *shape-simplicity* metric.

$$\textit{Shape Simplicity} = \frac{\text{shape}_{\text{area}}}{\text{convex-hull}_{\text{area}}} \times \frac{\text{shape}_{lenght}}{\text{convex-hull}_{lenght}}\ ,\ \in (0,1] \tag{5.1}$$

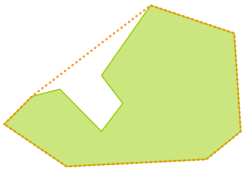


Figure 5.2: Example of an orange *convex-hull* geometry of a green geometry shape. A low area ratio between the two geometries — i.e. much white space — is defined to be more ’complex’.

Figure 5.3 shows an example of what shape-simplicity values for parcels results in. The goal is to filter out the difficult shaped areas, representing infrastructural areas and other peculiar shapes. Therefore, all parcels with a shape-simplicity lower than 0,7 were filtered out. To further reduce the number of objects to be reclassified, all parcels with an area smaller than 1.500 m^2 were filtered out. Figure 5.4 illustrates the filtered parcels from the earlier example. The result is a set of 80.992 parcels that will be reclassification by the VLM.

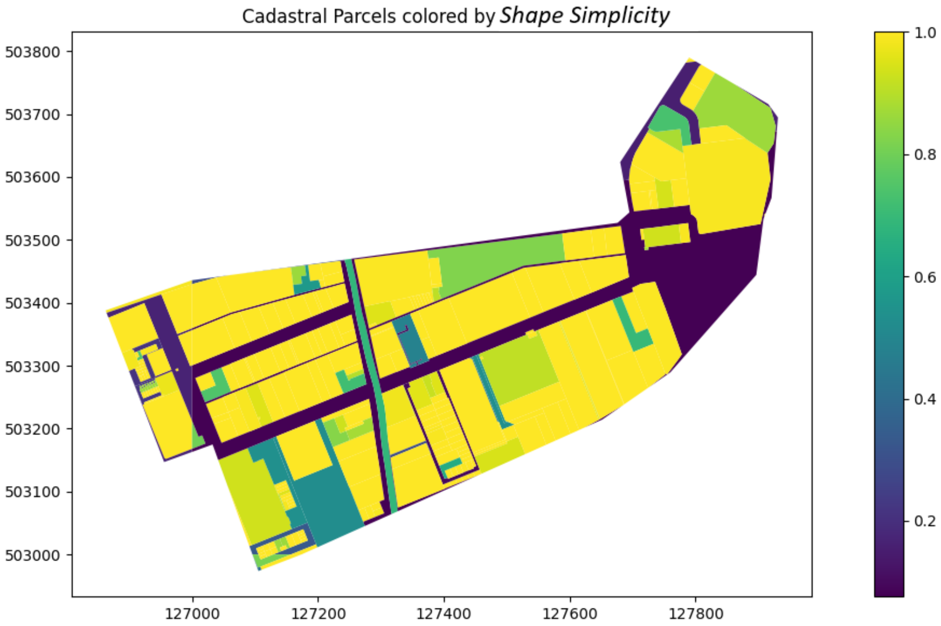


Figure 5.3: Example of what the *shape-simplicity* metric by Equation 5.1 results in for some parcels. The objective is to filter out low values having a more ’complex’ shape, which mainly affects public and infrastructural spaces. OSM id: 6320163.

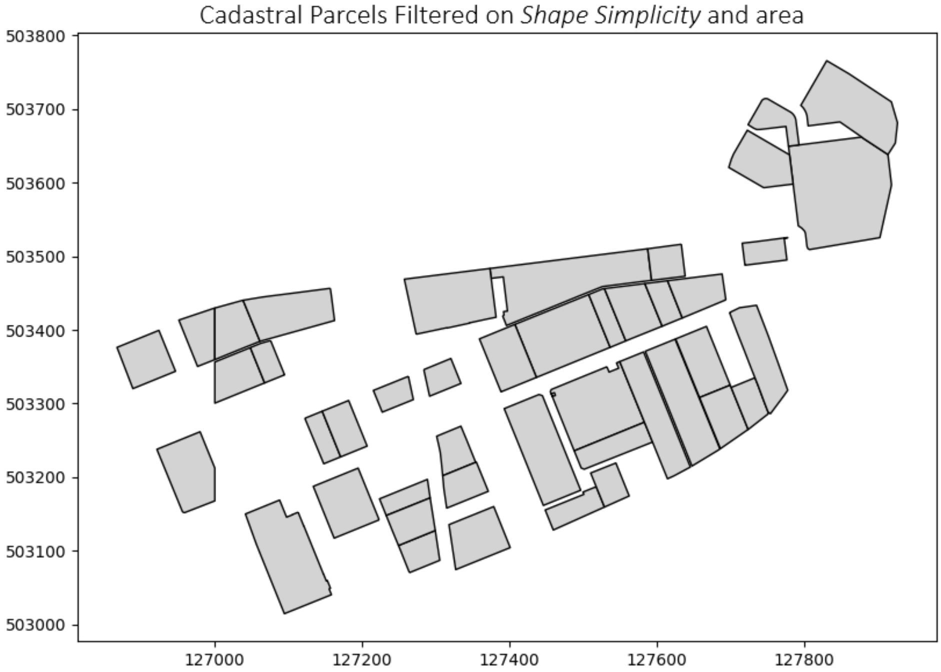


Figure 5.4: '*landuse:industrial*' cadastral parcels after filtering on *shape-simplicity* and area using thresholds of 0.7 and 1.500m^2 respectively. These parcels are used for reclassification by the VLM.

5.3.2. Aerial Image Creation

For aerial images, the PDOK '2022_ortho25' dataset [53] was used, containing raw images where each pixel size represents 25x25cm. It is decided to align the image input with the prompt engineering guidelines from Section 5.2.2. This means that the image will contain context with respect to the parcel, but not too extensive. Therefore, a *context-buffer* has been added for cases that a parcel's geometry is just at the edge of a tile, resulting in more tiles being retrieved and merged. Preventing too much context, a *blackout-buffer* is added, which masks out part of the image based on the parcel's simplified geometry shape. The current buffer is at 0,000.05 geometric degrees, a small shift to buffer and smooth out the geometric shape, only tested for the Netherlands. The final image is clipped to contain limited blacked out parts as possible.

In short, creating an aerial image of a parcel that is used for VLM inferencing involves the following steps:

1. compute, using coordinates, which files are required for the parcel,
2. retrieve tiles from directory or fetch using the API,
3. merge tiles together in one image,
4. draw simplified shape of the parcel's geometry,
5. draw *blackout-buffer* around the context window,
6. reshape image with minimal black area.

Figure 5.5 depicts the results of step 2, showing what tiles have been fetched. Figure 5.6 shows the result of steps 3 to 6. The latter image is input for the VLM which will be used to reclassify the area of the original parcel.

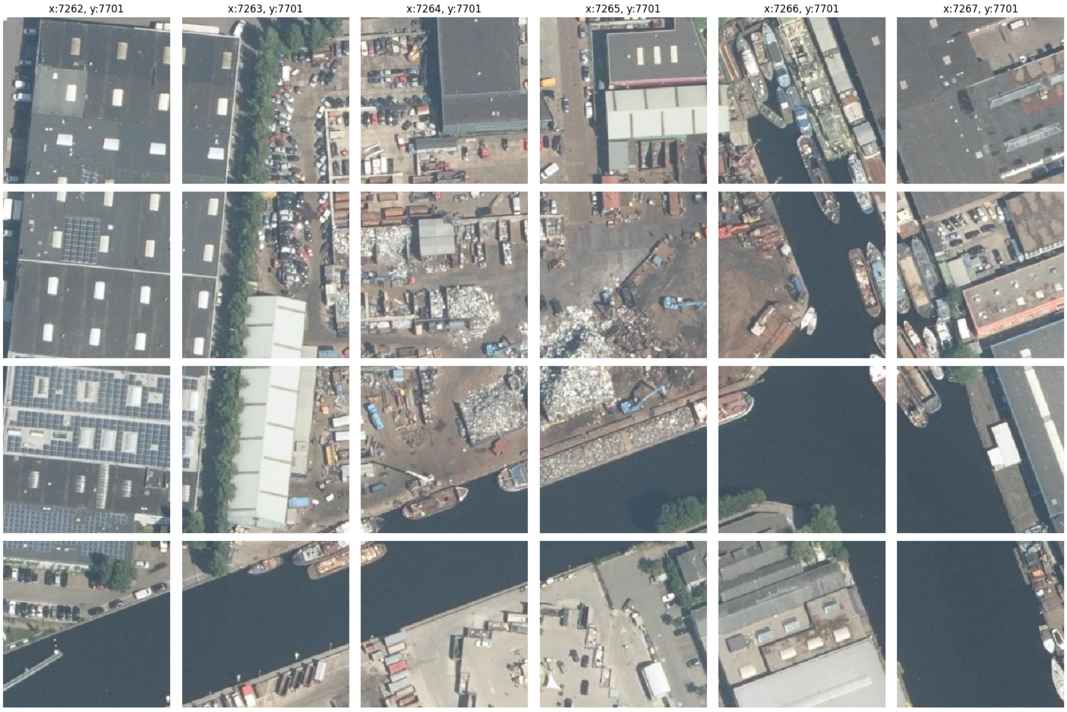


Figure 5.5: Tiles for a single parcel that will be merged. On top their corresponding tile-coordinates. OSM id: 223962763. Tiles source PDOK [53]

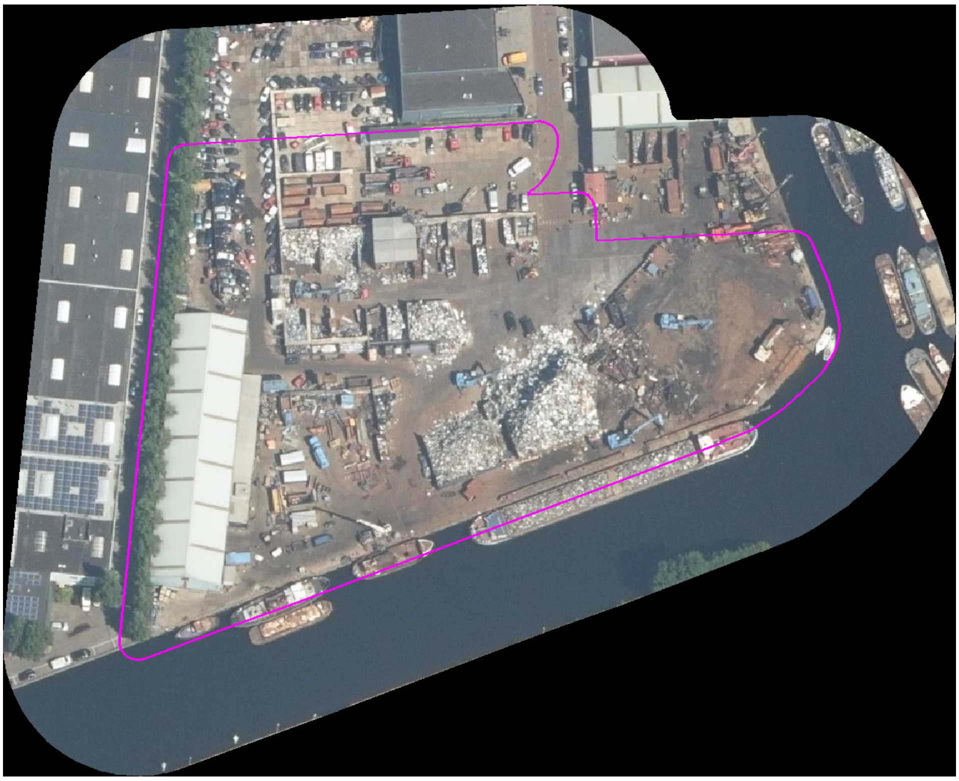


Figure 5.6: The resulting image of merging, drawing and reshaping that will be used as input for the VLM.

6 Results

First, results are presented on outputs of the VLM and how they differ in each of the prompt-temperature combinations. Then the results on the benchmark datasets are presented, to depict how the different model types perform. Finally, bootstrapping on the test set will reveal if any significant improvements have been made by 'enhancing' the benchmark dataset with a VLM.

6.1. Vision Model Output

Table 6.1 displays four different VLM outputs based on the prompts from Section 5.2.2 and the randomly selected aerial image in Figure 5.6. It is observed that the VLM is able to retrieve important information from the image and results in a classification of '*industrial:recycling*' in all four combinations. It must be noted that VLMs are stochastic and thus outputs change when re-inferencing, especially with higher temperature settings.

Prompt-Temperature	**Context**	**Parcel**	**Activity**	**Landuse Tag**	**NO_x Reasoning**	**NO_x Estimate ($kg/km^2/y$)**
direct_0.1	–	scrap metal piles, cranes, barges, warehouses, vehicles, storage tanks	–	industrial:recycling	–	2.234
direct_1.0	–	large metal scraps, several cranes, several warehouses, numerous vehicles, large piles of recycling materials, barges	–	industrial:recycling	–	3.128
steps_0.1	adjacent industrial buildings, road, boats, trees	scrap metal piles, several cranes, barges, warehouse, storage containers, vehicles	The parcel is used for metal recycling and storage, with loading and unloading activities from barges.	industrial:recycling	The site involves heavy machinery for moving and processing scrap metal, as well as transportation via barges, leading to moderate NO_x emissions.	3.128
steps_1.0	adjacent buildings, barges on the river, road	several piles of scrap metal, storage building, mobile cranes, trucks, barges loading/unloading	Metal recycling and processing, with loading and unloading activities via barges.	industrial:recycling	Recycling facility involves heavy machinery and transportation, but lacks the energy demand of smelting or chemical production, leading to moderately high NO_x emissions.	13.404

Table 6.1: The combinations of 'steps' and 'direct' approach with two different temperatures results in four VLM results on one aerial image, see Figure 5.6.

6.1.1. Emission estimates

Figure 6.1 displays for each of the prompt-temperature combinations a histogram on ’*NOx_emission_estimate*’ values for all the 80.992 parcels. To test whether there is significant ’bias’ in prompt-temperature predictions, a OLS model on all the $\log_{10}$ scaled estimates is fitted (`log10_nox_estimate = c(prompt_approach) * c(temperature)`). It is observed that for the ’direct’ prompt, changing the temperature setting does not significantly change the emission estimates. For the ’steps’ prompt changing the temperature results in significant changes: about -13% lower estimates on average for the ’steps_0_1’ combination compared to the other three combinations. This is visually confirmed by the green histogram, showing a higher count of lower emission estimates and lower maximum emissions estimates. This is probably because changing the temperature does affect the columns ’Activity’ and ’NO_x Reasoning’ (see Table 6.1) the most. Why this results in lower emission estimates on average is hard to determine without in-depth investigations. For the full OLS fit see Appendix D.

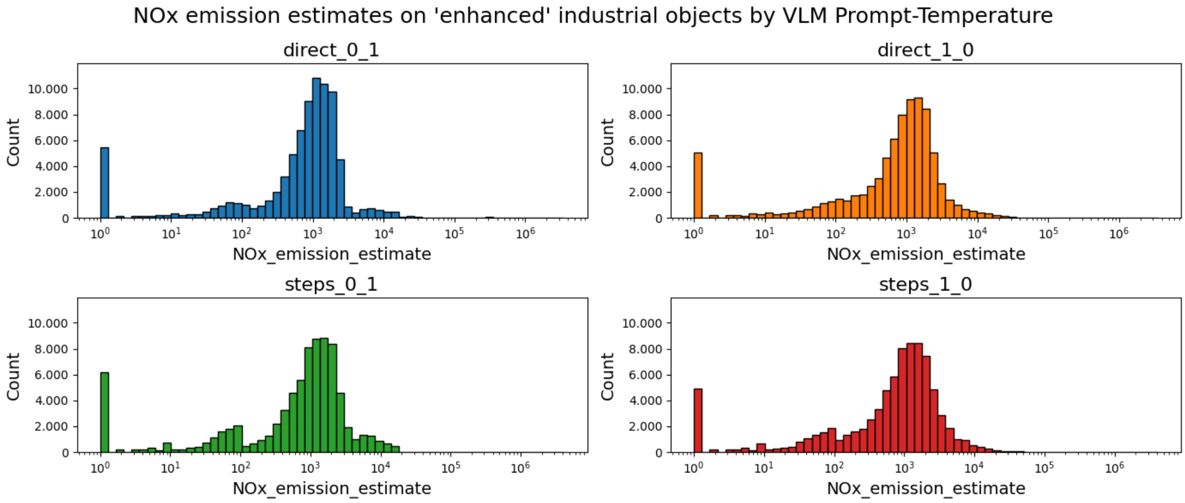


Figure 6.1: ’NOx_emission_estimates’ histogram by the different prompt-temperature combinations.

Additionally, by visual inspection, there seems to be a bias towards certain numeric values, as shown in Figure 6.2. There is a tendency towards numbers present in the prompts (1.000, 654.000, 4.138.000) and rounded numbers (1.000, 1.200, 12.000, 15.000). For both the results created by ’steps_0_1’ and ’steps_1_0’ the maximum emission estimate was 4.138.000, which was listed with ’*steel factory*’ in the prompt. It seems that a higher temperature results in a modest reduction in bias towards rounded numbers when comparing the regions of 15.000 ↔ 17.000.

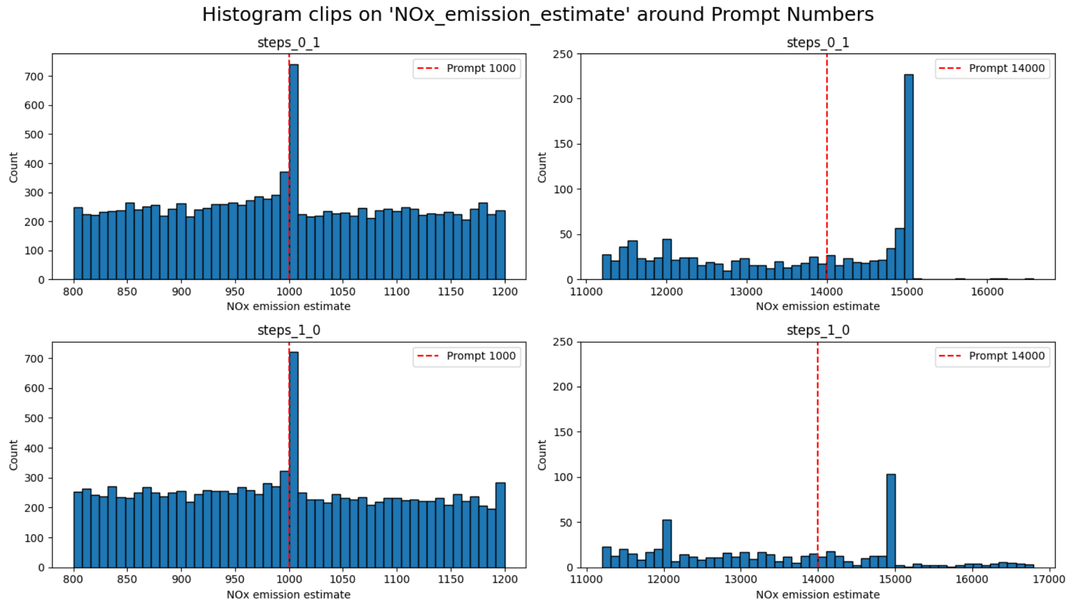


Figure 6.2: Four graphs displaying parts of the 'NOx_emission_estimate' histograms. *Prompt numbers* are present in the prompt used for instructing the VLM, for which the VLM has a tendency towards for selecting as emission estimates.

6.1.2. Tag Classifications

The classifications by the different prompt-temperature combinations results in some differences as could be seen in Figure 6.3. It is hard to uncover why the differences exist without verifying the ground truth of parcels' land-uses, and the fact that the experiments are performed one time on the same parcels with fixed prompt-temperature settings. From Figure 6.4 it is observed that he reclassified area is largest for general industrial land-uses, where '*industrial:logistics*', '*industrial:manufacturing*', and '*building:industrial*' account for 50% of the total. From Figure 6.5 and Figure 6.6 there are indications that the 'NOx Reasoning' step (see Table 6.1) has effect on estimating emissions with more consideration. A clear difference between the two prompting approaches is observed for the tags '*industrial:metal_construction*', '*building:industrial*', and '*industrial:cement*', with the variation likely reflecting closer alignment with actual emissions when evaluated in aggregate.

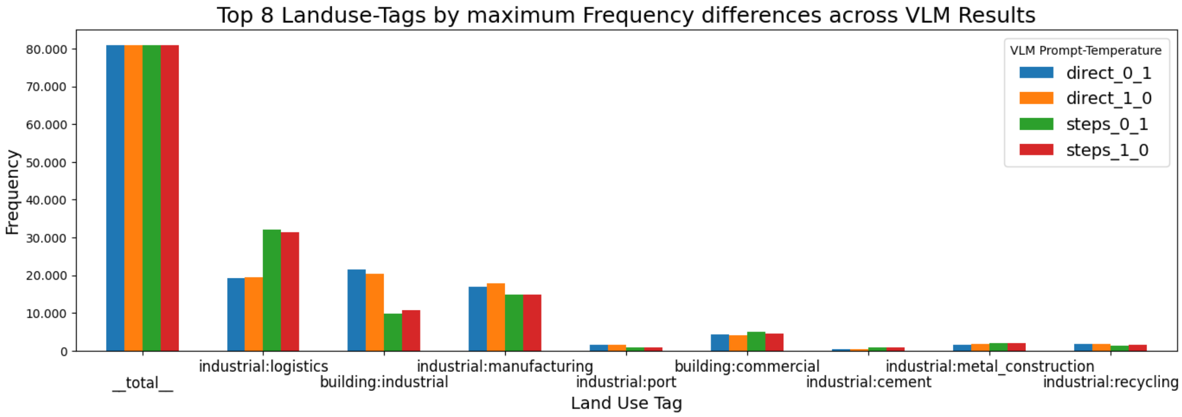


Figure 6.3: Displaying maximum differences between counts of land-use tags used by the four prompt-temperature settings, relative to the total number of classified parcels.

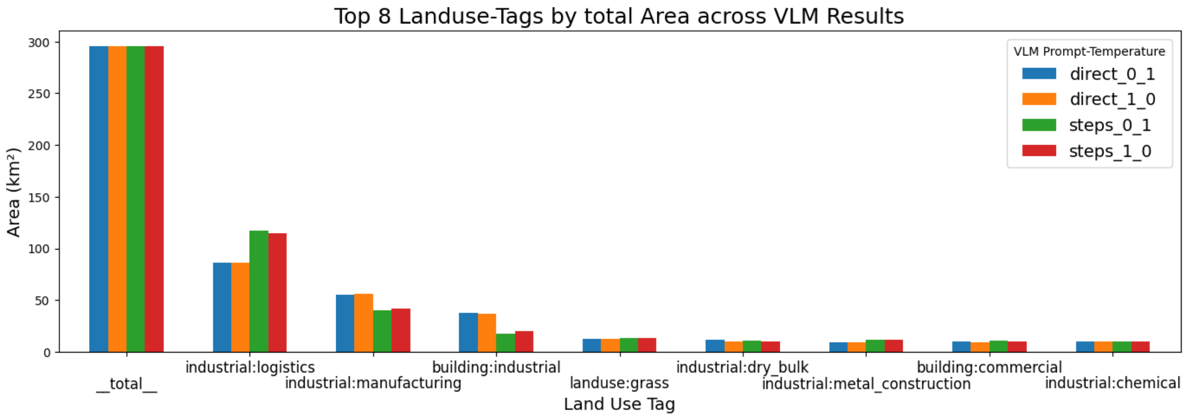


Figure 6.4: Displaying the total area being reclassified from '*landuse:industrial_area*' per land-use tag for each of the prompt-temperature setting.

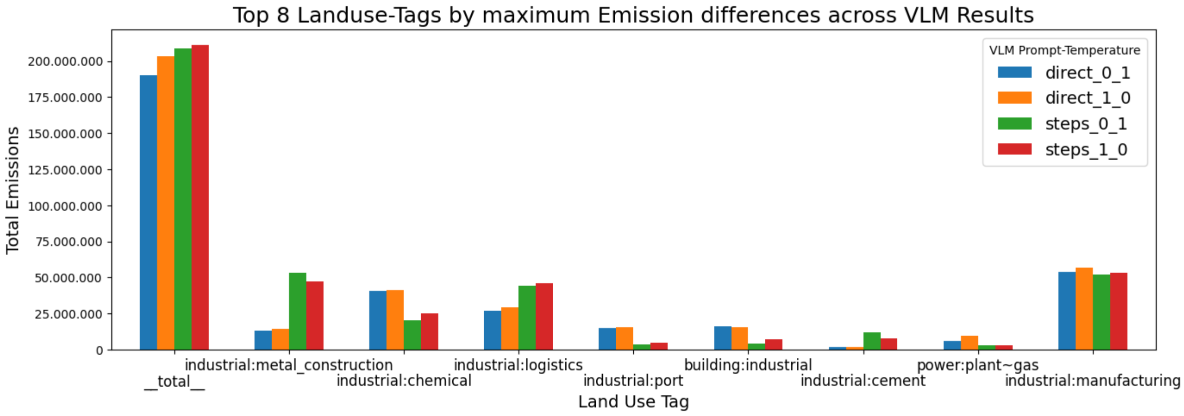


Figure 6.5: The total NO_x emissions being assigned to each of the land-use tags, relative to the total emissions on all the parcels reclassified per prompt-temperature setting.

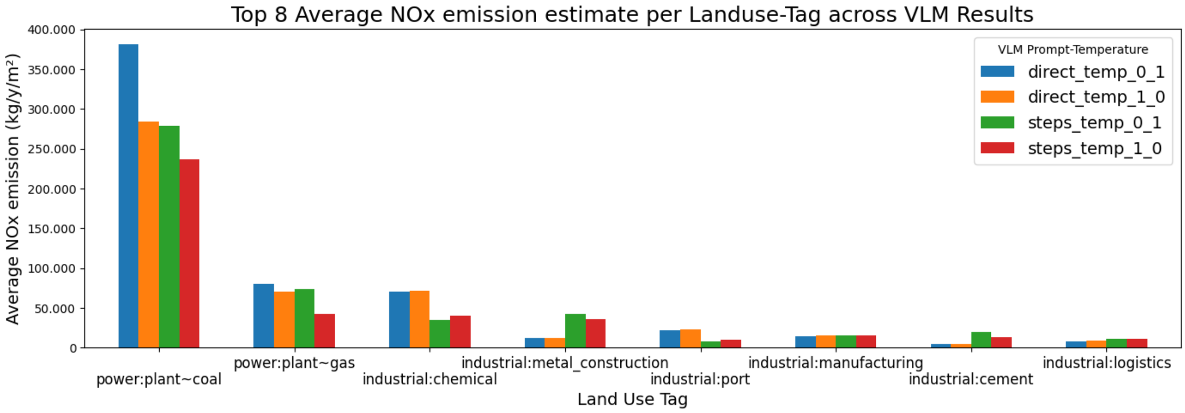


Figure 6.6: The average NO_x emissions being assigned to a single parcel of the land-use tags.

6.2. Models on Benchmark Dataset

All models have been optimised on the RMSE metric with the Optuna framework as described in Section 4.2, see Appendix B for hyperparameter tuning results.

Table 6.2 presents the results of the metrics for different models trained on the *benchmark* dataset. The HGBR model performs best on the evaluation of the pointwise test set (PW) on both the RMSE and R^2 metric and therefore is selected as a baseline. It is observed that models trained on a log-scaled target do not obtain good fits. Both the RF and EN models are statistically worse than the baseline, and the HGBR is almost.

It could not be concluded that the HGBR model trained on the original scale significantly outperforms the RF model, but the EN model also does not perform well. Thus, based on the $\text{RMSE}_{\text{mean}}$ metric, the HGBR model will represent the *benchmark* dataset fit when comparing with the 'enhanced' datasets.

Model	Target	RMSE PW	RMSE 95%-CI	RMSE mean	R^2 PW	R^2 95%-CI	R^2 mean	Sign.
HGBR	$\log_{10}$(nox)	20.593	[7.648 ; 33.433]	18.995	0,13	[0,05 ; 0,49]	0,21	0,87
RF	$\log_{10}$(nox)	21.759	[9.992 ; 34.012]	20.235	0,03	[0,01 ; 0,11]	0,04	worse^{1,0}
EN	$\log_{10}$(nox)	425.804	[163.787;646.141]	400.625	-373	[-2.308;-48]	-603	worse^{1,0}
HGBR	nox	<u>**20.027**</u>	[7.574 ; 32.713]	<u>**18.318**</u>	<u>**0,17**</u>	[0,06 ; 0,60]	<u>**0,27**</u>	baseline
RF	nox	20.053	[7.564 ; 32.821]	18.480	0,17	[0,06 ; 0,59]	0,26	0,58
EN	nox	23.207	[12.438 ; 32.974]	22.391	-0,29	[-2,16 ; 0,25]	-0,35	0,86

Table 6.2: Metrics of fitted models by RMSE optimisation on the **benchmark** dataset. Metrics on the kg/km^2/year scale. The significance levels pertain to both RMSE and R^2, 0,50 is similar to baseline.

Residual plots provide insight into the prediction behaviour of the models. From the graphs in Figures 6.7 and 6.8 it is observed that all models have an upward spread in residuals, showing underestimations on higher emission grids. The RF model trained on log-scale has constant underestimation of emissions, and the EN models have massive over-predictions on lower true emissions.

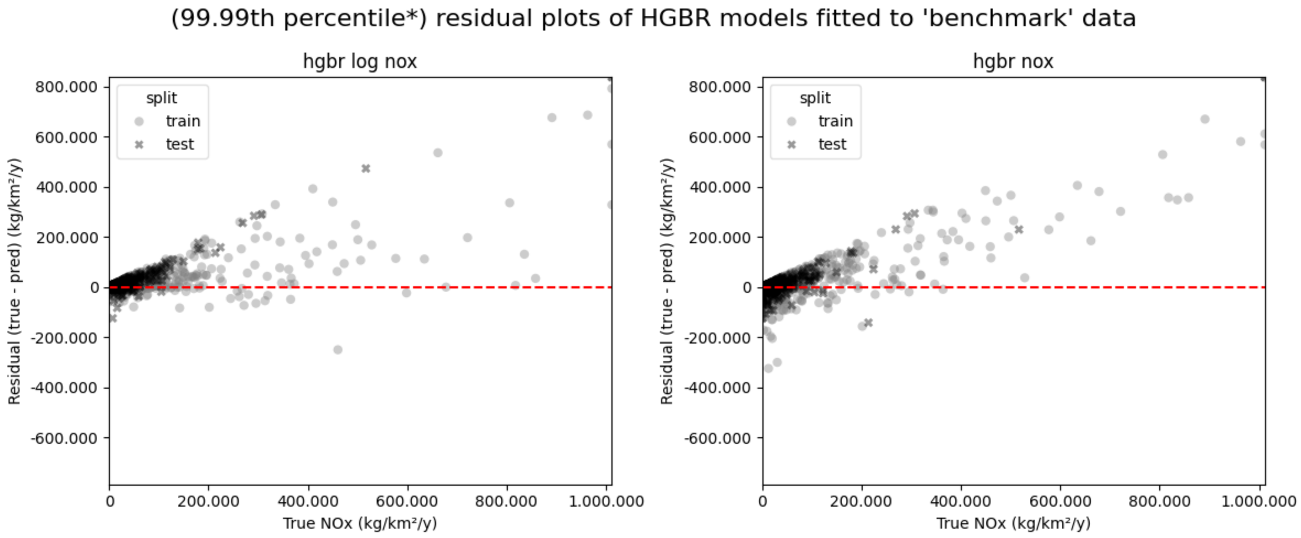


Figure 6.7: Residual graphs of the HGBR-target combinations on the **benchmark** dataset. A model of higher quality would have points closer to the red line, minimizing residuals. The spread of residuals increases with higher true emissions, showing underestimations for both models in high-emission areas.*The remaining percentile points have been clipped to the boundaries of the graph axes.

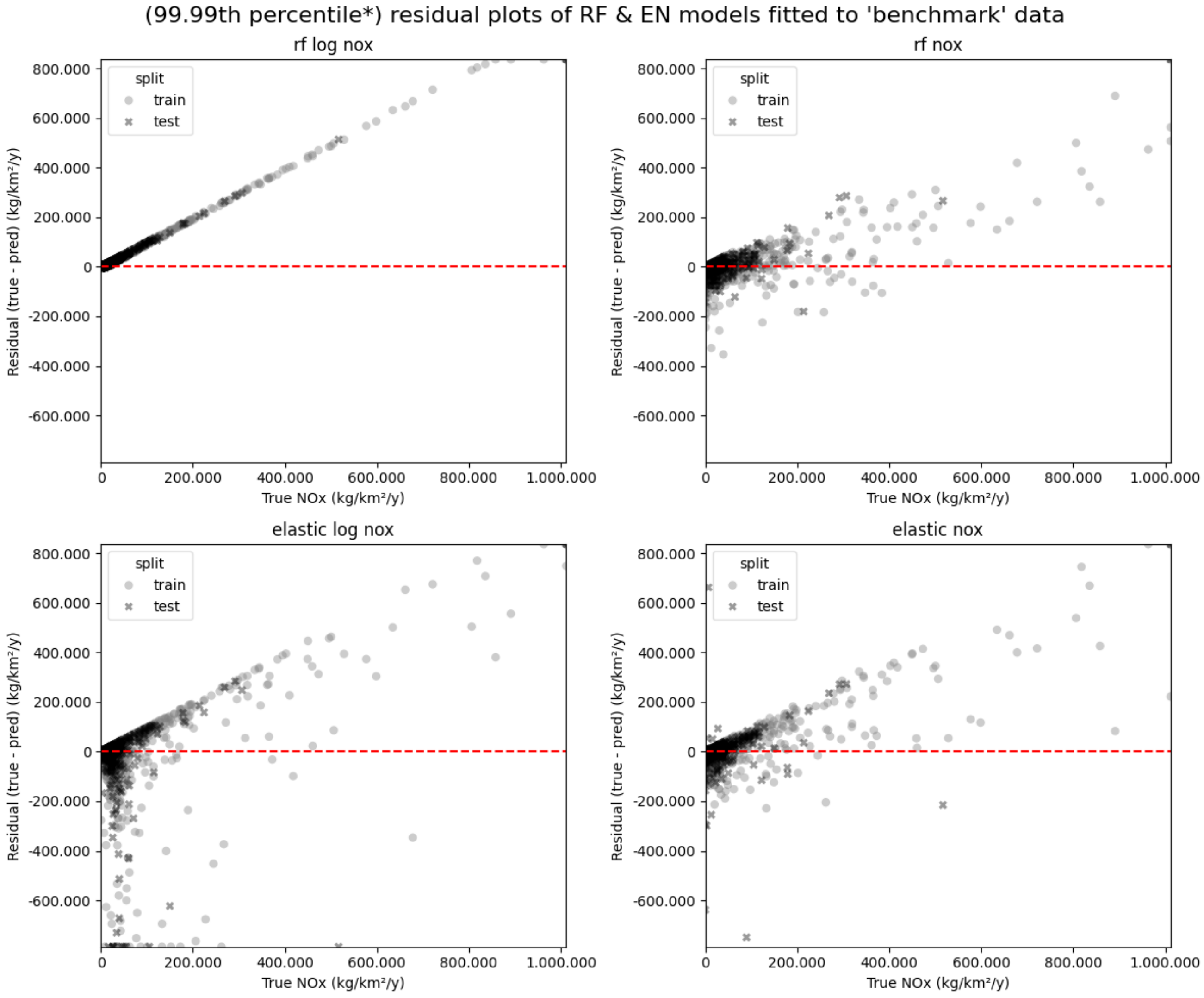


Figure 6.8: Residual graphs of the RF-target and EN-target combinations on the **benchmark** dataset. A model of higher quality would have points closer to the red line, minimizing residuals. The spread of residuals increases with higher true emissions, showing underestimations for both models in high-emission areas. *The remaining percentile points have been clipped to the boundaries of the graph axes.

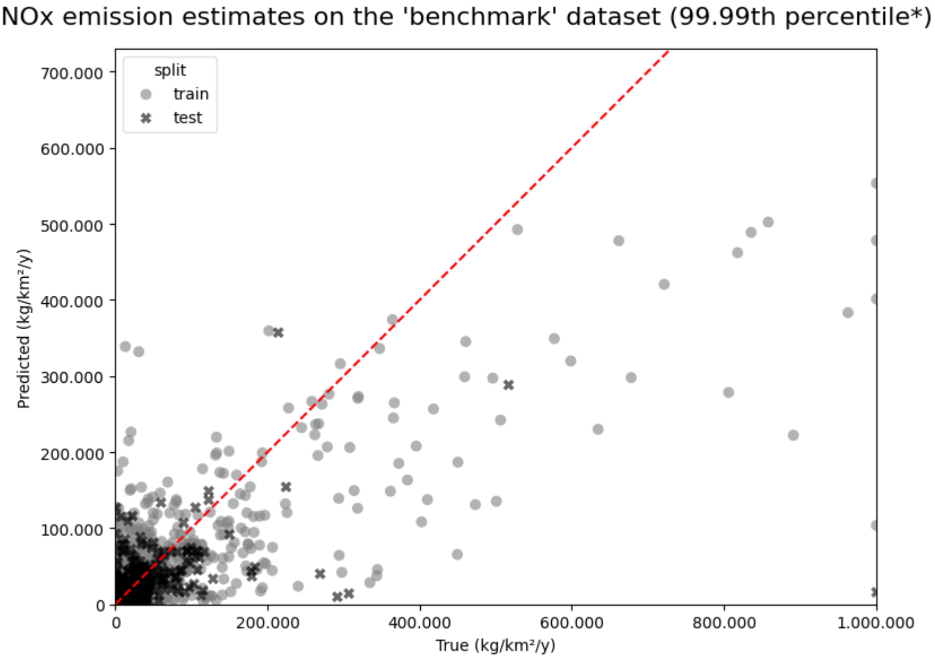


Figure 6.9: Predicted vs. True NO_x plot of the best fitting model on the **benchmark** dataset: HGBR trained on the NO_x scale. Points on the red line would be a perfect prediction.*The remaining percentile points have been clipped to the boundaries of the graph axes.

From the graph in Figure 6.9 it is even clearer that the best model structurally underestimates the high emission values. One reason for the underestimation is the prediction mechanism of Decision Trees, explained in Section 4.1. Since predictions are obtained by averaging the observed values within terminal nodes, the predictions are restricted to the range of its set of data points. This in combination with the sparsity of high-emission km^2 areas limits the likelihood of forming terminal nodes with strong predictive power for grids with exceptionally high emissions. The feature importances presented in Figure 6.10, provide some insight into the incomprehensible nature of tree ensembles. The feature '*landuse:industrial_area*' which will be 'enhanced' is in the top five important features. Furthermore, it is observed that negative features, with respect to direct NO_x emission, have influence on prediction, such as '*generator:electricity:wind*', '*natural:water*' and '*landuse:grass*'. The top 15 features account for 66% of accumulated feature importance.

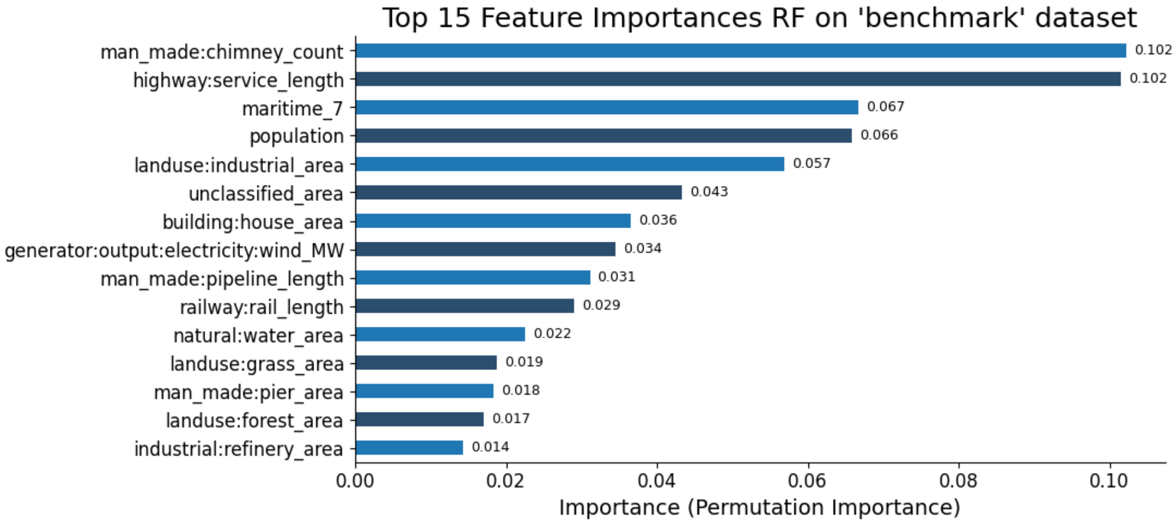


Figure 6.10: This graph shows the relative feature importances of the best-fitting model on the **benchmark** dataset, computed using permutation importance (features are randomly shuffled row-wise to measure the impact on predictive performance) and normalised across all features.

6.3. Models on VLM 'enhanced'Dataset

The training of models on the four newly created 'enhanced' datasets results in similar model fit characteristics as with the '**benchmark**' dataset. Therefore, only the results on the best performing models of each 'enhanced' dataset are presented in Table 6.3. Again, these models have been selected based on the lowest bootstrapped mean RMSE on the test-set of size 8.516. All selected models were trained on the kg/km^2/year scale.

None of the models trained on the 'enhanced' datasets significantly outperform the baseline HGBR model trained on the *benchmark* dataset. The best performing model is an HGBR model trained on the 'direct_1.0' dataset regarding the RMSE$_{mean}$ metric. There remains underestimation of higher emissions in all fits, see Figure C.1 in Appendix .

Dataset	Model	RMSE PW	RMSE 95%-CI	RMSE mean	R^2 PW	R^2 95%-CI	R^2 mean	Sign.
benchmark	HGBR	20.027	[7.574 ; 32.713]	18.318	0,17	[0,06 ; 0,60]	0,27	baseline
direct_0.1	HGBR	19.468	[11.081 ; 28.245]	18.760	0,22	[-0,81 ; 0,40]	0,08	0,47
direct_1.0	HGBR	19.345	[8.268 ; 30.872]	**17.762**	0,23	[0,15 ; 0,54]	**0,28**	0,36
steps_0.1	RF	**18.629**	[9.755 ; 28.860]	17.965	**0,29**	[-0,51 ; 0,44]	0,19	0,41
steps_1.0	HGBR	20.209	[7.801 ; 32.876]	18.426	0,16	[0,05 ; 0,57]	0,25	0,86

Table 6.3: Comparison between best fitted models by RMSE optimisation on different datasets. Dataset: either the benchmark or a VLM 'enhanced'dataset, Metrics on the kg/km^2/y scale.

Changes in the feature importances, see Figure 6.11, with the newly added feature '*NOx_emission_estimate*' prominently on seventh place. The cumulative importance of the top 15 features decreased from 66% to 55%, suggesting that the predictive power is more distributed between features. From Figure 6.13 it becomes clear that there have been great changes in feature importance order. The '*landuse:industrial*' feature has lost predictive importance, which might be caused by three features in green and blue which were altered or added by VLM changes. This is even more extensive on the RF model trained on the '*steps_0.1*' dataset, introducing five new VLM 'enhanced' features and '*NOx_emission_estimate*' topping the list.

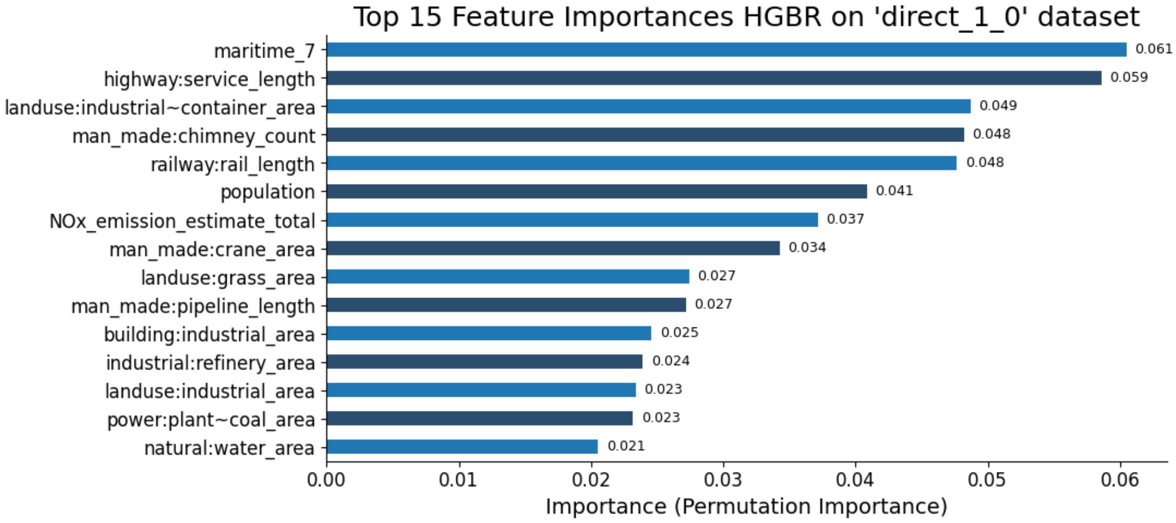


Figure 6.11: Relative feature importances of the best-fitting model, computed using the models built-in feature importance (importance derived from how much each split reduces prediction error during training) and normalised across all features.

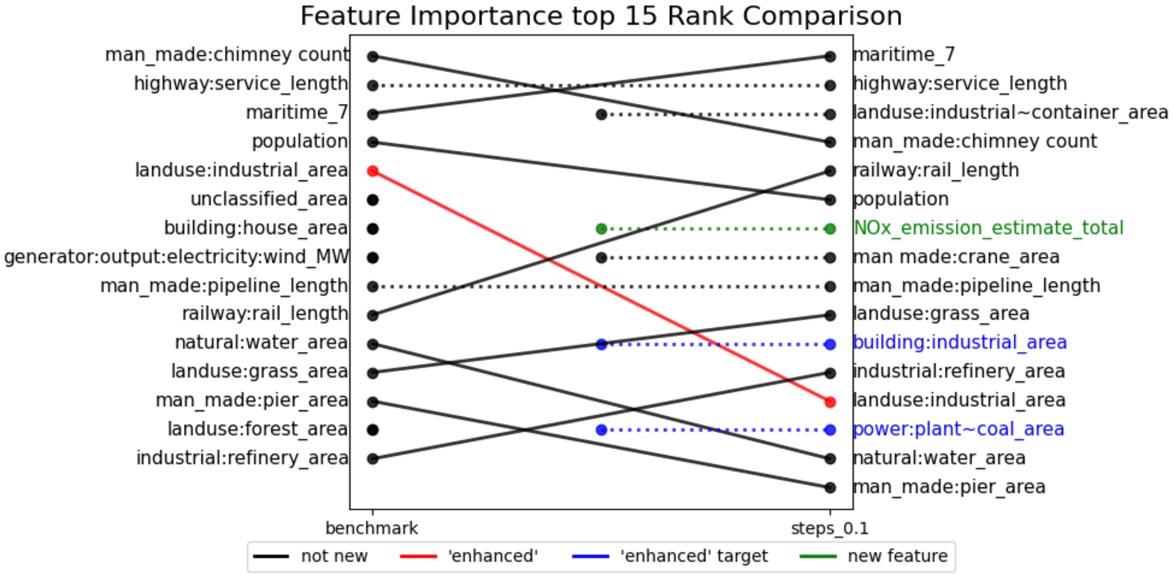


Figure 6.12: Comparison of the top 15 feature importances of the best-fitting models on the ’benchmark’ and ’direct_1.0’ datasets.

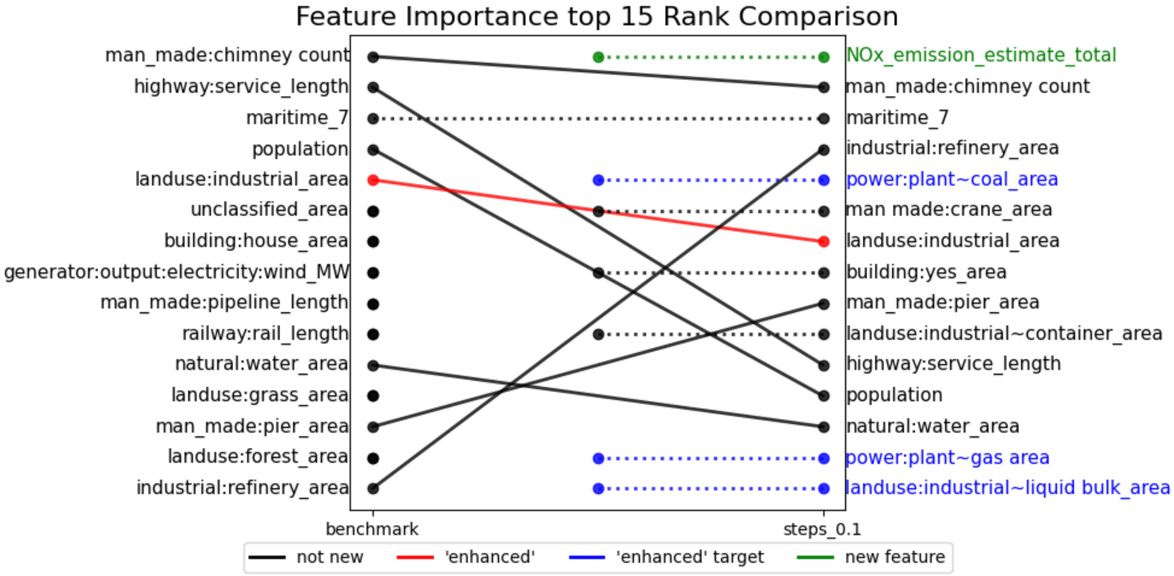


Figure 6.13: Comparison of the top 15 feature importances of the best-fitting models on the ’benchmark’ and ’step_0.1’ datasets.

6.4. Improvements

Table 6.4 shows if significant improvements have been made when computing metrics only for set of the 1 km^2 squares based on VLM ’enhancement’ or not. It is observed that gains, if there are any, of the models trained on the ’enhanced’ datasets with respect to the model trained on the ’benchmark’ dataset is driven by improvements on the ’unenhanced’ grids. It even turns out that the RF model trained on the ’steps_0.1’ dataset is significantly better than the ’unenhanced’ 1 km^2 squares baseline$_u$, decreasing the RMSE from 4.949 to 4.330 kg/km^2/year and increasing R^2 from 0,589 to 0,686. These observations suggest that only the ’unenhanced’ grids benefit from ’enhancing’ grids by the VLM.

		Bootstrapped Metrics					Pointwise Metrics			
Dataset-Model	**km^2 Set**	**RMSE**	**%Δ**	**R^2**	**%Δ**	**Sign.**	**RMSE**	**%Δ2**	**R^2**	**%Δ**
benchmark	ALL	18.318	–	0,269	–	BL$_a$	20.027	–	0,174	–
HGBR	ENH	48.138	–	0,184	–	BL$_e$	53.560	–	0,098	–
	UNE	4.949	–	0,589	–	BL$_u$	4.971	–	0,590	–
direct_0.1	ALL	18.760	-2,4%	0,078	-71,0%	0,47	19.468	2,8%	0,219	25,9%
HGBR	ENH	50.227	-4,3%	-0,208	–	0,47	52.118	2,7%	0,146	49,0%
	UNE	4.727	4,5%	0,626	6,3%	0,11	4.746	4,5%	0,626	6,1%
direct_1.0	ALL	**17.762**	**3,0%**	**0,283**	**5,2%**	0,36	19.345	3,4%	0,229	31,6%
HGBR	ENH	48.094	0,1%	0,166	-9,8%	0,36	51.789	3,3%	0,157	60,2%
	UNE	4.684	5,4%	0,632	7,3%	0,17	4.713	5,2%	0,631	6,9%
steps_0.1	ALL	17.965	1,9%	0,192	-28,6%	0,41	**18.629**	**7,0%**	**0,285**	**63,8%**
RF	ENH	47.840	0,6%	-0,049	–	0,42	49.990	6,7%	0,214	118,4%
	UNE	4.330	12,5%	0,686	16,5%	0,02	4.337	12,7%	0,688	16,6%
steps_1.0	ALL	18.426	-0,6%	0,253	-5,9%	0,86	20.209	-0,9%	0,159	-8,6%
HGBR	ENH	48.996	-1,8%	0,148	-19,6%	0,97	54.256	-1,3%	0,074	-24,5%
	UNE	4.642	6,2%	0,638	8,3%	0,23	4.657	6,3%	0,640	8,5%

Table 6.4: Comparison between the model trained on the benchmark dataset and the models trained on the VLM ’enhanced’datasets. The first two columns denote the dataset and model trained on the kg/km^2/year target scale, and set of the 1 km^2 grids: ALL are the 8.516 test set km^{2}s, ENH only the 1.127 *enhanced* km^{2}s, and UNE the 7.389 *un-enhanced* km^2 grids. The ’Point Metrics’ columns denote metrics, computed pointwise on the testset, and improvements with respect to the model trained on the *benchmark* dataset. ’Bootstrapped Metric’ columns denote metrics, computed on a set of metrics obtained by bootstrapping the testset, and improvements with respect to the model trained on the *benchmark* dataset. BL$_x$ mark the baseline on certain km^2 sets.

To better understand what might cause the effects, Table 6.5 shows the same columns but computed for the mean absolute error error (MAE). This is a confirmation of the same effect: ’enhanced’ grids being predicted worse than ’unenhanced’ grids. However, using this metric, there are two models trained on ’enhanced’ datasets statistically outperforming the model trained on the *benchmark* dataset.

The marginal improvements in RMSE of the HGBR model trained on the ’direct_1.0’ dataset after bootstrapping are 3,0% on the entire dataset, driven by 5,4% improvement on the ’unenhanced’ grids and only 0,1% improvement on the ’enhanced’ grids. Figures 6.14 and 6.15 show that the improvement in predictions occurs along the entire range of true emissions and is similar in specific regions.

		Bootstrapped Metrics			
Dataset-Model	**km^2 Set**	**MAE**	**%Δ**	**Sign.**	**Interpr.**
benchmark	ALL	2.749	–	BL_a	BL_a
HGBR	ENH	8.177	–	BL_e	BL_e
	UNE	1.919	–	BL_u	BL_u
direct_0.1	ALL	2.826	-2.8%	0,76	–
HGBR	ENH	10.031	-22.6%	0,99	worse
	UNE	1.734	9.7%	0,00	better
direct_1.0	ALL	**2.526**	**8.1%**	0,00	better
HGBR	ENH	9.098	-11.2%	1,00	worse
	UNE	1.534	20.1%	0,00	better
steps_0.1	ALL	2.602	5.3%	0,09	–
RF	ENH	9.179	-12.2%	0,91	–
	UNE	1.595	16.9%	0,00	better
steps_1.0	ALL	2.544	7.4%	0,00	better
HGBR	ENH	8.692	-6.3%	0,98	worse
	UNE	1.608	16.2%	0,00	better

Table 6.5: See caption Figure 6.4, now with the mean absolute error metric (MAE).

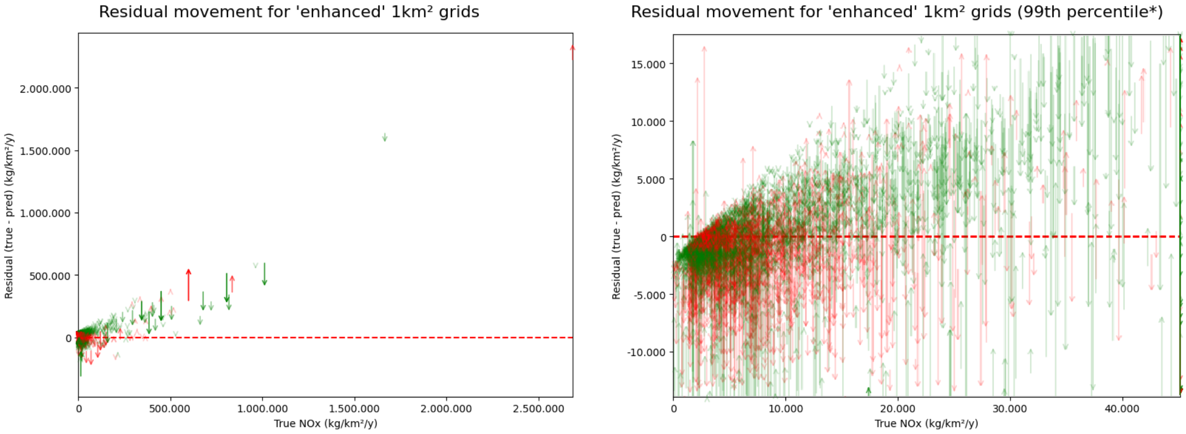


Figure 6.14: Both of these graphs are two residual plots combined of the best models trained on the '**benchmark**' and the '**direct_1,0**' datasets. The arrows denote the shifts in predictions ('benchmark' → 'direct_1,0') on each of the 'VLM-enhanced' 1 km^2 squares, where a green arrow denotes an improved prediction, and a red arrow is a worsened prediction.

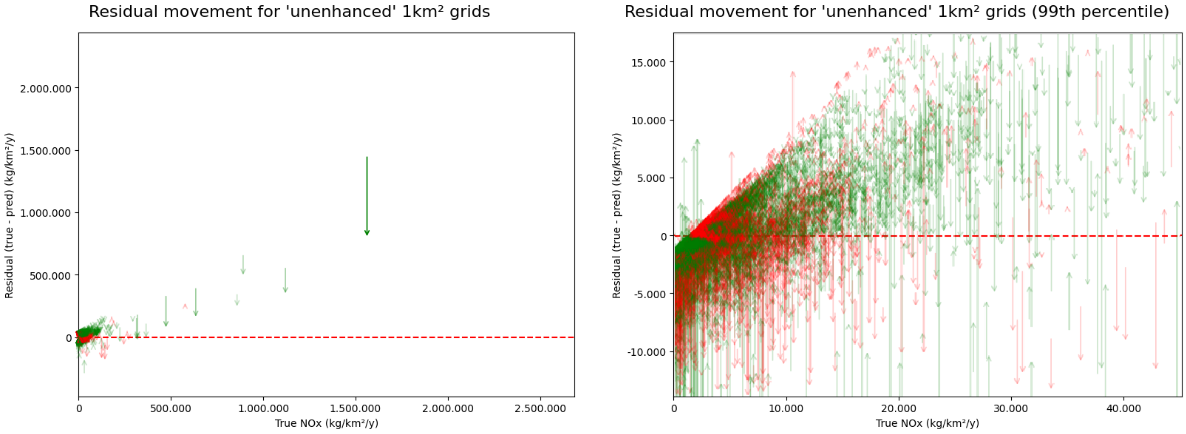


Figure 6.15: See caption of Figure 6.14, now on '*unenhanced*' grids.

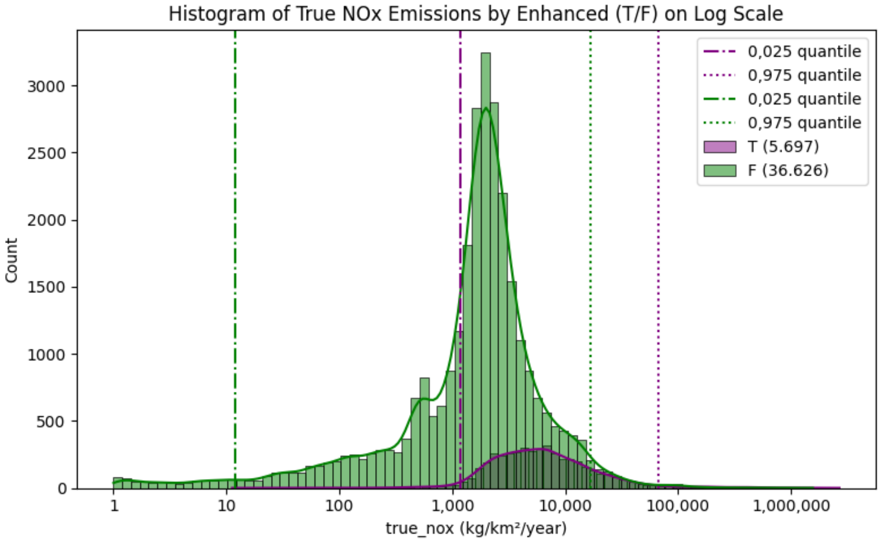


Figure 6.16: Histogram of target NO_x emissions split between the *enhanced* and *unenhanced* 1 km^2 grids.

The ’enhanced’ grids are substantially harder to predict. Across all trained models, their RMSE for ’enhanced’ grids is about 10 times higher than for ’unenhanced’ grids. Since higher emission values are harder to model, and Figure 6.16 shows the ’enhanced’ set contains mostly high emission values, the ’enhanced’ set results in much higher errors. This indicates that ’enhanced’ grids represent fundamentally different or more complex regions.

In addition, the ’enhanced’ subset is also much smaller (≈ 5,7k grids) compared to the ’unenhanced’ subset (≈ 36,6k grids). Since ’enhanced’ grids form a minority of the dataset, they have less influence during model training. Models optimise for global improvements and thus sacrifice accuracy on the ’enhanced’ subset over improvements on the ’unenhanced’ subset.

The effect is visible in RMSE and R^2. RMSE penalises large errors more, amplifying poor performance on ’enhanced’ grids. MAE, being less sensitive to large errors, amplifies the effect more clearly. Furthermore, this metric shows significant differences due to tighter confidence intervals, which confirms the effect.

Depending on the chosen metric, models trained on the VLM ’enhanced’ datasets show statistically significant improvements or not relative to the model trained on the benchmark dataset. Since this research focusses on the RMSE as the primary metric, no statistically significant improvement has been made.

7 Conclusion

This research demonstrates that enhancing the quality of open source data for Land Use Regression (LUR) by a Vision Language Model (VLM) is a promising approach. By aerial imagery analysis, the VLM successfully extracts meaningful features that enable tree ensembles to make improved predictions, as evidenced by the results on the RIVM NO_x 2022 dataset. Models trained on 'enhanced' datasets achieve improvements on pointwise RMSE, but lack statistical significance when bootstrapping the testset.

The investigation revealed several critical insights into behaviour of the VLM. Although these models have a robust capacity to interpret aerial images and infer critical context values, their decision-making processes remain largely opaque — making it challenging to pinpoint their reasoning driving predictions. Prompt engineering emerged as a critical factor, with subtle adjustments affecting outcomes massively. Introducing steps that generate reasoning-context enhanced understanding of emissions by the VLM, resulting in better predictions. Furthermore, a recurring bias toward round numbers or values explicitly mentioned in prompts was observed in numeric outputs, complicating proper emission estimation. Temperature parameter adjustments mitigated this bias to some extent and improved reasoning, though significant bias persisted.

A notable challenge arose in predicting high emission values, where all trained models constantly underestimated true high emissions. This underscores the inherent limitations in capturing extreme values with Decision Tree-based models. The research approach handled imbalance by creating a train-test split with stratified sampling, further strategies are needed to increase representation of high-emission grids.

This research lacks a thorough validation on the ground truth of land-use activities of the parcels, and therefore no strong claims could be made about the correctness of the VLM output. Nevertheless, the reclassification of objects from '*landuse:industrial*' to more specific land uses, together with the introduction of the new feature '*NOx_emission_estimate*' by the VLM, did influence the models trained on the enhanced data. Inspecting the feature importances shows that the contribution of the '*NOx_emission_estimate*' was greater than areal features. The impact of the reclassified area features may also have been affected by the the way overlap resolving was implemented. Overall, however, these features did not provide the models with sufficient predictive power for the 'enhanced' grids, which predominantly consisted of high-emission areas.

Although learning to replicate RIVM NO_x emissions remains a challenge, the research achieved meaningful results, demonstrating that open-source data enhanced by a VLM has the potential to complement LUR modelling with a statistically significant difference. The findings establish a promising new direction for accurate environmental modelling.

During the course of this research, a number of possible improvements were identified. The following list presents these, ordered from minor adaptations to substantial modifications and new approaches.

- Create a tailored list from which the VLM can select tags, some tags might be missing, too similar or redundant.
- Increase the size of parcels being reclassified by the VLM, by lowering the thresholds of filtering, or including more objects than '*landuse:industrial*'. Some outliers in OSM tags were observed.
- Retrieve more or different output features, adapt VLM parameters, and test different VLM models.
- Adapt each prompt per parcel to include additional available information.
- Enable Google Search grounding when inferencing with the VLM, which allows for searching of relevant information.
- Create multi-level prompts, allowing for follow-up questions based on certain results.
- Explore contributing to OSM, correcting mistakes or improving labels.
- Incorporate more data features emission estimation, like emission reports, weather, or energy usage.
- Get a theoretical understanding of VLM model outputs by testing on a 'ground-truth' dataset, for example the EPRTR dataset [10], and tune a model for specific tasks.
- Build an architecture for a data-driven method of the RIVM approach, including layers based on different models.

References

[1] W. H. O. R. O. f. Europe, “Review of evidence on health aspects of air pollution: REVIHAAP project: Technical report,” 2021, Accepted: 2021-06-10T12:33:47Z Number: WHO/EURO:2013-4101-43860-61757 Publisher: World Health Organization. Regional Office for Europe. [Online]. Available: https://iris.who.int/handle/10665/341712 (visited on 02/13/2025).

[2] C. M. Clark, Y. Bai, W. D. Bowman, *et al.*, “Nitrogen deposition and terrestrial biodiversity,” *In: Levin S.A. (ed.) Encyclopedia of Biodiversity, second edition, Volume 5, Waltham, MA: Academic Press. pp. 519-536*, vol. 5, pp. 519–536, 2013. DOI: 10.1016/b978-0-12-384719-5.00366-x. [Online]. Available: https://research.fs.usda.gov/treesearch/44835 (visited on 02/14/2025).

[3] B. J. Cardinale, J. E. Duffy, A. Gonzalez, *et al.*, “Biodiversity loss and its impact on humanity,” *Nature*, vol. 486, no. 7401, pp. 59–67, Jun. 2012, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature11148. [Online]. Available: https://www.nature.com/articles/nature11148 (visited on 02/14/2025).

[4] D. Tilman, P. B. Reich, and F. Isbell, “Biodiversity impacts ecosystem productivity as much as resources, disturbance, or herbivory,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. 10 394–10 397, Jun. 26, 2012, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1208240109. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1208240109 (visited on 02/14/2025).

[5] CBS. “Stikstofemissies naar lucht,” Centraal Bureau voor de Statistiek. Last Modified: 18-08-2025T11:43:36. (2025), [Online]. Available: https://www.cbs.nl/nl-nl/dossier/dossier-stikstof/stikstofemissies-naar-lucht (visited on 08/19/2025).

[6] H. RIVM, “Methodology for the calculation of emissions to air from the sectors energy, industry and waste,” 2022-0001, 2022. [Online]. Available: https://doi.org/10.21945/RIVM-2022-0001.

[7] RIVM. “Methodology for the calculation of emissions from agriculture. calculations for methane, ammonia, nitrous oxide, nitrogen oxides, non-methane volatile organic compounds, fine particles and carbon dioxide emissions using the national emission model for agriculture (NEMA).” (2023), [Online]. Available: https://rivm.openrepository.com/entities/publication/4b15e79e-aa60-4eb6-95c2-cb46b2254aca (visited on 02/19/2025).

[8] Rijksoverheid. “Alle emissiegegevens op één plek | emissieregistratie,” Emissieregistratie NL. (2025), [Online]. Available: https://www.emissieregistratie.nl/ (visited on 03/17/2025).

[9] G. Hoek, R. Beelen, K. de Hoogh, *et al.*, “A review of land-use regression models to assess spatial variation of outdoor air pollution,” *Atmospheric Environment*, vol. 42, no. 33, pp. 7561–7578, Oct. 1, 2008, ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2008.05.057. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231008005748 (visited on 01/06/2025).

[10] EU, *Industrial reporting under the industrial emissions directive 2010/75/EU and european pollutant release and transfer register regulation (EC) no 166/2006 - ver. 13.0 dec. 2024*, version 13, Type: dataset, Sep. 1, 2024. [Online]. Available: https://sdi.eea.europa.eu/catalogue/srv/api/records/ff47e25d-5d4c-491d-b9ce-de17ca61fe6d?language=all.

[11] W. W. Walters, B. D. Tharp, H. Fang, B. J. Kozak, and G. Michalski, “Nitrogen isotope composition of thermally produced NOx from various fossil-fuel combustion sources,” *Environmental Science & Technology*, vol. 49, no. 19, pp. 11 363–11 371, Oct. 6, 2015, ISSN: 1520-5851. DOI: 10.1021/acs.est.5b02769.

[12] M. R. Beychok, “Nox emission from fuel combustion controlled,” *Oil Gas J*, vol. 1, pp. 53–56, 1973.

[13] R. J. Wild, W. P. Dubé, K. C. Aikin, *et al.*, “On-road measurements of vehicle NO2/NOx emission ratios in denver, colorado, USA,” *Atmospheric Environment*, vol. 148, pp. 182–189, Jan. 1, 2017, ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2016.10.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231016308469 (visited on 03/04/2025).

[14] W. H. Organization, *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary.* World Health Organization, 2021, Accepted: 2021-09-22T05:28:17Z, ISBN: 978-92-4-003443-3. [Online]. Available: https://iris.who.int/handle/10665/345334 (visited on 02/14/2025).

[15] E. Council, “DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON AMBIENT AIR QUALITY AND CLEANER AIR FOR EUROPE,” The council European Union, Strasbourg, Oct. 23, 2024, p. 164. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2024/10/14/air-quality-council-gives-final-green-light-to-strengthen-standards-in-the-eu/ (visited on 02/14/2025).

[16] M. Zara, K. F. Boersma, H. Eskes, *et al.*, “Reductions in nitrogen oxides over the netherlands between 2005 and 2018 observed from space and on the ground: Decreasing emissions and increasing o3 indicate changing NOx chemistry,” *Atmospheric Environment: X*, vol. 9, p. 100 104, Jan. 1, 2021, ISSN: 2590-1621. DOI: 10.1016/j.aeaoa.2021.100104. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590162121000046 (visited on 02/12/2025).

[17] Luchtmeetnet. “Luchtmeetnet.nl,” Luchtmeetnet.nl. (2025), [Online]. Available: https://www.luchtmeetnet.nl/ (visited on 02/19/2025).

[18] RIVM. “Stikstofdioxide in lucht, 1992-2023 | Compendium voor de Leefomgeving.” (Jul. 22, 2024), [Online]. Available: https://www.clo.nl/indicatoren/nl023119-stikstofdioxide-in-lucht-1992-2023 (visited on 02/19/2025).

[19] F. Lautenschlager, M. Becker, K. Kobs, *et al.*, “OpenLUR: Off-the-shelf air pollution modeling with open features and machine learning,” *Atmospheric Environment*, vol. 233, p. 117 535, Jul. 15, 2020, ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2020.117535. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231020302703 (visited on 12/10/2024).

[20] A. Radford, J. W. Kim, C. Hallacy, *et al.*, *Learning transferable visual models from natural language supervision*, Feb. 26, 2021. DOI: 10.48550/arXiv.2103.00020. arXiv: 2103.00020[cs]. [Online]. Available: http://arxiv.org/abs/2103.00020 (visited on 02/26/2025).

[21] A. Krizhevsky, V. Nair, and G. Hinton, *CIFAR-10 and CIFAR-100*, Type: dataset, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[22] F. Liu, D. Chen, Z. Guan, *et al.*, “RemoteCLIP: A vision language foundation model for remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024, Conference Name: IEEE Transactions on Geoscience and Remote Sensing. [Online]. Available: https://ieeexplore.ieee.org/document/10504785 (visited on 02/28/2025).

[23] F. Pan, S. Jeon, B. Wang, F. Mckenna, and S. X. Yu, *Zero-shot building attribute extraction from large-scale vision and language models*, Dec. 19, 2023. DOI: 10.48550/arXiv.2312.12479. arXiv: 2312.12479[cs]. [Online]. Available: http://arxiv.org/abs/2312.12479 (visited on 02/25/2025).

[24] J. Roberts, K. Han, and S. Albanie, *SATIN: A multi-task metadataset for classifying satellite imagery using vision-language models*, Apr. 23, 2023. DOI: 10.48550/arXiv.2304.11619. arXiv: 2304.11619[cs]. [Online]. Available: http://arxiv.org/abs/2304.11619 (visited on 02/26/2025).

[25] M. Steininger, K. Kobs, A. Zehe, F. Lautenschlager, M. Becker, and A. Hotho, “MapLUR: Exploring a new paradigm for estimating air pollution using deep learning on map images,” *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 3, 19:1–19:24, Apr. 15, 2020, ISSN: 2374-0353. DOI: 10.1145/3380973. [Online]. Available: https://dl.acm.org/doi/10.1145/3380973 (visited on 12/13/2024).

[26] Christopher Frey, Jim Penman, Lisa Hanle, Suvi Monni, and Stephen Ogle, “IPCC guidelines for national greenhouse gas inventories,” I, 2006.

[27] Rijksoverheid. “Taakgroepen en verantwoordelijkheden | emissieregistratie,” Taakgroepen en verantwoordelijkheden. (2025), [Online]. Available: https://www.emissieregistratie.nl/over-emissieregistratie/organisatie/taakgroepen-en-verantwoordelijkheden (visited on 03/17/2025).

[28] RIVM. “Doorzoek alle documenten | emissieregistratie,” Documenten Emissieregistratie. (2025), [Online]. Available: https://www.emissieregistratie.nl/documentatie/doorzoek-alle-documenten (visited on 03/18/2025).

[29] RIVM. "Wegverkeer (exclusief bussen openbaar vervoer)," Wegverkeer (exclusief bussen openbaar vervoer). (2022), [Online]. Available: https://legacy.emissieregistratie.nl/erpubliek/documenten/07%20Ruimtelijke%20verdeling/Factsheets/Wegverkeer%20(exclusief%20bussen%20openbaar%20vervoer).pdf (visited on 03/18/2025).

[30] TNO, "Real-world fuel consumption of passenger cars and light commercial vehicles," TNO 2020 R11664, Oct. 30, 2020, p. 53.

[31] HDA. "Population 2022 netherlands," Humanitarian Data Exchange. Type: dataset Modified: 30 June 2022 Release 2022-06-30. (Mar. 2025), [Online]. Available: https://data.humdata.org/dataset/kontur-population-netherlands.

[32] Wikipedia. "Classification of european inland waterways," Wikipedia. (Mar. 20, 2025), [Online]. Available: https://en.wikipedia.org/wiki/Classification_of_European_Inland_Waterways.

[33] OSM. "OpenStreetMap," OpenStreetMap. (2025), [Online]. Available: https://www.openstreetmap.org/about (visited on 08/19/2025).

[34] OpenStreetMap. "Stats - OpenStreetMap wiki." (2025), [Online]. Available: https://wiki.openstreetmap.org/wiki/Stats (visited on 08/18/2025).

[35] S. A. Jochen Topf. "OSM_tag statistics," tagInfo. (2025), [Online]. Available: https://taginfo.openstreetmap.org/tags.

[36] Q. Zhou, S. Wang, and Y. Liu, "Exploring the accuracy and completeness patterns of global land-cover/land-use data in OpenStreetMap," *Applied Geography*, vol. 145, p. 102 742, Aug. 1, 2022, ISSN: 0143-6228. DOI: 10.1016/j.apgeog.2022.102742. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143622822001138 (visited on 08/18/2025).

[37] K. Zhang, S. Batterman, and F. Dion, "Vehicle emissions in congestion: Comparison of work zone, rush hour and free-flow conditions," *Atmospheric Environment*, vol. 45, no. 11, pp. 1929–1939, Apr. 1, 2011, ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2011.01.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231011000586 (visited on 04/07/2025).

[38] T. Johnson and A. Joshi, "Review of vehicle engine efficiency and emissions," *SAE International Journal of Engines*, vol. 11, no. 6, pp. 1307–1330, 2018, Publisher: JSTOR.

[39] B. Leo, R.A. Olshen, Charles J. Stone, and Jerome Friedman, *Classification and Regression Trees*, 1st Edition. New York: Chapman and Hall/CRC, 1984, 368 pp. [Online]. Available: https://doi.org/10.1201/9781315139470.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cambridge CB3 0FB, U.K.: Springer, Feb. 2006, 703 pp., ISBN: 0-387-31073-8. [Online]. Available: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/ (visited on 04/08/2025).

[41] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 1, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324 (visited on 04/08/2025).

[42] D. Fife and J. D'Onofrio, *Common, uncommon, and novel applications of random forest in psychological research*, Jun. 30, 2021. DOI: 10.31234/osf.io/ebsmr. [Online]. Available: https://osf.io/ebsmr_v1 (visited on 04/08/2025).

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 13, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754[cs]. [Online]. Available: http://arxiv.org/abs/1603.02754 (visited on 04/14/2025).

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 06/24/2025).

[45] P. Florek and A. Zagdaski, *Benchmarking state-of-the-art gradient boosting algorithms for classification*, May 26, 2023. DOI: 10.48550/arXiv.2305.17094. arXiv: 2305.17094[cs]. [Online]. Available: http://arxiv.org/abs/2305.17094 (visited on 06/24/2025).

[46] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, New York, NY, USA: Association for Computing Machinery, Jul. 25, 2019, pp. 2623–2631, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701. [Online]. Available: https://dl.acm.org/doi/10.1145/3292500.3330701 (visited on 06/11/2025).

[47] X. Yue, Y. Ni, K. Zhang, *et al.*, "MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI," in *Proceedings of CVPR*, 2024. [Online]. Available: https://mmmu-benchmark.github.io/#leaderboard.

[48] Community. "Hugging face leaderboard space," Hugging Face. (Feb. 27, 2025), [Online]. Available: https://huggingface.co/spaces?search=leaderboard.

[49] OpenVLM Leaderboard. "Open VLM leaderboard - a hugging face space by opencompass." (Feb. 18, 2025), [Online]. Available: https://huggingface.co/spaces/opencompass/open_vlm_leaderboard (visited on 02/26/2025).

[50] W.-L. Chiang, L. Zheng, Y. Sheng, *et al.* "Chatbot arena: An open platform for evaluating LLMs by human preference," Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots. _eprint: 2403.04132. (2024), [Online]. Available: https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard.

[51] Google Documentation. "Map tiles API style reference | google maps tile API | google for developers," Google for Developers. (2025), [Online]. Available: https://developers.google.com/maps/documentation/tile/style-reference (visited on 07/15/2025).

[52] G. Comanici, E. Bieber, M. Schaekermann, *et al.*, *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*, Jul. 11, 2025. DOI: 10.48550/arXiv.2507.06261. arXiv: 2507.06261[cs]. [Online]. Available: http://arxiv.org/abs/2507.06261 (visited on 07/14/2025).

[53] PDOK, *Kadastrale kaart PDOK*, 2025. [Online]. Available: https://www.pdok.nl/introductie/-/article/kadastrale-kaart (visited on 07/15/2025).

[54] Google AI. "Prompt design strategies | gemini API," Google AI for Developers. (Apr. 8, 2025), [Online]. Available: https://ai.google.dev/gemini-api/docs/prompting-strategies (visited on 06/03/2025).

[55] G. Boeing, *modeling and analyzing urban networks and amenities with OSMnx*. 2024. [Online]. Available: https://osmnx.readthedocs.io/en/stable/index.html.

[56] OSM. "OSM_overpass API," Overpass API. (2025), [Online]. Available: https://wiki.openstreetmap.org/wiki/Overpass_API.

A
Open Street Map tags

A.1. Collecting OSM data through an API

To build features for the entire grid of 1km^2 squares the python package OSMnx[55] is used, which allows for downloading geospatial features. The library can connect with different APIs of which the Overpass API is the official OSM read-only type and serves as an database over the web[56]. The Overpass Turbo application allows for testing Overpass Query-Language (OQL) for certain objects to get an understanding of what will be fetched. The OSMnx package simplified the retrieval of information by providing functions that transform a single line of code into the structured OQL and transform the return file in a GeoPandas 'geodataframe'. This is a format type that allows for analysis, transformation, and visualisation of the geometric objects queried.

A.2. Queried Tags

To query OpenStreetMap features using the `osmnx` library, one needs to provide a `dict` of '*key:value*' tags. The table below displays the '*key:value*' tags in an organised manner. The '*category*' column could be used to plot certain topics. The '*type*' column declares what the feature will be transformed to: area, count, or length. The '*min*' and '*max*' columns are used to filter out single objects based on their area or length. The '*expand-on-tag*' column is used for fine-tuning '*key:tags*' to '*key:tags<̃expand-on-tag*>', such as '*power:plantc̃oal.* The column '*weight-on-tag*' is used to multiply area or counts by the number of lanes, capacity or other relevant tags. Finally, the '*importance*' column notes the priority (lower) of that tag compared to others when an OSM object has multiple tags defined.

importance	**category**	**key**	**value**	**type**	**min**	**max**	**expand on tag**	**weight on tag**
2	industrial	industrial	refinery	area				
3	industrial	power	generator	area			generator:source	
4	industrial	power	plant	area			plant:source	
5	buildings	building	retail	area				
6	buildings	building	house	area				
7	buildings	building	apartments	area				
8	buildings	building	school	area				
9	buildings	building	greenhouse	area				
11	buildings	building	yes	area	100			
12	industrial	abutters	industrial	length				
13	industrial	man_made	silo	count				
14	industrial	man_made	chimney	count				
15	buildings	building	industrial	area	100			
16	buildings	building	commercial	area	100			
17	industrial	aeroway	aerodrome	area			aerodrome:type	
18	landuse	landuse	commercial	area				
19	landuse	landuse	construction	area				
20	landuse	landuse	residential	area				
21	landuse	landuse	retail	area				
22	landuse	landuse	institutional	area				
...	...	...	...	...	...	...	...	...

importance	category	key	value	type	min	max	expand on tag	weight on tag
...	...	...	...	...	...	...	...	...
23	landuse	landuse	allotment	area				
24	landuse	landuse	farmland	area				
25	landuse	landuse	farmyard	area				
26	landuse	landuse	forest	area				
27	landuse	landuse	meadow	area				
28	landuse	natural	scrub	area				
29	landuse	landuse	grass	area				
30	landuse	natural	wood	area				
31	landuse	natural	water	area				
32	landuse	natural	heath	area				
33	landuse	natural	wetland	area				
37	industrial	man_made	wastewater_plant	area				
39	industrial	landuse	industrial	area	50		cargo#product #industrial	
41	port	man_made	storage_tank	count				
42	port	man_made	pier	area	10			
43	port	man_made	crane	area				
44	port	amenity	ferry_terminal	count				
45	port	leisure	marina	area	100			
46	port	route	ferry	length				
47	port	attraction	boat_ride	count				
48	port	man_made	pipeline	length				
49	port	man_made	goods_conveyor	length				
50	port	industrial	port	area			cargo	
51	road_related	amenity	parking	area				capacity
52	road_related	amenity	parking_space	area				
53	road_related	amenity	fuel	area				
54	road_related	amenity	parking	count				
55	road_related	amenity	charging_station	count				capacity
56	road_related	highway	stop	count				
57	road_related	highway	traffic_signals	count				
58	road_related	highway	bus_stop	count				
59	road_related	highway	stop_position	count				
60	road_related	shop	car	count				
61	road_related	shop	car_repair	count				
62	road_related	highway	bus_stop	count				
63	roads	highway	motorway	length				lanes
64	roads	highway	trunk	length				lanes
65	roads	highway	primary	length				lanes
66	roads	highway	secondary	length				lanes
67	roads	highway	tertiary	length				lanes
68	roads	highway	residential	length				lanes
69	roads	highway	unclassified	length				lanes
70	roads	highway	motorway_link	length				lanes
81	roads	highway	trunk_link	length				lanes
82	roads	highway	primary_link	length				lanes
83	roads	highway	secondary_link	length				lanes
84	roads	highway	tertiary_link	length				lanes
85	roads	highway	service	length				lanes
86	roads	highway	road	length				lanes
87	roads	highway	busway	length				lanes
89	buildings	building	residential	area				
102	roads	railway	rail	length				
103	landuse	landuse	orchard	area				

A.3. Remapping of some Tags

The table below displays the grouping of tags from *remapped* to *target*, which removes duplicate feature names. The areas are summed to the target variable.

Target	**Remapped**
landuse:industrial~<...>_area	
meat	slaughterhouse
metal_construction	foundry
container	container_terminal
wellsite	wellsite;works
wastewater_plant	man_made:wastewater_plant, industrial:wastewater_plant
wood	woodworking, wood_processing, timber, sawmill
gas	natural_gas
oil	oil_depot
scrap_yard	auto_wrecker
dry_bulk	sand
recycling	waste_processing, junkyard
shipyard	boatyard
logistics	transport, distributor
medical_supply	pharmaceuticals
pumping_station	water_distribution, distributor_water
vehicle	passengers;vehicle
bakery	bakery_products
highway:unclassified_length	abutters:industrial_length

B

Optuna Hyperparameter Tuning Results

The tables below present the results of the hyperparameter-tuning by the Optuna framework of five the models trained on the different datasets. The best performing model is the model trained on the steps_0.1 dataset, which is a relatively smaller model, since it contains fewer trees, a ... The other models are relatively ... models with many

Parameter	**Range / Values**	**Notes**	steps_0.1
n_estimators	50 – 500	Number of trees	249
max_depth	3 – 20	Max tree depth	19
min_samples_split	2 – 20	Minimum samples to split	6
min_samples_leaf	1 – 20	Minimum samples per leaf	1
max_features	{sqrt, log2, None}	Feature subset strategy	None
max_leaf_nodes	200 – 100	Tree complexity control	791
min_impurity_decrease	0,0 – 1,0	Minimum impurity decrease	0,643594
bootstrap	{True, False}	Sampling with replacement	True
max_samples	0,5 – 1,0	If bootstrap=True	0,942533

Table B.1: Hyperparameter search space with results for RandomForestRegressor models labelled by dataset trained on.

Parameter	**Range/Values**	**Notes**	benchmark	direct_1.0	direct_0.1	steps_1.0
learning_rate	0,01 – 0,2	Step size, sample log scale	0,028869	0,030711	0,07775	0,053148
max_iter	50 – 500	Number of boosting trees	426	478	301	305
max_depth	6 – 15	Max depth single tree	10	11	12	8
max_leaf_nodes	200 – 1.000	Max leaves single tree	613	786	646	626
min_samples_leaf	1 – 50	Minimum samples per leaf	9	8	6	5
L2_regularization	0,0 – 1,0	L2 penalty	0,97405	0,155995	0,871578	0,197911
max_features	0,1 – 1,0	Fraction of features	0,591747	0,152275	0,443484	0,518431
max_bins	128 – 255	Histogram bins	150	238	225	187
early_stopping	True	Fixed (enabled)	True	True	True	True
validation_fraction	0,1 – 0,3	Validation split fraction	0,238082	0,220223	0,257581	0,185338
tol	10^{-7} – 10^{-4} (log scale)	Tolerance for stopping	0,000008	0,000013	0,000029	0,0
n_iter_no_change	10	Fixed	10	10	10	10

Table B.2: Hyperparameter search space with results for HistGradientBoostingRegressor models labelled by dataset trained on.

C

Model Fit Results

C.1. Enhanced LUR Models

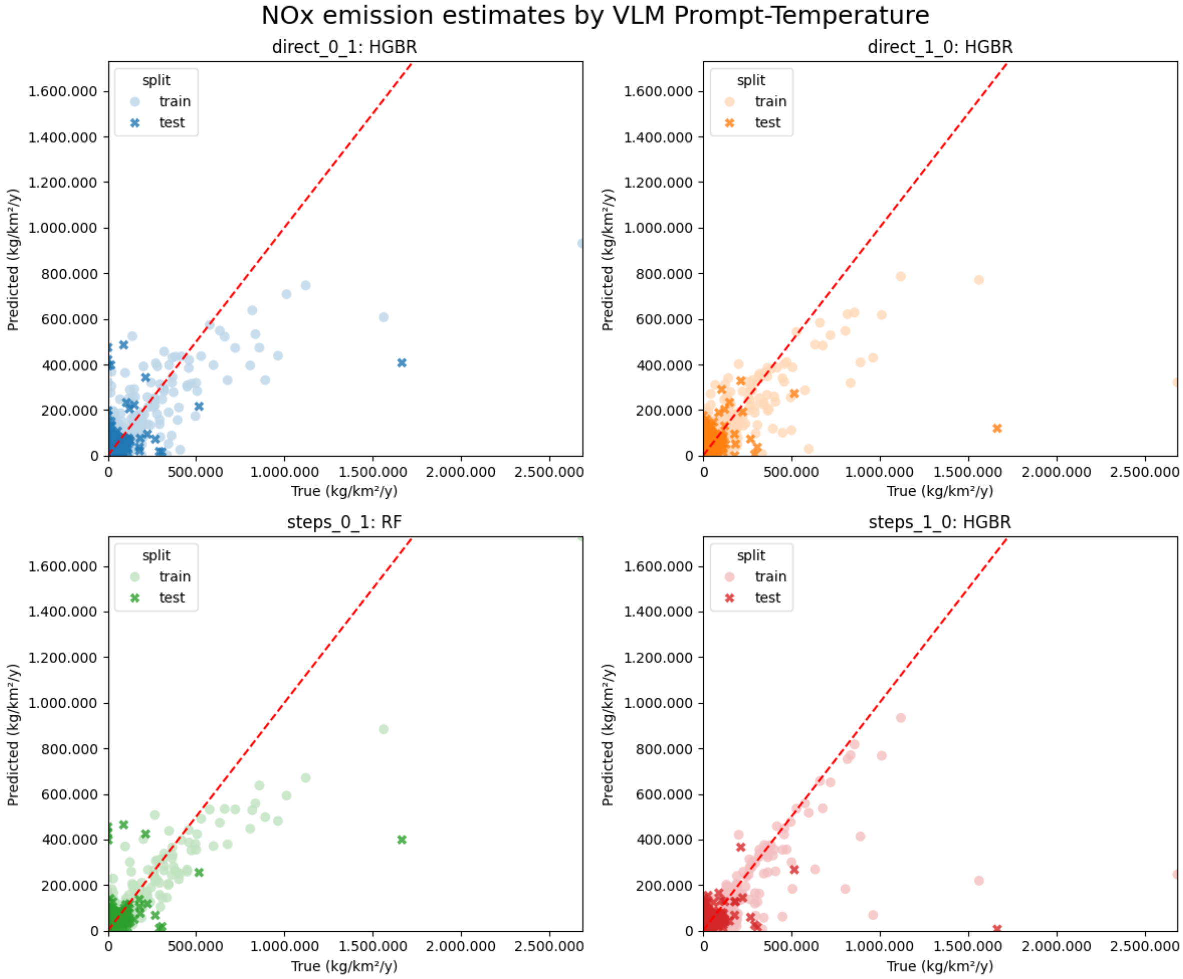


Figure C.1: Predicted vs. True NO_x plot of the best fitting model on the 'enhanced' datasets. Points on the red line would be a perfect prediction. The '**direct_1.0**' HGBR model performs best on $RMSE_{mean}$.

D

VLM output Bias

The following tables describe a OLS model fit on: `log10_nox_estimate = c(prompt_approach) * c(temperature)`. The goal is to find out how the prompt-temperature combinations differ in NO$_x$ emission estimates across the entire dataset.

Model fit statistics	
R^2	0,000
Adj, R^2	0,000
F-statistic	39,02
Prob(F)	$3,37 \times 10^{-25}$
Observations	323.977
Residual df	323.973
Log-likelihood	$-5,9513 \times 10^5$
AIC	$1,190 \times 10^6$
BIC	$1,190 \times 10^6$
Durbin–Watson	1.667

Term	**coef**	**std err**	**t**	**p>|t|**	**[0,025**	**0,975]**
Intercept	2,4713	0,005	463,026	0,000	2,461	2,482
c(prompt_name)[T,steps_nox_direct]	-0,0608	0,008	-8,060	0,000	-0,076	-0,046
c(temperature)[T,1,0]	0,0133	0,008	1,758	0,079	-0,002	0,028
c(prompt_name)[T,steps_nox_direct]:C(temperature)[T,1,0]	0,0485	0,011	4,545	0,000	0,028	0,069

Table D.1: OLS results for $\log_{10}$ NOx estimates with prompt × temperature interaction,

Dataset by:	Pred.	% Change	CI_low	CI_upper	CI Lower (%)	CI Upper (%)
direct_0,1	296,01	**baseline**	288,97	303,23	-2,38	2.44
direct_1,0	305,19	3,10	297,93	312,64	0,65	5,62
steps_0,1	257,32	-13,07	251,20	263,60	-15,14	-10,95
steps_1,0	296,66	0,22	289,60	303,89	-2,17	2,66

Table D.2: Back-transformed '*NOx_emission_estimate*' predictions in kg with percent changes relative to baseline. For the **direct** approach, there is a minimal effect of increasing the temperature value. For the **steps** approach, changing the temperature has large effects, where a lower temperature on average results in lower predictions.