Norwegian University
of Life Sciences

Master’s Thesis 2022 30 ECTS
School of Economics and Business

Past and future emissions in the Norwegian economy: A combined machine learning and index decomposition analysis

Benedikt Goodman
Master of Science in Applied Economics and Sustainability

Abstract

This thesis concerns the past and future emissions within the Norwegian economy. It employs retrospective analysis to analyse past sectoral emissions on a national level and supervised machine learning to predict emissions 5 years ahead for Norwegian municipalities. Policies addressing GHG-emissions have been on the agenda for over 30 years, yet sectoral emissions from Norwegian economic activity have risen 10.4% compared to the 1990-level. Most retrospective analysis carried out on emissions only analyses trends and tentative causes. Underlying drivers of emissions therefore remain unquantified, and their magnitude remains unknown. A logarithmic mean divisia index decomposition analysis is provided on sectoral emissions from 10 economic sectors in the period 1990 – 2019 alleviate the problem/provide answers on this area. The analysis shows that economic growth and worsening energy efficiency, particularly in the transport and petroleum & mining sectors, have contributed to a net increase of 6218 mktCO2e in emissions. Results also show that changes in economic structure, decreased usage of fossil fuels and increased carbon efficiency have worked as abating factors, but that they are outweighed by the factors which contribute to increase in emissions.

Given Norway's ambition of curbing its own emissions by a significant degree by 2030 and then net zero by 2050. While these goals are specific only to certain types of emissions, it is still a somewhat open question as to how emissions might develop in the near future. A supervised machine learning analysis was carried out on GHG emissions from 354 municipalities in the period 2009 – 2019. An architecture using univariate linear regression tests for variable selection and extreme gradient boosting for prediction on a panelised dataset of emissions provided the lowest prediction error and projects that emissions from Norwegian municipalities will fall, reaching a level of 33 470 mktCO2e in 2025.

Foreword

This has been a very fun and challenging project and the result of many hours of labour. It has however, been a labour of love and a very educating journey at that as it has really opened my eyes to the amount of ingenious computational tools out there. So, before I thank anyone else, the many computer scientists and academics that develop computational tools and frameworks that are open source and freely available to use for anyone ought to be given a moment of appreciation. There are many others that have helped me and deserves gratitude. First of all, Pernille Lappegård who has patiently suffered through months' worth of ramblings and concerns about the method and helped me proof-read. Christine Hung and Lars Vik from SINTEF has provided valuable guidance along the way in terms of structuring the thesis and provided pointers for content. My advisor, Professor Arild Angelsen also deserves thanks, for making me aware of decomposition analysis and providing me with detailed feedback during the writing process. I also wish to thank professor Ang from the university of Singapore who has worked tirelessly for many years to improve the LMDI-IDA method have made it easily accessible to learn. Xiwang Xiang, Xin Ma and their team also deserves thanks for building the PyLMDI functions I used to build the LMDI analysis module used to produce the results of the retrospective analysis. It would have taken me a long time to build everything from scratch, so I am eternally grateful for how this helped me. Lastly, I wish to thank Farhan Derakshan for providing me with trust and access to data on energy usage.

Data collection, preparation, and analysis was carried out using python. The machine learning in this thesis uses the Sci-kit learn library. The code and data used for this analysis can be made available upon request. All mistakes made are my own.

Contents

List of figures and tables

List of abbreviations

AR6 – Assessment report 6

EU ETS – European Union Emission Trading System

GHG – Greenhouse gas

IDA – Index Decomposition Analysis

IPCC – Intergovernmental Panel on Climate Change

LMDI – Logarithmic Mean Divisia Index

LULUCF – Land use, land use change and forestry

MAE – Mean Absolute Error

OOB – Out of bag

RF – Random Forest

REDD+ - Reducing emissions from deforestation and forest degradation

RMSE – Root Mean Squared Error

SDA – Structural Decomposition Analysis

UN – United Nations

XGboost – Extreme gradient boosting

1 Introduction

1.1 Introduction

The latest IPCC report, AR6 Climate Change 2022: Impacts, Adaptation and Vulnerability (IPCC, 2022a), was published on the 28th of February 2022, and summarises the observed and projected impacts and risks of climate change. The report marks the 32nd anniversary of the release the first report (IPCC, 2022d) and thus indicates how long climate change has been on the international agenda. The main finding of the report is summarised as follows:

"Climate change is affecting nature, people's lives and infrastructure everywhere. Its dangerous and pervasive impacts are increasingly evident in every region of our world. These impacts are hindering efforts to meet basic human needs and they threaten sustainable development across the globe" (IPCC, 2022c).

The reports' summary for policymakers highlights the negative effects across the vast majority of areas across all continents (IPCC, 2022b). The negative impacts relate to food security, access to drinking water and key economic sectors all have an alarming number of projected negative effects due to climate change (IPCC, 2022b). It is evident that climate change, in the words of Secretary-General of the UN security council Antonio Guterres, "Is the defining issue of our time" (UN Security Council SC/14445, 2021). By virtue of being an existential threat to society, climate change is also a threat to future prosperity and sustainable development.

Globally, climate change is driven by greenhouse gas (GHG) emissions. As such, Norway has like many other countries made policies that seek to curb emissions. Norwegian policies have aimed to lower emissions domestically and internationally and have been in place at least since 1989 (A. Gullberg & S. Aakre, 2015; Berg, 2015). The current set of goals contains a carbon budget set in cooperation with the EU. The target is a 40% reduction in non ETS-regulated emissions compared to 2005-levels (St. Meld. 13 (2020–2021)) along with the Paris agreement commitment of a 50% cut compared to 1990-levels by 2030. The cuts laid out by the EU counts cuts for all of Europe, abatement projects can therefore be carried out in other countries. The Norwegian government has expressed in their own climate strategy that there will be a focus on cuts domestically (St. Meld. 13 (2020–2021)). But despite having had policies targeted towards curbing emissions for the better part of the last 30 years, Norway has failed to reach the goal of keeping its domestic non-EU Emissions Trading System (EU ETS) regulated emissions to 48 million tonnes of CO2-equivalents (MtCO2e) by 2020.

The failure of reaching past targets for cuts in emissions and the ambitious goals set for the future raises two questions. First, why did Norway fall short of our targets despite 30 years of targeted policy? This is fundamentally a political question, but it also has an empirical aspect. (Norwegian Environment Agency, 2022). Following the literature review carried out for this thesis, most research conducted on Norwegian GHG emissions the last 20 years seems to focus on consumption based (indirect) emissions (Huang & Bohne, 2012; Peters & Hertwich, 2006; Steen-Olsen et al., 2016). These are studies where emissions embodied in the production and consumption of materials through their entire lifecycle are factored in. There are few recent economy-wide studies on production based (direct) emissions in Norway which quantifies the effect of drivers. Furthermore, reports on past emissions analyses by public bodies tend to restrict themselves to analysing trends (Norwegian Environment Agency, 2020). Consequently, the extent of how factors such as economic growth, carbon intensity and sectoral balances affect direct emissions remains partly unexplored. This thesis seeks to provide answers on this area. For this purpose, a retrospective study can provide answers. The second question relates to how emissions might develop in the future. This is especially relevant given the ambition of cutting domestic emissions by 2030.

1.2 Research questions

A substantial part of Norwegian climate policy has been aimed towards reducing emissions abroad (Berg, 2015) through programmes such as NICFI, which is a programme that aims to stop deforestation of rain forests internationally, and purchases of ETS quotas as outlined by T. Moe (2012). But, as made clear by white paper 13 (St. Meld. 13 (2020–2021)) strong commitments will also be made domestically in the coming decade. As 92.4% of domestic emissions stem from economic activity (Statistics Norway, 2021), the remaining few percentages within households and the effects of land absorption on emissions (LULUCF) falls beyond the scope of the retrospective analysis. Index decomposition analysis provides a way to quantify the effect of drivers on past emissions in order to study what effect they have had and will be the main tool for the retrospective analysis.

Regarding future emissions this thesis will, by means of supervised machine learning methods, seek to develop a model which can accurately predict future direct emissions with a prediction horizon of 5 years. This second part of the analysis is carried out with the additional goal of identifying covariate importance in relation to the development of future emissions. The predictive power of machine learning depends on the amount of data points for prediction available (Halevy et al., 2009). So, while it would be an interesting endeavour to predict emissions on a national- or county-wide level, municipal GHG emissions make it possible to form the largest dataset and henceforth potentially better predictions. As the subject matter of this paper concerns past and future emissions, the research questions are split into corresponding categories. The research questions for the purposes of this study are:

RQ 1: Retrospective analysis
What are the main drivers of change in sectoral emissions in Norway between 1990 and 2019?

RQ 2 Machine learning prediction:
How are direct GHG emissions on a municipal level likely to develop over the next 5 years? As sub-questions, I ask the following

a. (How) can a supervised machine learning model architecture predict future emissions reliably?
b. Does the winning architecture add predictive power compared to a naïve model?
c. Which variables are the most significant with regards to predicting future greenhouse gas emissions?

To carry out the retrospective analysis emissions are decomposed into subsequent drivers of emissions where each factor is selected based reviews of prior studies. Contributions of each driver are then measured over time using logarithmic mean divisa index decomposition analysis (LMDI-IDA). The method will be applied on 10 sectors in order to determine how the activity, structure and intensity-measures within the economy have affected direct emissions in the period of 1990 - 2019.

The predictive part of this thesis is carried out by first identifying relevant covariates through a literature review. Then, an architecture for prediction methods will be devised, and deployed via Python's sci-kit learn library, which offers access to a rich toolbox for supervised machine learning. The algorithm which exhibits the lowest amount of prediction error will then be selected to forecast future GHG emissions for all municipalities in Norway with a horizon of 5 years.

1.3 Structure of the thesis

Chapter 2 in this thesis provides exploratory analysis on both sectoral emissions on a national level, and total emissions on a municipal level in order to reveal trends and structure in the data. It also reviews existing analyses on Norwegian emissions. This is done with the purpose of providing context to the analysis in subsequent chapters. Chapter 3 provides an in-depth review of methods applied to retrospective analysis of emissions and which method is most applicable given the data at hand. It also outlines how the LMDI-IDA method has been implemented in the context of this thesis. Chapter 4 discusses and reviews machine learning methods and their application to predicting emissions. It also outlines the method implemented for prediction of future emissions in Norwegian municipalities.
Chapter 5 discusses results from the retrospective and predictive analysis, whilst chapter 6 contains concluding remarks.

2 Past emissions and trends in Norway

In order to provide context for both the retrospective index decomposition analysis and the machine learning methods, it is necessary to investigate historical emissions. There are several ways of making emission accounts in order to attribute them to their source. Before looking at historical emissions it is necessary to understand what is being counted in the data. On the national level, Statistics Norway provides two principal ways of categorising emissions (Statistics Norway, 2015). The first is sectoral emissions from economic activities in Norway and it follows the same classifications as GDP and the national energy accounts. It is thus highly compatible with GDP and energy data for analysis purposes. The second method ties emissions to emitting activity such as the production of different metals, types of vehicles and even fermentation in production of beer. The way these two are allocated geographically are different. Emissions tied to economic sectors follow economic entities just like GDP and energy does. As such, it also includes emissions from Norwegian economic activity abroad. Emissions grouped by emitting source includes all emissions made within the geographical boundaries of Norway, including emissions from foreign entities. Municipal GHG emissions are based on the latter category. This difference in geographical allocation explains why the two types of accounts might be different in absolute terms and the sizes of different sectors.

2.1 Norwegian emissions attributed to economic sectors

As evident from figure 1, sectoral carbon emissions in Norway have for the last 30 years fluctuated between 60 000 and 75 000 mktCO2e. Figure 1 shows that the largest emitting sectors in Norway are transport, manufacturing, and mining and petroleum. In the year 2020 these three accounted for approximately 72% of total emissions. In terms of structure the balance between sectors has been quite stable over time, although manufacturing has gone from being the largest polluting sector in 1990 to being the third most polluting sector in 2020. At the same time the share of emissions from mining and petroleum has grown, especially in the period leading up to 2005 before starting to decrease somewhat after 2015. Transportation has for most of the last 30 years been the largest polluting sector, see figure 1.

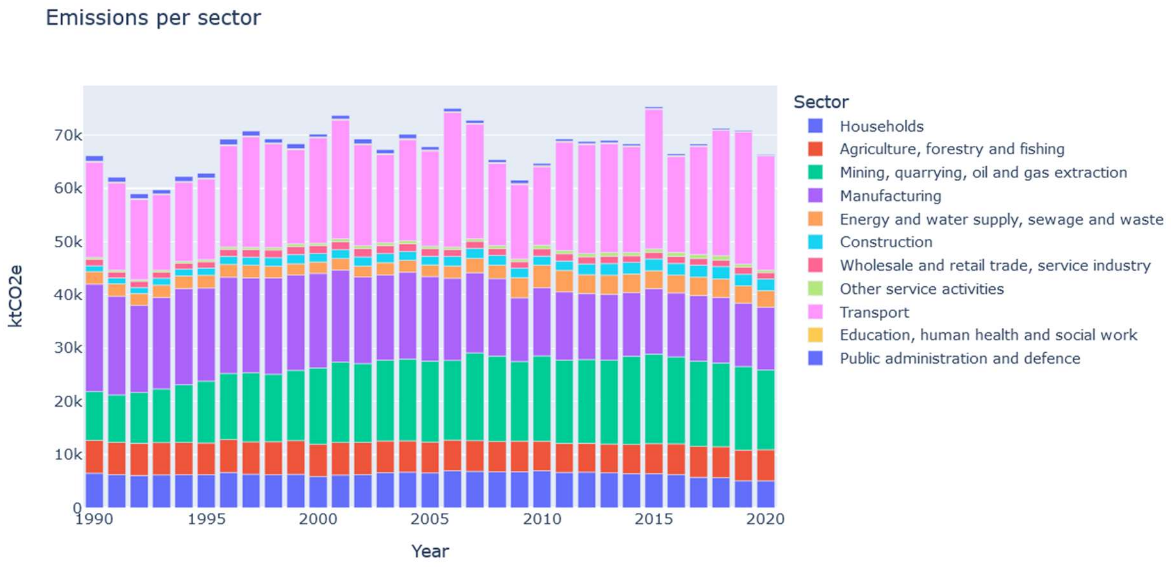


Figure 1 - Norwegian GHG emission per sector, millions of tons CO2e (SSB, 2021)

Looking at sub-sectors within the transportation sector in figure 2, two observations stand out. First, there is an upwards trend in emissions. Second, we see that the largest single source of emissions is ocean transport. Norway is, measured in terms of value, the fourth largest shipping nation in the world (Norwegian Shipowners' Association, 2021). Ocean transport accounted for in this sector thus includes emissions made abroad (Statistics Norway, 2015). Norwegian GDP and energy accounts are also adjusted to include this activity. Air transport is also another sector that is counted internationally, but subsidiaries registered abroad are left out. So, while Norway is home to large international flight companies such as Norwegian, they presumably register much of their activities abroad which do not count towards domestic emissions, despite air travel being a very carbon-intensive mode of transportation. This could explain while emissions from air travel are so low compared to sea transport. Another aspect to note about air transport is that the emissions from the sector were growing in the period 2011-2019 before lowering drastically in 2020. The latter is the effect of covid-19 related lockdowns and travelling restrictions. One last thing to note from emissions within subsectors of transportation is that emissions stemming from land transport are larger in 2020 than they were in 1990, despite the recent surge in adoption electric vehicles in private households (Statistics Norway, 2022b).

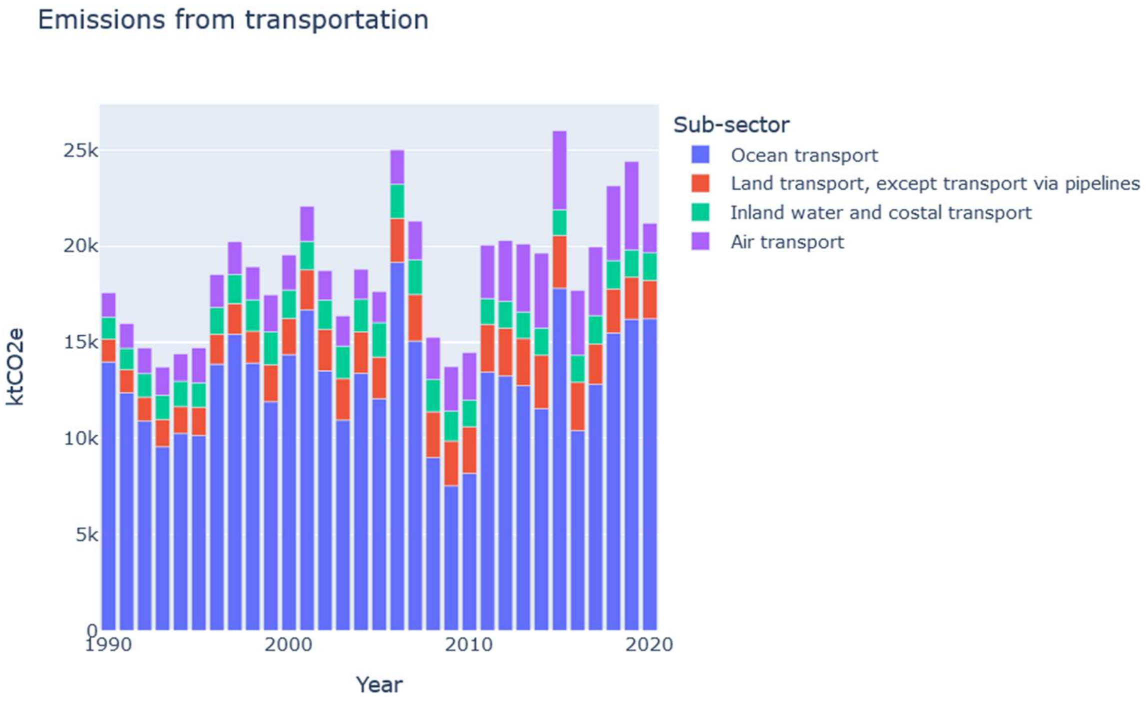


Figure 2 - Sub-sectoral emissions from the transport sector (SSB, 2021)

As evident from figure 3, in manufacturing, there has been a decline in emissions of almost 10 000 MKtCO2e. This is partially due to the production of basic metals emitting less from 2009 onwards. Refined petroleum products and pharmaceuticals has also decreased its emissions from over 6000 MKtCO$_2$ in 1990 to approximately 4700 MKtCO2e in 2020 thus lowering emissions. There has also been a steady decline in emissions from the production of wood products since the mid 90's.

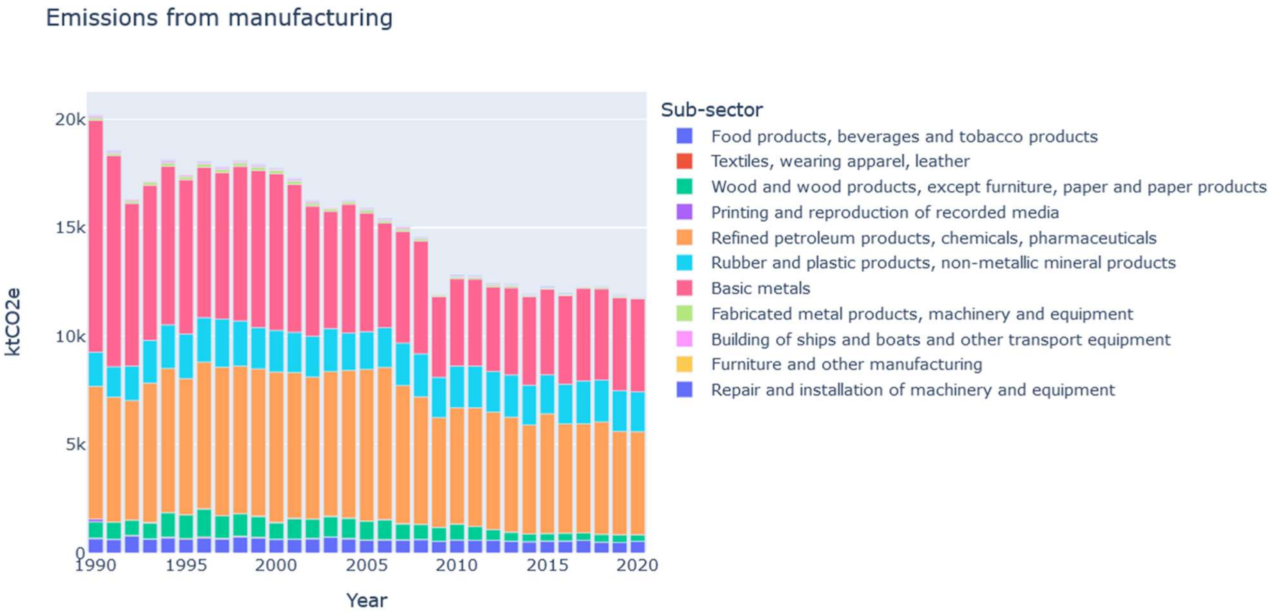


Figure 3 - Sub-sector emissions from manufacturing (SSB, 2021)

The last sub-sector of note is mining and petroleum extraction. Here emissions rose sharply in the 90's, from 8700 MKtCO2e in 1990 to over 14 000 MKtCO2e in 2000 and have stayed above that number ever since. Emissions reached a peak in 2015 at near 16 400 MKtCO2e and have towards 2020 decreased towards the 14-mark. The sector is completely dominated by the emissions from oil and gas extraction, accounting for nearly all of the sector's emissions, see figure 4.

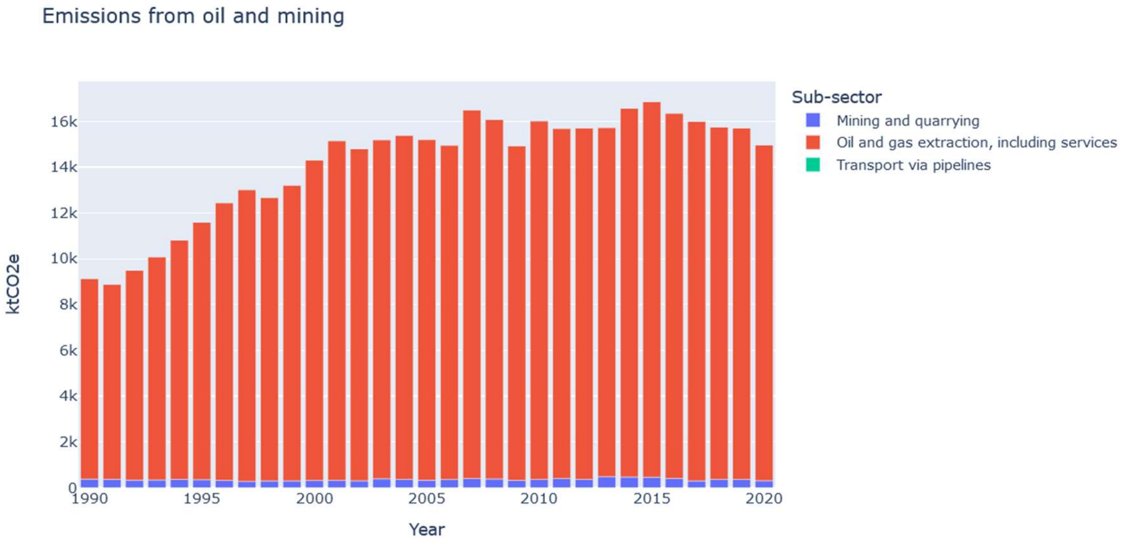


Figure 4 - Sub-sectoral emissions from mining, oil and gas extraction (SSB, 2021)

2.2 Municipal GHG emissions

The municipal GHG-emission accounts go back to 2009. At first, they were compiled by Statistics Norway, but have since 2015 been provided by Norwegian Environmental Agency. The emissions are, as mentioned, categorised according to emitting activity (Jacobsen & Lillesund, 2021). The accounts also contain economic sectors, but following correspondence with the Norwegian Environmental Agency (Seim, 2022), these sectors are differently categorised from the GDP accounts. Therefore, there is a different sectoral breakdown compared to national accounts. Geographically, emissions are allocated to the municipality they occur in. This means that activities offshore and aviation beyond a flight altitude of 914,4 meters are not included.

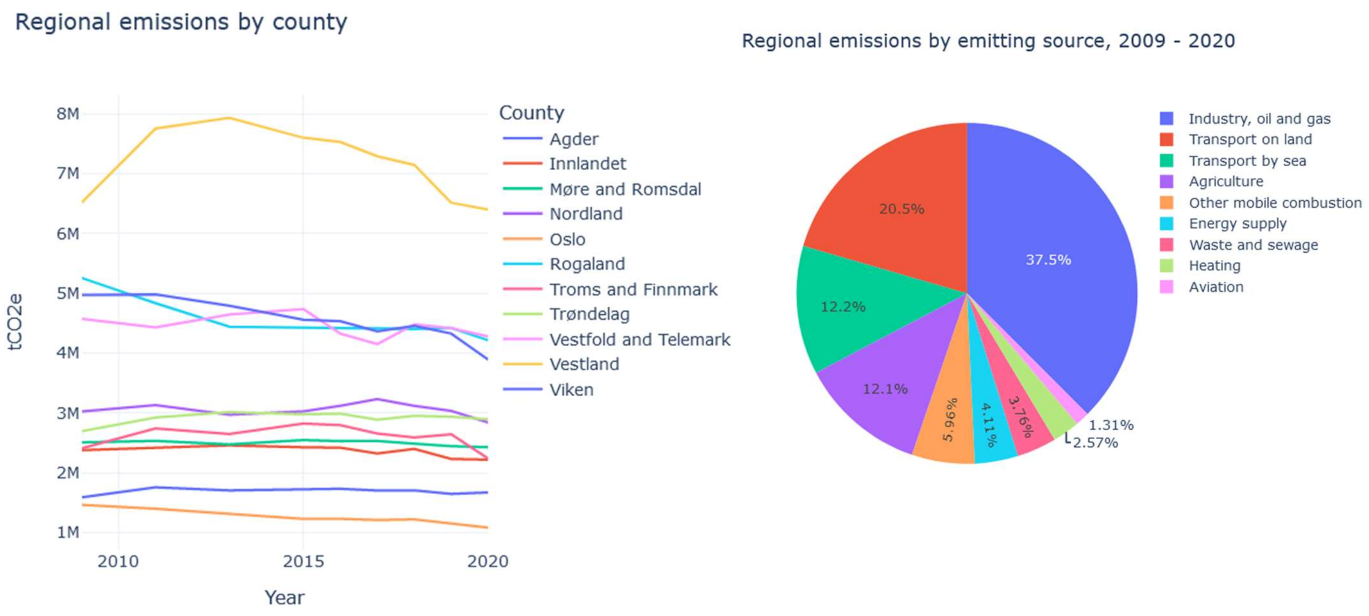


Figure 5 - Regional emissions by county and sector (Norwegian Environmental Agency, 2021)

The two largest polluting sectors in municipalities as a whole is the sector named "industry, oil and gas" along with transport on land which account for more than 50% of emissions since 2009. Runners up are the agriculture and sea transport sectors which contribute 12.1% and 12.2% respectively. In terms of development over time there seems to be a weakly decreasing trend in emissions across several regions, especially in Oslo, Viken, Rogaland, and Vestfold and Telemark. In the region of Vestland this trend is much stronger from 2013 onwards. There are also large regional differences in terms of the scale of emissions from municipalities. A scatterplot analysis reveals that some of this heterogeneity stems from municipalities that are home to emission intensive-industry. In the case of Vestland, Alver municipality, which is home to a large petroleum refinery, emits over a third of the county's total amount.

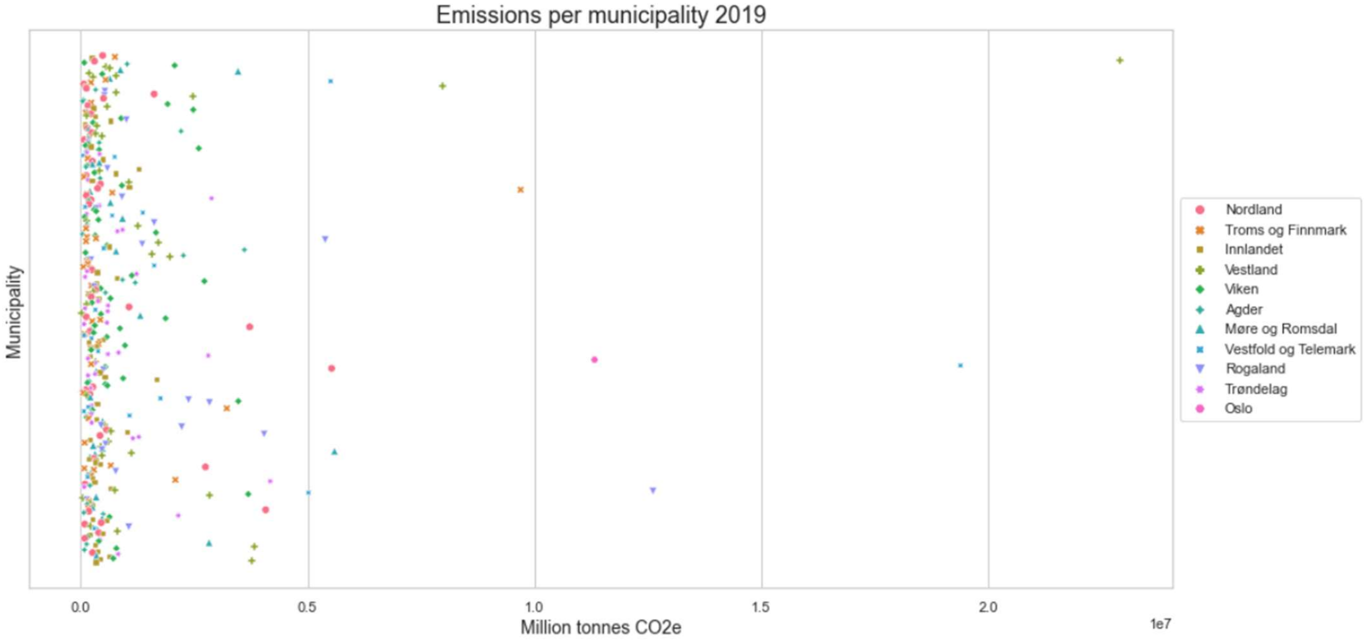


Figure 6 - Scatter plot of municipal emissions

2.3 Retrospective analysis of GHG-emissions in Norway

So far, it's evident that emissions in the Norwegian economy is subject to inertias, increases and reductions. With climate change being one of the most important issues of our time there is much research related to economic activity and emissions. With data from Norway being readily available there are several examples of studies which analyses direct emissions and quantifies drivers as part of an international analysis. Andreoni and Galmarini (2012) analyse emissions from sea- and air-transport from 14 countries, including Norway, and finds that increased economic activity has led to higher emissions in both sectors across the studied countries. Moutinho et al. (2018) studies Norwegian emissions as part of an international decomposition analysis featuring 23 countries through the period of 1985 to 2011. They find that the fossil intensity of fuels has on an aggregate level been falling steadily in the period, while the share of renewables has been increasing, thus creating downward pressure on emissions. However economic growth has to some degree counteracted the abatement. The same results can be found in Xu and Ang (2013) and a much earlier study by Hamilton and Turton (2002).

On the sub-national level the national inventory reports provide a thorough review of trends and developments (Norwegian Environment Agency, 2020). It shows that the diminishing emissions from sectors such as manufacturing, waste and industrial processes are offset by increased emissions in the petroleum and transport sectors. The reason stated for the increased emissions from the petroleum industry is "[…] explained by the increase of oil and gas production and the increase of energy demand in extraction, due to aging of oil fields and transition from oil to gas" (Norwegian Environment Agency, 2020, p. 38). While this report yields very good qualitative reasoning on why emissions change in its respective sector it does not yield any answers on the magnitude of the effects listed.

Other research related to Norwegian emissions seems to mainly focus on indirect emissions. That is, emissions embodied in consumption or usage of materials where impacts from the lifecycle of material flows are included. Yamakawa and Peters (2011) provide a structural decomposition analysis of Norwegian greenhouse gas emissions embodied in imports, exports and consumption for the time-period 1990 – 2002. Their results show that "70% of the growth in Norway's energy consumption and greenhouse gas emissions was caused by the production of exported products, in particular oil and gas production". Their study shows how Norway should make cuts in their petroleum-exporting industries if they are to contribute positively towards reaching stated goals for global emission cuts. As their main focus is on imports and exports, they group domestic sectors together. Hence, this study doesn't show how structural changes between economic sectors affect emissions over time domestically. Further studies documenting indirect emissions embodied in trade (Peters & Hertwich, 2006) and household consumption (Steen-Olsen et al., 2016) have also been carried out for the year 2000 and 1999 – 2012 respectively. Both show how indirect emissions were increasing for the time-period studied. Further studies analysing sector specific indirect emissions can be found in Huang and Bohne (2012), Ziegler et al. (2013), Hertwich and Roux (2011) and Sparrevik and Utstøl (2020). All these show how different drivers, economic activity and material consumption contributes towards indirect emissions in their respective sector.

Direct emissions seem to be a more sparsely studied area, at least in terms of retrospective analysis. Gavenas et al. (2015) studies direct emissions from the petroleum sector and find that emissions per unit of extraction rises as fields deplete. Simonsen et al. (2019) studies the cruise industry and show that 11.4% of their emissions happen in municipal ports. Aamaas (2019); Korsbakken (2020) and Korsbakken (2021) provide analysis which aims to create future projections of municipal emissions in Oslo, Bergen and Kristiansand. The models used in these studies decompose emissions across all sectors but use these to project future emissions according to given policy scenarios. As such they chart the tentative effects of drivers on emissions. Naturally, since these studies are carried out on a municipal level, they convey little information about the national level (nor were they meant to).

The only sector-wide study which decomposes and analyses the effect of drivers on emissions historically is Bruvoll and Larsen (2004). They decompose emissions into 9 different drivers across 8 different sectors. They find that emissions increased 15.5% in the time period studied. This was due to economic growth, population growth and structural change in the economy outweighing the efficiency gains from energy mixture, energy intensity and improvements in production methods. Overall, they also identified that emissions were falling per unit of GDP. The period studied was 1990 – 1999.

The literature review carried out for this thesis has only identified a few sector-specific studies on direct emissions (Gavenas et al., 2015). The only sectoral decomposition of direct greenhouse gas emissions seems to be Bruvoll and Larsen (2004). Their study provides a decomposition of emissions for the period 1990 – 1999 along with a general equilibrium simulation to identify the effect of carbon taxes.

The relatively small body of knowledge regarding how drivers of emissions have affected direct emissions within Norway the last 20 years should give pause. Questions regarding how shifts in economic structure, activity and efficiency of fossil fuel-use could benefit from more thorough documentation.

3 Retrospective analysis: Method and data

3.1 On retrospective analysis of emissions – a broad overview of viable methodologies

Retrospective analysis of GHG emissions is a diverse field with several of methodologies. In their literature review charting research related to greenhouse gas emissions and economic growth Mardani et al. (2019) identify several constellations of commonly used methodologies. These are Structural Decomposition Analysis (SDA), regression-based methods and Index Decomposition Analysis (IDA). Each of these have their respective strengths and weaknesses as well as data requirements. They are not mutually exclusive methodologies, as they can complement each other. Metcalf (2008)provides an evocative example of combining all three IDA and regression-based methods to identify the magnitude of drivers of emissions, and then which factors affect these drivers the most. For the sake of simplicity, a broad overview of the three methodologies and their application to retrospective analysis of emissions in isolation will be provided below.

Table 1 - Overview of methods used for quantification of drivers on emissions

	Structural decomposition analysis	*Regression-based analysis (STIRPAT)*	*Index Decomposition analysis*
Object of study	Indirect emissions and direct emissions	Direct emissions	Direct emissions
Estimator/technique	Non-parametric index & Leontief inverse	Regression methods	Non-parametric index
Data requirements	Input-output tables, emission factors	N << p	Data on drivers and emissions for at least two time periods

SDA-based methods are non-parametric and trace material flows both nationally and internationally. de Boer and Rodrigues (2020) point out that it is a method in which the link between impact and consumption activities is explored as the method can account for indirect emissions embodied in consumption of materials. As shown in chapter 3, Peters and Hertwich (2006) is an example of such analysis applied to Norway. Studies using these methods often emphasise the considerable effect of emissions embodied in high consumption of material- and energy-intensive goods

Regression based methods are widely used across many scientific fields and perhaps especially within the field of economics. Common for all is that they abide by the rules and conditions of their respective estimation technique. Broadly speaking, the majority of these are based on the law of large numbers (Hsu & Robbins, 1947) and asymptotic convergence of estimators (Wooldridge, 2013). According to York et al. (2003) the STIRPAT model is a common way of estimating the effect of drivers using conventional regression methods. Here, the IPAT identity, which according to Harrison and Pierce (2000, p. 7) stipulates that emissions (impact, I) are the product of population (P), affluence (A) and technology (T), is adapted into a structural equation for which parameters can be estimated. There are many more such models including various studies centred around proving or disproving the Environmental Kuznets Curve (Stern, 2003) as well as ex-post studies of effects related to environmental taxation.

Index decomposition analysis is the third methodology considered. According to Xu and Ang (2013) there were at least 80 papers published between 1991 and 2012. Further examples of usage can be found in Le Quéré et al. (2019), Trotta (2020) and O' Mahony et al. (2012). The latter two studies were applied on a sectoral level in Finland and Ireland in order to chart energy efficiency gains and contributors to carbon emissions. Subsequently they provide good proofs of concept for the purposes of this thesis. The method was initially developed to decompose and analyse changes in energy usage (Ang, 2004), as such that is also a field in which many studies using the methodology can be found in. Having said that, there are studies analysing drivers of emissions on an international level using the method as early as 1991 (Torvanger, 1991).

Fundamentally, the method allows for decomposition of an aggregate measure, like greenhouse gas emissions, into subsequent drivers and then measurement of how these affect the aggregate measure. The decomposition is first carried out via decomposing the aggregate through putting the drivers into an identity, then different kinds of index theory methods are applied to identify changes in drivers over time and how much they affect the aggregate measure. In their literary review de Boer and Rodrigues (2020) highlight how these methods are also used in the consumer price index, producer price index, producer price index and human development index. According to their article the earliest recorded use of an index analysis of price changes is the French economist Dutot who in 1738 looked at the price changes of several commodities vs incomes and concluded Louis XV was worse off compared to his ancestor Louis XII based on the difference in factors.

3.2 Why LMDI-IDA?

The task at hand is to decompose emissions and quantify the effect of drivers. For this we have seen that there are three separate families of methods that can help achieve this purpose. However, SDA methods are dependent on input-output data (Hoekstra & van den Bergh, 2003). These tables are only available from Statistics Norway in the same format in the period between 2012 and 2019. In relation to identifying the effect of drivers of emissions by using regression methods, we arguably face a lack of data. F. Harrell (2015) writes the following on page 72 about sample size: "[…] in many situations a fitted regression model is likely to be reliable when the number of predictors (or candidate predictors if using variable selection) p is less than m/10 or m/20, where m is the limiting sample size". The period 1990 – 2019 makes for 29 observations. At best this would make for a model with less than 3 variables if reliable results are to be considered, provided there are no multicollinearity issues. This is too restrictive in terms of modelling drivers of sectoral greenhouse gas emissions and not a viable methodology.

Hoekstra and van den Bergh (2003) as well as de Boer and Rodrigues (2020), point to how index decomposition analysis methods can provide analysis with few datapoints. In fact, the method only needs data from two points in time to work. If data exists over an interval of time, it is also possible to chain the analysis and sum the changes in order to get the effect of drivers. Ang and Goh (2019, p. 836) stress that chaining the analysis is always preferred because it reveals the year-on year changes and thus provides the best results. Unlike regression-based methods, having only 29 observations is not an obstacle for IDA-based methods. Thus, it is a suitable methodology for retrospective analysis of Norwegian greenhouse gas emissions.

3.3 On selection of index method, data and drivers

While the adequacy of index decomposition analysis for retrospective analysis of emissions has been established the matter of identifying the most suitable index method needs to be settled. There are numerous indexes available and Ang (2004) identifies the most commonly used index methods as Shapley-Sun, Logarithmic-mean divisia (LMDI), Arithmetic mean divisia, modified Fischer and Marshall Edgeworth. Between these index methods the multiplicative and additive version of LMDI come out on top based on the fact that they yield decompositions without residuals, yield easily interpretable results, can be adapted to handle zero-values and have *consistency in aggregation*. This means that year-on year results can be aggregated safely, and that results can be interpreted both on a sectoral level and an aggregate level. The only real drawback with the index method is that it is not robust against negative values. But in the context of decomposition of emissions, negative values are very rare as emissions, energy and GDP-data are always positive numbers. As such the LMDI decomposition method is the most suited for retrospective analysis of Norwegian emissions. The additive and multiplicative versions of the LMDI index have the same properties, aside from that the additive version yields results in the same unit as the aggregate size studied, while the multiplicative version yields percentages. Getting the results in terms of emissions have several advantages over percentages. First, it becomes much easier to check if there is a residual present in the results. Second, if the need arises for interpreting result as percentages, they can easily be converted.

3.3.1 Data and selection of drivers of emissions

Conceptually there is no limit to the number of drivers applied within an LMDI index decomposition analysis (LMDI-IDA) as long as the drivers can form an identity equation. However, which drivers to choose ultimately depends on the problem at hand and data availability. This study analyses sectoral emissions, hence, there’s a requirement that all other data can be separated into the same sectors as the emissions data. While energy and emissions data in Norway follow the same sectoral breakdown, sectors by GDP are categorised slightly differently. Fortunately, Statistics Norway provides correspondence-tables that associate the sectors in the energy, gdp and emissions data (Statistics Norway, 2009). It’s possible to categorise all three in such a way that they can be used together.

In their literature review on the application of LMDI-IDA methods applied to emissions, Xu and Ang (2013) commonly used drivers as overall activity in the economy (GDP), economic structure, energy intensity, fossil share of energy, and GHG-intensity. Equation 3.1 shows how this can be done on a sectoral level:

$$C \equiv \sum_i C_i \equiv \sum_i \sum_j A \frac{A_i}{A} \frac{E_i}{A_i} \frac{E_{ij}}{E_i} \frac{C_{ij}}{E_{ij}} \equiv \sum_i \sum_j A \times S_i \times EI_i \times F_{ij} \times U_{ij} \qquad (3.1)$$

Where C is aggregate emissions, i indicates sector, j fossil fuel type, C_i is emissions per sector, A is the activity level (total GDP), $S_i \left(= \frac{A_i}{A}\right)$ is sectoral GDP, EI_i $\left(= \frac{E_i}{A_i}\right)$ is sectoral energy use, F_{ij} $\left(= \frac{E_{ij}}{E_i}\right)$ is fossil energy use of type j, and U_{ij} $\left(= \frac{C_{ij}}{E_{ij}}\right)$ is the carbon intensity of fossil fuel j. As such, the model is able to capture the effect of shifts in sectoral balances (structural change), how much energy is consumed per economic output, the fossil share of energy used and how GHG-intense the fossil usage of energy is in the sector. The method implicitly assumes that emissions are the result of fossil fuels, this is a weakness when considering industries like agriculture where the emissions stem from animals. But as the data on sectoral emissions in chapter 2 have shown, most emissions stem from CO2, so this assumption is more correct than it is wrong.

Table 2 - Data used for LMDI-IDA analysis

Variable	Sectoral breakdown at collection	Unit
GDP	61	NOK, Value added at basic prices
Total energy use per sector	31	GWh, consumption for energy purposes
Fossil energy use per sector	31	GWh, consumption for energy purposes
Total emissions per sector	31	Megaton CO2e

The data on sectoral GDP (Statistics Norway, 2022a) sectoral energy use (Staistics Norway, 2021) and sectoral emissions (SSB, 2021) are all available for the time period 1990 – 2019. The GDP data uses real prices with 2015 as its reference year and measures “Value added and gross income generated from domestic production in an industry or sector […]”

(Statistics Norway, 2014). As mentioned in chapter 2, sectoral emissions include international shipping and flights from companies register in Norway. This is because it is not possible to separate domestic air transport and shipping from international the period 1990 – 2000 due to energy data not being available for this period. Furthermore, they are also an integral part of Norwegian economic activity as they operate out of Norway and are thus regulated by Norwegian policies. Thus, they were included.

Energy accounts are complex. This is because crude oil has more uses than just energy. As such consumption of energy have different categories in the Norwegian energy accounts. Consumption for energy purposes and total consumption, which includes petroleum turned into other products like plastics and so forth. For this study, consumption for energy purposes was chosen, as the petroleum only turns into a GHG once its combusted. The fossil energy category includes energy consumed from coal and its derivatives, natural gas, and crude oil. As waste is estimated to contain 20% plastics when its combusted, 20 % of the energy derived from waste is counted as fossil. Total energy consumption of energy products includes all energy consumed.

3.4 Application of additive LMDI-IDA

The additive LMDI-IDA is outlined by Ang (2015). It is a method which allows the decomposition of an aggregate measure into factors or drivers and then assess their impact over time. In an additive decomposition change in emissions (C) from year to year will then be expressed as such:

$$\Delta C_{Tot} = C^{T} - C^{0} = \Delta C_{factor\ 1} + \cdots + \Delta C_{factor\ n} \qquad 3.2$$

And the identity shown in 3.1 given the data selected is

$$ktCO2e_{s,t_{s,t}} \equiv GDP_t \frac{GDP_{s,t}}{GDP_t} \frac{GWh_{s,t}}{GDP_{s,t}} \frac{fGWh_{s,t}}{GWh_{s,t}} \frac{ktCO2e_{s,t}}{fGWh_{s,t}} \qquad 3.3$$

s = sector, t = time, C = emissions, fGWh = fossil energy use; measured in GWh, GWh = total energy use, measured in GWh

The generalised form for calculating the effect of drivers within a sector between two time periods can be written as:

$$\Delta C_i = \sum_i \left(\frac{C_i^T - C_i^0}{ln\ C_i^T - ln\ C_i^0} \right) \ln \left(\frac{F_{i,n}^T}{F_{i,n}^0} \right) \qquad 3.4$$

Where i indicates sector, C is emissions, F is factor and n is factor number and the superscripts T and 0 indicate time. Results will be in the number of units the individual factor adds or removes from the contributor from period to period. Results can then be graphed or put in tables for each sector. Xu and Ang (2013) provide a full overview of how this framework has been applied to studies in relation to emission studies.

3.5 Limitations of LMDI-IDA

- The method uses gdp -> inherits all the problems with it. It isn't really "output" we measure but we treat it as such. I.e. monetary measures are only a proxy of the actual volumes of materials and services
- Current model implicitly assumes that all emissions that come from economic activity are due to energy use. That isn't strictly true,
- Treats the Norwegian economy as a closed system. Structural shifts in and out of the country doesn't show up. This is somewhat alleviated that we're only looking at energy consumption but disregard energy production.
- Lumped all fossils together in one category, so results only give the degree of GWh that are derived from fossil fuels in general, and not per fuel. This was done for ease of interpretation of results.

4 Supervised machine learning: Method and data

Athey (2019) summarises supervised machine learning as a method which uses "a set of features or covariates (X) to predict an outcome (Y)". This could mean forecasting, or it could imply prediction of time-static phenomena. It is called *supervised* learning because the outcome variable Y guides the learning process which consists of describing associations and patterns in the input measures X, as well as how they relate to Y (Hastie et al., 2009).

In machine learning the goal is to minimise prediction error whilst treating the functional relationship of what is modelled as unknown (Breiman, 2001b). Strategies and models are then devised and deployed algorithmically in order to find the best compromise between bias and variance thus the best goodness of fit. It does not seek to estimate the true parameters of covariates like in econometrics. Put differently, machine learning methods can tell us whether a set of covariates adequately predict an outcome, and which model yields the best approximation of functional relationships. It is important to note that the functional relationship need not be true in the real world and that this restricts what we can learn from machine learning methods. The relationship to the real world ultimately depends on the data fed to the algorithm. For example, Athey (2019) points to how pianos can be reliable predictors of the presence of cats in pictures if the training & validation datasets associate the two. But that pianos have next to zero causal effect on turning animals in households into cats. So, to hammer this home, the goal of machine learning methods is not to find the true effect of a parameter, nor its causal powers. It finds predictors, and the robustness of these are ultimately a function of the model and the data used. Machine learning models might associate cats and pianos because people find it funny to have their cats walk on pianos. Not because pianos cause animals in their proximity to become cats. But, despite the possibility of relationships between factors in machine learning models being spurious, they might still tell us something. In the example of cats, it tells us that people like to take pictures of them in proximity to pianos – and from that we can learn something. So, while the importance of variables in a model making predictions of future greenhouse gas emissions might be spurious, they tell us that they have predictive power and are thus worthy of further investigation. This is a different approach to conventional econometric methods which seeks to estimate *ceteris paribus* (all else being equal) effects of covariates upon a dependent variable. Or put in terms of cats, the goal of a conventional econometric inquiry could for example involve estimating the propensity of pianos causing cat-ness in nearby animals.

The goal of the supervised machine learning methods employed in thesis is not to implement an estimation technique for measuring the effect of given parameters or policies, like in Bruvoll and Larsen (2004). Instead, it seeks to provide solid predictions, and to disclose which factors that have the strongest predictive power in order to forecast future emissions in Norwegian municipalities. There are potentially large differences between municipalities due to differences in geography, demography and economic structure. This could lead to different functional relationships in factors that predict future emissions. Machine learning methods specialises in dealing with complex functional relationships. As such it could yield highly accurate predictions. And while the functional relationship of variables might just be an approximation of real-world relationships, supervised machine learning methods could at the very least tell us which relationships the variables might *not* have. For example, if the functional relationship of factors proves to be poorly predicted by linear models, but well by models that are more robust to non-linearities and scores of interaction terms, then it is likely that the functional relationship in the data is not linear.

4.1 Machine learning and emissions prediction

There are several examples of studies employing supervised machine learning methods for prediction of localised greenhouse gas emissions from buildings and industrial processes. Examples can be found the literature reviews of Seyedzadeh et al. (2018) and Adams et al. (2020). Studies that seek to make predictions related to greenhouse gas emissions on regional or aggregate level seem to be fewer in number. Among these are Mardani et al. (2020) which employ artificial neural networks along with clustering techniques in order to predict emissions of the G20 nations for the same year. They find that their methodology can predict with a mean absolute error (MAE) of 0,065 thus making their method a viable tool for estimation of current emissions. Acheampong and Boateng (2019) forecast carbon intensities per unit of fossil energy consumed for 9 countries using a neural network and macroeconomic variables such as energy consumption, economic growth, financial development index, population and trademark applications. They use mean square error (MSE) as their prediction error measure and the model returns prediction errors in the order of $MSE < 0.01$. The MSE measure returns the prediction error as the squared value of the independent variable. In this study the CO2 intensity is given in kg per kg oil consumed. With such a low prediction error this goes to show how accurate predictions could be when using machine learning methods. Wei et al. (2018) provides the most interesting results applicable to this study. The study makes forecasts for the region of . Like other studies, their results are highly accurate, reporting a prediction error (RMSE) of only 0.002. Since the target variable in their study is 10 ktCO2e this amounts to an error 20 tonnes of CO2e compared to actual emissions. Their method uses more complex models than what is used in this thesis. Hence an explanation on how the models employed in their study work will not be provided here. Explanations of the models used by Wei et al. (2018), can be found in Huang et al. (2006), which gives an introduction to Extreme Learning Machines, Breiman (2001a), which outlines random forest models and Hussien et al. (2020), which systematically reviews the application of moth flame optimisation algorithms. For an introduction to neural networks in general, see Hastie et al. (2009, p. 389). The reasons for Wei et al. being relevant for the research problems in this thesis are the following:

1. The study provides a good overview of how machine learning methods can be used in selection of input variables from their dataset, and how different selection algorithms can improve the performance of their predictive models.
2. Their predictions vs the actual emissions are shown over a time period of 5 years where there is a change in trend of the actual emissions. Hence, giving an intuitive understanding of the quality of the predictions of future emissions. And how their models can predict changes in trend.
3. The study provides a detailed overview of variable importance for predictions. This gives an indication of which variables could be useful for predictions elsewhere.
4. Lastly, they provide a thorough presentation of their process and the order in which the models where implemented, thus making it easier to replicate similar process flows.

The variables in used in their study include public spending on infrastructure, sectoral GDP, population density, numbers of vehicles, R&D spending, coal and gas usage, production outputs and consumption data. The most reliable predictors in their study are shown to be total investments in fixed assets, sectoral GDP balances and population densities (Wei et al., 2018, p. 28997, tab. 3). The results from this table were used as inspiration in building the dataset for the predictions made in this thesis, more on that in chapter 4.3. While there are structural differences between the economy of Hebei and Norway there are also similarities. Roads are still used for driving, fossil fuels are still used as an energy source, and goods are transported for many of the same reasons in both regions. Consequently, variables should also carry across. Furthermore, the inclusion of machine learning methods for variable selection makes sure that only variables with the highest predictive outcome are selected for the predictive analysis. This solves the problem of using trial and error for variable selection and can improve model performance in scenarios where the predictive power of input variables is unknown.

There is one final thing worth discussing in the results from Wei et al.'s study. The variable importances are given in terms of mean decrease in gini impurity measures, which indicate how important a variable is in a random-forest model for making predictions (for an extended note on variable selection in random forest models, see appendix X). This measure only shows how important variables are for predictions in the given model and not their predictive power in general. It is notable that many variables have similar scores. In machine learning this could mean that many variables have weak predictive power and that robust predictions emerge from how well model manages to summarise and learn from weak predictors. The modelling strategy proposed in this thesis operates on this assumption, and this has led to the inclusion of a model designed to deal with this issue, namely XG boost.

It is evident that machine learning can provide accurate predictions of future emissions. While the number of studies existing on this area are few, the findings are promising. Machine learning can make robust predictions where the functional relationships between variables are unknown and can deal with scenarios where the predictive power of variables are potentially unknown. To this date there are no studies that have tried to apply machine learning methods in making predictions of future emissions in Norway. And while the level of this analysis was chosen based on the number of available data points for construction of the largest possible dataset, it has the added bonus of being a tool for forecasting emissions across all municipalities in Norway. In other words, if the predictions from the models used in this study are sound, it can function as a prototype for other, more extensive frameworks making predictions of future emissions, even in municipalities which lack the expertise for doing this manually.

4.2 Dataset, characteristics and created features

Raw data	**Variable name**	**Area**	**Unit**	**Source statistic, (SN = Statistics Norway)**
Surface area of agricultural land	acre_ag_area	Agriculture	Acres	SN table 06447
Surface area of buildings under, and finished construction	compl area nonhousing, constr area non housing	Construction	m2	SN table 05939
Number of existing buildings, excl housing	buildings	Construction	q	SN table 03173
Number of housing units	n_houses	Construction	q	SN table 03175
Population	Population	Demography	q	SN table 11342
Population density	pop/km2	Demography	pop/km2	SN table 11342
Agricultural land in use	acre_ag_area	Demography	Acres (?)	SN table 11342

Population growth	Population growth	Demography	%	Derived from SN table 11342
Imputation dummy, electricity. Indicates imputation where necessary in 2009	el_imp	Dummy	n.a	n.a
Imputation dummy, indicating imputed years in emissions data	em_imp	Dummy	n.a	n.a
Dummy indicating presence of large commercial port in municipality	port	Dummy	n.a	n.a
Employees in primary sector	Employees in primary sector	Economic measure of activity	q	SN table 07984
Employees in secondary sector	Employees in secondary sector	Economic measure of activity	q	SN table 07984
Employees in tertiary sector	Employees in tertiary sector	Economic measure of activity	q	SN table 07984
Municipal emissions in time t+5 (target variable)	tCO2e_t5	Emissions	log(tCO2e)	Norwegian Environmental Agency, Municipal emissions
Municipal emissions for time t and before	past emissions l_tCO2e	Emissions	log(tCO2e)	Norwegian Environmental Agency, Municipal emissions
Electicity consumption for 4 user groups. Mining & industry, households and agriculture, service industry, and in total	GWh usage total, GWh Mining and industry, GWh Service-sector, GWh Households and agriculture,	Energy	GWh	SN table 10314
Gross investments in public roads NOK per km	roadinv_NOK/km	Transport	NOK	SN table 11816
Shipping freight received and shipped pr port, international, domestic and total	sea freight dom (tons), sea freight int (tons), tot freight (tons)	Transport	Tons	SN table 03648
Number of ships arriving in ports per year per ship type & total	n [ship type]	Transport	q	SN table 08203
Tonnage of ships arriving in ports per year per type & total	tonnage [shiptype]	Transport	Ton	SN table 09518

Amount of registered vehichles per vehicle type. All vehicle types	n vehicles: [vehicle type]	Transport	q	SN table 07849
Amount of registered vehichles per fueltype. All fueltypes	n vehicle fueltype: [fueltype]	Transport	q	SN table 07849
% of electric vehicles registered	fuel: electricity %	Transport	%	Derived from SN 07849

The municipal reorganisation in 2020 complicated building the main dataset. While the emissions data for municipalities follows the new organisation with 365 municipalities, not all time-series from Statistics Norway have been merged into their new municipal units. To overcome this problem, a list from the Norwegian Mapping Authority was used (Norwegian Mapping Authority, 2020). This list shows which municipalities were merged and their new names post-merging and made it possible to merge old municipalities into new municipalities. Municipalities which were split had to be left out of the analysis, fortunately this only applied to 2 municipalities (Tysfjord and Snillfjord).

The master dataset contains 54 variables on 349 municipalities. The municipalities of Hammarøy, Tolga-Os, Hitra, and Narvik, Steinkjer and Heim lacked data to such a degree across many variables and were as such dropped from the dataset. The created features for this dataset were growth measures of existing data. Population growth in each municipality was derived from year-on-year changes in population and the number of electrical vehicles was derived from data on vehicle fuel types. Two dummy variables indicating where linear interpolation has been applied to fill in missing data. These two are applicable for municipal emissions and electricity consumption. Imputed variables are applicable for the years 2010, 2012 and 2014 for emissions and 2009 for electricity consumption.

4.2.1 Treatment of missing data

The emissions data from the Norwegian Environmental Agency lacks data for the years 2010, 2012 and 2014. To deal with this problem linear interpolation was used in order to create new datapoints. The reasoning behind using linear interpolation is that the emissions data in Norwegian municipalities seems on average to have a slow rate of change. Furthermore, since there are datapoints around the missing data, linear interpolation at least captures the trend in the data. In addition, a dummy variable was created indicating which years were imputed.

Missing data on the surface area allocated to agriculture was found in 16 municipalities. Most of the missing data were for singular years in timeseries with little change. These were imputed with linear interpolation. There were 4 municipalities which had no data for the entire time series, these were Berlevåg, Måsøy, Træna and Fedje. The amount of farmland here was assumed to be the average of all farmlands across Norway.

The source data on public expenditure per km of road from Statistics Norway is split two time-series. One pre-2015, and one post 2015. The time series overlap for the years 2015-16. These two had to be merged in order to be usable. Furthermore, the data doesn't follow the new organisation of municipalities, and so old municipalities had to be merged into new municipalities as well. It became evident that 14 municipalities suffered from error in reporting their data to Statistics Norway. These municipalities were found to only have data in either one or the other series. These were extracted and treated such that their expenditure from the old or the new time series (depending on where the data was reported) and merged into the usable time series in the master dataset.

4.3 Model architecture and methods applied

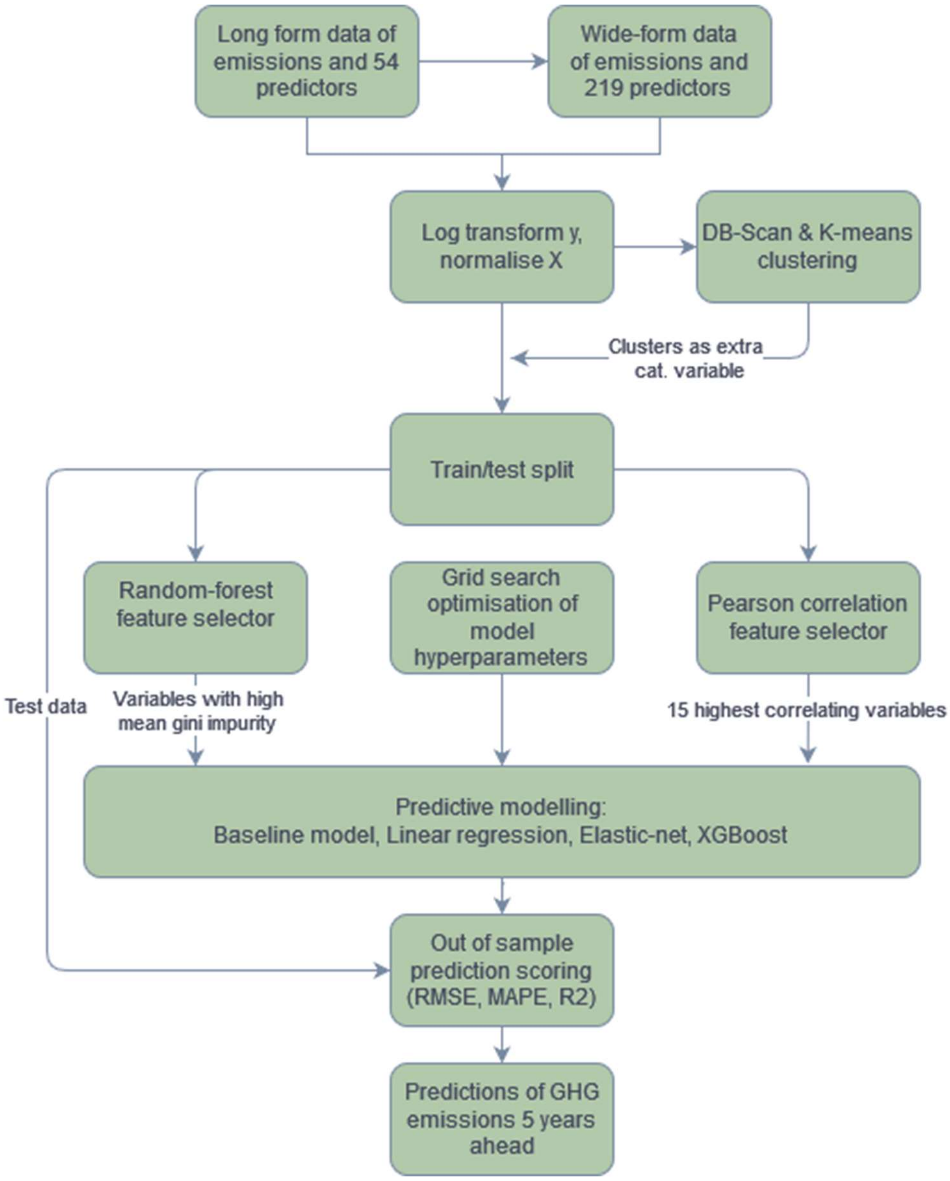

Long form data of emissions and 54 predictors
Wide-form data of emissions and 219 predictors
Log transform y, normalise X
DB-Scan & K-means clustering
Clusters as extra cat. variable
Train/test split
Random-forest feature selector
Grid search optimisation of model hyperparameters
Pearson correlation feature selector
Test data
Variables with high mean gini impurity
15 highest correlating variables
Predictive modelling: Baseline model, Linear regression, Elastic-net, XGBoost
Out of sample prediction scoring (RMSE, MAPE, R2)
Predictions of GHG emissions 5 years ahead

Figure 7 - Overview of machine learning architecture employed

The full model architecture is shown in figure 7 As machine learning ultimately comes down to trial and error in order to see which approach yields the best predictions, two versions of the master data set were created and tested in parallel. The prediction problem is the same for both - to predict emissions 5 years into the future. Both datasets contain the same information, just structured differently. In the wide-form dataset, each variable from a specific year in the long-form dataset becomes its own variable. For example, the number for registered vehicles in the long form dataset becomes registered vehicles in 2010, 2011, 2012 and so forth. This is why the variable count in the wide-form dataset is so high. The background for this approach came from Verenich et al. (2019) refers to this type of encoding the data as index encoding.

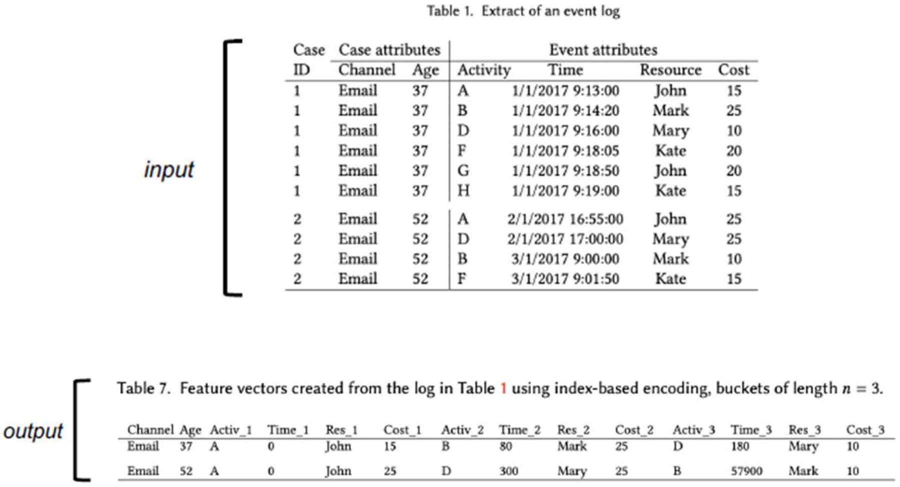

input
output

Table 1. Extract of an event log

Case	Case attributes		Event attributes			
ID	Channel	Age	Activity	Time	Resource	Cost
1	Email	37	A	1/1/2017 9:13:00	John	15
1	Email	37	B	1/1/2017 9:14:20	Mark	25
1	Email	37	D	1/1/2017 9:16:00	Mary	10
1	Email	37	F	1/1/2017 9:18:05	Kate	20
1	Email	37	G	1/1/2017 9:18:50	John	20
1	Email	37	H	1/1/2017 9:19:00	Kate	15
2	Email	52	A	2/1/2017 16:55:00	John	25
2	Email	52	D	2/1/2017 17:00:00	Mary	25
2	Email	52	B	3/1/2017 9:00:00	Mark	10
2	Email	52	F	3/1/2017 9:01:50	Kate	15

Table 7. Feature vectors created from the log in Table 1 using index-based encoding, buckets of length $n = 3$.

Channel	Age	Activ_1	Time_1	Res_1	Cost_1	Activ_2	Time_2	Res_2	Cost_2	Activ_3	Time_3	Res_3	Cost_3
Email	37	A	0	John	15	B	80	Mark	25	D	180	Mary	10
Email	52	A	0	John	25	D	300	Mary	25	B	57900	Mark	10

Figure 8 - example of index encoded data from longitudinal data (Verenich et al., 2019, tab. 1, tab 7)

What will now follow is a brief explanation of each node in the model architecture and its purpose in the process. In the case of the predictive models applied, an extended note on how these work in detail can be found in appendix A.

4.3.1 Variable transformations

The first step in the process was log-transform the emissions data. As seen in figure 6 in chapter 2, there are about 10 municipalities which dwarf all others in terms of GHG-emissions. Some of these house petroleum refineries, others heavy industry, and some are large cities. This led to the distribution of emissions to be heavily skewed, thus ruling out any form of regression-based methods unless rectified. Hence it was deemed necessary to log-transform emissions.

The second step implemented was normalising all the independent variables. This sets all the variables to the scale [0,1], without distorting the differences in their respective ranges or imposing much loss to the data. Normalisation is done via equation 4.1

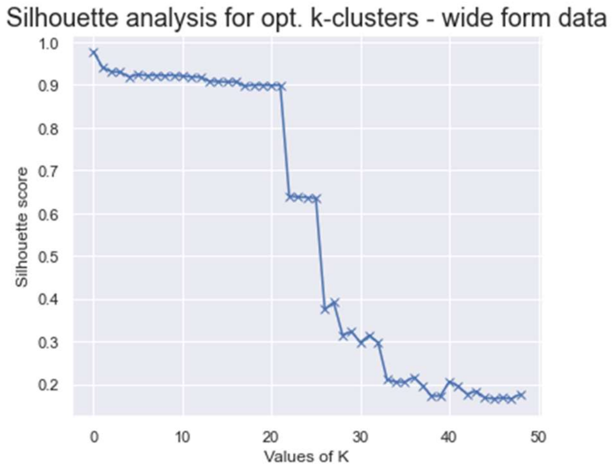
$$x_{norm} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \qquad 4.1$$

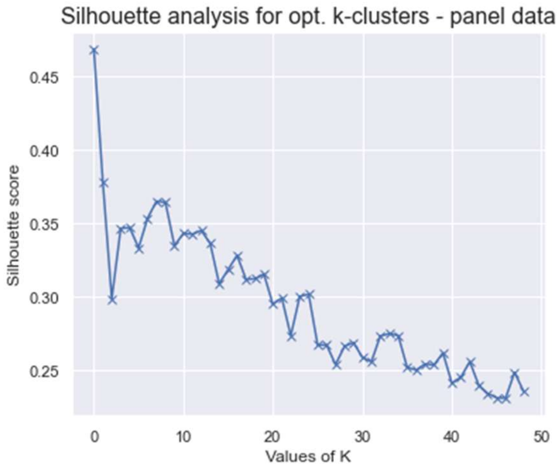
The normalisation was carried out due to the extreme difference in scale between variables. For instance, the amount of public funds allocated per km of road are on the order of $50\ 000\ NOK >$ while the number of new busses registered in a municipality per year is typically smaller than 50. The cost of doing this is that the effect of outliers is reduced, on the upside it enables the usage of clustering algorithms which are sensitive to the scales of data points when calculating the Euclidean distance between datapoints and assigning them to different clusters. Municipalities are different, and systemic differences are highly likely, thus this was deemed a necessary step.

4.3.2 Clustering method

K-means clustering is what's known as a clustering algorithm. It identifies *k* clusters within the data and then assigns each observation to a cluster based on a similarity measure between observations known as Euclidean distance. In layman's terms it is a method that can identify similar observations in a dataset and then attach a label to them indicating their respective categories. This method was employed to determine if there were any structural differences between observations. The method itself requires that the researcher sets the number of clusters. However, a method known as silhouette scoring can be used to determine the optimum number of clusters known as silhouette scoring. See appendix D for how silhouette scoring occurs. The number of clusters with the highest silhouette score is the optimum number of clusters. Surprisingly, the number of optimum clusters turned out to be 0 for both, thus indicating that all datapoints belong to the same cluster.

Figure 8 - Silhouette scores for optimal amounts of clusters





4.3.3 Splitting of data into training and test datasets

A fundamental principle of supervised machine learning is to split the data into training data and testing data. The idea is to validate the models' performances on unseen date in order to make sure that it can make solid predictions out of sample and *doesn't overfit* For the wide-form dataset this is entirely unproblematic. However, special care needed to be taken when splitting the panel data. This is because of the temporal dimension contained within the dataset. If one randomly selects a subset of data that includes a temporal dimension one might risk what's known as data leakage (Hannun et al., 2021). Which means that the model has access to data which it shouldn't have at that point in time. In other words, the model might have access to the data it is trying to predict. In the context of the data in this thesis such an example would be samples from 2015 being used to predict data from the same year. This inhibits any learning in the model and so data leakage makes any model useless in predicting anything out of sample. To get around this problem, a procedure developed by Griffin (2020) was deployed. The short version of how this was done is that instead of using randomly selecting samples for the testing and training datasets, the selection was done by selecting the ordered time series of municipalities at random. The proportion of training data vs test data is 75/25

4.3.4 Variable selection

Feature selection in the model was done in two ways. The first, was to use scikit learn's f_regression feature selector (Pedregosa, 2011a). This works the same as a regular pearson correlation score, but it returns all correlations as positive variables such that it's easier to select the variables with the strongest correlation. The method implemented selects the 15 most highly correlated variables. The formula applied for each variable inside the f_regression module is:

$$\frac{\mathrm{E}[(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_x \sigma_y} \qquad 4.2$$

The second way to choose input variables was achieved by using a random forest model. See appendix A for an extended note on how random forest models operate and how they are applicable to variable input selection. As the random forest model tests for how well a set of variables X predict an outcome Y implicitly makes it suited for variable selection. The fact that these models make no formal distribution assumptions about the data given and are non-parametric (Richmond, 2016) makes the model flexible in terms of what inputs it can handle. The number of variables this method outputs depends on how many variables exert predictive power on the dependent variable. Variables are selected on the basis of their mean decrease in gini impurity. Ultimately, gini impurity gives the probability for misclassifying an observation (Loazia, 2020). Tree-based models use this scoring to assess whether a split based on a variable was a good one or not. If the split is bad, the score is high and vice-versa. The mean decrease in gini impurity thus measures the quality of splits using a given variable in tree-based models. This is used to measure the predictive variables and is the reason why random forest models can be used for variable selection. The mean decrease in gini impurity scores always add up to 1. Hence, the importance scores from a random forest models will always be on the interval (0, 1). The background for its inclusion was inspired by Wei et al. (2018). Care needs to be taken when using this measure as it generally favours continuous variables and categorical variables of high ordinality, however since nearly all the variables in the dataset used are continuous variables this is deemed not to be a problem.

4.3.5 Models fitted

Once the most relevant variables for predicting future emissions were selected both datasets were passed into a battery of predictive models. The predictive models chosen were regular linear regression, elastic-net regression and XGboost[1]. Linear regression was included due to its ease of use and the fact that when used on panel data with cluster robust standard errors, the model becomes a pooled OLS model. Furthermore, if this model does well in making solid predictions on emissions from Norwegian municipalities it could be an indication that the functional relationship between explanatory variables is linear. The elastic-net model was included because that it performs further dimensionality reduction in the data given such that it only selects the most relevant input variables. It also tends to select the variable with the strongest predictive power when faced with two input variables with high collinearity. This deals with the problem of using the correlation-based f_regression module as a variable selector as it could potentially choose variables with strong collinearity. Lastly, the XG boost model was included due to its well documented performance on a multitude of prediction problems be it cross-sectional (Nielsen, 2016; Shwartz-Ziv & Armon, 2022) or even longitudinal data (Chen, 2021).

4.3.6 Optimisation of hyper-parameters

Each model has settings which in turn affects the learning process and the quality of predictions. For example, in the case of a random forest model these settings regulate the number of trees that the algorithm with estimate, or the maximum depth of each tree can have. In the field of machine learning these are often referred to as *hyper-parameters*. A challenge in any supervised machine-learning process is to find the set of hyper-parameters which provide the best possible predictions. Grid-search is a method for automating this process (Pedregosa, 2011b). According to scikit-learn's documentation on grid-search it is an exhaustive search of all hyper-parameters defined by the researcher. The machine learning models are then estimated for each combination of hyper-parameters iteratively, and then results are stored and selected based upon a score function. The type of score can vary based on the needs of the researcher. As this thesis seeks to build a model which makes accurate predictions, root mean square error was chosen as the scoring function.

4.3.7 Benchmark model

In order to see if any model provides predictive power, it is important to compare it to a naïve

[1] Appendix D provides an extensive account on how these models work, their strengths and their weaknesses. Linear regression was left out of this appendix as it is assumed that the reader is familiar with OLS regression.

benchmark model, The naïve models used for benchmarking the predictive models are derived from Wooldridge (2020, p. 376). The purpose of any naïve model is to provide a baseline forecast without any explanatory variables (i.e. simulating a basic guess). For both datasets the naïve model is implemented as follows:

$$y_{i,t+5} = y_{i,t} + \varepsilon_{i,t}$$

4.3.8 Model assessment, selection criteria and variable importance

Model performance in machine learning is measured in terms of predication errors. This thesis uses three measures root mean square error (RMSE), mean absolute percentage error (MAPE) and r2 which is included in order to see how much of the variance in emissions the models can pick up on. The RMSE score was selected because it returns prediction errors in the same unit as the target variable. I.e., production error will be returned in terms of tCO2e.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

Since it squares the distance between predicted and actual emissions, it will give a weight to large errors, which is desirable for the purposes of the prediction problem at hand. Because in addition to making accurate predictions, the model selected should also be correct in the magnitude of total emissions such that it can forecast not only on a local level but also on an aggregate level.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

The MAPE score is included not because it is the best measure for model accuracy, but because it gives an intuitive sense of how accurate model predictions are on average. As such, model performance is judged by its RMSE, while the MAPE score is only used to get an understanding of how good the predictions are in terms of average percentages.

5 Results and discussion

This chapter is divided into two parts. The first part contains discussion of the results from the retrospective LMDI-IDA analysis. The purpose of this part of the analysis was to identify the main drivers of sectoral emissions in Norway between 1990 and 2019. Aggregated results, with the effect of each driving factor of emissions, will thus be shown first. Following this, the analysis indicates that there are three sectors which have had the strongest effect on emissions. These are the transport, mining and petroleum, and industry sectors. These three will be discussed in detail considering their major impact on sectoral emissions. Detailed figures of results per driver can be found in appendix B.

The second part of this chapter shows the results from the machine learning predictions. The predictive performance of the different model-architectures are first assessed in terms of their model evaluation scores. Predicted versus actual values will be shown for the winning model along with its respective variable importances. A forecast of aggregate emissions will then follow as forecasts for all Norwegian municipalities are far too many to visualise.

5.1 Results of retrospective LMDI-IDA

$$C_{s,t} = ktCO2e_{s,t} \equiv GDP_t \frac{GDP_{s,t}}{GDP_t} \frac{GWh_{s,t}}{GDP_{s,t}} \frac{fGWh_{s,t}}{GWh_{s,t}} \frac{ktCO2e_{s,t}}{fGWh_{s,t}} \tag{5.1}$$

$$\Delta C_i = \sum_i \left(\frac{C_i^T - C_i^0}{ln\, C_i^T - ln\, C_i^0} \right) \ln \left(\frac{F_{i,n}^T}{F_{i,n}^0} \right) \tag{5.2}$$

The decomposition equation used for the LMDI index decomposition analysis (LMDI-IDA) is given in 5.1 above. The subsequent effect of the drivers over time is given by equation 5.2. The LMDI-IDA is meant to give a decomposition without any residual term and is consistent in aggregation. As such it is possible to assess the accuracy of the decomposition by taking the difference between the sum of changes in emissions over the period studies and compare it to the actual aggregate emissions from that period. This residual term should be 0, but can in reality be larger than that due to the division of factors that happens when the identity equation is formed. The residual in the results of this analysis is precisely $4{,}37 * 10^{-11}$. We see that some rounding error has happened, but the difference is negligible, and the LMDI-IDA analysis has thus been performed correctly. Furthermore, from the results of the sectoral decompositions it is evident that the analysis is very sensitive to industries with high GHG-intensity. Given that the LMDI-IDA model measures the effect of drivers *on emissions* this is to be expected. After all, a change in a factor in an industry which doesn't emit much will neither abate, nor pollute much. Since there are three large polluting sectors in the Norwegian economy, namely transport, industry, mining and petroleum this is where the effects of changes in drivers are the most extreme. In cases where efficiency gains have abated emissions, gains from these industries are generally large, this is likewise for cases where worsening efficiencies contribute to emissions.

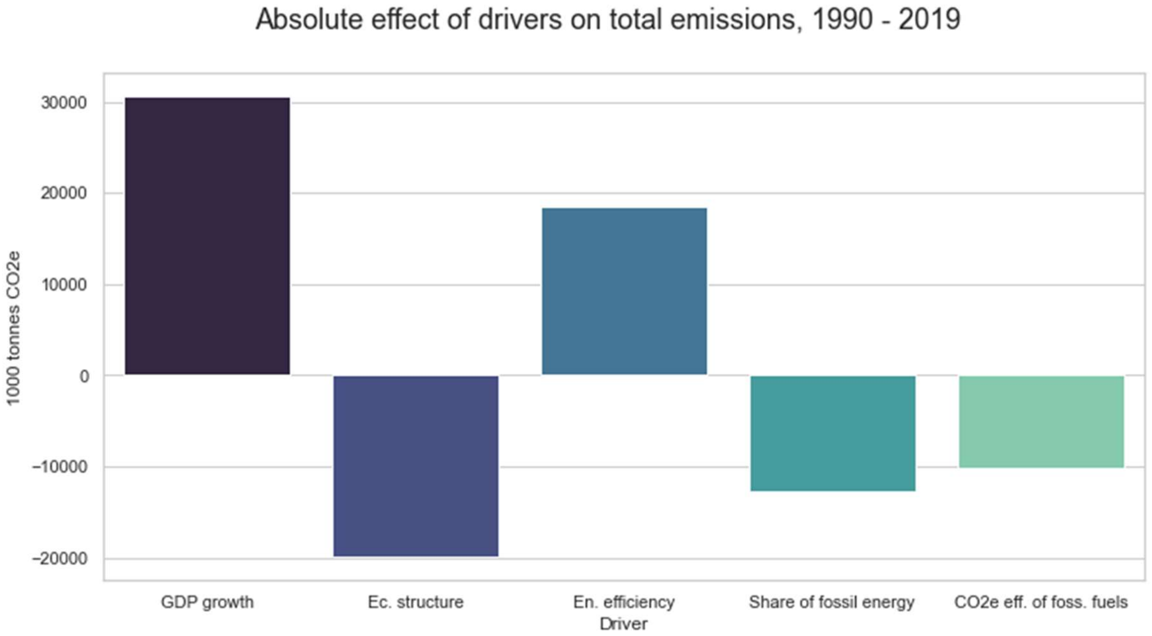


Figure 9 - Sum of effect of drivers on direct GHG emissions from economic sectors 1990 - 2019

The aggregated results of the analysis are shown in figure 8. We see that GDP growth and energy efficiency losses have contributed most strongly to increasing sectoral emissions. Further, changes in economic structure, falling share of fossil fuels, and increased carbon efficiency of fossils have had an abating effect. Ultimately, the drivers of increased emissions are of greater magnitude than drivers of lower emissions. Hence, there has been a 6218 mktCO2e increase in emissions since 1990. Over half of this stems from economic growth. The primacy of this factor is consistent with the findings in other sector-wide LMDI-IDA-based studies that apply similar factors as drivers (Andreoni & Galmarini, 2016; Kumbaroğlu, 2011; O' Mahony et al., 2012; Oh et al., 2010; Yao et al., 2015) and the analysis carried out by Bruvoll and Larsen (2004). Since structural changes in the economy have had a diminishing effect on emissions, this could be an indication of structural shifts towards industries with lower GHG-intensities. However, this analysis does not factor in international trade flows, thus it is unable to tell if the changes we see are due to shifts from high-polluting to lower polluting sectors, or if these GHG-intense sectors have been gradually outsourced. When it comes to the share of fossil fuels and carbon efficiency both have an abating effect which adds up to over 2000 mktCO2e. Surprisingly the results indicate that worsening energy efficiency has contributed strongly towards increasing emissions. This result undoes all the abatement stemming from structural changes in the economy. In other studies, this effect is usually opposite. For instance, Le Quéré et al. (2019) showed that increasing energy efficiency along with de-fossilisation of fuels was a leading cause of lower emissions in 18 developed countries.

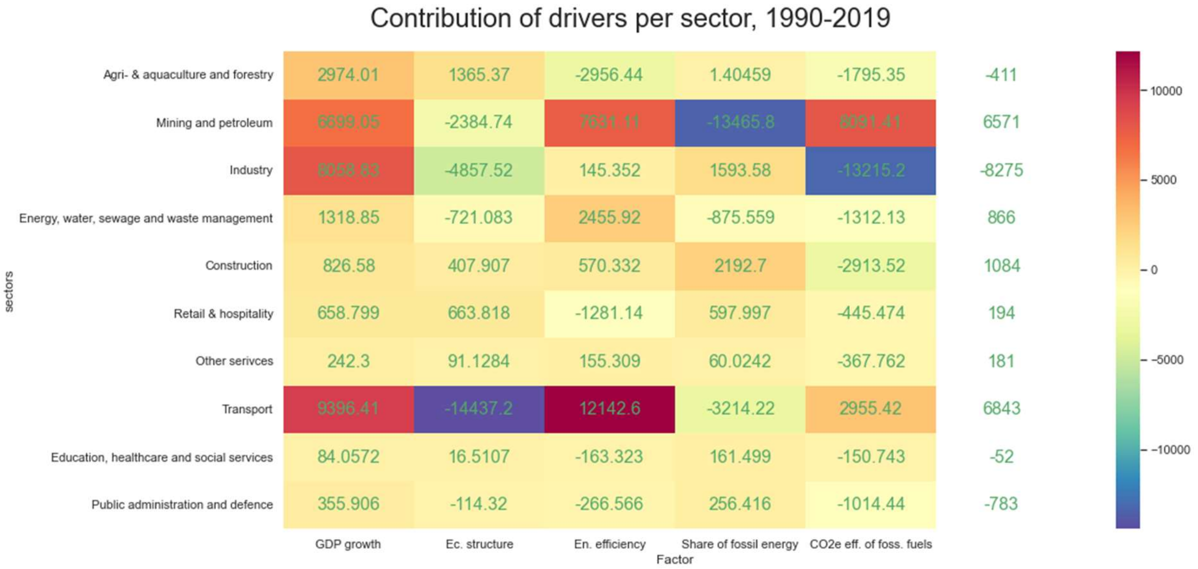


Figure 10 - Contributions towards emissions per sector in mkt CO2e

Figure 9 shows the contributions from each driver per sectors. The results reflect the development of sectoral emissions seen in chapter two. First of all, there are two sectors which dwarf all others in terms of contributions towards higher sectoral emissions. These two sectors are mining and petroleum, and transport. Each has contributed over 6000 mktCO2e towards Norwegian emissions in the studied time period. The reasons for this are their inherent GHG-intensity combined with increased economic growth in the economy combined with worsening energy efficiency and GHG-intensity per unit of fossil fuel expended. It is difficult to say exactly why the efficiency in these two sectors have fallen. Since the analysis uses sectoral GDP, it could be that it is sensitive to price variations in inputs and outputs such as the oil price. But while the price of brent crude oil has varied greatly the last 30 years as shown in figure 10, figure 11 shows that there is a trend in the worsening energy efficiency for both sectors from the mid 1990's and onwards. Besides, if oil prices go up so too do the revenues in the petroleum industry, sectoral GDP follows, and the resulting GWh/GDP ratio should improve. It is difficult to be conclusive without deeper analysis, which is beyond the scope of this thesis. However, figure 11 shows that the energy efficiency in the petroleum sector fell while the price of crude reached historical highs in real terms. This runs counter to the hypothesis of oil prices being a significant factor in explaining the worsening energy efficiency. Brandt (2011) shows how depleting oil and gas fields could offer an explanation for why it consumes more energy compared to its output over time. They show that as reservoirs gradually empty, more energy inputs are necessary for extraction. Likewise, Gavenas et al. (2015) also show that this effect has led to higher emissions from the petroleum industry in Norway. This effect could help explain the worsening energy efficiency and the significant contribution towards emissions from this sector.

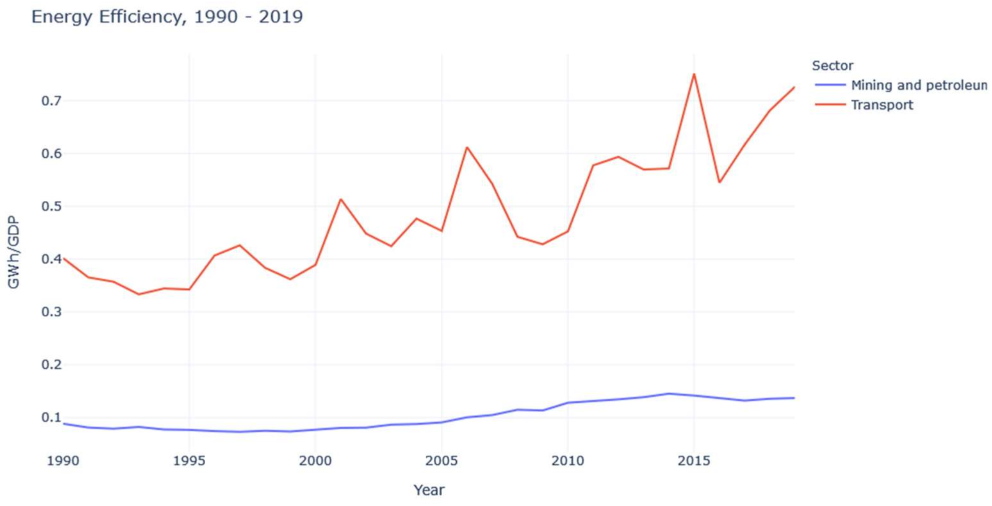


Figure 10 - Energy efficiencies of transport and, mining & petroleum sector.

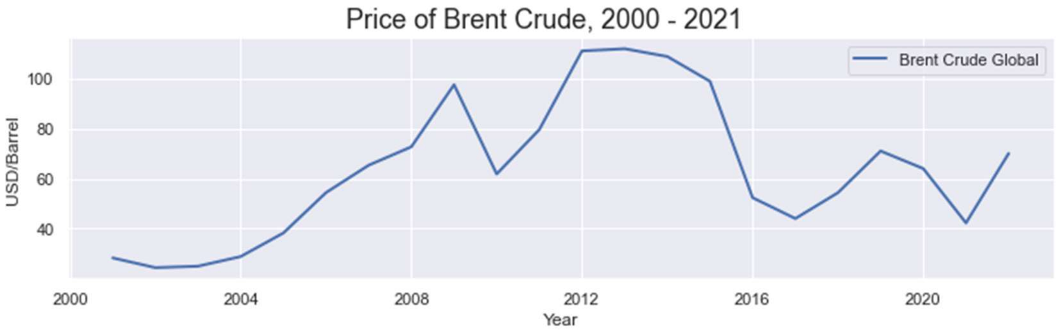


Figure 11 - Price of brent crude per barrel (PPI adjusted), (US EIA, 2022)

In the case of transport, results from this analysis are consistent with an analysis from Statistics Norway which documents worsening energy efficiency for the sector in the period 2010 – 2016 (Statistics Norway, 2017, Fig. 4). Poulsen et al. (2022) documents how shipping routes are planned to maximise economic gain but sacrifices both energy and carbon efficiency in order to achieve this. This could be a potential explanation for the worsening energy efficiency, given the size of the shipping sector in Norway and as seen in chapter 2, emissions from ocean transport constitute most of the emissions in the sector. Likewise, a study from the transport sector in Finland shows that energy efficiency in land transport increased in the period 1995 – 2002 but then decreased towards 2009 (Liimatainen & Pöllänen, 2010)1. A study on the sizes of smaller vehicles carried out by the International Council on Clean Transportation shows how vehicle sizes and weights have increased since 1980's (ICCT, 2017, Fig. 5, p. 50), but that the share of fossil fuels among small vehicles has been falling. If this also applies to other vehicles this could help explain why a worsening energy efficiency, a lower share of fossil fuels and worsening carbon efficiency per unit of fossil fuels occur at the same time. Furthermore, one must also remember the temporal dimension of this analysis. 10 years of increased efficiency measures might not outweigh the effect 20 years of worsening measures might have had on emissions. A final point regarding the transport sector worth discussing is the effect stemming from changes in economic structure. Transport is the sector in which this effect is the strongest, and in practical terms it implies that the size of the transportation sector has shrunk compared to others.

The results on the industry sector indicate that while economic growth has led to an increase of about 8000 mktCO2e, this is more than negated by the effects of structural change and the increased GHG-efficiency of fossil fuels. Overall, there has been a reduction of 8275 ktCO2e in this sector, which is by far the largest reduction out of all the sectors studied. The effect of structural change is easily explainable as it simply means that the sector has become smaller in relation to other sectors over time. Whether this is due to a substitution effect with other sectors in the Norwegian economy or if it is due to outsourcing cannot be determined in this analysis and could benefit from further study as it could shed light on the effectiveness of climate policy related to Norwegian industry the last thirty years. The Norwegian Water Resources and Energy Directorate suggests that most of the reductions stem from fewer process-related emissions, which constitutes most of the emissions from this sector (NVE, 2020, p. 12). In other words, most of the reduction has happened due to the implementation of production processes which emit less. Interestingly, according to their analysis most of these abatement measures were done in the period 1990 – 2010 and emissions have remained stagnant since. In the same report, the potential for increased electrification of the industry, which still uses a substantial amount of fossil energy, is assessed. They find that the potential for further abatement is considerable but contingent on expansion of the electric grid, as well as increased production of electricity if prices are to remain low.

Given the magnitude in the emissions coming from these two sectors just discussed, it follows that if Norway wishes to decrease its sectoral emissions, much could be gained from policies aimed at making structural shifts away from the transport and petroleum industry or increasing their energy efficiency. The Norwegian government has indicated that it wishes to electrify its petroleum sector which will likely lead to such an outcome. However, the current climate policy in Norway makes little mention of measures tied to international shipping. In the national climate plan for 2021 – 2030 it is only mentioned once in brief (St. Meld. 13 (2020-2021), p. 15). Moreover, only the domestic sea-transport sector is mentioned in the roadmaps for a green economy (Ministry of Climate and Environment, 2021). Given pledges given in the Paris agreement, perhaps more attention should be given to this sector.

5.2 Results of predictive analysis

Following the results of the estimation it is evident that the architecture developed for prediction municipal GHG emissions 5 years ahead succeeded. In fact, both the panelised data and the cross-sectional data enabled accurate predictions and the predictive models beat the baseline model with a wide margin. As the target variable was $log(tCO2e_{t+5})$ the RMSE is in log tonnes of CO2e. It surprising that the results from the cross-sectional dataset, which were estimated using data from 2015 and before, made accurate predictions for 2020. Two results are worth discussing, these two are the RF-LinReg (Random Forest – Linear Regression model), and the Corr-XGboost model on the panelised data. Predicted vs actual values will be shown for these and a forecast for emissions in 2025 will be made using the latter, considering this model provided the best performance.

Table 3 - Prediction results on testing data

Model	*Dataset*	*RMSE*	*MAPE*	*R2*
Baseline	Cross-sectional	1.77	0.83	0.00
RF – LinReg	Cross-sectional	**0.15**	**0.09**	**0.97**
RF – Elastic net	Cross-sectional	0.46	0.34	0.79
RF – Xgboost	Cross-sectional	0.21	0.01	0.96
Corr – LinReg	Cross-sectional	0.16	0.01	0.97
Corr – Elastic net	Cross-sectional	0.57	0.03	0.73
Corr – XGBoost	Cross-sectional	0.21	0.01	0.96
Baseline	Panel data	1.08	0.83	0.00
RF – LinReg (Pooled OLS)	Panel data	0.70	0.05	0.69
RF – Elastic net	Panel data	0.86	0.06	0.52
RF – XGboost	Panel data	0.14	0.01	0.98
Corr – LinReg (Pooled OLS)	Panel data	0.92	0.01	0.45
Corr – Elastic net	Panel data	0.61	0.04	0.76
Corr – XGBoost	Panel data	**0.13**	**0.08**	**0.98**
Trend model	Panel data	1.08	0.83	0.00

Figure 10 shows the RF-LinReg model's performance in estimating emissions for 2020 using data from 2015 and before. Both the estimated and predicted values were exponentiated post-estimation to show the true magnitudes of the values. It is evident that the model predicts very well on average and can even do well with outliers. The RMSE of this model is 0.15, which corresponds to an average error of $e^{0.15} = 1.16$ tonnes per year.

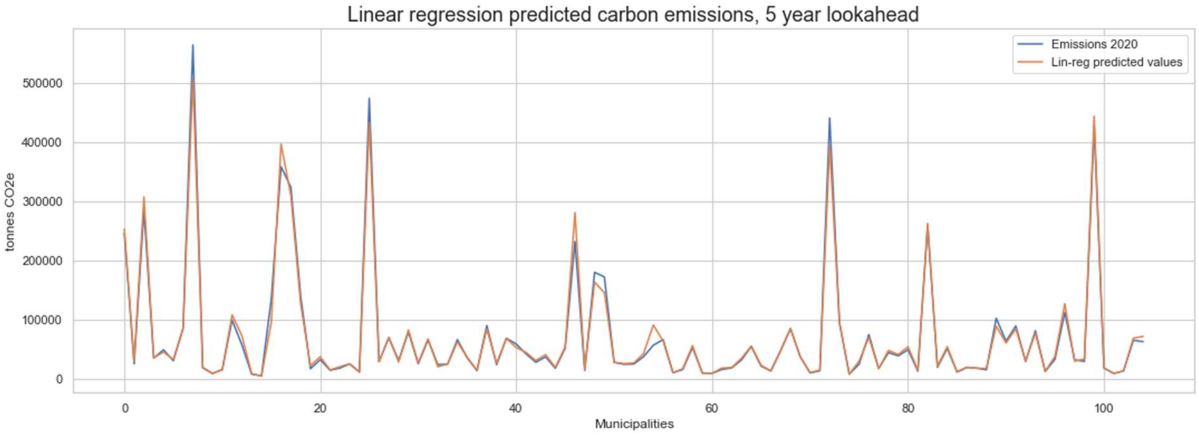


Figure 1211 – Predicted vs actual values. The numbers on the x-axis represents individual municipalities

The input variables in the linear regression model were selected by the random forest input selector. As covered in chapter 4, these scores always add up to 1, and show which variables are the most important in making predictions within the model. A higher score indicates higher importance, and vice-versa. Evidently, past electricity consumption from the years 2009-20015 are the variables with the highest predictive power. Considering that employees from the primary, secondary and tertiary also show up as contributing variables to predictions it seems that the random forest variable selector is sensitive to the economic structure of the municipalities it is applied to.

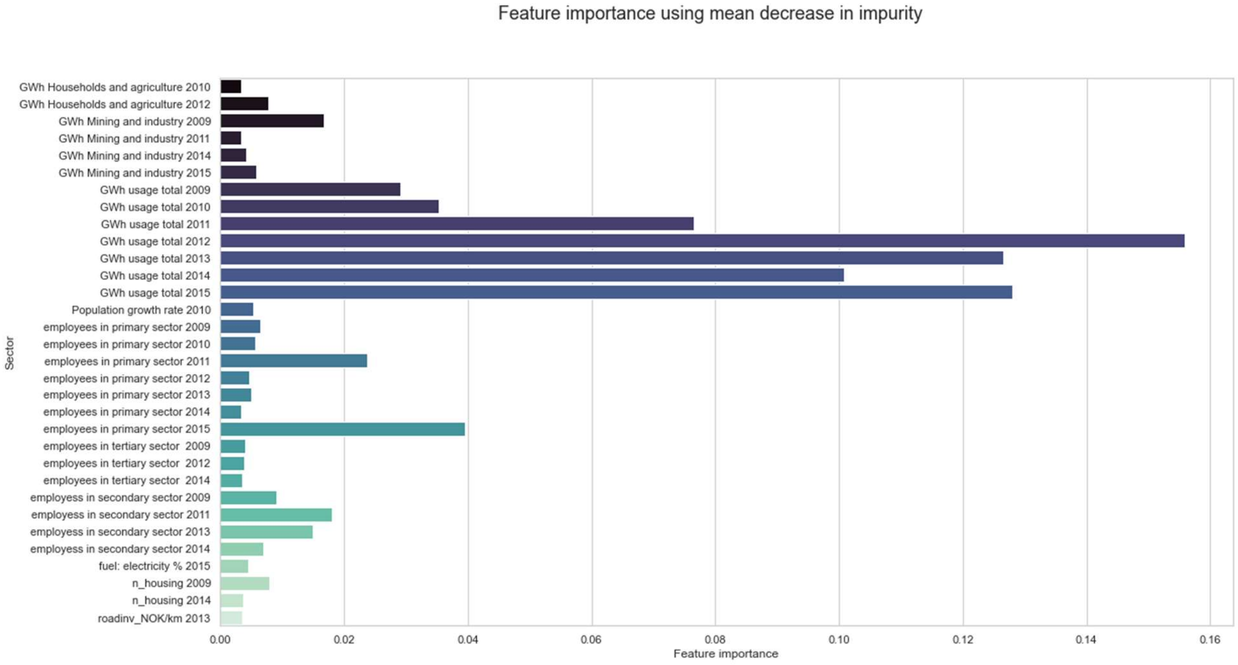


Figure 1312 - Variable importance of selected variables in the cross-sectional dataset

The winning model was the Corr-XGboost model applied on the panel dataset. Since the predictions were made on panel data the model can do dynamic predictions. For the training and test data, model's first prediction year is 2014 and the last is 2020. In figure 12 the performance of the model is plotted against the actual values of the testing data. Each dimple in the line presents one municipality over time. Noticeably the model manages to capture the magnitude for the emissions for each municipality. This is likely due to the inclusion of historical emissions as a variable. However, it seems to struggle with the trend in some of the outliers, and in the most extreme cases it severely underestimates the emissions. This could be because there are variables specific to these outliers are not present in the dataset. It could also have something to do with the normalisation applied to the data, as it gives outlier variables a maximum value of 1. However, for many municipalities, and especially those with fewer emissions, the model seems to perform better. Examples of predicted vs observed emissions can be found in appendix D.

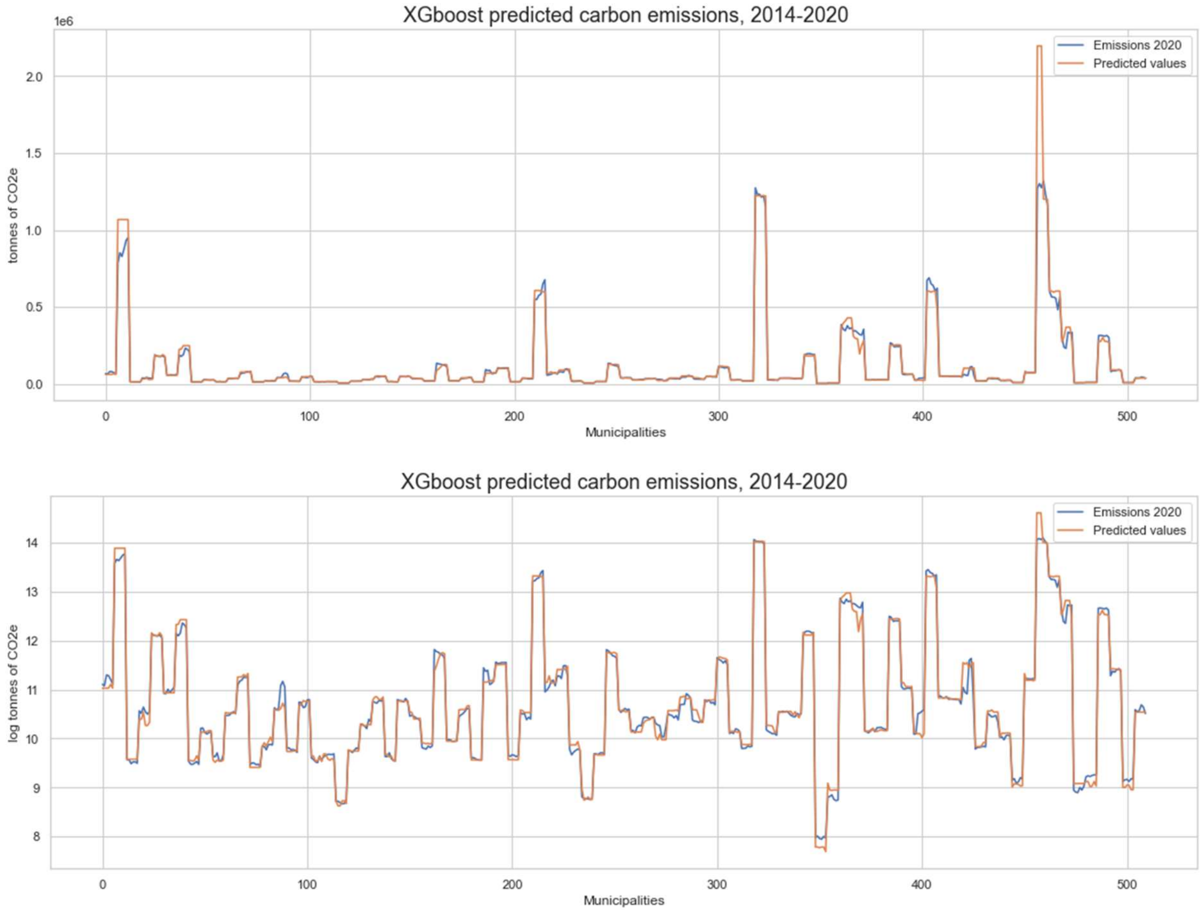


Figure 1413 - Predicted emissions between 2016 - 2020 with XG boost using data from 2015 and before. Each dimple in the line represents one municipality in the time-period 2014-2020

The input variables selected via sci-kit learn's f_regression module are different from the ones selected by the RF model on the cross sectional dataset. Since this module ranks the exactly the same as a regular pearson correlation matrix would, the variable importances were converted back to regular correlations for ease of interpretations. Using this method of input-selection, heterogeneity of variables have grown. We see that demographic data now plays a larger role with population, amounts of buildings (non-housing) and amount of housing units being included. Aside from electricity usage the number of different types of vehicles were also selected. This method of selecting variables does not seem sensitive to the economic structure of municipalities.

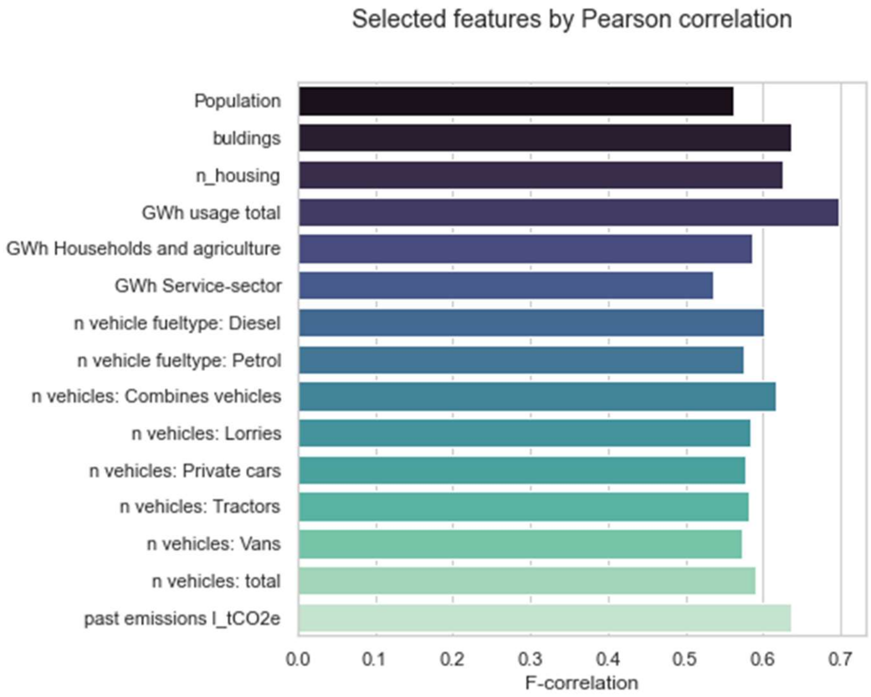


Figure 1514 - Pearson corelations with target variable

The XGboost weight measure, which calculates the amount of time a variable was used to split trees in the model during estimations, was used to assess the importances of predictors in the winning model. In figure 14 The past emissions l tCO2e variable which shows historical emissions is by far the most important variable. As such this is the most significant variable for predicting future emissions. Other factors include buildings, power usage and tractors. However, Since XG boost is a model which can combine many variables with weak predictive power, and then combine them into something with greater predictive power. It is likely that most of the variables shown in the feature weight diagram have little predictive power by themselves. This measure has the same problem as gini scores in that it measure the amount of splits on the trees the model fits to estimate the parameters. Again, as most variables are continuous this was judged not to be an issue.

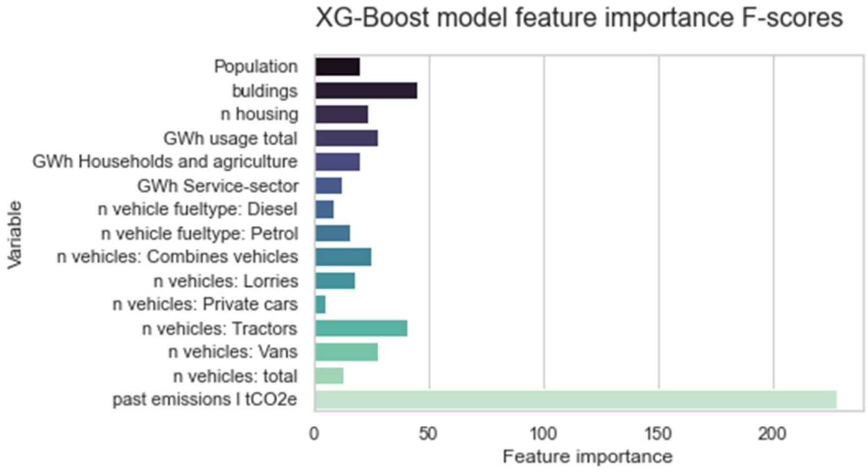


Figure 1615 - Variable importance of XGboost mode measured by feature-weight on panel data dataset

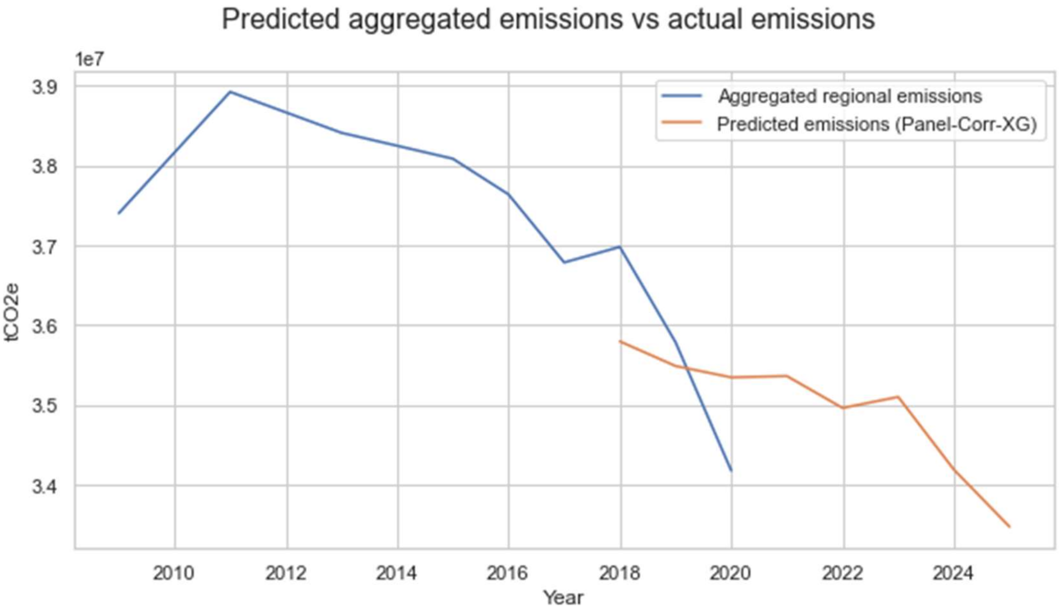


Figure 1716 - Aggregated emissions forecast

After finishing the training and testing process, new data was given to the corr-XGboost model to make forecasts. Since the model predicts 5 years ahead the last timestep is 2025. A, aggregate forecast was made by adding all the time-series in the forecasted dataset together. On aggregate the model seems to be able to capture the trend of the future emissions and indicates that they will lower towards the 3.35 mgtCO2e-mark in 2025. It seems that the model overestimates the emissions in 2020, due to the corona pandemic, and so the pace of the decreasing emissions is too slow compared to the current trend. However, given that 2020 was an outlier year in many ways, it would be interesting to see how well the predictions hold up given new data. Furthermore, the model is consistent in aggregation even despite not being able to accurately predict the behaviour of outliers. This implies that even the wrongful predictions weren't too far off the mark in absolute terms and that on average, the model is relatively precise.

Given the results of the analysis limitations of machine learning becomes clear. While the models used are able to provide accurate predictions, it is generally difficult to understand the effect of variables on predictions. This is because the measures only say how important predictions were inside the model. Extrapolating that to the real world is difficult and variable importances cannot be compared with findings from other, models. This sets a limit to what can be learned from the models used here. For example, while the number of tractors in a municipality today might contribute to predicting emissions in 5 years, it is unclear to what degree they contribute. This is because the functional relationship found by models such as XGboost remain inside the model.

On the positive side, while the predictions from the models used in this study are not perfect, they go a long way to show that even with relatively simple machine learning models, good predictions of future emissions are achievable. The models employed here can merely be seen as a prototype for more advanced architectures that can deal better with the entity-effects at play. The results also show how macro-economic data can be used for making forecasts of emissions with machine learning in Norway, thus meaning that more research using these methods might have something to add regarding questions of how future emissions might develop.

6 Conclusion

Climate change has been on the agenda for 30 years in Norway and ambitious goals for reducing emissions have been set for the next decade. This prompted two questions. First, what have been the main drivers of change in sectoral emissions in Norway for this period, and second, how are emissions on a municipal level likely to develop the next 5 years. To identify the main drivers of change of sectoral emissions, a retrospective analysis based on the LMDI-IDA framework was carried out. It showed that economic growth and worsening energy efficiency particularly in the transport, and mining and petroleum industry were the strongest factors in increasing emissions. Structural change, a lower share of fossil fuels and an increased carbon-efficiency of fossil fuels have all had abating effects on emission. Furthermore, the driving forces of emissions were of a greater magnitude than the abating factors leading to an increase in sectoral emissions of 6218 mktCO2e since 1990. Lastly, the industry sector was shown to be the largest abating factor, both due to structural change, and an increased carbon efficiency of fossil fuels. It is not clear why energy efficiency has dropped for the transportation, and mining and petroleum sector, and further studies are needed in order to shed light on these issues.

The predictive part of this this thesis employed machine learning methods to predict emissions on a municipal level 5 years into the future. It was shown that supervised machine learning methods can produce robust predictions, and that they add more explanatory power than naïve models. Two datasets were developed for the predictive analysis, one cross-sectional and one in a panel format. A random-forest feature selector combined with a linear regression model performed best on the cross-sectional dataset, while an XGboost model with a pearson correlation feature selector performed best on the panel data. The most reliable predictor of future emissions was shown to be past emissions. Variable importance otherwise varied across datasets. Electricity usage, and number of employees across sectors proved to be the most important variables for the cross-sectional dataset while buildings provided the highest feature importance scores in the panel data set. Predictions of future emissions in all Norwegian municipalities for the period 2020-2025 were made. The predictions were able to capture the trend of emissions but was not able to foresee the fall in emissions in 2020. The predictions indicate that the trend of lower emissions will continue towards 2025. It is difficult to say exactly how much emissions will continue to decrease given that the winning model estimated in this thesis couldn’t anticipate the effect of the pandemic. However, the results from the predictive analysis indicate that machine learning is a viable methodology for providing forecasts of municipal emissions, but further development of more accurate models are needed if they are to provide good answers to whether Norway is on track to reaching its climate policy goals or not.

7 References

A. Gullberg & S. Aakre. (2015). *Norsk klimapolitikk: 2030-målene og tilknytningen til EU.*: CICERO Senter for klimaforksning. Available at: https://pub.cicero.oslo.no/cicero-xmlui/handle/11250/284757 (accessed: 27.03).

Aamaas, B. K., Jan Ivar; Madslien, Anne. (2019). *Referansebane og framskrivning for Oslos klimagassutslipp mot 2030 - Revisjon mai 2019*. Oslo: CICERO Center for International Climate and Environmental Research - Oslo.

Acheampong, A. O. & Boateng, E. B. (2019). Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, 225: 833-856. doi: https://doi.org/10.1016/j.jclepro.2019.03.352.

Adams, D., Oh, D.-H., Kim, D.-W., Lee, C.-H. & Oh, M. (2020). Prediction of SOx–NOx emission from a coal-fired CFB power plant with machine learning: Plant data learned by deep neural network and least square support vector machine. *Journal of Cleaner Production*, 270: 122310.

Andreoni, V. & Galmarini, S. (2012). European CO2 emission trends: A decomposition analysis for water and aviation transport sectors. *Energy*, 45 (1): 595-602. doi: https://doi.org/10.1016/j.energy.2012.07.039.

Andreoni, V. & Galmarini, S. (2016). Drivers in CO2 emissions variation: A decomposition analysis for 33 world countries. *Energy*, 103: 27-37. doi: https://doi.org/10.1016/j.energy.2016.02.096.

Ang, B. W. (2004). Decomposition analysis for policymaking in energy:: which is the preferred method? *Energy Policy*, 32 (9): 1131-1139. doi: https://doi.org/10.1016/S0301-4215(03)00076-4.

Ang, B. W. (2015). LMDI decomposition approach: A guide for implementation. *Energy Policy*, 86: 233-238. doi: https://doi.org/10.1016/j.enpol.2015.07.007.

Athey, S. (2019). 21. The Impact of Machine Learning on Economics

The Economics of Artificial Intelligence: An Agenda. In Agrawal, A., Gans, J. & Goldfarb, A. (eds), pp. 507-552: University of Chicago Press.

Berg. (2015). *Norsk klimapolitikk 1987-2015*. Oslo, Norway: CICERO Senter for klimaforskning. Available at: https://www.cicero.oslo.no/no/posts/klima/norsk-klimapolitikk-1987-2015 (accessed: 23.02).

Brandt, A. R. (2011). Oil Depletion and the Energy Efficiency of Oil Production: The Case of California. *Sustainability*, 3 (10). doi: 10.3390/su3101833.

Breiman, L. (2001a). Random Forests. *Machine Learning*, 45 (1): 5-32. doi: 10.1023/A:1010933404324.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16 (3): 199-231.

Bruvoll, A. & Larsen, B. M. (2004). Greenhouse gas emissions in Norway: do carbon taxes work? *Energy Policy*, 32 (4): 493-505. doi: https://doi.org/10.1016/S0301-4215(03)00151-4.

Chen, J. M. (2021). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, 27 (1): 1-16. doi: 10.1007/s11294-021-09815-6.

de Boer, P. & Rodrigues, J. F. D. (2020). Decomposition analysis: when to use which method? *Economic Systems Research*, 32 (1): 1-28. doi: 10.1080/09535314.2019.1652571.

Gavenas, E., Rosendahl, K. E. & Skjerpen, T. (2015). CO2-emissions from Norwegian oil and gas extraction. *Energy*, 90: 1956-1966. doi: https://doi.org/10.1016/j.energy.2015.07.025.

Goh, T. & Ang, B. W. (2019). Tracking economy-wide energy efficiency using LMDI:

approach and practices. *Energy Efficiency*, 12 (4): 829-847. doi: 10.1007/s12053-018-9683-z.

Griffin. (2020). *Assigning Panel Data to Training, Testing and Validation Groups for Machine Learning Models*. Available at: https://towardsdatascience.com/assigning-panel-data-to-training-testing-and-validation-groups-for-machine-learning-models-7017350ab86e (accessed: 26.04).

Halevy, A., Norvig, P. & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24 (2): 8-12. doi: 10.1109/MIS.2009.36.

Hamilton, C. & Turton, H. (2002). Determinants of emissions growth in OECD countries. *Energy Policy*, 30 (1): 63-71. doi: https://doi.org/10.1016/S0301-4215(01)00060-X.

Hannun, A., Guo, C. & Maaten, L. v. d. (2021). *Measuring data leakage in machine-learning models with Fisher information*. Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, Proceedings of Machine Learning Research, pp. 760--770: PMLR.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, vol. 3: Springer.

Harrison, P. & Pearce, F. (2000). *AAAS atlas of population & environment*: Univ of California Press.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, vol. 2: Springer.

Hertwich, E. G. & Roux, C. (2011). Greenhouse Gas Emissions from the Consumption of Electric and Electronic Equipment by Norwegian Households. *Environmental Science & Technology*, 45 (19): 8190-8196. doi: 10.1021/es201459c.

Hoekstra, R. & van den Bergh, J. C. J. M. (2003). Comparing structural decomposition analysis and index. *Energy Economics*, 25 (1): 39-64. doi: https://doi.org/10.1016/S0140-9883(02)00059-2.

Hsu, P. L. & Robbins, H. (1947). Complete Convergence and the Law of Large Numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 33 (2): 25-31. doi: 10.1073/pnas.33.2.25.

Huang, G.-B., Zhu, Q.-Y. & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70 (1-3): 489-501.

Huang, L. & Bohne, R. A. (2012). Embodied air emissions in Norway's construction sector: input-output analysis. *Building Research & Information*, 40 (5): 581-591. doi: 10.1080/09613218.2012.711993.

Hussien, A. G., Amin, M. & Abd El Aziz, M. (2020). A comprehensive review of moth-flame optimisation: variants, hybrids, and applications. *Journal of Experimental & Theoretical Artificial Intelligence*, 32 (4): 705-725.

IPCC. (2022a). *AR6 Climate Change 2022: Impacts, Adaptation and Vulnerability*. Geneva: IPCC. Available at: https://www.ipcc.ch/report/ar6/wg2/ (accessed: 04.03).

IPCC. (2022b). *AR6 Climate Change 2022: Impacts, Adaptation and Vulnerability, Summary for Policymakers*. Geneva: IPCC. Available at: https://www.ipcc.ch/report/ar6/wg2/ (accessed: 06.04).

IPCC. (2022c). *FAQ 1: What are the new insights on climate impacts, vulnerability and adaptation from IPCC?*: IPCC. Available at: https://www.ipcc.ch/report/ar6/wg2/about/frequently-asked-questions/keyfaq1 (accessed: 16.03).

IPCC. (2022d). *History of the IPCC* IPCC. Available at: https://www.ipcc.ch/about/history/ (accessed: 22.03).

Jacobsen, J., Holmengen, Ekre, Rasch, Fluge & Lillesund, H., Seim, Gutterød. (2021). *Municipal emission inventory. Methodological documentation.* Oslo: Norwegian

Environmental Agency. Available at: https://www.miljodirektoratet.no/contentassets/684ed944b61948e8adbef6f3f5b699f7/dokumentasjonsnotat-versjon_5_2022.pdf/download (accessed: 01.05).

Korsbakken, J. I., Madslien, Anne Romundstad, Reidun Marie Aamaas, Borgar. (2020). *Bergens klimagassutslipp mot 2030 - Referansebane og mulighetsscenarier*. Oslo: CICERO Center for International Climate and Environmental Research - Oslo.

Korsbakken, J. I. R., Reidun Marie; Madslien, Anne. (2021). *Kristiansands klimagassutslipp mot 2030: Referansebane og tiltakspakker*. In Korsbakken, J. I. (ed.). Oslo: CICERO Center for International Climate and Environmental Research - Oslo.

Kumbaroğlu, G. (2011). A sectoral decomposition analysis of Turkish CO2 emissions over 1990–2007. *Energy*, 36 (5): 2419-2433. doi: https://doi.org/10.1016/j.energy.2011.01.027.

L. Breiman, J. F., R. Olshen, C. Stone. (1984). *Classification And Regression Trees* New York: Routledge. Available at: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-richard-olshen-charles-stone (accessed: 27.04).

Le Quéré, C., Korsbakken, J. I., Wilson, C., Tosun, J., Andrew, R., Andres, R. J., Canadell, J. G., Jordan, A., Peters, G. P. & van Vuuren, D. P. (2019). Drivers of declining CO2 emissions in 18 developed economies. *Nature Climate Change*, 9 (3): 213-217. doi: 10.1038/s41558-019-0419-7.

Liimatainen, H. & Pöllänen, M. (2010). Trends of energy efficiency in Finnish road freight transport 1995–2009 and forecast to 2016. *Energy Policy*, 38 (12): 7676-7686. doi: https://doi.org/10.1016/j.enpol.2010.08.010.

Loazia. (2020). *Gini Impurity Measure – a simple explanation using python*. Available at: https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33 (accessed: 05.05).

Mardani, A., Streimikiene, D., Cavallaro, F., Loganathan, N. & Khoshnoudi, M. (2019). Carbon dioxide (CO2) emissions and economic growth: A systematic review of two decades of research from 1995 to 2017. *Science of The Total Environment*, 649: 31-49. doi: https://doi.org/10.1016/j.scitotenv.2018.08.229.

Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M. & Cavallaro, F. (2020). A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production*, 275: 122942. doi: https://doi.org/10.1016/j.jclepro.2020.122942.

Metcalf, G. E. (2008). An Empirical Analysis of Energy Intensity and Its Determinants at the State Level. *The Energy Journal*, 29 (3): 1-26.

Ministry of Climate and Environment. (2021). *Veikart for grønn konkurransekraft*. Available at: https://www.regjeringen.no/no/tema/klima-og-miljo/innsiktsartikler-klima-miljo/veikart-for-gronn-konkurransekraft/id2604070/ (accessed: 10.05).

Morde. (2019). *XG Boost Algorithm: Long may she reign!* Available at: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

Moutinho, V., Madaleno, M., Inglesi-Lotz, R. & Dogan, E. (2018). Factors affecting CO2 emissions in top countries on renewable energies: A LMDI decomposition application. *Renewable and Sustainable Energy Reviews*, 90: 605-622. doi: https://doi.org/10.1016/j.rser.2018.02.009.

Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?*: NTNU.

Norwegian Environment Agency. (2020). *Greenhouse Gas Emissions 1990-2018, National*

Inventory Report. Available at: https://www.miljodirektoratet.no/publikasjoner/2020/april-2020/greenhouse-gas-emissions-1990-2018-national-inventory-report/ (accessed: 22/04).

Norwegian Environment Agency. (2022). *Norge skal fram til 2020 kutte i de globale utslippene av klimagasser tilsvarende 30 prosent av Norges utslipp i 1990*. Available at: https://miljostatus.miljodirektoratet.no/miljomal/klima/miljomal-5.1/ (accessed: 26.01).

Norwegian Environmental Agency. (2021). *Statistikk for utslipp av klimagasser for alle kommuner*. Oslo. Available at: https://www.miljodirektoratet.no/contentassets/684ed944b61948e8adbef6f3f5b699f7/utslippsstatistikk_alle_kommuner.xlsx/download (accessed: 06.01).

Norwegian Mapping Authority. (2020). *Kommune- og regionsendringer 2020*. Available at: https://www.kartverket.no/til-lands/kommunereform/tekniske-endringer-ved-sammenslaing-og-grensejustering/komendr2020 (accessed: 03.04).

Norwegian Shipowners' Association. (2021). *Norway is the world's fourth largest shipping nation measured by value* Norwegian Shipowners' Association. Available at: https://rederi.no/en/aktuelt/2021/norway-is-the-worlds-fourth-largest-shipping-nation-measured-by-value/.

O' Mahony, T., Zhou, P. & Sweeney, J. (2012). The driving forces of change in energy-related CO2 emissions in Ireland: A multi-sectoral decomposition from 1990 to 2007. *Energy Policy*, 44: 256-267. doi: https://doi.org/10.1016/j.enpol.2012.01.049.

Oh, I., Wehrmeyer, W. & Mulugetta, Y. (2010). Decomposition analysis and mitigation strategies of CO2 emissions from energy consumption in South Korea. *Energy Policy*, 38 (1): 364-377. doi: https://doi.org/10.1016/j.enpol.2009.09.027.

Pedregosa, V., Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay. (2011a). *sk-learn reference : sklearn.feature_selection.f_regression*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html (accessed: 27.04).

Pedregosa, V., Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay. (2011b). *Tuning the hyper-parameters of an estimator*. Available at: https://scikit-learn.org/stable/modules/grid_search.html#gridsearch-scoring (accessed: 27.04).

Peters, G. P. & Hertwich, E. G. (2006). Pollution embodied in trade: The Norwegian case. *Global Environmental Change*, 16 (4): 379-387. doi: https://doi.org/10.1016/j.gloenvcha.2006.03.001.

Poulsen, R. T., Viktorelius, M., Varvne, H., Rasmussen, H. B. & von Knorring, H. (2022). Energy efficiency in ship operations - Exploring voyage decisions and decision-makers. *Transportation Research Part D: Transport and Environment*, 102: 103120. doi: https://doi.org/10.1016/j.trd.2021.103120.

Richmond. (2016). *Algorithms Exposed: Random Forest*. Available at: https://bccvl.org.au/algorithms-exposed-random-forest/ (accessed: 24.04).

Schiltz, F., Masci, C., Agasisti, T. & Horn, D. (2018). Using Regression Tree Ensembles to Model Interaction Effects: A Graphical Approach. *Applied Economics*, 50. doi: 10.1080/00036846.2018.1489520.

Seim. (2022). *Email correspondence with Tomas Seim, advisor for emissions accounts, Norwegian Environmental Agency* (Email 25.03.2022).

Seyedzadeh, S., Rahimian, F. P., Glesk, I. & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering*, 6 (1): 5. doi: 10.1186/s40327-018-0064-7.

Shwartz-Ziv, R. & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84-90. doi: https://doi.org/10.1016/j.inffus.2021.11.011.

Simonsen, M., Gössling, S. & Walnum, H. J. (2019). Cruise ship emissions in Norwegian waters: A geographical analysis. *Journal of Transport Geography*, 78: 87-97. doi: https://doi.org/10.1016/j.jtrangeo.2019.05.014.

Sparrevik, M. & Utstøl, S. (2020). Assessing life cycle greenhouse gas emissions in the Norwegian defence sector for climate change mitigation. *Journal of Cleaner Production*, 248: 119196. doi: https://doi.org/10.1016/j.jclepro.2019.119196.

SSB. (2021). *Emissions to air - table 09288*: Statistics Norway,. Available at: https://www.ssb.no/en/statbank/table/09288/.

St. Meld. 13 (2020–2021). *Klimaplan for 2021–2030, Del 1, Kap 1.1*. Oslo, Norway: Klima- og miljødepartementet. Available at: https://www.regjeringen.no/no/dokumenter/meld.-st.-13-20202021/id2827405/ (accessed: 01.04).

Staistics Norway. (2021). *Production and consumption of energy, energy balance and energy account*. Oslo: Statistics Norway. Available at: https://www.ssb.no/en/statbank/table/11557/ (accessed: 31.04).

Statistics Norway. (2009). *Standard for næringsgruppering (SN)*. Available at: https://www.ssb.no/klass/klassifikasjoner/6/korrespondanser (accessed: 19.02).

Statistics Norway. (2014). *Concepts and definitions in national accounts*. Available at: https://www.ssb.no/en/nasjonalregnskap-og-konjunkturer/concepts-and-definitions-in-national-accounts#Value_added (accessed: 05.02).

Statistics Norway. (2015). *Hvilke utslipp dekkes av statistikkene?*: SSB. Available at: https://www.ssb.no/natur-og-miljo/artikler-og-publikasjoner/hvilke-utslipp-dekkes-av-statistikkene (accessed: 15.04).

Statistics Norway. (2017). *Norsk produksjon er blitt mer energieffektiv*: Statistics Norway. Available at: https://www.ssb.no/energi-og-industri/artikler-og-publikasjoner/norsk-produksjon-er-blitt-mer-energieffektiv (accessed: 31.04).

Statistics Norway. (2021). *Emissions to air*. Available at: https://www.ssb.no/en/statbank/table/09288/ (accessed: 07.01).

Statistics Norway. (2022a). *09170: Produksjon og inntekt, etter næring 1970 - 2021*. Available at: https://www.ssb.no/statbank/table/09170/ (accessed: 01.02).

Statistics Norway. (2022b). *To av tre nye biler er el-biler*. Available at: https://www.ssb.no/transport-og-reiseliv/landtransport/statistikk/bilparken/artikler/to-av-tre-nye-personbiler-er-elbiler (accessed: 01.05).

Steen-Olsen, K., Wood, R. & Hertwich, E. G. (2016). The Carbon Footprint of Norwegian Household Consumption 1999–2012. *Journal of Industrial Ecology*, 20 (3): 582-592. doi: https://doi.org/10.1111/jiec.12405.

Stern, D. I. (2003). The Environmental Kuznets Curve☆. In *Reference Module in Earth Systems and Environmental Sciences*: Elsevier.

T. Moe. (2012). *Norwegian Climate Policies 1990-2010: Principles, Policy Instruments and Political Economy Aspects*. Working paper: CICERO Center for International Climate and Environmental Research - Oslo. Available at: http://hdl.handle.net/11250/191854 (accessed: 18.03).

Torvanger, A. (1991). Manufacturing sector carbon dioxide emissions in nine OECD countries, 1973–87: A Divisia index decomposition to changes in fuel mix, emission coefficients, industry structure, energy intensities and international structure. *Energy Economics*, 13 (3): 168-186. doi: https://doi.org/10.1016/0140-9883(91)90018-U.

Trotta, G. (2020). Assessing energy efficiency improvements and related energy security and climate benefits in Finland: An ex post multi-sectoral decomposition analysis. *Energy*

Economics, 86: 104640. doi: https://doi.org/10.1016/j.eneco.2019.104640.

UN Security Council SC/14445. (2021). *Climate Change 'Biggest Threat Modern Humans Have Ever Faced', World-Renowned Naturalist Tells Security Council, Calls for Greater Global Cooperation*. Available at: https://www.un.org/press/en/2021/sc14445.doc.htm (accessed: 18/03).

US EIA. (2022). *Crude Oil Prices: Brent - Europe [ACOILBRENTEU]*: FRED, Federal Reserve Bank of St. Louis. Available at: https://fred.stlouisfed.org/series/ACOILBRENTEU (accessed: 13.05).

Verenich, I., Dumas, M., Rosa, M. L., Maggi, F. M. & Teinemaa, I. (2019). Survey and Cross-benchmark Comparison of Remaining Time Prediction Methods in Business Process Monitoring. *ACM Trans. Intell. Syst. Technol.*, 10 (4): Article 34. doi: 10.1145/3331449.

Wei, S., Yuwei, W. & Chongchong, Z. (2018). Forecasting CO2 emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine. *Environmental Science and Pollution Research*, 25 (29): 28985-28997. doi: 10.1007/s11356-018-2738-z.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*: South-Western Cengage Learning.

Xu, X. Y. & Ang, B. W. (2013). Index decomposition analysis applied to CO2 emission studies. *Ecological Economics*, 93: 313-329. doi: https://doi.org/10.1016/j.ecolecon.2013.06.007.

Yamakawa, A. & Peters, G. P. (2011). STRUCTURAL DECOMPOSITION ANALYSIS OF GREENHOUSE GAS EMISSIONS IN NORWAY 1990–2002. *Economic Systems Research*, 23 (3): 303-318. doi: 10.1080/09535314.2010.549461.

Yao, C., Feng, K. & Hubacek, K. (2015). Driving forces of CO2 emissions in the G20 countries: An index decomposition analysis from 1971 to 2010. *Ecological Informatics*, 26: 93-100. doi: https://doi.org/10.1016/j.ecoinf.2014.02.003.

York, R., Rosa, E. A. & Dietz, T. (2003). STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. *Ecological Economics*, 46 (3): 351-365. doi: https://doi.org/10.1016/S0921-8009(03)00188-5.

Ziegler, F., Winther, U., Hognes, E. S., Emanuelsson, A., Sund, V. & Ellingsen, H. (2013). The Carbon Footprint of Norwegian Seafood Products on the Global Seafood Market. *Journal of Industrial Ecology*, 17 (1): 103-116. doi: https://doi.org/10.1111/j.1530-9290.2012.00485.x.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2): 301-320. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

8 Appendix A: Note on the machine learning methods applied

Clustering algorithms applied (K-means and DB-scan)

Purpose of using these is to help find patterns in the data. K-means uses Euclidean distance functions to look for similarity between observations. Observations that form their own blob are labelled as a cluster by the algorithm.

- Show equations of model
- State usage of silhouette score for deciding on optimum amount of clusters
- This step is done as part of the

Elastic-net regularised regression

Regularised regression introduces penalty terms in the estimation of parameters. It is a group of methods where variable selection are built into the model. It seeks to lower both bias and variance in predictions by lowering the importance of coefficients for variables with low predictive power (Hastie et al., 2009). This is why models that applies regularised regression are often referred to as "shrinkage methods". The fundamental idea behind these types of models is that it introduces a penalty term λ into the estimation of parameters. This changes the way coefficients relate to each other.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} |\boldsymbol{y} - \boldsymbol{X\beta}|^2 + \lambda_1 \boldsymbol{\beta}^2 + \lambda_2 |\boldsymbol{\beta}| \qquad \text{(A.1)}$$

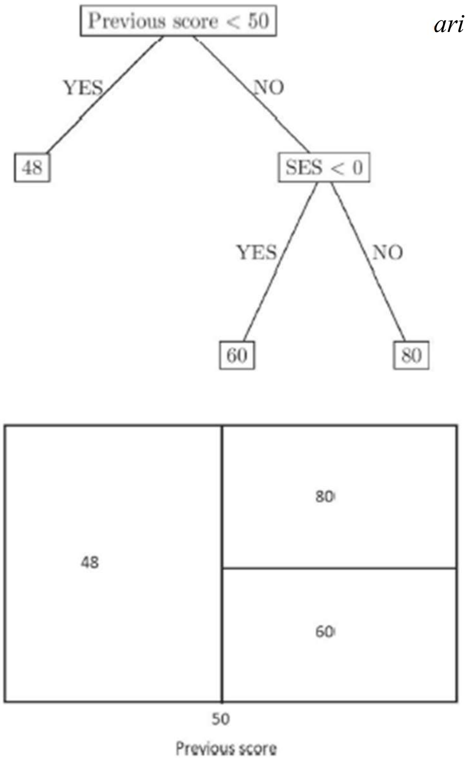
A popular rendition of these methods is the Elastic-net model. Equation 4.1 shows how the estimation of a beta parameter is changed by the two penalty terms applied. The two penalty terms are used in combination and were derived from the Lasso and Ridge regression models. Zou and Hastie (2005), which pioneered the model point to how the inclusion of these two terms together help overcome some of challenges associated with the usage of the terms on their own. This especially applies situations where there are strong correlations between variables but where a model needs implicit variable selection. These are scenarios where the lasso can be unstable since it will tend to select one of the two correlating variables. At the same time it solves the problem Ridge regression has when faced with many variables that do not add to the prediction as ridge regression penalty terms can never reach 0. How this happens becomes evident when considering the extended notation, the penalty term as the model creates a linear combination out of the two penalty terms based on values of α on the interval (0, 1). In the Elastic-net model both λ and α are hyperparameters which are set outside of the model and must thus be optimised for using grid-search.

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right) \tag{A.2}$$

The model tends to do well in scenarios where p >> n, where there are many correlating variables as “The second term encourages highly correlated features to be averaged, while the first term encourages a sparse solution in the coefficients of these averaged features” (Hastie et al., 2009, p. 662).

Random forest regressor

Fig *ariate*



In order to understand random forests, it is necessary to build a basic understanding of regression trees which form the basic building block of the model. The model was originally laid out in L. Breiman (1984) but the explanations here will be from Schiltz et al. (2018) and Hastie et al. (2009). Regression trees and forests are, often referred to as CART, are models that do not assume any specific functional form of the variables. The model fits a tree based on simple if-then rules that the model identifies in the data. Each covariate in the model is allocated a space “[…] where, the predicted value of the response variable within each region can be obtained as the mean of all the observations that belong to each region” (Schiltz et al, 2018, p. 3). I.e., the model splits the data into non-overlapping groups based on the mean within the respective group and then it arranges them in hierarchical form. Each group corresponds to a leaf in the tree. Each new branch in the tree is based on the threshold value which minimises the squared sum of errors for the mean of each group. There can only be two branches per group. Schiltz et al. (2018, p. 3) define the model of regression trees as

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \qquad \text{(A.3)}$$

Where R_m are the regions in the covariate space c_m is the mean of all observations belonging to that covariate space R_m. Splitting variables for each point are according to Hastie et al. (2009, p. 307) given by the model

$$R_1(j, s) = \{X|X_j \leq s\} \text{ and } R_2(j, s) = \{X|X_j > s\}. \qquad \text{(A.4)}$$

Which splits for the s that minimises the equation

$$\min_{j,\, s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right] \qquad \text{(A.5)}$$

where for any variable j and split point s the solution is given by

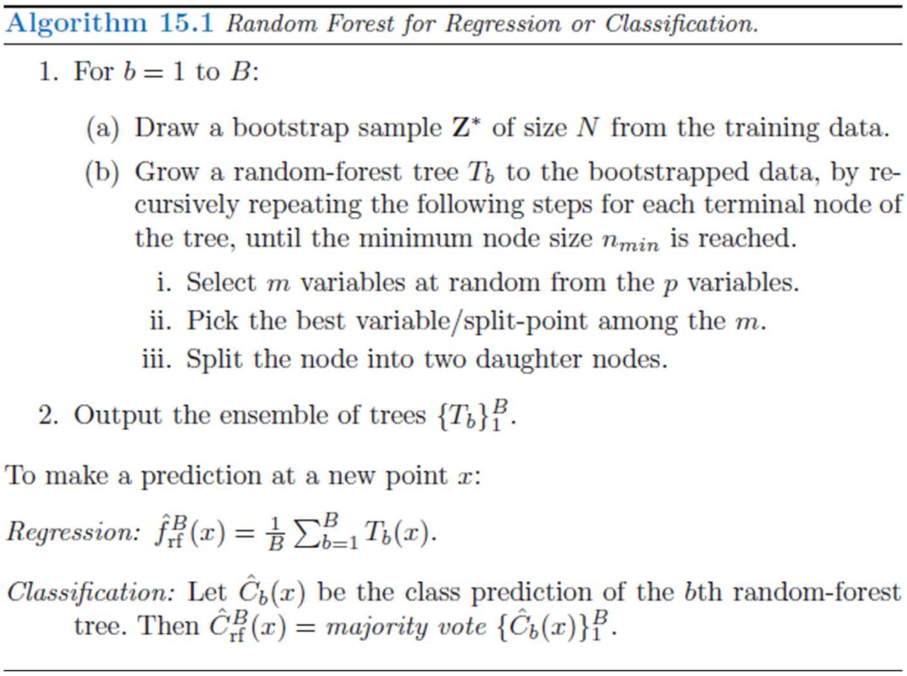
$$\hat{c}_1 = \text{ave}(y_i|x_i \in R_1(j, s)) \text{ and } \hat{c}_2 = \text{ave}(y_i|x_i \in R_2(j, s)) \qquad \text{(A.6)}$$

The size of a tree is governed by a cost-complexity pruning function which effectively works against how large the tree can grow. Mathematically the cost complexity criterion is given by

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|. \qquad \text{(A.7)}$$

Where $|T|$ is the number of terminal nodes in the tree, $Q_m(T)$ is the average sum of squares for the average of each leaf and N_m is the number of observations per node m in the tree. For any α, the branch $T_\alpha \subseteq T_0$ (for branch α belonging to tree T_0) one wants to minimise the cost complexity criterion. α is a hyperparameter set outside the model and it governs how large a tree can grow.

On their own individual regression trees often suffer from a propensity to overfit and small differences in the data can prompt new branches making the final fit of the model somewhat arbitrary. Furthermore, if the original split at the bottom of the tree is wrong, errors propagate through the entire tree. As such it is a method that has very high variance, but often leads to little bias. These problems can be solved by fitting many trees and then averaging the results according to some rule. This process relies on *bagging* (selection of subsamples with replacement) and gives rise to the *Random Forest* model. Since the model fits multiple models and then averages results from each fit it becomes less prone to overfitting and can thus handle more input variables as well as yielding a higher degree of accuracy. The full algorithm for the Random Forest model is given by:



Algorithm 15.1 *Random Forest for Regression or Classification.*

```
1. For $b = 1$ to $B$:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
   (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by re-
       cursively repeating the following steps for each terminal node of
       the tree, until the minimum node size $n_{min}$ is reached.
         i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
       iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Figure 18 - Random Forest algorithm according to by Hastie et al. (2009, p. 588)

The expected value of the forest is the same as each individual tree, this ensures that the model keep the low bias inherent to regressor trees. In order to minimise the variance the model randomly selects a subset of the input variables before it fits each tree such that after B trees the random forest regressor is according to Hastie et al (2009, p. 589)

$$\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b). \qquad \text{(A.8)}$$

$T(x;\Theta_b)$ is the formal expression of the tree and Θ are the parameters which the trees estimate. Finally, the random forest model leaves a proportion of the dataset out when fitting its trees. This is referred to as the out of bag (OOB) sample. Once the Random Forest model has finishes fitting its trees it runs the OOB samples through the created regression trees in order to validate its build. The prediction error is given by the amount of OOB samples which are correctly estimated by the individual trees and yields the stopping criterium for ending the cycle of fitting the model. In other words, the estimation stops once the OOB error stabilises (Hastie et al., 2009, p. 593).

Feature importance in Random Forests

The mean decrease in impurity is a way to measure variable importance in tree-based models like Random Forest and XGboost and is Sci-kit learn’s default inbuilt method for seeing which variables good contributors to predicting the target variable successfully. The number is calucated by the formula

$$G = \sum_{i=1}^{n_c} p_i(1-p_i) = 1 - \sum_{i=1}^{n_c} p_i^2$$

This measures the risk of misclassification of groups inside the model. In effect it is a measure that measures how well variables cause splits in the random forest model when it splits trees. The inclusion of this measure enables the RF model to be used as a feature selector because utltimately, many variables will cause very few splits in the model, while others will cause many. As such one can select the n most predictive variables from the model as features used inside other models. As Random Forest regressor model is built on the regression tree model it doesn’t making any assumptions about the distribution or relationships within the data either.

XGboost

XGboost is a random forest model taken further. It is a very complicated model so this text will only seek to build an intuitive understanding of what the model does. The approach is the same where many weak models are put into an ensemble, and together they provide much stronger predictions than on their own. Like in the random forest, these weak learners are regression trees. Each weak learner has a prediction rate which is only slightly better than the expected value and the learns iteratively by employing gradient descent to minimise a given loss function (usually prediction error measures). This means that it remembers the error given when fitting a specific variable. This means the if a variable is proven to give a large error, it is weeded out from the prediction of the algorithm. It also has inbuilt regularisation for the fitting of trees (like the Elastic-net model). This is why the model is providing good results in nearly all instances. Like the random forest the model doesn't make any distributional assumptions and have been used to make predictions on cross-sectional (Shwartz-Ziv & Armon, 2022), panel (Chen, 2021) and time-series data(Nielsen, 2016)

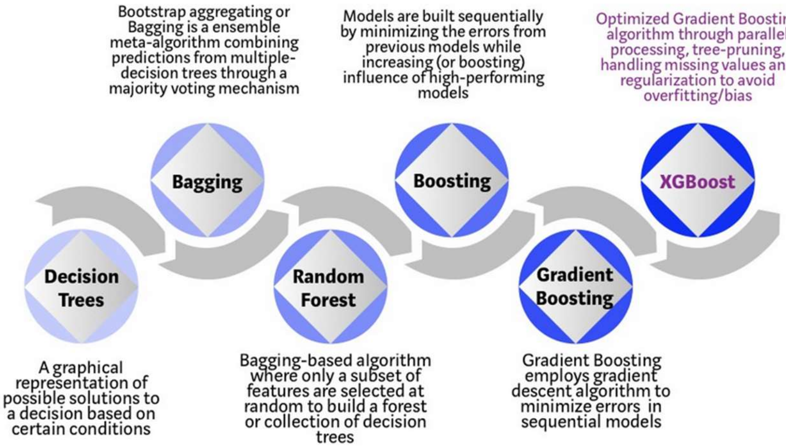
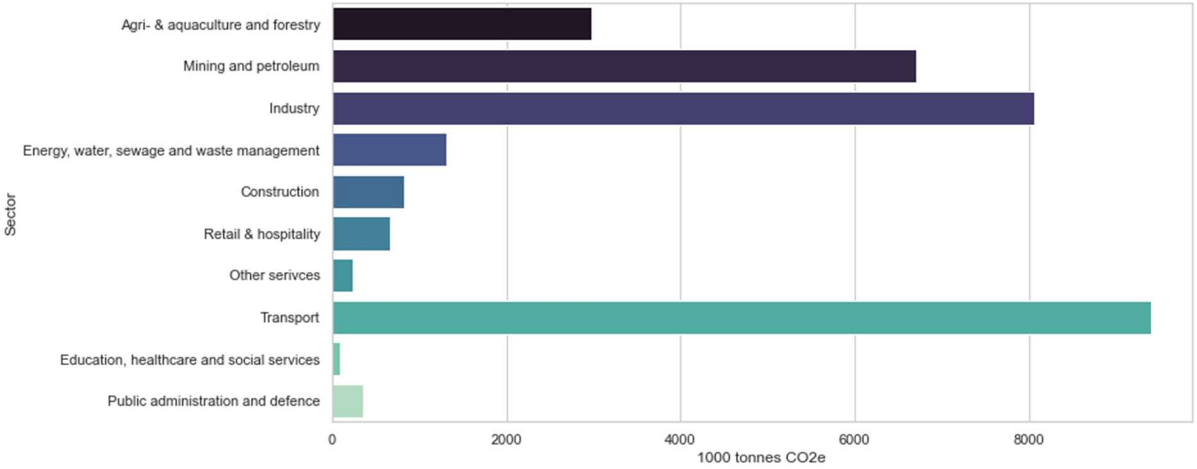

Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism
Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models
Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias
Bagging
Boosting
XGBoost
Decision Trees
Random Forest
Gradient Boosting
A graphical representation of possible solutions to a decision based on certain conditions
Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees
Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models

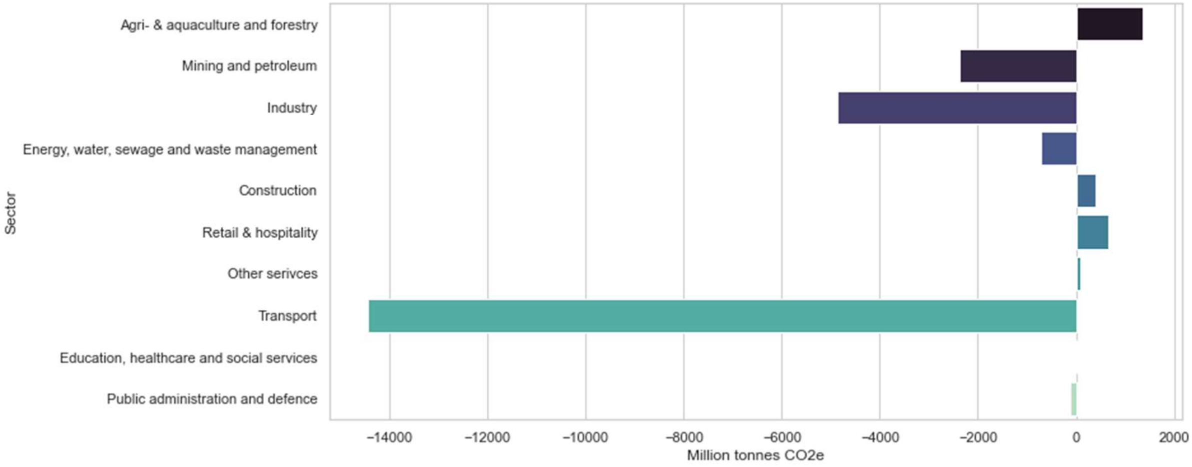
Figure 19 – Evolution of tree based models (Morde, 2019)

9 Appendix B: Sectoral results from LMDI-IDA analysis

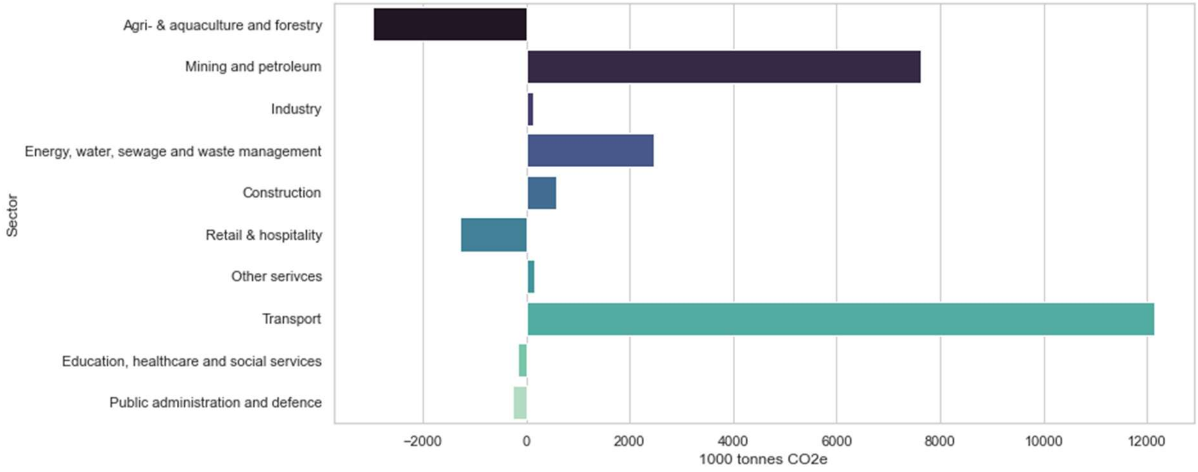
Effect of total GDP-growth on sectoral emissions, 1990-2019

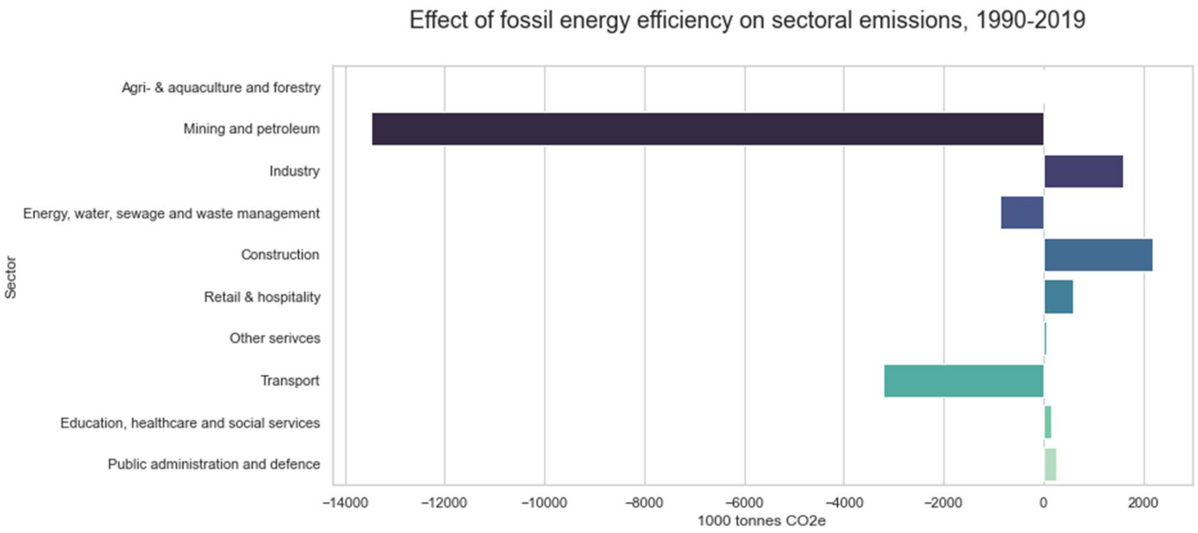


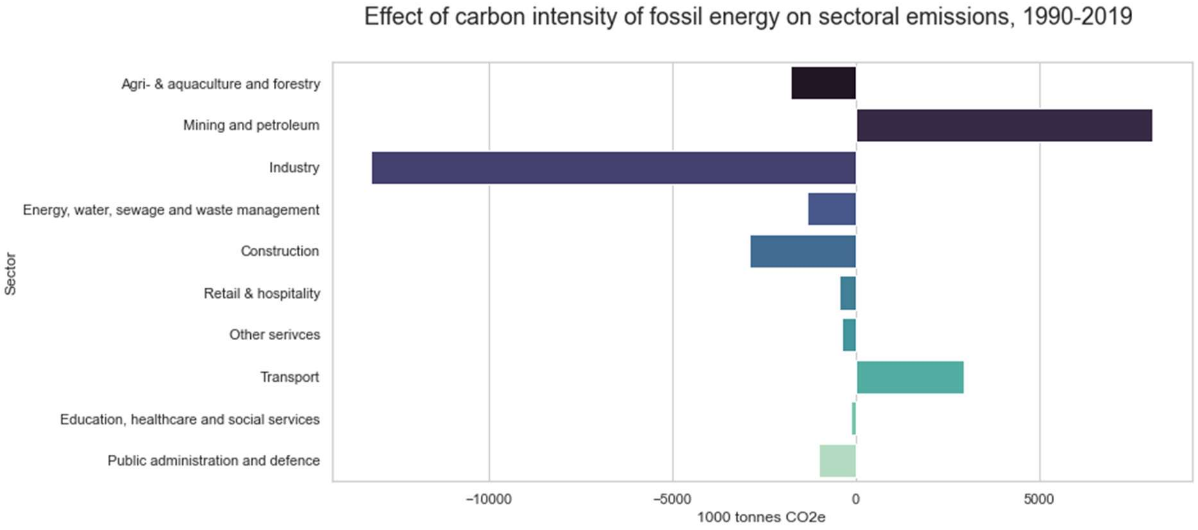
Effect of structural change on sectoral emissions, 1990-2019



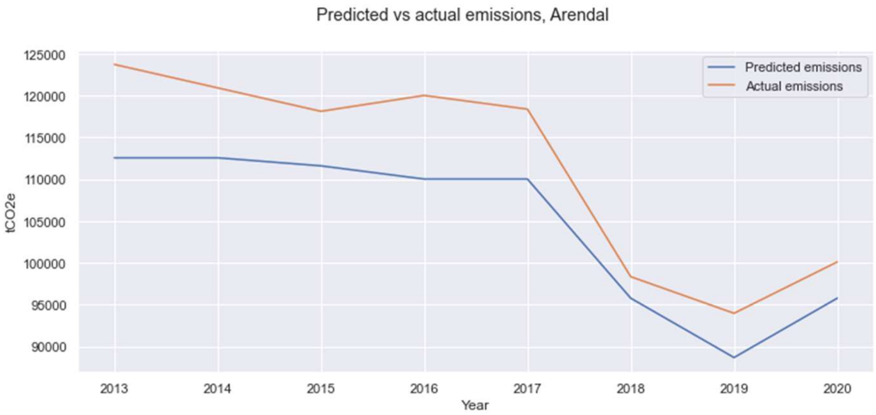
Effect of energy intensity on sectoral emissions, 1990-2019

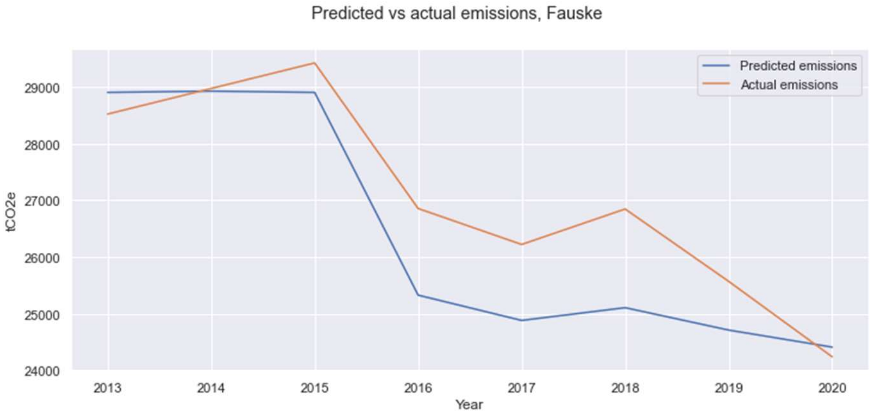


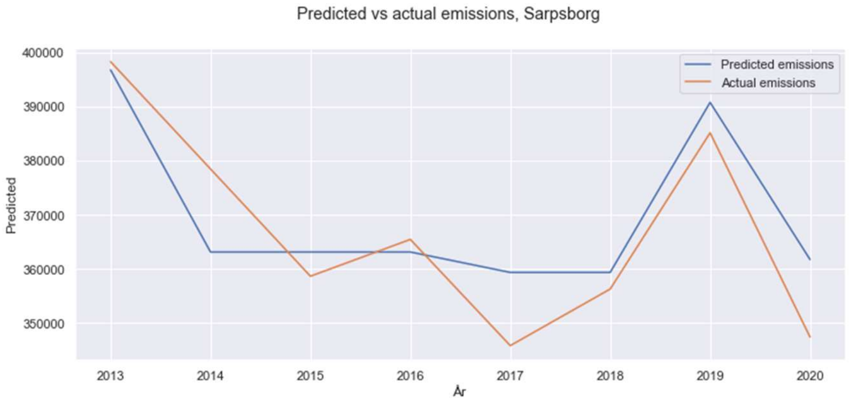

Effect of fossil energy efficiency on sectoral emissions, 1990-2019
Sector
Agri- & aquaculture and forestry
Mining and petroleum
Industry
Energy, water, sewage and waste management
Construction
Retail & hospitality
Other serivces
Transport
Education, healthcare and social services
Public administration and defence
−14000
−12000
−10000
−8000
−6000
−4000
−2000
0
2000
1000 tonnes CO2e

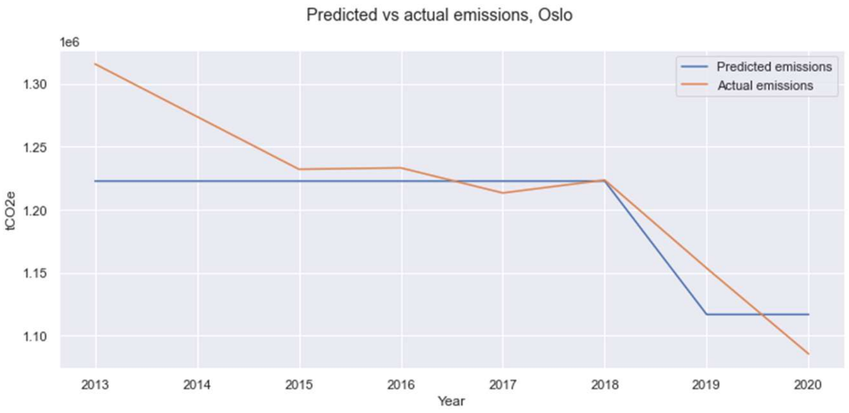

Effect of carbon intensity of fossil energy on sectoral emissions, 1990-2019
Sector
Agri- & aquaculture and forestry
Mining and petroleum
Industry
Energy, water, sewage and waste management
Construction
Retail & hospitality
Other serivces
Transport
Education, healthcare and social services
Public administration and defence
−10000
−5000
0
5000
1000 tonnes CO2e

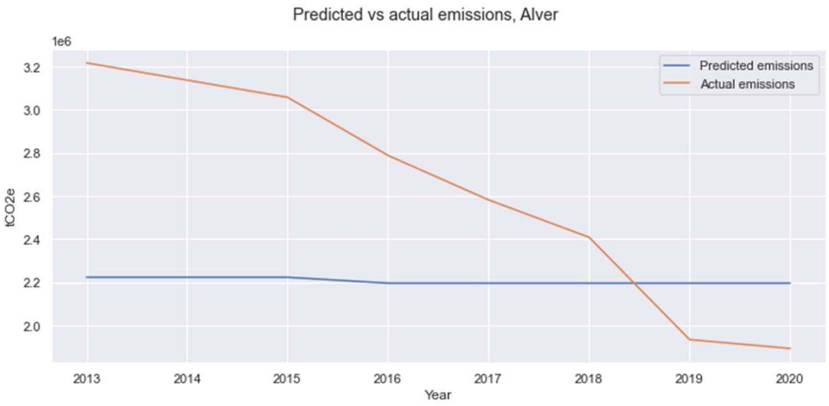
10 Appendix D: Predicted vs observed results of individual municipalities










Predicted vs actual emissions, Alver
1e6
3.2
3.0
2.8
2.6
2.4
2.2
2.0
tCO2e
2013
2014
2015
2016
2017
2018
2019
2020
Year
Predicted emissions
Actual emissions

11 Appendix D: Hyper parameters for models estimated

Model	*Parameter combination*
RF feature selector	n_estimators: 40
F_regression feature selector	Select 15 strongest correlations
XG-boost (cross sectional)	Learning rate: 0.05 Gamma: 0.05 Max depth: 60 N_estmators: 600
Elastic net	Alpha: 0.04 L1 ratio: 0.47 Tolerance: 10
XG-boost (Panel)	Learning rate: 0.05 Gamma: 0.05 Max depth: 60 N_estmators: 200
Elastic net	Alpha: 0.02 L1 ratio: 0.05 Tolerance: 5

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway