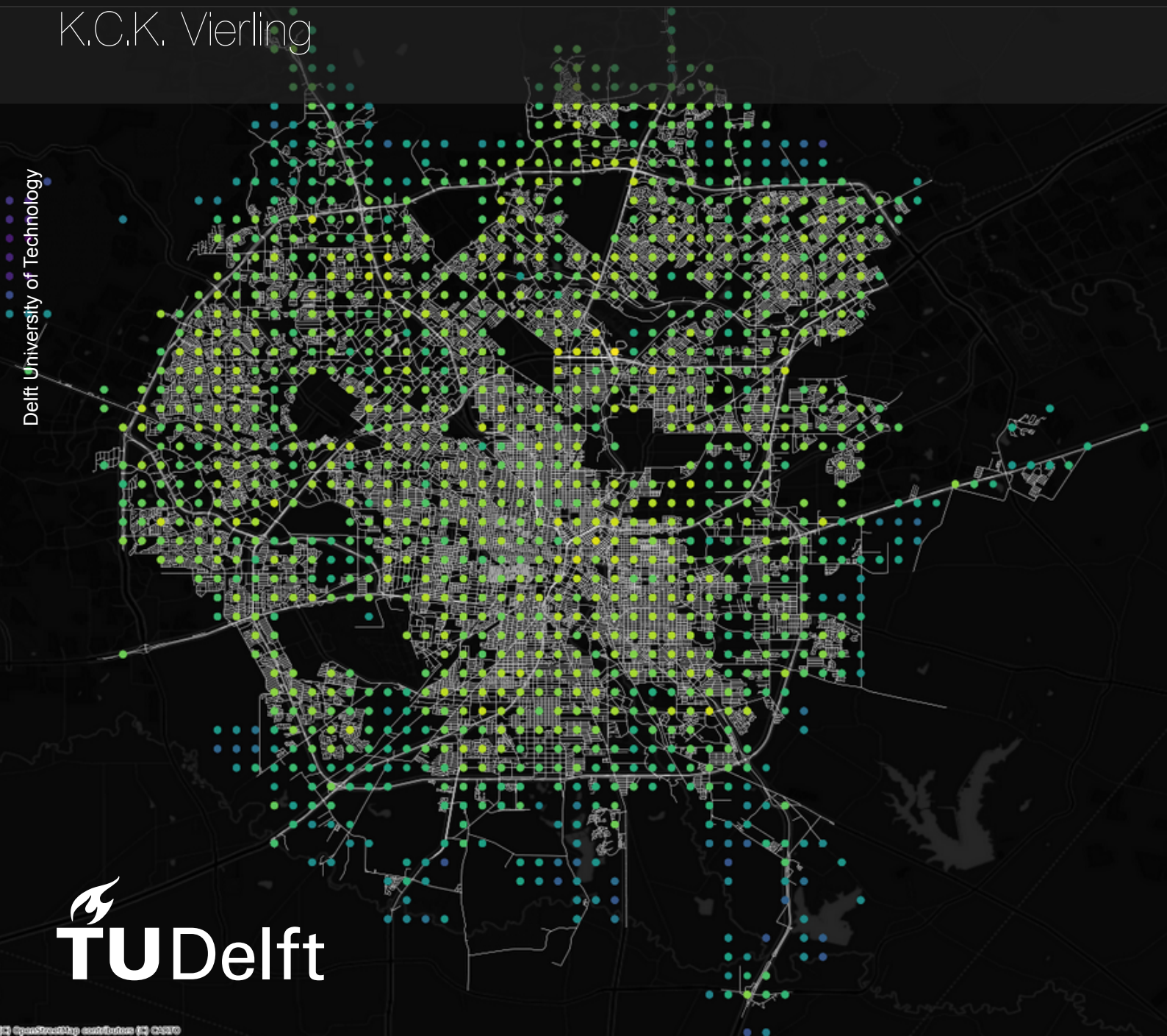
Quantifying and visualising anthropogenic emissions of CO2 in the urban environment

Exploring trends in the spatial distribution of emissions

K.C.K. Vierling



Delft University of Technology

TU Delft

Quantifying and visualising anthropogenic emissions of CO2 in the urban environment

Exploring trends in the spatial distribution of emissions

by

K.C.K. Vierling

Master thesis submitted to Delft University of Technology,
in partial fulfilment of the requirements for the degree of
MASTER OF SCIENCE
in *Engineering Policy Analysis*
Faculty of Technology, Policy and Management
to be defended publicly on Friday September 16, 2022 at 10:30 AM.

Student number:	4532635	
Project duration:	Aprill 11, 2022 – September 16, 2022	
Thesis committee:	Dr. ir. S. van Cranenburgh,	TU Delft, chair
	Dr. N. Goyal,	TU Delft, supervisor

Cover:	Model output for San Antonio, combined with a road network graph from OSMnx and background chart retrieved from CartoDB
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Executive summary

About 75 percent of CO_2 emissions caused by the combustion of fossil fuels stem from cities. With rapid global urbanisation, this fraction is likely to increase. Climate policy minimising the amount of emitted carbon can not be evaluated without accurate quantification of CO_2 emissions.
Not only the amount of emitted carbon is of interest, but also being able to place emissions in both the spatial as well as the temporal context is necessary. This gives policy makers and scientist the ability to compare the climate impact of neighbourhoods and cities.
For doing so, it is necessary to have a rasterised estimate with a high resolution. This is commonly achieved by using consumption statistics, combined with elaborate atmospheric and traffic models. Due to the resource intensity however, these highly accurate emission data products, named bottom-up, have very limited spatial coverage. Other estimates, relying on remotely sensed proxies, divide national consumption statistics over a country. These data products are a reliable first estimate, and available worldwide. But have been found to lack resolution to evaluate policy, also their accuracy around urban cores is known to be limited.
This is the main tension point, unless a more elaborate method is used, the spatial distribution of CO_2 emissions is unknown for many of the cities worldwide. The aim of this study is therefore to tackle the problem of gridded urban emission data products. This is done by exploring a new approach, which combines world wide remotely sensed covariates and bottom-up emission data products using a machine learning approach. This model can then be used to produce bottom-up estimates for cities that where previously not in the dataset.
For this the Hestia dataset is used, which has emission estimates at a high resolution for three cities in the United States. Through a literature review important spatial covariates where selected to train the model with. These covariates include data retrieved by satellite, such as night time lights. Or other data products, such as population statistics and road networks.

Analysis and findings

To solve the issue of accurately mapping emission in the urban context a Random Forest model is trained and analysed. After the literature review, in which the necessary input data is selected. The model is build, this involves the following key issues: algorithm selection, feature selection, hyper-parameter tuning method of cross validation and calculating scoring statistics.
Central in the study is the pipeline used to produce a reliable and stable approach for the estimation of emissions. This consists of a pipeline which is depicted in figure 1. As depicted the emission observation data is clustered spatially, so that the points that are either in the test or train set are geographically related. This pipeline gives a stable result and helps to evaluate the model about the generalisability. After model training and tuning, the model relevant features and errors are analysed more thoroughly.
Once a model has been developed a secondary aim is to show a use of the model. The model estimates of new cities are used to provide insight in the spatial distribution of emissions of eight US cities.

To assess the spatial generalisability of the model, the emission observations of the Hestia dataset are compared against model estimates. Performance metrics are used to quantify the quality of the estimate. An overview of these is given in table 1. As can be inferred, the model only works when log transforming the emission observations. The explained variance and R^2 is the highest in log transformed form at $\sim$0.47. The mean relative absolute error, which indicates by what percentage the model is off indicates that the model is off by about seven percent per grid cell. In non transformed form, the model performs comparatively worse with cells averaging 172 percent. The features and hyper-parameters selected through this method are found to be stable under a variety of model specifications. This provides evidence that the presented model performance generalisability are the that could be achieved given the current data.

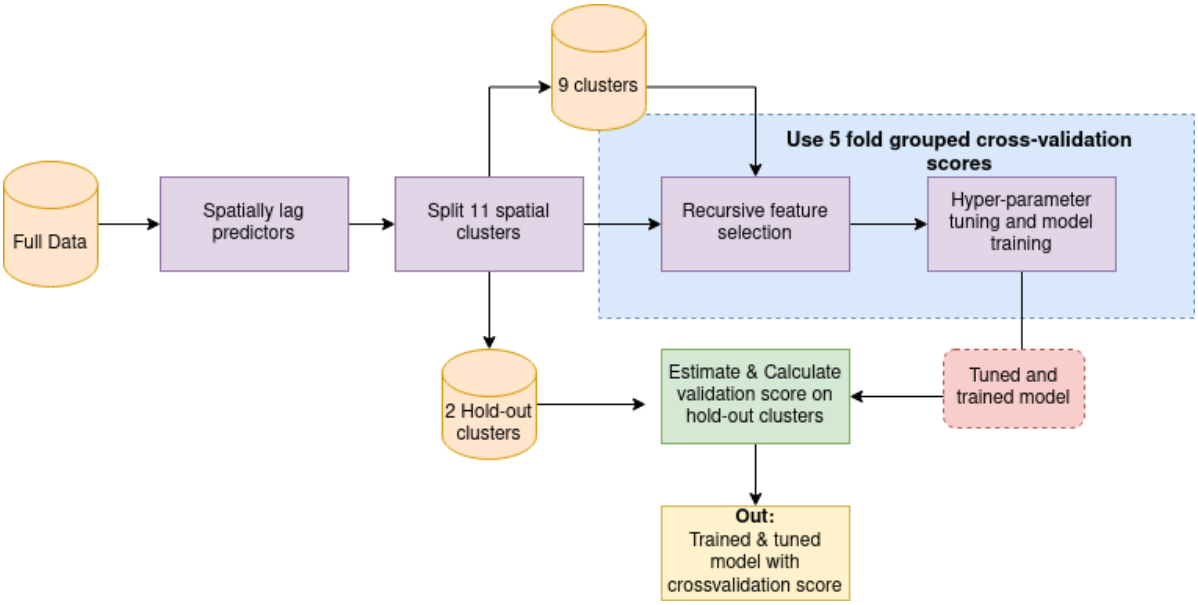


Figure 1: Model design steps followed in this study, which is done recursively to also take account of the effects which data observations ends up in the training and in the test set and how this affects the pipeline.

Cluster	MAE	MRAE	R^2	R	Kolmogorov stat	Explained Variance	Log Transformed
3	0.96	0.07	0.47	0.69	0.12	0.47	Yes
3	2.40E+06	1.72	0.01	0.16	0.12	0.03	No

Table 1: Model performance, both in log transformed form and without. The model performance deteriorates in non transformed form. While the performance metrics of a single cluster is presented here, these results are stable for different geographical regions.

In figure 2 a visualisation of the model estimate is compared with the observed results of a bottom-up estimate. While the distinction between high and low emission cells is lost, the general profile and distribution of emissions is roughly estimated. One can see that the model mainly fails to predict the maximum and minimum value cells.
The relevant covariates for the model are the following, in order of importance:: Nightly radiance, intersection density, population, GDP, PM25 and Pedestrian intersections. Inspection of the partial dependence shows that each of these variables, except PM25, have a positive increasing relationship with the models emission estimate. This relationship quickly converges for higher values. The highest relative errors mainly occur around the peripheries of cities, for cells with low emission values. The relationship between the covariates and the model estimate is not linear, but more complex.
The next step is to generalise the model. Thus, for eight of the most populous cities in the United States an inter quartile range coefficient is calculated using the emission distribution estimate produced by the model. Four analysis years have been analysed, in these years no trends regarding the inter quartile range coefficient are found. Findings do suggest that the coefficient diverges between cities. Cities with a lower coefficient tend to have a single cluster of high emissions in the centre. Whereas cities with a higher coefficient have multiple of such high emission clusters.

Recommendations & implications

The main limitation of this study is the limited spatial and temporal coverage of the input data. Using this model the insights from bottom-up emission data products can be generalised to other cities. More data regarding the spatial distribution of emissions within cities is necessary.
When evaluating the performance of spatial emission estimates it is important to employ grouped cross validation with the groups being geographically correlated points. The estimate performance needs to be evaluated per spatial group. Various algorithms for combining bottom-up data sources with re-

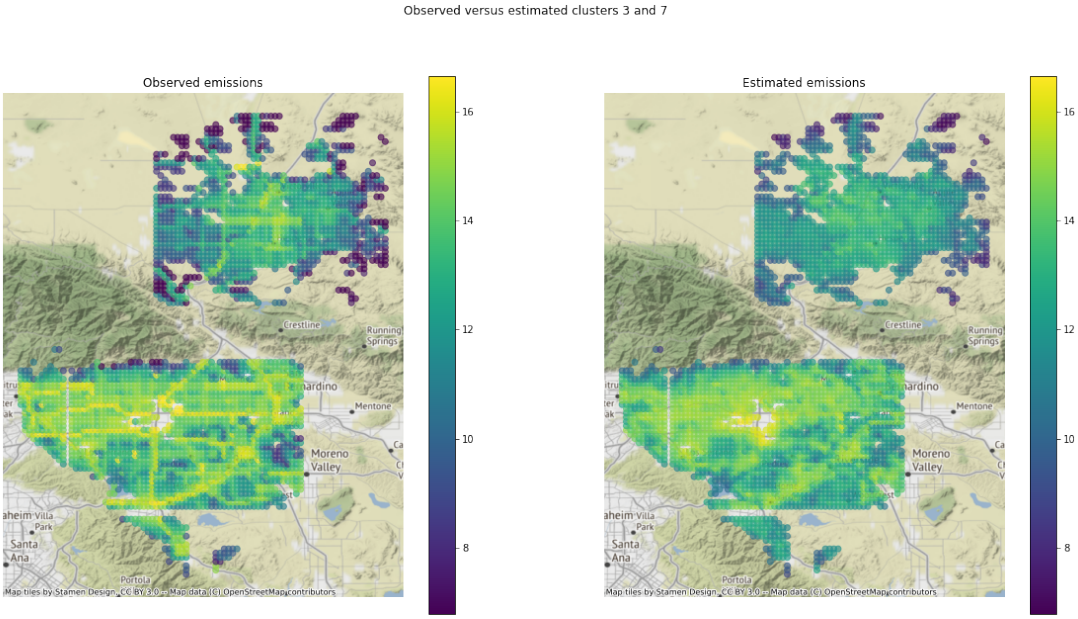


Figure 2: Plot depicting the observed emission values versus estimates made by the model. While for the top area the distribution seems to be lost, for the bottom the model is able to follow trends in emissions better

motely sensed estimates have been tested. These tests indicate that ensemble models outperform other machine learning approaches, such as linear and support vector regression. Highlighting that the relationship between the spatial covariates and emissions is not linear but more complex.
The presented model does not include point source data, while the importance of this is stressed in the literature. Creating an ensemble of a model estimating point source emissions with this could solve this issue. As under current model design, including point source features did not increase the reliability of the model estimate.
More-over the relevant covariates, pedestrian and car road networks where found to be important, the model could be extended by including traffic and congestion data, as the presence of the road does not necessarily correlate with the use of it. The issue of estimating emissions along the peripheries of cities remains, as cells with lower observed emissions typically have a higher relative error. This means that in for further improvements of such models a focus on estimating the emissions around the peripheries of cities should boost model performance
While this study is not aimed to be an improvement of the top-down methods, it would be interesting to see a comparison study, with the other prominent emission data products. This creates insight in whether the issue of accurately distributing urban emissions has been pushed in the right direction.

This study lays forward a new scalable approach to estimate emissions and the spatial distribution in urban areas. Currently the study data is limited, but as more data becomes available a Random Forest regression approach could prove to produce accurate insights. Also the findings in model design and evaluation could be used in future studies. For city planners the results can be used to gain a first insight in the spatial distribution of emissions within their city. Pinpointing high emission zones or comparing different urban areas. Knowledge of the spatial distribution of cities can be a pathway of understanding how to develop less carbon intensive cities.

Contents

1. Introduction

Anthropogenic emissions of CO_2 are the prime drivers of climate change. The ever rising emissions of CO_2 caused by fossil fuel combustion have detrimental effects on our living environment. According to IPCC (2014), nearly three quarters of emissions stem from cities. This means that decisions at the city level could have considerable impact on climate change mitigation.

A key point for climate mitigation policies is limiting and reducing the amount of emitted carbon. For this it is necessary to quantify the amount of carbon emitted, as this gives insights in trends. And policy makers the ability to quantitatively assess the effectiveness of interventions. The need for quantifying the amount of carbon emissions is illustrated by initiatives such as thenet-zero cities, that aim to be carbon neutral in 2030 partially by offsetting carbon emissions (Seto et al., 2021), or for example the covenant of mayors where city decision makers have pledged to create emission inventories, among other policy changes (Kona et al., 2021; Nangini et al., 2019).
Not only the quantification, but also being able to place emissions in both a spatial and temporal context, lends insight to the amount of emissions, as well as how they are spatially distributed. For this, rasterised gridded emission maps are necessary. Which gives policy makers the ability to not only compare cities, but neighbourhoods, zones or even the effects of certain road designs (Zhou and Gurney, 2010). This requires increasingly accurate and finer CO_2 emission quantification data products.

Creating such rasterised high resolution mappings commonly require rigorous approaches that combine ground measurements and consumption data with statistical and atmospheric models (Chen et al., 2020; Gurney et al., 2019a, 2009). Emission data products retrieved like this are named bottom-up and are considered a reliable estimate. Doing so however requires sufficient statistical infrastructure, therefore the amount of cities covered by such a mapping is very limited.
A solution to the limited coverage is to use worldwide available remotely sensed data, retrieved through satellites, as a proxy for emissions. And coming up with a distribution of emissions like that. Studies leveraging this (see for example Oda et al. (2018) or Asefi-Najafabady et al. (2014)) thus rely on a spatial co-variate to distribute emissions spatially. These methods rely on downscaling emissions, entailing spatially distributing emissions of a known sum total. The aim of these models is to distribute emissions at the national or even global scale. As a consequence, models leveraging remotely sensed data are known to have higher uncertainty in urban cores (Chen et al., 2020; Gurney et al., 2019a).
The objective of this study is to change this trade-off between accuracy and coverage. This is done by presenting a new scalable ”hybrid” approach using a Random Forest model. The idea is to train a statistical learning algorithm using bottom-up emissions estimates and remotely sensed covariates. This extends the coverage off the bottom-up CO_2 emission mappings, and shifts the accuracy of the remotely sensed estimates. In addition to introducing the hybrid approach, the new model results will be used to calculate emission distributions.
The remainder of this chapter will first cover a broad overview of this topic. This then flows into the research problem and gap. Followed by the introduction of the research aims and questions and the study contribution. Then the study limitations and a further structural outline for this thesis is discussed.

Existing models based on remotely sensed data, often named top-down methods in the literature, create an estimation that is approximately correct. And generally cover a great area producing insights with a relatively low amount of data and assumptions. While these top down methods are found to be a reliable first guess, higher resolution and accuracy is needed for the analysis of climate mitigation policies (Chen et al., 2020; Oda et al., 2018). The majority of the top-down methods rely on population density and nightly radiance as proxy for emissions (Asefi-Najafabady et al., 2014; Oda et al., 2018; Ou et al., 2015). An important caveat of this is that it is known to create a bias when estimating emissions within cities. A limitation of nightly radiance as a spatial proxy is that the sensors measuring the light are easily over-saturated in highly illuminated areas, such as urban areas (Shi et al., 2014). Consequentially, data products using night time lights tend to overestimate emissions in urban cores, and

misallocate them in less densely populated areas, missing for example big point sources, for example electricity generation (Zhou and Gurney, 2010).

Using a bottom-up approach, great accuracy could be achieved by combining fossil fuel consumption data with intricate transport or atmospheric models. As explained these models could for example use ground-truth observations, with gathered data obtained by flight campaigns or flux tower samples (Chen et al., 2020; Gurney et al., 2012, 2019a; Yadav et al., 2022). An example of a model that maps emissions down to a street and building level is Hestia. To achieve high accuracy extensive data is required. The coverage is thus limited to four US cities (Gurney et al., 2012). So, while highly accurate, this method requires considerable statistical infrastructure and data quality which is not feasible everywhere. For those places top-down using remote sensing data would be of value, which is able to produce global results in one blow, using strikingly less resources.

This summarises the research gap, unless a more elaborate method is used, the spatial distribution of emissions in the urban context is not accurate if only relying on downscaling emissions through a top-down method. Contrastingly, cities with bottom-up data products have better insight (Gurney et al., 2019a). As explained, the inaccuracies resulting from using solely night time lights as a spatial proxy result in uncertainty regarding the spatial emissions and distribution thereof in cities. This is troublesome as insight in the spatial distribution of emissions would be off great value for climate mitigation policies, for example lending insight in how the distribution affects the total emission sum. For this however, more accuracy and more coverage specifically in urban cores is necessary, as currently remotely sensed estimates do suffice as a first estimate but are not accurate enough for more in-depth analysis (Chen et al., 2020). Therefore the aim of this thesis is to to lay forward a new approach to estimate the distribution of urban emissions, using remotely sensed covariates, based on the more accurate bottom-up approaches laid forward in the studies of Zhou and Gurney (2010); Gurney et al. (2012). Doing so increases confidence in the estimate, substantiating a new ”hybrid” method for estimating emissions in the urban area. This model can then be used for the secondary aim, which to use the new estimate and to create new insights regarding the spatial distribution of emissions in a variety of cities.
For this implementation the earlier introduced Hestia dataset will be used. As this data product covers four cities in the US and the emissions down to the building level are likely other cities in the US. Due to time constraints for now only eight of the most populous cities will be considered, this as more people likely imply more emissions hence picking the cities likely the most impact.

This leads to the following main research question: *Using a new hybrid approach, how are emissions spatially distributed within eight of the most populous cities in the US?*
To answer this, and to improve where other emission data products fall short, a new hybrid model is developed, based on a Random Forest, combining properties of both the bottom-up and top-down methods. Once trained, the new model is used to produce estimates, which are assessed and generalised . This provides insight in the performance of the method.

The aim of the main question is two fold. The first main problem is to understand how this new hybrid method performs, and gaining confidence in the model output. Then, a pilot, or proof of concept is presented, extending coverage to new cities, using this new model.
For understanding the new model performance, firstly, the choice of covariates and bottom-up data product to base the model on is important. Then the aim is to know how this model generalises and how this is affected by design choices made during implementation. And to understand how to distribute emissions more accurately. Knowing the generalisability and robustness of these results, the aim shifts towards understanding where the model is lacking. This is mainly aimed at understanding the caveats and understanding where the model could be improved, by inspecting where the model makes the greatest relative mistakes.

After confidence in the results of this first method is build, the model will be used to estimate the emission distribution in eight of the most populous cities. This is a pilot showing how this method produces new results for unseen examples. Form this, it is possible to quantify the spatial distribution of emissions, and compare the cities with one another, as well as say something about the temporal trends.

Having covered the aims of this research, the following sub-questions will explain this aim.

- **SQ A:***Which spatial covariates and bottom-up data products to use for the new hybrid model?*
- **SQ B:***Using the hybrid approach, how well does the model spatially generalise on a test set, how is this affected by various model designs?*
- **SQ C:***Where does the model have the five percent highest residuals error and what are the properties of these cells?*
- **SQ D:***What are the trends in the spatial distribution of emissions in cities?*

The contribution of this study is hence to lay forward this new hybrid approach and assessing whether this new hybrid model is a worth-wile pursuit. This study will give insight in whether this hybrid approach further extends the bottom-up approach. The covariates effecting the spatial distribution of emissions is relevant gives insight in which covariates are use full to pursue in later studies. In short, this research is aimed to solve the problem of the accuracy of emission distributions within the urban environment.
While not a direct aim of thisof this study, these higher accuracy emission maps can then be used to assess how this would affect the total emission sum. Earlier studies have suggested that the urban form, zoning and compact development could lead to decreased fossil fuel use (Creutzig et al., 2016). As discussed earlier, the current emission data products only have moderate accuracy in the urban environment (Hutchins et al., 2017; Gurney et al., 2019a). Having access to this information could lead to a better understanding of the causes of urban emissions, and how effective urban planning could influence them. Doing so ties into the societal grand challenge of decreasing carbon intensity in the urban landscape.

Given time and computational resources it is not possible to compare the new method against other data products. The contribution of this study hence mainly lays in creating this new hybrid method and assessing the feasibility. In the current research design the Hestia data set is used and assessed on the spatial and temporal generalisability. Other bottom-up estimates, such as for example Vulcan could also be used. This however is beyond the current scope, only one bottom-up data product will be used for this assessment. In a similar vain, the scope for the included covariates is limited as they need to be openly available and have a world-wide coverage. If having more data and time resources, it would be possible to use more spatial covariates.
Model estimates intend to place the emissions where they occur. Therefore it is possible that the emission profile of certain cities does not reflect the real amount of emissions, as only emissions within the administrative boundary of a city are covered. For example power plants beyond the administrative boundaries are therefore not included in the model estimate.
Usage of Hestia also limits both the spatial and temporal scope. The model will only be used for the years Hestia has coverage for. Spatially the model is not intended to be generalised beyond the USA, until more data could provide insight in the model performance there. This limitation is due to the fact that the properties of energy use and emissions for other countries are uncertain.

For the remaining chapters first a literature review is performed, to gain insight in which spatial covariates have a connection in the literature with CO_2 emissions. Then the methodology section will first describe the Hestia data set, then the study area used for training the model. This then flows into co-variate pre-processing, model selection and performance assessment.
The result section then covers in how the new model generalises, where uncertainty is located and how the model estimates are produced. This section concludes with analysis of the spatial distribution of emissions in the eight most populous cities in the US. The study is wrapped up with a discussion section and conclusion giving light to the importance of the findings.

2. Methodology

The following sections will discuss the approach used to achieve the research aims. The main research question is: *Using a new hybrid approach, how are emissions spatially distributed within eight of the most populous cities in the US?* The main problem to solve is the issue of estimating urban emissions and the distributions thereof. For this a new method, the ”hybrid” method is used. This method relies on remotely sensed covariates, a bottom up estimate and an algorithm to combine them.
To guide the reader through the methodology section an issue tree is presented in figure 2.1. This diagram is intended to give an overview of the design choices in the model. The first main split, is the data choice. Each of the successive problems is independent. First this section will thus cover the data aspect of the hybrid model, the selection, preprocessing and exploratory data analysis.
The method en robustness branch presents issues and design choices, in a sense they are independent, for example, hyper-parameter tuning will always have to happen regardless of which algorithm is used. Following this structure and solving issues helps bring forward new knowledge on how to solve the issue of estimating emissions in cities. The same format and order will be used in the results section.
Having done so, will give light to the secondary aim, how are emissions distributed in eight of the most populous cities, according to this newly developed approach.

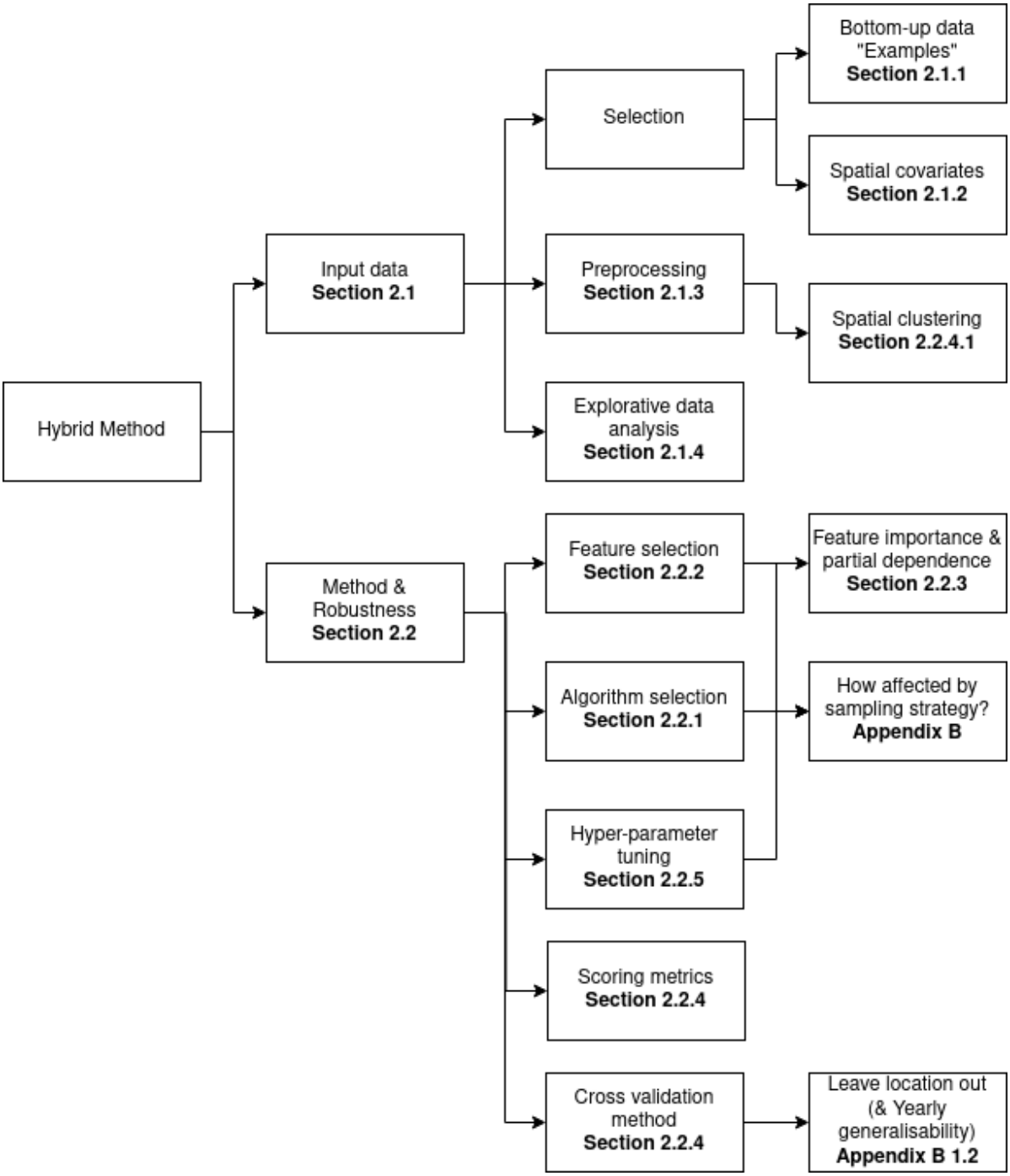


Figure 2.1: Issue tree, with key issues to solve as well as the sections in which these choices are covered

2.1. Data

As explained, the hybrid approach relies on co-variate and a bottom-up emission inventory. Therefore the first part of this research will discuss and argue for the reason of including certain covariates. This is achieved using a literature review.

2.1.1. Bottom-up emission inventory

To understand why the Hestia data set is used for creating a hybrid method, first a short background on bottom-up emission inventories is given. This section then provides an argument on why Hestia is used and is central in this study.

Overview of bottom-up emission inventories
A wide body of literature and approaches for reaching (national) emission estimates exist. Generally, these inventories rely on consumption statistics in combination with emission factors to calculate the amount of CO_2 emitted for various activities and categories. This is known as a bottom-up method.
As synthesised by Asefi-Najafabady et al. (2014), prominent and reliable whole nation bottom-up fossil fuel CO_2 emission estimates are formulated by about five organisations. The first to be discussed; the IEA (International Energy Agency), which tracks emissions per economic sector, using surveys based on internationally agreed IPPC guidelines (IEA, 2021; Eggleston et al., 2006). Likewise, British Petroleum (BP) also produces emission inventories, using consumption statistics following the IPCC (British Petroleum, 2021). The Energy Information Administration (EIA) of the US, produces emissions estimates per fuel type, using fossil fuel consumption statistics. Similarly, CDIAC also produces emissions by fuel type, based on guidelines constituted by the UN (Marland and Rotty, 1984) and spatially distributes emissions using population density obtained by satellite imagery(Hutchins et al., 2017). This makes that CDIAC is more of a hybrid method, as it combines properties of both bottom-up and downscaling. While CDIAC was discontinued since 2012, it forms the basis of the top down method presented in Oda et al. (2018), using extrapolation for unavailable years.

While these national inventories are independently produced, they often share aspects of the same data, such as the survey used (Macknick, 2011).

A bottom-up data product considered as the most accurate in the US is Vulcan (Hutchins et al., 2017). Which uses geolocated bottom-up calculations by the Environmental Protection Agency (EPA). Non-point sources are down-scaled by administrative regions, and mobile emissions using a road model (Gurney et al., 2009). Vulcan has a high resolution at $1\times$1km and has data of the years 2010-2015 (Gurney et al., 2020), this method however is considered less accurate around urban centres (Gurney et al., 2019b).
The Hestia dataset is an extention of the Vulcan dataset, that takes geolocated consumption data down to the street and building level for one specific county, and is aimed to remain linkage to the Vulcan project (Zhou and Gurney, 2010). The Hestia method uses Vulcan data as input, and further enhances said data with location specific information such as building data and surveys (Gurney et al., 2019b).

Summarising, at the national scale, plenty of bottom-up emission estimates are available. At the urban scale however, this information is more sparse. For the United States, Hestia and Vulcan are considered the most accuracte (Hutchins et al., 2017). For emission quantification spatially and temporally mapping uban areas using a bottom-up approach only the Hestia method is available and to the authors knowledge uncontested (Gurney et al., 2012; Hutchins et al., 2017; Chen et al., 2020). The high accuracy and sparse availability is why Hestia is chosen as the bottom-up inventory behind the newly developed method.

Hestia background
So, the focus of this study is the Hestia dataset (Zhou and Gurney, 2010; Gurney et al., 2012). The reason for choosing the Hestia dataset is that high resolution gridded emission maps with reliable estimates of fossil fuel CO_2 emissions of urban centres are sparse. The next best product with gloabal coverage is ODIAC with the same resolution. At a coarser resolution of 1°$times$1°($\sim$ 11 km at the equator) FFDAS and GRACE are available. These resolutions are quite coarse for the urban environment, comparing neighbourhoods would not be possible. Therefore the decision falls on using Hestia to use and to fit a model to. This inherently means that in this study the Hestia results will be regarded as ground truth.

The Hestia dataset covers four areas in the US, Baltimore, Indianapolis, the Los Angeles Basin and Salt Lake City. All areas, except Salt Lake City have coverage for the years 2010 until 2015. As Salt Lake city only includes the years 2002, 2010 and 2011. Because of this temporal mismatch, for which the spatial covariates have no coverage, the Salt Lake City area is omitted from this analysis.
The Hestia dataset is a fossil fuel emissions inventory dataset with the intention to map emissions where they occur, at a fine spatial resolution of maximally $sim1 \times 1$km. In terms of emission scope, this means that the intention of this method is to map scope-1 fossil fuel emissions. Scope-1 emissions

are all emissions that fall within a certain administrative area Seto et al. (2021).
The Hestia method combines consumption data of on-road emissions, industrial and point source data (for example from power plants, airports etc.) combined with ground level observations. This results in a fossil fuel emission inventory with a fine temporal resolution, giving insight in the spatial distribution of emissions within a city (Gurney et al., 2019b).

When comparing the Hestia data set with existing remotely sensed estimates the median difference per grid cell range from 47 to 84% (Gurney et al., 2019a) and whole city relative differences between -1.5 up to 20.8%. The authors attribute this difference between Hestia and the remotely sensed emission inventories to misallocation of point-sources and relatively high allocation in urban centres. This is inline with findings of Chen et al. (2020), that compared various remotely sensed inventories.
This means that it is likely that the Hestia data set approaches the actual distribution of emissions better then existing remotely sensed inventories for the areas Hestia has coverage for.

2.1.2. Spatial covariates

To be considered as a reliable proxy, the spatial covariates to be used for emission estimation need have a connection in the literature with emissions. Therefore first a literature review will be conducted to find the relationship between emissions and the spatial covariates.
In addition to the connection in the literature, the proxy datasets have to meet certain requirements to be included in the model. As the study is ought to be reproducible, all used data is publicly accessible and available in a gridded format. The spatial resolution of the covariates is minimally 1×1 km, preferably smaller as otherwise these proxies will not contain enough information to base the algorithm on. In terms of the temporal resolution, Hestia covers the years of 2010 until 2015, therefore the proxies have to cover these years as well.

For integration and access to publicly available satellite data Google's Earth Engine will be used. Which has an API for accessing various remote sensing data, as well as the ability to import rasterised files.

Night time lights
A much reported proxy introduced by Oda and Maksyutov (2011) is using night time lights. The assumption behind this proxy is the following; anthropogenic fossil fuel emissions occur where there are people. Hence, these places will be illuminated at night. An example of a global data-product leveraging this proxy is the ODIAC data product. However, often night time radiance is combined with other proxies. For example in the the Fossil Fuel Data Assimilation System (also known as FFDAS) which also uses night time lights in combination with gridded population density maps (Asefi-Najafabady et al., 2014; Rayner et al., 2010). Methods for predicting emissions using night time lights lead to reliable results when compared to bottom-up approaches (Chen et al., 2020; Gurney et al., 2019a), especially when combining them with other proxies (Ou et al., 2015; Geng et al., 2017). There are multiple versions of night time lights available used in various social studies. Introduced in 1973 Croft (1978) the DMSP Operational Line Scan (OLS), is currently used for in most data products to date. However a new data-product exist: VIIRS-DNB, which was introduced around 2012. This data has a higher resolution, greater dynamic range and solves many issues of the original night time lights data product (Elvidge et al., 2013).
Concluding, night time lights are a reliable proxy used in most data-products relying on remotely sensed data (Ou et al., 2015; Chen et al., 2020; Gurney et al., 2019a). However a limitation is that these methods rely on older sensors used for night time lights. In this study however the newer VIIRS-DNB data product will be used, leveraging the better dynamic range to have more clear differences in urban cores.

Population density
A closely related proxy to night time lights is population density. This spatial statistic is important in identifying urban cores, or industrial zones. An notable emission inventory relying on this proxy is CDIAC. Which uses national consumption statistics and spatially down-scales emissions to finer resolution using population density (Hutchins et al., 2017). EDGAR (Emissions Database for Global Atmospheric Emissions) also relies on population density (Janssens-Maenhout et al., 2017). And the data products of Ou et al. (2015) & FFDAS (Asefi-Najafabady et al., 2014) and combine this proxy with night time lights.

While the idea that densely populated areas have more emissions intuitively seems correct, high emitting industrial zones are often not inhabited by people. For example Ou et al. (2015) finds that population density is the least correlated with emissions, compared to night time lights and distance to roads. Data products hinging on night time lights and population density often fail to estimate the magnitude of industrial or electricity production point sources(Gurney et al., 2019a; Chen et al., 2020). This is inline with findings of a comparison study Hutchins et al. (2017). In which the authors conclude that EDGAR and FFDAS fail to discriminate predicting low emission levels. This likely means that population density is not the best proxy for emissions. However, this proxy still is a statistical significant predictor in many studies, especially in combination with other proxies (Ou et al., 2015; Asefi-Najafabady et al., 2014; Geng et al., 2017).

Air quality maps

Fine particulate matter, also known as $PM_{2.5}$, could also serve as a proxy to carbon emissions. Like CO_2, these particles are a by-product of combustion. Therefore, global air quality maps therefore could serve as a proxy for fossil fuel CO_2 emissions, to the authors knowledge there is no CO_2 emission data-product leveraging this co-variate. In this study a global gridded dataset constituted by van Donkelaar et al. (2021) will be used. These datasets are the result of geophysical hybrid models that combine chemical transport, satellite data and ground calibration approaches to produce global high resolution air quality maps.

While the same authors also van Donkelaar et al. (2021) produce North American estimates using a regional model, the global model is used in this study. While the North American regional model likely is more accurate, data with global coverage is preferred for plausible geographical generalisability beyond the USA.

In a recent study, conducted by Anenberg et al. (2019), found that there is no significant correlation between CO_2 emission rates and $PM_{2.5}$ within cities. This is attributed to the fact that most developing nations tend to have a tendency to aim policy at $PM_{2.5}$ as these byproducts are mitigated relatively easy compared to carbon emissions (e.g. using air filters). Furthermore, they argue that in developed countries the more polluting industries are moved outside of cities, or even to developing nations with less strict environmental protection laws.

Byproducts of combustion, such as NO_2, are also often used a spatial proxy for CO_2, as these are easily observable using sensors onboard satellites, see (Dou et al., 2022; Berezin et al., 2013). The EDGAR CO_2 emission map uses the same methodology as with CO_2 as with their global $PM_{2.5}$ and NO_2 data-products (Janssens-Maenhout et al., 2019). A potential proxy could be the Sentinel 5P TROPOMI NO_2 satellite data. However the coarse spatial resolution (3.5km $\times$ 7 km) (Sneep, 2021) makes this dataset to coarse for the current task. And therefore omitted for now.

Regional GDP

The Regional gross domestic product could also have influence the amount of emissions. A known relationship is that a higher GDP in an area means more energy consumption and hence more emissions (Chen et al., 2019). This proxy also indicates industrial areas, especially in combination with other predictors.

GDP and socio-economic status has shown to be an important driver for carbon footprints (Nielsen et al., 2021). The difficulty with this predictor is to know whether urban centre GDP reflects household income. When comparing cities world-wide the trend in general is that a higher income results also in higher emissions. Some studies however point out that for very high GDP this relationship decouples (Haberl et al., 2020). Nangini et al. (2019) finds something similar, cities in the highest quantiles of GDP tend to have less emissions within their administrative boundary, the author however argues that these cities likely have exported significant part of their emissions. This could be the cause for the low correlation found between emissions and GDP in Hsu et al. (2022).

Global Power plant database

As discussed before, using only population density and nightly radiance tends to miss allocate emissions around point source polluters. For example caused during electricity generation (Gurney et al., 2019a; Zhou and Gurney, 2010). Knowing the location and installed capacity of power plants is therefore also an important spatial co-variate for estimating emissions.

The World Resources Institute keeps a database of the location of power plants, as well as the installed and estimated capacity, commissioning year and fuel type (Byers et al., 2021). Using this data could thus provide additional insight in the emissions of point sources and provide additional information for the model on emissions.

Road networks
One of the much reported limitations of remotely sensed data products compared to bottom-up estimates are on-road emissions (Gurney et al., 2019a). Solely using night time radiance does not cover important properties that explain the heterogeneity of emissions on roads. Bottom-up approaches sometimes use road networks for making an estimate (Gurney et al., 2020), or even use congestion data using user data from navigation systems (Nangini et al., 2019).
Not only the presence and amount of roads is important. But additionally a growing body of literature suggests that the urban fabric, which could be modelled by using intersection and building density influences transport emissions (Newman and Kenworthy, 2015; Ewing and Cervero, 2010).
In particular intersection density is used as an indicator for pedestrian friendly street design. Recent case studies for example have found that pedestrian friendliness and having a well connected street network are associated with using less carbon intensity travel modes, such as cycling or using public transport (Christiansen et al., 2016; Liu et al., 2017). This association means that these predictors contain information regarding preferred mode of transport and urban emissions. This predictor works in two ways, the presence of roads for cars has a positive relationship between on road emissions. However intersection density or the presence of pedestrian roads influences the walk ability of a city giving inhabitants access to less carbon intensive modes of transport.

Spatially lagged covariates
In the current context the above mentioned covariates will be transformed to a rasterised co-variate giving the possibility to relate emissions with the proxy to estimate emissions.
However for some co-variates, the spatially lagged co-variate could also contain information for the emission model.
In this context spatially lagging is defined as the average value of the cells around it. For example for nightly radiance knowing that a certain raster cell has higher radiance than the cells around it contains information about the spatial properties of that cell, perhaps the cell is at the edge of a city, or in a secluded industrial area. This does not only work for nightly radiance, but also for for pedestrian and motor roads, and intersections. As especially these predictors have a lagging property, as it is likely that the road is not only present in a single cell but multiple cells around it.
In similar vain, GDP, PM25 and population will also be lagged. As having the lagged value also contains information about the spatial distribution of the co-variate and therefore inherently with emissions.

2.1.3. Data preprocessing

The spatial covariates come in two shapes, remotely sensed data in a rasterised format and some data in the form of a point or polygon. The rasterised data format has a continuous value per cell corresponding to the value on the ground in the specified resolution. The geometries indicate the presence and shape of an entity, for example a power plant or a road. Most co variate data is in the rasterised format, the preprocessing steps for that data is found in section 2.1.3. With the final goal to estimate emissions on a 1×1km raster grid it is necessary to convert geometries to a raster format, the method for doing this is discussed in the section 2.1.3.
After this the processing for each co-variate is discussed.

Raster data preprocessing
All raster data, both covariates and the observed emission values are re-projected to be in the WGS84 projection.
To create the covariate data, it is necessary to link each observed emission value with the geopsatial covariates. This creates a record of entries containing emission data and the predictor data. Each observation, covariate or target variable, is discretisised as a point in the centre of the grid-cell. This means that it is assumed that for each square kilometer the emissions and co-variate data is the uniformly the same across that cell and. Doing so creates a grid with points spaced 1 kilo meters apart. Any point with any missing data both in target or predictor variable is omitted from the analysis.

Geometry data preprocessing
Most predictor data are available in a rasterised format. However, the power plant and road data are either points (for example the location of a power plant) or a polygon (the roads).
The power plant data of the WRI dataset has point data with latitude and longitude values. To couple this data with a grid cell from the raster (as obtained described in 2.1.3) first the geographical projection of both the rasterised data and the power plant data is transformed to a ”flat” projection in metres. Then each point is matched to the nearest point with available in the raster format, with a maximum distance of 1 km. It is possible for a raster point to have multiple power plant points associated with it. This property is asymmetric, it is thus not possible for a power plant point to be associated with multiple points from the raster. Once matched with the closest raster point the properties of the power plant, such as estimated capacity are then added to that cell.

For the road data, retrieved from the OSMnx python package (Boeing, 2017). There are two types of geometries to convert to a rasterised format. First are the polygons of the roads. These are converted to points using the centroid of each polygon. The intersections are in a point format with a longitude and latitude value, like the power plant data. The obtained points from intersections and road centroids are then matched to observed emissions points like the power plant data, again with a maximum distance of 1 km. This creates a count of intersections and road centroids per grid-cell. The variable is interpreted as follows, if the value is for example 25 this means that 25 road centroids fall within this raster cell.
Observations of the Hestia data set without roads or power plants associated with it get a value of zero, meaning no power plant or roads present in that raster cell.

Hestia data
The relevant Hestia data for the three study areas is available in a gridded format. Each area is converted to the WGS84 projection and if a smaller resolution then 1×1 km is available re-sampled to be in that resolution.
To prepare the data to be fed to the algorithm any raster cells with a zero emission value are omitted from the analysis, as these are likely caused by either a sampling error or because of missing values.
As explained in the section 2.2.4 and during initial exploratory data analysis input Hestia data in the various spatial contexts is highly variable and skewed. Therefore before using this variable for training of the model, the emission data is log transformed. As during testing it was found that doing so created a leap in model performance in terms of explained variance. Similar behaviour is observed in Stevens et al. (2015) where the authors argue that log transforming the target variable likely helps the model in finding better splits, as the target is more uniformly distributed after log transformation.

Power plant data
As explained in section 2.1.3 power plant data, such as for example the installed capacity are assigned to the raster cells. Cells without a power plant associated with get the same property but set as zero.
From the Byers et al. (2021) dataset the following properties are used: GWh, Capacity in MW, Estimated GWh in 2016, fuel type and the commissioning year. Where the former two are used as a filter. This filter works as follows, when data is necessary for the year 2013, a power plant with commissioning year 2014 will be removed from the dataset. This means that it is assumed that there are no emissions prior to the commissioning year related to the power plant. The fuel type used in the power plant also works as a filter power plants with fuel types, hydro, solar or wind as either primary or secondary fuel source are excluded from analysis. This involves the assumption that power plants with these fuel types do not result in higher emissions.

Population and nightly radiance
The night time lights data available is uncorrected for stray light. Stray light means that some cells without radiance might also capture part of the radiance of the neighbour. Stray light corrected data is only available from 2014 onward, therefore uncorrected data will be used for now.
The VIIRS night time light data available through earth engine is a monthly average radiance value. To produce an annual composite, the sum of monthly average radiance is used. The idea is that the sum of a grid cell in a year captures the temporal variations of radiance more accurately. Rather than taking the average value of each month and creating a yearly composite image like that.
For the year 2012 the first available data comes from April. This creates an issue when comparing the

2012 composite to the the other years. To correct for the first three missing months the sum composite of 2012 is increased by 25% percent. This means that it is assumed that the growth in the first months without data is the same as the other months.
The original resolution is ~ 460 m, therefore the data has to be re sampled to a 1$\times$1 km resolution.

Population data is all ready in a rasterised format per year. Therefore this data does only need to be re-sampled to the correct resolution and is without processing an annual composite.

PM2.5
The air quality data is available in a rasterised format and forms an annual composite. Therefore, preprocessing of this variable consists of re-projection, and aligning resolution. As discussed before, any raster cells with missing values are omitted.

GDP
Like the PM2.5 data this data is in a rasterised format. Like PM2.5, only missing value removal, re-sampling the resolution and re-projection is necessary.
This data is only available for the years of 2000 and 2015. Therefore, the data for other years has to be imputed. As GDP and economic growth is often captured and described by percentage change, the assumption here is that GDP follows exponential growth. As there is no additional data available for now, it is assumed that the growth is constant over these 15 years.
Computation of GDP in a specific year is done by first calculating a growth year per cell. This is done using the following formula: $GrowthRate_{i,j} = (\frac{GDP_{i,j,2015} - GDP_{i,j,2000}}{GDP_{i,j,2000}})^{\frac{1}{15}}$ For each cell with coordinates i, j a growth rate $GrowthRate_{i,j}$ is calculated over all years. To arrive at an annual growth rate the fifteenth order root is taken of this number.
Then to find the value of GDP in a specific year for cell i, j the value of GDP in 2000 is multiplied by the yearly growth rate by the power of years difference between 2000 and the year needed. $GDP_{i,j,year} = GDP_{i,j,2000} \times GrowthRate_{i,j}^{year-2000}$. Arriving at a GDP estimate per cell for the desired year.
Other ways to temporally interpolate GDP could be preferred and lead to more accurate results(Chen et al., 2007) . This however lays beyond the scope of the current model.

Spatially lagging variables
Spatially lagging a variable entails including information about the value of that variable for the cells around it. Some variables have high spatial auto-correlation, for example the presence of a road that stretches through multiple cells. Therefore not only including the value of the variable in that particular cell, but also the value of the co variate in the cells around it contains information for the model.
In more detail, the spatial lag of a cell is defined as the average value of the neighbour cells. The definition of a neighbouring cell however could be operationalised in multiple ways. In the centre of the raster the neighbour concept is relatively easy to understand, each cell would have 8 neighbours, as the cells diagonally are also included, each getting equal weight. This concept, is known as queen neighbours, the centres of the raster points would be approximately 1 km apart. However using this method there is no maximum distance for neighbours to be considered out of range. And as city boundaries and coast lines are not strictly rectangular, this method was found to be less reliable.
Therefore a distance band is used, where a circle with a radius of 1 km is drawn around each point. Any point that falls within this radius is considered a neighbour. During robustness test this method results in less neighbours on average (3.4 neighbours versus 7.7), but slightly higher model accuracy. Unless the maximum distance to be defined a neighbour is set each spatially lagged variable except the power plant data (such as estimated capacity) is significantly spatially auto-correlated.

2.1.4. Exploratory data analysis
The results of the exploratory data analysis will be covered in the results section. This first consist of evaluation and the presentation of the descriptive statistics for the created data set.
The data analysis are only presented for cells with non zero emission observation. This as the statistics for the data should be presented in both non transformed form, and log transformed form. Log transformation of zero observations results in infinite values introducing error in the summary statistics.
Out of the total amount of observations, N = 68607, only 468 (that is less then 0.07 percent of the total)

observations have zero observed emission values.

Beside the summary statistics a pair plot for each combination of variables will be presented. This plot shows an histogram, along with an bi-variate kernel density plot providing a quick overview between relationships of these variables. Finally to give the reader a sense in how these variables are spatially distributed the spatial covariates along with the observed emission variables will be plotted on a map.

The appendix will cover descriptive statistics for all spatial covariates, including the few that where not included in the final model. And show the earlier described pair-plot for the different analysis years and for 5 different locations.

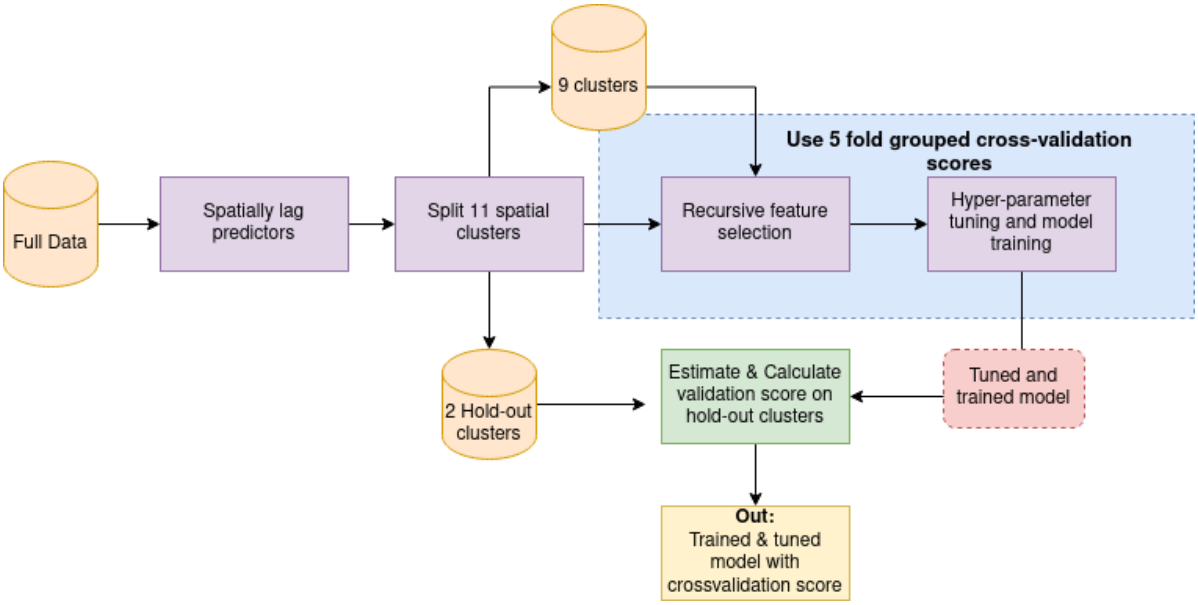


Figure 2.2: The steps used to obtain the optimal hyper-parameters and features for the model

2.2. Model and selection of best model

In the following section the method and steps to construct the model are discussed. As an outline, first the underlying statistical model, the Random Forest, will be discussed in section 2.2.1. Having the full data set it is possible to find a robust model that performs best given the current data availability. To achieve this a pipeline is setup which is depicted in figure 2.2. After the calculation of spatial lag for all variables the data is split into eleven unique spatial clusters, this is done according to the description in 2.2.4. Two spatial cluster are used as a hold-out set to asses the final performance of the tuned and trained model. The remaining observations are first used for recursive feature elimination and hyper-parameter optimisation. The explanation for recursive feature elimination is described in section 2.2.2. And explanation of hyper parameter selection is found in 2.2.5.
Performance assessment during recursive feature elimination and hyper-parameter tuning is done using grouped cross validation, which is described in section 2.2.4. This section also contains explanation about the scoring metrics used to evaluate the model during the various steps in the process.
The result of recursive feature elimination are a selection of features that are found to have most explanatory power. The obtained features, will then be used as features during hyper-parameter tuning. The purpose of these two methods is thus similar, search the space of features and hyper-parameters for optimal cross-validation scores. Then using the optimal parameters and features, a model is trained on all clusters except the hold-out cluster. Then grid cell values of emissions from the hold-out set are estimated using the found optimal model from which the final model validation score can be calculated. Which gives quantitative insight in model performance and generalisation.

2.2.1. Statistical learning algorithm

For this study a Random Forest regressor is used. In previous studies ensemble methods are often used because of their flexibility and generalisability (Stevens et al., 2015; Hsu et al., 2022). Ensemble methods rely on building multiple models and using a voting mechanism to reach a single prediction. This increases the generalisability of the model (Zhou, 2021), which is often the main purpose of using a statistical learning algorithm in a spatial context (Meyer et al., 2018).

The Random Forest model fits this description. The random forest regressor is an ensemble method, it constructs multiple decision trees, each trained on another part of the data. Using a sampling technique called ”bagging”. Then this panel of multiple decision trees, hence ”forest”, is used to make an estimate. To reach a consensus vote out off these trees, a voting mechanism has to be used. For regression this would be taking the mean of all the individual decision trees.

Bagging, also known as bootstrap aggregation, entails taking random samples of the observed values. This prevents over-fitting, decreasing the variance without introducing extra bias (Breiman, 2001). This can be explained as each tree is trained with a different part of the data.

The understand the working and classification of a decision tree imagine the task of identifying whether a piece of fruit is an apple or an orange. A single decision tree would ask, is the colour orange? For all answers ”yes” the decision tree would predict it is an orange, else (if the answer is ”no”) an apple. The training a single decision tree thus means determining which question is the most informative. To start it would do this with the full data set, and will recursively split the data using the most informative splits. For classification problems the Gini index (measuring entropy abd information gain) is used to determine the most informative split. For regression, the current task, real mean square error (RMSE) is commonly used (Burkov, 2019).
The following parameters, further discussed in section 2.2.5, can affect the training process. Firstly, the minimum samples per split constrain certain splits, that for example split 99 of 100 data entries to a certain side. This constraint therefore prevents overfitting. The same goes the parameter min samples per leaf. The leaf, the end of the decision tree limits the splits in the end stages of training. The max tree depth limits the amount of decions each tree makes. In Stevens et al. (2015) the authors find that taking the mean of all regression trees results in the lowest error, in this study therefore the consensus voting mechanism will be similar.

To verify the choice of statistical model, multiple test where ran using different statistical learning algorithms using the default parameters using cross-validation scores, full results are included in appendix C.1. Another ensemble method tested is the XGBoost algorithm, which is used in comparable studies (see Hsu et al. (2022)). However with similar hyper parameters the Random Forest performance is quite similar to the that of XGBoost. Not only ensemble methods where tested but also Support Vector Machines (SVM), Elastic Net, Lasso and linear regressions where compared to the Random Forest outcomes. It was found that the Random Forest outperforms each of these statistical on various metrics, such as mean absolute error, mean absolute percentage error except for explained variance where SVM scores best. The generalisation and build in methods to prevent overfitting are likely important properties for the performance of the Random Forest in this case.

2.2.2. Feature selection

In addition to leave location out cross validation, iterative feature selection also reduces the risk of overfitting a spatial Random Forest model (Meyer et al., 2018). When including all predictors, including the lag of predictors there are 26 plausible variables. Running recursive feature elimination decreases this number to six use full predictors. Using fewer covariates reduces model complexity. And is aimed to find the most relevant predictors for optimal model performance while decreasing risk of overfitting (Kuhn et al., 2013).

The recursive feature elimination method used in this study is the algorithm proposed in Guyon et al. (2002), which avoids. The algorithm is a wrapper-method and hence trains and assesses a new model in each iteration. The algorithm starts with all features and recursively eliminates features with the least feature importance. Using the cross-validation splitting and scores described in 2.2.4 the model with the best cross validation score is selected and is the result of the process. For this operation the untuned random forest is used with 100 trees. Running this operation it is important to optimise only on the training set.

2.2.3. Feature importance and effects

Once the features are selected the importance relative to each other also contains information. To do this permutation feature importance is used. Using the hold-out test set the increased error after removing a feature is measured. This is compared to the baseline model, the fully trained and tuned model with all features. As the Random Forest is non deterministic, this operation has to be repeated multiple times. Calculating the mean and standard deviation of these iterations give a sense of certainty and magnitude of the feature importance.

In addition to the feature importance, partial dependence plots are used to gain an insight in how the covariates affect the models estimate (Friedman, 2001). Doing so also gives insight in the type of relationship (linear or more complex) between co-variate and estimated emissions.
An important limitation of this technique is that it assumes that the input variables are independently distributed from each other. Initial exploratory data analysis showed the opposite. Nonetheless, in combination with feature importance it helps with the interpretability of the model and does not assume a linear relationship, unlike using partial correlation for example.

2.2.4. Model performance assessment

Once a model has been trained, it can be used to produce an estimate. To understand how the model is performing a series of test is designed. These test entail comparing observed values,from the Hestia data set, with model outcomes.
Being a spatial model, the sampling strategy, which data entry ends up in the training and which in the test set, and validation method are crucial. The dataset created according to the section 2.1.2 will act as both a train and test set in cross validation, how it is repeatedly split between train and test, is described in section 2.2.4. For the cross validation strategy however, the data set needs to be spatially clustered, how this is done is described in detail in section 2.2.4.
Spatial clustering and using grouped cross validation is necessary because the dataset contains multiple observations for the same location. So ensuring that a certain location is not in both the test and train set helps give better insight in how the model spatially generalises.
This would work for points, however the goal of the final model is to be able to predict entire cities, which have might be spatially correlated. How this spatial clustering of the Hestia data set observations is performed is explained in the subsection 2.2.4.
To quantitatively compare the results of different model outcomes a variety of scoring metrics have been used, these are described in section 2.2.4. In general, these metrics quantify the difference between the test set observations and model estimates.

Spatial clustering

The Hestia dataset used in this study covers three geographically divergent areas: Baltimore (Zhou and Gurney, 2010), Indianapolis (Gurney et al., 2012) and the LA Basin (Gurney et al., 2019b) located in the United States.
The coverage of Hestia in the LA Basin covers a large area, containing also non urban areas. Therefore, nine boxes where drawn on the map, this is visualised in figure 2.3. This covers the important urban centres in Hestia data of the LA Basin.

For validation it is important to have several unique groups in terms of location (see section 2.2.4 and (Meyer et al., 2018)). Therefore observations from the LA Basin are spatially clustered using a KMeans algorithm. The two areas outside of the LA Basin, Indianapolis and Baltimore, both are in a separate spatial cluster. Because these nine areas are regarded as a urban centres in the LA Basin, the amount of clusters for the KMeans algorithm is set at nine as well. The spatial coverage of these clusters in LA are visualised on the map in figure 2.3. Test with other amounts of cluster made the inter cluster ranges bigger or more sporadic.

With nine spatial clusters in the LA area and one for Indianapolis and one for Baltimore there are eleven unique clusters in total. In appendix A the similarity of the distributions of emissions is analysed. Using a two sample Kolmogorov-Smirnov test it is possible to statistically test whether the distribution of emissions is similar for the different spatial clusters. Each possible pair of spatial clusters is tested, however none of them are statistically significant similar when testing at a 95% confidence level. The Kolmogorov-Smirnov statistic ranges between 0.10 and 0.91, where a lower value means that the distributions are more similar and a value of 1 means that the distributions are dissimilar.
To test if the mean of observations of the various clusters various significantly a two sample T-test is preformed. This finds that the mean of 5 clusters are similar. The spatial clustering hence does not work ideal, but some similarity between input observations is found. The main conclusion of running this analysis is that the input target data is highly variable in the different spatial contexts.
This means that, as expected, the input groups are heterogeneous variable which could help with generalisation of the model.

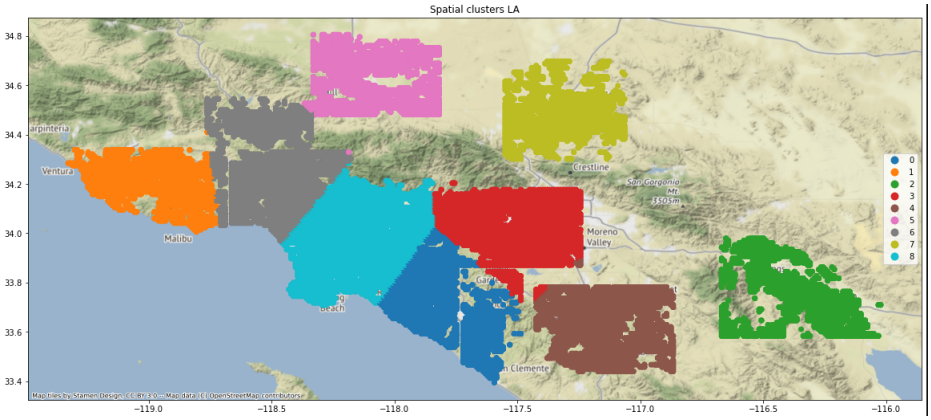


Figure 2.3: Showing the results of the spatial clustering through KMeans in the LA Basin

Cross validation

To validate the model, to select features and to asses the generalisability of the Random Forest Model grouped k-fold cross-validation will be used. Cross validation entails training the model on a train set and hereafter assessing performance on the test set with for the model unseen examples. Regular, or random, five-fold cross validation would be creating for instance five folds and holding out randomly 20% of observations for the test set.
However, it has been empirically proven that for spatial regression models using random k-fold cross-validation leads to over optimistic model performance scores Meyer et al. (2018). The outcomes of the aforementioned study plead for using either ”Leave Location Out Cross Validation” (LLOCV) and ”Leave Time Out Cross Validation” (LTOCV). As this would lead to more reliable view of model performance.

Therefore, Leave Location Out Cross validation is used for model assessment. This is achieved by grouping each observation into one of 11 unique spatial clusters of observations, described in 2.2.4, and using one or multiple of these clusters as a hold-out set. This means that the remaining groups are used as the training data. This is done k times so that each group served once as a test set.
This is known as k fold group cross validation. Using this method ensures that each observation of a spatial cluster can never be both in the test and train set. This is use full compared to non-grouped cross validation as now the observations are spatially comparable. Which is similar to the final model purpose where the goal is to predict entire cities at once. Other grouped cross validation methods have been tried, such as using a group shuffle split strategy. However, doing so is not aligned with the goal to predict entire spatially common groups (read new cities) at once. As a certain set amount of observations (30% for example) is placed in the test set per group. Reducing the spatial similarity of points.

Using k fold grouped cross validation, also means that k unique scores between test and train set are obtained. The cross validation score is the average of those k scores. In this study multiple scoring metrics are used, which are described in section 2.2.4.

There are eleven unique spatial groups, however during the tuning and feature selection of the model, one cluster is held out to be used as an external validation set of the tuned model for final assessment of model performance, this is illustrated in figure 2.2.
The effect of changing the amount of folds (k), has also been tested. Using fewer folds prevents the model from overfitting.

Performance metrics

For quantitatively assessing the model performance and calculating cross-validation scores to use during model feature selection and tuning, scoring metrics will be used. These metrics express the difference between y, the observed value of emissions and the estimated $\hat{y}$. When calculating cross validation scores the following five metrics are used: two sample Kolmogorov-Smirnov, Mean absolute

error (MAE), R^2, Mean absolute percentage error (MAPE) and proportion of explained variance.
The two sample Kolmogorov-Smirnov test is used to test whether two sets of observations come from the same distribution. The Kolmogorov-Smirnov statistic measures statistical divergence between the two distributions, a value closer to zero means they are more similar. Using this test a significance level can also be calculated. If the p value is greater then 0.05 it is possible to reject the null hypothesis, that the samples come from different distributions. During model assessment however this happens rarely. Still, this statistic is use full as it can be used as a measure to assess how similar the distributions are, with a value closer to zero meaning more similarity. If model A has a Kolmogorov-Smirnov score of 0.5 and the B of 0.05, the distribution of model B is better.

Mean absolute error is the mean absolute difference between observed and estimated. This metric can be misleading, as it can be skewed upwards by large observations, therefore the mean absolute percentage error is also calculated. The absolute difference between observed and estimated is then divided by the observed value, giving insight in relative deviation. Finally both R^2 and explained variance are used. These two metrics differ slightly, as the R^2 is the sum of squared residuals over the the total observed variance. Explained variance is the model predicted variance subtracted by the observed variance. Both give insight into the explanatory value of the model.

2.2.5. Hyper-parameter tuning

Having obtained the relevant features, the hyper-parameters for the random forest need to be optimised. Again the purpose here is two-fold. Firstly it ensures model performance and secondly some parameters, such as the minimum amount of samples per split could prevent overfitting and thus help generalisation. The hyper-parameters are not ”learned” or optimised during model training and are set in advance.
An overview of the available hyper-parameters and the space searched is found in table 2.1. For the Random Forest mainly the amount of trees and the number of features considered when creating a tree split are the most important (Breiman, 2001). Minimum samples per split, minimum leaf samples and maximum tree depth are parameters that could prevent overfitting. The discrete bootstrap parameter controls the method of sampling when building a tree. When false all observations are used to build a tree.

To search this space of parameters a randomised search with 75 iterations has been used. This is done as the search space of the hyper-parameters is quite large and as it was unclear how these parameters would affect model performance. The assessment of model performance is using a cross-validation score (See section2.2.4). Having multiple scoring metrics it is possible to assess and refit the models on the best score. For this, a single scorer needs to be selected. Experimentation lead to the conclusion that using the mean absolute error leads to more stable results also for the other hyper-parameters. The best estimator scores highest on both MAE and MAPE and is close to the highest explained variance score. Similair behaviour, in terms of scoring, is observed in when comparing the effect of 10 fold cross validation. In both cases mainly the estimator with optimal parameters for the Kolmogorov score, do particularly bad on other metrics.
A plausible improvement could be to use a grid-search, so extensively trying each possible combination of hyper-parameters, however due to the computational intensity of this task a randomised search was used in this study.

For this process the results of the ten and five fold grouped cross validation is tested. On the log transformed scores the five fold grouped cross validation outperforms the ten fold by a long shot. The hyper-parameters mostly affected by this change are the amount of estimators and the minimum samples per leaf, which are both higher in the case of using five fold group cross validation.
When comparing the cross validation scores during hyper parameter optimisation, the ten fold grouped cross validation outperforms the five fold, on the test set metrics however, this is reverse. This could be a sign of overfitting, which thus seems to reduce when using five fold grouped cross validation.

Parameters	Search space
N estimators	[50;500] with steps of 50
Max features	Auto, Sqrt, log2 and None
Min samples split	[2;16] with steps of 2
Min samples leaf	[1;8] doubling each time
Max tree depth	[10:210] in 10 steps including no max depth
Bootstrap	True or False

Table 2.1: Table with the hyper-parameter space searched using random grid search

2.3. Model generalisation

As discussed, once confidence in and knowledge regarding the model performance has been achieved the model estimates can be generalised to new cities. This sheds light on the final sub-question: *What are the trends in spatial distribution in cities?* For this the full model, trained with all available data and the found optimal hyper-parameters is used.
This model is then used to estimate emissions in the eight cities. For each of the four analysis years an estimate is made using the model, from these estimates it is then possible to calculate the IQR coefficient. In the result section only the aggregate results will be presented. For full analysis per city please refer to the appendix D.

2.3.1. Analysed cities

As the used Hestia dataset only covers three cities in the United States, the generalisation of the newly developed model will also remain conservative. Therefore only cities in the US are considered for analysis. Due to time constraints, only eight of the top ten most populous cities are analysed. These are New York, Chicago, Houston Phoenix, Philadelphia, San Antonio, San Diego and San Jose. Omitted from this list are Los Angeles, as this city is in the Hestia Data set. Dallas is omitted because of technical limitations.

2.3.2. Quantifying emission distributions

Having an emission estimate of a city it is possible to visually inspect how emissions are distributed. But a natural question that arises is how equal are the emissions distributed within the city. For this the inter quartile range coefficient, IQR coefficient hereafter, is used. The IQR coefficient is a measure of statistical dispersion and applied to the current case reflects the distribution of emissions within a city. The coefficient is calculated as follows: $IQR_{coef} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$. Where Q_1 and Q_3 stand for the value of estimated emissions in the first quartile (25 percentile) and the third quartile (75 percentile) respectively. Importantly, this number falls between zero and one, and thus can be compared across cities.
As explained the emission target variable is log transformed during model training. This means that the model output, the emission estimate, is also still in the log transformed form. Experimentation with these results show that in log transformed form, the IQR coefficient estimated is too low, see C.3.1. Therefore, the model output needs to be ”back transformed”, from log space to the original space. If Y are the original emission observations, the model is trained using the natural logarithm $Y_{log} = ln(Y)$. The model estimate, $\hat{Y_{log}}$ can thus be transformed to an untransformed estimate, $\hat{Y}$, by doing $\hat{Y} = e^{\hat{Y_{log}}}$. This measure is thus used on the back transformed emission estimates. As, after log transformation the distribution of emissions are remain stable. This behaviour is discussed in more detail and with empirical data in appendix C.3.1 . The effects of using other measures of statistical dispersion, such as the Gini-coefficient or standard deviation have been tested. However, during experimentation it was found that the coefficient of sta- tistical dispersion is the most stable compared to the Gini-coefficient. For full results please refer to appendix C.3.1 The IQR has smaller absolute relative differences between estimated and observed.
The IQR will give an insight in how emissions are dispersed in various cities over the analysis years.

This metric outlining how the IQR changed over the years. This can then be interpreted along with a visual distribution of emissions for each of the ten cities.

3. Results

In the following chapter the results are presented. This will use the following structure: first the input data for the model will be described in the form of exploratory data analysis. The chapter then flows into the discussion of model performance compared to the observed values. This will then cover the robustness of these results when using different design choices. Having discussed the robustness towards design choices, the selected features and their model importance and partial dependence will be discussed. Followed by this is the inspection of cells with high absolute and relative errors. The section ends with the spatial trends found when generalising the model.

3.1. Input data & Exploratory Data Analysis

As covered in the methodology section the choice of input data for the model is fairly important. Once the full data set has been created the first step is hence recursive feature elimination. Not all input features will hence be discussed, only the six relevant features of the model

Recursive feature elimination trains and evaluates models in succession and removes the least important feature from the analysis. Starting at 22 variables and ends with one. The model with the best mean absolute error cross validation score is used. This results in the following six features nightly radiance, gdp, pm25, population, pedestrian and intersection density. Remarkably, none of the lagged variables where selected, also not when defining the neighbours in different manners. The stability of these results was tested by comparing 5 and 10 fold grouped cross validation. And by changing the spatial cluster to hold-out. In general, less stable models with worse metrics also have more features, of which the importance is often low. This is taken as a sign that using recursive feature elimination prevents overfitting and that the aforementioned features contain the most stable and important information for generalisability.

Having insight in the six important features the descriptive statistics of these covariates are presented, see table 3.1. For full descriptive statistics, of all 22 initial variables, please refer to the appendix.
From these statistics it is clear that the data is highly variable, for almost all of the variables the observed standard deviation is higher then the mean. Also the mean is commonly higher then the median of values, meaning a skewed distribution.

Having discussed the descriptive statistics, the relationships between the input variables are inspected. This is visualised in figure 3.1. Starting at the top of the figure, the distribution of nightly radiance is visualised. As also inferred from the descriptive statistics, this distribution is skewed and not normally distributed. This seems to be the case for all other variables except PM25 and Logged Emissions, for the former this makes sense, as this is the operation log transformation of a variable achieves.
Nightly radiance seems to have a positive increasing relationship with the other variables. As the magnitude of nightly radiance increases, the amount of observations become less. Therefore the density and relationship between nightly radiance hence also decreases quite a bit and is less similar.
GDP shows similar behaviour, albeit more linear behaviour for example between GDP and population.
PM25 Does not show this positive increasing relationship, likely as the variable is more normally distributed For PM25 in general there seems to be some centre of gravity and for higher values of one of

N = 68139	Nightly Radiance [$nW/cm^2/sr$]	GDP [$]	Pm25 [μ/m^3]	Population [#]	Intersection density [#]	Pedestrian intersection [#]	Emissions [kgC]	Logged Emissions [-]
mean	223.5	5.109E+07	10.2	9.4	63.8	223.9	2.500E+06	12.8
std	304.5	7.193E+07	2.6	13.7	64.3	298.0	1.697E+07	2.4
min	1.2	5.759E-02	3.1	0.0	0.0	0.0	3.552E+00	1.3
25%	18.9	1.237E+06	8.2	0.1	4.0	24.0	7.576E+04	11.2
50%	128.2	2.045E+07	10.3	3.2	44.0	114.0	5.590E+05	13.2
75%	346.1	7.647E+07	11.9	14.3	112.0	318.0	2.340E+06	14.7
max	25861.1	9.378E+08	20.1	220.5	661.0	3446.0	1.475E+09	21.1

Table 3.1: Descriptive statistics for the selected features, zero emission values are omitted, note that descriptive statistics for both emissions and the log transformed form are presented, of which only the former is used in model training.

the other variables does not increase much.
Population shows a similar distribution as nightly radiance and GDP. Both pedestrian and car road intersection density show a positive relationship with the other variables. The pedestrian roads however increase less quickly, visually the relationship seems more flat, compared to the car road intersection density. Pedestrian intersection density is however more skewed and has a higher mean median and maximum value. What happens is thus that this spatial variable is very high in certain grid cells, likely in urban cores where the other variables have all-ready achieved their maximum value.

Perhaps the most interesting relationship is the row of plots at the bottom, showing the relationship between logged emissions and the spatial proxies. These relationships doe not seem linear. Rather, positively increasing and then quickly flattening, or converging for higher values of the other spatial covariates. PM25 the only other variable not normally distributed does not show this converging behaviour.

In addition to these plots, to get a feel for the data it is use full to plot the covariates and their magnitude in the the spatial context. This is visualised in figure 3.2. Especially nightly radiance is highly concentrated in the middle. The This seems to be roughly where the highest population density and pedestrian intersection density is found. This likely indicates the centre of the city. Intersection density also seems to be correlated with population. PM25 mainly is concentrated around the coast and not like the other variables. The logged emissions seem to be spread quite equally over the spatial extent of Baltimore. This is mainly because of log transformation for which the values are closer together. There are breaks in the emission distribution of Baltimore. The relationship with the covariates is not directly visually discernible, however the high concentration zone seems to fall in the same area as the areas with the highest nightly radiance.

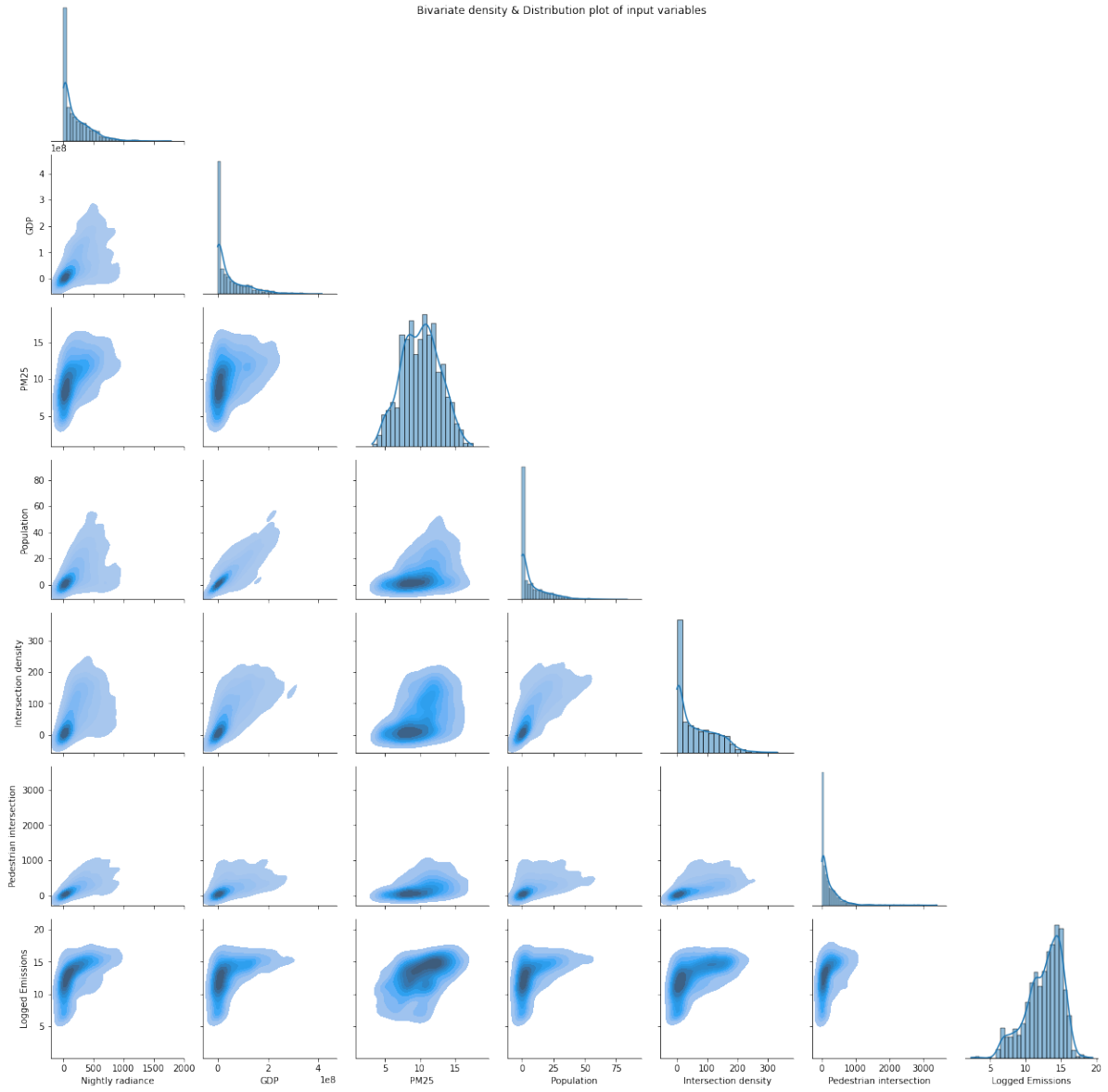


Figure 3.1: Paitplot showing the Bivariate density and distributions for each variable combination, also the target variable, logged emissions is included.

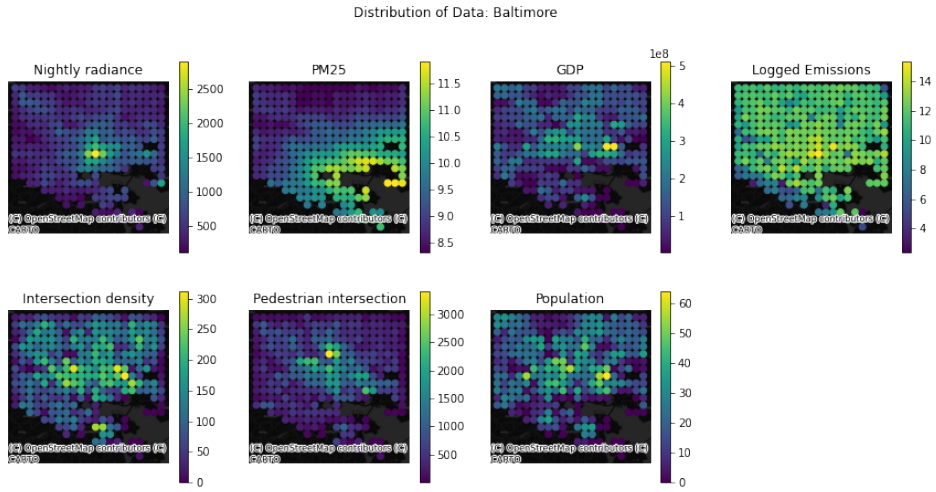


Figure 3.2: Plot showing the distribution of various proxies for the year 2015 in Baltimore, mainly emissions are spread all over the city, where the spatial proxies seem to be more concentrated around certain areas.

3.2. Model performance

Having covered to input data, the model can be put together and trained using this data. The presentation and analysis of model performance has the following structure: first the test set performance both in transformed form and original space will be covered. These findings suggest that the model only works in the log transformed space. Then the section will cover whether back transformed target regression could solve these issues. Then the results for will be disaggregated per year, giving insight in the yearly generalisability. And then the robustness of these results will be assessed by testing how these results are affected by the hold-out data set. This helps prove the point that these model results are robust and that this is the maximum performance that could be achieved using a Random Forest and this data.
The section hereafter , see sectopn 3.4, will further inspect the model error, partial dependence and relative feature importance of the most robust model.

3.2.1. Test-set performance assessment

To asses the spatial generalisability of the fully tuned and trained model, the model is used to produce estimates on the hold out set observations. This gives the ability to compare these estimates against the observed values and to calculate the performance metrics as described in section 2.2.4. Additional descriptive statistics for the covariates, observed emission values, as well as estimated can be found in the appendix C.2.1. These results are presented in two ways, both in the log transformed space and secondly in the original space. Please refer to table 3.2 for test set results. The test set observations come from different spatial clusters and years. The reason for presenting the results per spatial cluster and year is that when not splitting it up the results could appear better then they are for some of the spatial clusters. Further discussion of this is found in the appendix C.2.2 and results are found in table C.4. This is inline with the findings of (Meyer et al., 2018). Therefore the results are presented here per spatial cluster in table 3.2.

Looking at table 3.2, the metrics in log transformed space indicate a moderately well performing model. In log transformed form, the mean absolute error for cluster three is 0.96 and 1.23, which means that in log space, the prediction is off on average by $e^{1.23}$ Easier to interpret is the mean relative absolute error, which is a relative percentage error calculated per cell. The mean of this metric gives an indication how far the model is off on average per cell. In log transformed space this value falls between 7 and 12 percent for these two clusters. When looking at the original space however this variable this error becomes 171 and 416 percent on average per raster cell.

	MAE	MRAE	R^2	R	Kolmogorov stat	expl_var	Log Transformed
Cluster							
3	0.9630868	0.072965	0.468048	0.69392	0.119924	0.472708	Yes
3	2404580	1.718238	0.012429	0.164933	0.119924	0.027151	No
7	1.236299	0.127662	0.559271	0.776239	0.139814	0.582796	Yes
7	1215962	4.156965	-0.002452	0.029731	0.139814	0.000632	No

Table 3.2: Test set performance metrics for the two spatial clusters, both in log transformed space and back transformed, note the deterioration of these metrics in untransformed space.

The R^2 in log transformed space indicates that the model is able to explain about 50 percent of the model variance. In non transformed space this number shrinks substantially, even becoming negative indicating that a base regressor predicting the mean of observations of each cell would have given a better estimate. This deterioration between normal space and log space is also observed for the explained variance.
The spatial correlation coefficients are all significant, interesting again is the high spatial correlation (the predicted value of a raster cell is high if the observed value is high) that decreases in the original space, however remains significant. It could be possible that this number is influenced by some cells with relatively high emissions, as has been a known problem of emission inventories (Hutchins et al., 2017).
The single metric that does not deteriorate after log transformation is the Kolmogorov-Smirnov statistic. Indicating that the samples in both log space and original space have the same chance of being from the same distribution. These values however are not significant, but the fact that this statistic is closer to zero indicate that the divergence between the two distributions are smaller.

In short, each of these metrics thus indicate that the model has better accuracy in log transformed form. It is possible that these results give a highly positive view of performance as essentially the space to evaluate the model in has shrunk. However even only having emission estimates in log space is still valuable. As this provides insight about the distribution of emissions, which according to the Kolmogorov-Smirnov statistic indicators remain stable.
Back transforming the results to kgC however results in very low explained variance and large MEA and MRAE errors.
The performance only in log transformed space suggest that back transformed target regression. This entails training the Random Forest in log transformed space but evaluating the model tuning process and feature selection in the original space, which could potentially increase model accuracy. This is tested in appendix C.2.2, doing this test however yields no additional promising results, parameters and features do not change, therefore this path was not pursued further.

Yearly generalisability
Table 3.3 shows the hold-out performance metrics isolated per year in log transformed space. Inferred from this table is that the performance metrics remain stable over the analysis years. With a maximum absolute error in 2013 and a minimum in 2012, the maximum difference between said years is about seven percent. The conclusion is that the model generalises with similar confidence over the analysis years, without much error increase as the analysis years change.

The figure 3.3 depicts the kernel density function of both observed and estimated distribution over the years. It seems that the Hestia data does not vary much on a yearly basis. This is unlike the kernel density plots that show more yearly variation. While the peak and the distributions look approximately correct, the models show a bimodal distribution which is not observed in the Hestia data, this is likely a sign of overfitting on the train set.

Effect of hold out set on features and tuning
The hold-out set is now picked arbitrarily a natural question is then to see how the hold out set affects the model selection and performance. When doing five fold grouped validation, two spatial clusters have to be picked as hold-out test set.

	MAE	MRAE	R^2	R	Kolmogorov stat	expl_var
Year						
2012	1.030	0.090	0.677	**0.824**	0.075	0.677
2013	1.104	0.099	0.630	**0.801**	0.091	0.632
2014	1.068	0.095	0.659	**0.817**	0.080	0.663
2015	1.099	0.097	0.652	**0.813**	0.085	0.653
Total	1.075	0.095	0.655	**0.813**	0.077	0.650

Table 3.3: Table showing the metrics over a variety of years, these results remain stable for the analysis years

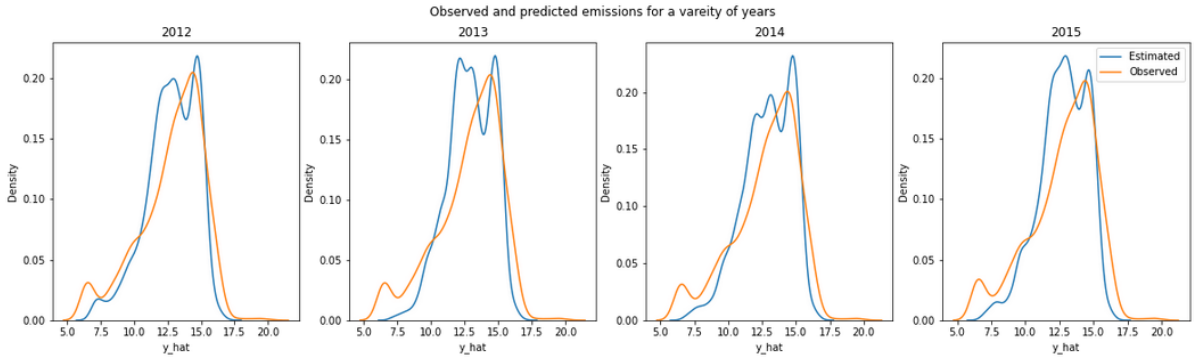


Figure 3.3: Distribution of observed and predicted values in log space, it is visible that the model distribution changes considerably over the years the observed values do not show this behaviour

Referring back to the pipeline depicted in figure 2.2 and the issue tree in figure 2.1. There are several results that could be affected by changing the training data, these are the selected features, hyper-parameters and the performance scores.

Following the pipeline depicted in figure 2.2 the first step is recursive feature elimination. The effect of picking different hold-out sets on the recursive feature elimination process is not immediately clear as can be seen in table 3.4. But the choice of hold-out sets determine the train set that the recursive feature elimination, RFE hereafter, bases its results on. In the three experiments the outcome of the RFE seem to be relatively stable.
The second question that arises is how the results of hyper-parameter optimisation change under different training and hold-out sets. The results of said experiment are given in table C.6, in the appendix. Going through the pipeline, the cross-validation scores remain relatively stable for different hyper-parameter sets, the effects of tuning is thus very marginal on the final score. The parameter search results however are a bit different. Therefore in the appendix C.2.2 an additional test is ran to compare how these hyper-parameter sets affect the cross validation scores.

The main conclusion of these experiments is that while there is some variation in the found optimal hyper-parameters, the results are quite robust and do lead to the best possible generalisation performance. The test set performance is only marginally dependent on the hyper-parameters. The fact that the same features are selected also for different hold-out sets give confidence to the fact that the same parameters would be found for other data.

Effect of hold-out on performance
Knowing that the optimal hyper-parameters and features only marginally depend on the spatial clusters in the hold-out set. A secondary question that arises is whether the reported test-set performance metrics, given the optimal parameters, are dependant on the choice of hold-out sets. As at the moment the two spatial cluster three and seven used for inferring conclusions are picked arbitrarily. The effect

Test set	Predictors RFE
[0,1]	No change
[8,5]	lagged variables of GDP and PM25 included
[10,1]	No change

Table 3.4: Results of RFE for different hold-out sets. As one can see the result of this process is relatively stable. However during these experiments in one instance the selected features where slightly different.

of the hold-out sets is tested in a similar manner as before in section 3.2.1.

To perform this experiment the hyper-parameters are kept constant. For each iteration in of the experiment two spatial clusters are picked randomly to be the hold-out test set. A model is trained on the remaining training set. Then for each of the two clusters the test set performance is calculated independently and results are stored. This is done a total of 35 times. This gives 70 unique results per metric which can be grouped per spatial cluster.
This is visualised in figure 3.4. Starting at the top-left plot, depicting the MAE and working left to right towards the bottom, each of these metrics will be discussed. The MAE seems to be comparatively stable for clusters 0-8, with values between 1 and 1.4 typically. The MAE for clusters nine and ten are worse, with a value of about 3.5. Looking at the MRAE similar behaviour is observed, where the first and third quantile of the box-plot are between 0.07 and 0.15. Which means that per raster cell on average the prediction is between 7 and 15 percent off in half off the cases.
Again, looking at the first and third quantile, R2 scores are typically between 0.18 and 0.49 for the different hold-out sets. Execeptions to this are spatial clusters nine and ten which have big negative R2 scores even in log standardised space. For explained variance and R this difference is smaller where only cluster ten remains troublesome to predict. The Kolmogorov-Smirnov statistic behaves in a similar manner as the MAE and MRAE where spatial clusters nine and ten have comparatively worse performance.

Zooming out it seems that for some clusters the test performance deteriorates. This is true for clusters nine and ten. Which are Baltimore and Indianapolis, the two spatial clusters that where not in the LA-Basin data. It is very well possible that as the model has mainly trained on Hestia data from the LA area the generalisation performance to other geographical context deteriorated.
It seems that the performance on MAE, MRAE, R2 and Kolmogorov-Smirnov are correlated, the patterns between these metrics look similair in the plot(for example when comparing clusters 0 and 1, the MAE and MRAE for spatial cluster 1 are higher). The spatial correlation coefficient (R) and explained variance seem to behave a little different.

3.2.2. Key takeaways test-set performance

The model is trained and assessed according to the key issues depicted in the figure 2.1. To test whether the model is robust, test regarding the algorithm selection are covered in appendix C.1. The Random Forest was a clear winner. The robustness of other points, such as feature selection and the result of hyper-parameter tuning have also been tested, of which the results are foung in section 3.2.1. No big differences in model performance nor in the outcomes are found when varying the training data for the model.
A final test, varying the cross validation method (amount of folds) and hold out test set confirm that the reported scores do not vary significantly.

This therefore substantiates the presented performance scores in table 3.2. Having tested many differ-

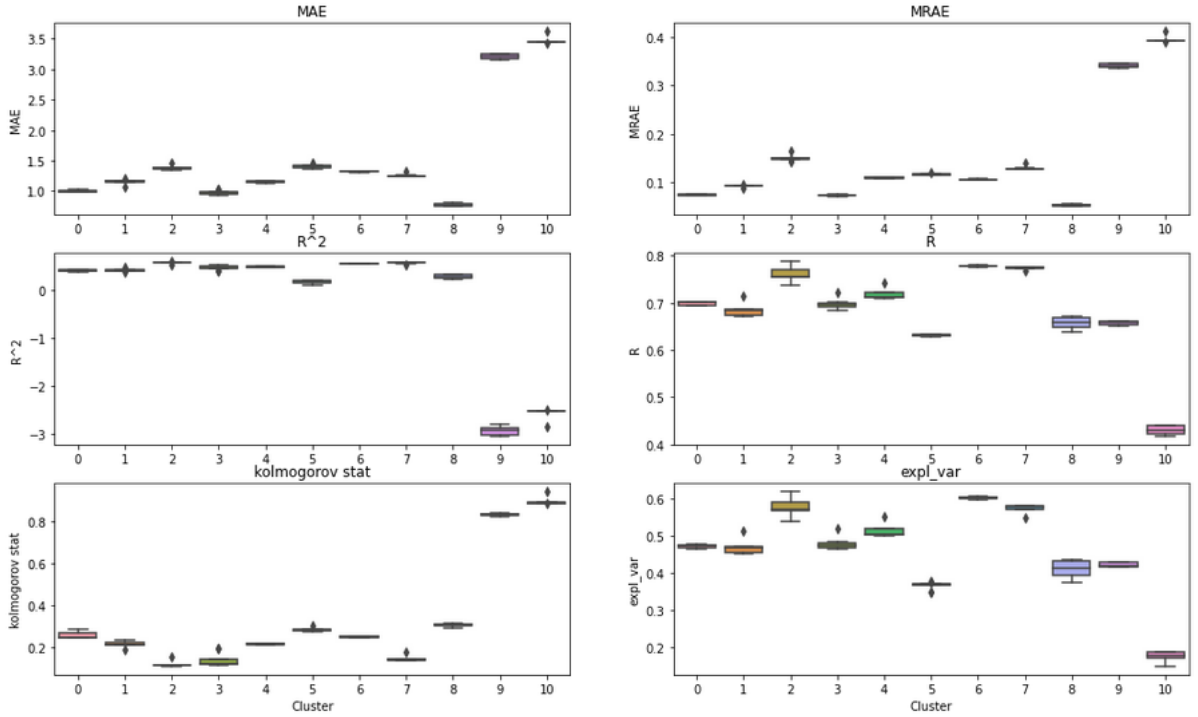


Figure 3.4: The test set performance metrics for different clusters visualised. The model performs comparatively worse on spatial clusters nine and ten, for the other clusters the performance seems to be relatively stable.

ent model specifications, using different designs all result in similar performance scores.
The conclusion therefore is that the design choices during model development are correct, and that other designs would not lead to significantly better results. A key issue to work with is that the model only performs in the log transformed space. Even back transformed target regression did not yield in different performance results.

3.3. Inspection of selected features

As discussed, the process of recursive feature elimination results in six features relevant for the model. As explained in section 3.1 these are the following features: nightly radiance, GDP, PM25, population, pedestrian and intersection density. These where not all covariates fed to the model. As each of them also have a lagged specification. However, in the various runs the spatially lagged variables are never included. A plausible reason for this is that the spatial resolution is to coarse for this variable to have sufficient explanatory power. Another reason for this could be because of misspecification of this variable. However, as discussed in the section 2.1.2 other specifications of ”neighbouring” cells have been tested, which did not result in significant changes. Another suspicion on why spatially lagged variables are never included is that lagged estimates at the edges of cities does not exist. It would be straightforward to test this, extending the rectangle of spatial coverage by 1 km on each side. However, due to the fact that many spatial clusters all-ready border each other, there would not be a substantial amount that is helped by this.

So, the spatially lagged variables are never included, other covariates not included are the power-plant data. This is likely caused as to few cells actually have power-plant data, making no significant changes in the model performance.
Interestingly the intersection density for both pedestrian roads and car roads are more relevant to the model then the road centroids. The intersection density co-variate thus conveys more information regarding emissions to the model. In other studies, such as Creutzig et al. (2016) the operationalisation for the presence of roads is also the intersection density.
It is likely that at intersections cars and other traffic spend a longer time with lower speeds, or even idling. And that therefore the emissions around intersections are higher.

3.3.1. Feature importance & partial dependence

Having covered the selected, and covariates not picked up by the model, it is important to discuss their relative importance. This helps understanding how the model makes an estimate, and hence which data is interesting in later models.
To do so permutation feature importance is calculated for the features in the Random Forest. The result of this analysis is visualised in figure 3.5. The new model mainly hinges on nightly radiance. A second co-variate important for the model is the intersection density. The other features seem to have comparatively less effect, and have roughly the same value for the mean accuracy decrease.

To asses the relationship between the model estimate with the covariates a partial dependence plot is used, which is depicted in figure 3.6. The blue line depicts the partial dependence over increasing values for the covariates. To also provide insight in the distribution of said co-variate this is also plotted in the form of a histogram. All variables except PM25 have a clear positive increasing relationship. These positive relationships increase rapidly for the lower values of the input features, and often quickly converge. An exception is nightly radiance that keeps increasing. The only variable not strictly showing a positive relationship is PM25, which has a dip in the middle.
Interestingly most variables, except for PM25 have a very skewed distribution, exactly where the model assumes the strongest relationship between feature and model output. Referring back to the exploratory data analysis, it seems that the shape of the partial dependence plot line quickly increases and then converges. Which is also observed in the input data in figure 3.1
In conclusion, especially for the lower end of co-variate data the model makes hard distinctions.
An important caveat to using partial dependence plots, is that it assumes that the input variables are independently distributed. This is a troublesome assumption to make as it is likely that these variables are not independent. Isolating the effects of one variable like so might mask some of the relationships in the model, as certain values might occur in high conjunction with each other.

So in conclusion, when looking at the covariates nightly radiance and intersection density have the highest relative importance in the model. The other covariates have similar impacts. Almost all increase show a positive relationship between input magnitude and model output.

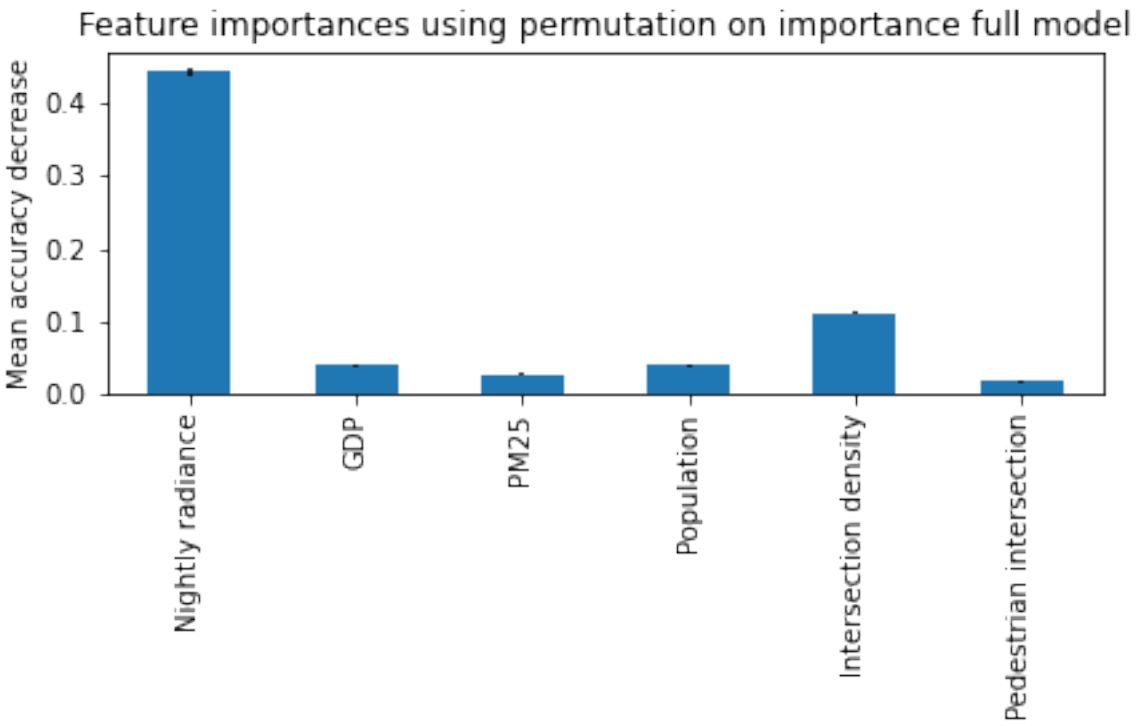


Figure 3.5: Feature importance of the Random Forest model, the model mainly relies on nightly radiance and road intersection data

3.4. Model inaccuracies

Having interpreted and analysed the difference between observed emission values and estimated values by the model, these estimates can be placed in a spatial context and compared against the observed values. This is depicted in figure 3.7. As can be seen the model mainly misses the extreme values, predicting average emissions for cells with very low values. And roads, which are clearly traceable in the original observed emission plots are lost.
To get a better understanding of where the model makes the most errors a second plot is presented, see figure 3.8.
Looking at the absolute residual plot, the original road pattern is still visible and is underestimated generally speaking. The residuals appear to be smaller within the city centres. Looking at the relative residual plot (on the right), a different picture emerges. The roads are still visible, however this is not where the model makes the most percentage wise mistakes. It is evident that the highest percentage wise errors are at the edges of cities, with compared to the other observations lower emissions.
The fact that there seems to be some spatial dependency in the residuals is a sign that the selected covariates fail to explain spatial variance of the response variable.

To further explore the properties of the cells with high relative errors the summary statistics of the spatial covariates are retrieved for observations which have an relative absolute error bigger then the 95% of all the values.
To test and to explore the properties of the cells with a high absolute relative error, the cells that fall in the upper 95 percent quantile are selected. The relationship between estimated and observed, as well as the absolute relative error is visualised in figure 3.9. The selected cells in the upper quantile are marked in red. The first table presented see table 3.5 presents the summary statistics for the hold out set. The second table, 3.6 shows the summary statistics for the 5% of cells with the highest relative absolute error. All of the spatial co-variates are lower. The mean and deviation of emissions however is higher. Interestingly, the values of the first and third quarter are considerably lower than the average. This means that likely there is a single point source predicted wrongly with relatively low values for the

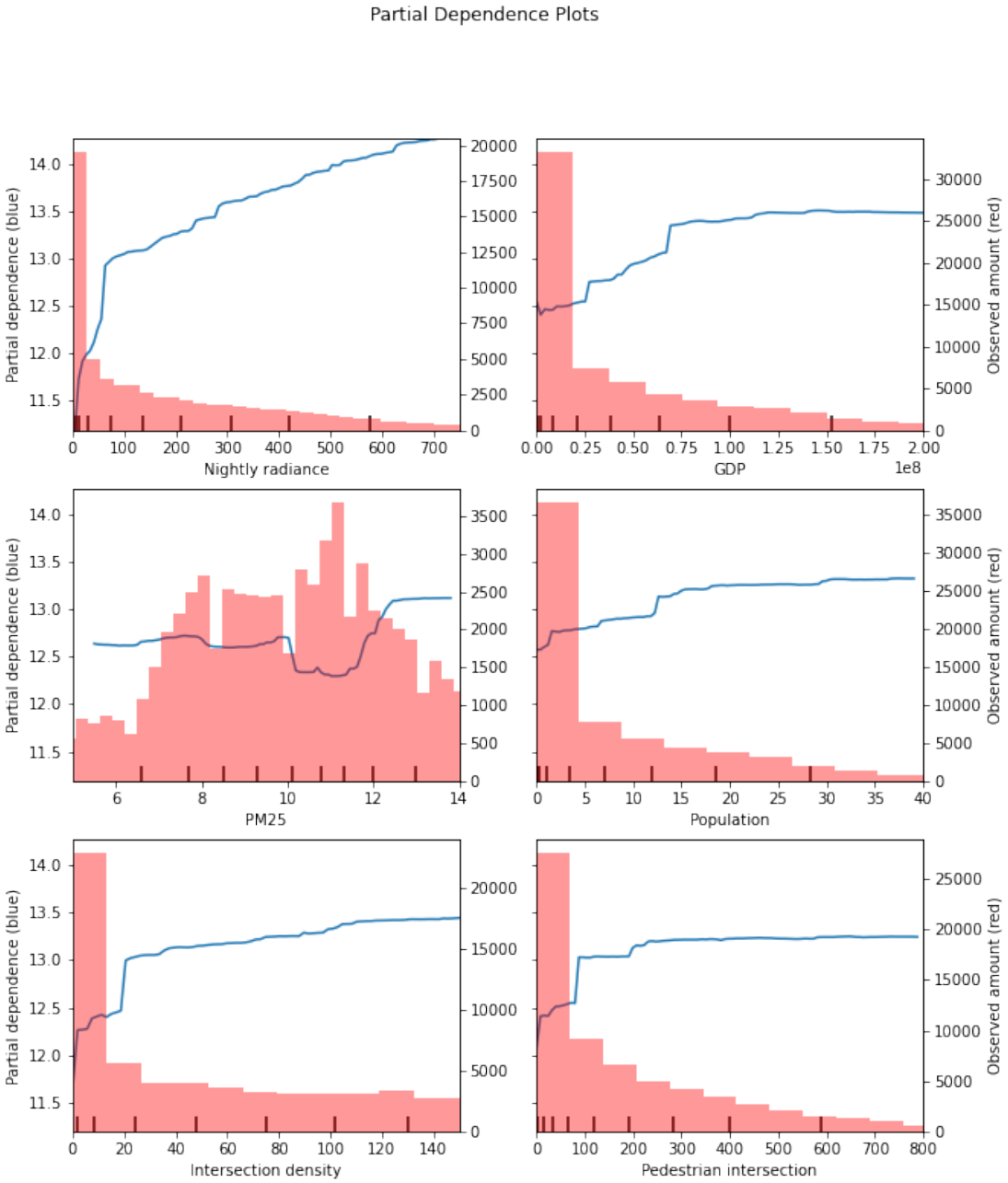


Figure 3.6: Partial dependence between estimate and covariates, most have an increasing relationship

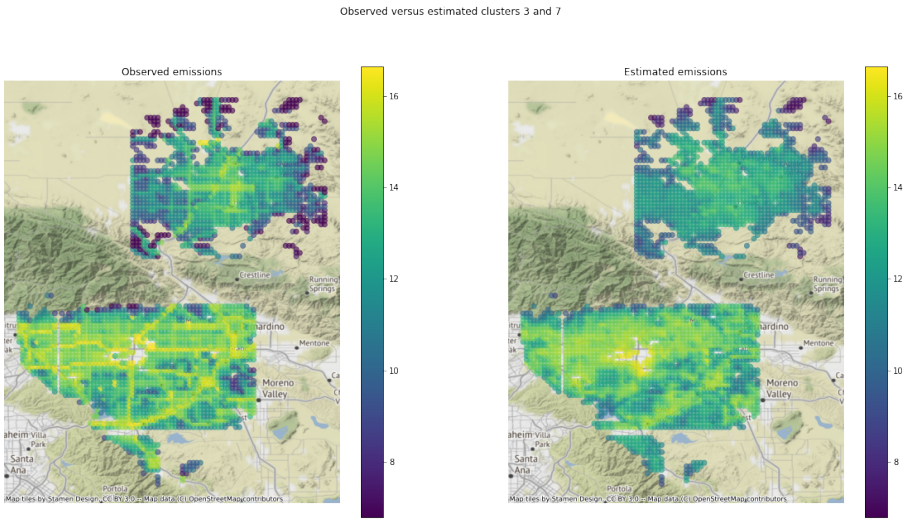


Figure 3.7: Plot depicting the observed emission values versus estimates made by the model. While for the top area the distribution seems to be lost, for the bottom the model is able to follow trends in emissions better

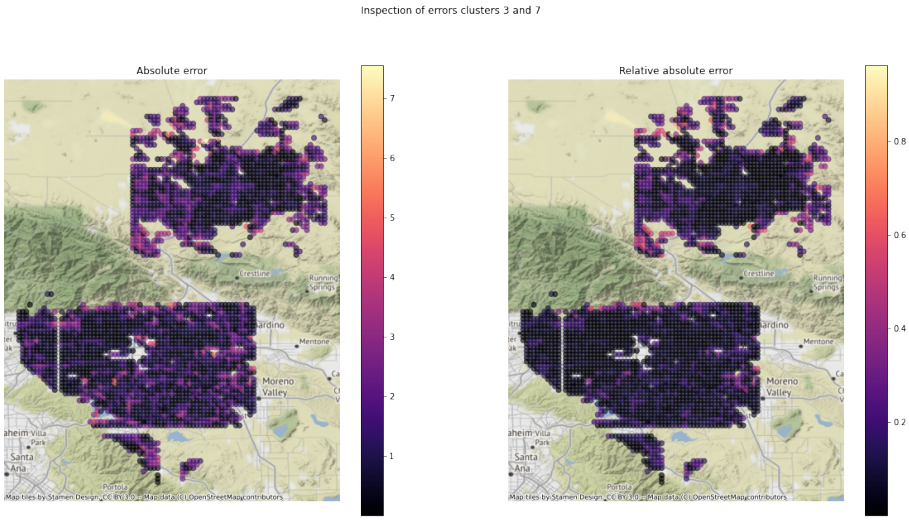


Figure 3.8: Plot showing the absolute error on the left and the relative error on the right. While the highest absolute error is spread more randomly through the city, the highest relative errors are mainly situated around the edges of the city, for lower emission cells.

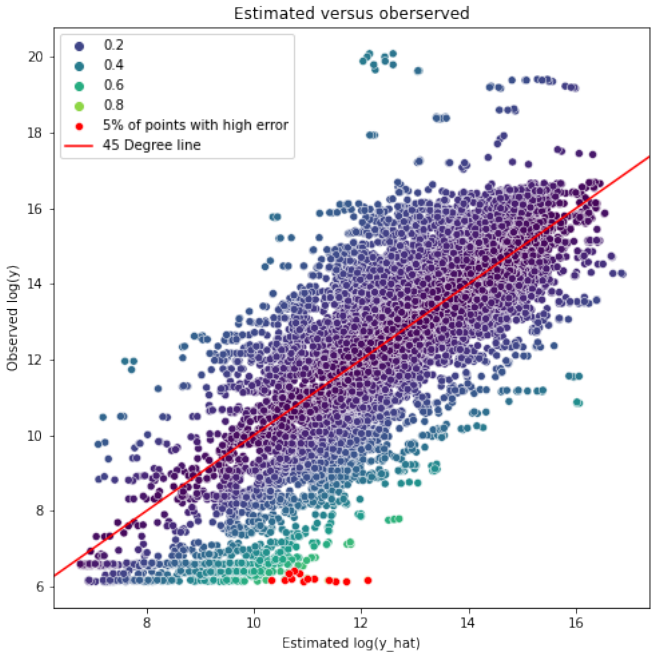


Figure 3.9: Scatter-plot showing the relationship between observed and estimated in log space, the absolute relative error is represented through the colours of the points, the highest absolute errors are situated at cells with low observed values.

N= 13398	Nightly Radiance [$nW/cm^2/sr$]	GDP [$]	Pm25 [μ/m^3]	Population [#]	Intersection density [#]	Pedestrian intersection [#]	Emissions [kgC]
mean	193	4.03E+07	11.2	7.1	53	189	2.47E+06
std	235	4.77E+07	2.9	9.2	56	247	1.57E+07
25%	20	1.67E+06	8.6	0.1	4	36	8.73E+04
50%	108	1.96E+07	11.1	2.6	34	98	5.76E+05
75%	282	6.62E+07	13.5	11.1	88	242	2.10E+06

Table 3.5: Summary statistics for the predictors and emission observation of test set cells

covariates.
These summary statistics help provide evidence that the model performs poorly in cells on the edge of a city. Plotting the spatial locations of these points in figure 3.10 show that indeed, these points are situated near the boundary.
Concluding, the model produces relatively the biggest errors on the edges of cities. For cells with comparatively less emissions the model estimations are less accurate. This is likely caused by the fact that there are not enough of these cells in the training data set with low emissions. Especially point sources in areas with low values for the spatial covariates could lead to relatively big absolute percentage errors.

N=670	Nightly Radiance [$nW/cm^2/sr$]	GDP [$]	Pm25 [μ/m^3]	Population [#]	Intersection density [#]	Pedestrian intersection [#]	Emissions [kgC]
mean	53	2.44E+06	9.5	0.3	6	28	7.92E+06
std	138	7.44E+06	2.4	1.1	17	54	5.65E+07
25%	4	4.67E+04	7.6	0.0	0	2	6.06E+02
50%	5	1.76E+05	9.1	0.0	0	10	1.05E+03
75%	25	7.43E+05	10.6	0.1	2	32	5.84E+03

Table 3.6: Summary statistics for the 5% of test cells with the highest relative absolute error, most notably each of the predictors have considerably lower means and quartile values. The first and second quartile are also lower, the mean of emissions however is very high compared to full test metrics.

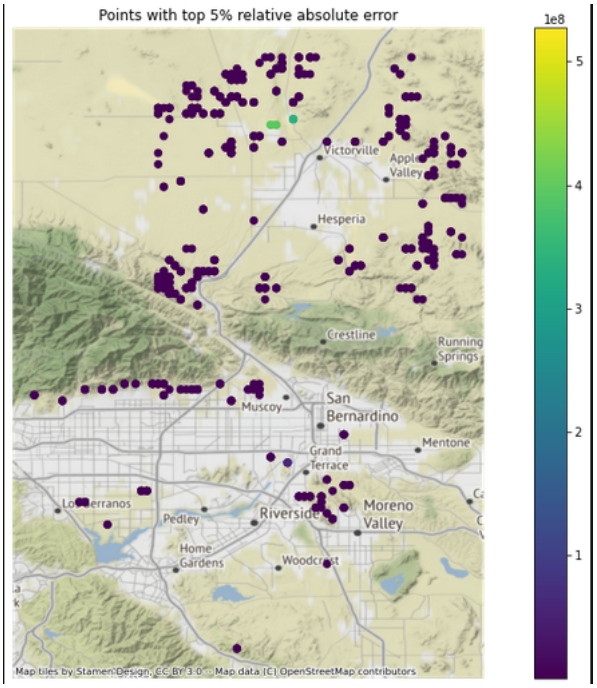


Figure 3.10: Plot showing the spatial location of the top 5% of points with the highest relative absolute error, most points are situated at the edge of cities, the gravest errors are situated within an industrial zone.

3.5. Spatial trends and distribution

To shed light on the research aim what are trends in the spatial distribution of emissions in cities, the generalised model will be used. As the model uncertainties when generalising to new cities are known and discussed in section 3.2.1. To analyse the spatial distribution of emissions in a city the inter quartile range coefficient (IQR coefficient) is used. As this indicator for statistical dispersion was found to be estimated most accurately with relatively the smallest errors, see appendix C.3.1 for the full analysis of the coefficient.
An additional robustness test is done, to understand the dependence and relative error of the estimate of the IQR coefficient under different spatial clusters as test set. Full results can be found in the appendix C.3.2, the relative absolute error is small and on average 13 percent over the different clusters. Except for one spatial cluster the results are stable across. And found to be particularly more stable then for example a Gini-coefficient. Likely as this coefficient is more affected by extreme values, which the model fails to predict, see the tables in appendix C.2.

The IQR is interpreted as follows, a higher IQR coefficient means a less equal distribution of emissions in a city. From analysis, described in appendix D.9 it was found that cities with a high IQR coefficient have more variability and a few clusters with very high emissions. Where cities with a lower IQR have a more even spread.

The IQR coefficient is computed for eleven cities in total. Three of these cities are Indianapolis, Los Angeles and Baltimore, from the Hestia data. For these cities hence the original Hestia data is used. The other eight analysed cities are eight of the most populous cities in the USA, minus Los Angeles (as covered in the Hestia data) and Houston as the spatial extend of this city is to large in the current implementation. To arrive at these estimates a ”final” model is trained using the full data-set and found features and hyper-parameters. For the eight new cities, the spatial covariates are retrieved, and the model is used to create an estimate. As the estimate is still in log space, the estimate is back transformed and the IQR coefficient is calculated over the analysis years. Back transformation is possible as it was found that this does not increase relative error, see appendix C.3.1

For full results analysed and split per city please refer to the appendix D. The results presented in this

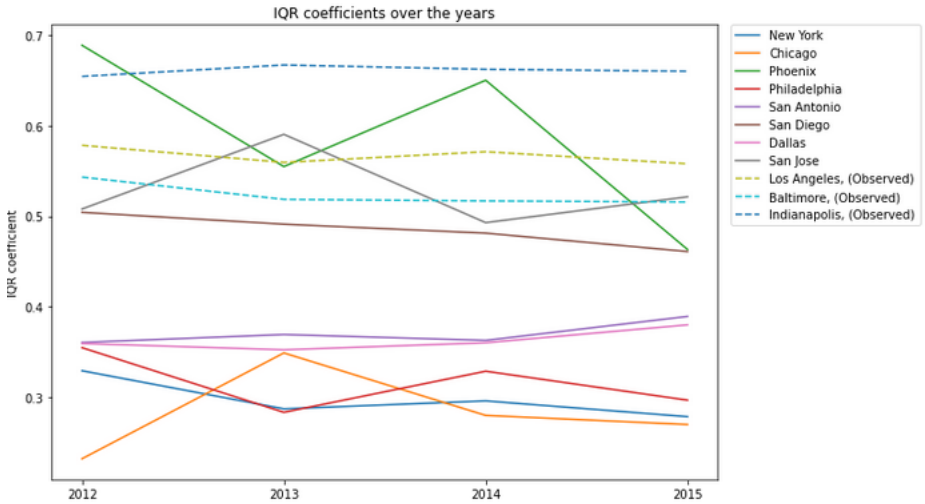


Figure 3.11: The IQR coefficient over the analysis years, the Observed IQR coefficients, from the Hestia set, are also included, these seem to be higher than most estimates. For most estimates the IQR seems to be on a weakly decreasing trend, except for San Antonio and Dallas.

section cover only the aggregate results. These are visualised in figure 3.11.
For the observed data of Indianapolis, Los Angeles and Baltimore the IQR coefficient does not show a strong trend, but shows a decrease of about 3 percent, this might be a consequence of uncertainty in the Hestia data set it self. The IQR coefficient is the highest for Indianapolis at $\sim$0.66. And the lowest for Baltimore at $\sim$0.51.
For the newly analysed cities three have similarly high IQR coefficients: Phoenix, San Jose and San Diego. For these cities the IQR coefficient also seems to be in a decreasing trend. The other analysed cities: San Antonio, Dallas, Philadelphia, Chicago and New York. All have a considerably lower IQR coefficient. For these cities the distribution and variability of emissions is lower and hence the model estimates that it is more equally divided here. While model results and errors could lead to errors in the approximation of the IQR it seems that the trend is moving down ward. Except for San Antonio and Dallas. Given the model performance this trend is not clearly discern able, and likely more an effect of random variation.
Comparing cities with a higher IQR coefficients to the ones with lower IQR coefficients, the cities with higher IQR coefficients often have a cluster of high emission points located somewhere within the bounds of the city. Cities with lower IQR coefficients however, have a more equal distribution and a cluster of high emissions are often more central in the city. This is illustrated in figure 3.12
The analysed cities located around the coast often have higher emission values along the coast that decrease while going inland. Other cities, located inland typically show high emissions in the centre, with decreasing intensity towards the edges. Some cities, for example Phoenix and San Jose, have breaks in the high intensity emission cells in the centre. This generally increases the IQR. These relatively lower intensity cells are often hills and other breaks in the urban fabric. This indicates that possibly adding for example the average elevation of a grid cell could potentially explain some model variance. Similarly, perhaps land use classification could have explanatory power.

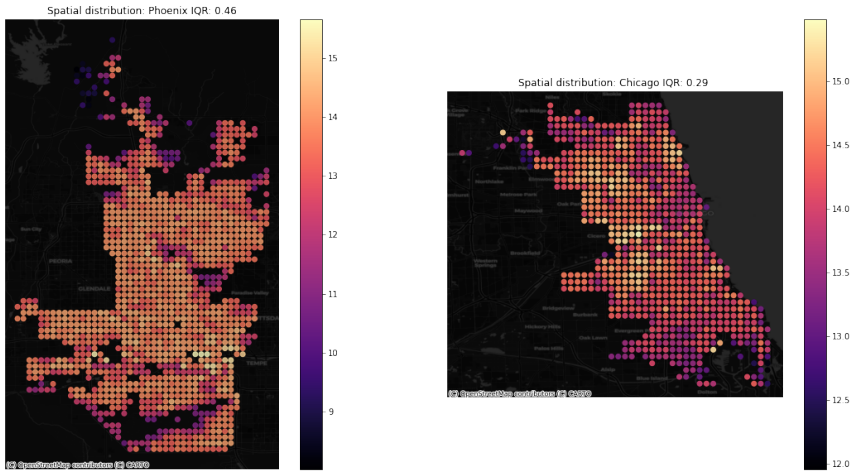


Figure 3.12: Spatial distribution of emissions of Phoenix, the city with the highest IQR coefficient, compared to Chicago, the city with the lowest IQR coefficient. high emission cells are more clustered and in the centre. For Phoenix the high emissions cells are less clustered, resulting in a higher IQR coefficent.

4. Discussion & Conclusion

The main aim of this study is to answer the following research question *Using a new hybrid approach, how are emissions spatially distributed within eight of the most populous cities in the US?*. To do so first this study covered the development and design of the new approach. Which uses a combination of bottom-up emission inventory data with remotely sensed covariates through a Random Forest model. The next step is using the newly made model estimates. Calculating the inter quartile range coefficient for a number of cities and years. However, the main focus is inspection whether the newly produced model using a Random Forest fits and how this method could be further developed.

To provide an answer to said research question the following four sub-questions have been answered.

- **SQ A:***Which spatial covariates and bottom-up data products to use for the new hybrid model?*
- **SQ B:***Using the hybrid approach, how well does the model spatially generalise on a test set, how is this affected by various model designs?*
- **SQ C:***Where does the model have the five percent highest residuals error and what are the properties of these cells?*
- **SQ D:***What are the trends in the spatial distribution of emissions in cities?*

Findings of this study suggest that the method to optimise the model is found to be stable for different geographical locations. The method is scalable and can be used and extended as more data becomes available. Multiple performance and robustness tests have been performed, and no differences in model performance have been found. Suggesting that given the current data no further improvement could be achieved. An important caveat of the model is that the estimates are still in log transformed form. During training it was found that the model has better generalisability performance to other locations.
Inspection of the relevant features suggest that for the model the following features are relevant and in order of importance: Nightly radiance, intersection density, population, GDP, PM25 and Pedestrian intersections. Inspection of the partial dependence shows that each of these variables, except PM25, have a positive increasing relationship with the models emission estimate. This relationship quickly converges for higher values. The highest relative errors mainly occur around the peripheries of cities, for cells with low emission values. The relationship between the covariates and the model estimate is not linear, but more complex.
Knowing these caveats, the generalised model is used to estimate emissions in the ten most populous cities in the USA. No indication of a yearly trends of spatial variability is found. The model however grants new insight in the spatial distribution of emissions in these cities.

Answering sub-question A is done by performing a literature review. As discussed the new model relies on a previously made bottom-up emission inventory. Apart from various self-reported emission inventories, produced by cites, bottom-up estimates with a resolution around 1×1 km are sparse. Hestia is one of the only bottom-up inventories that covers urban centres (Gurney et al., 2019b, 2012; Zhou and Gurney, 2010), therefore this data product is chosen. For the spatial covariates the following where selected for the model learning process: Nightly radiance (Oda et al., 2018), GDP (Chen et al., 2019), PM25, population (Ou et al., 2015; Asefi-Najafabady et al., 2014), power-plant point source data (Zhou and Gurney, 2010), road centroids & intersection density and the same for pedestrian roads (Newman and Kenworthy, 2015; Ewing and Cervero, 2010). Each of these variables, except power-plant point source data also gets a lagged variable, as for example a road would not occur only in a single cell, but also in the cells around it. The idea is that spatially lagging would provide extra information to the machine learning model, such as whether a cell is situated in an urban centre or at the periphery. Previous studies mainly use nightly radiance and population as spatial proxy. However, GDP is often associated with higher fuel consumption as well. PM25 is also not commonly included, however an important by product of fossil fuel combustion. Roads, and in particular intersections and pedestrian roads have a known connection in the literature with emissions, however these are rarely included in

top-down emission inventories.
The literature calls for the inclusion of point source data, and road and transport models. However as will be discussed for sub-question C, point source data provides no significant improvement to the hybrid model.
To answer sub-question B, the generalisation performance, especially in the original space is marginal with R^2 scores around zero . In agreement with earlier findings, log transformation of the dependent variable however helps model generalisation (Stevens et al., 2015), which boosts the R^2 towards $\sim$0.5. This has the downside of losing interpretability. While in the absolute sense (how much kilograms of emissions in a city or grid cell) the model is inaccurate, the distribution curve of emissions is preserved. The model generalises with similar confidence over the analysis years. Experiments with leaving location and or time out cross validation, confirmed earlier findings of Meyer et al. (2018). Not properly dis-aggregating regions results in a more positive view of cross validation scores. When evaluating the observed versus estimated value of emitted carbon, it is therefore important to practice leave location out cross validation . In this analysis, the performance metrics remain stable over the included years. These findings hence suggest that calculating the performance metrics per year is less important, when dis-aggregating the regions in a grouped form.
Tests with different spatial clusters where ran, to test whether the performance metrics presented in this study are dependant on the observations in the training an test set. This is done to find out whether the presented pipeline in this study is robust, and leads to stable results. This pipeline consist of first running recursive feature elimination and then hyper-parameter tuning.
These test point out that both the tuned hyper-parameters and selected features remain stable for different data observations in the training set. On the performance metrics however, certain spatial locations result in high model estimate errors independent of model specification, illustrating the trouble of spatial generalisation. The fact that most data points for the trained model come from the Los Angeles Basin and that Baltimore has the highest standard errors suggests that the model did not have enough observations for different spatial areas. This means that a model like this can not be evaluated using a single test set, one should use (grouped) cross validation scores to report and asses model performance to maximise plausible generalisation.
Various machine learning models are tested, ensemble methods outperform other regression techniques, such as various linear regression forms: Lasso, Elastic net and OLS, as well as Support Vector regression. Random Forest was used in this study, but the findings also suggest that XGBoost would be a worthwhile path to pursue albeit more computationally expense.
While not pursued in the current study, it would be interesting to validate and compare the model estimates against other existing data products. As conducted in (Hutchins et al., 2017; Chen et al., 2020; Gurney et al., 2019a) especially indicating and comparing the estimates in urban cores. which the model is aimed to improve.

To answer sub-question C the descriptive statistics of the cells with the highest relative absolute error are inspected and compared to the descriptive statistics of the original set. It becomes evident that most errors occur in grid cells with lower emissions and lower co-variate values. Exploring the spatial distribution of these points, these are often located around the edges of cities. This suggests that the model might not have had enough examples or observations of such low value points.
Interesting is that the uncertainty around city boundaries is not uncommon in other remotely sensed data products. When comparing emission inventories the biggest relative errors occur for cells with low emission values (Hutchins et al., 2017). This error could be attributed to using night time lights as a spatial proxy as studies find higher uncertainty around city boundaries and areas with lower night time light levels Chen et al. (2020); Asefi-Najafabady et al. (2014); Gurney et al. (2019a).

When inspecting the model feature importance it is found that the model mainly relies on night time lights and pedestrian intersections. Some studies argue a need for inclusion of point source data to improve model estimates (Oda et al., 2018; Oda and Maksyutov, 2011; Gurney et al., 2019a). In the current study however the feature selection algorithm consistently excludes the WRI power plant data. This might be a result of relatively few power plants in comparison to other cells, hence the variable achieves a relatively low importance. A performance increase could potentially be achieved by building an ensemble model with a specific power-plant emission model, something as developed by AlKheder and Almusalam (2022). In a sense this is troublesome as often the majority of a countries CO_2 emissions

stem from electricity generation (Zhang et al., 2013). For a developed country for the USA however, it is possible that in the context of urban in boundary emissions, point sources play a less big role then initially thought. As the majority of electricity generation likely happens outside of the urban zone.
The inclusion of road networks, in particular that of intersections and pedestrian roads, validates earlier findings how urban form and street network design relates to emissions (Ewing and Cervero, 2010; Liu et al., 2017). The current model however, has a positive association that between intersections, pedestrian intersections and emissions. The difference is that for pedestrian intersections the relationship becomes less positive for very high values. For car intersection density the relationship converges less. This positive increasing relations is likely caused by the fact that there are more intersections in high density, high emission urban cores. The strongest relationship is found for lower values of both these covariates, which is where the most data is available. It would therefore be of value to zoom into more high density urban cores, to further understand this relationship.
Another surprising finding regarding the covariates is that the air quality has a relatively small effect compared to the other covariates. As suggested by Anenberg et al. (2019) perhaps the issue of fine particulate matter are mitigated easier using filters. Which hence likely results in a weaker association between this co-variate and emissions.

In addition to model developing a model and providing ways of model evaluation. The model has been generalised to eight of the most populous cities in the USA. While in the absolute sense the accuracy of model results are rather limited, the distribution of emissions was found to be stable. Therefore, using this property, an inter quartile range coefficient is calculated for these cities over the years 2012 until 2015.
This gives insight in the spatial dispersion of emissions in these cities. Over the years no obvious trends in the inter-quartile range coefficient are found. Findings however suggest that some cities have considerably lower inter quartile range coefficients. Of the analysed cities the city of Phoenix has the highest IQR and Chicago the lowest. Visual inspection suggest that cities with multiple clusters of high emissions cells have a higher IQR coefficient. Breaks in the urban fabric, in terms of hills parks or other natural barriers often are a contributor to this.
Combining this insight with the total emissions within a city would be an interesting pursuit. As Creutzig et al. (2016) found that urban form can significantly affect the CO_2 emissions within a city. Further research in how the inter quartile range is affected by urban form and what would be optimal is necessary.

The main aim of this research is: *How are emissions spatially distributed within eight of the most populous cities in the US, and what are the trends?* This is achieved by utilising a Random Forest model to estimate emissions.
In the analysed years no convincing trends regarding the inter quartile range coefficient where found. However, using the developed model it is possible to create new insights, regarding the spatial distribution of emissions. Some have a more obvious centre cluster and some have emissions scattered around.
Creation a model using bottom up mappings results in marginal generalisation, as both emissions and spatial covariates are highly heterogeneous. Log transformation of the emissions dependent variable helps model generalisation. Basing an urban emission model using data from a bottom-up approach could be a solution to covering other areas not covered by a bottom up mapping. For this to work however, more data in a variety of regions is necessary. Using the developed model it is possible to create new insights, regarding the spatial distribution of emissions. The finding show that the spatial distribution of emissions differ in the analysed cities. Some have a more obvious centre cluster and some have emissions scattered around.

4.1. Study limitations & recommendations

An important limitation of following the supervised emission model approach, is that the model can never be more accurate than the data used to train the model. In this research the Hestia data set is taken as ground truth, this data-product however, is also an approximation of reality.More-over, the Hestia dataset is limited to three cities situated in North America. And given the need for spatial and temporal overlap in the co-variate data. The spatial and temporal coverage are therefore limited. As a

consequence the spatiotemporal generalisability of the current model is marginal. To be able to further extend the findings of such models, it would be valuable to also produce bottom-up estimates of other cities, in different contexts. For example in European cities, which are often less car centric compared to the US, or urban areas in developing countries.
The experiments with different hold out sets suggest that for certain spatial clusters the estimation is off quite a bit. The fact that this cluster is Baltimore suggests that the model has had most training examples of the Los Angeles Basin. This design choice results in more spatial points included but with increased chance of overfitting for a certain area.

In terms of yearly generalisability the model has not been explicitly tested to generalise in terms of both years and location. The study of Meyer et al. (2018) suggest that spatial random forest models could be evaluated by using either leave location out, which is applied in this thesis, or leave time out. If not done correctly it could give an overly positive view of model results. While the error across years is similar, a combination of both leave location and time out cross validation could yield different results.
Regarding the temporal generalisability, under the current model implementation the road network is fixed for the analysis years. Perhaps as roads are an important predictor, this attributes to the consistent errors across the analysis years. It would be valuable to include traffic or even congestion data from navigation systems such as suggested by Nangini et al. (2019).
As especially the peripheries of cities result in higher errors it would be interesting to also include extra spatial covariates. Perhaps a digital terrain model, with data about raster cell elevation or incline would provide insight in the use of lands. In the same topic, using a land use classification map could also be valuable information for the model.

It is found that log transforming helped the target variable, in the sense that the statistics around explained variance, spatial correlation and the distribution statistics became better. A limitation regarding this is losing interpretability, especially as back transforming resulted in the a performance decrease of these results. Also, as the target variable is log transformed, it is possible that the performance statistics give a more positive view then the case in reality. A back transformed target regression approach was considered and tried, but did not yield in substantially changing results. Time limitations did result in not pursuing this path further, while it might very well be a favourable path as the model is then evaluated in the original space.

Given the current model performance the IQR coefficient is typically off by about thirteen percent. One of the analysis cities however, Baltimore, has a considerably higher inaccuracy at around 50 percent. This creates uncertainty in the IQR coefficient, where for the newly analysed cities it is plausible that the estimate is quite far off. Under current research design however it is not possible to test this. It is therefore uncertain whether the found trends are applicable and are just random variation, hence the conclusion that there are no obvious trends in the IQR coefficients.

Given the limited spatial extent of this study. An interesting path to follow would be to extent the spatial coverage and study the IQR coefficient for more cities worldwide under different environmental climates. This could lead insight to the question of emissions would ideally or fairly be distributed. This could lead to a typology of emission distributions, and as more yearly data becomes available, perhaps more clear trends could be discerned.
To increase model performance the question of how to include point source data remains. But perhaps an ensemble model, where at least one model is specifically aimed at estimating power plant emissions could improve model accuracy.
To further strengthen findings and usability of the current model the inclusion of carbon sinks within the urban area. And how sinks would affects the total amount of emissions and distribution would also be an interesting path to pursue.
Given the new model it would be interesting to validate the model against ground samples and perform an in-depth case study of a certain city, assessing the accuracy of the model (also relative to the other existing data products) and understanding how neighbourhoods compare.
To create more insight in model performance a comparison study, using the other prominent emission data products is necessary.

4.2. Study contribution

Concluding, while model performance is marginal this study brings a scalable approach that can be used and extended as more data becomes available. Findings also suggest that the Random Forest dasymetric approach is the most likely candidate to produce accurate insights, with the advantage of being efficient in terms of computational resources. It also substantiated earlier findings, that leave location out cross validation is important for model evaluation. Additionally this study substantiates findings that cells with relatively lower emissions in the urban area are particularly hard to estimate.
The study in this context is a pilot study, with the aim of solving the problem of urban emission distributions. The pipeline to develop a model, which consist of recursive feature elimination and hyperparameter tuning is found to be stable and results in better explanatory power. During these steps it is important to use grouped cross validation, as this yields a realistic view of model performance.
The relationship between observed emissions and the covariates used in this study is not linear and more complex, to build a spatial emission model using a supervised approach, an ensemble model therefore outperforms other methods. Findings in the current study suggest that point source emissions are not all that important in the context of urban emissions. The issue brought forward by many studies of estimation of emissions around the peripheries is not solved through this approach.
In terms of insights regarding the emission distributions, no trends in the inter quartile range coefficient are found. But a first insight of this coefficient is given in eight of the most populous cities in the USA. Indicating that cities with a higher inter quartile range coefficient have multiple clusters of high emission cells, spread out within the city boundary.
These findings, that the spatial distribution of emissions in cities world wide is largely unknown, is an important implication for policy makers. More knowledge regarding this is necessary. As this helps evaluate how differences in the spatial distribution affects the total emission sum of a city and how this could be influenced.

Bibliography

AlKheder, S. and Almusalam, A. (2022). Forecasting of carbon dioxide emissions from power plants in kuwait using united states environmental protection agency, intergovernmental panel on climate change, and machine learning methods. *Renewable Energy*, 191:819–827.

Anenberg, S. C., Achakulwisut, P., Brauer, M., Moran, D., Apte, J. S., and Henze, D. K. (2019). Particulate matter-attributable mortality and relationships with carbon dioxide in 250 urban areas worldwide. *Scientific reports*, 9(1):1–6.

Asefi-Najafabady, S., Rayner, P., Gurney, K., McRobert, A., Song, Y., Coltin, K., Huang, J., Elvidge, C., and Baugh, K. (2014). A multiyear, global gridded fossil fuel CO2 emission data product: Evaluation and analysis of results. *Journal of Geophysical Research: Atmospheres*, 119(17):10–213.

Berezin, E., Konovalov, I., Ciais, P., Richter, A., Tao, S., Janssens-Maenhout, G., Beekmann, M., and Schulze, E.-D. (2013). Multiannual changes of co 2 emissions in china: indirect estimates derived from satellite measurements of tropospheric no 2 columns. *Atmospheric Chemistry and Physics*, 13(18):9415–9438.

Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

British Petroleum (2021). Methodology for calculating CO2 emissions from energy use. `https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2021-carbon-emissions-methodology.pdf`. Online; accessed on 3-3-2022.

Burkov, A. (2019). *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada.

Byers, L., Friederich, J., Henning, R., Kressig, A., Li, X., McCormick, C., and Malaguzzi Valeri, L. (2021). A global database of power plants. `https://files.wri.org/d8/s3fs-public/2021-07/global-power-plant-database-technical-note-v1.3.pdf?VersionId=KNA6zn0E2HgUcEsXhtZuvfAlIqWOjLib`. Online; accessed on 21-2-2022.

Chen, B. et al. (2007). *An empirical comparison of methods for temporal distribution and interpolation at the national accounts*. BEA.

Chen, G., Shan, Y., Hu, Y., Tong, K., Wiedmann, T., Ramaswami, A., Guan, D., Shi, L., and Wang, Y. (2019). Review on city-level carbon accounting. *Environmental science & technology*, 53(10):5545–5558.

Chen, J., Zhao, F., Zeng, N., and Oda, T. (2020). Comparing a global high-resolution downscaled fossil fuel CO2 emission dataset to local inventory-based estimates over 14 global cities. *Carbon Balance and Management*, 15(1):1–15.

Christiansen, L. B., Cerin, E., Badland, H., Kerr, J., Davey, R., Troelsen, J., Van Dyck, D., Mitáš, J., Schofield, G., Sugiyama, T., et al. (2016). International comparisons of the associations between objective measures of the built environment and transport-related walking and cycling: Ipen adult study. *Journal of transport & health*, 3(4):467–478.

Creutzig, F., Agoston, P., Minx, J. C., Canadell, J. G., Andrew, R. M., Quéré, C. L., Peters, G. P., Sharifi, A., Yamagata, Y., and Dhakal, S. (2016). Urban infrastructure choices structure climate solutions. *Nature Climate Change*, 6(12):1054–1056.

Croft, T. A. (1978). Nighttime images of the earth from space. *Scientific American*, 239(1):86–101.

Dou, X., Wang, Y., Ciais, P., Chevallier, F., Davis, S. J., Crippa, M., Janssens-Maenhout, G., Guizzardi, D., Solazzo, E., Yan, F., et al. (2022). Near-real-time global gridded daily co2 emissions. *The Innovation*, 3(1):100182.

Eggleston, H., Buendia, S., Miwa, L., Ngara, K., and Tanabe, K. (2006). 2006 IPCC Guidelines for National Greenhouse Gas Inventories.

Elvidge, C. D., Baugh, K. E., Zhizhin, M., and Hsu, F.-C. (2013). Why viirs data are superior to dmsp for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network*, 35(0):62.

Ewing, R. and Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American planning association*, 76(3):265–294.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Geng, G., Zhang, Q., Martin, R. V., Lin, J., Huo, H., Zheng, B., Wang, S., and He, K. (2017). Impact of spatial proxies on the representation of bottom-up emission inventories: A satellite-based analysis. *Atmospheric Chemistry and Physics*, 17(6):4131–4145.

Gurney, K. R., Liang, J., O'Keeffe, D., Patarasuk, R., Hutchins, M., Huang, J., Rao, P., and Song, Y. (2019a). Comparison of global downscaled versus bottom-up fossil fuel CO2 emissions at the urban scale in four US urban areas. *Journal of Geophysical Research: Atmospheres*, 124(5):2823–2840.

Gurney, K. R., Liang, J., Patarasuk, R., Song, Y., Huang, J., and Roest, G. (2020). The vulcan version 3.0 high-resolution fossil fuel CO2 emissions for the United States. *Journal of Geophysical Research: Atmospheres*, 125(19):e2020JD032974.

Gurney, K. R., Mendoza, D. L., Zhou, Y., Fischer, M. L., Miller, C. C., Geethakumar, S., and de la Rue du Can, S. (2009). High resolution fossil fuel combustion co2 emission fluxes for the united states. *Environmental science & technology*, 43(14):5535–5541.

Gurney, K. R., Patarasuk, R., Liang, J., Song, Y., O'keeffe, D., Rao, P., Whetstone, J. R., Duren, R. M., Eldering, A., and Miller, C. (2019b). The hestia fossil fuel co 2 emissions data product for the los angeles megacity (hestia-la). *Earth System Science Data*, 11(3):1309–1335.

Gurney, K. R., Razlivanov, I., Song, Y., Zhou, Y., Benes, B., and Abdul-Massih, M. (2012). Quantification of fossil fuel CO2 emissions on the building/street scale for a large US city. *Environmental Science & Technology*, 46(21):12194–12202.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

Haberl, H., Wiedenhofer, D., Virág, D., Kalt, G., Plank, B., Brockway, P., Fishman, T., Hausknost, D., Krausmann, F., Leon-Gruchalski, B., et al. (2020). A systematic review of the evidence on decoupling of gdp, resource use and ghg emissions, part ii: synthesizing the insights. *Environmental Research Letters*, 15(6):065003.

Hsu, A., Wang, X., Tan, J., Toh, W., and Goyal, N. (2022). Predicting European cities' climate mitigation performance using machine learning.

Hutchins, M. G., Colby, J. D., Marland, G., and Marland, E. (2017). A comparison of five high-resolution spatially-explicit, fossil-fuel, carbon dioxide emission inventories for the United States. *Mitigation and Adaptation Strategies for Global Change*, 22(6):947–972.

IEA (2021). Greenhouse gas emissions from energy: Overview. https://www.iea.org/reports/greenhouse-gas-emissions-from-energy-overview. Online accessed on; 1-3-2021.

IPCC (2014). Impacts, adaptation, and vulnerability. *Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1132.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., Bergamaschi, P., Pagliari, V., Olivier, J. G., Peters, J. A., et al. (2019). Edgar v4. 3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012. *Earth System Science Data*, 11(3):959–1002.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., Bergamaschi, P., Pagliari, V., Olivier, J. G. J., Peters, J. A. H. W., van Aardenne, J. A., Monni, S., Doering, U., and Petrescu, A. M. R. (2017). EDGAR v4.3.2 Global Atlas of the three major Greenhouse Gas Emissions for the period 1970–2012. *Earth System Science Data Discussions*, 2017:1–55.

Kona, A., Monforti-Ferrario, F., Bertoldi, P., Baldi, M. G., Kakoulaki, G., Vetters, N., Thiel, C., Melica, G., Lo Vullo, E., Sgobbi, A., et al. (2021). Global Covenant of Mayors, a dataset of greenhouse gas emissions for 6200 cities in Europe and the Southern Mediterranean countries. *Earth System Science Data*, 13(7):3551–3564.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Liu, Z., Ma, J., and Chai, Y. (2017). Neighborhood-scale urban form, travel behavior, and co2 emissions in beijing: implications for low-carbon urban planning. *Urban Geography*, 38(3):381–400.

Macknick, J. (2011). Energy and CO2 emission data uncertainties. *Carbon Management*, 2(2):189–205.

Marland, G. and Rotty, R. M. (1984). Carbon dioxide emissions from fossil fuels: a procedure for estimation and results for 1950-1982. *Tellus B: Chemical and Physical Meteorology*, 36(4):232–261.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 101:1–9.

Nangini, C., Peregon, A., Ciais, P., Weddige, U., Vogel, F., Wang, J., Bréon, F.-M., Bachra, S., Wang, Y., Gurney, K., et al. (2019). A global dataset of CO2 emissions and ancillary data related to emissions for 343 cities. *Scientific Data*, 6(1):1–29.

Newman, P. and Kenworthy, J. (2015). The theory of urban fabrics. In *The End of Automobile Dependence*, pages 105–140. Springer.

Nielsen, K. S., Nicholas, K. A., Creutzig, F., Dietz, T., and Stern, P. C. (2021). The role of high-socioeconomic-status people in locking in or rapidly reducing energy-driven greenhouse gas emissions. *Nature Energy*, 6(11):1011–1016.

Oda, T. and Maksyutov, S. (2011). A very high-resolution (1 km$\times$ 1 km) global fossil fuel co 2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*, 11(2):543–556.

Oda, T., Maksyutov, S., and Andres, R. J. (2018). The Open-source Data Inventory for Anthropogenic CO 2, version 2016 (ODIAC2016): a global monthly fossil fuel CO 2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data*, 10(1):87–107.

Ou, J., Liu, X., Li, X., and Shi, X. (2015). Mapping global fossil fuel combustion CO2 emissions at high resolution by integrating nightlight, population density, and traffic network data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(4):1674–1684.

Rayner, P., Raupach, M., Paget, M., Peylin, P., and Koffi, E. (2010). A new global gridded data set of CO2 emissions from fossil fuel combustion: Methodology and evaluation. *Journal of Geophysical Research: Atmospheres*, 115(D19).

Seto, K. C., Churkina, G., Hsu, A., Keller, M., Newman, P. W., Qin, B., and Ramaswami, A. (2021). From low-to net-zero carbon cities: The next global agenda. *Annual Review of Environment and Resources*, 46(1):377–415.

Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., Chen, L., and Wu, J. (2014). Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing*, 6(2):1705–1724.

Sneep, M. (2021). Sentinel 5 precursor/TROPOMI KNMI and SRON level 2 input output data definition. `https://sentinel.esa.int/documents/247904/3119978/Sentinel-5P-Level-2-Input-Output-Data-Definition`. Online; accessed on 23-5-2022.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one*, 10(2):e0107042.

van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., et al. (2021). Monthly global estimates of fine particulate matter and their uncertainty. *Environmental Science & Technology*, 55(22):15287–15300.

Yadav, N., Sorek-Hamer, M., Von Pohle, M., Asanjan, A. A., Sahasrabhojanee, A., Suel, E., Arku, R., Lingenfelter, V., Brauer, M., Ezzati, M., et al. (2022). Deep transfer learning on satellite imagery improves air quality estimates in developing nations. *arXiv preprint arXiv:2202.08890*.

Zhang, M., Liu, X., Wang, W., and Zhou, M. (2013). Decomposition analysis of co2 emissions from electricity generation in china. *Energy policy*, 52:159–165.

Zhou, Y. and Gurney, K. (2010). A new methodology for quantifying on-site residential and commercial fossil fuel co2 emissions at the building spatial scale and hourly time scale. *Carbon Management*, 1(1):45–56.

Zhou, Z.-H. (2021). Ensemble learning. In *Machine learning*, pages 181–210. Springer.

A. Appendix: Spatial Clusters

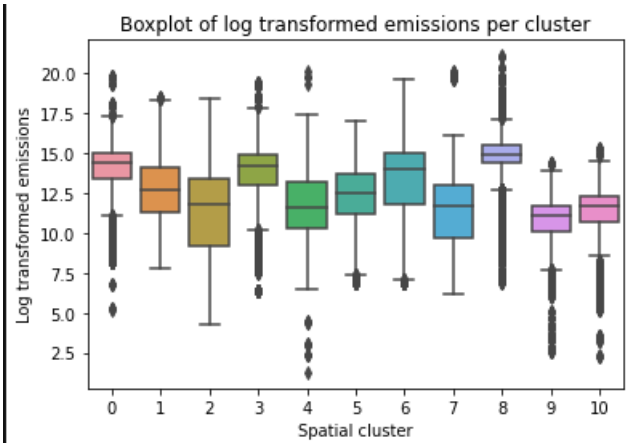


Figure A.1: Boxplot showing the log transformed distributions of emissions in the 11 unique spatial clusters

To test how the classified spatial clusters of the KMeans algorithm compare to each other, two statistical tests are performed. First a Manwhitney U test to understand whether the two groups are different on the emission observation variable. The variable of interest is skewed and continuous. Two groups are created, one is a single spatial cluster and the other the rest of the observations. Secondly a student t-test is used to understand if the means of different clusters are similar.

First a Manwhitney U test is performed. For each test one spatial clusters is compared to the population mean (the population consists of the other observation clusters). When running two sample t-test for each combination of clusters you find that the sample means of clusters 0 and 3, 4 and 2, 4 and 7, 4 and 5, 7 and 1 are statistically similar.
The conclusion of this analysis is that the input example sets are different. This indication is positive, as this means the observations used to train the observations are different, and hence a better generalisation could be achieved.

B. Appendix: Exploratory Data Analysis

After having collected, created and aggregated the data an exploratory data analysis is performed. This gives the reader a feeling of the properties of the data set and how the variables relate.
Important, to note here is that initially there are eleven unique covariates, when also including each lagged variable, this amount to 22 variables. Initially all data is presented. However, only a few variables are in the end included in the model, due the recursive feature elimination process. Thus after having presented the full summary statistics for all included variables, only the for the model relevant variables are described.

B.1. Descriptive statistics for all variables

N = 68139	avg_rad	gdp	pm25	population	capacitymw	gwh_2016	gwh_estimt	road_centroids	intersection_density	pedestrian_roads	pedestrian_intersection	emissions
mean	223.5	5.11E+07	10.2	9.4	1.0	1.3	3.1	25.2	63.8	79.3	223.9	2.50E+06
std	304.5	7.19E+07	2.6	13.7	30.6	51.7	109.2	25.2	64.3	100.2	298.0	1.70E+07
min	1.2	5.76E-02	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.55E+00
25%	18.9	1.24E+06	8.2	0.1	0.0	0.0	0.0	2.0	4.0	10.0	24.0	7.58E+04
50%	128.2	2.05E+07	10.3	3.2	0.0	0.0	0.0	18.0	44.0	44.0	114.0	5.59E+05
75%	346.1	7.65E+07	11.9	14.3	0.0	0.0	0.0	43.0	112.0	114.0	318.0	2.34E+06
max	25861.1	9.38E+08	20.1	220.5	1922.0	4890.8	7909.4	216.0	661.0	1218.0	3446.0	1.48E+09

Table B.1: Descriptive statistics for full dateset, initially fed to model, the spatially lagged statistics are included in another table. Note that the power plant data is sparse and skewed

B.2. Descriptive statistics for lagged variables

N = 68139	avg_rad_lag	population_lag	gdp_lag	pm25_lag	road_centroids_lag	intersection_density_lag	pedestrian_roads_lag	pedestrian_intersection_lag
mean	224.0	9.5	5.16E+07	10.1	25.4	64.4	79.5	224.4
std	257.1	11.8	6.39E+07	2.6	21.1	54.4	83.5	249.6
min	0.0	0.0	0.00E+00	0.0	0.0	0.0	0.0	0.0
25%	25.1	0.5	3.29E+06	8.2	5.0	12.0	15.0	38.0
50%	148.4	5.6	2.90E+07	10.3	23.3	57.0	54.0	144.0
75%	356.6	14.1	7.67E+07	12.0	41.3	106.5	122.0	343.5
max	13070.0	142.6	6.89E+08	19.3	164.0	462.0	895.7	2904.7

Table B.2: Descriptive statistics for the spatially lagged variables, which seem to mimic the properties, in terms of standard deviation mean, of the original variables they are based on.

B.3. Visualisation per year

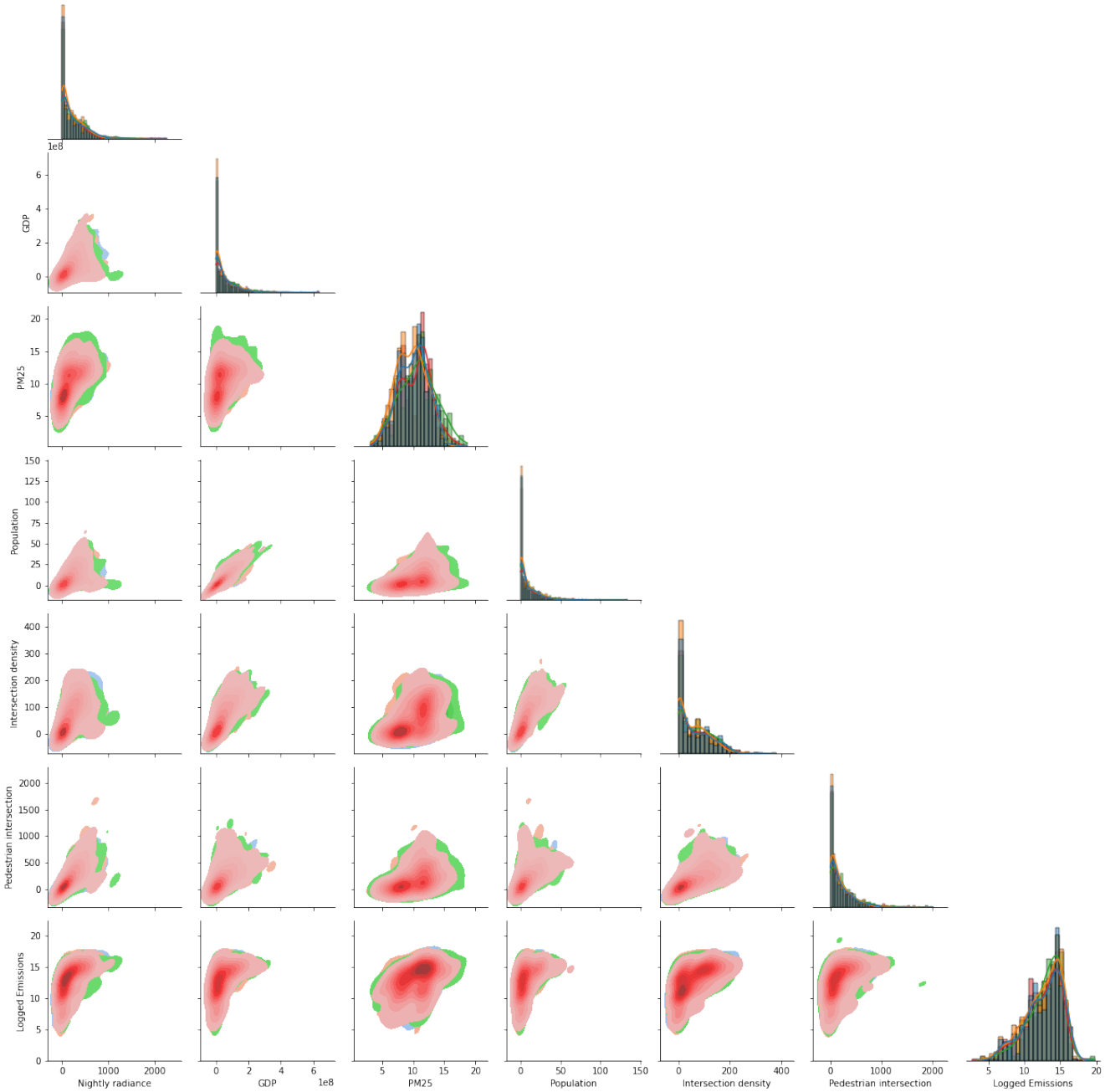


Figure B.1: Pair-plot showing the relationships and distributions of spatial covariates, conditioned on the year. This plot shows that there are no big differences between the analysis years in terms of distribution nor relationships between the variables.

C. Appendix: Results

C.1. Choice of algorithm

For the hybrid method the algorithm that combines the spatial covariates with the bottom-up emission inventory results is important. Multiple machine learning techniques exist that could help with the current regression task. Earlier studies, and for example Stevens et al. (2015); Meyer et al. (2018) commonly use a Random Forest model. Other ensemble methods such as XGBoost, relying on gradient tree boosting, also are valid contenders.
To asses which model would produce the best result a test is designed. As with the main model five fold grouped cross validation is used. For the scoring metrics the same are used to evaluate each model. An overview of the results of this experiments can be found in the table C.1. The reported scores are the average scores of the five folds.

Perhaps counter intuitively, the metrics in table C.1 reflect that the estimates by XGBoost are more accurate. However, in this study the Random Forest model is further used. Inspection of the individual test set scores reveal that the difference between XGBoost and Random Forest is rather small, for some folds the Random Forest is better and for some XGBoost. In general however XGBoost outperforms the Random Forest. There are however two additional reason to pick the Random Forest, A has a lower computational time, and B is used more often. Because the difference is very small the decision is made to use the Random Forest further in this study.

	Random Forest	XGBoost	SVM	Elastic Net	Lasso	OLS
MAE	1.35	1.34	1.49	1.49	1.50	1.49
R2	0.37	0.38	0.25	0.24	0.23	0.26
MAPE	0.12	0.12	0.14	0.14	0.14	0.14
Expl_var	0.42	0.43	0.31	0.28	0.27	0.31
KS-score	0.13	0.13	0.28	0.24	0.28	0.23

Table C.1: Comparing the algorithm performance, Random Forest and XGBoost are close, the non ensemble based models have far worse performance. In this test XGBoost outperforms the Random Forest

C.2. Test set performance

C.2.1. Descriptive statistics test set

The table C.2 gives an overview of the descriptive statistics of the input and emission observations and estimates of the model. The four most right columns provide the most information regarding model output and observed values. Looking at the difference between columns $ylog$ and $yhat$, of which the first is the emissions observation column in log transformed space and the former one the estimated output. The means are similar, however the standard deviation of the observed emission values is considerably higher then estimated. This is also partly reflected in the min and max values that are higher and lower for the observed values. The values for the first, second and third quantiles are more similar between estimated and observed.
In the untransformed space, looking at the emissions column and the column $yhatunlog$ the estimates seem to be far off in absolute sense. The means are vastly different. Like observed in log space, the standard deviation of observed values is greater then the model estimation.
The input variables seem to have a higher inter-quartile range coefficient compared to the observed and estimated emissions. Especially the population and GDP covariates have higher dispersion compared to the other variables.

	avg_rad	gdp	pm25	population	intersection_density	pedestrian_intersection	emissions	y_hat_unlog	y_log	y_hat
count	13398.0	13398.0	13398.0	13398.0	13398.0	13398.0	13398.0	13398.0	13398.0	13398.0
mean	192.6	40275160.0	11.2	7.1	53.4	189.2	2469796.0	1161718.0	12.8	12.8
std	234.6	47687120.0	2.9	9.2	56.3	247.0	15697920.0	1650761.0	2.4	1.8
min	1.4	63.1	5.8	0.0	0.0	0.0	464.0	875.8	6.1	6.8
25%	19.5	1668393.0	8.6	0.1	4.0	36.0	87330.9	126385.3	11.4	11.7
50%	108.5	19624710.0	11.1	2.6	34.0	98.0	575727.3	442017.0	13.3	13.0
75%	281.6	66239150.0	13.5	11.1	88.0	242.0	2102308.0	1802777.0	14.6	14.4
max	2418.7	259271300.0	20.1	62.3	327.0	3122.0	527446600.0	21203820.0	20.1	16.9
IQR Coefficent	0.87	0.95	0.22	0.98	0.91	0.74	0.92	0.87	0.12	0.10

Table C.2: Descriptive statistics for the covariates, observed emission values and predictions in log standardised form and in the original space.

C.2.2. Test set performance

The table C.3 gives an overview of the set set results both in transformed and untransformed space. These numbers are achieved by first comparing observed and estimated in log transformed space, this is reflected in the left column. The untransformed metrics are found in the second column.
The final column, named back transformed target regression, is a robustness test in a way. Given the detoriation of the explained variance metric, R2 and Spatial Correlation in the untransformed space the idea was to use a back tranformed target regression. The same method for recursive feature elimination and hyper-paramter tuning was used to asses whether this would improve for example the explained variance. The difference between back transformed target regression and the current method is subtle. the original model transforms the target variable, emissions in this case, to log space and the Random Forest trains and evaluates the error in log space. When using the approach of a transformed target regression still trains in the transformed space, however model evaluation happens in the original space. This is use full As the model is more interpretable in the original space. However, the results of doing so led to the same hyper-parameters and features and hence without a significant performance increase. Therefore it was decided against doing so. Similarlily, the hyperparameters and selected features did not change. This gives reason to believe that the current approach, evaluating the model in log space does not lead to significantly different results. The current approach, using log transformed emissions values is the method to derive most value from the current dataset. Gaining a model that is able to produce insights, but only in log space.

Score	Log transformed	Reversely transformed	Back transformed target regression
MAE	1.075	1.916e6	1.961e6
MRA E	0.095	2.718	2.992
Expl variance	0.6556	0.013	0.019
R2	0.6546	0.006	0.136
R	**0.8131**	**0.1154**	**0.1364**
Kolmogorov	0.077	0.077	0.077

Table C.3: Test set performance in log space and original space. Note the accuracy decrease in non log space, even after using back transformed target regression. Kolmogorov scores stay similair, showing that the prediction of the distribution remains stable.

Test set performance per spatial cluster

The metrics for the two spatial cluster in the hold out set are given in table C.4. One can see that the metrics tend to vary between the different clusters. A notable difference between cluster three and seven is that while MAE and MRAE are lower for cluster three, the explained variance, R2 and spatial correlation indicate better results for cluster 7. Another observation is that the Kolmogorov-Smirnov statistic for the individual clusters (at 0.120 and 0.140) is higher then the combined score ($\sim$ 0.077), see table 3.3 and C.3. The other metrics that take a hit when splitting out the results over the clusters are the R2, R and explained variance. These scores do not reflect the average of these metrics for the two clusters. For MAE and MRAE the result seems to be more of an average score of the two clusters.

	MAE	MRAE	R^2	R	kolmogorov stat	expl_var
Cluster						
3	0.963	0.073	0.468	0.694	0.120	0.473
7	1.236	0.128	0.559	0.776	0.140	0.582

Table C.4: Metrics divided per cluster, especially explained variance, R square and spatial correlation coefficients give a less optimistic view when separating the cluster performance.

Test set performance per year and spatial cluster
A final robustness test is splitting out the observations and estimates in both spatial and temporal dimensions. The detailed results are found in table C.5. The finding is remarkably similar as reflected in table 3.3. Namely, the yearly metrics do not vary much from each other. The biggest differences of the performance metrics occur when looking at other clusters, and hence occur in the spatial dimension. The results in the temporal dimension hence are conlcuded to be stable over the analysis years.

	MAE	MRAE	R^2	R	kolmogorov stat	expl_var	year
Cluster							
3	0.936	0.071	0.482	0.706	0.119	0.488	2012
3	0.968	0.073	0.455	0.684	0.121	0.460	2013
3	0.940	0.072	0.491	0.706	0.114	0.491	2014
3	1.008	0.076	0.446	0.683	0.131	0.459	2015
7	1.164	0.117	0.601	0.786	0.115	0.609	2012
7	1.301	0.137	0.508	0.762	0.171	0.543	2013
7	1.251	0.129	0.551	0.773	0.150	0.577	2014
7	1.230	0.128	0.576	0.793	0.144	0.606	2015

Table C.5: Showing the performance metrics dis-aggregated over years and spatial clusters. Grouping all together might provide an overly optimistic view of model results.

Effect of hold-out on model selection
A question that arises is how the results of hyper-parameter optimisation change under different training and hold-out sets. The results of said experiment are given in table C.6. The pipeline to select the best model remains unchanged across these experiments. The only difference is the amount of search iterations (50 versus 75 for computational reasons) during the randomised hyper-parameter grid search. It seems that the cross-validation scores remain relatively stable for different hyper-parameter sets, the effects of tuning is thus marginal on the final score. The search results however are quite different. The first two columns, with hold-out sets 1 and 10 & 0 and 1 result in the same optimal hyper-parameter set. This set is labelled as set A. The originally found set, as described in the main report using cluster three and seven is labelled set B.

To asses which of these hyper-parameters sets result in better test set performance an additional test was ran. For each iteration in this test, two random hold-out sets where picked. Then the test set performance of two independently trained models are calculated. First the model is trained on the training set (without the randomly picked hold-out set), this model is then used to estimate emissions in the performance set. The two models are trained like this, the difference between these two models however, are the hyper-parameters, testing first the hyper-parameter set A and set B.
To have a change to see different combinations of the 100 possible hold-out set combinations, this experiment runs for 35 iterations. The average of the scores is calculated and given in table C.7. Surprisingly the hyper-parameters set B generalises best with a lower average than set B. The original hyper-parameter set found with more iterations thus outperforms the one found under different hold-out sets.

The main conclusion of these experiments is that while there is some variation in the found optimal hyper-parameters, the results are quite robust and do lead to the best possible generalisation performance. The test set performance is only marginally dependent on the hyper-parameters. The fact that the same features are selected also for different hold-out sets give confidence to the fact that the same parameters would be found for other data.

Test set	[0,1]	[10,1]	[8,5]	[3,7]
N estimators	250	250	100	450
Min samples split	4	4	4	16
Min samples leaf	2	2	8	8
Max features	sqrt	sqrt	auto	sqrt
Max depth	10	10	10	10
Bootstrap	TRUE	TRUE	TRUE	FALSE
CV Score (MAE)	1.349	1.206	1.459	1.394
Search Iterations	50	50	50	75
Hyper-parameter set	A	A	-	B

Table C.6: Results of hyper-parameter tuning process for randomly picked hold-out sets. These hyper-parameters seem to vary a little bit between samples. However, the cross-validation score seems to be relatively stable, the effect of the hyper-parameters is marginal.

	Hyper-parameter set	**MAE**	**MRAE**	**R^2**	**R**	**kolmogorov stat**	**expl_var**
Mean Score	A	1.576	0.151	-0.142	0.676	0.334	0.457
Mean Score	B	1.509	0.148	0.020	0.674	0.286	0.456

Table C.7: Mean test set scores of the model with different hyper-parameter sets, as can be inferred from the mean scores, set B leads to better performance in terms of MAE and r2.

C.3. Inter quartile range robustness

This section is aimed to understand the robustness of the inter quartile range coeficient results. Knowing the robustness of these results and relative error to expect helps when generalising the model.

C.3.1. Inter quartile range coefficient and transformation

To generalise findings and to analyse trends, the inter quartile range coefficient could be used. Before generalising however, the accuracy of this coefficient has to be assessed. This gives more confidence when generalising the model.

This is done by computing the IQR coefficient for both the observed and estimated emission values. This is done twice, once in log space and once in the original space, to see how this coefficient is affected by log transformation. The table given in table C.8 shows the observed and estimated values of the inter quartile range coefficient. The relative difference is also given, an interesting property is that this relative difference stays the same before and after log transformation. Giving confidence in the utility and accuracy of this indicator after back transformation. In this sample The relative difference is between 2 and 21 percent. Like with the other metrics, these relative differences are smaller (and

Cluster	Year	Observed Gini	Estimated Gini	Relative Difference [%]	Observed Gini (log)	Estimated Gini (log)	Relative Difference [%]
3	2012	0.665329	0.517511	0.222172	0.068304	0.057425	0.222172
3	2013	0.668836	0.519433	0.223377	0.068304	0.055968	0.223377
3	2014	0.67168	0.49967	0.256089	0.069068	0.055771	0.256089
3	2015	0.672821	0.52776	0.215601	0.070553	0.057272	0.215601
7	2012	0.918135	0.568458	0.380856	0.119527	0.079326	0.380856
7	2013	0.914492	0.549848	0.398739	0.119496	0.064823	0.398739
7	2014	0.911483	0.566166	0.378851	0.119889	0.071813	0.378851
7	2015	0.908129	0.565462	0.377333	0.121973	0.074729	0.377333

Table C.9: Gini-coefficients both estimated and observed, split out per spatial cluster and year. Like the conclusion for the IQR coefficient, the relative differences remain stable after log transform, these relative differences however are comparatively bigger then for the IQR coefficient.

hence a better prediction) for spatial cluster three. These differences or relative errors seem reasonable given the model performance.

The models estimate of the Gini-coefficient is also tested, with the same method as for the IQR coefficient. These results could be found in table C.9. Again, like for the inter quartile ranges relative differences, the relative difference remains the same before and after log transformation. However, for the Gini-coefficient the relative differences are considerably bigger, between 21 and 37 percent. It seems that this indicator is more sensitive to mis-estimates of the model, likely as the model estimation misses extreme values and generally has a lower standard deviation. Therefore it was decided against using the Gini-coefficient.

Cluster	Year	Observed IQR coef	Estimated IQR coef	Relative Difference [%]	Observed IQR coef (log)	Estimated IQR coef (log)	Relative Difference [%]
3	2012	0.73	0.75	0.02	0.07	0.07	0.02
3	2013	0.74	0.77	0.04	0.07	0.07	0.04
3	2014	0.75	0.72	0.03	0.07	0.07	0.03
3	2015	0.76	0.77	0.02	0.07	0.07	0.02
7	2012	0.93	0.77	0.17	0.15	0.09	0.17
7	2013	0.93	0.74	0.21	0.15	0.08	0.21
7	2014	0.93	0.76	0.18	0.15	0.09	0.18
7	2015	0.93	0.78	0.16	0.15	0.09	0.16

Table C.8: Analysis in the behavior of the IQR coeficient. For the analysis years and spatial clusters, the relative difference between estimated and observed is generally small and does not change after log transformation.

C.3.2. IQR performance under different hold-out sets

Using a similar method as described in section 3.2.1 the effect of the hold out set on the inter quartile range coefficient and the relative error is determined. Doing so gives insight in the robustness of the results of the IQR coefficient estimate. And is aimed to illustrate that this is coefficient is not super dependent on the test set cluster

The result of this experiment is visualised in figure C.1. The absolute relative difference typically lays between 5 and 16 percent, which seems acceptable given the other performance metrics such as the MRAE which almost falls within the same confidence interval. On average the model has a relative absolute difference of 13 percent.

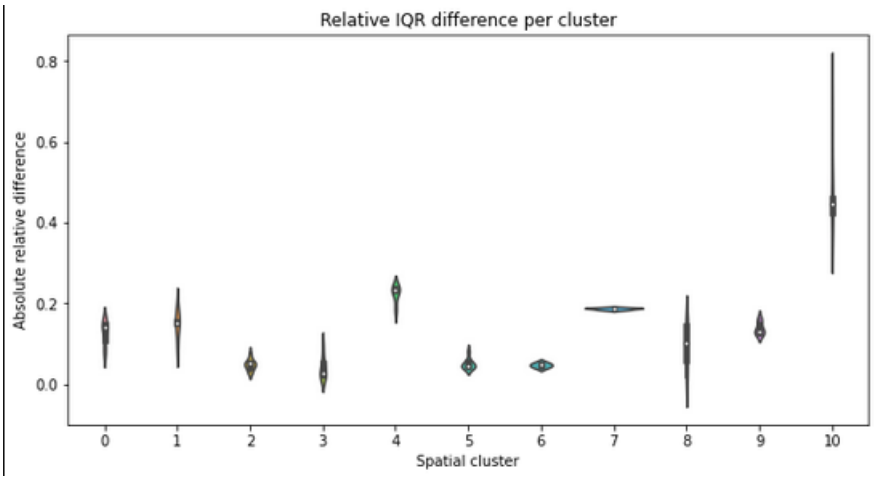


Figure C.1: Violin plot depicting the absolute relative difference between the observed and estimated IQR coefficient. Except for cluster ten, the values lie between 5 and 16 percent, which seems acceptable given the model performance

D. Appendix: Emission estimates for ten largest cities

The following appendix consist of analysis per city of model output and interpretation of the spatial distribution of emissions in that city. The end section D.9 combines the insights of these per city analyses.

D.1. New York City

New York in the state of New York is ranked as the most populous city in the USA. The model results, see figure D.1, indicate less variability of emissions with a lower inter quartile range coefficient. This is also reflected in the distribution plot where the plot becomes more ”steep” as the years increase. The map from 2015, visualised in figure D.2, confirms this finding. The emissions seem to be quite evenly spread. With a cluster of emissions around Staten Island. In general it seems that emissions are concentrated around the coast

D.2. Chicago

Ranked the third most populous city in the USA, situated in Illinois. The model results,figure D.3,in terms of distribution and IQR coefficient seem less stable when compared to other cities. The emission distribution curves seem to be similar for the years 2012 and 2014. The years 2015 and 2013 show different behaviour. There is no obvious tendency of the IQR coefficient.
Looking at the figure D.4, emissions in Chicago are less situated at the coast. The distribution here is high emissions in the centre of the city, lower on the outskirts and around the coast.

D.3. Phoenix

Phoenix, located in Arizona. Looking at the model results, visualised in figure D.5, there seems to a tendency over the years to have more density around the median, that slowly decreases. This is also reflected by it strongly decreasing IQR coefficient.
The high IQR coefficient is also reflected on the map, visualised in figure D.6, there seems to be ”breaks” in the form of parks or edges around and within the city in the emission pattern. In general however, the centre of the map is a large cluster of high emission cells. There is also a cluster of high emission cells to the south east.

D.4. Philadelphia

The predictions for Philadelphia visualised in figure D.7 show that the distribution curve changes considerably over the years. The peak and the form shifts around quite a bit, seemingly making the predictions unstable. As a consequence the IQR coefficient also shifts around. It however seems to show a slight decreasing tendency.
The IQR coefficient is also quite low compared to the other cities, this is also reflected in the map in figure D.8, showing a quite even distribution across the city, but with a cluster of high emission cells in the south. As indicated by the low IQR coefficient relatively few cells fall within the top ranges of emissions.

D.5. San Antonio

The estimates for San Antonio, Texas, are visualised in figure D.9. The results seem quite similar and constant over the analysis years. The IQR coefficient has an upward tendency over the years. Meaning that over these years, less cells fall within the first and third quantiles, getting more extreme values..
Looking at the map, visualised in figure D.10, the emissions seem to be mainly concentrated around the centre. Going outward emissions generally decrease.

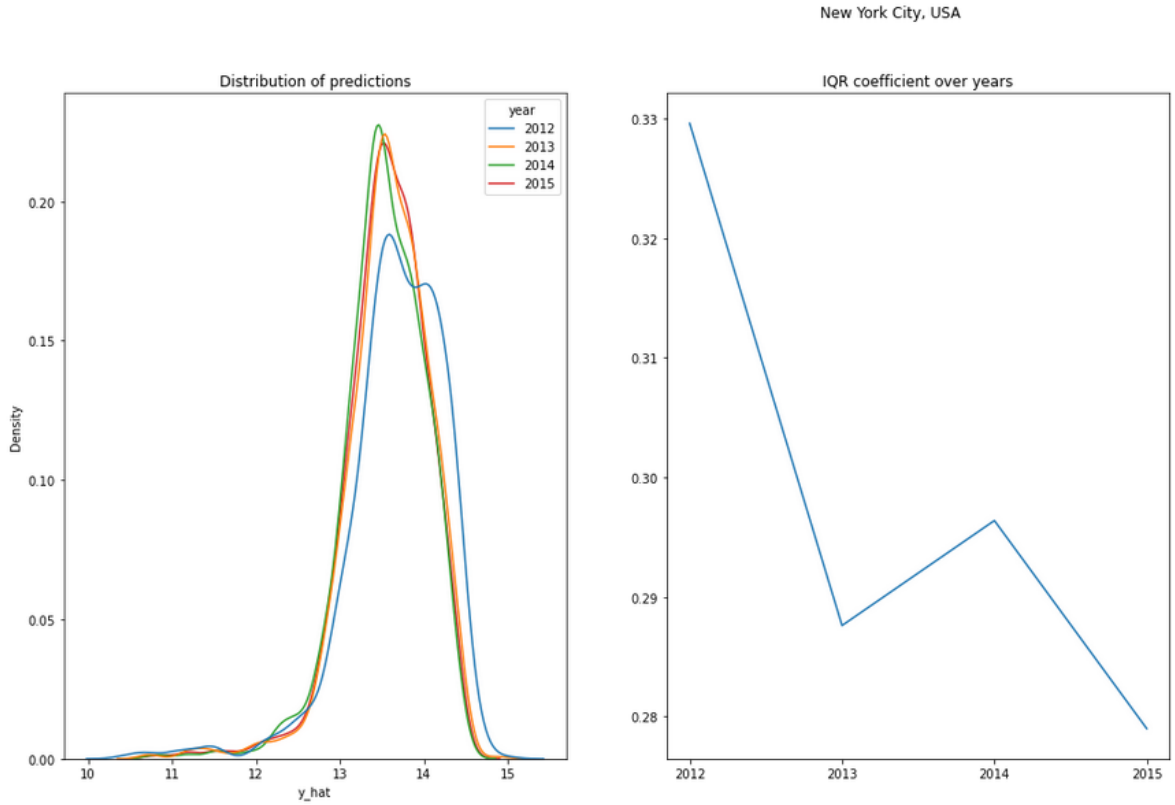


Figure D.1: Model estimates for the distribution of emissions in New York, the cureve seems to shift left as the years progress, the IQR shows decreasing behaviour

D.6. San Diego

The estimates for San Diego seem to be relatively stable over the years, see figure D.11. The concentration of the median predicted values seems to increase, this is also reflected by the decreasing IQR coefficient over the years. The increase around the median along with a decreasing IQR coefficient means a rising trend over the years.
Looking at the map of San Diego, see figure D.12, emissions seem to be concentrated around a centre on the coast. There is also an area to the south with a cluster of very high emissions.

D.7. Dallas

The model estimate, visualised in figure D.13, seems to be stable over the years. The value with highest density seems to have shifted towards slightly more frequent an lower. The IQR coefficient seems to slightly increase over the analysis years and is around 0.36 on average.
The map, figure D.14, reveals no obvious clusters of emissions in or around the centre, as the IQR suggests, emissions seem to be relatively evenly dispersed over the extend of Dallas.

D.8. San Jose

Except for the year of 2013 the model estimates for San Jose, California, visualised in figure D.15 seem to not differ much. The IQR coefficient is around 0.5 on average. This is slightly higher compared to the other cities analysed.
Looking at the map, in figure D.16, this is reflected, as the higher emission cells are mainly clustered around the centre. Moving outwards of the city, generally the emissions become lower. The centre of San Jose also shows breaks in high emission clusters, which is again reflected by the IQR coefficient.

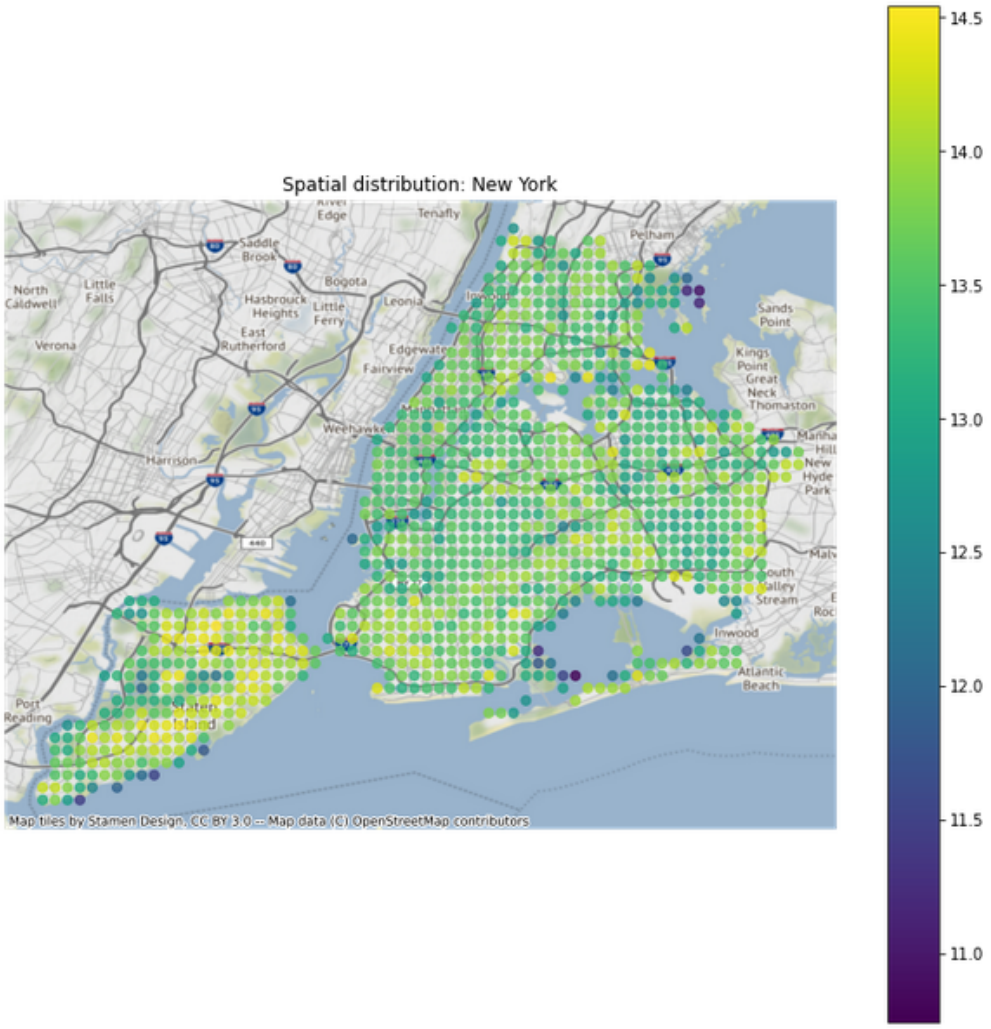


Figure D.2: Emission map of NY in 2015, emissions seem to be relatively even distributed

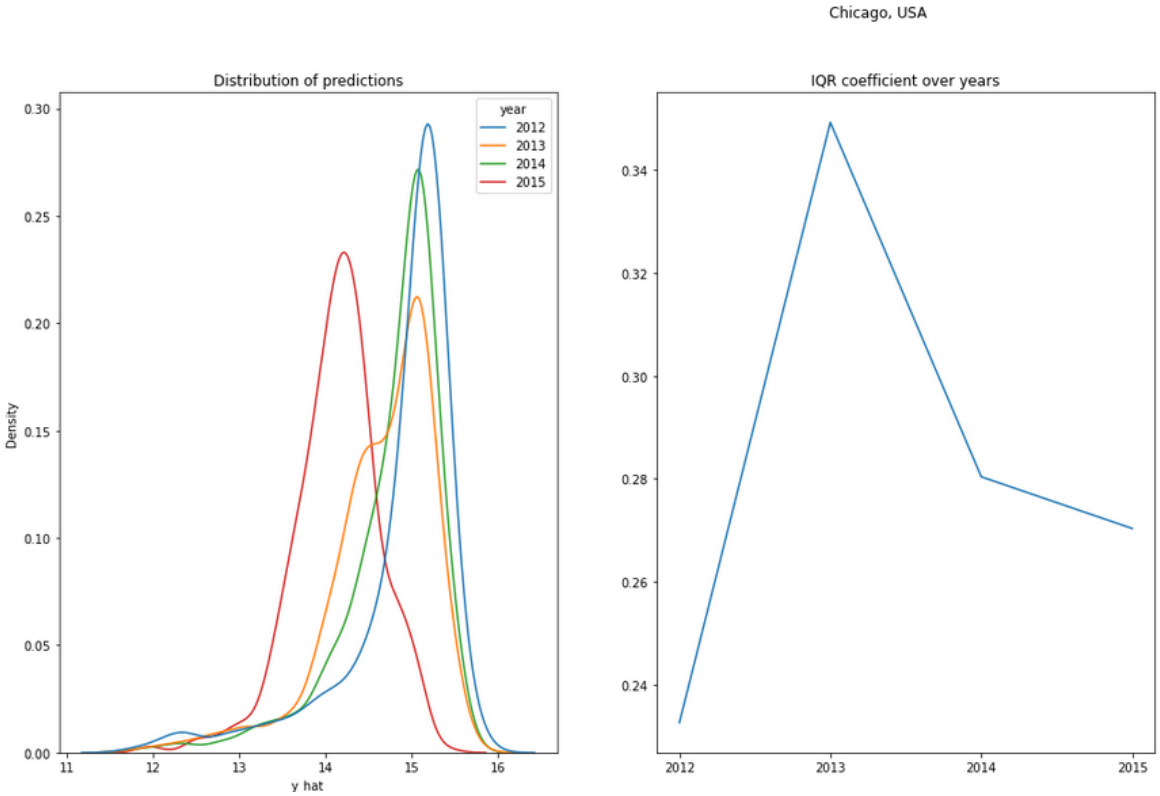


Figure D.3: Density and IQR coefficient plot of emissions for Chicago, the yearly variability is seems larger and less stable, inferring a trend is difficult, but especially for the year 2015 the emissions are lower

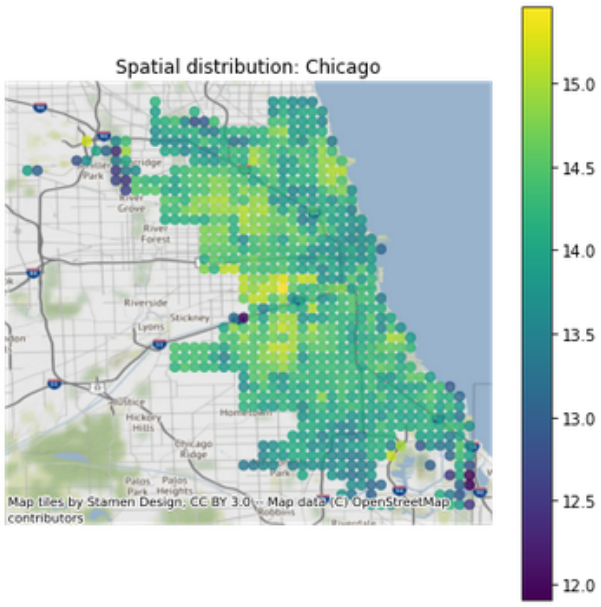


Figure D.4: The emission distribution in Chicago visualised, there is a cluster of high emission cells roughly in the centre. Unlike other cities on the shore, less emissions are situated along the coast.

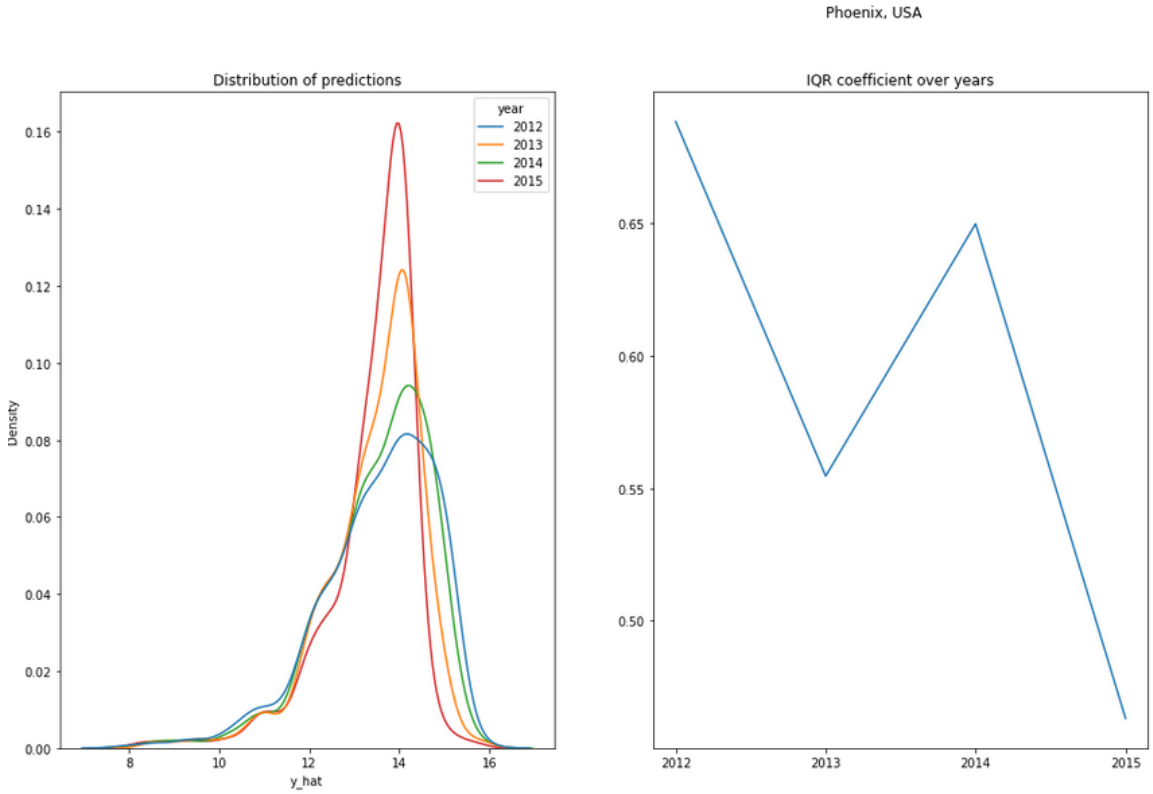


Figure D.5: Model results over the years for Phoenix city, showing a stronglt decreasing IQR coefficient trend. The estimates over the years seem to change considerably, shifting towards more density around the median.

D.9. Aggregate results

Combining the results mentioned above it is possible to make statements about the model results over the ten most populous cities in the USA. The IQR coefficient is analysed for the both the model estimations (which cover eight of the ten most populous cities). As wall as for the observed IQR coefficient in the original Hestia data.

The observed data has an IQR coefficient that is comparatively stable over the years. The IQR for the observed data varies between 0.55 and 0.66 for these cities.
For the new cities, where the IQR is achieved through model estimates, three cities lay within this realm namely: Phoenix, San Jose and San Diego. The IQR seems to decrease a little bit meaning less dispersion of emissions in a city.
The other cities, San Antonio, Dallas, Philadelphia, Chicago and New York have a considerably lower IQR coefficient. For these cities the distribution and variability of emissions is lower and hence the model estimates that it is more equally divided here.
Comparing cities with a higher IQR coefficients to the ones with lower IQR coefficients, the cities with higher IQR coefficients often have a cluster of high emission points located somewhere. Where cities with lower IQR coefficients, seem to have a more equal distribution and a cluster of high emissions more centrally.
The analysed cities located around the coast often have higher emission values along the coast that decrease while going inland. Other cities, located inland typically show high emissions in the centre, with decreasing intensity towards the edges. Some cities, for example Phoenix and San Jose, have breaks in the high intensity emission cells in the centre. This generally increases the IQR. These relatively lower intensity cells are often hills and other breaks in the urban fabric. This indicates that possibly adding for example the average elevation of a grid cell could potentially explain some model variance. Similarly, perhaps land use classification could have explanatory power.

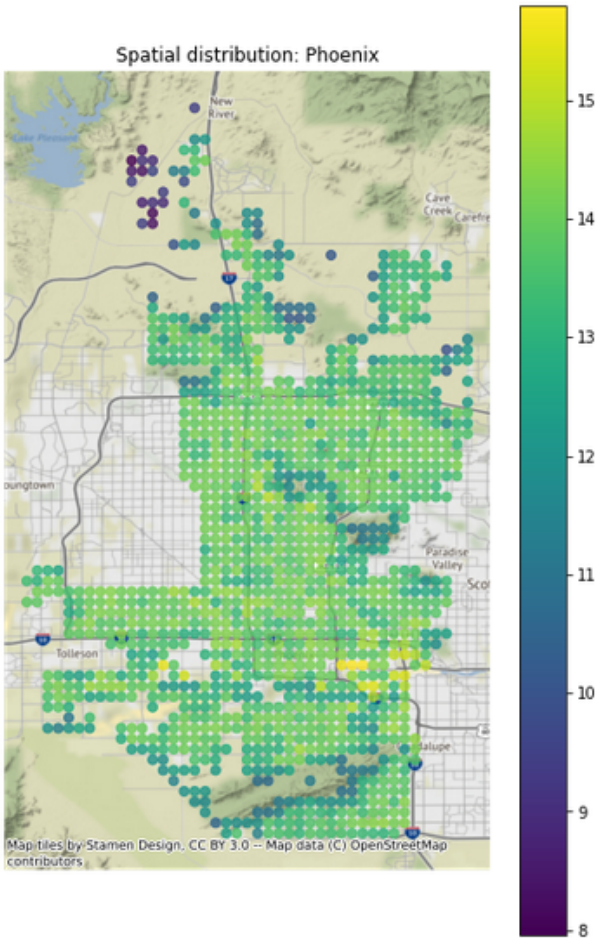


Figure D.6: Spatial distribution of emissions in Phoenix, multiple clusters of high emission areas can be observed, this is slightly different then the standard distribution, there also seem to be breaks in emissions, in the form of hills.

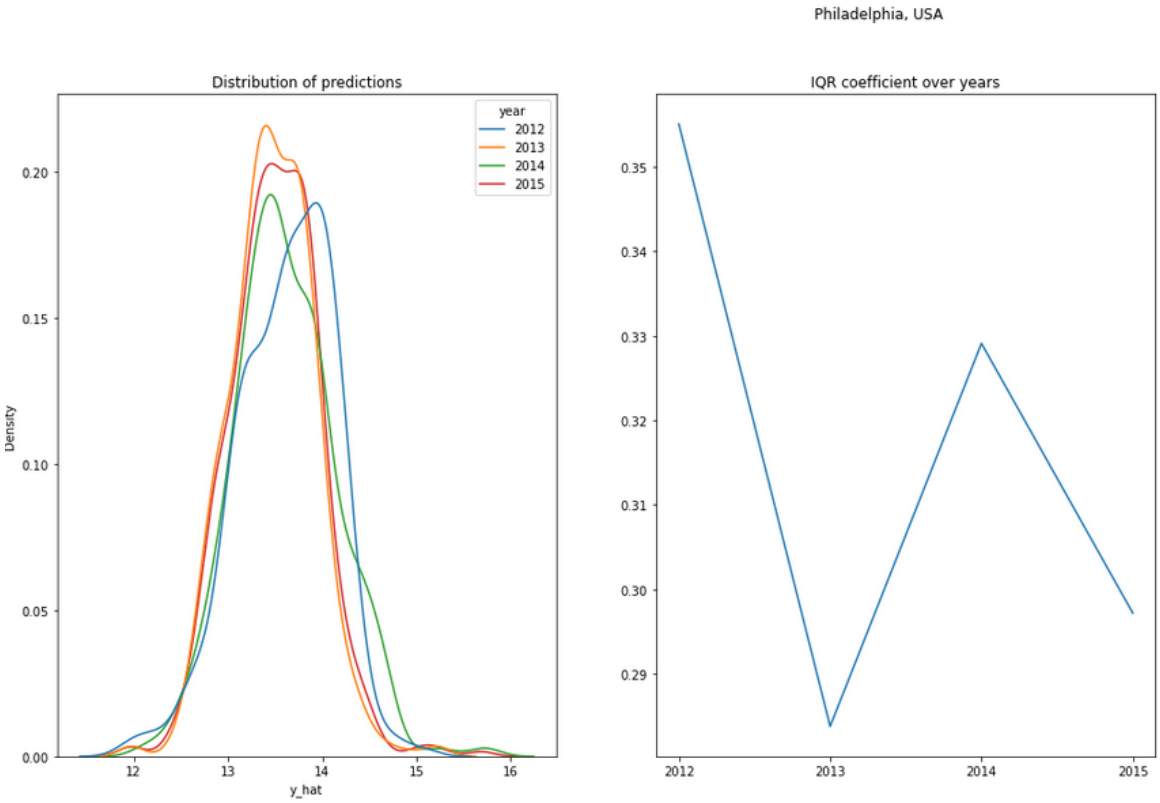


Figure D.7: The distribution of emissions in Philadelphia, with the peak shifting around considerably over the analysis years, as a consequence the IQR coefficient also shifts around more

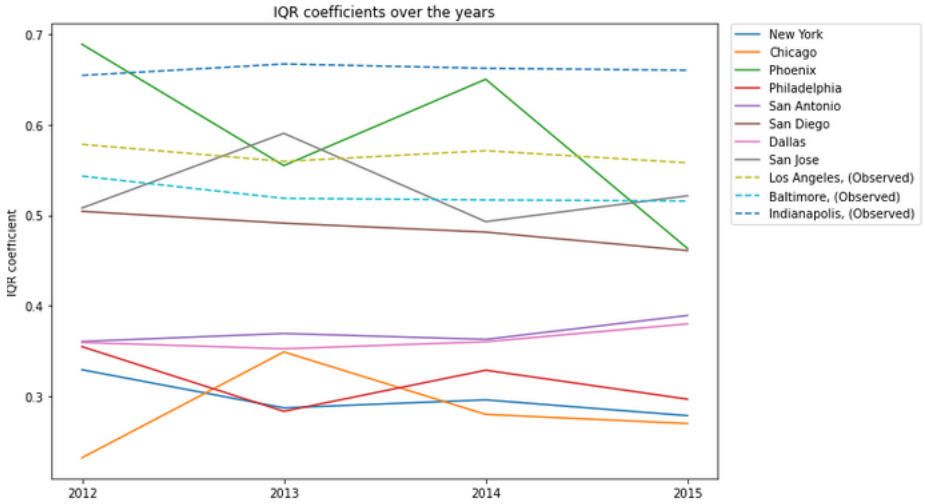


Figure D.17: The IQR coefficient over the analysis years, the Observed IQR coefficients are also included, these seem to be higher then most estimates. For most estimates the IQR seems to be on a decreasing trend, except for San Antonio and Dallas.

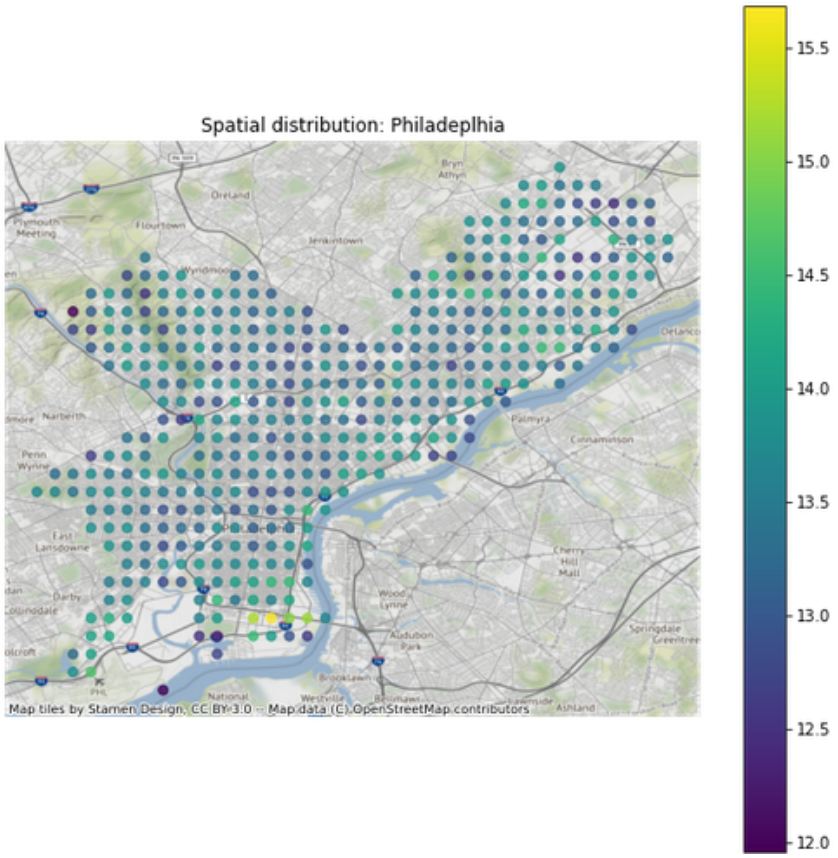


Figure D.8: Map of Philadelphia, showing a quite even spread of emissions in the city. To the south there is cluster of high emission cells, but over the rest of Philadelphia the emissions are spread evenly.

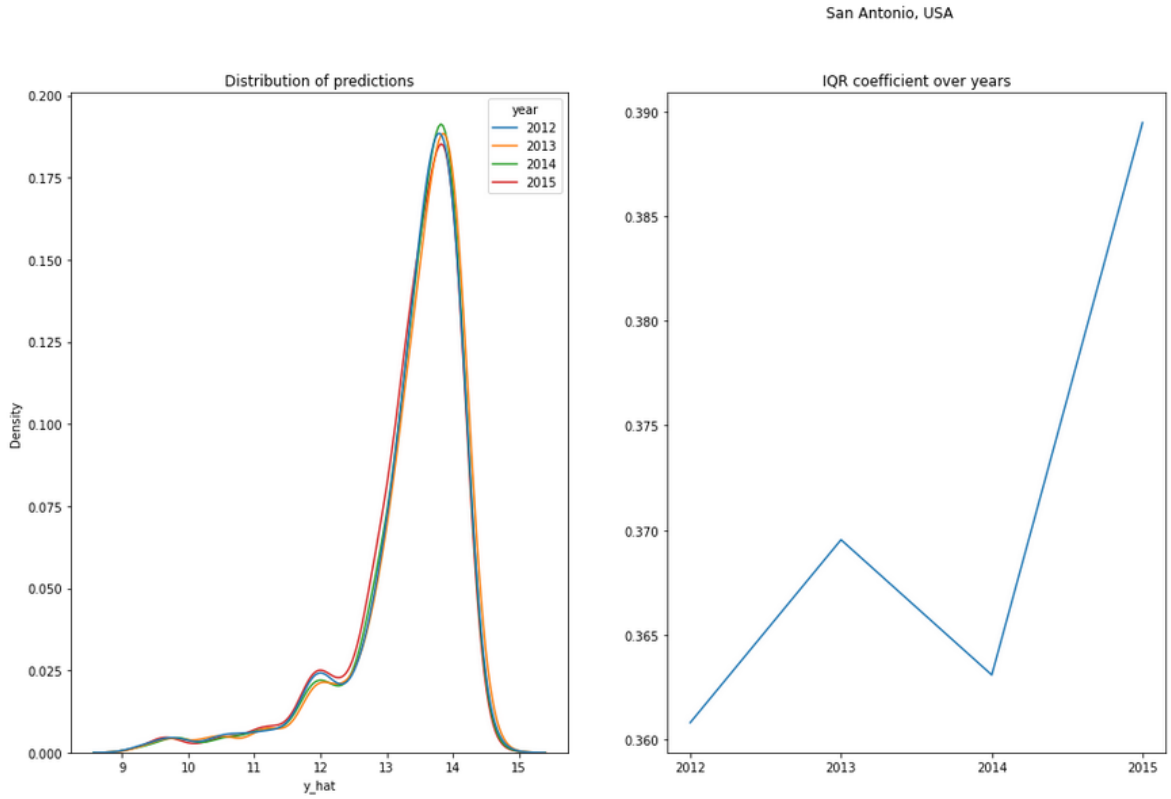


Figure D.9: The distribution plot of San Antonio seems stable over the analysis years, the IQR however seems to show an increasing trend.

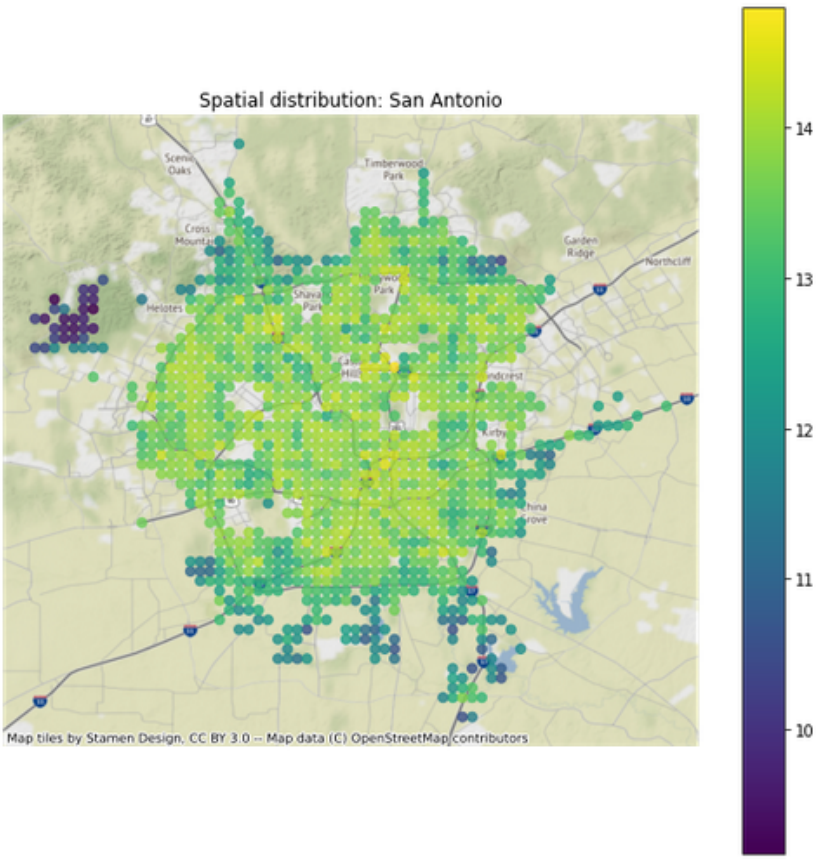


Figure D.10: Emissions on a map in San Antonio, which shows a centre of high emissions. emissions decrease moving outwards

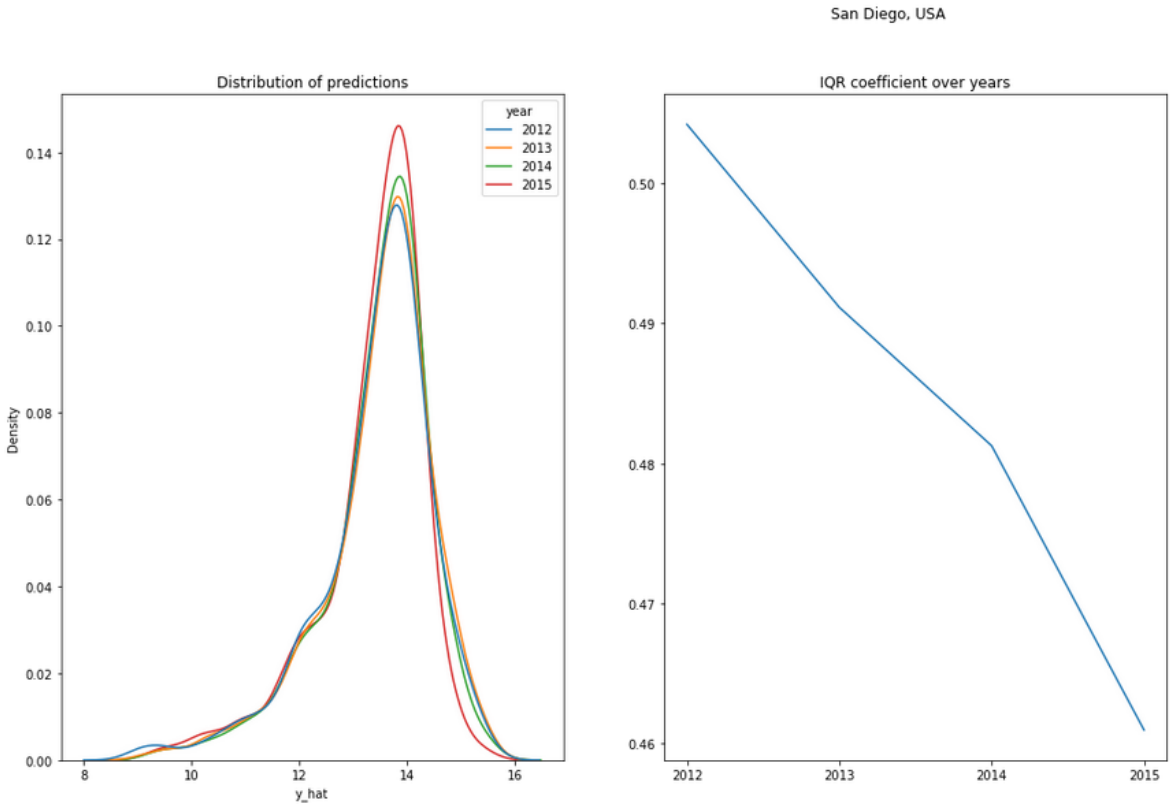


Figure D.11: Distribution of emissions, which seem relatively stable over the years. The IQR coefficient shows a decreasing tendency.

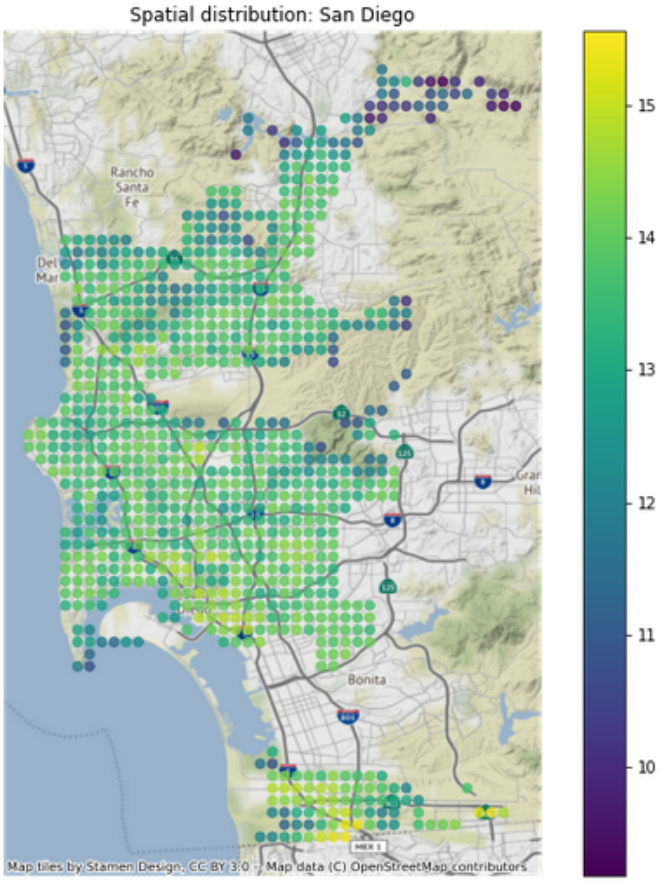


Figure D.12: Map of emissions of San Diego, showing multiple high emission clusters, resulting in a higher IQR, emissions are concentrated around the coast and decrease moving in lands.

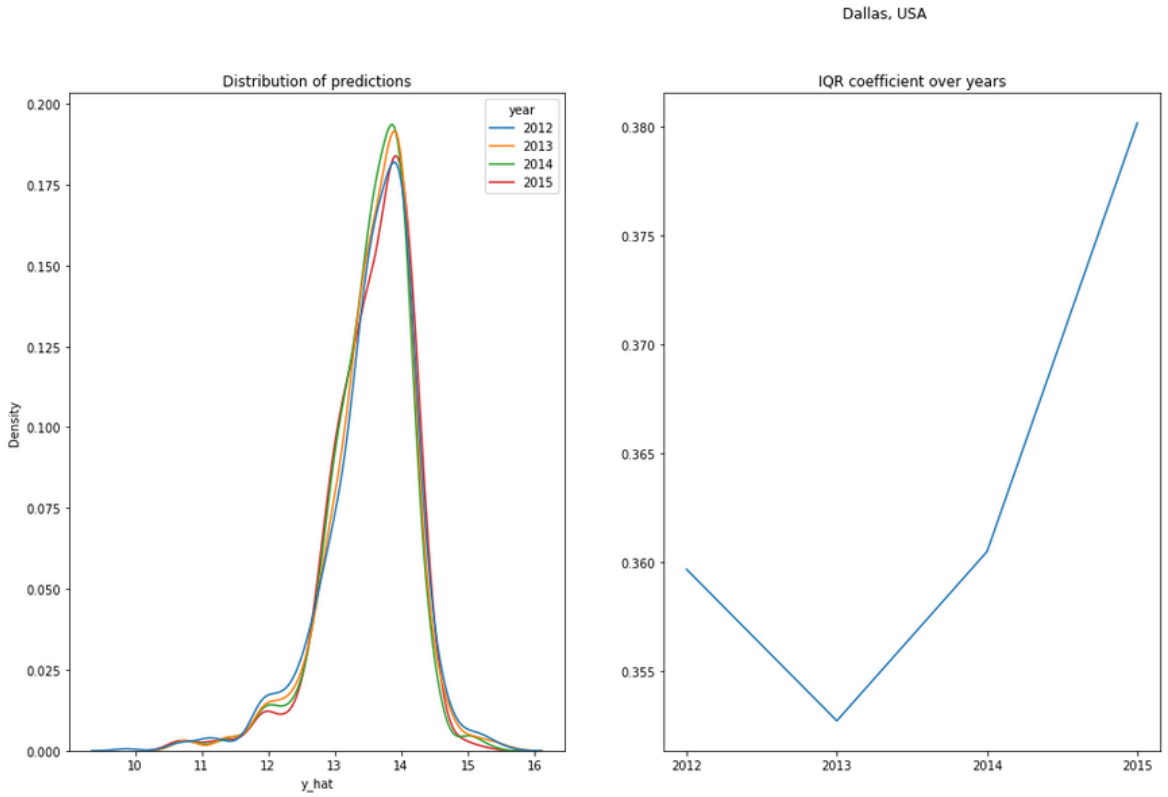


Figure D.13: Distribution of emission for Dallas, results are stable over the analysis years, the IQR coefficient shows an increasing trend.

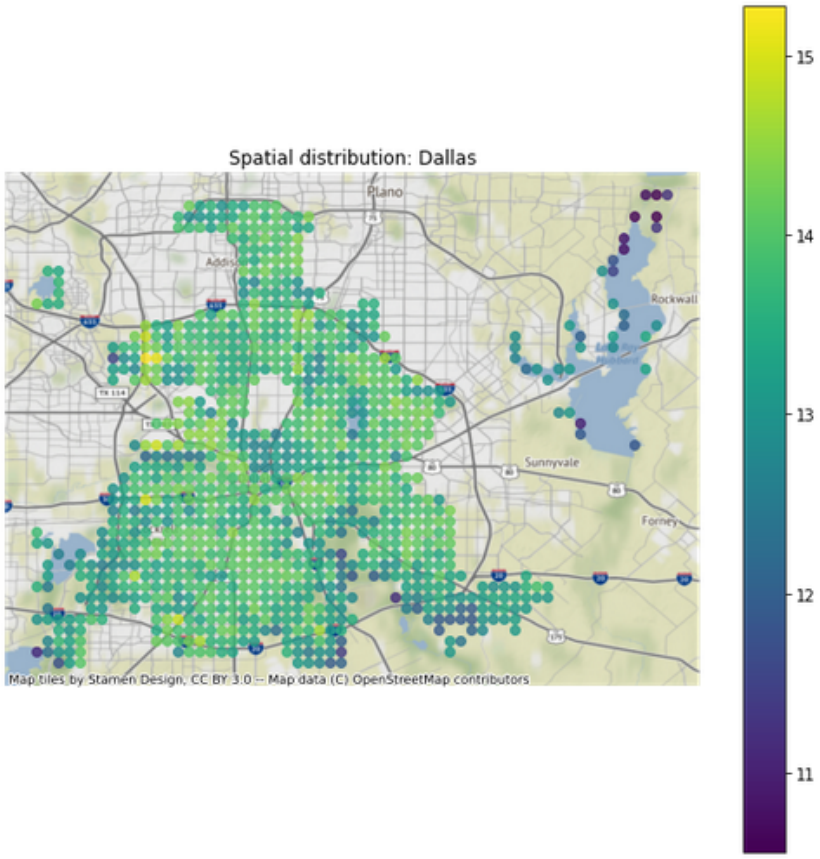


Figure D.14: Map of emission in Dallas, without obvious clusters of emissions in the centre, a few high emission cells are situated at the edge of the city boundaries

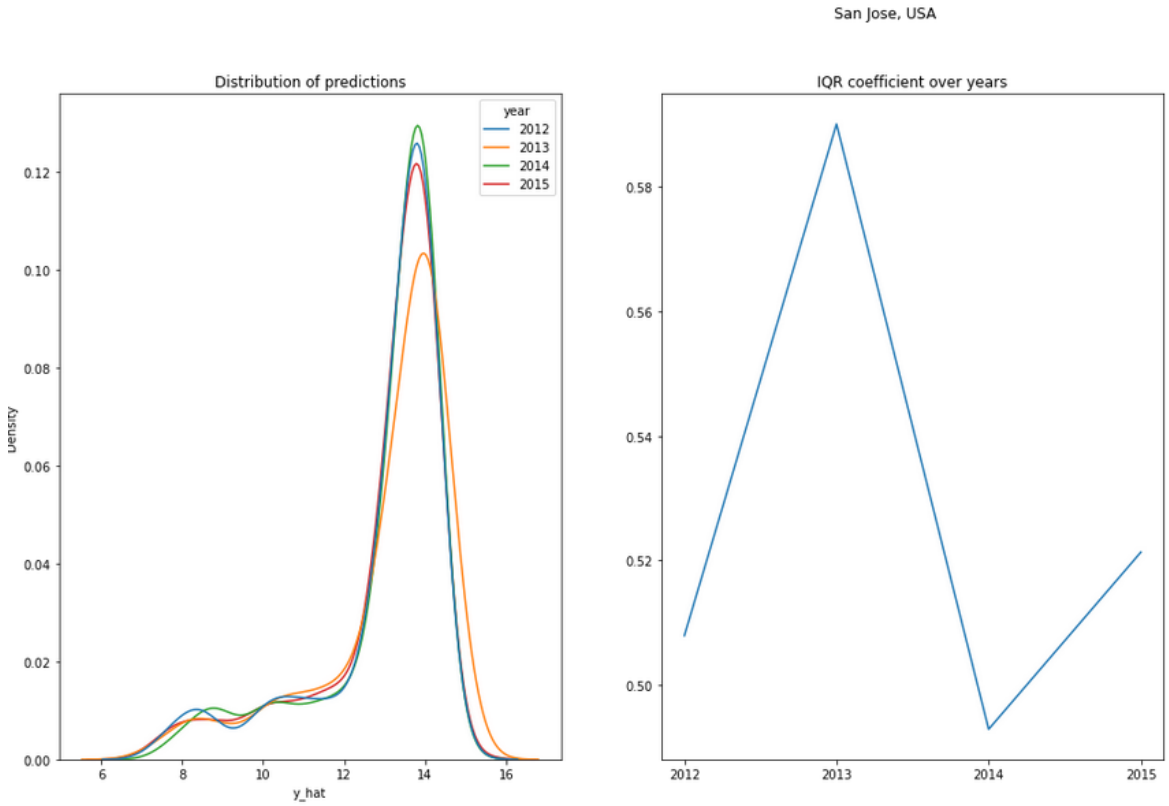


Figure D.15: Distribution of emissions for San Jose, which except for the lower end, looks stable, the IQR coefficient as a result bounces around a bit.

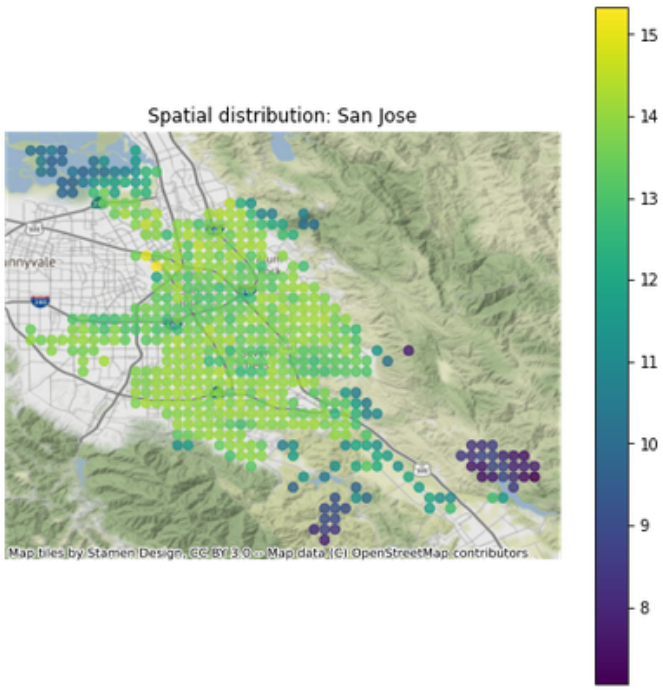


Figure D.16: Map of emissions for San Jose, showing a few clusters of high emission cell in the centre, this pattern is broken by areas with lower emissions also in the centre.