Development of an emissions forecasting model for South African industrial facilities

S Viljoen

orcid.org/0000-0002-9036-6609

Dissertation accepted in fulfilment of the requirements for the degree *Master of Engineering in Mechanical Engineering* at the North-West University

Supervisor: Dr J Pascoe

Graduation: July 2022

Student number: 26985993

ACKNOWLEDGEMENTS

I would like to express my thanks to the following people who made a significant contribution to the success of this study:

- Thanks to Mine Clever (Pty) Ltd and its sister companies for providing financial support to complete this study.
- Thanks to my study leader, Dr Janine Pascoe, for your assistance, guidance, and valuable inputs.
- A special thanks to Dr Andries Gous. Without your inputs and mentorship, none of this would be possible.
- Thanks to my mom and dad for always supporting me and encouraging me to reach for my dreams.
- Thanks to all my co-workers and friends. Your support means more than you could imagine.
- Finally, I would like to thank God for providing me with the strength and perseverance to complete this study.

ABSTRACT

Title: Development of an emissions forecasting model for South African industrial facilities

Author: S Viljoen

Supervisor: Dr J Pascoe

Keywords: mitigation, net-zero emissions, carbon budget, emissions forecasting, emissions reduction

Greenhouse gas (GHG) emissions are one of the main contributors to climate change.

Several mitigation measures such as carbon tax, carbon budgets and sectoral emissions targets (SETs) are used to reduce GHG emissions. These mitigation measures require long-term plans, detailing how reduction outcomes or targets will be achieved.

Carbon emissions forecasting is vital to ensure effective carbon mitigation measures. Forecasted emissions give a good indication of whether a facility's emissions targets will be met and if more stringent mitigation measures are required.

Several forecasting models have been developed to predict emissions at a country-specific level. However, mitigation policies aim to reduce carbon emissions at a facility level. Thus, a forecasting model is required to predict carbon emissions for South African facilities.

An emissions forecasting model was developed based on the basic forecasting model development steps; focussing on problem definition, data identification, preliminary analysis, choosing and fitting models, and using and evaluating the model.

Different forecasting models were evaluated, whereafter the models best suited for the data were used. The time-series variables, material input and production data, were estimated using the error-trend-seasonality (ETS) forecasting model, the activity data was calculated using multivariable linear regression, and carbon emissions were predicted using linear regression. The developed forecasting model is then validated and tested on an industrial facility in South Africa.

After testing and validating the model, the model is used to predict the baseline material input, production, activity data and ultimately the emissions. The developed model forecasted the fuel combustion and process emissions within a 6% error. The fugitive emissions were forecasted within a 23% error.

Using the baseline forecast, the facility's emission roadmap was incorporated into the forecast to predict the emissions when mitigation measures are implemented. Next, the forecasted emissions are compared to the allowable emissions to determine whether the facility is likely to meet its emission targets.

The chosen facility aims to reduce its Scope 1 and Scope 2 emissions by 30% by 2030. To achieve this target, an emissions reduction of 18.5% is required by 2025. By only implementing a 10% reduction in material input, a 9% reduction in Scope 1 emissions is forecasted. Thus, the implementation of additional mitigation measures are required to ensure the facility meet its emission reduction targets.

In addition to using long-term forecasts to determine if a facility will meet its emission targets, short-term forecasting can be used for setting up carbon budgets. During the first carbon tax phase, companies can receive an additional 5% allowance on their carbon tax liability when participating in the carbon budget system. This amounts to a combined total of about R 2.3-billion in avoided carbon tax liability when considering six South African industries.

Ultimately, the developed forecasting models can be used to develop accurate mitigation plans for any South African industrial facility.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS I

ABSTRACT II

TABLE OF CONTENTS IV

LIST OF FIGURES VI

LIST OF TABLES X

NOMENCLATURE XII

LIST OF ABBREVIATIONS XIV

CHAPTER 1 – INTRODUCTION AND LITERATURE 1

1.1 Preamble 1

1.2 Protocols and policies 1

1.3 GHG mitigation and sustainability strategies 6

1.4 Emissions forecasting and forecasting models 11

1.5 Significance of the study 17

1.6 Problem statement and research objective 18

1.7 Scope of investigation 19

1.8 Dissertation layout 19

CHAPTER 2 - METHODOLOGY 21

2.1 Preamble 21

2.2 Gathering information 22

2.3 Preliminary analysis 29

2.4 Choosing and fitting models 33

2.5 Using and evaluating the forecasting model 41

2.6 Conclusion 47

CHAPTER 3 – RESULTS AND DISCUSSION 49

3.1 Preamble 49

3.1 Case study selection 50

3.2 Application of emissions forecasting model on industrial facility 51

3.3 Conclusion 87

CHAPTER 4 - CONCLUSION 89

4.1 Preamble 89

4.2 Summary of the study 89

4.3 Limitations and recommendations 94

4.4 Meeting the required objectives 94

4.5 Closure 95

BIBLIOGRAPHY 96

APPENDIX A - SELECTED FORECASTING MODELS 107

APPENDIX B – MODEL DEVELOPMENT AND MODEL VALIDATION 117

LIST OF FIGURES

Figure 1-1: National and international climate change protocols and policies. 2

Figure 1-2: Overview of carbon emission categories across the value chain (adapted from the Greenhouse Gas Protocol). 9

Figure 1-3: Breakdown of forecasting models used in literature. 17

Figure 2-1: Basic steps for forecasting model development. 21

Figure 2-2: Sub-steps for the gathering information step. 22

Figure 2-3: Missing data in a time series. 25

Figure 2-4: Statistical interpolation techniques. 26

Figure 2-5: Time series interpolation techniques. 27

Figure 2-6: Sub-steps in the preliminary analysis step. 29

Figure 2-7: Horizontal pattern. 30

Figure 2-8: Trend pattern. 30

Figure 2-9: Seasonal pattern. 30

Figure 2-10: Cyclic pattern. 30

Figure 2-11: Additive and multiplicative seasonality. 31

Figure 2-12: Trend removal. 33

Figure 2-13: Sub-steps when choosing and fitting models. 34

Figure 2-14: Forecasting model selection. 38

Figure 2-15: Model development steps. 40

Figure 2-16: Illustration if k-fold cross-validation. 40

Figure 2-17: Sub-steps in using and evaluating the forecasting model. 41

Figure 2-18: Mean absolute error for testing forecasting accuracy. 44

Figure 2-19: Allowable emissions in the target year. 46

Figure 2-20: Validate whether emissions targets will be met. 46

Figure 2-21: Steps and sub-steps for forecasting model development. 47

Figure 3-1: Steps for forecasting model development (repeat of Figure 2-1). 49

Figure 3-2: Contribution of the facility's emissions. 50

Figure 3-3: Sub-steps for the gathering information step. 51

Figure 3-4: Time series graphs used for identifying missing or incorrect values. 55

Figure 3-5: Sub-steps in the preliminary analysis step. 57

Figure 3-6: Material input time series data. 58

Figure 3-7: Material input data for three consecutive years. 59

Figure 3-8: Production time series data. 59

Figure 3-9: Production time series data for three consecutive years. 60

Figure 3-10: Relationship between the different variables. 61

Figure 3-11: Trend removal of material input data. 63

Figure 3-12: Trend removal of production data. 64

Figure 3-13: Sub-steps when choosing and fitting models. 65

Figure 3-14: Forecasting model selection. 66

Figure 3-15: Model development steps. 67

Figure 3-16: 5-fold cross-validation illustration. 67

Figure 3-17: Forecasted material input data. 69

Figure 3-18: Forecasted production data. 69

Figure 3-19: Fuel combustion emission 2 linear regression model. 71

Figure 3-20: Forecasting accuracy of the material input and production data. 73

Figure 3-21: Accuracy of the forecasted carbon emissions. 76

Figure 3-22: Accuracy of forecasted emissions. 76

Figure 3-23: Sub-steps in using and evaluating the forecasting model. 77

Figure 3-24: Comparison of ETS forecasting using Excel and Python. 79

Figure 3-25: Contribution of emissions. 83

Figure 3-26: Forecasted carbon emissions for 2019 to 2025. 86

Figure 3-27: Steps and sub-steps for forecasting model development. 87

Figure 4-1: Basic forecasting model development steps and sub-steps. 90

Figure 4-2: Forecasting model development steps. 91

Figure 4-3: Accuracy of forecasted emissions. 92

Figure 4-4: Total forecasted Scope 1 emissions. 93

Figure A-1: Line of best fit. 109

Figure A-2: Linear regression model with a perfect linear relationship. 110

Figure B-1: Linear relationship between the different variables. 117

Figure B-2: A data distribution plots. 120

Figure B-3: Autocorrelation test for the activity data. 121

Figure B-4: Homoscedasticity test for the activity data. 122

Figure B-5: Regression model outlier analysis. 123

Figure B-6: Linear regression models for the fuel combustion emissions. 124

Figure B-7: Fuel combustion emissions distribution plots. 125

Figure B-8: Autocorrelation test for the fuel combustion emissions. 126

Figure B-9: Homoscedasticity test for fuel combustion emissions. 126

Figure B-10: Regression models for the process emissions. 127

Figure B-11: Process emissions distribution plots..... 129

Figure B-12: Homoscedasticity test for the process emissions. 130

Figure B-13: Comparison of the forecasted activity data using Excel and Python. 131

Figure B-14: Comparison of the forecasted emissions using Excel and Python. 134

LIST OF TABLES

Table 1-1: Adaptation and mitigation objectives and interventions of SA's NCCRP..... 4

Table 1-2: Risks associated with an increase in global temperatures..... 7

Table 1-3: Summary of emissions forecasting models available in literature. 13

Table 1-4: Impact of carbon budget allowances on the total carbon tax liability..... 18

Table 2-1: Data quality dimensions. 23

Table 2-2: Advantages and disadvantages of different forecasting models. 35

Table 2-3:Advantages and disadvantages of Python. 43

Table 3-1: Summary of data variables. 52

Table 3-2: Breakdown of data points collected. 52

Table 3-3: Summary of outliers. 56

Table 3-4: Material input trendline parameters..... 62

Table 3-5: Production data trendline parameters..... 64

Table 3-6: Equations used for forecasting the activity. 70

Table 3-7: Equations used to forecast the fuel combustion emissions. 71

Table 3-8: Equations used to forecast the process emissions. 71

Table 3-9: Equation used to forecast the fugitive emissions. 72

Table 3-10: Comparison of the average 2025 input data using Excel and Python..... 80

Table 3-11: Comparison of the average 2025 activity data using Excel and Python. 80

Table 3-12: Comparison of the total 2025 emissions using Excel and Python. 80

Table 3-13: Input and activity data forecasting accuracy statistics..... 81

Table 3-14: Fuel combustion emissions forecasting accuracy statistics. 81

Table 3-15: Process emission forecasting accuracy statistics. 82

Table 3-16: Fugitive emissions forecasting accuracy statistics. 82

Table 3-17: Total emissions forecasting accuracy statistics. 83

Table 3-18: Reduction in emissions with the implementation of mitigation measures. 86

Table 4-1: Forecasting accuracy statistics. 92

Table A-1: Output matrix of the LINEST function. 111

Table A-2: Statistical parameters from output matrix. 111

Table B-1: Multivariable linear regression parameters for the activity data 118

Table B-2: Linear regression parameters for the fuel combustion emissions. 124

Table B-3: Constant multipliers used for the fuel combustion emissions. 125

Table B-4: Linear regression parameters for the process emissions 128

Table B-5: Constant multipliers used for the process emissions. 128

Table B-6: Multicollinearity test for the process emissions. 129

Table B-7: Constant multipliers used for calculating the fugitive emissions. 130

Table B-8: Comparison of the average 2025 activity data using Excel and Python 132

Table B-9: Comparison of the total 2025 emissions using Excel and Python. 134

NOMENCLATURE

Terminology	Description
Activity data:	Data on an activity that results in emissions during a given period.
Business-as-usual:	Normal operations of a company to continue with the execution of standard functions.
Carbon budget:	The amount of GHG emissions permitted.
Carbon dioxide equivalent:	A metric measure used to compare GHG emissions based on the global warming potential.
Carbon tax:	Taxation mechanism to reduce GHG emissions.
Emissions roadmap:	A flexible plan that sets out the emissions reduction targets of a company including financial instruments needed to reach these targets.
Garbage in garbage out:	The output of the system is only as good as the data that is put into the system.
Materiality:	Concept to evaluate the importance of an identified misstatement and the effect on the overall data. The materiality limit as stated in the National Greenhouse Gas Emission regulations is 5%.
Mitigation measure:	Means to prevent, reduce or control GHG emissions.
Mitigation plan:	Plan detailing how the facility's emission targets will be met.

Polluter pays principle:	Those who pollute should bear the cost of managing pollution to prevent damage to human health or to the environment.

LIST OF ABBREVIATIONS

ANNs	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average
BECCS	Bioenergy carbon capture and storage
CO_2	Carbon dioxide
CO_2e	Carbon dioxide equivalent
CSV	Comma separated value
DFFE	Department of Forestry, Fisheries, and the Environment
EPI	Environmental Performance Indicator
ETS	Error-Trend-Seasonality
GDP	Gross Domestic Product
GHG	Greenhouse gas
GIGO	Garbage In, Garbage Out
$GtCO_2e$	Gigaton carbon dioxide equivalent
INDC	Intended Nationally Determined Contributions
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquartile range
LEDS	Low Emissions Development Strategy
LT-LEDS	Long-term low greenhouse gas emission development strategies
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error

MVRA	Multivariable regression analysis
$MtCO_2e$	Megaton carbon dioxide equivalent
NCCAS	National Climate Change Adaptation Strategy
NCCRP	National Climate Change Response Policy
NDC	Nationally Determined Contributions
RMSE	Root Mean Squared Error
SETs	Sectorial Emissions Targets
UNFCCC	United Nations Framework Convention on Climate Change
TSV	Tab separated value
XLSX	Microsoft Excel Open XML Spreadsheet
XML	Extensible markup language

CHAPTER 1 – INTRODUCTION AND LITERATURE

1.1 Preamble

Climate change is one of the most significant environmental challenges that threaten the earth's sustainability. One of the main contributors to climate change is anthropogenic greenhouse gas (GHG) emissions [1]. Since the start of the industrial era, carbon dioxide (CO_2) concentrations in the atmosphere have increased over 30%, resulting in an approximate 0.6 °C rise in the global surface temperature [3].

As the surface temperature increases, rainfall patterns become more variable and unpredictable. These changes in rainfall patterns impact water sources and the production of crops and increase the vulnerability of impoverished communities [2]. As a developing country, South Africa is very vulnerable to the effects of climate change [4].

According to Climate Watch Data[a], the global GHG emissions increased from 3.26 $GtCO_2e$[b] emissions to 4.89 $GtCO_2e$ between 1990 and 2018. In 2018, South Africa was ranked as the 13th highest carbon dioxide emitter globally, emitting 0.46 Gt carbon dioxide emissions annually[c].

The energy sector is the largest contributor to South Africa's total annual emissions (excluding emissions from Forestry and Other Land Use), with a contribution of 79% in 2017 [5]. South Africa has a high dependency on fossil fuels for energy generation with about 86% of the electricity generated in South Africa originating from coal-fired stations [6].

To mitigate GHG emissions and adapt to the effects of climate change, several protocols and policies were implemented.

1.2 Protocols and policies

Several international policies and agreements have been developed at a global level to prevent and alleviate the risks associated with global climate change. These policies include

[a] https://www.climatewatchdata.org/data-explorer/historical-emissions

[b] The carbon dioxide equivalent (CO_2e) is the number of metric tonnes of CO_2 emissions with the same global warming potential as one metric ton of another greenhouse gas

[c] https://www.ucsusa.org/resources/each-countrys-share-co2-emissions

the United Nations Framework Convention on Climate Change (UNFCCC), the Kyoto Protocol, and the Paris Agreement which ensure that all countries play their part in reducing GHG emissions.

South Africa's commitments under the Paris Agreement necessitated the development of its Low Emissions Development Strategy (LEDS), which outlines a set of actions up to 2050 [2]. To realise the goals outlined in LEDS, several additional policies were developed focussing on growth, climate change response, climate change legislation as well as current and updated emissions results. Practical execution realised from the mitigation measures introduced under the policies include price on carbon (Carbon Tax), emissions management that places limits on sectors, and plans on how to achieve the mitigation measures at a facility level.

The goals and outcomes of these levels (as illustrated in Figure 1-1) will be discussed in further detail.

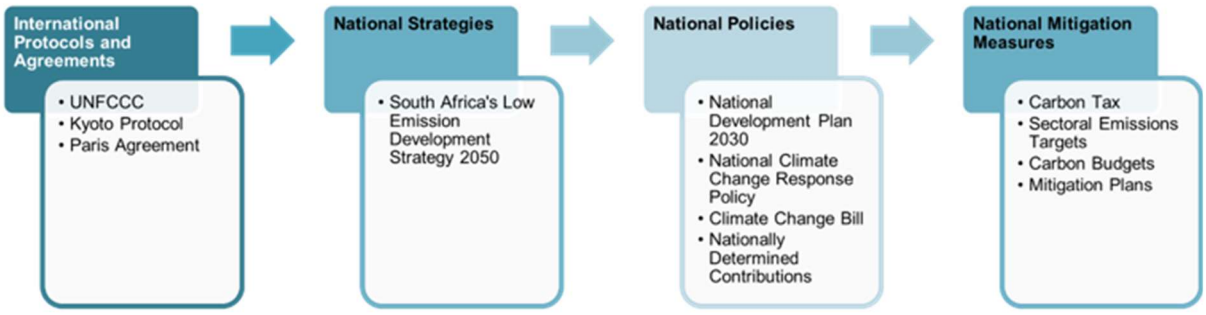


Figure 1-1: National and international climate change protocols and policies.

1.2.1 International protocols and policies

In March 1994, the UNFCCC came into effect [7]. The UNFCCC mandates that member nations stabilise GHG concentrations in the atmosphere at a level that will not cause environmental harm [1]. The UNFCCC established the following steps to reduce GHG emissions [8]:

1. Reduce GHG emissions below 1990 levels by means of an agreement (Kyoto Protocol),
2. Establish a market for trading carbon credits (a permit that allows a company to emit a specified amount of GHG emissions), and

3. Enable developing countries to move towards a low-carbon economy by means of financial aid from developed countries.

The UNFCCC did not identify any specific environmental change initially; however, it established a process through which agreements could be made to control GHG emissions [8]. One of these subsequent agreements to the UNFCCC is the Kyoto Protocol [9].

The Kyoto Protocol was adopted in December 1997. The aim of this protocol is to legally obligate developed countries to reduce their GHG emissions by setting internationally binding emissions reduction targets [10]. During the first commitment period (2008 – 2012), countries committed to reduce their GHG emissions to an average of 18% against 1990 levels [9]. During the second commitment period (2013 – 2020), countries committed to reduce their GHG emissions by at least 18% below 1990 [11].

South Africa ratified the Kyoto Protocol in 2002 pledging to reduce its GHG emissions by 34% by 2020, and 42% by 2025 compared to "business-as-usual" [12]. South Africa's commitments under the Kyoto Protocol form the basis of its national emissions reduction targets submitted to the Paris Agreement [2].

The Paris Agreement governs emission reductions following the second commitment period of the Kyoto Protocol. The agreement states that a more than 2°C increase in the global average temperature (compared to the pre-industrial era), poses a significant threat to sustainable development [1]. Unlike the Kyoto Protocol, which only requires developed countries to reduce their emissions, the Paris Agreement encourages all countries to set emission reduction targets [1, 9].

1.2.2 National strategies

The Paris Agreement encourages countries to submit long-term low greenhouse gas emission development strategies (LT-LEDS), to better frame what actions are required to limit the global surface temperature increase at 1.5 °C. LT-LEDS are submitted voluntarily [1].

Countries under the Paris Agreement are required to submit Nationally Determined Contributions (NDCs) on a 5-year basis. NDCs enable countries to define their own pledges regarding emissions targets based on the circumstance, resources, and abilities of the country. NDCs express actions taken by each country to reduce their national emissions and adapt to the effects of climate change [1]. These actions include, but are not limited to the development of an early warning, vulnerability, and adaptation monitoring system, and the development of a National Adaptation Plan [2].

1.2.3 National protocols and policies

As a member of the Paris Agreement, South Africa is required to submit its NDC every five years [1]. On 25 September 2015, South Africa submitted its intended nationally determined contribution (INDC) which became their first NDC upon the ratification of the Paris Agreement. The NDC outlines how South Africa intend to reduce its GHG emissions and adapt to the impacts of climate change while ensuring sustainable development [13].

The NDC consists of two components: a mitigation component, and an adaptation component [13]. The mitigation component focusses on the country's commitments towards reducing GHG emissions (to reduce annual GHG emissions to the range of 398-510 $MtCO_2e$ by 2025, and 398-440 $MtCO_2e$ by 2030). The adaptation component, however, explains South Africa's goals for adaptation towards mitigating the impact of climate change [14].

In 2019, the South African government published the National Climate Change Adaptation Strategy (NCCAS), which supports the country's obligations in terms of the Paris Agreement [15]. The NCCAS serves as South Africa's National Adaptation Plan and forms the basis for meeting the adaptation objectives as outlined in the NDC. While the NCCAS focusses primarily on the adaptation of climate change, it acknowledges that mitigation and adaptation have an impact on each other [2].

The NCCAS is derived from several of South Africa's adaptation policies which include the National Climate Change Response Policy (NCCRP), published in 2011. Unlike the NCCAS, the NCCRP prioritises both climate change mitigation and adaptation. The NCCRP envisions an effective climate change response and a transition to a climate resilient carbon economy and just society [2].

The objectives and interventions regarding climate change mitigation and adaptation, specifically focussing on carbon pricing, water, agriculture and commercial forestry, health, biodiversity and ecosystems, human settlements, disaster risk reduction and management are summarised in Table 1-1 [15].

Table 1-1: Adaptation and mitigation objectives and interventions of SA's NCCRP.

Component	Objective	Intervention
Adaptation	Offset South Africa's vulnerability to climate change.	Extend health protection and health promotion measures.
		Manage water resources and plan for contingencies.
		Adapt rangeland practices.

Component	Objective	Intervention
		Adaptation in agriculture.
		Implement changes in forestry practices.
		Protect plant biodiversity.
		Protect animal biodiversity.
		Develop actions that will offset South Africa's economic vulnerability to climate change response measures.
Mitigation	Create a national GHG mitigation plan that promote sustainable development in South Africa.	Promote a sustainable development path through coordinated government policy.
		Control exhaust emissions from road-going vehicles through a joint implementation strategy.
		Implement a transport sector mitigation programme.
		Develop and implement a mitigation programme aimed at the coal-mining sector.
		Implement sustainable development through coordinated policies, strategies, and incentives.
		Reduce GHG emissions in the agricultural sector.
		Facilitate the establishment and extension of forest schemes.
		Optimise waste management practice to minimise the emissions of GHGs and develop a government position to implement a waste sector mitigation programme.

The mitigation components of the NCCRP are further captured in the Climate Change Bill which forms the legislative foundation for climate change adaptation and mitigation response in South Africa [2].

The Climate Change Bill introduces several mitigation measures to meet the mitigation targets as stipulated in South Africa's updated NDC. These mitigation measures include carbon budgets, sectorial emission targets and the phase down and phase out of synthetic GHG emissions [2, 16].

The mitigation measures mentioned in the Climate Change Bill are also mentioned as cross-cutting measure in South Africa's Low Emissions Development Strategy (LEDS), which was published in February 2020 (in accordance with Article 4 of the Paris Agreement).

LEDS is developed based on three key climate policy documents: The National Development Plan, the National Climate Change Response Policy, and the Climate Change Bill. The strategy focuses primarily on GHG mitigation; however, the strategy also considers South Africa's vulnerability to the impacts of climate change and the need to build resilience to these impacts. [2].

1.2.4 National mitigation measures

To support South Africa's transition towards a low carbon economy, national mitigation measures such as carbon tax, sectoral emissions targets (SETs), carbon budgets, and mitigation plans were implemented, or are in the process of being implemented [2].

South Africa's Carbon Tax Act came into effect on 1 June 2019. The Act is based on the "polluter pays principle" and aims to put a price on carbon, adopting the negative costs of emitting GHG emissions. The 2021 tax rate is set at R134 per tonne CO_2e. The objective of the Carbon Tax Act is to encourage companies towards adopting cleaner technologies in the future [17].

Carbon tax consists of two phases: the first phase spans from 1 June 2019 to 31 December 2022, and the second phase from 2023 to 2030 [18]. The impact of carbon tax will be reviewed before the commencement of the second phase. The review will account for the progress made in GHG emissions reduction as stated in the NDC [19].

In terms of the first carbon tax phase, taxpayers that participate in the voluntary carbon budget system are eligible for a 5% carbon budget allowance [17]. The voluntary carbon budgets are only applicable until 31 December 2020 [18]. Upon the promulgation of the Climate Change Bill, carbon budgets will become mandatory, and penalties will be imposed if a company exceeds its allocated carbon budget [16].

1.3 GHG mitigation and sustainability strategies

1.3.1 Carbon budgets

Reduction in GHG emissions is crucial to mitigating climate change. One way to limit future GHG emissions is by the adoption of carbon budgets [20]. The global carbon budget represents the maximum amount of GHG emissions that humankind can emit to keep the rise in global temperatures to 2°C, or preferably well below 1.5°C, compared to pre-industrial temperatures [20].

A 1.5°C increase in global temperatures will impact health, livelihoods, food security, water supply, human security and economic growth [21]. Some of the risks associated with climate change can be found in Table 1-2 [21].

Table 1-2: Risks associated with an increase in global temperatures.

Impacted area	Associated impacts
Coastline	Due to thermal expansion of oceans, sea-levels can rise in the range of 20 and 80 cm during the next century. The increase in sea-levels decreases the effectiveness of coastal defences against floods and storms which in turn impacts the instability of beaches and cliffs.
Storms and floods	Storms and floods are some of the leading natural hazards. During periods of accelerated climate change, weather patterns are more erratic and an increase in the number of storms can be observed. An increase in global temperatures can also result in an increase in heavy rainfall in specific geographic location resulting in floods.
Drinking water	Rising temperatures can impact the access to fresh drinking water. An increase in evaporation has an impact on water supplies and can also contribute to the salinisation of irrigated agricultural lands. Groundwater supplies as well as yields of rivers and reservoirs are affected by changes in river run-offs.
Health	The transmission of infectious disease is directly affected by climatic factors such as an increase in temperatures.
Biodiversity	A threat to certain fauna and flora is posed due to these species' inability to migrate in response to climate change. This could be due to the species' geographical location or due to human activity such as farming and urbanisation.
Agriculture	Increased temperatures can result in a decrease in food production in both developed and developing countries.

Table 1-2 clearly describes the immense threat climate change is to society. To limit the increase in global temperatures to 1.5°C, it is important to stay within the total global carbon budget [21]. GHG mitigation mechanisms are mostly implemented at country level, thus it is necessary that carbon budgets are also divided into country carbon budgets [20].

In September 2020, carbon budgets and SETs were approved by the South African Government to meet its absolute reductions in terms of the LEDs [2]. The Minister of the Department of Forestry, Fisheries, and the Environment (DFFE) is responsible for allocating carbon budgets at facility level [16].

Carbon budgets are aligned with the national GHG emissions trajectory and are allocated to any person that conducts an activity listed in terms of section 24(2) of the Climate Change Bill [16]. A carbon budget must span three successive five-year periods and specify the maximum amount of GHG emissions that a person may emit during the first five-year period [16].

Companies to whom carbon budgets are allocated are required to submit a GHG mitigation plan (previously known as pollution prevention plans), detailing what mitigation actions are

required to ensure compliance with the carbon budget. Companies that do not comply with their allocated carbon budget are required to provide a description on the mitigation measures required to achieve compliance [16]. For emissions exceeding the carbon budget, a higher tax rate will apply which can be up to five times the carbon tax rate[d].

The first phase of carbon budgets (under the Carbon Tax Act) came to an end on 31 December 2020. During this first phase, companies submitted carbon budgets on a voluntary basis [17]. However, upon the promulgation of the Climate Change Bill, participation in the second and subsequent carbon budget phases will become mandatory [16].

Carbon emissions are divided into three categories also known as scopes [22]. Currently, carbon budgets are only applicable to Scope 1 emissions [17]. The scopes are categorised as follows [22]:

- *Scope 1*: Scope 1 emissions are the direct emissions resulting from activities over which the company has operational control. These emissions are further subdivided into combustion, process, fugitive, and waste emissions.
    - *Combustion emissions*: Emissions resulting from the combustion of fuels in stationary source such as turbines, boilers, and furnaces.
    - *Process emissions*: Emissions resulting from the processing of materials and chemicals such as ammonia production.
    - *Fugitive emissions*: Emissions resulting from intentional or unintentional releases such as equipment leaks.
    - *Waste emissions*: Emissions resulting from waste processing.
- *Scope 2:* Scope 2 emissions include all indirect emissions from electricity and/or heat purchased by the company.
- *Scope 3:* Scope 3 emissions include all other indirect emissions resulting from activities over which the company has no operational control. Scope 3 emissions are therefore emissions resulting from business travel, procurement etc. and form the greatest share of a company's total emissions.

An overview of scope and emissions in the value chain can be found in Figure 1-2 [22].

[d] https://static.pmg.org.za/210302CarbTax_Sys_in_South_Africa..pdf

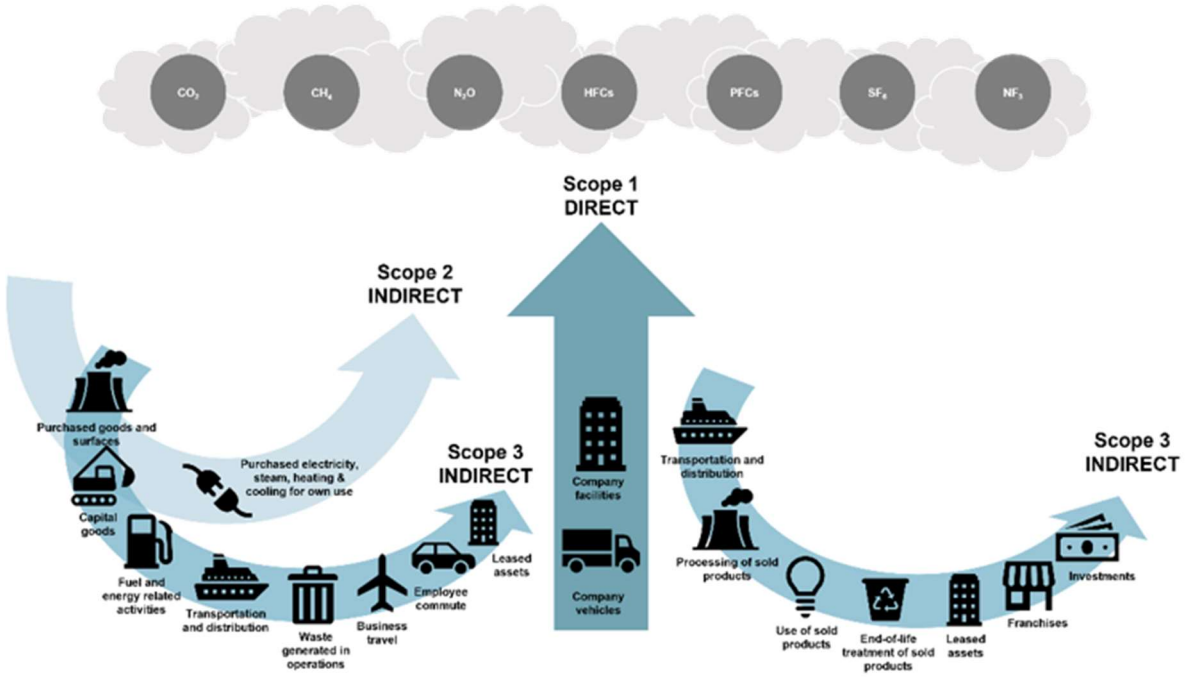


Figure 1-2: Overview of carbon emission categories across the value chain (adapted from the Greenhouse Gas Protocol).

Figure 1-2 illustrates the different emission scopes as well as the seven main GHG emissions resulting from the indicated activities. The global standard for the measuring and management of GHG emissions is known as the Greenhouse Gas Protocol. To ensure compliance with the Greenhouse Gas Protocol, companies are required to report both Scope 1 and Scope 2 emissions [22]. However, carbon tax and carbon budgets are currently only applicable to Scope 1 emissions (emissions owned by the company). Thus, this study only focusses on developing a forecasting model based on Scope 1 emissions.

It is important to note that carbon emissions are generally calculated using emission factors. The GHG Protocol defines an emission factor as “a factor allowing GHG emissions to be estimated from a unit of available activity data (e.g. tonnes of fuel consumed, tonnes of product produced) and absolute GHG emissions” [22]. Emission factors are usually expressed as the weight of the pollutant divided by a volume, weight, distance, or duration of the emission source. In most cases, emission factors are taken as averages of available data and is assumed to represent a long-term average for all facilities in the source category.

The IPCC Guidelines use a tiered approach to describe the different methods for calculating GHG emissions. There are three levels (referred to as tiers) available to calculate emissions: Tier 1, Tier 2, and Tier 3. Each tier level is associated with and increasing level of accuracy. The three levels of tiers are known as Tier 1, Tier 2, and Tier 3 and is each associated with

an increasing level of detail and accuracy, with Tier 3 calculations being the most accurate. The tiers can be described as follows [23, 24]:

- *Tier 1:* This method is the simplest method and has the highest uncertainty. IPCC default emission factors are used for most of the activities and based on international standards or estimates.
- *Tier 2:* This method is similar to the Tier 1 method; however, country specific emissions factors are used for most activities. Tier 2 calculations have a lower uncertainty than Tier 1 calculation but a higher uncertainty than Tier 3 calculations.
- *Tier 3:* Emissions are calculated using process models, direct measurements, mass balances etc. Tier 3 calculations have the lowest level of uncertainty.

To reduce these GHG emissions and ensure companies are within their carbon budgets, emissions mitigation measures are required [20]. Two climate change mitigation approaches investigated by the Intergovernmental Panel on Climate Change (IPCC) are conventional mitigation technologies such as renewable energy, and negative emission technologies such as afforestation and reforestation [21, 25].

1.3.2 Carbon mitigation strategies

As of 2017, renewable energy accounted for 18% of the total final energy consumption and, as of 2018, 26% of the global electricity purchased originated from renewable energy sources [26]. Decarbonisation efforts through renewables include thermal and photovoltaic solar energy and can be deployed in the industry, transportation and building sectors [25].

A potential decarbonisation approach that can be applied in the power and industrial sectors is carbon capture, storage, and utilisation. During carbon capture, storage and utilisation, CO_2 gases are separated and captured from processes that are reliant on fossil fuels such as coal, oil, or gas. The main objective of carbon capture, storage, and utilisations is to reduce emissions levels while utilising fossil fuel sources [25].

For fossil fuel dependant countries such as South Africa, switching from coal to gas can pave the way to a low carbon and potentially zero-carbon economy and is applicable to the power, industry, transportation and building sectors [27-29]. Fuel efficiency gains also play a prominent role in emissions mitigation efforts. Efficiency gains in the power sector can be achieved by enhancing fuel combustion efficiency and by improving turbine generator efficiencies [26].

Where conventional mitigation strategies are not effective enough to remove all emissions, negative emissions technologies can be used to remove the remainder of emissions [30]. One of the most prominent negative emission technologies is bioenergy carbon capture and storage (BECCS) [21]. During BECCS, biomass captures CO_2 via photosynthesis which is used for energy production through combustion. The resulting CO_2 emissions are then captured and stored in geological reservoirs [31, 32]. BECCS can result in a significant reduction in GHG emissions due to the removal of CO_2 emissions from the atmosphere [25].

Another negative emissions technology is afforestation and reforestation. Afforestation involves introducing trees and trees seedlings in an area with no previous tree growth, while reforestation consists of replanting deforested areas. During tree growth, atmospheric CO_2 is captured and stored in biomass, organic matter, and soil where it can remain for a long period [25].

For carbon capture on a permanent basis, biochar can be used. Biochar is produced from biomass during pyrolysis [33-36]. Carbon is captured by biomass through CO_2 uptake during plant growth. The carbon is then processed into char which can be applied to soils for extended periods. During the conversion process, biomass carbon is stored in a stable form which is resistant to decomposition [37, 38].

The reduction of GHG emissions through emissions mitigation can result in increased efficiency and cost effectiveness. The successful management and mitigation of a company's exposure to climate change related risks, while ensuring the company remains profitable, can provide a competitive advantage over rival companies. The reduction of GHG emissions is also seen as a form of intangible value. Thus, if a company can successfully reduce its GHG emissions, it could reflect in the company's stock market performance [39].

1.4 Emissions forecasting and forecasting models

The reduction of GHG emissions is important to mitigate the effects of climate change. To ensure a company's emissions do not exceed its carbon budget, emissions forecasting can be used. Emissions forecasting can also be used for setting up GHG mitigation plans and to explain trends in GHG emissions as required during the validation and verification process [40].

Emissions forecasting is vital for decision making regarding climate change policies and the development of medium- and long-term emissions reduction strategies [41]. According to Auffhammer and Carson [42], emissions forecasting can be used for future global warming

predictions, for estimating the costs of emission reductions and for predicting the expected benefits of climate change mitigation. The latter is especially important when setting up mitigation plans.

There are several types of forecasting models available in literature. Abdullah and Pauzi [43] reviewed literature on forecasting methods as well as the estimation of CO_2 emissions. According to their research, the most popular forecasting methods are neural networks, grey models, computer-based models, and optimal growth models. Seventeen studies related to the topic were considered. A summary of these studies can be found in Table 1-3.

Table 1-3: Summary of emissions forecasting models available in literature.

Literature model	Forecasting model	Purpose of model	Emissions forecasting level	Input variables
Literature model 1: Radojevic et al [44]	Neural network	To estimate the CO_2 emissions of Serbia for predicting environmental indicators of sustainable development.	Country	• Share of renewable energy sources • Gross Domestic Product (GDP) • Gross energy consumption • Energy intensity
Literature model 2: Liu et al [45]	Neural network	To forecast CO_2 emissions in China.	Country	• GDP • Exports • Consumer price index • Investment in fixed assets • Population
Literature model 3: Yap and Karri [46]	Feedforward neural networks	To predict vehicle emissions using a two-stage emissions predictive model.	Sectoral	• Engine forces • Attractive forces
Literature model 4: Behrang et al [47]	Integrated multi-layer perceptron neural network and Bees Algorithm	Forecast global CO_2 emissions.	Global	• Population • GDP • Oil trade movements • Natural gas trade movements • Demand for fossil fuels • Primary energy demand

Literature model	Forecasting model	Purpose of model	Emissions forecasting level	Input variables
Literature model 5: Li et al [48]	Radial Basis Function neural network combined with time series	Predict CO_2 emissions in China.	Country	• Energy consumption
Literature model 6: Sözen et al [49]	Artificial neural networks	Predict GHG emissions in Turkey.	Country	• Energy consumption
Literature model 7: Pao et al [50]	Nonlinear Grey Bernoulli	Predict three indicators: carbon emissions, energy consumption, and real output in China.	Country	• Energy consumption • CO_2 emissions • Energy intensity • GDP data
Literature model 8: Pau and Tsai [51]	Grey prediction model	Forecast CO_2 emissions, energy consumption, and real GDP to investigate the dynamic between variables for Brazil.	Country	• CO_2 emissions • Energy consumption • Real GDP
Literature model 9: Lin et al [52]	Grey forecasting model	Estimate future CO_2 emissions in Taiwan.	Country	CO_2 emissions related to consumption of: • Coal • Petroleum • Natural gas
Literature model 10: Lu et al [53]	Grey model	Capture the development trends of the number of vehicles, vehicular energy consumption, and CO_2 emissions in Taiwan.	Country	• Fuel consumption • Fuel density

Literature model	Forecasting model	Purpose of model	Emissions forecasting level	Input variables
Literature model 11: Feng et al [54]	Dynamics based computer simulation model	Estimate and predict urban energy consumption trends and CO_2 emissions for Beijing.	City	• Energy demand • Carbon emission factors
Literature model 12: Hui-ya et al [55]	Optimal economic growth model	Predict net carbon emissions of Yunnan, China, from 2007 to 2050.	City	• Energy consumption
Literature model 13: Huang and Wang [56]	Optimal growth model	Investigate Shanghai's energy consumption and carbon emissions.	City	• Economic growth • GDP • Energy consumption
Literature model 14: Wu et al [57]	Novel multi-variable grey model	Examine the relationship between energy consumption, urban population, economic growth, and CO_2 emissions in BRICS countries.	Country	• Energy consumption • Urban population • Economic growth • CO_2 emissions
Literature model 15: Ma et al [58]	Multivariate grey model, firefly algorithm, and the association rule algorithm	Develop a hybrid-carbon emissions forecasting model to be used in the formulation of emissions reduction policies of China.	Country	• Energy consumption • Economic growth • Industrial structure • Foreign direct investment • Urbanisation
Literature model 16: Wang et al [59]	Metabolic non-linear grey model, Autoregressive Integrated Moving Average, and Back Propagation Neural Network (BPNN)	Develop more effective carbon mitigation strategies for the forecasting of carbon emissions in China, US, and India.	Country	• Energy consumption • Population • GDP

Literature model	Forecasting model	Purpose of model	Emissions forecasting level	Input variables
Literature model 17: Nong [60]	Global computable general equilibrium model	Enhance the capacity and accuracy for climate change and energy policy assessment by analysing the impacts of the carbon tax policy on the economy of South Africa.	Facility	• Carbon price

As seen in Table 1-3, out of the seventeen studies considered, sixteen of the studies focussed on predicting CO_2 emissions. Most of the studies developed emissions forecasting models using neural networks (six out of seventeen). A breakdown of the forecasting models used can be found in Figure 1-3.

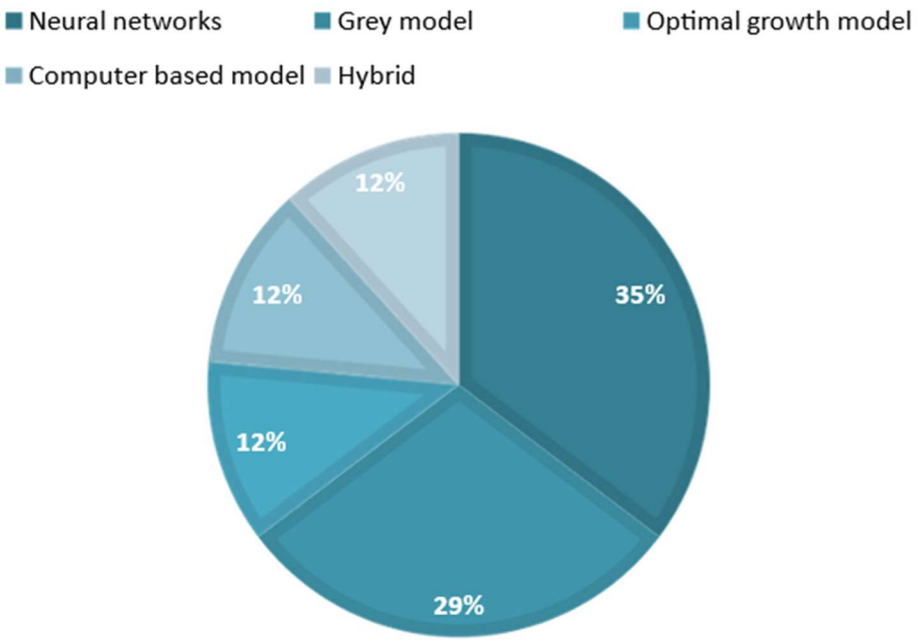


Figure 1-3: Breakdown of forecasting models used in literature.

Only one of the studies listed in Table 1-3 mentions the development of a forecasting model at facility level. However, the model focusses on analysing the impacts of carbon tax on the economy and not on emissions forecasting. Eleven of the studies focus on emissions forecasting at a country level. Three studies focus on the development of an emissions forecasting model at a city level, and one study considers emissions forecasting model at sectoral, global, and facility level, respectively. Thus, currently emissions forecasting is rarely done at facility level. This makes it difficult for facilities to ensure that they will remain within their carbon budget.

1.5 Significance of the study

In this section, the significance of the study is highlighted by evaluating the carbon tax liability of six South African industries that participate in the carbon budget system. The industries include electricity generation, petrochemical production, mining and minerals, iron and steel production, ferrochrome production, and cement production.

During the first phase of carbon tax, the forecasting of GHG emissions enables facilities to set up their carbon budgets. By participating in the carbon budget system, taxpayers can receive a 5% tax free allowance on their total carbon tax liability. The total liability as well as the avoided liability for the six industries are summarised in Table 1-4 [61].

Table 1-4: Impact of carbon budget allowances on the total carbon tax liability.

Industry	Total emissions	Total liability	5% of total emissions	Liability of 5% of total emissions
	$MtCO_2e$	R (million)	$MtCO_2e$	R (million)
Electricity generation	223.00	R 29 882	11.15	R 1 494
Petrochemical production	59.23	R 7 937	2.96	R 397
Mining and minerals	30.91	R 4 142	1.55	R 207
Iron and steel production	11.56	R 1 549	0.58	R 77
Ferrochrome production	6.99	R 937	0.35	R 47
Cement production	3.75	R 503	0.19	R 25
Total	**335.44**	**R 44 949**	**16.77**	**R 2 247**

**Carbon tax liability calculated at R 134 per tonne CO_2e emissions*

A total of 16.77 $MtCO_2e$ emissions of the six reported entities are exempted from carbon tax due to the facilities' participation in the carbon budget system. This amounts to a R 2.3-billion avoided carbon tax liability.

In the second phase of carbon tax, carbon budgets will become mandatory and will be allocated by the governmental structures. For emissions exceeding the carbon budget during the second carbon tax phase, a higher tax rate, which is proposed as R 600 per tonne CO_2e emissions, will be applied. GHG emissions forecasting can then be used to ensure that the facilities' emissions are within their carbon budget.

1.6 Problem statement and research objective

GHG emissions are one of the main contributors to climate change. To reduce emissions and minimise the risks associated with climate change, several national and international policies were implemented.

Carbon budgets are one of the mitigation measures implemented by the South African government to reduce GHG emissions. Based on the reported emissions of six South African

entities, the participation in the carbon budget system could result in an avoided carbon tax liability of R 2.3-billion.

With the promulgation of the Climate Change Bill, carbon budgets will become mandatory. Facilities to which carbon budgets were allocated are required to submit a GHG mitigation plan detailing how they intend to reduce their emissions.

South African facilities need to forecast their emissions at a facility level to ensure that effective carbon mitigation measures are in place. No facility-level forecasting models are currently available in literature. Most models focus on national or global emissions predictions. Thus, there is a need for facility-based emissions forecasting models.

The objective of this study is to outline the development of forecasting models that can be used to predict the carbon emissions for South African facilities. The developed forecasting models can then be used to:

- Identify potential emissions reduction opportunities,
- Determine the effect of different mitigation measures on the company's total GHG emissions, and
- Determine whether emissions targets will be met with the mitigation measures already in place.

1.7 Scope of investigation

This study specifically focusses on the process to develop emissions forecasting models for South African facilities. Emissions will be forecasted at facility level and only Scope 1 emissions will be considered when developing the models.

Some of the forecasting models considered in this study include neural networks, grey models, linear regression as well as multivariable linear regression which are further investigated in Chapter 2.

1.8 Dissertation layout

Chapter 1: Introduction and literature

Chapter 1 presents a brief introduction to the study and introduces relevant literature regarding existing emissions forecasting models. This chapter further highlights potential gaps in literature and emphasises the need for the study which is presented in the form of a problem statement. Objectives to be met throughout this study are also discussed in Chapter 1.

Chapter 2: Methodology

Chapter 2 presents a brief overview on the different forecasting models available in literature. This chapter also presents the model development steps, focussing on data preparation, model development, and model validation.

Chapter 3: Results and discussion

Chapter 3 presents the development and verification of the emissions forecasting model using an actual case study. The three different emissions categories (combustion, process, and fugitive) were considered for the case study.

Chapter 4: Conclusion

Chapter 4 presents the conclusions made regarding the study. The chapter also serves as proof that all objectives stated in Chapter 1 are met and proposes recommendations for further studies regarding the study topic.

CHAPTER 2 - METHODOLOGY

2.1 Preamble

The literature in this chapter is used to identify the steps needed to develop an emissions forecasting model for South African facilities. According to Hyndman and Athanasopoulos [62], the basic steps for forecasting development are: problem definition, gathering information, preliminary (exploratory) analysis, choosing and fitting models, and using and evaluating the forecasting as illustrated in Figure 2-1.

Figure 2-1: Basic steps for forecasting model development.

Problem definition is often regarded as one of the most difficult steps in the development of a forecasting model. The reason for this is because it requires the understanding of the following [62]:

- The way in which the forecasts will be used.
- The people requiring the forecasts.
- The way in which the forecast will be incorporated in the organisation.

The first step (problem definition) has already been addressed in Chapter 1. The problem definition is as follows: South African facilities need to forecast their emissions at facility level; however, no facility-level forecasts are available in literature. This chapter will focus on the last four steps.

During the gathering information step, the desired data is identified, collected, and inspected for any missing or incorrect values. Next, the data is analysed to identify any patterns or outliers in the data which will have an influence on the type of forecasting model used. Based on information gathered in the preliminary analysis, a suitable forecasting model is chosen and fitted to the available information. After developing the model, the desired data is forecasted, and the model is evaluated to determine the accuracy of the forecasted data.

2.2 Gathering information

The first step in forecasting model development, after problem definition, is data gathering. For this step, the following information is required: statistical data, and the accumulated expertise of people who collect and forecast the data [62]. During the data gathering step, data is identified, collected, and inspected to ensure there are no missing or incorrect values. These sub-steps are illustrated in Figure 2-2.

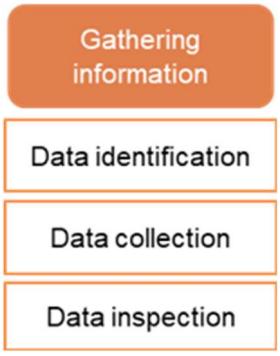


Figure 2-2: Sub-steps for the gathering information step.

The following paragraphs discuss each sub-section in detail and explain how it will be utilised in the development of the forecasting models.

2.2.1 Data identification

Data identification is defined as the process that highlights the data sets required as well as their desired properties. These properties include sample period, data format, and accuracy [63].

Before looking at the properties mentioned above, the type of data required to forecast the GHG emissions needs to be identified. GHG emissions are calculated using activity data. However, it is not always possible to estimate the activity data for all emissions. Hence, the relationship between emissions and influencing factors such as production data etc. needs to be determined. This further allows for estimating how changes in the influencing factors will affect the carbon emissions.

Next, the desired sample period for the data needs to be determined. The sample period can be daily, weekly, monthly, or annually. Time series data is usually available in daily intervals, which can be aggregated to a monthly basis, while emissions data is calculated annually according to the National Greenhouse Gas Emissions Reporting Regulations (NGER) [64].

The data format should also be considered during data identification. Data format plays a fundamental role in understanding the data, representing the data, and determining the

storage space needed for the data [65]. Examples of different data formats include comma separated values (CSV), tab separated values (TSV), text format, and extensible markup language (XML) [65].

The format in which the data is stored should [65]:

- Be compact enough to store large data sets,
- Be readable by various software and tools,
- Represent categorical and numerical features,
- Be local, to transfer data independently, and
- Not be changed by tools and software.

All the data used in this study is available in an XLSX format (a zip compressed XLM spreadsheet file) which is the current default extension used for a Microsoft Excel file.

The quality of data used in forecasting model development has a significant impact on the forecasting accuracy. The forecasting model can only be as good, in terms of accuracy, as the quality of data provided to the model. It is important to prepare a consistent data set to ensure high quality data is used in model development [66]. Data quality depends on the type of data used, where the data is obtained from, as well as what is included and excluded from the data set [67].

Some of the data quality characteristics to consider during data gathering are summarised in Table 2-1.

Table 2-1: Data quality dimensions.

Data quality dimensions	Description
Accessibility	Data should be available to be used by anyone. Data that cannot be accessed serves no purpose [68].
Accuracy	Data should give a true representation of reality [69]. The two most important characteristics of data accuracy are form and content [70].
Trustworthiness	Data should originate from a trustworthy source [68]. The trustworthiness of data is dependent on four key components: credibility, transferability, dependability, and conformability [71]. Data needs to be obtained from a reliable source to ensure good quality data [72].
Timeliness	Indicates the time expectation for the accessibility of data (is the data available when needed and expected?). Timeliness is an important aspect in database management and refers to the availability and accessibility of data [68].

Data quality dimensions	Description
Reliability	Data conveys the right information and can be traced and connected to other data [38].
Consistency	Data is equivalent across different data table sources or systems [68].
Completeness	Data is comprehensive and all the required data is available. Data considered complete given that all expectations are met [68].
Relevance	Collected data should be relevant to its intended purpose and meet current as well as potential future needs. Is the collected data relevant to its intended purpose and can the objective be met [68]?
Representability	Data should be representable as it gives assurance of all properties included.

From Table 2-1 it is evident that there are several characteristics that determine the quality of data. To ensure the collected data is of a high quality, data should be easy to access, the available data should be accurate, and data should be obtained from a trustworthy source. It is also important to ensure that the data conveys the correct information, and that the data is consistent over different sources and systems. Lastly, it is important to ensure that the data set is representable and complete, and that only relevant data is collected.

2.2.2 Data collection

After the data required is identified, the identified data needs to be collected. According to Robertson [69], data collection is a crucial stage in the model development process. Data collection is based on the GIGO (Garbage In, Garbage Out) principle, i.e. the output of the system is only as good as the data that is put into the system [73-76].

2.2.3 Data inspection

Once a complete data set is collected, data should be inspected to ensure that no missing or incorrect values are used and that it reflects the truth [77].

Missing data

Missing data is defined as the data value that is not stored for a variable in the observation of interest [78]. Missing data is a common occurrence in research and has a significant effect on the results of the study since data processing and data analysis require a complete data set [79, 80]. Missing data in a time series is illustrated in Figure 2-3.

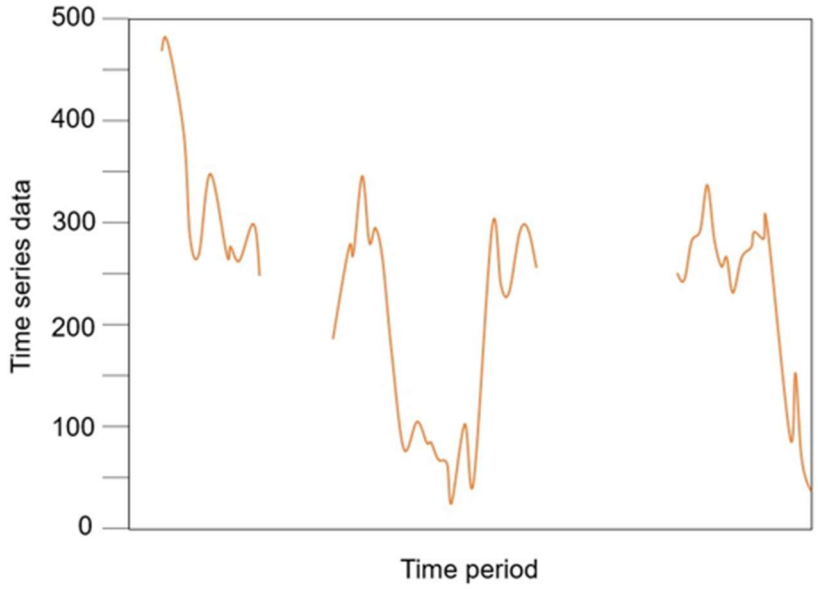
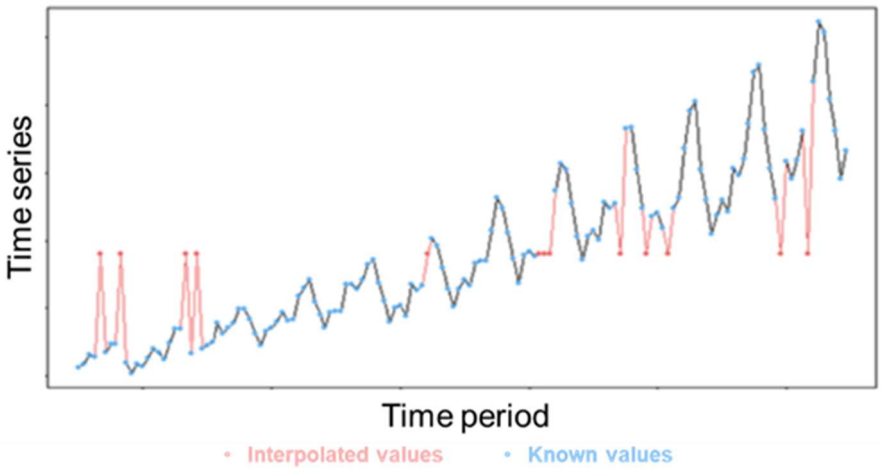


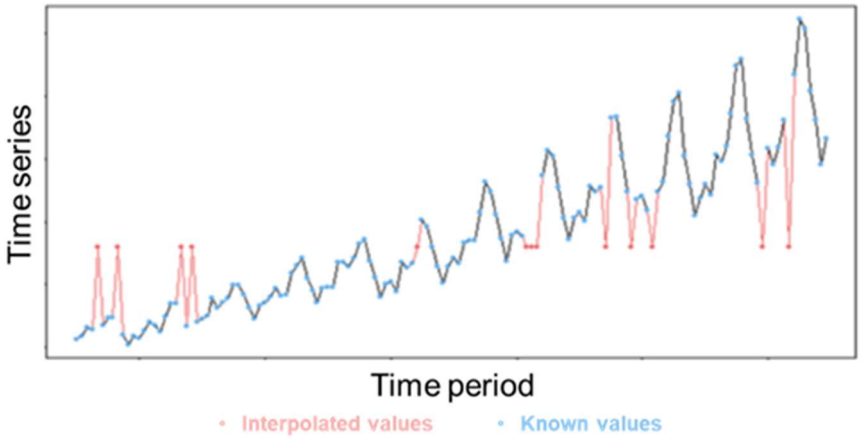
Figure 2-3: Missing data in a time series.

From Figure 2-3 it is evident that missing data has a significant impact on model development since no information on data patterns, trends, increasing and decreasing values etc. are available.

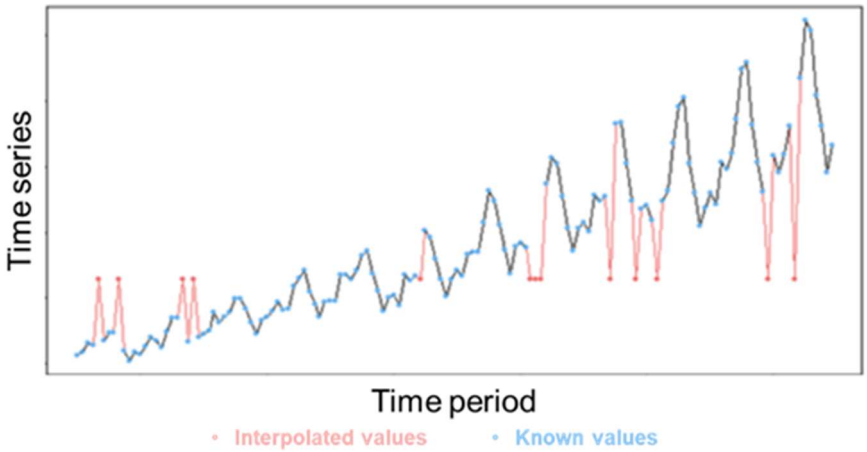
To ensure a data set is complete, the missing values need to be replaced with reasonable values which can be determined using interpolation [81]. Interpolation is known as the process of estimating and inserting an unknown value between known values [82]. Statistical interpolation techniques include mean, median, and mode interpolation [81]. These techniques are illustrated in Figure 2-4 (a to c).



(a) Mean interpolation



(b) Median interpolation



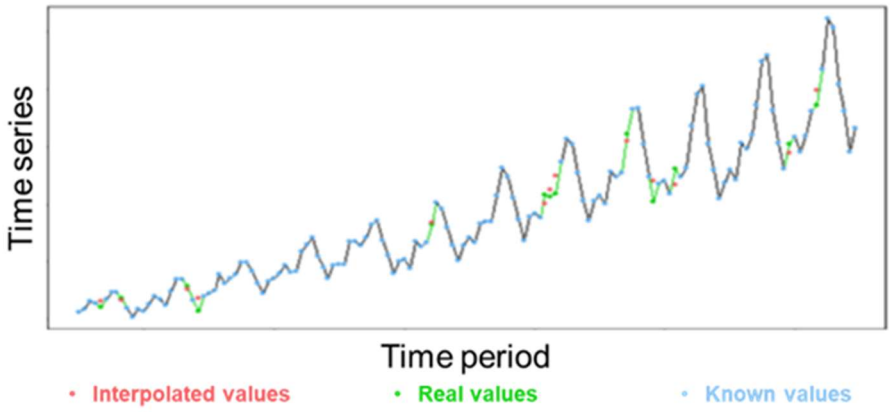
(c) Mode interpolation

Figure 2-4: Statistical interpolation techniques.

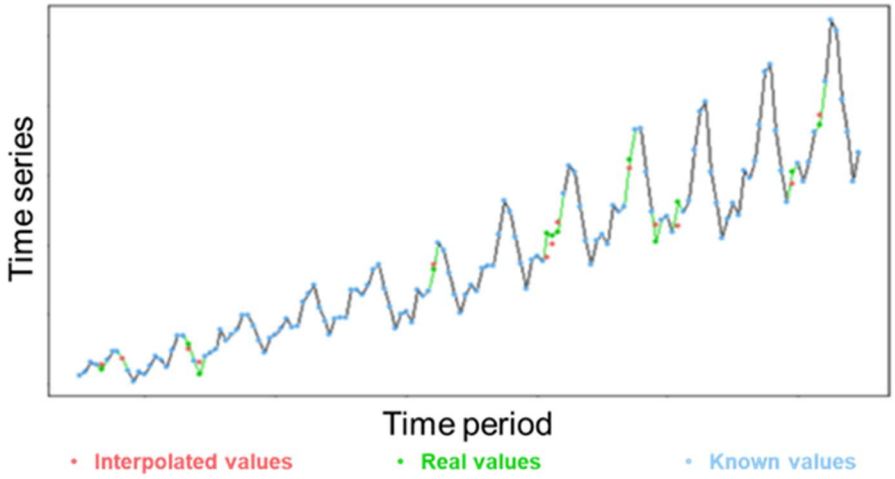
When using mean interpolation, Figure 2-4 (a), the missing values are filled with the mean of the observed values. The same is true for median and mode interpolation; however, the mean is replaced by the median and mode respectively (Figure 2-4 (b) and (c)).

From Figure 2-4 it is evident that the mean, median, and mode interpolation have a significant impact on the trend, cyclical, and seasonal components of the time series. To prevent influencing these occurrences, linear and spline interpolation is more suitable [83].

For linear interpolation, a straight line is created between the two points around the missing values and the point midway between these two values is used to fill the missing values. The last interpolation technique is spline interpolation. Spline interpolation is defined by piecewise polynomials [83]. The linear (a) and spline (b) interpolation techniques are illustrated in Figure 2-5.



(a) Linear interpolation



(b) Spline interpolation

Figure 2-5: Time series interpolation techniques.

From Figure 2-5 it is evident that linear or spline interpolation give a more reasonable value for time series data with strong trend, cyclical, or seasonal components. Thus, for this study, missing values will be filled using linear interpolation.

Outlier investigation

Another important pre-processing step is the detection and/or removal of outliers to ensure the correct data is used [84]. An outlier (also known as an anomaly) is defined as a datapoint that does not conform with the normal patterns in the data set [85]. Outlier identification is used for cleaning data by removing noisy or erroneous data [86]. By removing noisy and erroneous values, the performance of the forecasting model can be greatly improved [84].

There are various outlier detection methods available that can efficiently detect outliers [87]. However, to ensure successful outlier removal, it is necessary to recognise why outliers are identified and how they deviate from the normal data [86].

According to Santoyo [88], some of the most popular outlier detections methods are:

- Z-Score or Extreme Value Analysis,
- Probabilistic and Statistical Modelling,
- Linear regression models,
- Proximity based models,
- Information theory models,
- High dimensional outlier detection methods, and
- Interquartile range method [89].

The interquartile range (IQR) method is simple to use, has a low sensitivity to distortion due to outliers, and can be used for normal and non-normal distributed data. Thus, this outlier detection method will be used in this study.

The IQR method considers the range between the 25th (Q_1) and 75th (Q_3) percentiles of the sample data and covers the central 50% of the data. Two types of outliers can be defined by the IQR method: mild outliers and extreme outliers. Mild outliers are described as values beyond the inner limits described in Equation (2-1) and Equation (2-2), while extreme outliers are described as values beyond the outer limits as described by Equation (2-3) and Equation (2-4).

$$lower\ inner\ limit = Q_1 - 1.5 \times IQR \quad (2\text{-}1)$$

$$upper\ inner\ limit = Q_3 + 1.5 \times IQR \quad (2\text{-}2)$$

$$lower\ outer\ limit = Q_1 - 3 \times IQR \quad (2\text{-}3)$$

$$upper\ outer\ limit = Q_3 + 3 \times IQR \quad (2\text{-}4)$$

One of the disadvantages of the IQR method is that it does not consider sample size; thus, legitimate values are discarded in large data sets. However, a modified IQR method can be used where sample size is accounted for [89].

Sample size

Other than inspecting the data set for missing or erroneous data, the data sample needs to be of a representative size to ensure the data set is complete [90]. For time series data, Hanke et al [91] suggests a minimum sample size of double the seasonality of the data. For regression and multiple linear regression, Green's method can be used [92]. This method is based on the number of predictable variables. The sample size can be calculated using Equation (2-5).

$$n = 50 + 8p \quad (2\text{-}5)$$

Where:

- n is the sample size,
- p is the number of predictors.

Another method for calculating the required sample size is the Cornish method. The Cornish method is a precision-based approach where the precision is known, and the formula is used to solve the standard error [93]. The formula for calculating the standard error contains the sample size as one of the variables; thus it can be used to calculate the sample size, should the standard error be available. The Cornish method is not an easy method to use since various assumptions are required during the calculation process [93].

After gathering a complete set of good quality data, the initial analysis can now be done.

2.3 Preliminary analysis

During the preliminary analysis stage of the forecasting model development process, patterns in the data can be identified by graphing the data [62]. For data where multiple patterns are observed, time series decomposition and trend removal can be used to highlight specific patterns in the data. Thus, the preliminary analysis step consists of the identification of data patterns, time series decomposition, and trend removal as illustrated in Figure 2-6.

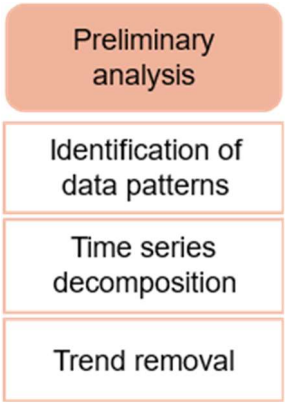


Figure 2-6: Sub-steps in the preliminary analysis step.

The following paragraphs discuss each sub-section in detail and explain how it will be utilised in the development of the forecasting models.

2.3.1 Data pattern identification

Patterns in time series data include horizontal, trend, seasonal, and cyclic patterns [62]. These patterns are illustrated in Figure 2-7, Figure 2-8, Figure 2-9, and Figure 2-10 respectively.

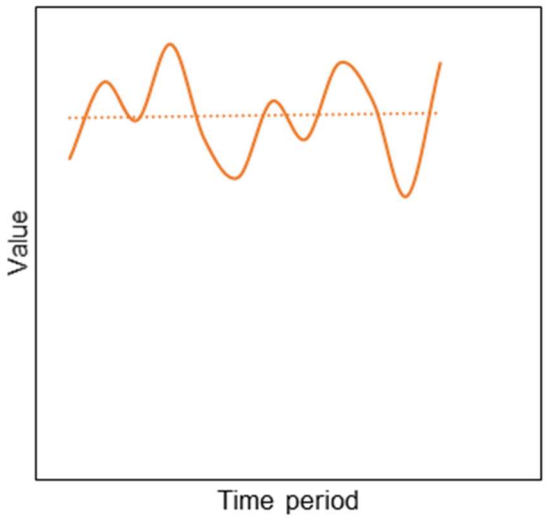


Figure 2-7: Horizontal pattern.

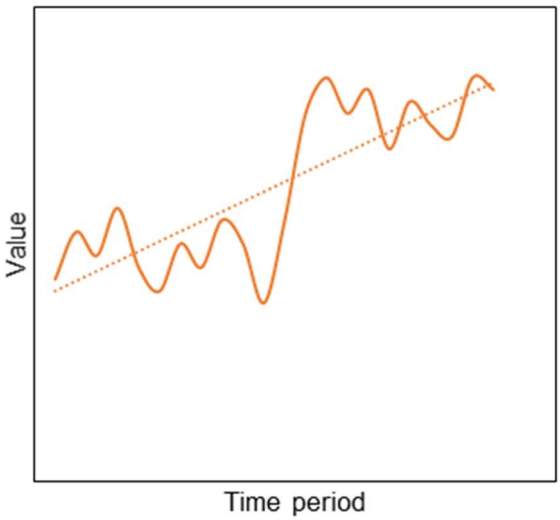


Figure 2-8: Trend pattern.

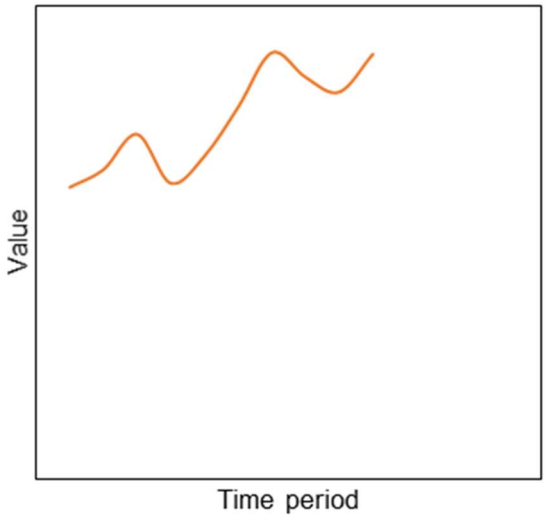


Figure 2-9: Seasonal pattern.

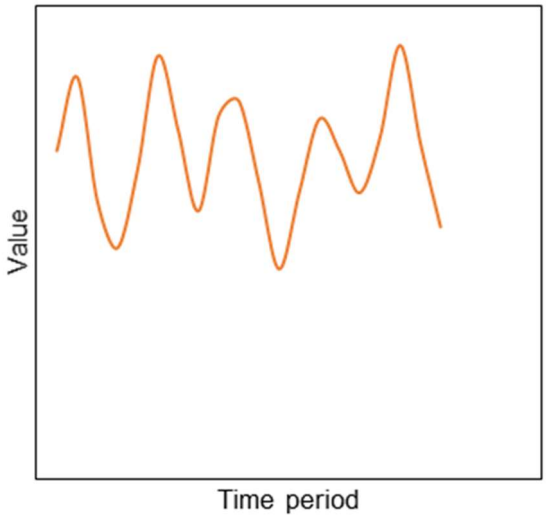


Figure 2-10: Cyclic pattern.

A horizontal pattern is observed in Figure 2-7, where the data fluctuates around a constant mean which is illustrated by the straight, horizontal trendline. In Figure 2-8, a constant increase in the time series data can be observed. This is known as a trend pattern. Trend patterns are defined as a long-term increase or decrease in the observed data.

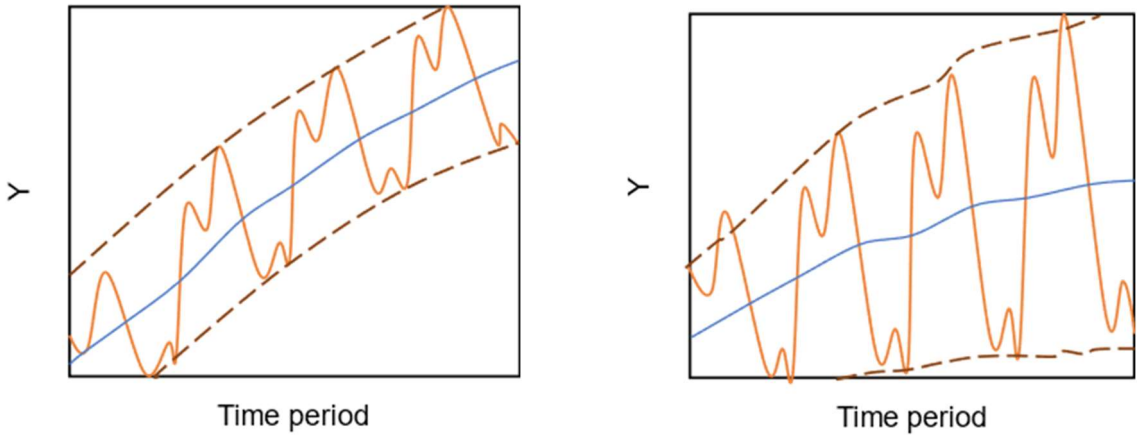
In Figure 2-9, a spike and dip in the data can be observed which are followed by another dip and spike at roughly the same interval. This is known as seasonality. A seasonal pattern can be observed when a time series is influenced by factors such as the time of year or day or week. It is important to note that seasonality occurs at a known and fixed frequency.

In Figure 2-10, fluctuations in the data can be observed without a discernible pattern. This is known as a cyclic pattern. Cyclic patterns display infrequent falls and rises in the observed data. This is usually due to economic conditions.

It is important to identify time series patterns in data before deciding which forecasting model to use [62].

2.3.2 Time series decomposition

Because time series data can exhibit a variety of patterns, it is helpful to split time series into several components. This process is known as time series decomposition. The two types of time series decompositions are known as additive and multiplicative decomposition. Additive decomposition is generally used when there are changes in magnitude of seasonal fluctuations. Multiplicative decomposition is used when there is a variation in seasonal patterns proportional to the level of time series [62, 94]. This is illustrated in Figure 2-11 (a and b).



a) Additive seasonality b) Multiplicative seasonality

Figure 2-11: Additive and multiplicative seasonality.

In Figure 2-11 (a and b), the blue line represents the level of the time series. For the additive time series, the difference between the maximum point and the level of the time series is relatively constant. However, for the multiplicative time series the magnitude of the seasonal components grows as time goes on.

Additive and multiplicative decompositions can also be described by Equation (2-6) and Equation (2-7) respectively.

$$y_t = S_t + T_t + R_t \tag{2-6}$$

$$y_t = S_t \times T_t \times R_t \tag{2-7}$$

Where:

- y_t is the times series value at time t,
- S_t is the seasonal component at time t,
- T_t is the trend component at time t,
- R_t is the remainder component (error) at time.

The decomposition of a time series allows for a better understanding of the data and helps to explore different ways to model the data. After separating the data into the different time series components, the trend can now be removed from the data.

2.3.3 Trend removal

According to Chan [95], trend removal, also known as detrending, is often desirable in time series analysis. Trend removal renders a time series stationary, allows for the identification of cyclic and other patterns, highlights potential sub-trends, and places the focus on fluctuations in the time series and other important factors [96].

There are several methods available for detrending a time series. One of these methods is known as the least squares linear trendline method [97]. With this method, a trendline is constructed in the form of Equation (2-8).

$$T_t = mx_t + C_t \tag{2-8}$$

Where:

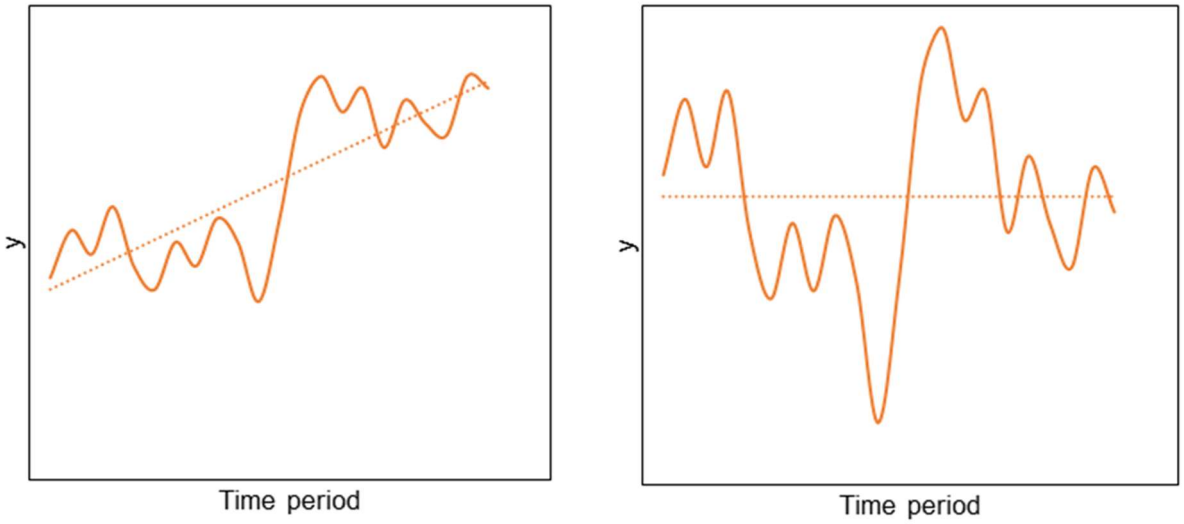
- m is the slope,
- x_T is the independent variable,
- C_t is the intercept.

The data is then detrended by dividing or subtracting with the trend component for multiplicative and additive decomposition, respectively. This is calculated with Equation (2-9) and Equation (2-10).

$$y_{DT} = y_t - T_t = S_t + R_t \tag{2-9}$$

$$y_{DT} = \frac{y_t}{T_t} = S_t \times R_t \tag{2-10}$$

The trendline of the resulting time series should consist of a horizontal line as illustrated in Figure 2-12.



(a) Before trend removal (b) After trend removal

Figure 2-12: Trend removal.

In Figure 2-12 (a), time series data with a constant increase in the data can be observed. After removing the trend, the data still follows the original pattern; however, a horizontal trendline can now be observed (Figure 2-12 (b)).

Once the preliminary analysis for the data has been done, it is time to choose the best-suited forecasting model and fit the data to the model.

2.4 Choosing and fitting models

Three things should be considered when choosing and fitting forecasting models. The first is the availability of historic data, the second is the strength of relationships between forecasting and explanatory variables, and the last is how the forecasted data will be used [62].

There are several different forecasting models available in literature. It is important to consider the advantages and disadvantages of the different models to choose the model(s) best suited for the data. After the model(s) is selected, the forecasting model needs to be developed. The sub-steps in choosing and fitting models as described are illustrated in Figure 2-13.

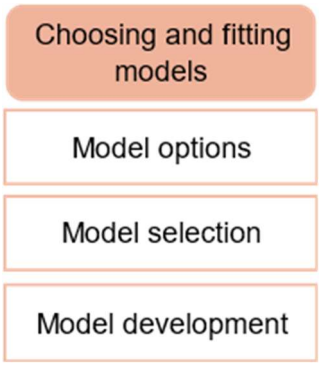


Figure 2-13: Sub-steps when choosing and fitting models.

2.4.1 Model options

As mentioned before, there are several different forecasting models available in literature. These models include, but are not limited to, artificial neural networks, Grey predictive models, Autoregressive Integrated Moving Average (ARIMA) models, and exponential smoothing.

The advantages and disadvantages of some of the different forecasting models available in literature are summarised in Table 2-2.

Most of the models discussed in Table 2-2 are relatively easy to implement and do not require a high skill level to apply. However, there are some exceptions. ANNs require great computational resources and the reliability of the ARIMA model is dependent on the skill level of the forecaster. Although a small sample size is sufficient for the Grey predictive model, the model does not consider new information and cannot reflect the characteristics of the current situation.

For models such as linear regression, and the Holt-Winters method, outliers can have a significant effect on the forecasted results; thus outlier detection/removal is crucial for these models. Autoregressive models, exponential smoothing, and the Holt-Winters method can all be used where trend, seasonality or any other recurring patterns are visible in the data. All these advantages and disadvantages should be considered when determining which forecasting model(s) to use for model development.

Table 2-2: Advantages and disadvantages of different forecasting models.

Model	Description	Advantages	Disadvantages
Artificial Neural Networks (ANNs) [98, 99]	ANNs are mathematical tools or physical devices that function similarly to biological neural systems.	• Less formal statistical training is required to develop • Can implicitly detect complex non-linear relationships between dependent and independent variables • Can detect all possible interactions between predictor variables • Can be developed using multiple different training algorithms	• "Black box" model, thus their ability to explicitly identify casual relationships are limited • Modelling requires greater computational resources • Prone to overfitting • Model development is empirical, and there are still methodological issues to resolve
Grey predictive model [100]	A model where some of the information is known and some of the information is unknown.	• A small sample size is sufficient for using the Grey predictive model • Model is simpler and more suitable to use than the general econometric models • Can be used for short-term prediction activities	• Model ignores new information and cannot accurately reflect the characteristics of the current situation
ARIMA model [101]	The ARIMA model consist of two components: an autoregressive component (time series is expressed in terms of past values of itself) as well as a moving average component (the current and lagged values of 'white noise' error term (the moving average component).	• When creating short-term forecasts, the model is relatively robust • ARIMA models can usually outperform more sophisticated structural models in short-term forecasting ability [101]	• The reliability of the ARIMA model is dependent on the skill and experience of the forecaster • The ARIMA is a "backward looking" model. "Backward looking" models are poor at predicting turning points unless turning point represents a return to a long-run equilibrium [101]

Model	Description	Advantages	Disadvantages
Linear regression	Models the relationship between two variables by fitting a linear equation to the observed data[5].	• Simple to implement and easy to interpret • Susceptible to over-fitting. This can be avoided using dimensionality reduction techniques, regularization techniques and cross validation	• Outliers can have an impact on the forecasted results • Assumes there is a linear relationship between the dependent and independent variable. Thus, independence between attributes is assumed
Multiple linear regression analysis (MVRA) [102]	MVRA is used to examine the relationship between various independent variables and a dependent variable.	• MVRA enables the determination of the relative influence of one or more predictor variables to the criterion variable • Can be used to identify outliers, or anomalies	• Incomplete or erroneous data can result in a false conclusion that a correlation is a causation
Autoregressive model [62], [103]	A statistical model that predicts future values based on past values.	• Can identify if there is a lack of randomness in the data • Can forecast recurring data patterns • Less information required	• An autocorrelation coefficient that cannot be less than 0.5 is required to ensure accurate results
Moving average [104]	Uses averages of different subsets of the data set.	• Can be used for measuring the trend of any series • Appliable to linear as well as non-linear trends	• The trend can neither be described by a straight line nor a standard curve • The model cannot extend the trend to forecast future values • Not applicable to short time series[6]
Exponential smoothing [105]	Exponentially decreasing weights are assigned for newest to oldest observations.	• Fully automatic forecasting method • Low level of skill required from the forecaster, and it is quick to implement	• Not based on a formal theory or statistical model • Modelling process less informative than that of linear regression

[5] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[6] https://www.emathzone.com/tutorials/basic-statistics/merits-and-demerits-of-moving-average-method.html

Model	Description	Advantages	Disadvantages
		• Can model seasonal variations, trend as well as autoregressive, and moving average processes • Effective to use when model parameters change over time	• Provides less accurate forecasts when strong seasonal or cyclic variations are present
Holt-Winters Method (Triple exponential smoothing) [106]	An exponential smoothing method for handling seasonal data.	• Effective for short-term forecasts • Forecasted results are easy to interpret • Easy to implement • Can adapt to changes in trend and seasonal patterns	• Outliers can significantly distort forecasted results • Cannot be used in a multivariate approach • Can only account for a single seasonal pattern • Only forecast the past information of the forecasting variable

2.4.2 Model selection

Forecasting emissions are dependent on multiple steps. This requires input or activity data that determines the trend. This trend should then be converted into emissions by determining the relationship between the emissions and the input/activity data. This is illustrated in Figure 2-14.

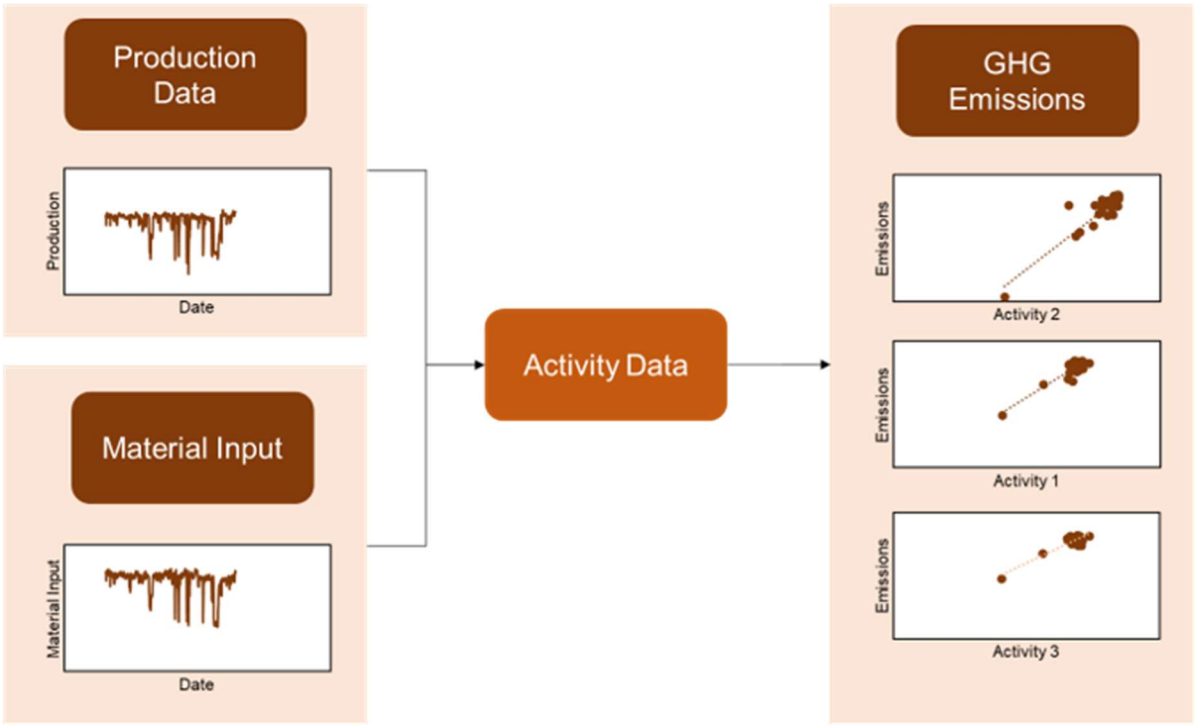


Figure 2-14: Forecasting model selection.

Based on the readily available data, two different forecasting models are required: one to forecast time series data such as production and material input, and another to forecast the emissions based on the activity data.

To forecast the production and material input data, a model is required that is easy to implement and can predict seasonality. Thus, based on the information from Table 2-2, exponential smoothing can be used to predict the time series data. This study will focus on a specific type of exponential smoothing known as the Error-Trend-Seasonality (ETS) model which is also known as triple exponential smoothing.

As with all forecasting models, the ETS model has a few limitations. In order to use this model, the data must have both trend and seasonality elements. It should also be noted that the model relies heavily on historic data and does not consider any elements that could influence future production data. Although, the ETS model can be used accurately for relatively short-

term forecasts, long-term forecast using this model is not advised [62]. For carbon emissions, a short-term forecasting period can be considered as five-years or less [107]. Further information on this model can be found in Appendix A.

To forecast the emissions, a model that predicts the relationship between the time series variables and the emissions is required. Linear regression and multivariable linear regression are both suitable forecasting models to use. It should be noted that the five assumption of linear regression analysis need to be considered before the regression models can be considered valid. The five assumptions are [108]:

1. *Linear relationship:* Described a straight-line relationship between two variables.
2. *Multivariate normality:* Any linear combination of multiple normally distributed variables is also normally distributed. A quick test for multivariate normality is distribution graphs.
3. *No or little multicollinearity:* Multicollinearity can be defined as the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.
4. *No autocorrelation or independence:* The residuals are independent of each other; thus there is no correlation between consecutive terms in the time series data.
5. *Homoscedasticity:* The variance of the residual in a regression model is constant. Thus, the error term does not vary much as the value of the predictor variable change

Further information on the regression models can be found in Appendix A.

2.4.3 Model development

After the correct and accurate data is obtained and the type of forecasting model has been decided on, the actual model can be developed. The model development steps are as follows: split data into modelling and testing data, fit the data, forecast the training data, determine the forecasting accuracy of the fitted data, and use and evaluate the model. These steps are illustrated in Figure 2-15.

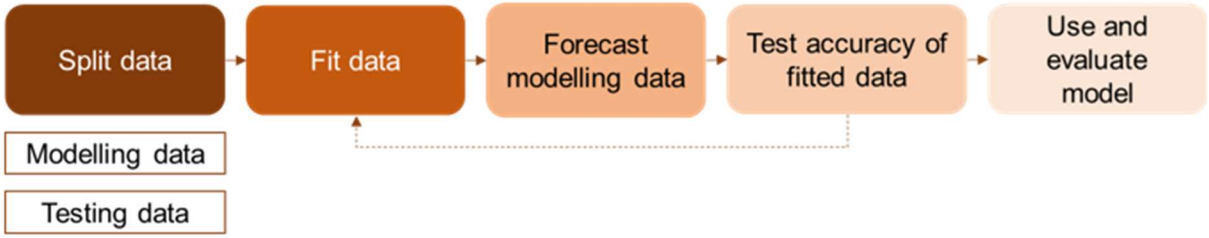


Figure 2-15: Model development steps.

When developing the forecasting model, data is separated into two sets: training and testing data. Training data is used to estimate the parameters of the chosen forecasting method while test data is used to evaluate the accuracy of the forecasting model. This is important to ensure that imported data can be validated since the same data cannot be used for both modelling and validation [62]. The total sample typically consists of 80% training data and 20% testing data.

The approach of splitting data into training (modelling) and testing data is known as cross-validation [109, 110]. One of the disadvantages of cross-validation is that a larger training set will result in a smaller testing set and vice versa. To overcome this disadvantage, an improved version of cross-validation known as k-fold cross-validation (illustrated in Figure 2-16) can be used [111-114].

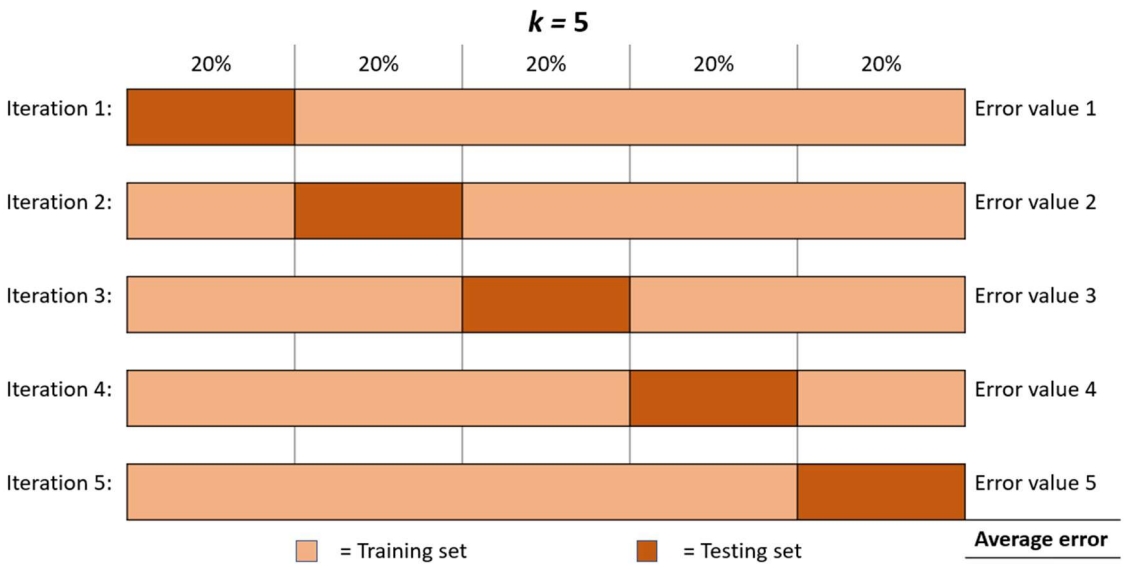


Figure 2-16: Illustration if k-fold cross-validation.

In k-fold cross-validation, the data is split into *k* equally sized data sets after undergoing a random shuffling. In Figure 2-16, a *k*-value of 5 is used thus the iterative process will have 5

repetitions. Each iteration uses a different set of testing data with the remaining data sets combined to form the training set. A chosen error metric is then calculated for each iteration which is ultimately averaged for all iterations in the end.

Common practice suggests a *k*-value of 5 or 10 should be used [115] resulting in 20% and 10% testing data, respectively. Seeing as the data is split into 80% modelling data and 20% testing data, a *k*-value of 5 is chosen.

Next, the training (modelling) data needs to be fitted to the selected forecasting models. This is done by minimising the mean absolute error. After the training data has been fitted to the forecasting model, the resulting model is used to forecast the training data.

Thereafter, the accuracy of the fitted data is determined by comparing the actual and forecasted training data. If the accuracy of the predicted training data is deemed inaccurate, the data needs to be refitted until a sufficient accuracy is obtained.

Once the forecasting accuracy of the fitted data is deemed sufficient, the model can be used and evaluated.

2.5 Using and evaluating the forecasting model

After choosing and fitting the forecasting model, it is necessary to forecast emissions and evaluate the model. To determine whether the study objective from Chapter 1 is met, it is necessary to validate the model, test the forecasting accuracy and to validate if the emissions targets are met. The sub-steps of this final step in the methodology are illustrated in Figure **2-17**.

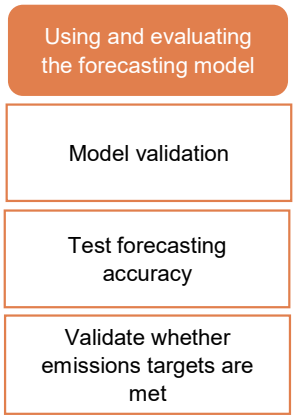


Figure 2-17: Sub-steps in using and evaluating the forecasting model.

As illustrated in Figure 2-17, the model first needs to be validated, whereafter the forecasting accuracy needs to be tested using the validation data. Next, the requirement for the emissions

targets needs to be established, whereafter the forecasted emissions are used to determine if these targets are met and if adequate mitigation measures are in place to ensure that the target emissions are met.

2.5.1 Model validation

After developing the model and determining the forecasting accuracy, the developed models need to be validated. Machine learning can be used as an alternative to Excel for forecasting values using exponential smoothing as well as linear and multivariable linear regressions.

Machine learning is a branch of artificial intelligence and computer science that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so[7]. Machine learning algorithms require historical data as input to predict new output values. There are four types of machine learning; supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. A short description of each machine learning type are as follows [116]:

- *Supervised learning:* The use of labelled data sets to train algorithms to accurately predict outcomes.
- *Unsupervised learning:* Algorithms are used to analyse and cluster unlabelled data sets without the need for human intervention.
- *Semi-supervised learning:* A mixture of supervised and unsupervised learning where a small amount of labelled data and a large amount of unlabelled data is used.
- *Reinforcement learning:* Machine learning models are trained to make a sequence of decisions.

Some of the programming languages used for machine learning are listed below.

- Python
- R Programming
- JavaScript/Java
- C/C++

[7] https://www.ibm.com/cloud/learn/machine-learning

In this study, the programming language, Python, will be used due to its simplicity. Some of the advantages and disadvantages of Python is summarised in Table 2-3 [8] [9].

Table 2-3:Advantages and disadvantages of Python.

Advantages	Disadvantages
The language is easy to learn and to use.	Python has a low execution speed.
Python is a productive language due to the simple nature of the programming language.	Python has a high memory consumption, thus it is not recommended if the user wants to optimise memory.
The Python language is flexible and does not restrict the user like other programming languages.	Although Python provides easy programming, the database access layer of Python is primitive and underdeveloped.
Python has an extensive library.	Python is dynamically typed, thus changes in the data type of a variable can occur at any time which can result in run time errors.
The Python community is very supportive; thus guides, tutorials, and documentation are freely available.	Due to the simplicity of Python, users have difficulty in learning ither programming languages.

The Python models will be developed using the *ETSModel* class and *LinearRegression* class for the ETS and the regression models, respectively. More details on the classes used can be found in Appendix A.

2.5.2 Forecasting accuracy

To ensure that the target year emissions are not under- or over-predicted to a great extent, it is important to evaluate the accuracy of the forecasting model.

Forecasting accuracy can be measured using four tests: the mean absolute error, the mean square error, the mean absolute percentage error, and the root mean squared error.

The mean squared error (MSE) takes the average of the sum of squared errors [117] as described in Equation (2-11). By squaring the errors, all negative signs are removed, and larger differences are given more weight. To improve the forecasting accuracy, the MSE needs to be minimised. The MSE is often used with regression models [118].

$$MSE = \frac{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}{n} \quad (2\text{-}11)$$

Where:

- n is the number of data points,

[8] https://www.aplustopper.com/advantages-and-disadvantages-of-python/
[9] https://understandingdata.com/advantages-and-disadvantages-of-python/

- x_i is the observed value, and
- $\hat{x}_i$ is the forecasted value.

The mean absolute error (MAE) is another model evaluation generally used for regression models to measure the variability of the model [119]. The MAE is computed by averaging all the absolute errors [75], as illustrated in Equation (2-12).

$$MAE = \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|}{n} \tag{2-12}$$

To better interpret the MAE of the forecasting model, it can be divided by the average to get a %MAE as illustrated in Equation (2-13).

$$\%MAE = \frac{MAE}{\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{2-13}$$

The MAE between the actual and forecasted values is illustrated in Figure 2-18. The closer the dots are to the line, the more accurate the forecasting model is.

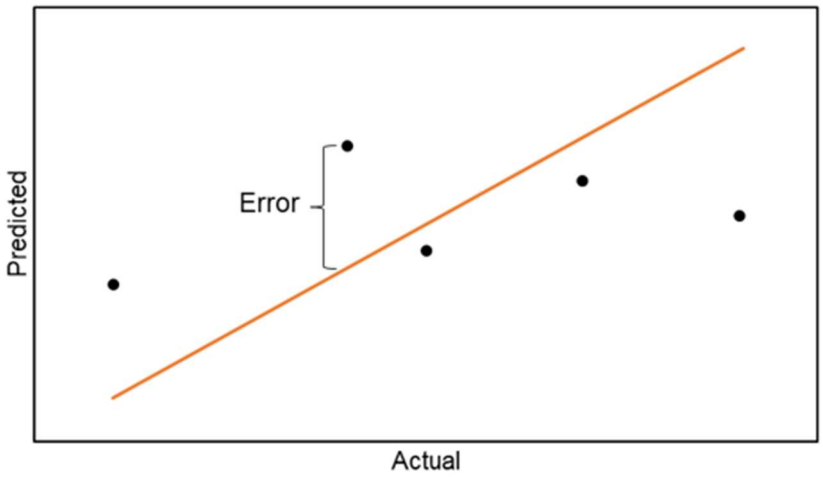


Figure 2-18: Mean absolute error for testing forecasting accuracy.

The mean absolute percentage error (MAPE) is frequently used to compare the forecasting performance between data sets [62]. The MAPE takes the average of the absolute values of the percentage forecasts as illustrated in Equation (2-14).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{2-14}$$

The MAPE is the most common measure used to evaluate the forecasting accuracy, and usually works best when no extremes or zeros are present in the data [105].

The root mean squared error (RMSE) measures the standard deviation of the residuals. Thus, the RMSE gives an indication of how spread out these residuals are. The RMSE and %RMSE can be calculated using Equation (2-15) and Equation (2-16) respectively.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \hat{x}_i)^2}{n}} \quad (2\text{-}15)$$

$$\%RMSE = \frac{RMSE}{\frac{1}{n}\sum_{i=1}^{n} x_i} \quad (2\text{-}16)$$

According to Willmott and Matsuura [40], the RMSE is not a good indicator of average model performance as it might indicate a misleading average error. A better metric to use is the MAE.

For GHG emissions validation and verification, a materiality threshold of 5% is considered based on the total GHG inventory. Thus, if the reported emissions differ by 5% of the actual emissions, the error is seen as materially misleading [120]. The same rule of thumb will be used for the %MAE and %MAPE. If the calculated %MAE and %MAPE are less than 5%, the forecasting model is deemed suitable to be used in industry.

After determining the accuracy of the forecasting model, the baseline emissions can now be forecasted, and the facility's emissions roadmap can be incorporated into the forecasting model. The resulting forecasts can then be used to validate whether the facility will meet its emission reduction targets.

2.5.3 Validate whether emissions targets will be met

To forecast the change in emissions over time, the emissions roadmap first needs to be determined. An emissions roadmap can be defined as a flexible plan that sets out the emissions reduction targets of a company. The emissions roadmap also includes the actions and financing instruments needed to reach the target emissions[10].

To determine if the emissions targets are met, the accountable emissions should be less or equal to the allowable emissions. If the accountable emissions are more than the allowable

[10] https://architecture2030.org/publications-2/roadmap-to-zero/

emissions, the emissions targets are not met, and more stringent mitigation measures are required [121]. This occurrence is visually illustrated in Figure 2-19.

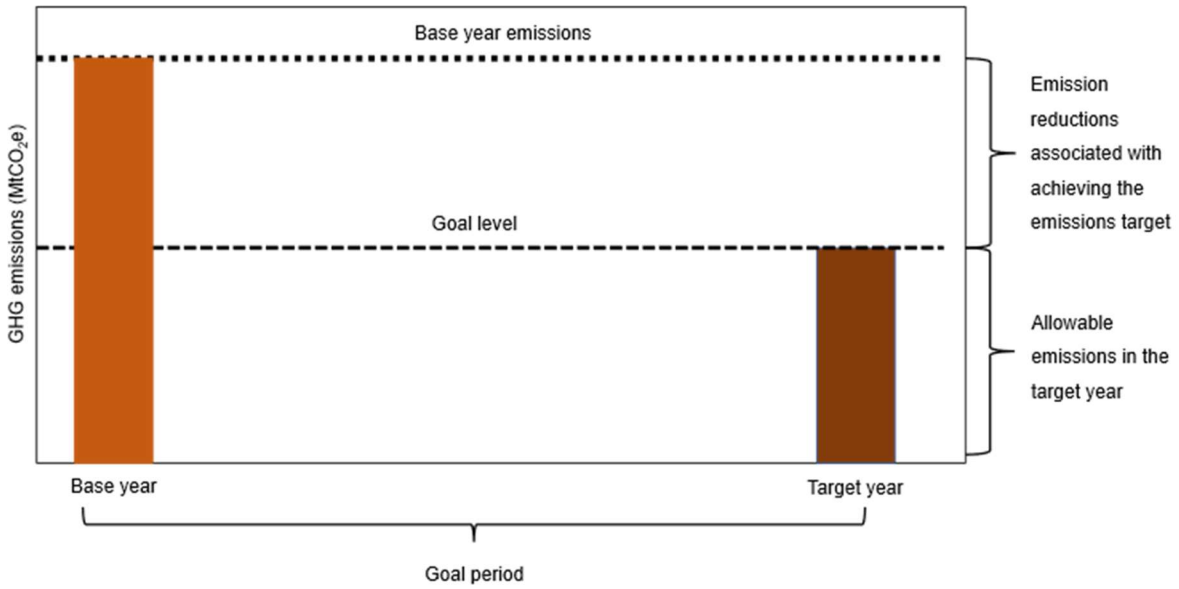


Figure 2-19: Allowable emissions in the target year.

Figure 2-19 presents two bar charts each indicating a respective amount of GHG emissions. The base year is the period used as a baseline while the target year represents the period at which the goal is predicted to be met. If the actual emissions in the target year are the same as the allowable emissions, it indicates that the emissions target is met. If the target year emissions are greater than the goal level emissions, the emissions target is not met (the actual emissions are greater than the allowable emissions), and more rigorous mitigation measures are required. A practical example is illustrated in Figure 2-20.

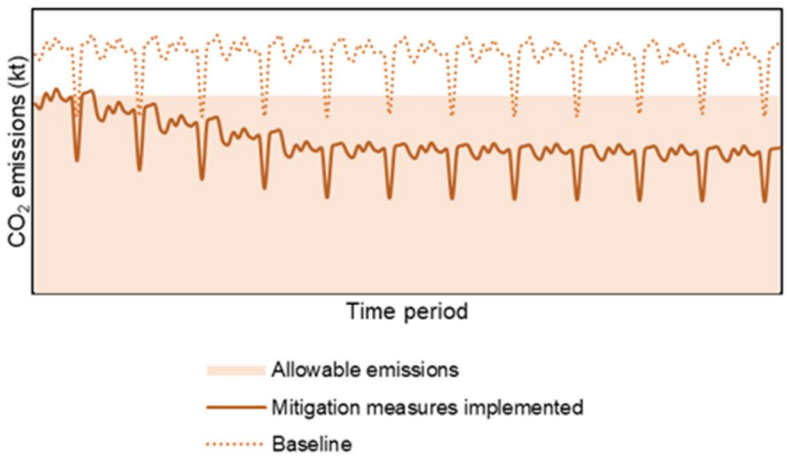


Figure 2-20: Validate whether emissions targets will be met.

The baseline emissions (orange dotted line) indicate the estimated emissions when no mitigation measures have been implemented. From Figure 2-20 it is evident that the emissions target (shaded area) will not be met if no mitigation measures are implemented. If mitigation measures are implemented (solid orange line), the emission target will be met. If this was not the case, more stringent measures are required to ensure that the actual emissions are well below the target emissions.

2.6 Conclusion

In this chapter, the steps required to develop an emissions forecasting model were discussed. The basic steps in the development are problem definition, gathering information, preliminary analysis, choosing and fitting the model, and using and evaluating the forecasting model.

The problem definition for the forecasting model was defined in Chapter 1 and is as follows: South African facilities need to forecast their emissions at facility level; however, only a few facility-level forecasts are currently available in literature (seventeen studies were identified). The rest of the steps applicable to Chapter 2 are illustrated in Figure 2-21.

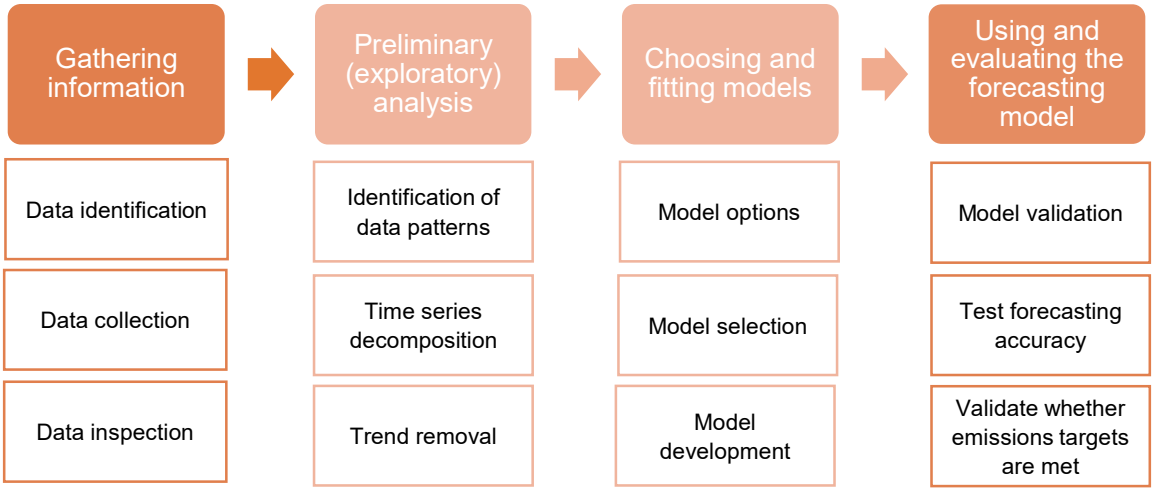


Figure 2-21: Steps and sub-steps for forecasting model development.

In the information gathering step, good quality data is identified, sufficient data is collected, and data is inspected to check for missing or incomplete data.

During the preliminary analysis, data is graphed to identify horizontal, trend, seasonal, and cyclic patterns. These patterns need to be considered when deciding which forecasting model or models to use. It is also important to identify and remove outliers to increase the performance of the developed forecasting model.

Various forecasting models are considered. Based on the advantages and disadvantages of the different models, the forecasting model(s) best suited for the data is determined.

After developing the models in Excel, the models need to be validated using alternative programming languages. For this study's model validation, the programming language, Python was used.

The forecasting accuracy of the chosen model is determined using tests such as MAE, MAPE, MSE, and RMSE. To validate whether the emissions target is met, the accountable emissions should be less or equal to the allowable emissions. If the accountable emissions are more than the allowable emissions, the emissions target is not met, and more stringent mitigation measures are required.

CHAPTER 3 – RESULTS AND DISCUSSION

3.1 Preamble

A general approach for the development of an emissions forecasting model was established in Chapter 2. The approach is based on five steps, namely: problem definition, gathering information, preliminary analysis, choosing and fitting models, and using and evaluating the forecasting model. The five main steps are illustrated in Figure 3-1.

Figure 3-1: Steps for forecasting model development (repeat of Figure 2-1).

The first step was addressed in Chapter 1 as the objective of the study. The purpose was defined as the development of an emission forecasting model for South African industrial facilities. In the second step, the type and quality of the data required are identified whereafter the data is collected and inspected to ensure there are no missing or incomplete values. Next, the data is analysed to identify any recurring patterns or trends in the data and the trend is removed.

In the choosing and fitting models' step, specifically the model development sub-step, data is divided into modelling and validation data. The modelling data is used to develop a representable model which is then validated using the validation data. Model validation is necessary to ensure that the model can estimate future emissions.

To verify the forecasting model development approach, the model is applied to the Scope 1 emissions of an industrial facility based in South Africa. The Scope 1 emissions of the facility consist of fuel combustion, process, and fugitive emissions; hence the specific facility was chosen.

To validate the models developed in Excel, the programming language, Python is used.

3.1 Case study selection

A large industrial facility that consists of various energy intensive activities was selected as the case study. The facility produces a variety of products through integrated plants and manufacturing processes. The facility also produces electricity and steam for on-site consumption and support of the manufacturing processes.

The chosen industrial facility forms part of a large South African company with various facilities located nationally and internationally. The company is committed to reducing its carbon footprint by pursuing gas as an energy source, renewables, and energy efficiency (EE) projects. The company has set a global EE improvement target of 30% by 2030 (compared to their 2017 performance). Thus, to ensure the company reaches its target, an emissions reduction of 18.5% is required by 2025 (assuming a reduction of 2.3% per year). Each facility of the company is expected to contribute to achieving the emissions reduction target.

Emissions that contribute to the facility's carbon footprint include fuel combustion, process, and fugitive emissions. The fuel combustion emissions further consist of 6 fuel combustion activities while the process and fugitive emissions consist of 5 and 1 activity for the respective emissions. The contribution of the various emissions is illustrated in Figure 3-2.

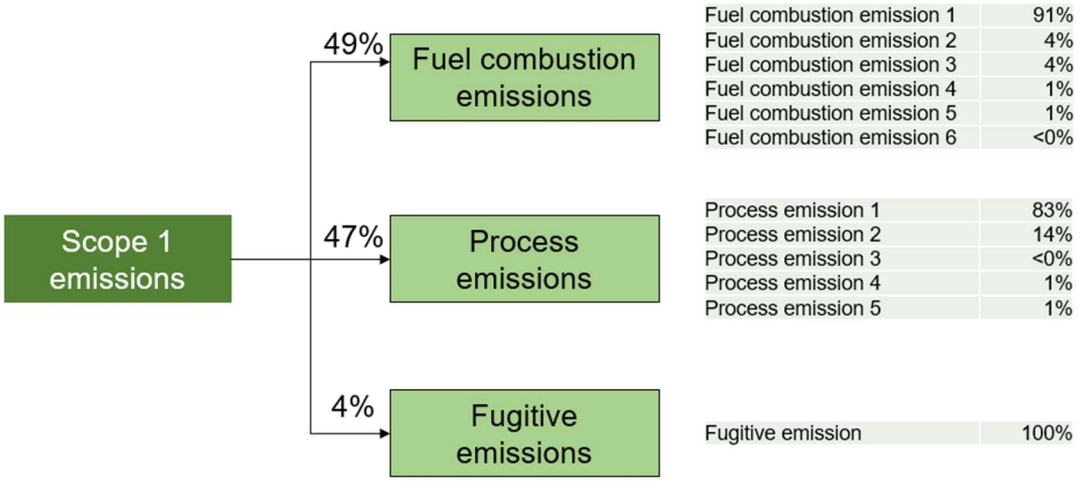


Figure 3-2: Contribution of the facility's emissions.

Fuel combustion emissions (emissions from the combustion of fossil fuels for energy generation) contribute to approximately 49% of the facility's emissions. The fuel combustion

emissions further consist of emissions from 6 different activities where *fuel combustion emission 1* has the largest contribution (approximately 91%).

Process emissions (emissions from the unintended release of gases during the production process) contribute to about 47% of the facility's emissions. The process emissions further consist of emissions from 5 different activities where *process emission 1* has the largest contribution (approximately 83%).

Fugitive emissions (emissions emitted to maintain process stability and safety) have the smallest contribution, contributing to about 4% of the total emissions. The fugitive emissions consist of emissions from only one activity.

3.2 Application of emissions forecasting model on industrial facility

3.2.1 Gathering information

The sub-steps included in the gathering information step are data identification, data collection, and data inspection. These sub-steps are illustrated in Figure 3-3.

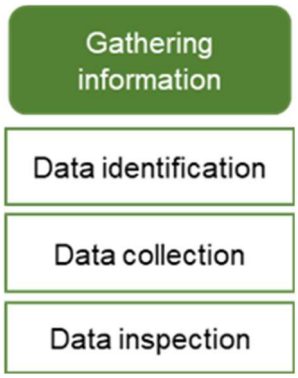


Figure 3-3: Sub-steps for the gathering information step.

During data identification, variables that have an influence on the carbon emissions were identified. The data was then collected from reliable and trustworthy sources whereafter the data was inspected to ensure there are no missing or incomplete data.

Data identification

To determine potential GHG emissions, activity data is required. However, it is not always possible to estimate the activity data for all emissions. Thus, the relationship between emissions and influencing factors is required.

Two factors were identified which influence the carbon emissions of the case study facility. These factors are the raw material input data and production data.

Once the influencing factors have been identified, the desired sample period is determined. The chosen facility's carbon emissions are calculated monthly while the time series data is available daily, which is then aggregated to a monthly total. A summary of the data variables is presented in Table 3-1.

Table 3-1: Summary of data variables.

Variable
Material input
Production data
Activity data 1
Activity data 2
Fuel combustion emissions
Process emissions
Fugitive emissions

The Scope 1 emissions of the chosen facility consist of fuel combustion, process, and fugitive emissions. To determine the values of these emissions, activity data is required. The activity data is calculated using material input and production data. Generic names and descriptions are provided to protect the privacy and anonymity of the facility.

Data should be obtained from trustworthy sources to ensure that only high-quality data is collected. All time-series data can be obtained from the facility's Environmental Performance Indicator (EPI) spreadsheets that can be tracked back to calibrated measurement points. Emissions data can be obtained from the Annexure 3 submissions to the Department of Forestry, Fisheries, and the Environment (DFFE). These data usually undergo stringent evaluations and various audits which increase their reliability.

Data collection

After identifying the desired data, the data needs to be collected. A total of 11 723 data points were collected over a three-year period (January 2017 to December 2019). The breakdown of the data points is provided in Table 3-2.

Table 3-2: Breakdown of data points collected.

Variable	Data frequency	Data points
Material input	Daily	2 190
Production data	Daily	1 095

Variable	Data frequency	Data points
Activity data 1	Daily	5 846
Activity data 2	Daily	2 190
Fuel combustion emissions	Monthly	186
Process emissions	Monthly	180
Fugitive emissions	Monthly	36

The *material input* consists of two data tags, the *production data* of one and *activity data 2* of two tags where each tag consist of about 1 095 data points, over the three-year period. *Activity data 1* consists of 5 data tags with approximately 390 data points for each tag for the same period. The *fuel combustion, process, and fugitive emissions* consist of 6, 5, and 1 activity respectively with 36 data points for each activity over the three-year period.

The collected data is then divided into modelling and testing data, where 7 564 data points (65%) are used during the model development, and 4 159 data points (35%) are used to validate the model.

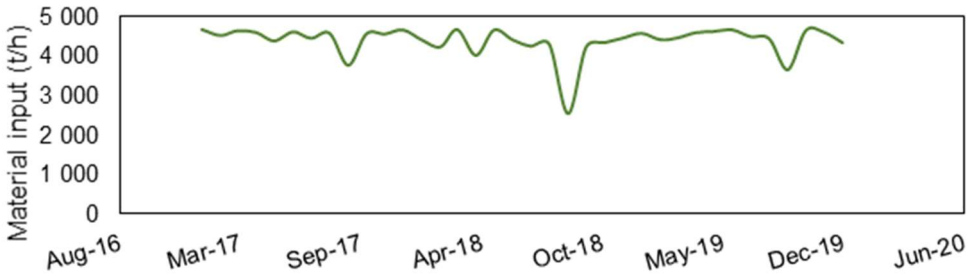
The data was obtained from 3 different sources. The influencing factors and activity data were obtained from the EPI sheets where the data is reported daily and aggregated monthly. The emissions data was obtained from the Annexure 3 submissions to DFFE which are available monthly.

Data inspection

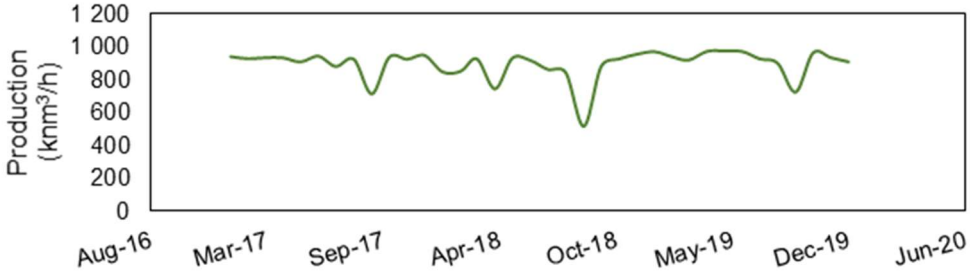
The final step in gathering information is to ensure the data set is complete. The data should therefore be inspected for any missing or incorrect values. Missing data is replaced with values which are determined using linear interpolation; outliers are identified using the IQR method as discussed in Chapter 2.

Missing or incorrect data

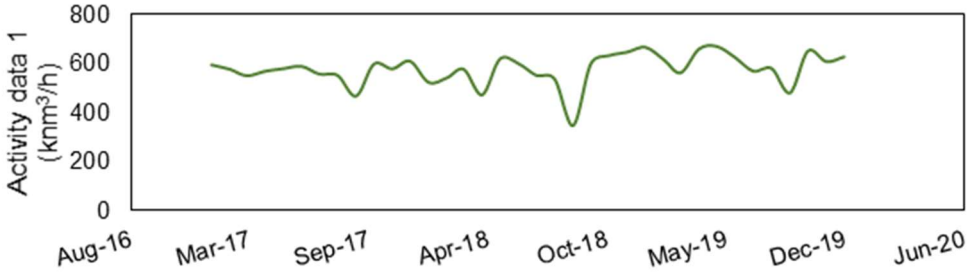
Missing data is a common occurrence in research and has a significant effect on the results of the study, since data processing and data analysis require a complete data set. To identify any missing or incorrect values in the collected data, the time series of the variables were inspected over the three-year period. The time series graphs of the seven respective variables (a to g) can be found in Figure 3-4.



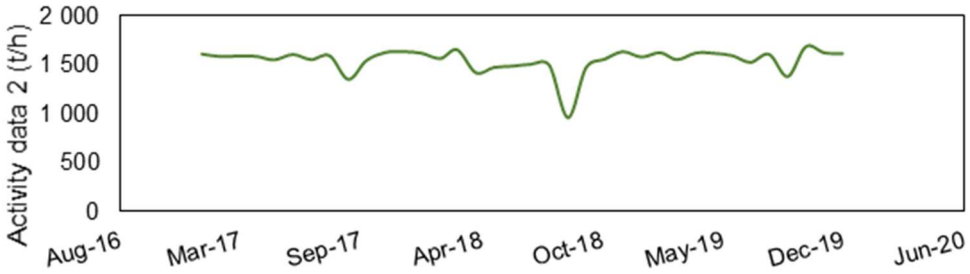
a) Material input data



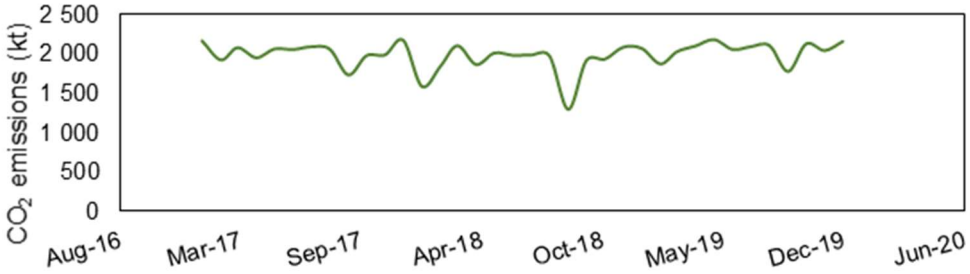
b) Production data



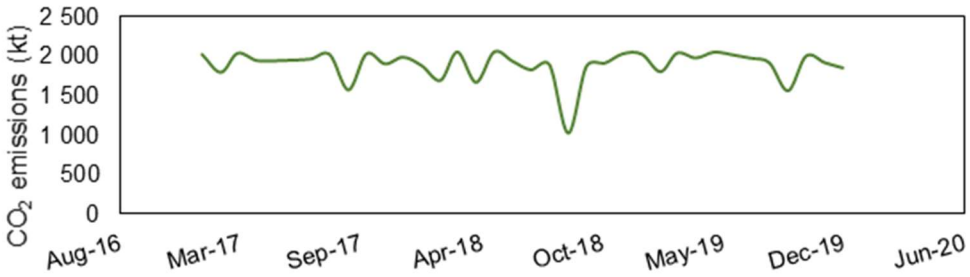
c) Activity data 1



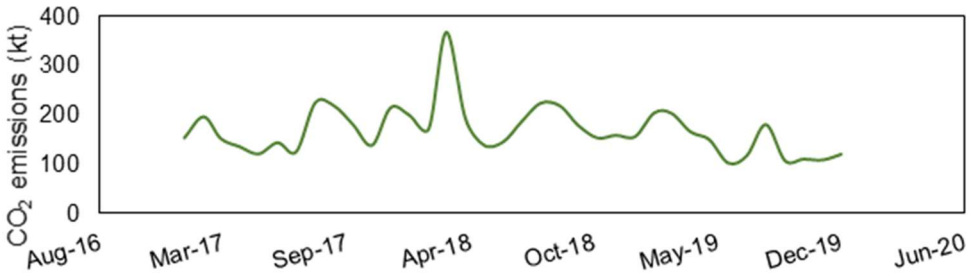
d) Activity data 2



e) Fuel combustion emissions



f) Process emissions



g) Fugitive emissions

Figure 3-4: Time series graphs used for identifying missing or incorrect values.

From Figure 3-4, no missing or incorrect data is visible in the time series graphs for the respective data sets.

Outlier investigation

Outliers in this study were identified using the interquartile range (IQR) method. Mild and extreme outliers were identified using Equations (3-1) to (3-4).

$$lower\ inner\ limit = Q_1 - 1.5 \times IQR \quad (3\text{-}1)$$

$$upper\ inner\ limit = Q_3 + 1.5 \times IQR \quad (3\text{-}2)$$

$$lower\ outer\ limit = Q_1 - 3 \times IQR \quad (3\text{-}3)$$

$$upper\ outer\ limit = Q_3 + 3 \times IQR \quad (3\text{-}4)$$

A total of 263 outliers were identified using the IQR method. These flagged outliers were further inspected and compared with previous values using regression graphs or intensity analyses to determine whether they are true outliers. This evaluation was based on a visual inspection of the flagged values against other values at similar conditions. The material input data was plotted against the production data and the flagged points were evaluated against similar input and production values. Values that differed significantly from other observations and showed no correlation to the other variables were classified as true outliers. Similarly, the

activity data (1 and 2), and emissions were plotted against the production data to visually inspect the flagged values against similar conditions. For the production data, the data from one year was plotted against the data from another year. The regressions graph was then visually inspected to determine if the flagged values can be classified as true outliers.

Only 54 of the initial flagged outliers were identified as true outliers and were subsequently removed using the visual inspection method. The breakdown of these outliers is summarised in Table 3-3.

Table 3-3: Summary of outliers.

Variable	Data frequency	Outliers identified	Outliers removed
Material input	Daily	51	9
Production data	Daily	78	13
Activity data 1	Daily	63	15
Activity data 2	Daily	30	10
Fuel combustion emissions	Monthly	6	3
Process emissions	Monthly	7	3
Fugitive emissions	Monthly	1	1

Table 3-3 indicates that the most outliers were identified and removed for the *activity data 1* variable. This is most likely due to *activity data 1* having the most data points (5 846 data points over three years in daily intervals). The least number of outliers were identified and removed from the *fugitive emissions'* variable data sets which have the least number of data points (36 data points over three years in monthly intervals).

Sample size

The data sample needs to be of a representative size to ensure the data set used for all the evaluations is complete. Time series data requires a sample size of two to six times the seasonality. Time series data is available daily and a seasonality of 365 days can be observed. Thus, a sample size between 730 and 2 190 data points is required (or sample size of 24 to 72 if monthly data is used). Both time series data sets have a sample size of 1 095 (36 for monthly data), thus the sample size is sufficient.

For linear and multivariable linear regression, the sample size is determined using Green's method as described in Equation (3-5).

$$n = 50 + 8p \tag{3-5}$$

Where:

- n is the sample size, and
- p is the predictor.

Linear regression only has one predictor; thus 58 data points are required. Multivariable linear has two predictors; thus 66 data points are required. For linear and multivariable linear regression respectively, a sample size of 72 and 108 data points are available for the variables. Thus, the sample size is sufficient.

Once the identified data is collected and inspected for any missing or incomplete data, the data can be analysed to determine if there are any trends or recurring patterns in the data.

3.2.2 Preliminary analysis

The preliminary analysis of data includes the identification of various data patterns, decompositions of the time series, and removing the trend as illustrated in Figure 3-5.

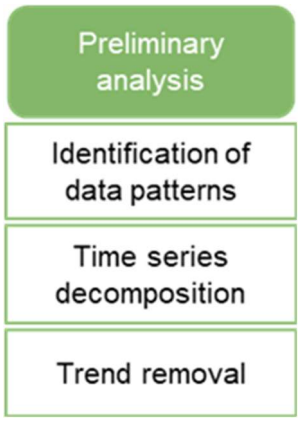


Figure 3-5: Sub-steps in the preliminary analysis step.

The first sub-step involves the identification of data patterns such as trend, seasonality, and cyclical patterns which need to be incorporated into the development of the forecasting model. To further highlight cyclical or seasonal patterns in the data, the time series is decomposed, and the trend is removed.

Data pattern identification

As mentioned in Chapter 2, the following patterns can be identified in time series data: horizontal, trend, seasonal, and cyclical patterns. The data pattern identification for material input and production data can be found in Figure 3-6 and Figure 3-7 as well as Figure 3-8 and

Figure 3-9 respectively. Since these two variables are used to determine the activity data, and subsequently the emissions, only these variables are discussed and analysed in detail.

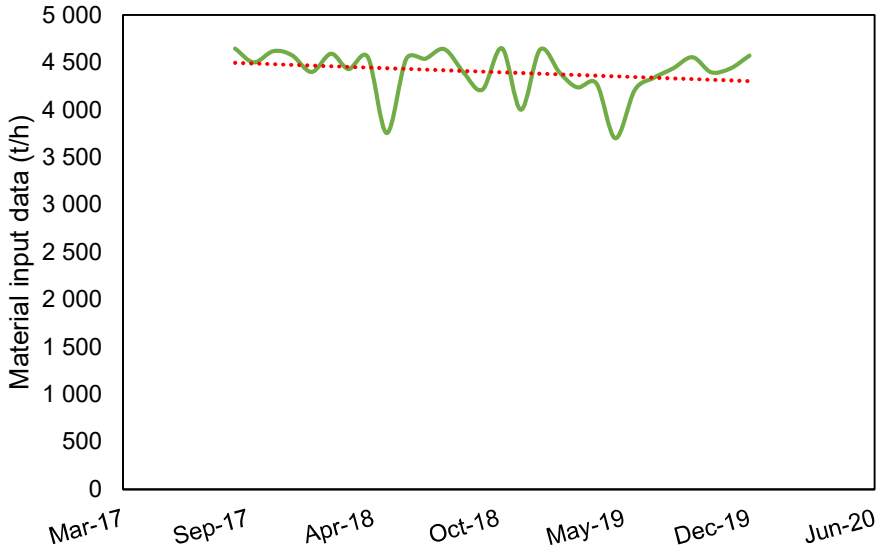


Figure 3-6: Material input time series data.

Figure 3-6 indicates a slight downward trend in the material input over a three-year period. In September of each year, a prominent dip is observed in the material input data. This ultimately indicates seasonality since the pattern occurs at a constant frequency but just varies in magnitude. The dips and rises that do not occur at a constant frequency are indicative of cyclical patterns in the data.

To better differentiate between cyclical and seasonal patterns in the material input data, the data for the respective years are plotted against each other on the same graph as illustrated in Figure 3-7.

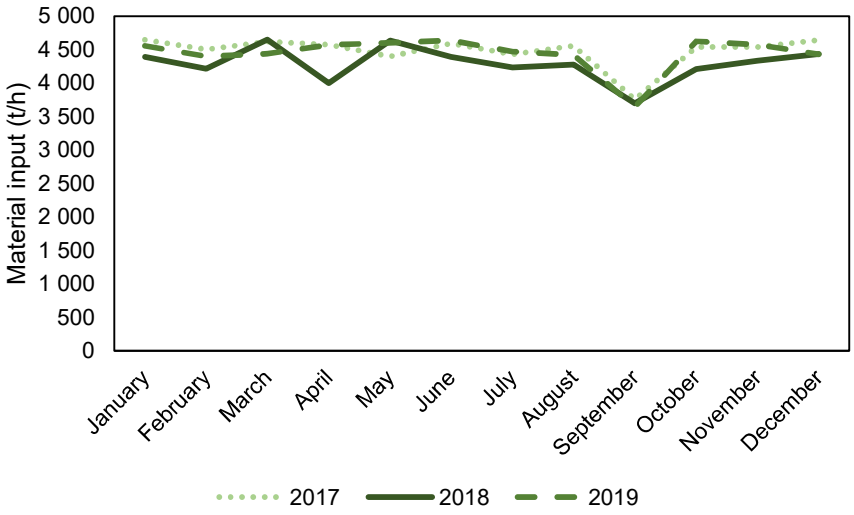


Figure 3-7: Material input data for three consecutive years.

The seasonality in September that was observed in Figure 3-6 is also observed in Figure 3-7. An additional seasonality pattern that was identified for all three years is a dip in the data for February. A cyclical pattern in the data can be observed for March 2018 to May 2018. This cyclical behaviour can be attributed to economic or process instabilities at the facility.

The production time series is illustrated in Figure 3-8.

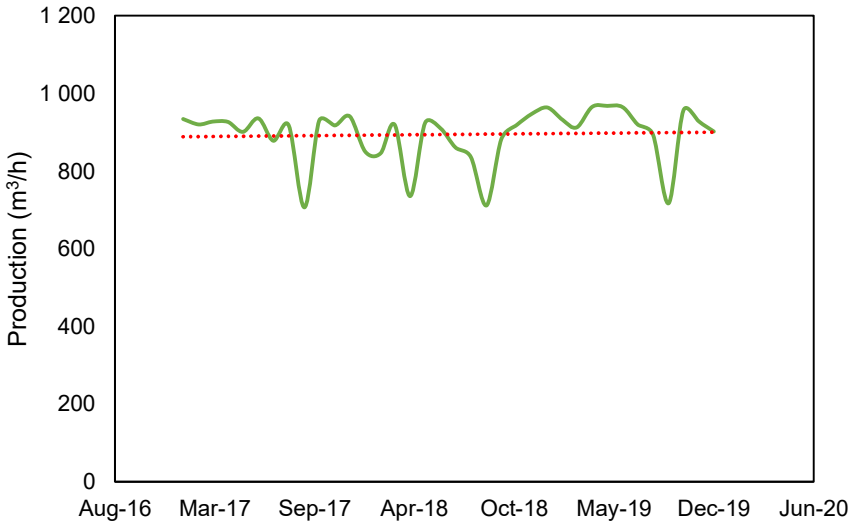


Figure 3-8: Production time series data.

From Figure 3-8, the production remains approximately constant for all three years as indicated by the horizontal trend line. As with the material input data, a prominent dip can be observed in September for each year. To better differentiate between cyclical and seasonal

patterns, the production data for the respective years are plotted against each other as illustrated in Figure 3-9.

For the production data, the 2017 and 2019 data show a slight dip in March while the 2018 data shows an increase for the same month. The 2018 data shows a prominent dip in the April data whereas the 2017 data is slightly decreasing and there is a small peak for the 2019 data. Like the material input data, the production data also shows a prominent dip for September of all years, thus indicating a seasonality of 12 months due to normal operational fluctuations.

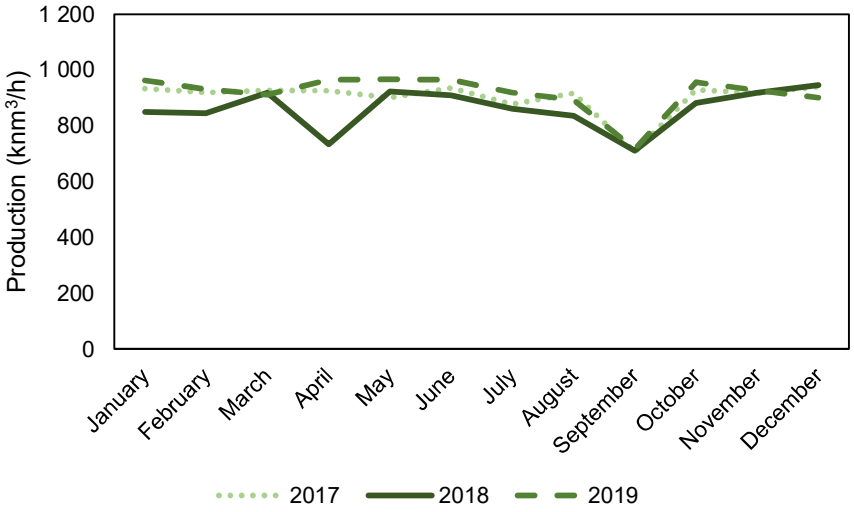


Figure 3-9: Production time series data for three consecutive years.

Other than identifying patterns in the data, the relationship between the variables should also be investigated. The different variables used in the development of the forecasting model are plotted in Figure 3-10. The variables are investigated over the three-year period.

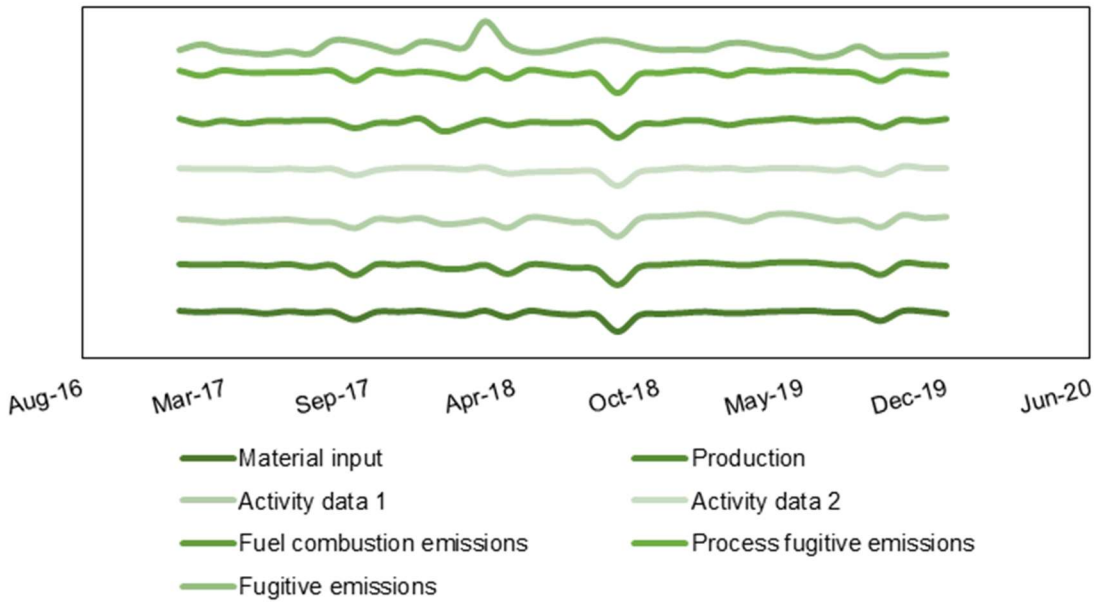


Figure 3-10: Relationship between the different variables.

From Figure 3-10, a similar data pattern can be observed between the various variables with a slight variation visible for the fugitive emissions. Prominent dips in the data can be observed for September month of all the variables for 2017, 2018, and 2019. The variation in the fugitive emissions is considered normal, due to their inherent nature. These emissions are mostly uncontrolled or used to ensure safe operations such as flaring and venting of pressure vessels or harmful gasses. The relationship between different variables can further be illustrated using regression plots. These plots can be found in Appendix B.

Since a similar data pattern can be observed for most of the variables, linear regression and multivariable linear regression are considered for the forecasting model development.

Time series decomposition and trend removal

As seen in the previous section, time series data exhibit a variety of patterns. To highlight these different patterns, the time series can be split into its different components by means of time series decomposition.

In Chapter 2, two types of time series decompositions were identified, namely additive and multiplicative decomposition. Additive decomposition is used when there are changes in the magnitude of the seasonal fluctuations while multiplicative decomposition is used when the variation in the seasonal pattern is proportional to the level of the time series.

Thus, for both the material input and the production data, multiplicative decomposition is used. Since the material input and production data are used to calculate the activity data and subsequently the carbon emissions, only these variables are treated as time series variables and thus decomposed into the respective components.

After decomposing the time series data, the trend can now be removed. Trend removal is often desirable in time series analysis. The trend for both the material input and production data is removed using the methodology discussed in Chapter 2 (Section 2.3.3). The trendline is first calculated using Equation (3-6).

$$T_t = mx_T + C_T \tag{3-6}$$

The trendline is calculated separately for the respective years to ensure the forecasted trend also remains horizontal. The slope (m) and the intercept (C_T) for the material input and production data can be found in Table 3-4 and Table 3-5 respectively.

After the trendline is calculated, the detrended time series can be determined using Equation (3-7), which should result in a horizontal trendline.

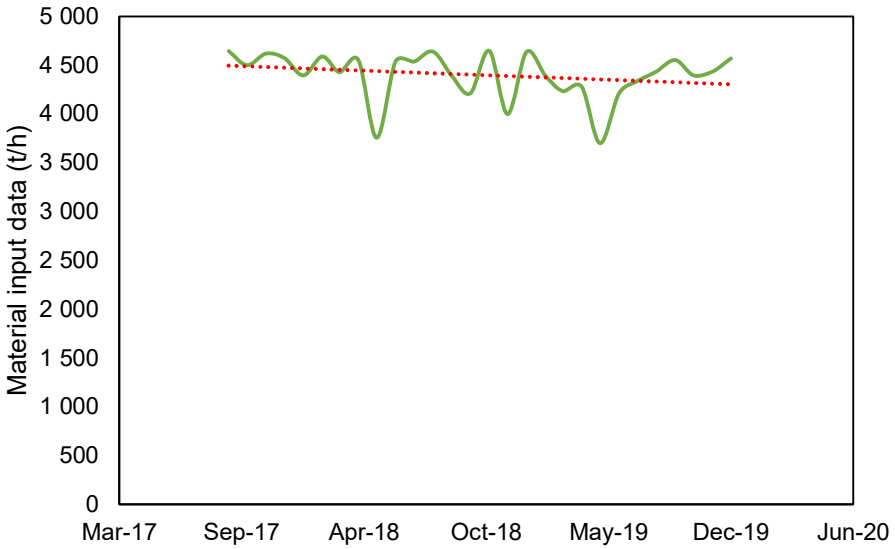
$$y_{DT} = \frac{y_t}{T_t} \tag{3-7}$$

The trendline parameters were calculated using the 2017, 2018, and 2019 data. The slope and intercept for the material input data can be found in Table 3-4.

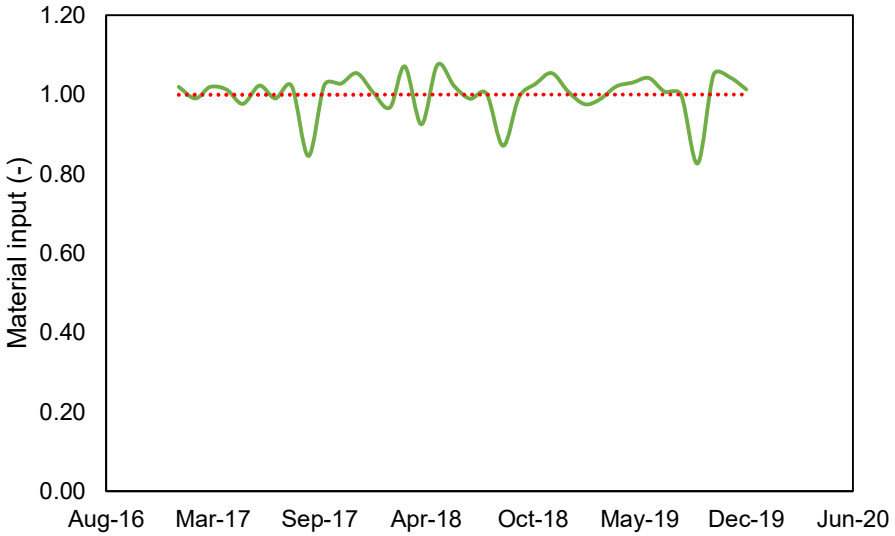
Table 3-4: Material input trendline parameters.

Material input			
Year	2017	2018	2019
$\boldsymbol{m}$	-0.47	-49	-0.44
$\boldsymbol{C_T}$	24 692	25 575	23 575

Using the parameters summarised in Table 3-4 as well as Equations (3-6) and (3-7), the trendline is removed as illustrated in Figure 3-11.



a) Material input data before trend removal



b) Material input data after trend removal

Figure 3-11: Trend removal of material input data.

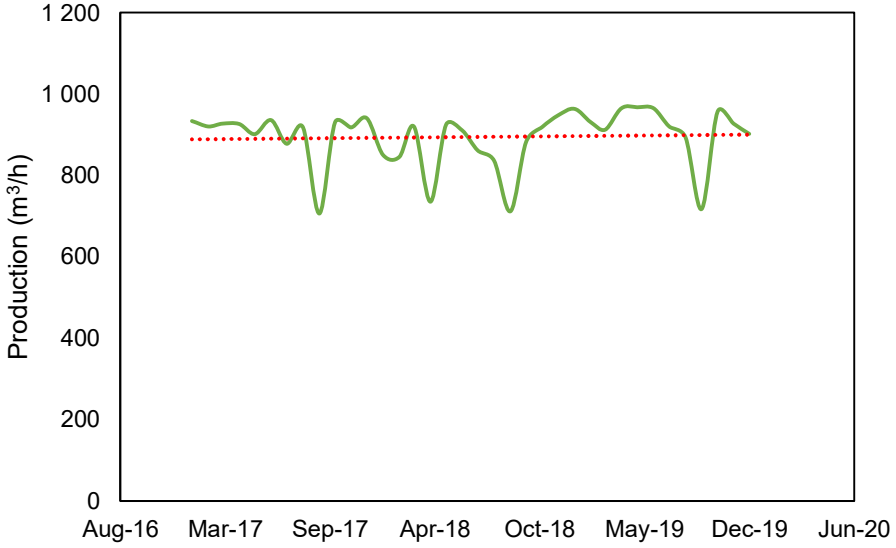
Before removing the trend for the material input data, a decrease in the data can be observed over a two-year period as illustrated in Figure 3-11 (a). Atter the trend is removed, a horizontal trendline is observed in Figure 3-11 (b).

The slope and intercept for the production data are summarised in Table 3-5.

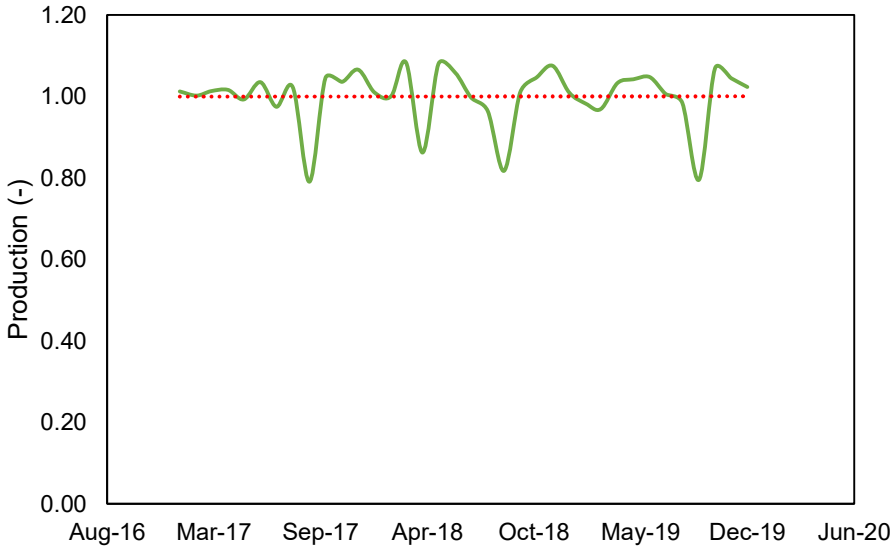
Table 3-5: Production data trendline parameters.

Production data			
Year	2017	2018	2019
m	-0.12	0.12	-0.22
C_T	6 054	-4 277	10 581

The production data before and after trend removal is illustrated in Figure 3-12.



a) Production data before trend removal



b) Production data after trend removal

Figure 3-12: Trend removal of production data.

As with the material input data, a downward trend in the production data can be observed as illustrated in Figure 3-12 (a). A horizontal trend can be observed after removing the trend as illustrated in Figure 3-12 (b).

After the trend for the material input and production data is removed, the model can now be developed.

3.2.3 Choosing and fitting models

Before the development of the emissions forecasting model, a forecasting model(s) needs to be selected as illustrated in Figure 3-13.

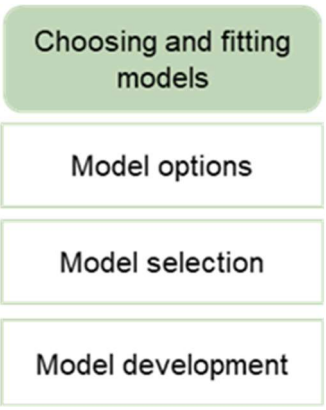


Figure 3-13: Sub-steps when choosing and fitting models.

In Chapter 2 various models such as the ARIMA model, exponential smoothing, and the Grey prediction model were considered for the development of the emissions forecasting model. Based on the advantages and disadvantages of the various models, the models best suited for the data were identified and the models were developed using historic data.

Model selection

Emissions forecasting is dependent on multiple steps. First input or activity data is required which determines the trend. This trend is then converted into emissions by determining the emissions and the input/activity data as illustrated in Figure 3-14.

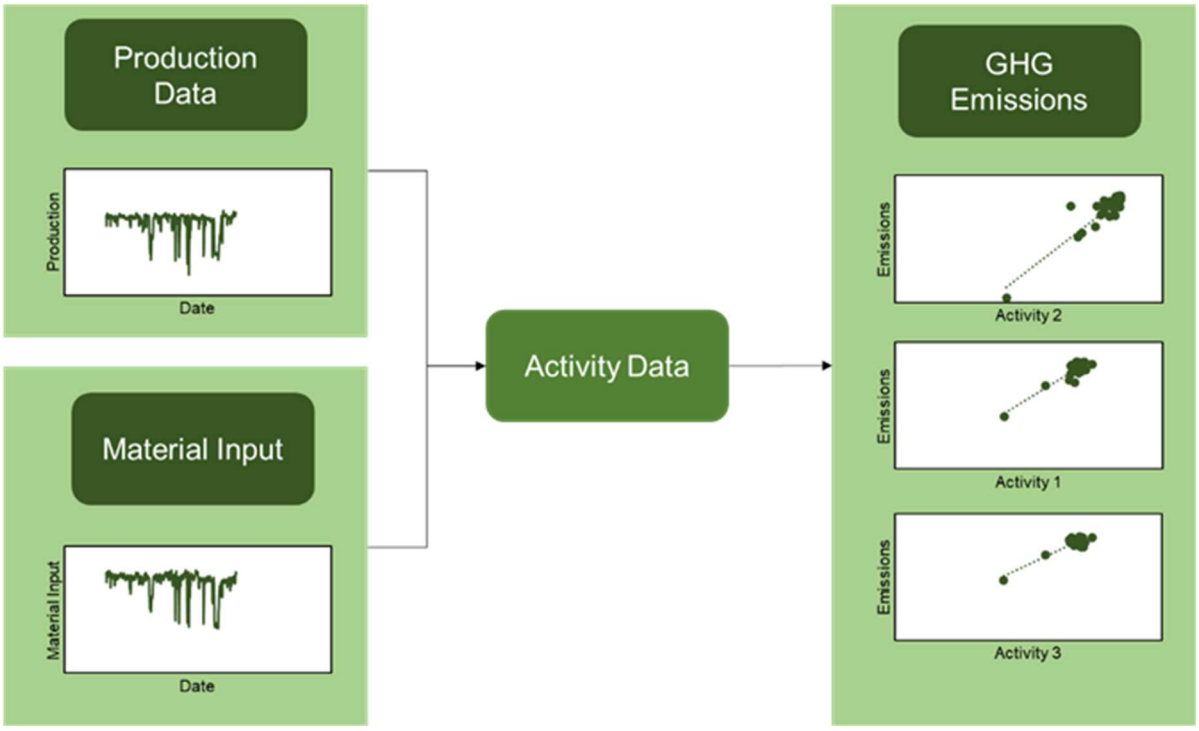


Figure 3-14: Forecasting model selection.

Various prediction models are required to ultimately forecast the facility's carbon emissions. The first is a model that can forecast the material and production data. The forecasted input data is then used within the second model to estimate activity data. Finally, the estimated activity data is used to forecast the carbon emissions of the facility.

For the time series forecasting model (the model used to predict the material and production data), a model is required that is easy to implement, can predict seasonality, and can forecast emissions for the long-term. Based on these criteria, the exponential smoothing model, also known as the error-trend-seasonality (ETS) model, was selected.

To calculate the activity data or the potential emissions, the relationship between the variables needs to be considered. Linear and multivariable linear regression were selected as the best models to evaluate relationships against each variable. After selecting the models best suited for the data, the forecasting models can be developed.

Model development

The model development process consists of five steps. The first step is to split the data into modelling and testing data. Secondly, the data should be fitted, and thirdly, the modelling data is predicted. The fourth step is to determine the accuracy of the fitted data. Lastly, the

developed model can be used and evaluated to forecast the facility's emissions. These steps are illustrated in Figure 3-15.

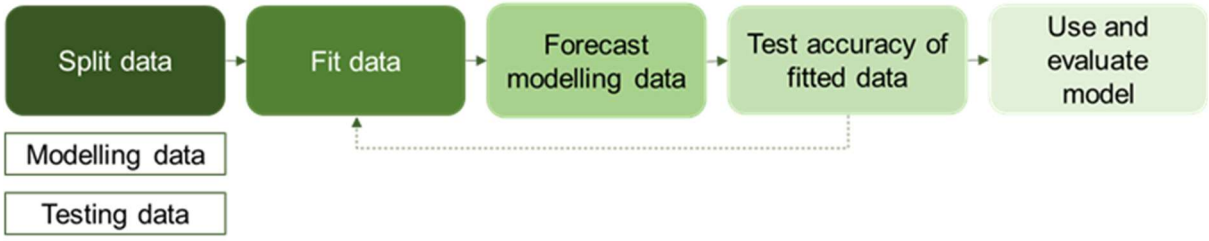


Figure 3-15: Model development steps.

The data used for the development of the forecasting model needs to be split into training and testing data. The training data is used to estimate the parameters of the forecasting model while the testing data is used to validate the accuracy of the forecasting model. This is known as cross-validation. To validate the accuracy of the regression models, an improved version of cross-validation, known as *k*-fold cross validation, is applied. This is illustrated in Figure **3-16**.

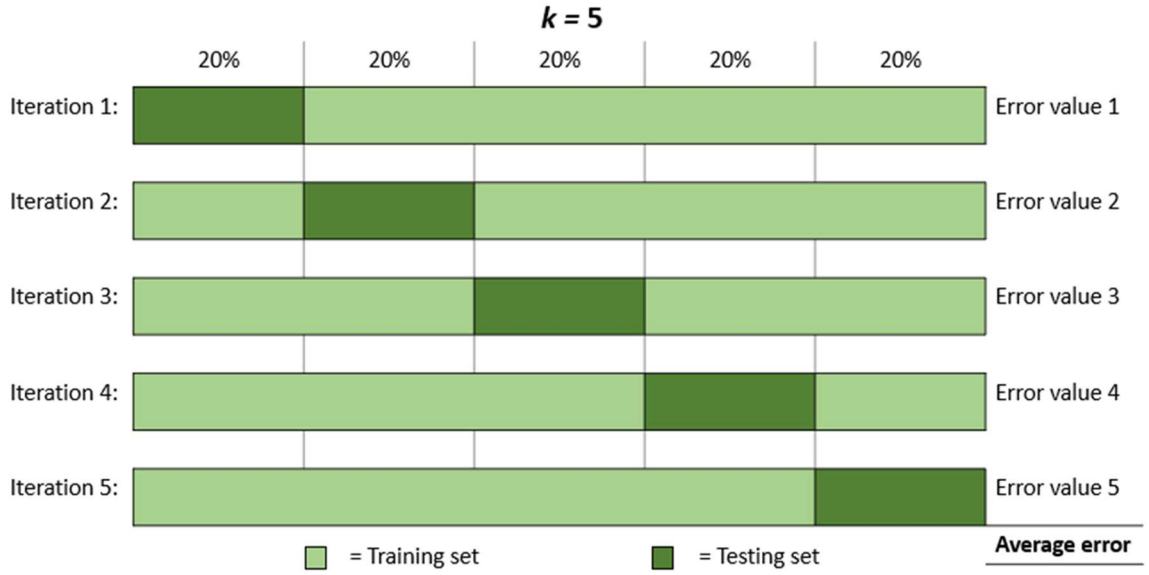


Figure 3-16: 5-fold cross-validation illustration.

To validate the regression models, 5-fold cross validation is used. The data is split into 5 equally sized data sets after undergoing random shuffling. Each iteration uses a different set of testing data and the remaining data sets are combined to form the training set. The iterative process will have five repetitions in total. For each iteration the %MAE and MAPE is then calculated, whereafter the error metrics are averaged for all iterations in the end.

After the data is split into modelling and testing data, the data is fitted to the model and the model parameters are determined.

The material input and production data are both determined using exponential smoothing. Excel has a built-in function to forecast values for a specific future target date using exponential smoothing. The forecasting function is presented by Equation (3-8).

$$\text{FORECAST.ETS}(\text{target_date}, \text{values}, \text{timeline}, [\text{seasonality}], [\text{data_completion}], [\text{aggregation}]) \quad (3\text{-}8)$$

The facility aims to reduce its Scope 1 and Scope 2 emissions by 30% by 2030. Thus, to achieve this target, an emissions reduction of 18.5% is required by 2025. To ensure the emission forecasts are aligned with the Carbon Budgets, the target date used in the function is taken as January 2019 to December 2025. For accurate forecast aimed at the 2030 to 2050 horizon, qualitative forecasting methods can be applied. However, these models are expensive to develop and require inputs from experts in various fields.

In the second argument, the material input and production data for the period January 2017 to December 2019 is inserted, with their corresponding dates selected for the third argument. It is important to note that the input variables as well as the forecasted data are monthly values.

For the seasonality argument, a seasonality of 12 is used as previously observed. For the last two arguments in the forecasting function, the Excel default values were used. The data completion argument has a default value of 1 which indicate that missing values are replaced using interpolation. The alternative is to replace the missing values with zero which does not give an accurate representation of the data.

For the aggregation argument, seven different logics can be chosen (average, count, counta, max, median, min, and sum). For this model, the average logic was selected which results in the formula averaging values with the same timestamp. The resulting time series forecasts for the material input and production data are illustrated in Figure 3-17 and Figure 3-18 respectively.

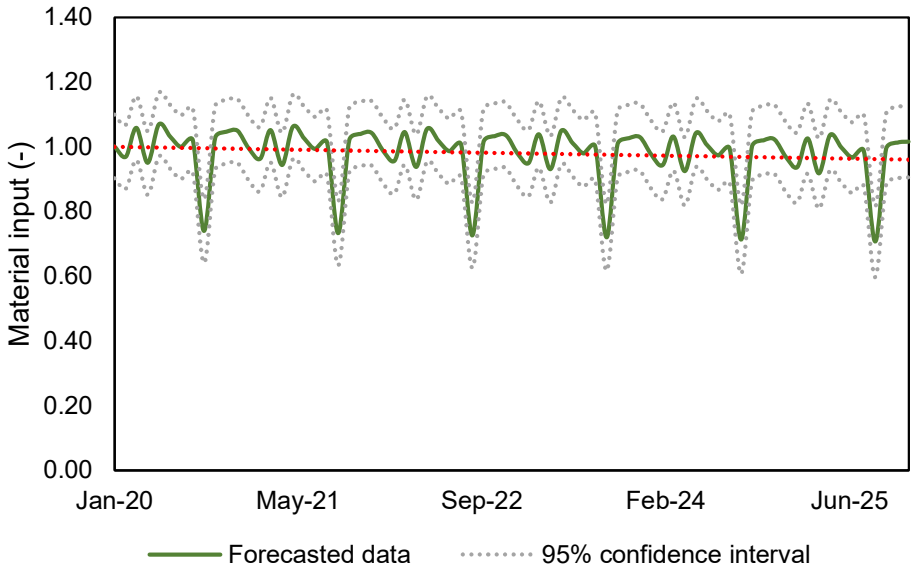


Figure 3-17: Forecasted material input data.

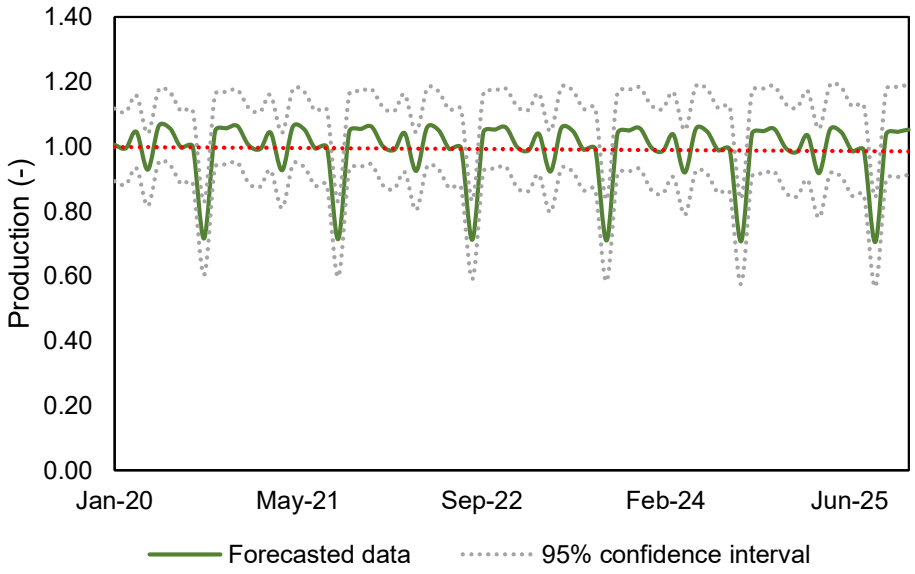


Figure 3-18: Forecasted production data.

The forecasted data illustrated in Figure 3-17 and Figure 3-18 shows a very slight downward trend in the material input data and production data respectively. The forecast illustrates the data when no mitigation measures are implemented on the variables to reduce GHG emissions. These forecasts are used as the baseline material input and production data.

After fitting the parameters for the material input and production data, the activity data can be determined using multivariable linear regression whereafter the emissions data can be

determined using linear regression, or in some cases, a constant multiplier. The IPCC Guidelines contain emission factors for nearly all industrial activities. However, the use of these default factors is associated with a tier 1 methodology, which is associated with a high uncertainty. The chosen facility uses a tier 3 methodology (which has the lowest uncertainty); thus to limit the extent of over or under estimating the emissions, linear regression is used.

The activity data can be determined using multivariable linear regression with the material input (x_M) and production data (x_P) as input variables. The equations used to forecast the activity data can be found in Table 3-6.

Table 3-6: Equations used for forecasting the activity.

Variable	Unit	Forecasting equation
Activity data 1	knm^3/h	$y_{AC_1} = -256.59x_M + 778.52x_P + 55.34$
Activity data 2	t/h	$y_{AC_2} = 157.86x_M + 784.28x_P + 610.60$

The fuel combustion emissions are determined using linear regression with either activity data or production data as input variables. Some fuel combustion emissions cannot accurately be predicted since they are largely affected by instabilities in the combustion process. One of these emissions is *emission* 2 of the fuel combustion emissions.

In Figure 3-19 it is evident that no correlation exists between the dependent and independent variables for *emission 2* as indicated by the almost horizonal trendline. For these emissions, a constant multiplier was used which was calculated by dividing the carbon emissions by a dependent variable (see Appendix B).

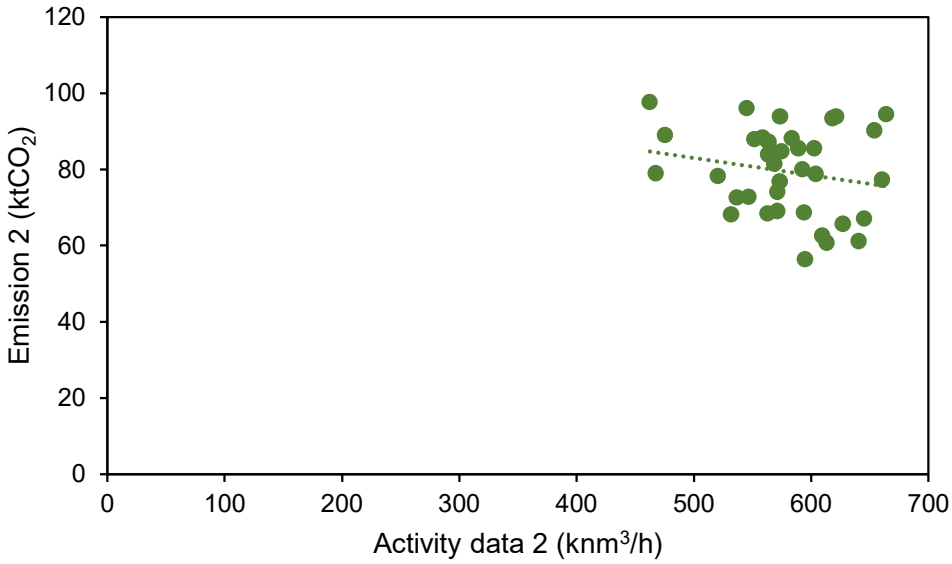


Figure 3-19: Fuel combustion emission 2 linear regression model.

A summary of the equations used to estimate the fuel combustion emissions can be found in Table 3-7.

Table 3-7: Equations used to forecast the fuel combustion emissions.

Variable	Unit	Forecasting equation
Fuel combustion emission 1	tCO_2e	$y_{FC_1} = 1\,061 y_{AC_2} + 169\,873$
Fuel combustion emission 2	tCO_2e	$y_{FC_2} = 138.83 y_{AC_1}$
Fuel combustion emission 3	tCO_2e	$y_{FC_3} = 55.81 y_{AC_2} - 10\,031$
Fuel combustion emission 4	tCO_2e	$y_{FC_4} = 15\,066 x_P$
Fuel combustion emission 5	tCO_2e	$y_{FC_5} = 7.04 \times 10^{-3} y_{FC_1}$
Fuel combustion emission 6	tCO_2e	$y_{FC_6} = 2.17 \times 10^{-3} y_{FC_1}$

The process emissions were determined using linear regression with material input data, production data or another process emission as input variables. As with the fuel combustion emissions, *process emission 5* was calculated using a constant multiplier which can be found in Appendix B. The equations used to forecast the process emissions can be found in Table **3-8**.

Table 3-8: Equations used to forecast the process emissions.

Variable	Unit	Forecasting equation
Process emission 1	tCO_2e	$y_{PF_1} = 1\,759\,988 x_M - 185\,023$
Process emission 2	tCO_2e	$y_{PF_2} = 398\,679 x_M - 129\,585$
Process emission 3	tCO_2e	$y_{PF_3} = 1\,271 x_M + 995.48$

Variable	Unit	Forecasting equation
Process emission 4	tCO_2e	$y_{PF_4} = 2.25 \times 10^{-2} y_{PF_1} + 10\,541$
Process emission 5	tCO_2e	$y_{PF_5} = 1.25 \times 10^{-2} y_{PF_1}$

The fugitive emissions also cannot be forecasted accurately due to fluctuations in the emissions; thus a constant multiplier was used. This constant was determined as 8.84×10^{-2} and was calculated by dividing *fugitive emission 1* with *activity data 1*. The equation used to estimate the fugitive emissions can be found in Table 3-9.

Table 3-9: Equation used to forecast the fugitive emissions.

Variable	Forecasting equation
Fugitive emission 1	$y_{F_1} = 8.84 \times 10^{-2} y_{AC_2}$

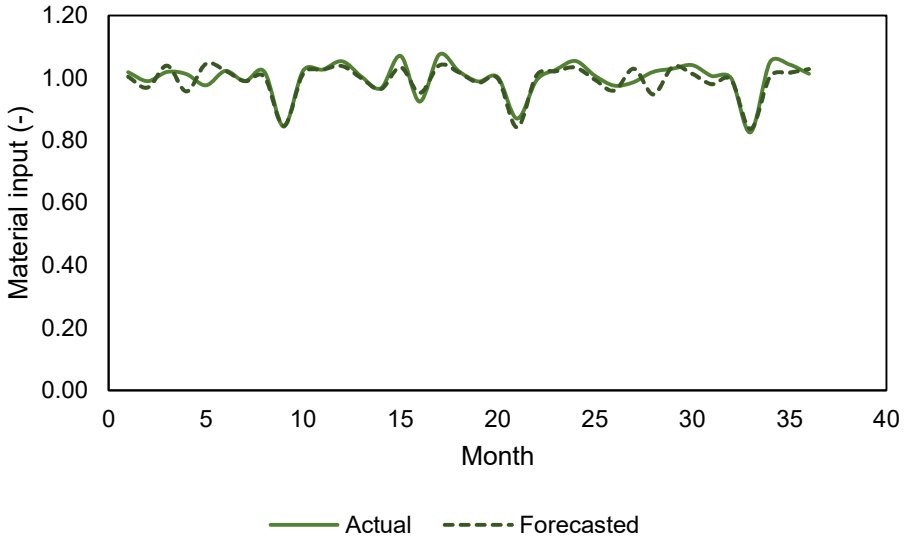
As previously mentioned, the five assumptions of linear regression need to be considered before the regression models can be considered valid. The analysis of these assumptions can be found in Appendix B.

After determining the forecasting parameters using the modelling data, the data is forecasted for the same period whereafter the forecasting accuracy of the fitted data is determined.

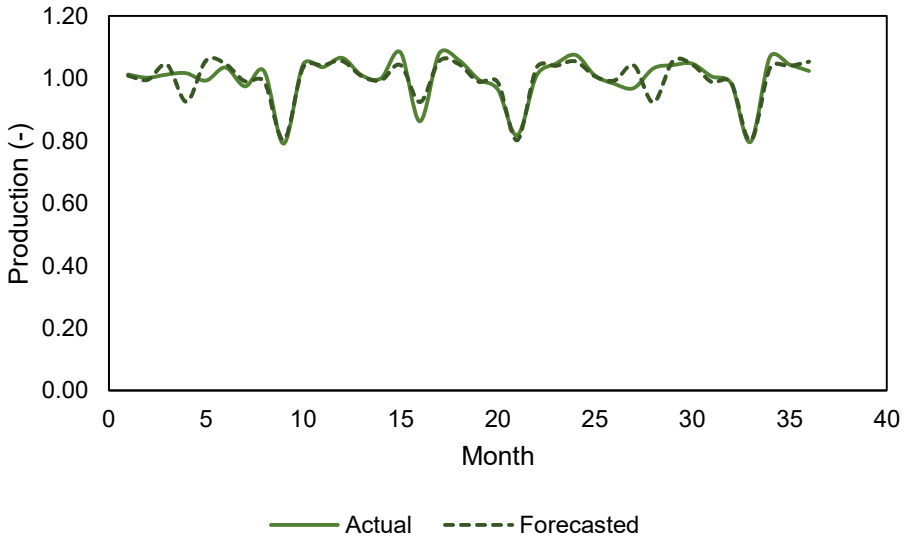
It is important to test the accuracy of the model before it can be used to forecast emissions. The accuracy of the developed model is tested by using the modelling data. This is done by comparing the actual and forecasted modelling data, whereafter parameters such as the mean absolute percentage error (MAPE) and the percentage of mean absolute error (%MAE) are calculated. The MAPE is used to compare the forecasting performance between the data sets while the %MAE is used to evaluate the measure of variability of the model.

For the time series data, however, the forecasted and actual modelling data cannot be compared due to the nature of the forecasting model used. The exponential smoothing model uses the weighted averages of the historical data to forecast the data. Thus, to validate the time series models, the average of the actual 2017 and 2018 data is compared to the forecasted 2019 data.

Since the exponential smoothing model uses weighted averages to forecast the data, the forecasted data is expected to be approximately the same as the historic data. A comparison of the actual 2017, 2018, and 2019 data and the forecasted 2020, 2021, and 2022 data is illustrated in Figure 3-20 (a) and Figure 3-20 (b) for the material input data and production data respectively.



a) Forecasting accuracy of material input data

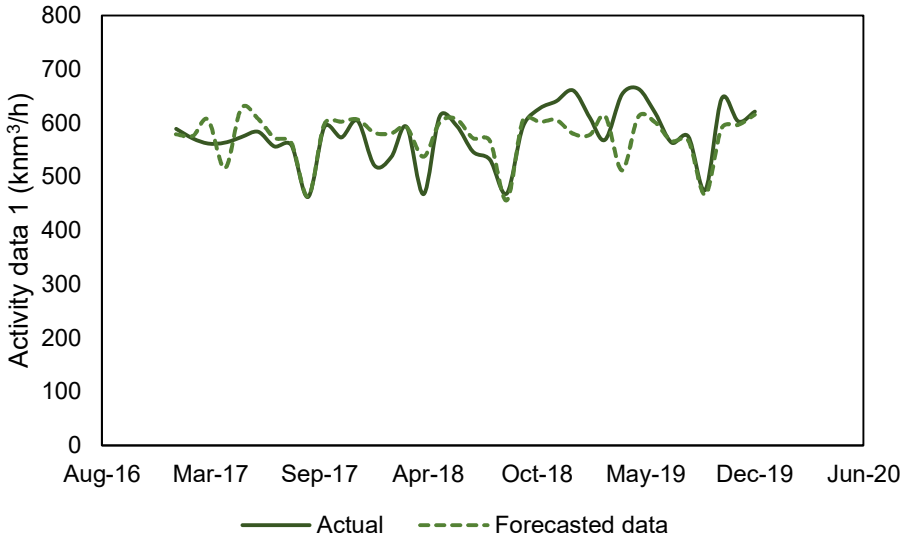


b) Forecasting accuracy of production data

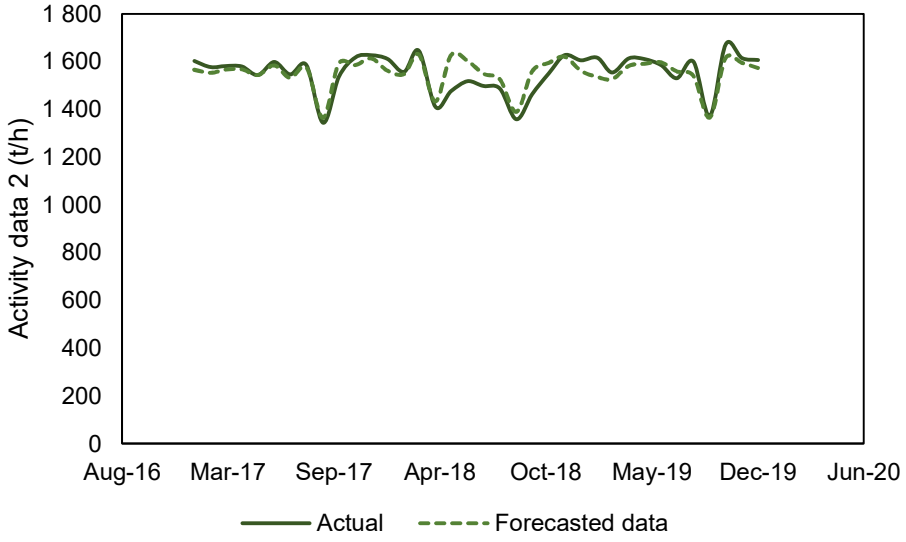
Figure 3-20: Forecasting accuracy of the material input and production data.

Figure 3-20 indicates that the historic data for three year and forecasted data for three years is approximately the same, which is expected, with a similar pattern observed between the actual and forecasted data.

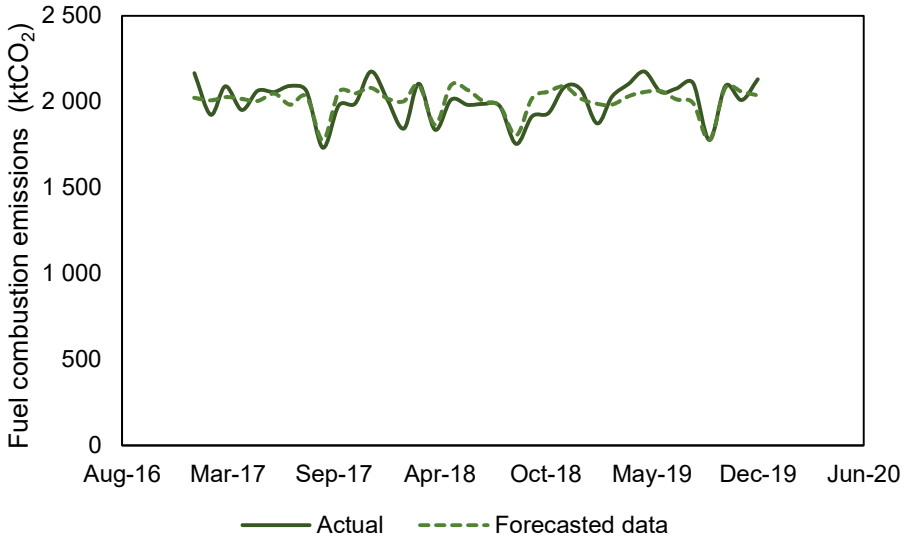
Using the time series forecast the activity data and carbon emissions are subsequently calculated using multivariable regression and linear regression respectively. The resulting forecast for the respective activity data (a and b) and emissions (c to e) are illustrated in Figure 3-21.



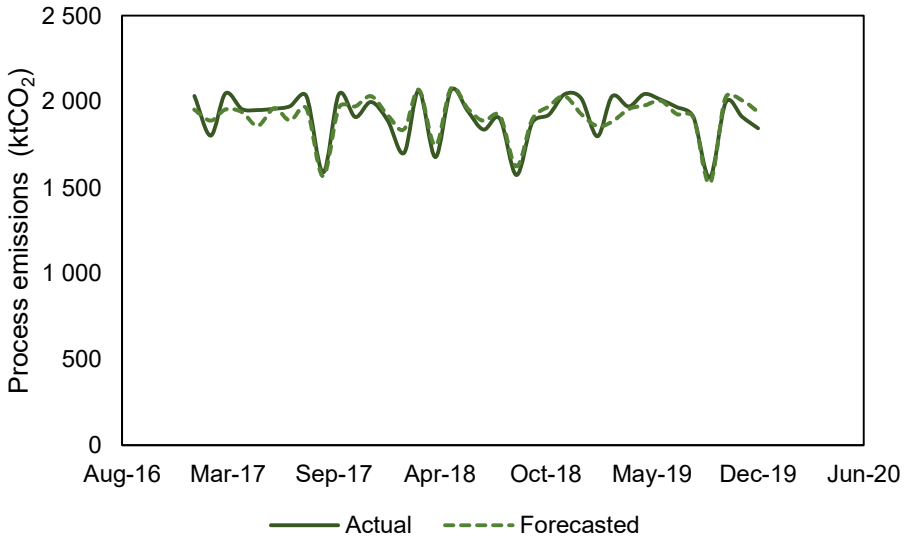
a) Forecasting accuracy of activity data 1



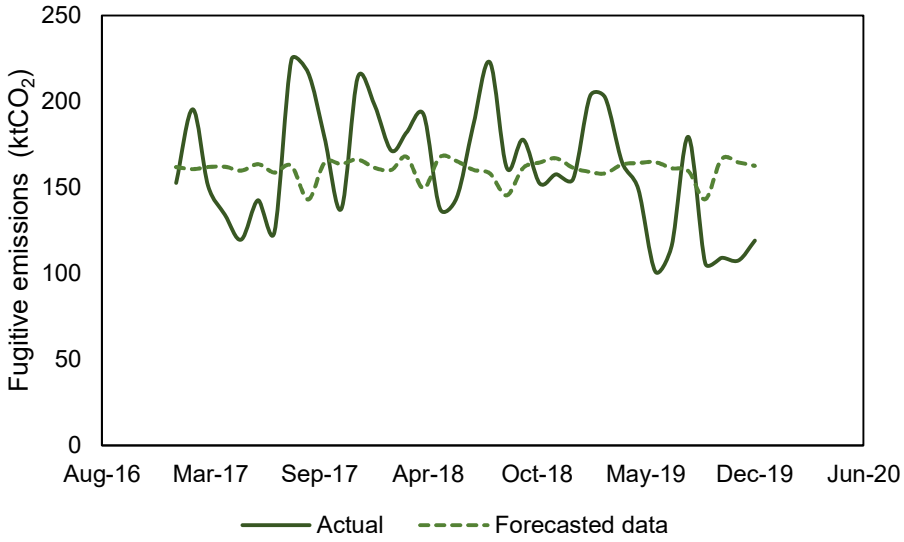
b) Forecasting accuracy of activity data 2



c) Forecasting accuracy of the fuel combustion emissions



d) Forecasting accuracy of the process emissions



e) Forecasting accuracy of the fugitive emissions

Figure 3-21: Accuracy of the forecasted carbon emissions.

From Figure 3-21, the forecasted data for the activity data, the fuel combustion emissions and the process emissions show a good correlation with the actual data. However, a good correlation is not present for the fugitive emissions due to the unpredictability of the emissions data. This is further emphasised by the MAE and %MAE as illustrated in Figure 3-22**.**

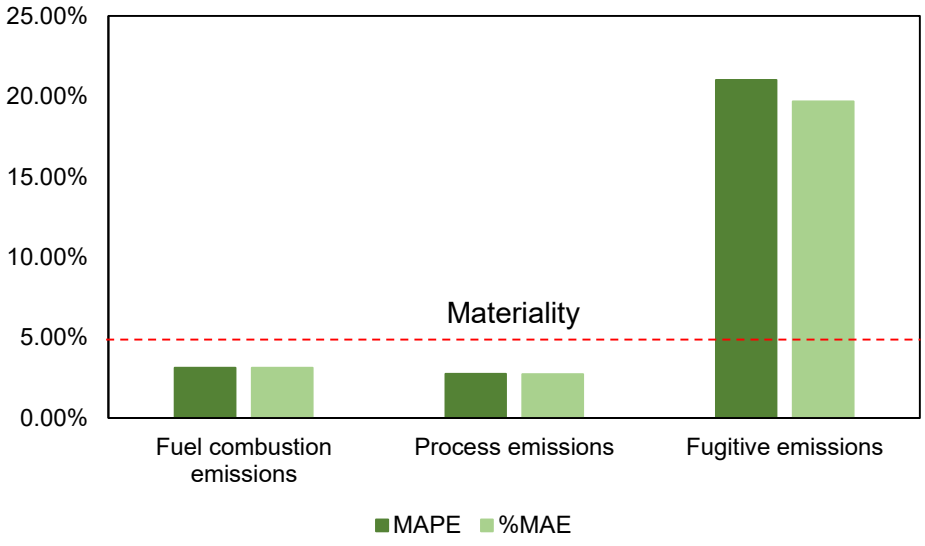


Figure 3-22: Accuracy of forecasted emissions.

Figure 3-22 indicates that the MAPE and %MAE for both the fuel combustion and process emissions are below the materiality of 5%. However, the MAPE and %MAE for the fugitive emissions are greater than the materiality threshold. This is expected since the fugitive

emissions are not highly dependent on either the production or material input. These emissions are emitted to maintain process stability and safety.

However, the fugitive emissions contribute only about 4% of the total emissions. This would mean that if the 20% MAPE of the fugitive emissions were added to the total, the total emissions would only increase by approximately 0.6%. This means that the high MAPE and %MAE for the fugitive emissions are still acceptable when their significance is included [61]. Overall, the forecasting accuracy of the respective emissions is sufficient, and the model can now be used and validated.

3.2.4 Using and evaluating the forecasting model

The forecasting model developed in the previous three steps presents an acceptable accuracy to proceed with the actual emission predictions. After developing and using the forecasting model, it is necessary to test the forecasting accuracy and to determine if the model is reasonable by validating whether the emissions targets will be met as illustrated in Figure 3-23.

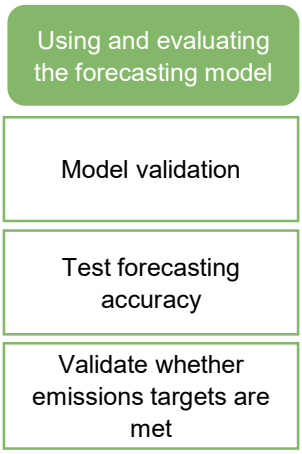


Figure 3-23: Sub-steps in using and evaluating the forecasting model.

The emissions first need to be forecasted using the developed model before the forecasting accuracy of the model can be tested. The forecasting accuracy can then be determined by comparing the forecasted emissions with the validation data. The validation data was not used during model development which indicates how well the model can forecast new emissions.

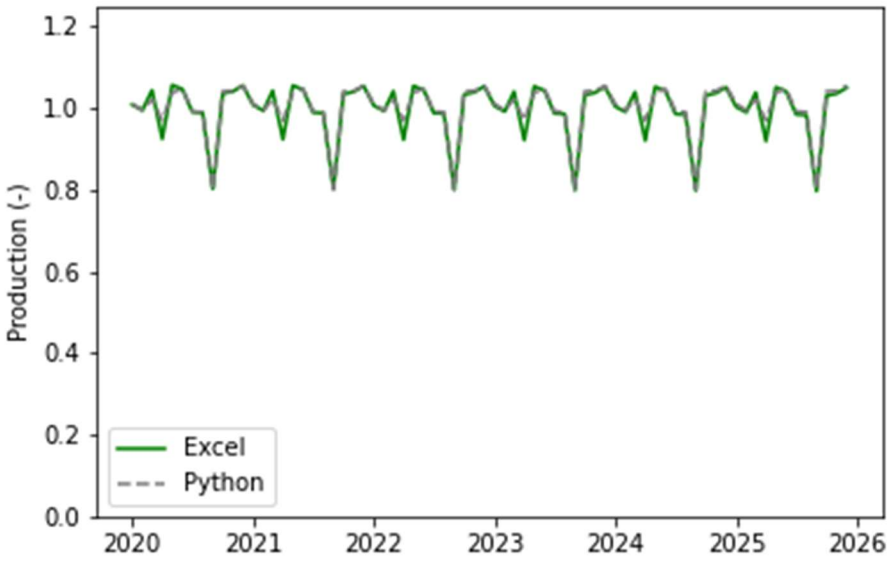
After the forecasting accuracy of the model was deemed sufficient, the facility's emissions roadmap can be incorporated into the forecast. The facility can then validate if its emission targets will be met when implementing the proposed mitigation measures.

Model validation

Machine learning can be used as an alternative to Excel for forecasting values using exponential smoothing as well as linear and multivariable linear regressions. Thus, to validate the Excel models, the programming language, Python, was used. The classes and coding used to develop the Python models can be found in Appendix A.

Both the material input and production data were forecasted using the ETS forecasting model with a seasonality of 12 months. The forecasted values, up to 2025, using Excel and Python for the material input and production data are illustrated in Figure 3-24 a) and b) respectively.

a) Comparison of forecasted material input data using Excel and Python



b) Comparison of production forecasting model using Excel and Python

Figure 3-24: Comparison of ETS forecasting using Excel and Python.

In Figure 3-24, the solid green line represents the values forecasted using Excel and the grey dashed line represents the values forecasted using Python. A similar pattern can be observed between the Excel model and the Python model for the material input data as well as the production data.

For the material input data (Figure 3-24 (a)), the Excel and Python values are similar for the first few years whereafter a small deviation can be observed. While the forecasted values in Python show a horizontal trend, a slight downward trend in the Excel model can be observed. This can be explained by the slight downward trend in the data used to develop the model.

A similar observation is made with the production data (Figure 3-24 (b)). The Excel, and Python values are comparable for the first few years, whereafter a deviation can be observed. However, the deviation in the production data is much smaller than the deviation observed in the material input data.

To determine the magnitude of the deviation between the Excel and Python models, the 2025 average values for the respective models were compared. The results are summarised in Table 3-10.

Table 3-10: Comparison of the average 2025 input data using Excel and Python.

Input data	2025 Excel average	2025 Python average	Difference
Material input (-)	0.9721	1.0000	2.79%
Production (-)	0.9938	1.0000	0.66%

The difference between the average 2025 values using Excel and Python is calculated as 2.79% and 0.66% for the material input and production data, respectively. This insignificant difference between the two models ultimately validates the Excel model to be used for further forecasting. For additional validation, Python was also used to forecast the two activity datasets as well as the emissions. The results are summarised in Table 3-11 and Table 3-12. The graphs and detailed discussions are available in Appendix B.

Table 3-11: Comparison of the average 2025 activity data using Excel and Python.

Data description	2025 Excel™ average	2025 Python average	Difference
Activity data 1 (knm^3/h)	579.28	577.58	0.34%
Activity data 2 (t/h)	1 543.13	1 552.74	0.62%

The difference between the average 2025 values using Excel and Python is calculated as 0.34% and 0.62% for Activity data 1 and Activity data 2, respectively.

Table 3-12: Comparison of the total 2025 emissions using Excel and Python.

Data description	2025 Excel™ total	2025 Python total	Difference
Fuel combustion emissions (ktCO$_2$)	23 942	23 108	3.61%
Process emissions (ktCO$_2$)	22 200	22 943	3.24%
Fugitive emissions (ktCO$_2$)	1 917	1 928	0.56%
Total emissions (ktCO$_2$)	48 059	47 980	0.16%

From Table 3-12, it is evident that the fuel combustion emissions have the largest difference (3.61%) when comparing the Excel model to the Python model, while a difference of 3.24% and 0.56% was observed for the process and fugitive emissions, respectively. When combining all the emissions, a difference of 0.16% can be observed between the Excel model and the Python model.

None of the differences presented in Table 3-10, Table 3-11, Table 3-12 and exceeds the materiality threshold of 5% (as indicated in the National Greenhouse Gas Emission regulations).Thus, the Excel models used to predict the input data, activity data and emissions are successfully validated.

It should, however, be noted that for short-term forecast, a good correlation between the Excel and Python model can be observed which indicates that Excel can effectively forecast data for up to roughly five years. However, to ensure long-term accurate forecasts (five years and more), qualitative forecasting methods are required. Unfortunately, these models are expensive to develop and require input from experts in various fields.

Test forecasting accuracy

The forecasting accuracy of the model was tested by comparing the detrended forecasted 2019 data with the validation data, which consists of the actual 2019 data for the various variables.

The forecasting accuracy statistics for the material input, production and activity data can be found in Table 3-13.

Table 3-13: Input and activity data forecasting accuracy statistics.

Variable	MAPE	%MAE
Material input	2.00%	2.01%
Production	2.37%	2.36%
Activity data 1	4.58%	4.58%
Activity data 2	2.50%	2.48%

From the table it is evident that the MAPE and %MAE for all the variables are below the materiality threshold of 5%.

To determine the accuracy of the forecasted fuel combustion emissions, the weighted MAPE, MAE, and %MAE were calculated using Equations (2-12), (2-13) and (2-14). The forecasting accuracy statistics are indicated in Table 3-14.

Table 3-14: Fuel combustion emissions forecasting accuracy statistics.

Variable	Contribution	MAPE	%MAE
Fuel combustion emission 1	90.01%	3.26%	3.21%
Fuel combustion emission 2	4.09%	15.92%	15.03%
Fuel combustion emission 3	4.62%	11.04%	10.60%
Fuel combustion emission 4	0.63%	43.04%	36.73%
Fuel combustion emission 5	0.49%	86.42%	37.13%
Fuel combustion emission 6	0.15%	332.91%	48.33%
Overall (weighted)	**100%**	**5.77%**	**4.57%**

Fuel combustion emissions 1 is the largest contributor to the total fuel combustion emissions and also has the lowest MAPE and %MAE of the emissions in the table. Although the MAPE

for both *fuel combustion emission 4, 5* and *6* is significantly larger than 5%, it is still acceptable since all three emissions contribute to less than 1% of the total fuel combustion emissions.

The statistics for the overall fuel combustion emissions were calculated by comparing the total actual emissions with the total forecasted emissions. From Table 3-14 it is evident that the forecasting accuracy of the fuel combustion emissions is acceptable since the overall (weighted) MAPE and %MAE are either close to or below the materiality threshold.

The process emissions can secondly be calculated using the material input and production. As with the fuel combustion emissions, the forecasting accuracy of the process emissions was determined using the MAPE, MAE, and %MAE. A summary of these statistics can be found in Table 3-15.

Table 3-15: Process emission forecasting accuracy statistics.

Variable	Contribution	MAPE	%MAE
Process emission 1	82.23%	3.04%	3.02%
Process emission 2	13.90%	6.37%	6.26%
Process emission 3	2.57%	10.03%	9.57%
Process emission 4	0.12%	5.97%	5.58%
Process emission 5	1.18%	23.37%	22.06%
Overall (weighted)	**100.00%**	**3.93%**	**3.87%**

As with the fuel combustion emissions, the forecasting accuracy of the process emissions is acceptable since both the MAPE and %MAE are below 5%. *Process emission 1* has the largest contribution to the total process emissions as well as the lowest MAPE and %MAE. Although *process emissions 5* has the largest MAPE and %MAE, it is still acceptable since it only contributes to about 1% of the total process emissions.

The last evaluation is for the fugitive emissions. These emissions are calculated using a constant multiplier (indicated in Table 3-9). The forecasting accuracy of the fugitive emissions is given in Table 3-16.

Table 3-16: Fugitive emissions forecasting accuracy statistics.

Variable	MAPE	%MAE
Fugitive emission 1	23.37%	22.06%

Both the calculated MAPE and %MAE of the fugitive emissions are significantly larger than the materiality thresholds of 5%. However, since the fugitive emissions are small compared to

the fuel combustion and process emissions as illustrated in Figure 3-25, the MAPE and %MAE are deemed acceptable.

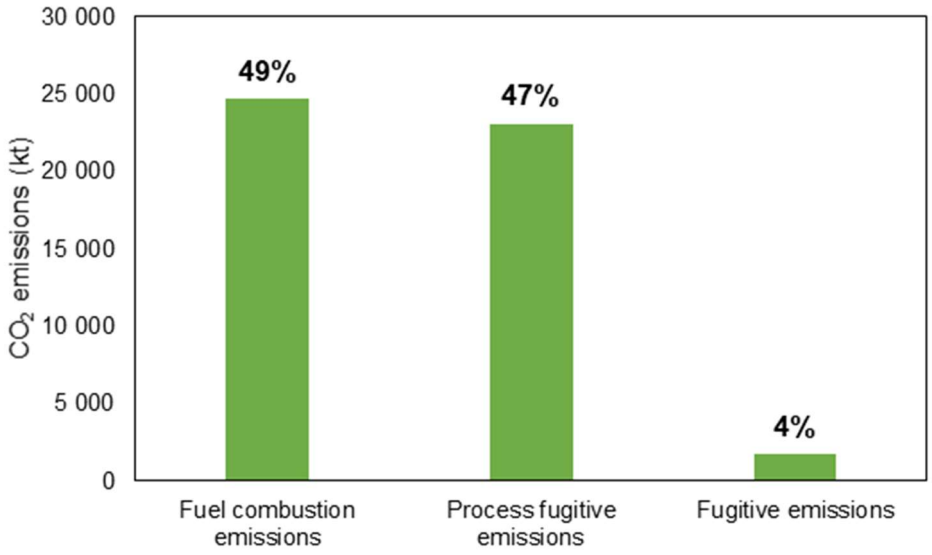


Figure 3-25: Contribution of emissions.

As seen in Figure 3-25, the fugitive emissions contribute to about 4% of the total emissions. This indicates that the variance within the results will only have a minor impact on the final outcomes.

The forecasting accuracy of the total emissions are summarised in Table 3-17.

Table 3-17: Total emissions forecasting accuracy statistics.

Variable	Contribution	MAPE	%MAE
Fuel combustion emissions	49%	5.77%	4.57%
Process emissions	47%	3.93%	3.87%
Fugitive emissions	4%	23.37%	22.06%
Overall (weighted)	**100.00%**	**5.61%**	**4.94%**

As seen in Table 3-17, the MAPE and %MAE calculated for the total emissions are 5.61% and 4.94% respectively. The calculated MAPE and %MAE for the total emissions is acceptable since it is either close to, or below 5%.

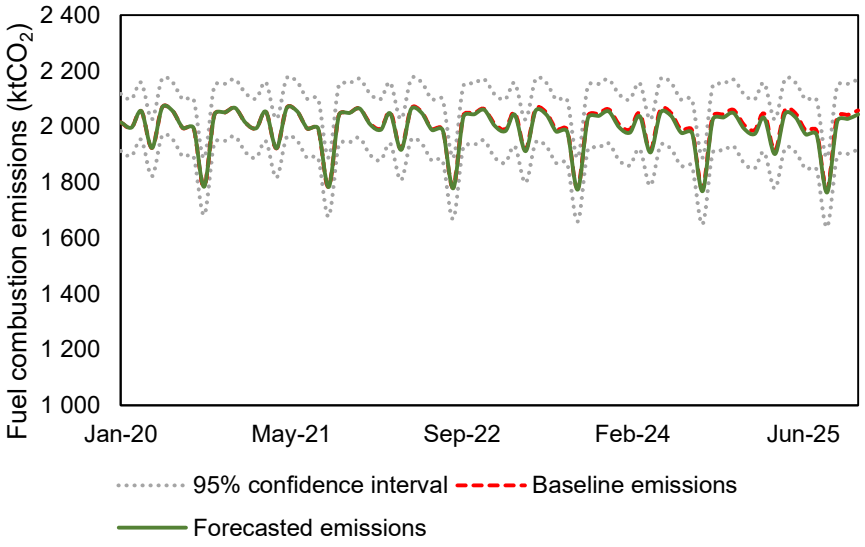
After testing the forecasting accuracy of the developed forecasting model, the proposed mitigation measures can be implemented to see whether the emissions targets will be met.

Validate whether emissions targets are met

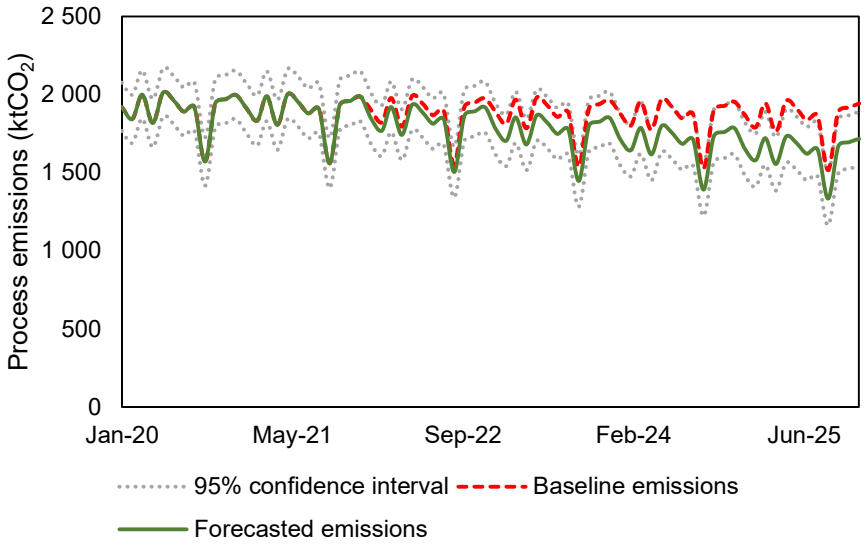
The facility used as a case study aims to reduce its baseline carbon emissions by 30% by 2030, thus an emissions reduction of 18.5% is required by 2025.To achieve this target, several mitigation measures were proposed regarding their Scope 1 and Scope 2 emissions.

One of the mitigation measures proposed to reduce their Scope 1 emissions is a total 22.5% reduction in their material input for the 2021 to 2030 period. Thus, to reduce their emissions by 18.5% by 2025 a 10% reduction in the material input is required by 2021 (assuming a 2.5% reduction per year). As this is one of the facility's biggest projects, this was used to validate whether the facility will reach its emission targets. It is assumed that the production will stay constant.

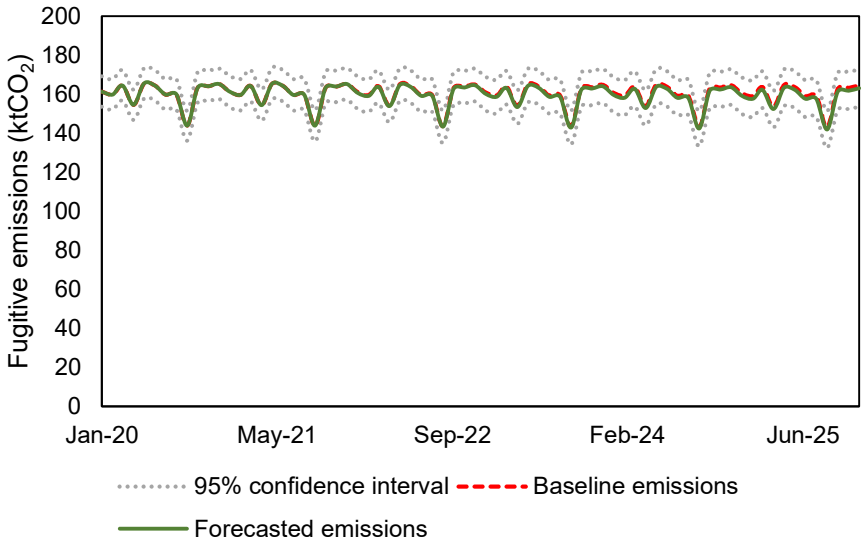
To determine whether the facility's emission targets will be met, the reduction in material input was applied to the baseline forecasts and the carbon emissions were forecasted up until 2025. The resulting forecasted emissions are illustrated in Figure 3-26.



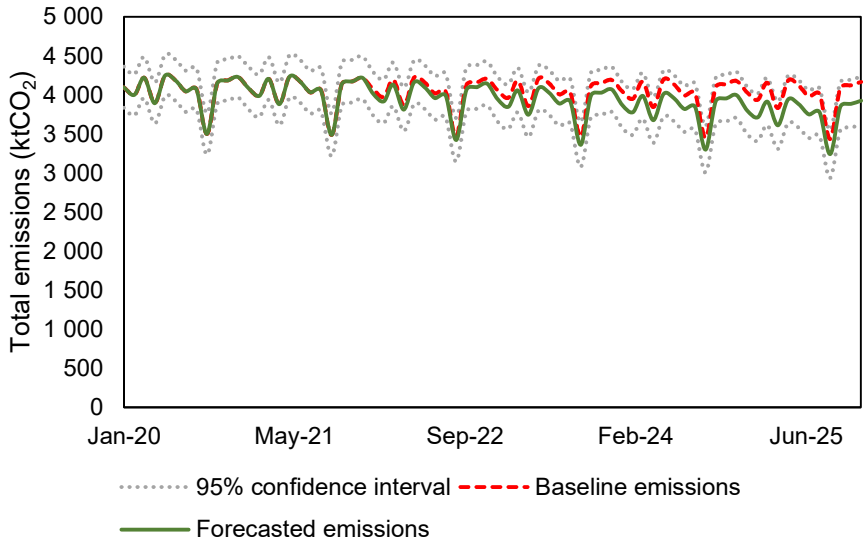
a) Forecasted fuel combustion emissions for 2019 to 2025



b) Forecasted process emissions for 2019 to 2025



c) Forecasted fugitive emissions for 2019 to 2025



d) Total forecasted emissions for 2019 to 2025

Figure 3-26: Forecasted carbon emissions for 2019 to 2025.

From Figure 3-26, a reduction in all emissions can be observed when reducing the material input. The process emissions show the largest decrease in carbon emissions which is expected since about two thirds of the material input contribute to the process emissions. The reduction in the respective emissions is summarised in Table 3-18.

Table 3-18: Reduction in emissions with the implementation of mitigation measures.

Emission type	Baseline emissions	Forecasted 2025 emissions	Target 2025 emissions	Reduction in emissions	Margin of error
	$ktCO_2$	$ktCO_2$	$ktCO_2$	%	%
Fuel combustion	24 275	23 777	20 485	2%	6%
Process	23 287	19 608	19 651	19%	10%
Fugitive	1 991	1 900	1 680	5%	6%
Total	**49 552**	**45 285**	**41 817**	**9%**	**8%**

As is evident from Table 3-18, only the forecasted process emissions are below their 2025 emissions reduction target and thus on track to reaching their 2030 target. Based on the forecasted trajectory, the fuel combustion and fugitive emissions will most likely not meet their individual emissions reduction targets. By only implementing a 18.5% reduction in their material input for the 2021 to 2025 period, the total forecasted emissions are 9% below the 2025 target. Thus, without the implementation of the additional mitigation measures (not evaluated in this study, the facility will likely not meet its emissions targets as specified.

3.3 Conclusion

In this chapter, the steps discussed in Chapter 2 were used to develop an emission forecasting model. The steps used (as illustrated in Figure 3-27) are as follows: gathering information, preliminary analysis, choosing and fitting models, and using and evaluating the forecasting model.

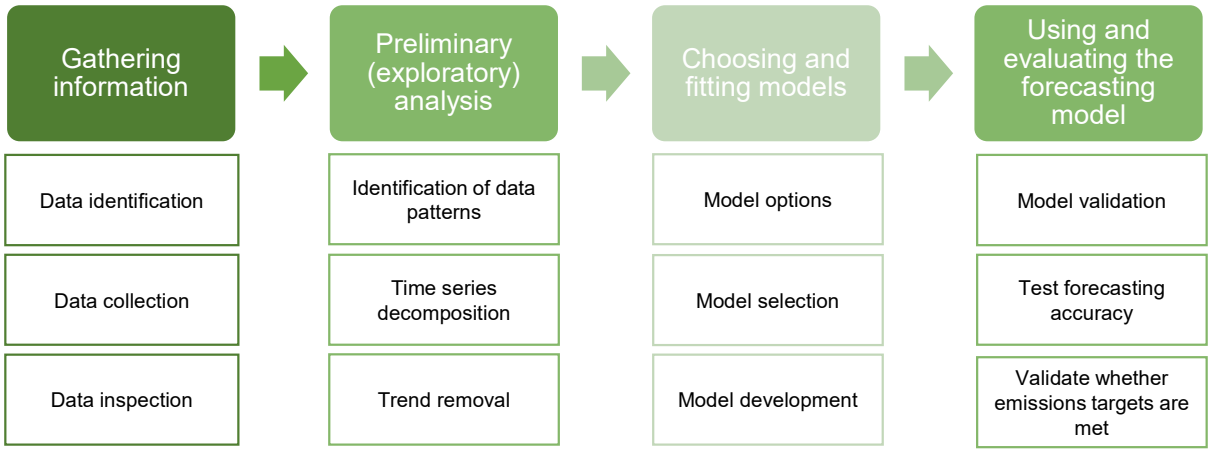


Figure 3-27: Steps and sub-steps for forecasting model development.

The data required to forecast the carbon emissions was identified as the material input and production data. The required data was then collected from 3 different sources after which it was inspected for any missing or incorrect data. No missing values were identified; however, 263 potential outliers were identified of which 54 were identified as true outliers and subsequently removed.

In the second step, trends, seasonality, and cyclical patterns were identified for the time series data. The time series were then decomposed into their respective components whereafter the trend was removed to render the time series stationary.

In the third step, the modelling data was fitted to the different forecasting models. First, the material input data and production data were determined using the error-trend-seasonality model, also known as the exponential smoothing model. Next, the activity data was forecasted using multivariable linear regression with material input data and production data as input variables. Finally, the emissions were calculated using linear regression or, in some cases, constant multipliers.

To validate the models developed in Excel, the programming language Python was used. The Python models gave very similar results compared to the Excel models, thus the developed models are validated.

To determine the accuracy of the developed models, the data was divided into training and testing data and validated using five-fold cross validation with MAPE and %MAE as the error metric. An acceptable MAPE and %MAE were calculated for all the variables.

After the forecasting accuracy was determined, the mitigation measures as specified in the emissions roadmap were implemented in the forecasting model. Based on the forecasted emissions, by only reducing the material input by 18.5% by 2025, the facility's emissions targets for the Scope 1 emissions will not be met. Additional mitigation measures are required to ensure the facility meets its 2025 emission targets and subsequently its 2030 target.

Ultimately, the developed emissions forecasting model can be used by any South African industrial facility to forecast its Scope 1 carbon emissions and to validate whether its emissions targets are likely to be met.

CHAPTER 4 - CONCLUSION

4.1 Preamble

In this chapter the conclusions made and the main findings regarding the study are discussed (Section 4.2). Limitations found and recommendations for further studies regarding the study topic are proposed in Section 4.3. The chapter also serves as proof that all objectives stated in Chapter 1 are met (Section 4.4) before the document is finally concluded in Section 4.5.

4.2 Summary of the study

4.2.1 Chapter 1: Introduction and literature

GHG emissions are one of the biggest contributors to climate change.

To reduce GHG emissions, and subsequently alleviate the impact of climate change, several national and international mitigation policies were implemented. International policies include the UNFCCC, Kyoto Protocol, and the Paris Agreement. National policies include South Africa's low-emissions development strategy, the Climate Change Bill as well as the Nationally Determined Contributions which are required in terms of the Paris Agreement.

From the national and international policies mentioned above, mitigation measures were developed. These mitigation measures include carbon tax, carbon budgets as well as mitigation plans.

To determine whether a facility's emissions are within its allocated carbon budget, emissions forecasting is required. Emissions forecasting can also be used to set up a GHG mitigation plan.

South Africa's current mitigation policies require the forecasting of emissions at facility level. However, emission forecasting models that are available in literature are mainly used for national or regional areas. Thus, an emissions forecasting model is required that can predict carbon emissions at facility level. Therefore, the objective of this study is to identify the steps to develop emission forecasting models for South African industrial facilities that can:

- Identify potential emissions reduction opportunities,
- Determine the effect of different mitigation measures on the facility's total GHG emissions, and

- Determine whether emissions targets will be met with the mitigation measures already in place.

The significance of the study is highlighted by evaluating the carbon tax liability of six South African industries that participate in the carbon budget system. These industries include electricity generation, petrochemical production, mining and minerals, iron and steel production, and cement production. By participating in the carbon budget system, taxpayers can receive a 5% tax free allowance on their total carbon tax liability. A total of 16.77 $MtCO_2e$ emissions of the six reported entities are exempted from carbon tax due to the facilities' participation in the carbon budget system. This amounts to a R 2.3-billion avoided carbon tax liability based on the 2021 carbon tax rate of R 134/tCO_2e.

4.2.2 Chapter 2: Methodology

In Chapter 2, the methodology used to meet the study objectives as outlined in Chapter 1 are discussed and an overview of the forecasting models available in literature is given.

The emissions forecasting models were developed using the following development steps: problem definition, gathering information, preliminary analysis, choosing and fitting models, and using and evaluating the forecasting model. The steps as well as their sub-steps are illustrated in Figure 4-1.

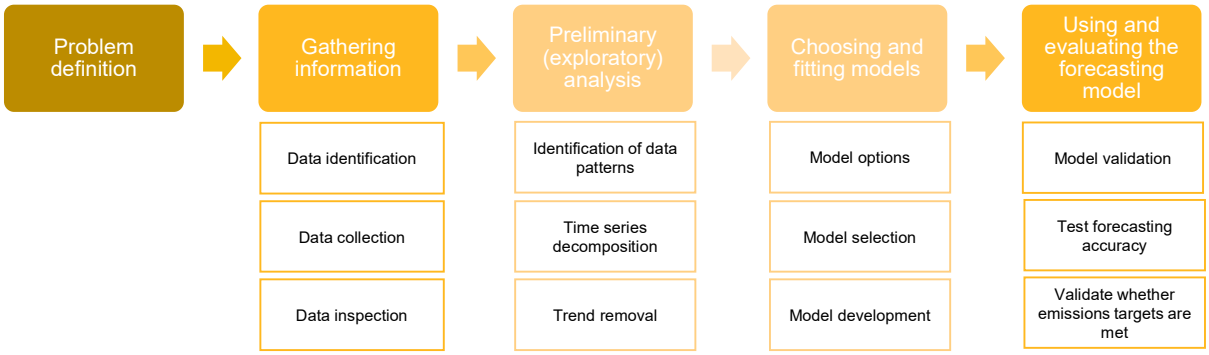


Figure 4-1: Basic forecasting model development steps and sub-steps.

During the gathering information step, the variables that are required to forecast the carbon emissions were identified. Data was also checked to ensure good quality data from reliable sources was obtained. After collecting the necessary data, the data was inspected for any missing or incomplete values whereafter any missing values were estimated using linear interpolation.

During the preliminary step, various time series data patterns were identified. The identified patterns include trend, seasonal, and cyclical patterns. The time series was then decomposed into the different components whereafter the trend was subsequently removed.

Various forecasting models were considered for the development of the emissions forecasting model. These models include the Grey Bernoulli model, the ARIMA model, the exponential smoothing model as well as linear and multivariable linear regression.

The forecasting models best suited for the data was then selected. To forecast the time series data, exponential smoothing (error-trend-seasonality model) was used. Next, the activity data was determined using multivariable linear regression whereafter linear regression was used to determine the carbon emissions.

After selecting the forecasting models best suited for the data, the model is developed using the steps illustrated in Figure 4-2.

Figure 4-2: Forecasting model development steps.

During the forecasting model development, the data was first split into modelling and testing (validation) data and validated using five-fold cross-validation. After fitting the data to the model, the model was validated, and the forecasting accuracy was determined whereafter the data was forecasted using the developed model.

4.2.3 Chapter 3: Results and discussion

In Chapter 3, the forecasting model development steps (developed in Chapter 2) were applied to a large industrial facility as a case study.

Before using the developed model to forecast new emissions, the forecasting accuracy of the modelling data was tested by calculating the mean absolute percentage error (MAPE) and the mean absolute error percentage (MAE%). For the material input and production, the three years of historic data was compared to three years of forecasted data. The forecasted data provided similar results to the historic data, which is expected. The MAPE and %MAE of the respective emissions is given in Figure 4-3.

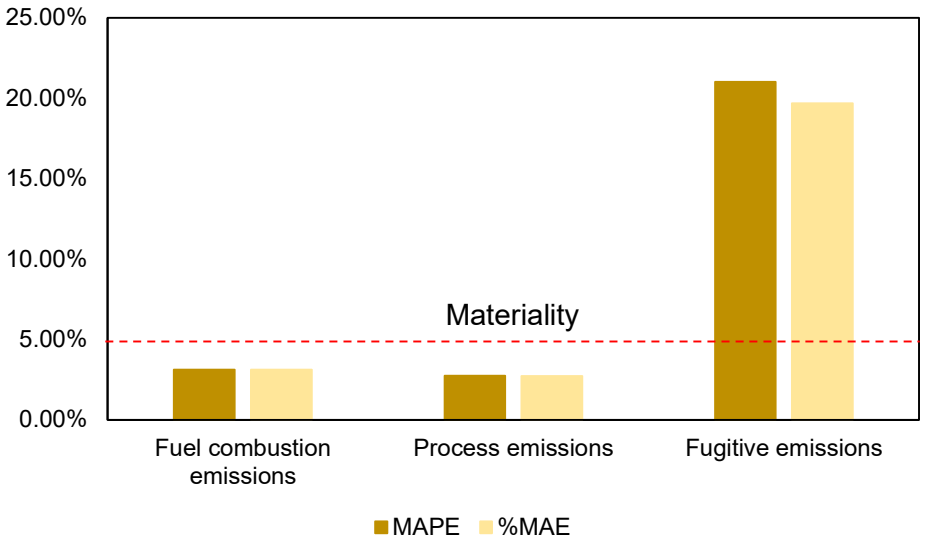


Figure 4-3: Accuracy of forecasted emissions.

The MAPE and %MAE for both the fuel combustion and process emissions are below the 5% materiality threshold while the fugitive emissions are greater than the threshold as illustrated in Figure 4-3. Due to the small contribution of the fugitive emissions to the total emissions, the forecasting accuracy of the respective emissions is sufficient, and the model was subsequently used and evaluated.

To validate the forecasting model, the model was used to forecast the testing data, whereafter the actual and forecasted testing data was compared. The forecasting parameters for the validation data can be found in Table 4-1.

Table 4-1: Forecasting accuracy statistics.

Variable	MAPE	%MAE
Material input	2.00%	2.01%
Production	2.37%	2.36%
Activity data 1	4.58%	4.58%
Activity data 2	2.50%	2.48%
Fuel combustion emissions	5.77%	4.57%
Process emissions	3.93%	3.87%
Fugitive emissions	23.37%	22.06%
Total emissions	**5.61%**	**4.94%**

From the table it is evident that the MAPE and %MAE for all the variables except *activity data 1* and the *fugitive emissions*, are close to, or well below the materiality threshold. Thus, the developed model can accurately predict potential carbon emissions.

After validating the emissions forecasting model, emissions until 2025 were forecasted. The resulting emissions have a horizontal trend and were thus used as the baseline emissions, which indicate the future emissions when no mitigation measures have been implemented.

To validate whether the facility will meet its emissions targets, the facility's emissions roadmap was implemented on the forecasted data.

The facility in the case study aims to reduce its Scope 1 and Scope 2 emissions by 30% in 2030. Thus, an emissions reduction of 18.5% is required by 2025. One of the Scope 1 mitigation measures proposed to reach this target is a 10% reduction in its material input for the 2021 to 2025 period. Figure 4-4 illustrates the forecasted carbon emissions before and after implementing one of the proposed mitigation measures.

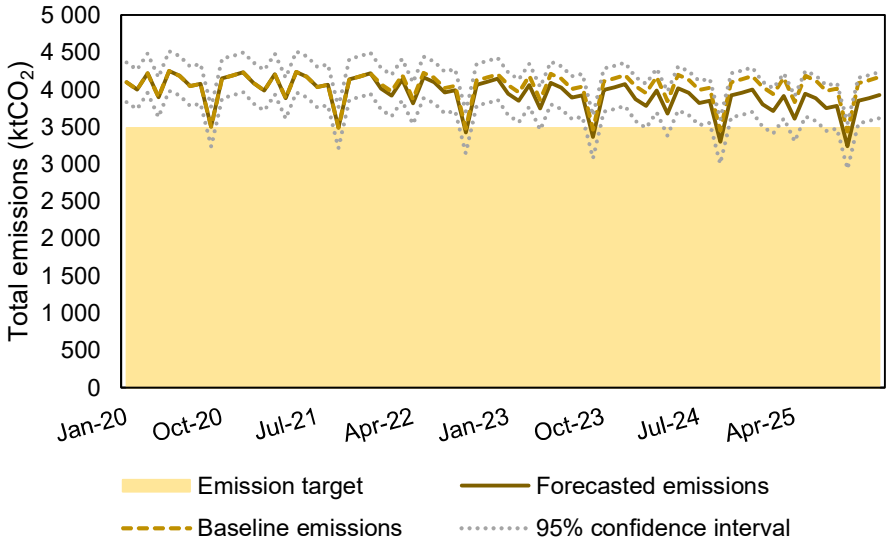


Figure 4-4: Total forecasted Scope 1 emissions.

After the mitigation measures were implemented in the developed forecasting model, it was determined that the facility is likely to reduce its Scope 1 emissions by 9% by 2025. However, this is still below its target of 18.5% as indicated in Figure 4-4. Thus, with the implementation of the additional mitigation measures (not evaluated in this study), the facility is likely to meet its emissions targets as specified.

It is assumed that the facility's targets were developed using detailed evaluations to ensure that the target is achievable. A similar conclusion, that targets will be met, is obtained by using the forecasting models, showing that the models can be used to forecast the emissions of a facility.

4.3 Limitations and recommendations

The emissions forecasting model developed in this dissertation only focusses on predicting Scope 1 emissions since they are currently the only scope that is subjected to carbon tax during the first phase. During the second phase, Scope 2 emissions are likely to become liable to carbon tax either directly or indirectly; thus, alternative models will be required to forecast Scope 2 emissions for facilities.

It is further recommended to develop a model that can forecast the financial impacts of exceeding the allocated carbon budget.

Since more companies are starting to report their Scope 3 emissions as well, an additional forecasting model can be developed to predict these emissions. Additionally, a model can be developed that can predict a company's entire carbon footprint.

The developed forecasting model only forecasts emissions for industrial facilities in South Africa. It is recommended that the model be extended to include additional sectors, such as the mining sector, as well.

For this study, a simplified forecasting model was developed using Excel and validated using the programming language, Python. However, to improve the accuracy of the forecasted emissions, advanced machine learning techniques are recommended.

4.4 Meeting the required objectives

As mentioned in Chapter 1, the developed forecasting model should be able to:

- Identify potential emission reduction opportunities,
- Determine the effect of different mitigation measures on the company's total GHG emissions, and
- Determine whether the 2025 emission target will be met with the mitigation measures in place.

With the implementation of a 10% reduction in material input over the 2021 to 2025 period, reductions of 12%, 19% and 5% in the fuel combustion, process, and fugitive emissions respectively were observed. Since the reduction in fuel combustion and fugitive emissions are still below the target of 18.5%, further reduction opportunities are available for these emissions.

A 18.5% reduction in the facility's overall Scope 1 emissions was predicted for 2025 when the only mitigation measure in place was a 10% reduction in the material input. Thus, additional mitigation measures are required to ensure that the facility will meet its 2025 emission target.

4.5 Closure

The developed forecasting model was able to reasonably predict the facility's emissions up until 2025. For the total emissions, the model indicated a forecasting accuracy of 5.61% and 4.94% when using the MAPE and %MAE as error statistics. By implementing the proposed mitigation measure into the forecasting model, a 9% reduction in the facility's carbon emissions is predicted, which is below the 18.5% target. Thus, additional mitigation measures are required to ensure the facility meet its 2025 emissions target.

In addition to using long-term forecasts to determine if a facility will meet its emission targets, short-term forecasting can be used for setting up carbon budgets. During the first carbon tax phase, companies can receive an additional 5% allowance on their carbon tax liability when participating in the carbon budget system. For six South African industries, this amounts to a total of about R 2.3-billion in avoided carbon tax liability.

Emissions forecasting plays an important role in both mitigation and financial planning for industries. This is not only important for an operational facility, but also provides industries with information on probable future scenarios. Thereby, assisting facilities and industries to adapt to the effects and alleviate potential impacts of climate change.

BIBLIOGRAPHY

[1] United Nations Framework Convention on Climate Change, "Paris Agreement," *21st Conference of the Parties,* p. 25, 2015.

[2] Republic of South Africa, "South Africa's Low-Emission Development Strategy 2050," 2020.

[3] RSA Department of Environmental Affairs and Tourism, "A National Climate Change Response Strategy," pp. 1-48, 2004.

[4] S. Kreft, D. Ekstein, and I. Melchior, "Global Climate Risk Index 2017," 2017.

[5] Republic of South Africa, "Draft 7th National Greenhouse Gas Inventory Report for the Republic of South Africa for Public Comment," 2020.

[6] E. Graham, "Global Electricity Revue 2021 South Africa," Ember, Accessed: 23 June 2021. [Online]. Available: https://ember-climate.org/wp-content/uploads/2021/03/Global-Electricity-Review-2021-South-Africa.pdf

[7] United Nations Framework Convention on Climate Change, "Status of Ratification of the Convention," Accessed: 26 June 2021. [Online]. Available: https://unfccc.int/process-and-meetings/the-convention/status-of-ratification/status-of-ratification-of-the-convention.

[8] United Nations Framework Convention on Climate Change, "United Nations Framework Convention on Climate Change," 1992.

[9] United Nations Framework Convention on Climate Change, "Kyoto Protocol," 1998.

[10] United Nations Framework Convention on Climate Change, "The Kyoto Protocol - Status of Ratification," Accessed: 26 June 2021. [Online]. Available: https://unfccc.int/process/the-kyoto-protocol/status-of-ratification.

[11] United Nations Framework Convention on Climate Change, "What is the Kyoto Protocol?," Accessed: 26 June 2021, doi: https://unfccc.int/kyoto_protocol.

[12] RSA Department of Environmental Affairs, "National Climate Change Response White Paper," 2010.

[13] Republic of South Africa, "South Africa's Intended Nationally Determined Contribution," 2015.

[14] Republic of South Africa, "Proposed Updated Nationally Determined Contribution," 2021.

[15] RSA Department of Environment, Forestry and Fisheries, "National Climate Change Adaptation Strategy," 2019.

[16] RSA Department of Environmental Affairs, "Climate Change Bill," *Gov. Gaz.,* vol. 636, no. 41689, 2018.

[17] RSA National Treasury, "Carbon Tax Act of 2019," *Gov. Gaz.,* vol. 647, no. 42483, 2019.

[18] RSA National Treasury, "Publication of the 2019 Carbon Tax Act," ed. South Africa, 2019.

[19] RSA National Treasury, "Gazetting of Trade Exposure and Greenhouse Gas Benchmark Regulations and Renewable Energy Premium Notice in terms of the Carbon Tax Act," ed. South Africa, 2020.

[20] J. Moss and R. Kath, "Historical Emissions and the Carbon Budget," *J. Appl. Philos.,* Journal Article vol. 36, no. 2, pp. 268-289, 2019. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pif&AN=PHL2387834&site=eds-live.

[21] Intergovernmental Panel on Climate Change, "Special Report: Global Warming of 1.5 °C," 2018.

[22] World Resources Institute, "The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard," 2015. [Online]. Available: https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf.

[23] Department of Environment, Forestry and Fisheries,, "Draft Methodological Guidelines ofr Quantification of Greenhouse Gas Emissions," *Gov. Gaz.,* vol. 43962, no. 1283, 2020. [Online]. Available: https://www.gov.za/sites/default/files/gcis_document/202012/43962gon1283.pdf.

[24] Intergovernmental Panel on Climate Change, "2006 IPCC Guidelines for National Greenhouse Gas Inventories," 2006.

[25] S. Fawzy, A. I. Osman, J. Doran, and D. W. Rooney, "Strategies for mitigation of climate change: a review," *Environ. Chem. Lett.,* vol. 18, no. 6, pp. 2069-2094, 2020, doi: 10.1007/s10311-020-01059-w.

[26] REN21, "Renewables 2019 - global status report," 2019. Accessed: 27 June, 2021. [Online]. Available: https://www.ren21.net/wp-content/uploads/2019/05/gsr_2019_full_report_en.pdf

[27] N. Victor, C. Nichols, and C. Zelek, "The U.S. power sector decarbonization: Investigating technology options with MARKAL nine-region model," *Energy Econ.,* Article vol. 73, pp. 410-425, 2018, doi: 10.1016/j.eneco.2018.03.021.

[28] Z. A. Wendling, "Bridges beyond renewable energy: Decarbonizing the global electricity sector under uncertainty," *Energy Res. Soc. Sci.,* vol. 48, pp. 235-245, 2019, doi: 10.1016/j.erss.2018.09.020.

[29] G. Pleßmann and P. Blechinger, "How to meet EU GHG emission reduction targets? A model based decarbonization pathway for Europe's electricity supply system until 2050," *Energy Strategy Rev.,* vol. 15, pp. 19-32, 2017, doi: 10.1016/j.esr.2016.11.003.

[30] A. C. Lin, "Carbon Dioxide Removal after Paris," *Ecol. Law Q.,* vol. 45, no. 3, pp. 533-582, 2018. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edshol&AN=edshol.hein.journals.eclawq45.31&site=eds-live.

[31] J. C. M. Pires, "Negative emissions technologies: A complementary solution for climate change mitigation," *Sci. Total Environ.,* vol. 672, pp. 502-514, 2019, doi: 10.1016/j.scitotenv.2019.04.004.

[32] Royal Society, "Greenhouse gas removal," 2018. [Online]. Available: https://royalsociety.org/-/media /policy/projects/green house -gas-removal/royal-society-greenhouse-gas-removal-report-2018.pdf.

[33] D. Matovic, "Biochar as a viable carbon sequestration option: Global and Canadian perspective," *Energy,* vol. 36, no. 4, pp. 2011-2016, 2011, doi: 10.1016/j.energy.2010.09.031.

[34] B. A. Oni, O. Oziegbe, and O. O. Olawole, "Significance of biochar application to the environment and economy," *Ann. Agric. Sci.,* vol. 64, no. 2, pp. 222-236, 2019, doi: 10.1016/j.aoas.2019.12.006.

[35] A. I. Osman *et al.*, "Upcycling brewer's spent grain waste into activated carbon and carbon nanotubes for energy and other applications via two-stage activation," *J. Chem. Technol. Biotechnol.,* vol. 95, no. 1, pp. 183-195, 2020, doi: 10.1002/jctb.6220.

[36] A. I. Osman, D. W. Rooney, C. Farrell, J. Harrison, and A. H. Al-Muhtaseb, "The production and application of carbon nanomaterials from high alkali silicate herbaceous biomass," (in English), *Sci. Rep.,* vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-59481-7.

[37] A. I. Osman *et al.*, "Production and characterisation of activated carbon and carbon nanotubes from potato peel waste and their application in heavy metal removal," *Environ. Sci. Pollut. Res.,* vol. 26, no. 36, p. 37228, 2019, doi: 10.1007/s11356-019-06594-w.

[38] H. Chen, A. I. Osman, C. Mangwandi, and D. Rooney, "Upcycling food waste digestate for energy and heavy metal remediation applications," *Resour. Conserv. Recycl.,* vol. 3, 2019, doi: 10.1016/j.rcrx.2019.100015.

[39] K. Nishitani and K. Kokubu, "Why Does the Reduction of Greenhouse Gas Emissions Enhance Firm Value? The Case of Japanese Manufacturing Firms," *Bus. Strategy Environ.,* vol. 21, no. 8, pp. 517-529, 2012, doi: 10.1002/bse.734.

[40] J. Willmott Cort and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.,* vol. 30, no. 1, pp. 79-82, 2005. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.24869236&site=eds-live.

[41] X. Zhao and D. Du, "Forecasting carbon dioxide emissions," *J. Environ. Manage.,* vol. 160, pp. 39-44, 2015, doi: 10.1016/j.jenvman.2015.06.002.

[42] M. Auffhammer and R. T. Carson, "Forecasting the path of China's CO2 emissions using province-level information," *J. Environ. Econ. Manage.,* vol. 55, no. 3, pp. 229-247, 2008, doi: 10.1016/j.jeem.2007.10.002.

[43] L. Abdullah and H. M. Pauzi, "Methods in forecasting carbon dioxide emissions," *J. Teknol.,* 2015, doi: 10.11113/jt.v75.2603.

[44] D. Radojevic, V. Pocajt, I. Popovic, A. Peric-Grujic, and M. Ristic, "Forecasting of Greenhouse Gas Emissions in Serbia Using Artificial Neural Networks," *Energy Source Part A,* vol. 35, no. 8, pp. 773-740, 2013.

[45] P. Liu, G. Zhang, and S. Cheng, "Carbon Emissions Modeling of China Using Neural Network," presented at the Fifth International Joint Conference on Computational Science and Optimization, 2012.

[46] W. K. Yap and V. Karri, "Emissions Predictive Modelling by Investigating Various Neural Network Models," *Expert Syst. Appl.,* vol. 39, no. 3, pp. 2421 - 2426, 2012.

[47] M. A. Behrang, E. Assareh, M. R. Assair, and G. A, "Using Bees Algorithm and Artificial Neural Network to Forecast World Carbon Dioxide Emission," *Energy Source Part A,* vol. 33, no. 19, pp. 1747 - 1759, 2011.

[48] S. Li, R. Zhou, and X. Ma, "The Forecast of CO2 Emissions in China based on RBF Neural Networks," presented at the Industrial and Information Systems (IIS), 2nd International Conference, 2010.

[49] A. Sözen, Z. Gülseven, and E. Arcaklioğlu, "Forecasting based on Sectoral Energy Consumption of GHGs in Turkey and Mitigation Policies," *Energy Policy,* vol. 35, no. 12, pp. 6491 - 6505, 2007.

[50] H. T. Pao, H. C. Fu, and C. L. Tseng, "Forecasting of CO2 emissions, energy consumption and economic growth in China using an improved grey model," *Energy,* vol. 40, no. 1, pp. 400 - 409, 2012, doi: 10.1016/j.energy.2012.01.037.

[51] H. T. Pao and C. M. Tsai, "Modelling and Forecasting the CO_2 Emissions, Energy Consumption, and Economic Growth in Brazil," *Energy,* vol. 36, no. 5, pp. 2450 - 2458, 2011, doi: 10.1016/j.energy.2011.01.032.

[52] C.-S. Lin, F.-M. Liou, and C.-P. Huang, "Grey forecasting model for CO2 emissions: A Taiwan study," *Appl. Energy,* vol. 88, no. 11, pp. 3816-3820, 2011, doi: 10.1016/j.apenergy.2011.05.013.

[53] I. J. Lu, C. Lewis, and S. J. Lin, "The forecast of motor vehicle, energy demand and CO2 emission from Taiwan's road transportation sector," *Energy Policy,* vol. 37, no. 8, pp. 2952-2961, 2009, doi: 10.1016/j.enpol.2009.03.039.

[54] Y. Y. Feng, S. Q. Chen, and L. X. Zhang, "System Dynamics Modelling for Urban Energy Consumption and CO2 Emissions: A Case Study of Beijing, China," *Ecol. Modell.,* vol. 252, pp. 44 - 52, 2013.

[55] L. Hui-ya, W. Zheng, M. Xiao-zhe, and Z. Yong-bin, "Prediction on Net Carbon Emissions of Yunnan under Emissions and Forest Carbon Sinks," presented at the 2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, 2011.

[56] R. Huang and Z. Wang, "Prediction on Shanghai's Energy Consumption Trend and Carbon Emission Peak," presented at the Measuring Technology and Mechatronics Automation (ICMTMA), Third International Conference, 2011.

[57] L. Wu, S. Liu, D. Liu, Z. Fang, and H. Xu, "Modelling and forecasting CO2 emissions in the BRICS (Brazil, Russia, India, China, and South Africa) countries using novel multi-variable grey model," *Energy,* vol. 79, pp. 489-495, 2015, doi: 10.1016/j.energy.2014.11.052.

[58] X. Ma, P. Jiang, and Q. Jiang, "Research and application of association rule algorithm and an optimized grey model in carbon emissions forecasting," *Technol. Forecast. Soc. Change,* vol. 158, 2020, doi: 10.1016/j.techfore.2020.120159.

[59] Q. Wang, S. Li, and Z. Pisarenko, "Modeling carbon emission trajectory of China, US and India," *J. Clean. Prod.,* vol. 258, 2020, doi: 10.1016/j.jclepro.2020.120723.

[60] D. Nong, "Development of the electricity-environmental policy CGE model (GTAP-E-PowerS): A case study of carbon tax in South Africa," *Energy Policy,* vol. 140, 2020, doi: 10.1016/j.enpol.2020.111375.

[61] A. G. S. Gous, "A practical approach to quantify carbon tax in South Africa," PhD, Department of Mechanical Engineering, North-West University, Potchefstroom Campus, 2018.

[62] R. J. Hyndman and G. Athanasopoulos, *Forecasting : principles and practice*. OTexts, 2014.

[63] B. S. S. Onggo and J. Hill, "Data identification and data collection methods in simulation: a case study at ORH Ltd," *J. Simul.,* vol. 8, no. 3, pp. 195-205, 2014, doi: 10.1057/jos.2013.28.

[64] RSA Department of Environmental Affairs, "National Greenhouse Gas Emission Reporting Regulations," *Gov. Gaz.,* vol. 622, no. 40762, pp. 4-24, 2017.

[65] S. Ahmed, J. Ferzund, A. Rehman, M. Ali, M. A. Sarwar, and A. Mehmood, "Modern Data Formats for Big Bioinformatics Data Analytics," *Int. J. Adv. Comput. Sci. Appl.,* vol. 8, no. 4, pp. 366-377, 2017.

[66] W. Booysen, "Measurement and verification of industrial DSM projects," PhD, Electrical Engineering, North-West University, Potchefstroom Campus, 2014. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat01185a&AN=nwu.b1988792&site=eds-live

[67] C. Van Aarde, "A general approach to develop and assess models estimating coal energy content," Masters, Mechanical Engineering, North-West University, Potchesfstroom Campus, 2018.

[68] R. Mahanti, *Data quality: Dimensions, measurement, strategy, management and governance*. Milwaukee: ASQ Quality Press, 2019.

[69] N. Robertson and T. Perera, "Automated data collection for simulation?," *Simul. Pract. Theory,* vol. 9, pp. 349-364, 2002.

[70] J. E. Olson, *Data Quality : The Accuracy Dimension* (The Morgan Kaufmann Series in Data Management Systems). San Francisco, Calif: Morgan Kaufmann (in English), 2003.

[71] Y. S. Lincoln and E. G. Guba, *Naturalistic inquiry*. Sage Publications, 1985.

[72] R. Xu, "Modeling Process: Predictive modeling in insurance," *Forecasting Future,* vol. 11, no. 8, 2013.

[73] W. D. Kelton, R. P. Sadowski, and D. T. Sturrock, *Simulation with Arena*, 4th ed. McGraw-Hill Higher Education, 2007.

[74] K. N. H. P. Liyanage, "Methodology for input data modelling in the simulation of manufacturing systems," PhD, School of Engineering, Sheffield Hallam University,

1999. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ddu&AN=449790038FB7243D&site=eds-live

[75] L. A. Oakshott, *Business Modelling and Simulation*. UK: Pitman Publishers, 1997.

[76] S. Robinson and V. Bhatia, "Secrets of Successful Simulation Projects," presented at the Proceedings of the 1995 Winter Simulation Conference, Arlington, USA, 1995.

[77] W. Hamer, W. Booysen, and E. H. Mathews, "A practical approach to managing uncertainty in the measurement and verification of energy efficiency savings," *S. Afr. J. Ind. Eng.,* vol. 28, no. 3, pp. 128-146, 2017.

[78] K. Hyun, "The prevention and handling of the missing data," *Korean J. Anesthesiol.,* vol. 64, no. 5, pp. 402-406, 2013, doi: 10.4097/kjae.2013.64.5.402.

[79] J. W. Graham, "Missing Data Analysis: Making It Work in the Real World," *Annu. Rev. Psychol.,* vol. 60, pp. 549-576, 2009. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsovi&AN=edsovi.00000689.200960000.00021&site=eds-live.

[80] P. More. "An introduction to missing value imputation in univariate time series." https://medium.com/@poojamore282/an-introduction-to-missing-value-imputation-in-univariate-time-series-7739a34e87e3 (accessed 23 August, 2021).

[81] J. Singh. "Missing data in Time Series." https://jagan-singhh.medium.com/missing-data-in-time-series-5dcf19b0f40f (accessed 8 August, 2021).

[82] R. Sheposh, "Interpolation (mathematics)," ed: Salem Press, 2019.

[83] N. J. van den Berg *et al.*, "Implications of various effort-sharing approaches for national carbon budgets and emission pathways," *Clim. Change,* vol. 162, no. 4, pp. 1805-1822, 2020, doi: 10.1007/s10584-019-02368-y.

[84] J. Ha, S. Seok, and J.-S. Lee, "Robust outlier detection using the instability factor," *Knowl. Based Syst.,* vol. 63, pp. 15-23, 2014, doi: 10.1016/j.knosys.2014.03.001.

[85] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.,* vol. 41, 2009, doi: 10.1145/1541880.1541882.

[86] X. Zhao, W. Cui, Y. Wu, H. Zhang, H. Qu, and D. Zhang, "Oui! Outlier Interpretation on Multi-dimensional Data via Visual Analytics," *Comput. Graph. Forum,* vol. 38, no. 3, pp. 213-224, 2019, doi: 10.1111/cgf.13683.

[87] H. Jiawei, P. Jian, and K. Micheline, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[88] S. Santoyo. "A Brief Overview of Outlier Detection Techniques: What are outliers and how to deal with them?" https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561 (accessed 14 July, 2021).

[89] G. Barbato, E. M. Barini, G. Genta, and R. Levi, "Features and performance of some outlier detection methods," *J. Appl. Stat.,* vol. 38, no. 10, pp. 2133-2149, 2011, doi: 10.1080/02664763.2010.545119.

[90] R. Cornish, "Statistics : An introduction to sample size calculations," 2006.

[91] J. E. Hanke, A. G. Reitsch, and D. W. Wichern, *Business forecasting*, 7th ed. Prentice Hall, 2001.

[92] Q. Cui, M. O. Ward, E. A. Rundensteiner, and J. Yang, "Measuring Data Abstraction Quality in Multiresolution Visualizations," *IEEE Trans. Vis. Comput. Graph.,* vol. 12, no. 5, pp. 709-716, 2006, doi: 10.1109/TVCG.2006.161.

[93] L. A. Botes, "Objective evaluation of industrial energy efficiency models for the RSA Section 12L tax incentive," Masters, Mechanical Engineering, North-West University, Potchefstroom Campus, 2018. [Online]. Available: https://nwulib.nwu.ac.za/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat01185a&AN=nwu.b2212834&site=eds-livehttp://orcid.org/0000-0003-3194-7224http://hdl.handle.net/10394/30627

[94] A. Plummer. "Different Types of Time Series Decomposition." https://towardsdatascience.com/different-types-of-time-series-decomposition-396c09f92693 (accessed 15 June, 2021).

[95] K. H. Chan, C. H. Jack, and J. K. Ord, "A Note on Trend Removal Methods: The Case of Polynomial Regression versus Variate Differencing," *Econometrica,* vol. 45, no. 3, pp. 737-744, 1977, doi: 10.2307/1911686.

[96] W. Kenton. "Detrend." Investopedia. https://www.investopedia.com/terms/d/detrend.asp (accessed 15 July, 2021).

[97] D. Bourg, *Excel Scientific and Engineering Cookbook*. O'Reilly Media, Inc, 2006.

[98] J. Frankenfield. "Artificial Neural Network (ANN)." https://www.investopedia.com/terms/a/artificial-neural-networks-ann.asp (accessed 15 July, 2021).

[99] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.,* vol. 49, no. 11, pp. 1225-1231, 1996, doi: 10.1016/S0895-4356(96)00002-9.

[100] W. Zhou and D. Zhang, "An Improved Metabolism Grey Model for Predicting Small Samples with a Singular Datum and Its Application to Sulfur Dioxide Emissions in China," *Discrete Dyn. Nat. Soc.,* 2016, doi: 10.1155/2016/1045057.

[101] A. Meyler, G. Kenny, and T. Quinn, "Forecasting Irish inflation using ARIMA models," 1999.

[102] D. Weedmark. "The Advantages & Disadvantages of a Multiple Regression Model." https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html (accessed 17 July, 2021).

[103] J. Gordon. "Autoregression - Definition." https://thebusinessprofessor.com/research-analysis-decision-science/autoregression-definition (accessed 17 July, 2021).

[104] J. Fernando. "Moving Average (MA)." https://www.investopedia.com/terms/m/movingaverage.asp (accessed 17 July, 2021).

[105] S. Glen. "Exponential Smoothing: Definition of Simple, Double and Triple." https://www.statisticshowto.com/exponential-smoothing/ (accessed 17 July, 2021).

[106] H. Arsham. "Time-Critical Decision Making for Business Administration." http://home.ubalt.edu/ntsbarsh/business-stat/stat-data/forecast.htm#rhomave (accessed 17 July 2021).

[107] H. Yang and J. F. O'Connell, "Short-term carbon emissions forecast for aviation industry in Shanghai," *J. Clean. Prod.,* Article vol. 275, 12/01/December 2020 2020, doi: 10.1016/j.jclepro.2020.122734.

[108] Inspection STC Working Group, "Guidelines for reporting uncertainty in measurement and verification," 2017.

[109] C. Turhan, K. E. Ekmen, G. G. Akkurt, T. Kazanasmaz, and I. E. Uygun, "Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation," (in English), *Energy and Buildings,* Article vol. 85, pp. 115-125, 01 / 01 / 2014, doi: 10.1016/j.enbuild.2014.09.026.

[110] O. De Jeses and M. T. Hagan, "Backpropagation through time for a general class of recurrent network," vol. 4, ed. Piscataway, NJ, USA, USA: IEEE, 2001, p. 2638.

[111] C. V. Gallagher, K. Leahy, P. O'Donovan, K. Bruton, and D. T. J. O'Sullivan, "Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0," (in English), *Energy and Buildings,* Article vol. 167, pp. 8-22, 05 / 15 / 2018, doi: 10.1016/j.enbuild.2018.02.023.

[112] C. V. Gallagher, K. Leahy, D. T. J. O'Sullivan, and K. Bruton, "The suitability of machine learning to minimise uncertainty in the measurement and verification of

energy savings," (in English), *Energy and Buildings,* Article vol. 158, pp. 647-655, 01 / 01 / 2018, doi: 10.1016/j.enbuild.2017.10.041.

[113] M. Mittal, B. Bora, S. Saxena, and A. M. Gaur, "Performance prediction of PV module using electrical equivalent model and artificial neural network," *Solar Energy,* Article vol. 176, pp. 104-117, 12/01/December 2018 2018, doi: 10.1016/j.solener.2018.10.018.

[114] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy,* Article vol. 123, pp. 168-178, 06/15/June 2014 2014, doi: 10.1016/j.apenergy.2014.02.057.

[115] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013.

[116] P. Flach, *Machine learning: The art and science of algorithms that make sense of data*. New York, NY: Cambridge University Press, 2012.

[117] D. R. Anderson, D. J. Sweeney, and T. A. Williams, *Modern business statistics with Microsoft Excel*. South-Western, 2003.

[118] "Mean Squared Error," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., ed. Boston, MA: Springer US, 2010, pp. 653-653.

[119] "Mean Absolute Error," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., ed. Boston, MA: Springer US, 2010, pp. 652-652.

[120] Republic of South Africa, "Draft Technical Guidelines for Validation and Verfication of Greenhouse Gas Emissions," 2020.

[121] World Resources Institute, "Mitigation Goal Standard," 2014.

[122] NIST/SEMATECH., *e-Handbook of Statistical Methods*. 2012.

[123] G. Shmueli and K. C. Lichtendahl, *Practical time series forecasting with R : a hands-on guide*, 2nd ed. Axelrod Schnall Publishers, 2016.

[124] M. Luiz and D. H. Graeme, *The SAGE Dictionary of Quantitative Management Research*. Los Angeles: SAGE Publications Ltd (in English), 2011.

[125] D. H. Graeme and S. Nick, *The Multivariate Social Scientist : Introductory Statistics Using Generalized Linear Models*. London: SAGE Publications Ltd (in English), 1999.

[126] A. Kamer-Ainur and M. Marioara, "Errors and limitations associated with regression and correlation analysis," *Stat. Comput.,* vol. 709, 2007.

[127] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied statistics for the behavioral sciences*, 5th ed. Wadsworth Cengage Learning, 2003.

[128] S. J. Sheather, *A modern approach to regression with R* (Springer texts in statistics). Springer, 2009.

APPENDIX A - SELECTED FORECASTING MODELS

A.1 Exponential smoothing

One of the forecasting models chosen in Chapter 2 is the exponential smoothing model. The exponential smoothing model has been around since the late 1950s and is still one of the most popular forecasting models to use [62].

Data is forecasted by using weighted averages of past observation where the weights decay exponentially as the observations get older [62, 122].

Exponential smoothing models are known to generate quick, reliable forecasts which is advantageous and important for application in industry.

There are three types of exponential smoothing: single, double, and triple exponential smoothing [122, 123].

The single exponential smoothing model is a time series forecasting method for univariate data without trend and seasonality. Single exponential smoothing requires a single parameter, α, which is also known as the smoothing factor or smoothing coefficient [122, 123].

The α – parameter controls the rate at which the influence of observation decays exponentially. A large α value indicates that the forecasting model uses the most recent past observation to predict future values. A smaller α value indicates that historic data has a larger influence when predicting future data [122, 123].

Double exponential smoothing, also known as Holt's linear trend, is an extension to exponential smoothing and is suitable for predicting data with trends in the univariate time series. In addition to the α parameter the β smoothing parameter is introduced to control the decay of influence of change in the trend [123].

Triple exponential smoothing, also known as Holt-Winter's exponential smoothing, is suitable to predict a univariate time series with trend and seasonality. An additional smoothing coefficient, γ, is introduced which controls the influence on the seasonal component [122, 123].

Excel has a built-in function to forecast values for a specific future target date using exponential smoothing. This function is described in Equation (A-1).

FORECAST.ETS(target_date, values, timeline, [seasonality], [data_completion], [aggregation]) (A-1)

The arguments for the Excel function described above are as follows:

- *Target date*: The data point for which a value is predicted.
- *Values*: An array or range of the numeric data to be predicted.
- *Timeline:* An independent array or range of numeric data. The timeline requires a consistent step between data, which cannot be zero.
- *Seasonality:* An optional value for indicating the length of the seasonal pattern. When left blank, the seasonality is automatically detected.
- *Data completion:* An optional value for handling missing data. Missing values can either be calculated using interpolation (this is the default) or missing values are treated as zeros.
- *Aggregation:* An optional value for aggregating multiple values with the same time stamp. The default aggregation method is averaging values.

The forecasting function is based in the AAA version of the Exponential Smoothing Algorithm. Thus, the function is based on additive seasonality, additive trend, and additive error. Equation (A-2) can be used to calculate the forecasted emissions while Equations (A-3), (A-4) and (A-5) can be used to calculate the level, trend and seasonality respectively.

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + e_t \quad \text{(A-2)}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha e_t \quad \text{(A-3)}$$

$$b_t = b_{t-1} + \beta e_t \quad \text{(A-4)}$$

$$s_t = s_{t-m} + \gamma e_t \quad \text{(A-5)}$$

Where:

- y_t is the forecasted value
- l_t is the level
- b_t is the trend
- s_t is the seasonality
- α, β and γ are smoothing parameters with $0 < \alpha \leq 1$, $0 < \beta \leq 1$ and $\alpha + \gamma \leq 1$
- m is the seasonality period

A.2 Linear regression and multi-variable linear regression

Ordinary least squares (OLS) regression, also known as linear regression, is a general linear modelling technique that is used to model a single response variable. Linear regression can be applied to both single and multiple explanatory variables [124].

The relationship between the response variable (y variable) and the explanatory variable (x variable) is represented by using the line of best fit which is illustrated in Figure A-2 [125].

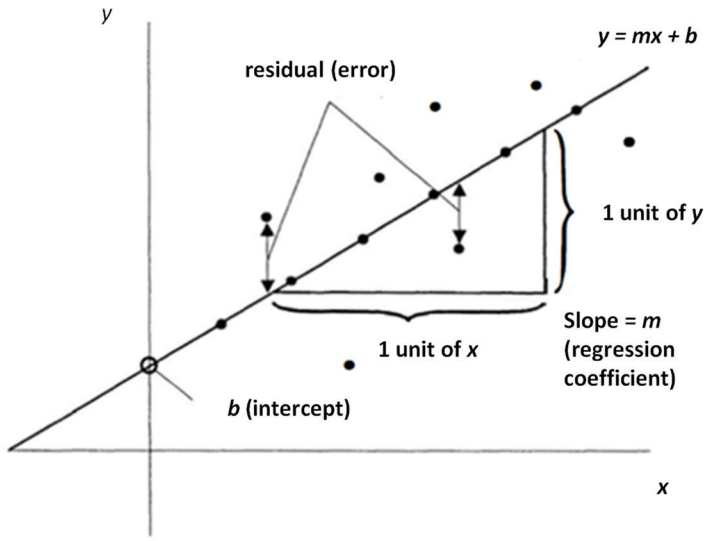


Figure A-1: Line of best fit.

If the relationship between the two variables is linear, it can be described by a straight-line (Equation (A-6)).

$$y = mx + b \tag{A-6}$$

Where:

- y is the response variable
- m is the regression coefficient or slope
- x is the explanatory variable
- b is the intercept

A linear regression with a perfect linear relationship is illustrated in Figure A-2 [125].

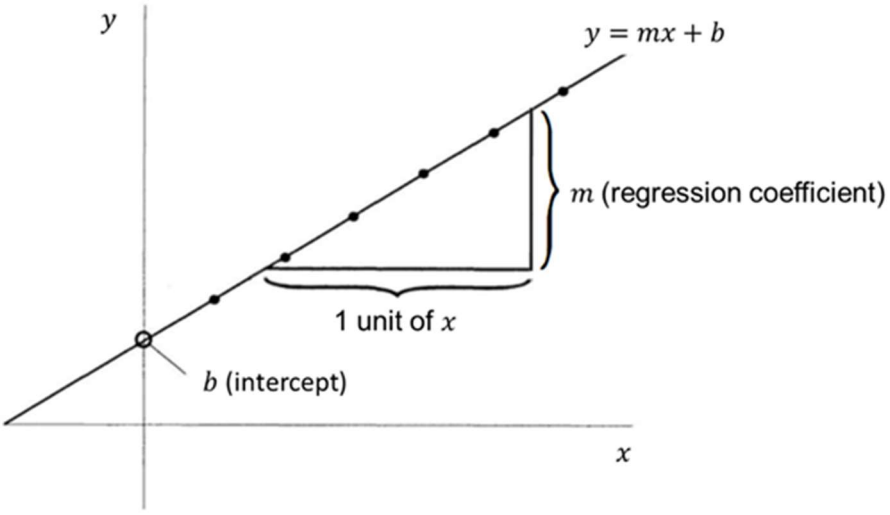


Figure A-2: Linear regression model with a perfect linear relationship.

OLS can be extended to include multiple explanatory variables by adding additional variables to the equation as illustrated in Equation (A-7).

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \cdots + m_n + x_n + b \quad \text{(A-7)}$$

Excel has a built-in function that can be used to determine the straight-line equation for a response variable with a single or multiple explanatory variables. The formula is illustrated in Equation (A-8).

$$\text{LINEST(known_ys,[known_xs],[const],[stats])} \quad \text{(A-8)}$$

The arguments for the function described above are as follows:

- *Known y-values:* A set of y-values where the $y = mx + b$ relationship is known.
- *Known x-values:* An optional variable where the x-values in the $y = mx + b$ relationship is already known.
- *Constant:* A logical value. To calculate the constant, b, as normal the argument is omitted or set as TRUE. If b is equal to zero, the argument is set as FALSE.
- *Statistics:* A logical value. When set to TRUE, additional regression statistics are returned. When set to FALSE or if the argument is omitted, the m-coefficients and the constant, b, are returned.

The output of the LINEST function is a matrix that includes the regressor, residual error term, and additional statistics. The output matrix is illustrated in Table A-1.

Table A-1: Output matrix of the LINEST function.

x_4	x_3	x_2	x_1	C
SE (x_4)	SE (x_3)	SE (x_2)	SE (x_1)	SE (C)
R^2	$RMSE$	#N/A	#N/A	#N/A
F	ν_2	#N/A	#N/A	#N/A
SS_{reg}	SS_{res}	#N/A	#N/A	#N/A

The first row of the output matrix contains the regressors for each independent variable and the required residual error term (C).

It is necessary to evaluate the regressor(s) of the model to see if the resulting model behaves as expected. If all the regressors are equal to zero, the output is completely dependent on the residual error value (C). Regressor evaluation is also indicative of the influence each regressor has on the output of the model (positive or negative).

The second row of the matrix contains the standard error of each regressor and indicates the error amount that the associated regressors add to the standard error of the model [67].

Additional statistical information given by the output matrix involves the predictability and accuracy of the data which is given in Table A-2 [67, 126].

Table A-2: Statistical parameters from output matrix.

Parameter	Formula	Application
Coefficient of determination (R^2)	$\frac{\sum_{k=1}^{p}(\widehat{y_k}-\hat{y})^2}{\sum_{k=1}^{p}(y_k-\hat{y})^2}$	The coefficient of determinations is in the range $0 \leq R^2 \leq 1$. A R^2 close to 1 is indicative of a good model.
Root mean square error $(RMSE)$	$\sqrt{\frac{SS_{reg}}{\nu_2}}$	A smaller value indicates the predicted data points lie close the actual data points.
F-value (F)	$\frac{\frac{SS_{reg}}{\nu_1}}{\frac{SS_{res}}{\nu_2}}$	The regression is justified with an $\alpha -$ confidence level if F is greater than $F_{critical}$.
Degree of freedom for residuals (ν_2)	$P - n$	An equality instead of an approximation is required if the degree of freedom is zero.

Parameter	Formula	Application
Sum of squares of regressor errors (SS_{reg})	$\sum_{k=1}^{P}(\widehat{y_k} - \hat{y})^2$	Indicative of the variability of regressed data.
Sum of squares of residual errors (SS_{res})	$\sum_{k=1}^{P}(\widehat{y_k} - y_k)^2$	Indicative of the variability of the data.

A.3 Python models

Machine learning can be used as an alternative to Excel when developing forecasting models. In this study, the models developed in Excel, were validated by using the programming language Python.

Statsmodels is a Python module that provides classes and functions which can be used for estimating various statistical models, as well as for conducting statistical tests, and statistical data exploration. One of the classes in this module is the *ETSModel*. Unlike the Excel function which is based on additive trend, additive seasonality, and additive error, the Python class can have either additive or multiplicative components.

To import *ETSModels* to Python, Equation (A-9) can be used.

```
from statsmodels.tsa.exponential_smoothing.ets import ETSModel
```
(A-9)

Equation (A-10) describes the arguments required to develop the ETS model as well as their default values, where applicable.

$$\begin{aligned}&\text{ETSModel}(\text{endog}, \text{error} = \text{'add'}, \text{trend} = \text{None}, \text{damped_trend} = \text{False}, \text{seasonal} = \\ &\text{None}, \text{seasonal_periods} = \text{None}, \text{initialization_method} = \\ &\text{'estimated'}, \text{initial_level} = \text{None}, \text{initial_trend} = \text{None}, \text{initial_seasonal} - \\ &\text{none}, \text{bounds} = \text{None}, \text{dates} = \text{None}, \text{freq} = \text{None}, \text{missing} = \text{'none'})\end{aligned} \quad \text{(A-10)}$$

Where:

- endog is an array consisting of the observed time-series process (y)
- error is the error component which can be additive ('add') or multiplicative ('mul')
- trend is the trend component which can be 'add', 'mul', or None
- damped_trend describes whether the trend component is damped or not
- seasonal is the seasonal component which can be 'add', 'mul', or None
- seasonal_periods which describes the number of periods in a complete seasonal cycle for seasonal models
- initialization_method is the method for initialisation of the state base model which can be 'estimate', 'heuristic' or 'known'
- initial_level is the initial component level which is required if the initialisation is 'known'
- initial_trend is the initial trend component which is required if the initialisation is 'known'
- initial_seasonal is the initial seasonal component which is required if the initialisation is 'known'
- bounds which is a dictionary with parameter names as keys and the respective bounds intervals as values

To validate the Excel model using Python, the following code can be used:

```
1 ETS_model = ETSModel(
2 Data,
3 error = 'add',
4 trend = 'add',
5 damped_trend = False,
6 seasonal_periods = 12,
7 )
8 Model_fit = ETS_model.fit()
```

(A-11)

In Equation (A-11), "Data" refers to the variable for which the model is developed. In order to compare the Python model to the Excel model, the trend, error, and seasonality components are taken as additive, and the trend is assumed to be undamped (which is the default). Based on the preliminary analysis done in Section 3.2.2, a seasonality of 12 is used. In the last line (line 8), the data is fitted to the ETS model.

As with the ETS model, machine learning can be used to develop regression models. *Sklearn* is a Python module which integrates classic machine learning algorithms into a scientific Python package. One of the classes available in this module is the *LinearRegression* class.

The *LinearRegression* class fits a linear model with the coefficients $w = (w_1, \ldots, w_p)$ to minimise the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

To import *LinearRegression* to Python, Equation (A-12) can be used.

```
from sklearn.linear_model import LinearRegression
```
(A-12)

Equation (A-13) describes the arguments required to develop the linear regression models as well as their default values, where applicable.

$$\text{LinearRegression}(*, \text{fit_intercept} = \text{True}, \text{normalize} = \text{'deprecated'}, \text{copy_X} = \text{True}, \text{n_jobs} = \text{None}, \text{positive} = \text{False}) \quad \text{(A-13)}$$

Where:

- fit_intercept can be either True or False. If True, the intercept for the model will be calculated. If False, no intercept will be used in the calculations
- normalize - when the fit_intercept, is set False, the parameter will be ignored. If the parameter is set to True, the regressors X will be normalised before regression by subtracting the mean and dividing by the 12-norm
- copy_X – if set to True, X will be copied otherwise it may be overwritten
- n_jobs indicate the number of jobs use for the computation
- positive – when set to True, the coefficients are forced to be positive

To validate the regression models in Excel, a multivariable linear regression and linear regression definitions were created in Python. These definitions are described by Equation (2-14) and Equation (2-15), respectively.

```
1 def MVRA(x1, x2, y):
2     X = np.vstack((x1, x2)).T
3     Y = y.to_numpy(np.float64)
4     model = LinearRegression().fit(X, Y)
5     intercept = model.intercept_
6     slope = model.coef_
7     return intercept, slope
```
(A-14)

```
1 def Linear_Regression(x, y):
2     X = (x.to_numpy(np.float64)).reshape((-1,1))
3     Y = y.to_numpy(np.float64)
4     model = LinearRegression().fit(X, Y)
5     intercept = model.intercept_
6     slope = model.coef_
7     return intercept, slope
```

(A-15)

The definition for multivariable linear (Equation (A-14))(2-15) has three inputs: the two independent variables (x_1 and x_2) and the dependent variable (y). The definition for the linear regression model has two inputs: the independent variable (x) and the dependent variable. In the second and third lines the input data is converted to the format required by the *LinearRegression* class. The data is fitted to the model in line 4 whereafter the intercept and slope is defined in line 5 and line 6, respectively. The definition then returns the intercept and slope of the regression model which can be used to predict the dependent variable.

APPENDIX B – MODEL DEVELOPMENT AND MODEL VALIDATION

B.1 Linear regression and multivariable linear regression model development

During the preliminary analysis step in Section 3.2.2, a similar pattern between all variables (with fugitive emissions being the exception) were observed. The relationship between the variables can further be illustrated using regression plots as seen in Figure B-1.

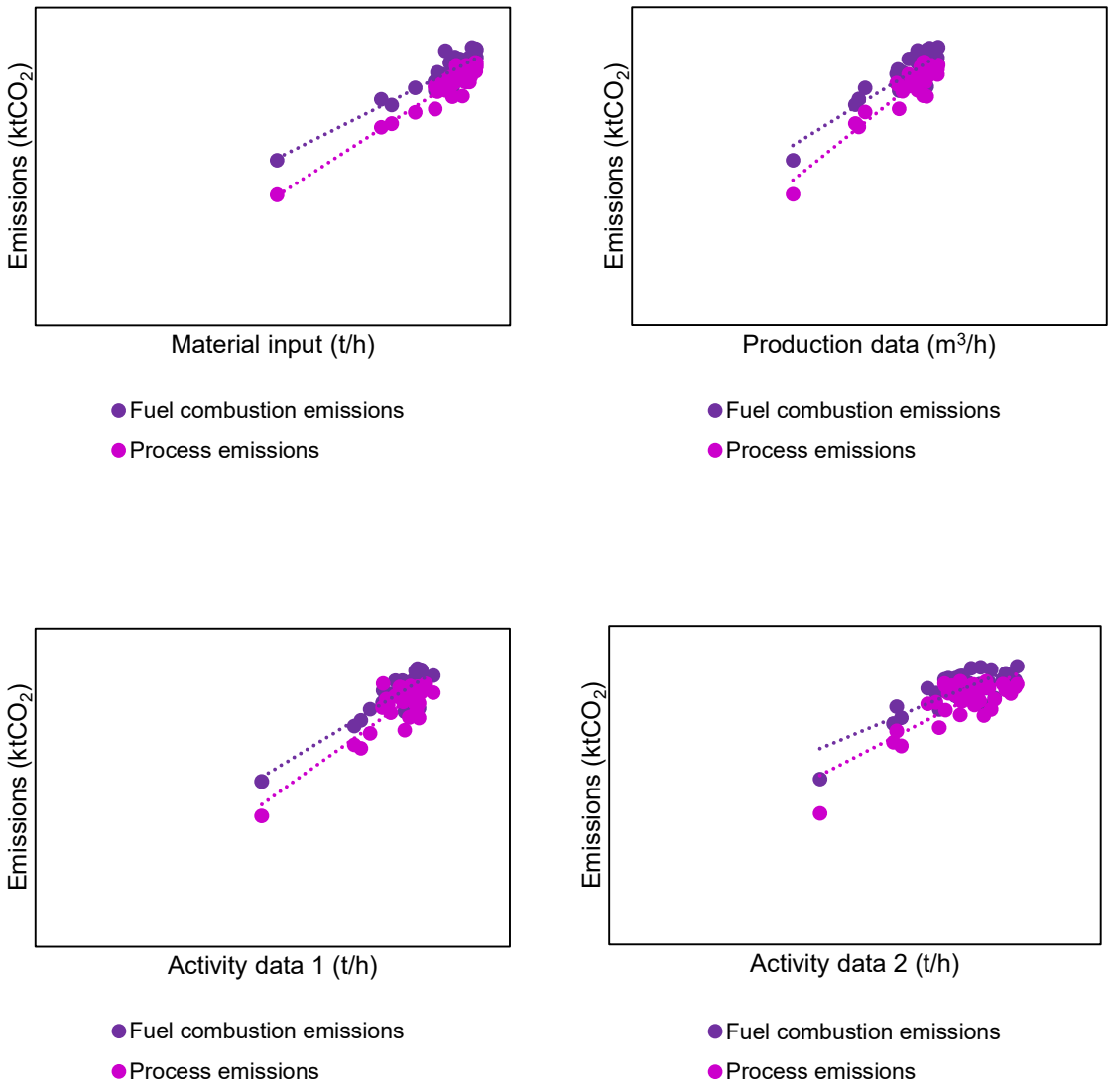


Figure B-1: Linear relationship between the different variables.

In Figure B-1, a linear relationship can be observed between the input data (material input and production data) and the emissions (fuel combustion and process emissions) as well as the activity data and the emissions.

In Appendix A, statistical parameters regarding the regression models were briefly discussed, while the regression models were developed in Chapter 2. The parameters for these models are given in Table B-1, Table B-2, and Table B-4 for the activity data, fuel combustion emissions, and process emissions respectively. However, it should be noted that the five underlying linear regression analysis assumptions need to be addressed before the statistical analysis of the variables can be considered valid. The five assumptions are [108]:

1. *Linear relationship:* Described a straight-line relationship between two variables.
2. *Multivariate normality:* Any linear combination of multiple normally distributed variables is also normally distributed. A quick test for multivariate normality is distribution graphs.
3. *No or little multicollinearity:* Multicollinearity can be defined as the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.
4. *No autocorrelation or independence:* The residuals are independent of each other; thus there is no correlation between consecutive terms in the time series data.
5. *Homoscedasticity:* The variance of the residual in a regression model is constant. Thus, the error term does not vary much as the value of the predictor variable change.

The activity data was determined using multivariable linear regression with the material input and production data as the independent variables. The regressions parameters for the activity data are summarised in Table B-1.

Table B-1: Multivariable linear regression parameters for the activity data.

Parameter	Activity data 1	Activity data 2
x_1 (Material input)	-256.59	157.86
x_2 (Production data)	778.52	784.28
C	55.34	610.60
SE (x_1)	305.82	482.80
SE (x_2)	403.39	636.82
SE (C)	127.24	200.87
R^2	0.68	0.67
$RMSE$	30.60	48.30
F	35.14	32.78
$F_{critical}$	3.28	3.28
ν_2	33	33
SS_{reg}	65 780	152 976
SS_{res}	30 891	76 990

From Table B-1 a R^2 of 0.68 and 0.67 can be observed for *activity data 1* and *activity data 2*, respectively. The R^2 value represents the proportion of variance for a dependent variable which is explained by an independent variable or variables in a regression model[11]. An R^2 value of 0.5 indicates that approximately half of the observed variation can be explained by the model's inputs. Furthermore, a weak relationship is implied by an R^2 between 0.1 and 0.5 while a strong relationship is implied when R^2 > 0.5 [127].

Thus, seeing as the R^2 for both *activity data 1* and *activity data 2* is more than 0.5, a strong relationship exists between the dependent and independent variables for the activity data. This indicates that the fist assumption of multivariable linear regression (that a linear relationship exists between the variables) is valid.

The variables ν_2, SS_{reg}, and SS_{res} are used to calculate the F-value. If the F-value is greater than $F_{critical}$, the regression is justified as mentioned in Appendix A. To determined $F_{critical}$, the Excel function given in Equation (B-1) can be used.

$$=\text{F.INV.RT}(\text{probablity},\text{deg_freedom1},\text{deg_freedom2}) \quad \text{(B-1)}$$

Where:

- probability is the significance level used
- deg_freedom1 is the numerator degrees of freedom
- deg_freedom2 is the denominator degrees of freedom

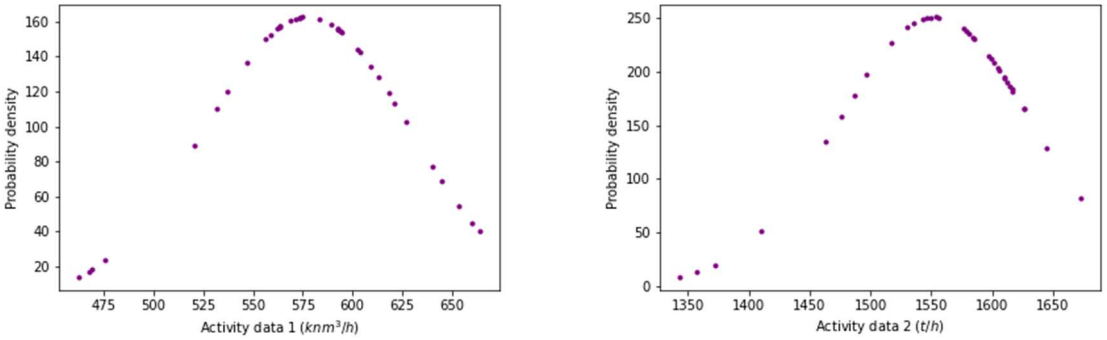
To calculate $F_{critical}$ a significance level of 0.05 was used[12].

For both *activity data 1* and *activity data 2*, $F > F_{critical}$ as given in Table B-1. Thus, the regression is justifiable. However, before the model can be considered valid, the remaining four assumptions of multivariable linear regression needs to be addressed.

To test for multivariate normality, distribution plots can be used. These plots are illustrated in Figure B-2.

[11] https://www.investopedia.com/terms/r/r-squared.asp

[12] https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/



a) *Activity data 1* distribution plot b) *Activity data 2* distribution plot

Figure B-2: A data distribution plots.

From Figure B-2, it is evident that *activity data 1* follows a normal distribution; however, the distribution plot for *activity data 2* is slightly skewed to the left. For cases where the assumption of multivariate normality is violated 1) verify that there are no extreme outliers present that result in the assumption being violated 2) apply a nonlinear transformation to the response variable[13].

The next assumption to be tested is the multicollinearity function. Multicollinearity can be calculated using the *Variance Inflation Factor* (VIF) [An introduction to statistical learning], which can be calculated as follows:

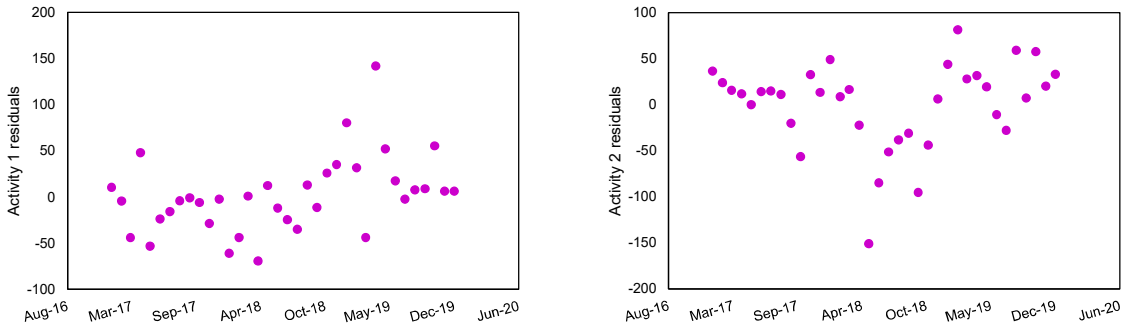
$$VIF = \frac{1}{1 - R^2} \tag{B-2}$$

Using the VIF test as described in Equation (B-2), values were calculated as 3.12 and 2.99 for *activity data 1* and *activity data 2*, respectively. For a VIF value of less than 4, no multicollinearity is assumed. For values higher than 10, significant multicollinearity is assumed [128]. The VIF value for both *activity data 1* and *activity data 2* is less than 4, thus no multicollinearity assumed.

The third assumption, no autocorrelation, can be tested by plotting the residuals over time[14]. If a noticeable pattern in the data can be observed, autocorrelation is present. The graphs used to determine if autocorrelation is present for the activity data is illustrated in Figure B-3.

[13] https://www.statology.org/multiple-linear-regression-assumptions/

[14] https://online.stat.psu.edu/stat501/lesson/14/14.3



a) *Activity data 1* residual plot b) *Activity data 2* residual plot

Figure B-3: Autocorrelation test for the activity data.

From Figure B-3, no noticeable pattern can be observed in the residual plots for either *activity data 1* or *activity data 2.* This it is likely that no autocorrelation is present between the variables. The following options can be considered when autocorrelation is present[15]:

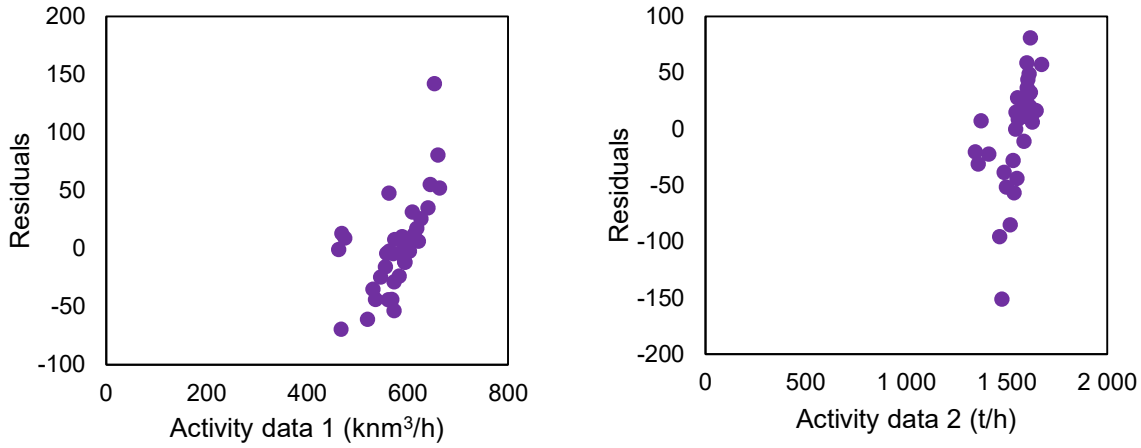
1. For positive correlation, the lags of the dependent and/or independent variable can be added to the data.
2. Fore negative correlation, ensure that none of the variables are over differenced.
3. For seasonal correlation, seasonal dummy variable can be added to the model.

A formal statistical test to determine if autocorrelation exists between variables is the Durbin-Watson test.

The last assumption of linear regression is homoscedasticity, which can be tested by plotting the residuals against the fitted data as illustrated in Figure B-4. If heteroscedasticity exist, a funnel-shaped plot will be observed[16].

[15] https://www.statology.org/multiple-linear-regression-assumptions/

[16] https://www.upgrad.com/blog/assumptions-of-linear-regression/



a) *Activity data 1* residual plot

b) *Activity data 2* residual plot

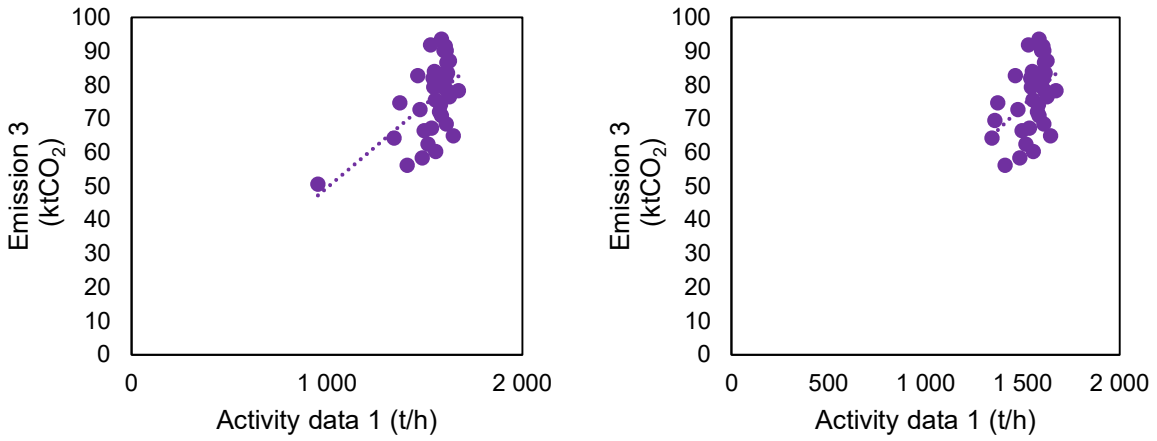
Figure B-4: Homoscedasticity test for the activity data.

From Figure B-4, a linear relationship can be observed between the fitted values and the residuals for both *activity data 1* and *activity data 2*. No funnel-shaped pattern can be observed in the residual plots; thus it is likely that heteroscedasticity is not present in the data. When the assumption of homoscedasticity is violated, the following options are available:

- Transform the response variable by taking the log, square root, or cube root of all the variables.
- Redefine the response variable to use a rate and not the raw value.
- Use weighted regression.

All five assumptions of linear regression are valid, thus the regression model used to determine *activity data 1* and *activity data 2* are valid.

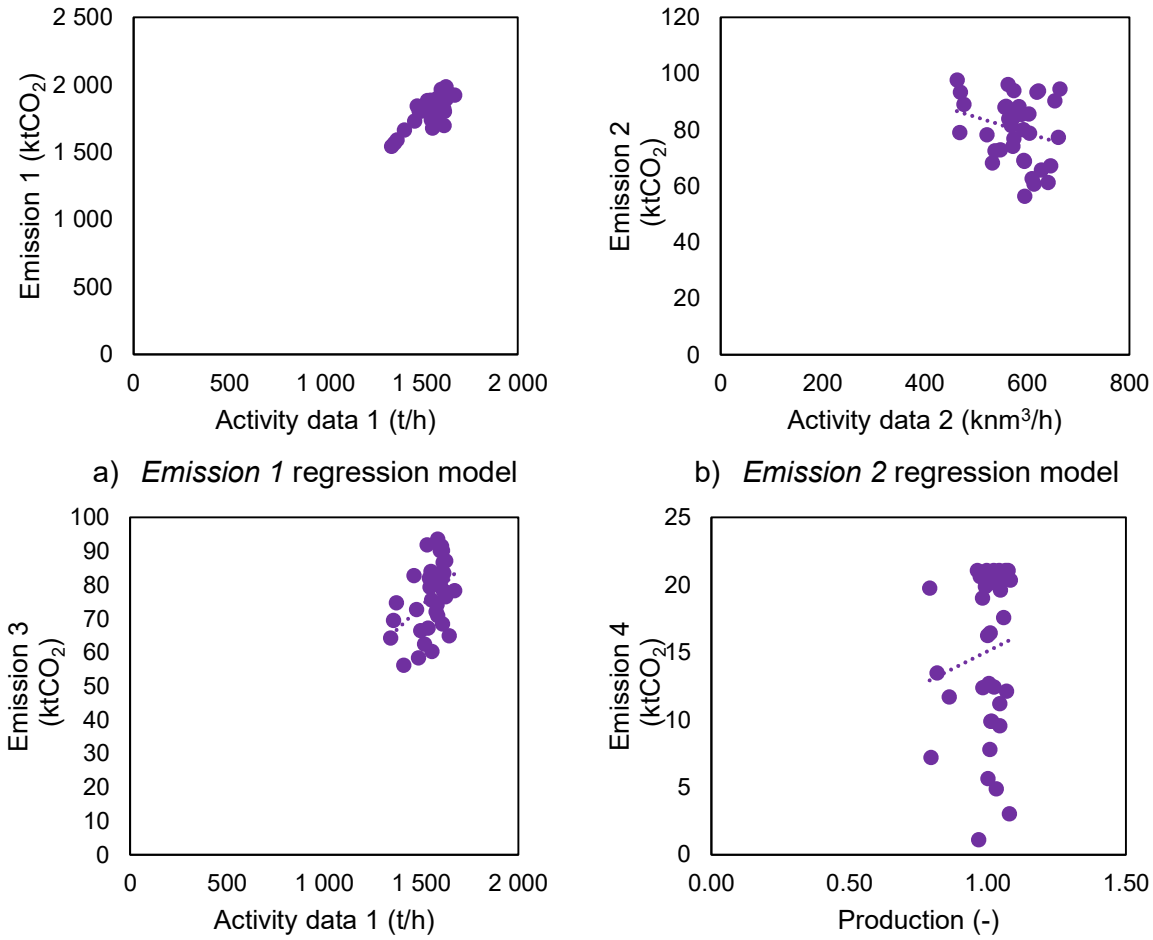
The fuel combustion emissions were calculated using linear regression with the activity data or production as the independent variable. To ensure that all outliers have been removed, an outlier analysis was done for each regression model. The regression models were visually inspected, and any outliers present in the data was subsequently removed as illustrated in Figure B-5.



a) Before outlier removal b) After outlier removal

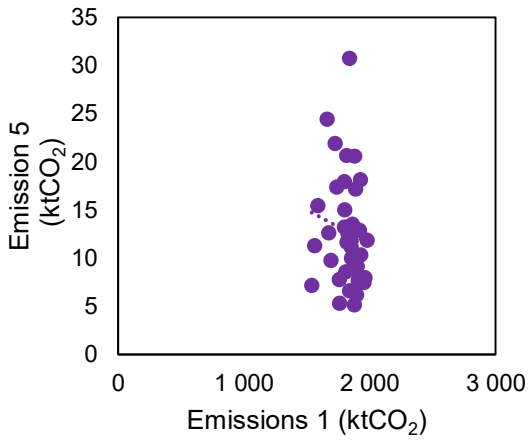
Figure B-5: Regression model outlier analysis.

The linear regression models for the fuel combustion emissions (with the outliers removed) are illustrated in Figure B-6**.**

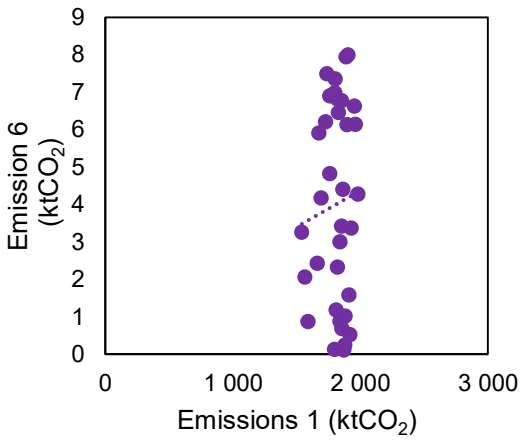


a) *Emission 1* regression model b) *Emission 2* regression model

c) *Emission 3* regression model d) *Emission 4* regression model



e) *Emission 5* regression model



f) *Emission 6* regression model

Figure B-6: Linear regression models for the fuel combustion emissions.

From Figure B-6, only the linear regression models for *emission 1* and *emission 3* show a correlation between the dependent and independent variables. The observation is further emphasised by the regression parameters for the fuel combustion emissions given in Table B-2.

Table B-2: Linear regression parameters for the fuel combustion emissions.

Parameter		Emission 1	Emission 2	Emission 3	Emission 4	Emission 5	Emission 6
		tCO_2	tCO_2	tCO_2	tCO_2	tCO_2	tCO_2
x_1	Variable	Activity data 2	Activity data 1	Activity data 2	Production	Emission 1	Emission 1
	Value	1 061	-56.53	55.81	10 271	-21.06	2.03
C		169 873	112 778	-10 031	4 795	45 481	1 809 235
SE (x_1)		144.08	36.01	18.92	14 140	11.74	7.09
SE (C)		224 009	20 868	29 417	14 177	18 253	33 536
R^2		0.61	0.07	0.20	0.02	0.09	0.00
$RMSE$		69 092	11 195	9 073	6 137	5 630	111 163
F		54.23	2.47	8.70	0.53	3.22	0.08
$F_{critical}$		4.13	4.13	4.13	4.13	4.13	4.13
ν_2		34	34	34	34	34	34
SS_{reg}		2.59×10^{11}	3.09×10^{8}	7.16×10^{8}	1.99×10^{7}	1.02×10^{8}	1.01×10^{9}
SS_{res}		1.62×10^{11}	4.26×10^{9}	2.80×10^{9}	1.28×10^{9}	1.08×10^{9}	4.20×10^{11}

From Table B-2, it is evident that only *emission 1* has a R^2 greater than 0.5. The R^2 for *emissions 2, 4, 5*, and *6* implies that there is no relationship between the dependent and independent variables, while a weak relationship is implied for *emission 3*.

For *emission 1* and *emission 3*, $F > F_{critical}$ while the opposite is true for *emission 2* and *emission 4*. This implies that only the regression for *emission 1* and *emission 3* is justified; thus constant multipliers are used for *emission 2, 4, 5*, and *6*. The constant multipliers can be calculated using Equation (B-3).

$$Constant\ multiplier = \frac{Forecasted\ Variable}{Dependent\ variable} \quad \text{(B-3)}$$

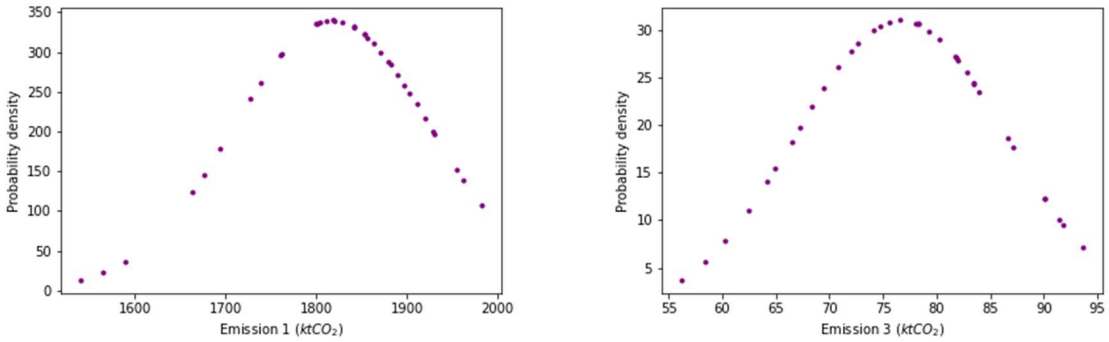
The constant multipliers for *emission 2, 4, 5*, and *6* of the fuel combustion emissions can be found in Table B-3.

Table B-3: Constant multipliers used for the fuel combustion emissions.

Forecasted Variable	Dependent variable	Constant multiplier
Emission 2	Activity data 1	138.83 tCO_2 knm^{-3} h
Emission 4	Production	15 066 tCO_2
Emission 5	Emission 1	7.04×10^{-3} tCO_2
Emission 6	Emission 1	2.17×10^{-3} tCO_2

To justify the regression model used to calculate *emission 1* and *emission 3*, remaining four assumptions of linear regression need to be addressed.

As with the activity data, the multivariate normality for the fuel combustion emissions can be tested using distribution plots (Figure B-7).



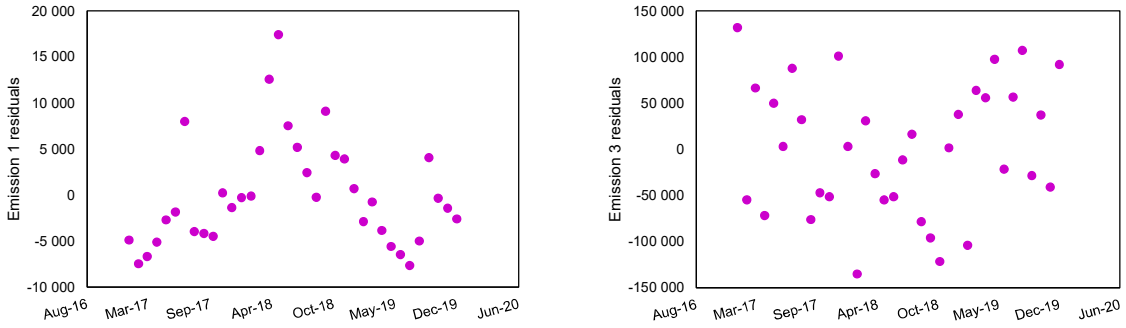
a) *Emission 1* distribution plot b) *Emission 3* distribution plot

Figure B-7: Fuel combustion emissions distribution plots.

Both *emission 1* and *emission 3* have a fairly normal distribution plot; however, *emission 1* is slightly skewed to the right.

The VIF value for *emission 1* and *emission 3* were calculated as 2.59 and 1.26, respectively. Since both these values are less than 4, no multicollinearity is assumed.

To test for autocorrelation between variables, the residuals are plotted against time as illustrated in Figure B-8.

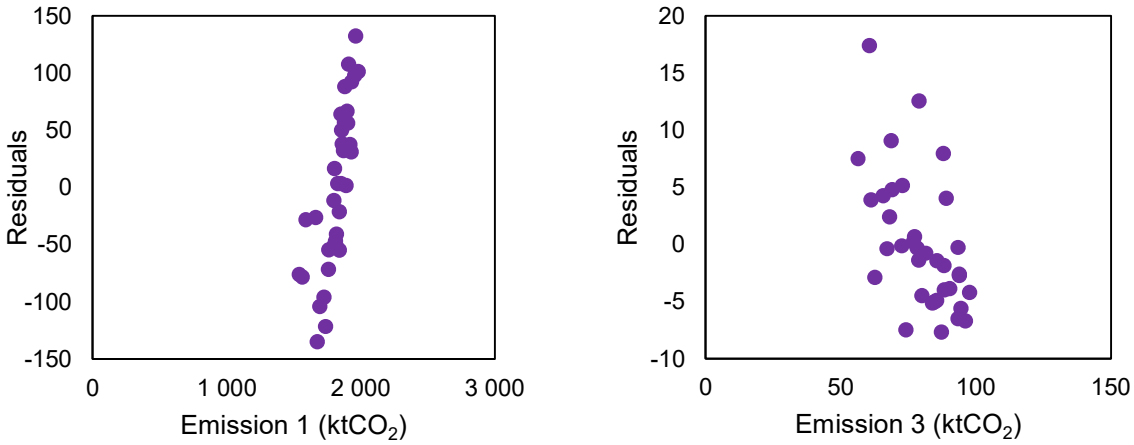


a) *Emission 1* residual plot b) *Emission 3* residual plot

Figure B-8: Autocorrelation test for the fuel combustion emissions.

As with the activity data, no noticeable pattern can be observed for *emission 1* and *emission 3*; thus it is likely that no autocorrelation exist between the variables.

The residual plots used to determine the homoscedasticity of the variables are illustrated in Figure B-9.



a) Emission 1 b) Emission 3

Figure B-9: Homoscedasticity test for fuel combustion emissions.

For both *emission 1* and *emission 3*, a linear relationship can be observed between the fitter values and the residuals. Furthermore, no funnel-shape patterns can be observed for both the variable; thus heteroscedasticity is likely not present in the data.

All five assumptions of linear regression are valid for *emission 1* and *emission 3*, thus the use regression models are valid.

The process emissions can be calculated using linear regressions with the material input data or *emission 1* of the process emissions as the independent variable. Similar to the fuel combustion models, and outlier analysis was done on all the regression models. The resulting linear regression models for the process emissions can be found in Figure B-10.

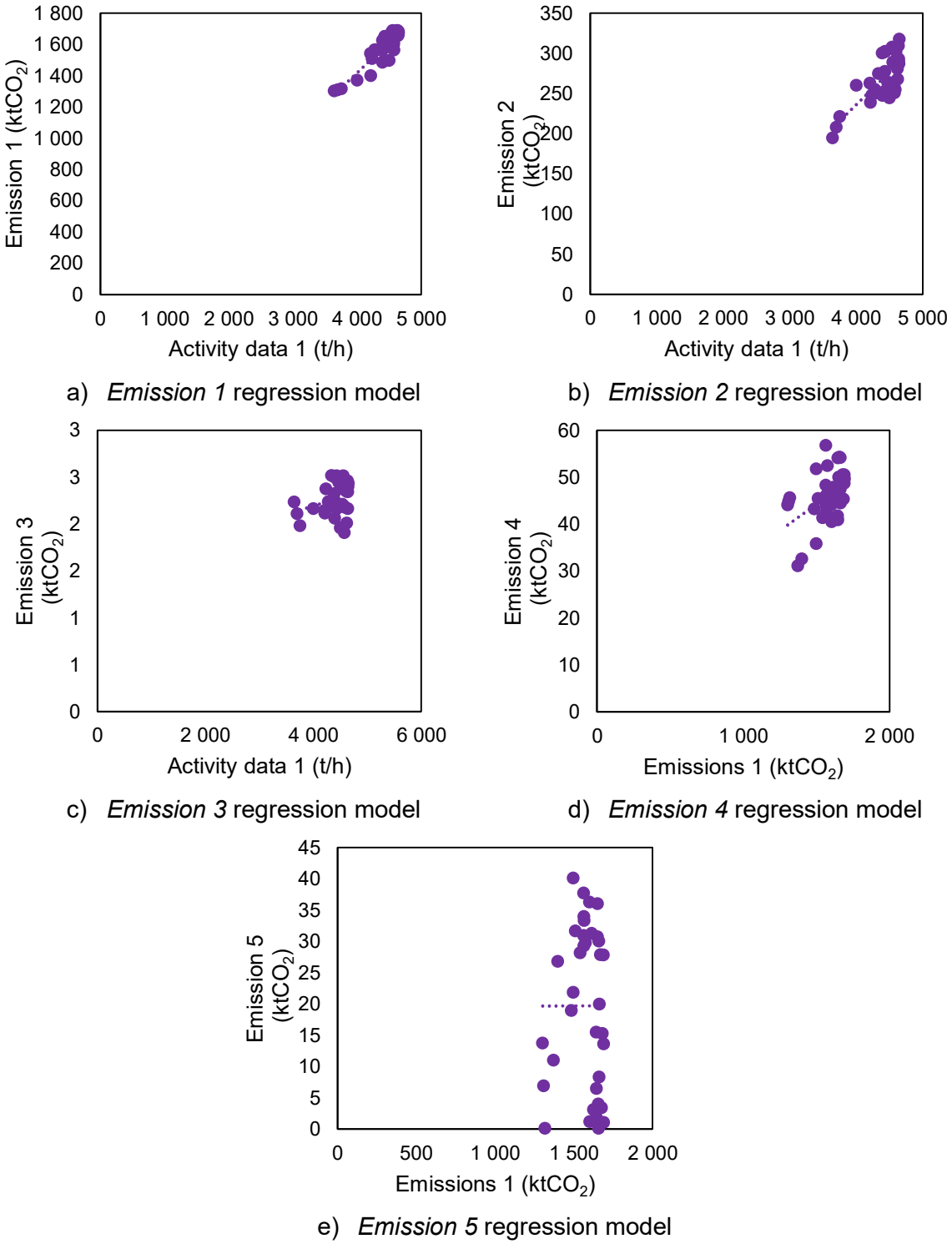


Figure B-10: Regression models for the process emissions.

From Figure B-10, only *emission 5* shows no correlation between the dependent and independent variable. This observation is further emphasised by the regression parameters for the fugitive emissions given in Table B-4.

Table B-4: Linear regression parameters for the process emissions.

Parameter		Emission 1	Emission 2	Emission 3	Emission 4	Emission 5
		tCO_2	tCO_2	tCO_2	tCO_2	tCO_2
x_1	Variable	Material input data	Material input data	Material input data	Emission 1	Emission 1
	Value	1 759 988	398 679	1 271	2.25×10-2	4.16×10^{-5}
C		-185 023	-129 585	995	10 541	19.625
SE (x_1)		174 827	54 988	481.05	0.01	0.02
SE (C)		175 090	55 071	481.77	12 104	31 290
R^2		0.75	0.61	0.17	0.20	0.00
$RMSE$		57 526	18 094	158.29	5 130	13 623
F		101.34	52.57	6.98	8.62	0.00
$F_{critical}$		4.13	4.13	4.13	4.13	4.13
ν_2		34	34	34	34	34
SS_{reg}		3.35×10^{11}	1.72×10^{10}	1.75×10^{5}	2.27×10^{8}	7.76×10^{2}
SS_{res}		1.13×10^{11}	1.11×10^{10}	8.52×10^{5}	8.95×10^{8}	5.98×10^{9}

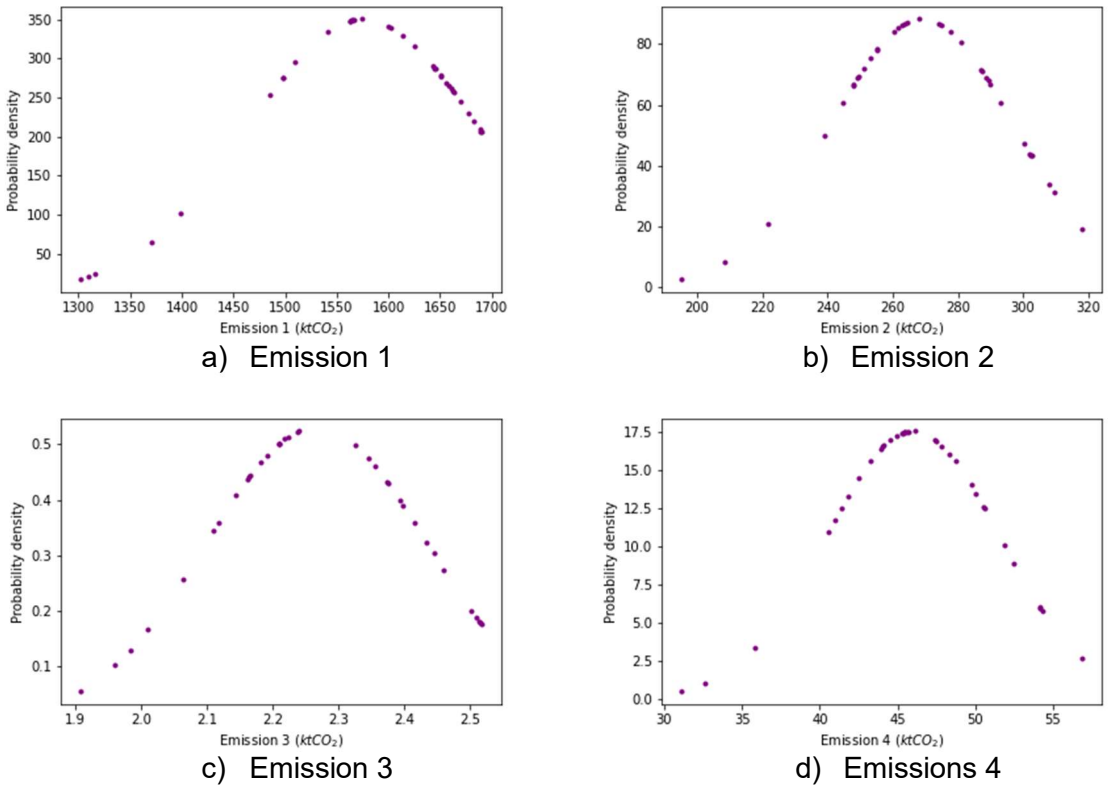
From Table B-4, it is evident that *emission 1* and *emission 2*, have a strong relationship seeing that $R^2 > 0.5$. *Emission 3* and *emission 4* have a weak relationship while *emission 5* has almost no relationship. The regression models for all the process emissions, except *emission 5*, are justified since $F > F_{cricitical}$. Thus, constant multipliers were used to calculate *emission 5*. These constant multipliers can be found in Table B-5.

Table B-5: Constant multipliers used for the process emissions.

Variable	Dependent variable	Constant multiplier
Emission 5	Emission 1	1.25×10^{-2}

To justify the regression models used to the process emission, the remaining four assumptions of linear regression needs to be addressed.

The distribution plots used to determine if the multivariate normality assumption is met are illustrated in Figure B-11.



a) Emission 1 b) Emission 2

c) Emission 3 d) Emissions 4

Figure B-11: Process emissions distribution plots.

From Figure B-11, a normal distribution can be observed for *emission 2, 3,* and *4.* For *emission 1,* the data is slightly skewed to the left.

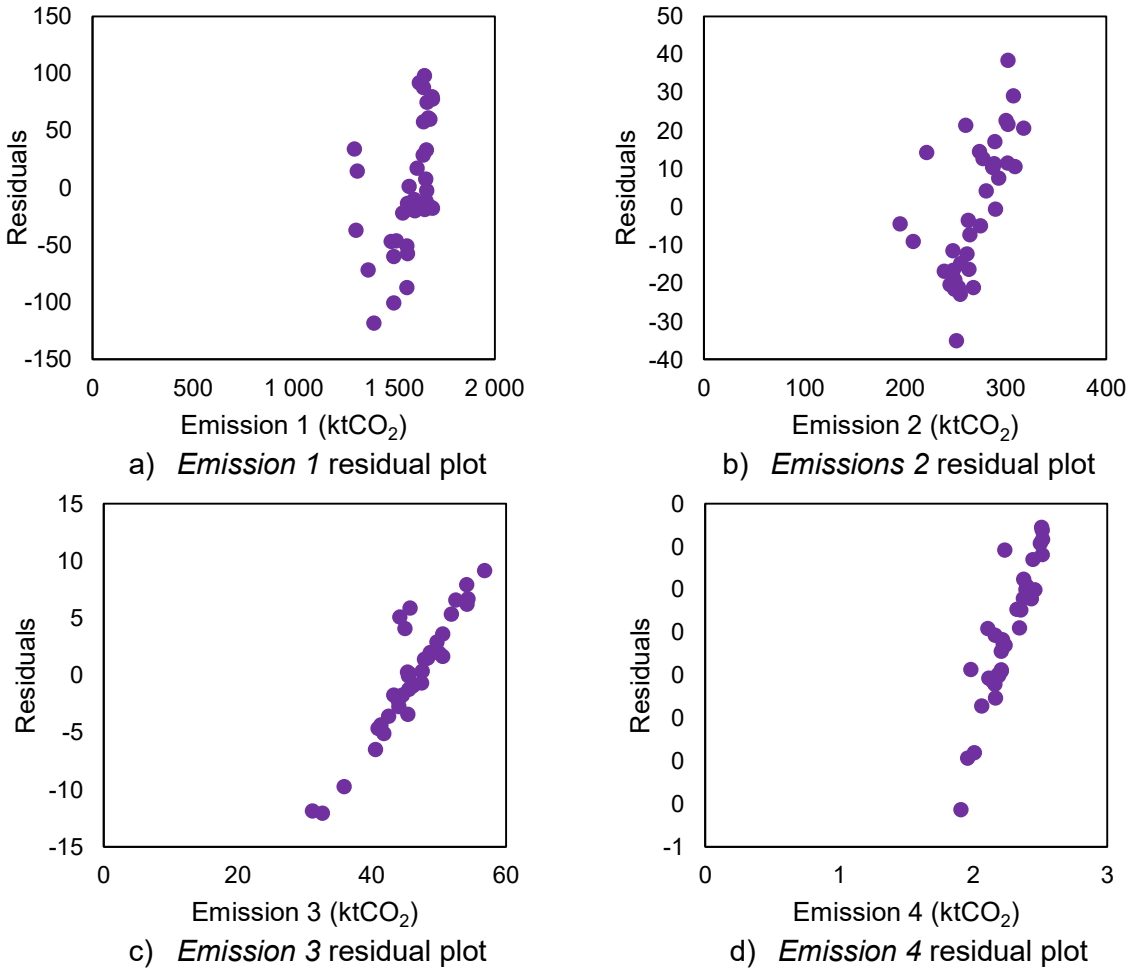
To determine if multicollinearity exist between the variables, the VIF value is calculated. The VIF values for the process emissions can be found in Table B-6.

Table B-6: Multicollinearity test for the process emissions.

Variable	VIF value
Emission 1	3.98
Emission 2	2.54
Emission 3	1.21
Emission 4	1.25

All the values in Table B-6 are less than 4; thus is assumed that no multicollinearity is present.

The residual plots used to determine the homoscedasticity of the variables are illustrated in Figure B-12.



a) *Emission 1* residual plot
b) *Emissions 2* residual plot
c) *Emission 3* residual plot
d) *Emission 4* residual plot

Figure B-12: Homoscedasticity test for the process emissions.

For *emission 1* and *emission 2,* heteroscedasticity might be present; however more data points are required to determine if this is the case. For *emission 3* and *emission 4* a linear relationship can be observed between the fitted values and the residuals; thus heteroscedasticity is likely not present. For the process emissions, all five assumptions of linear regression are confirmed, thus the linear regression models are valid.

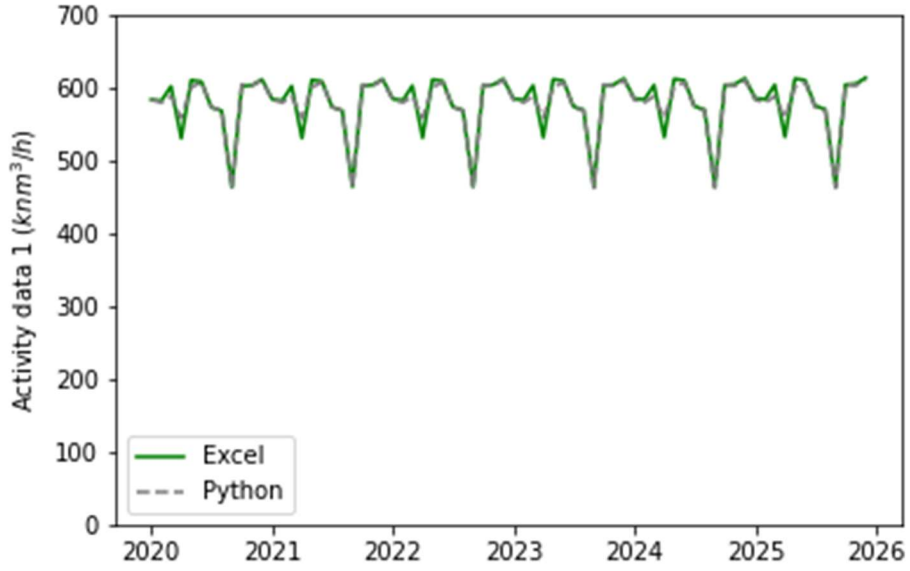
Since the fugitive emissions include the routine and emergency flaring or venting of gases, these emissions are calculated using a constant multiplier with *activity data 2* used as the dependent value. The constant multiplier used to determine the fugitive emissions can be found in Table B-7.

Table B-7: Constant multipliers used for calculating the fugitive emissions.

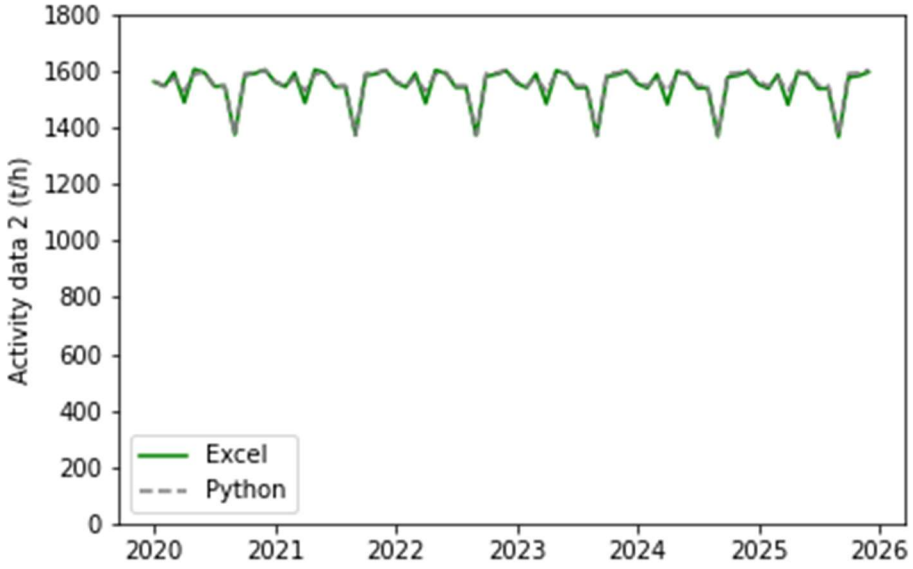
Variable	Dependent variable	Constant multiplier
Emission 1	Activity data 2	8.84×10^{-2}

B.2 Validation of Excel models using Python

With the forecasted input data, the activity data is calculated using multivariable linear regression. The results obtained using Excel and Python, respectively are illustrated in Figure **B-13**.



a) Comparison of the forecasted *activity data 1* using Excel and Python



b) Comparison of the forecasted *activity data 2* using Excel and Python

Figure B-13: Comparison of the forecasted activity data using Excel and Python.

From Figure B-13 (a), it is evident that the forecasted values for Activity data 1 are very similar when using Excel and Python, respectively. For Activity data 2 (Figure B-13 (b)), the Excel and Python models provide similar results in the beginning, whereafter a slight deviation in the values can be observed. This is most likely due to the influence of the material input data being larger than the influence of the production data.

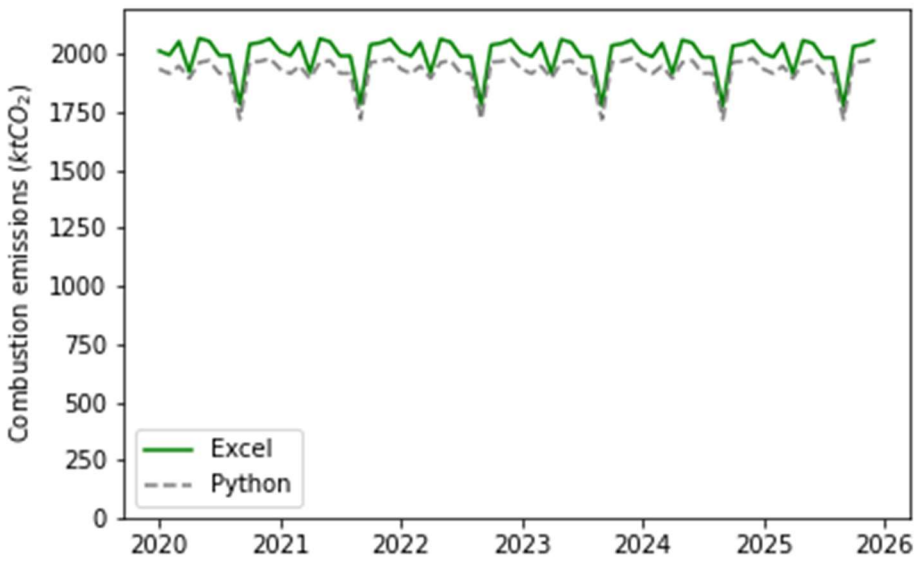
The magnitude if the deviation between the Excel and Python models were determined by comparing the average 2025 values for the activity data. The results are summarised in Table **B-8**.

Table B-8: Comparison of the average 2025 activity data using Excel and Python.

Activity data	2025 Excel average	2025 Python average	Difference
Activity data 1 (knm^3/h)	579.28	577.58	0.34%
Activity data 2 (t/h)	1 543.13	1 552.74	0.62%

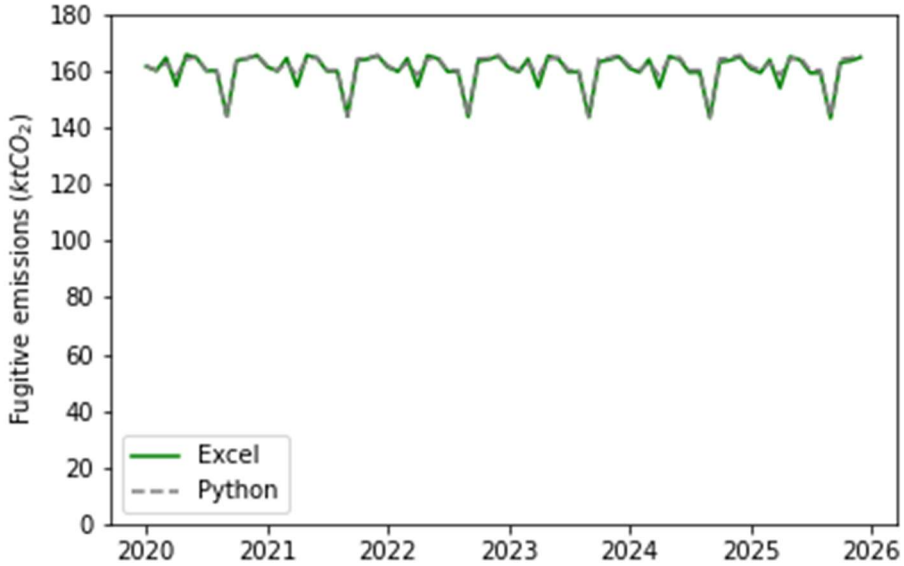
The difference between the average 2025 values using Excel and Python is calculated as 0.34% and 0.62% for Activity data 1 and Activity data 2, respectively. Thus, the Excel model can be validated since both the calculated differences are less than 5%.

With the activity data, the emissions are calculated using linear regression. The results obtained using Excel and Python, respectively are illustrated in Figure B-14.

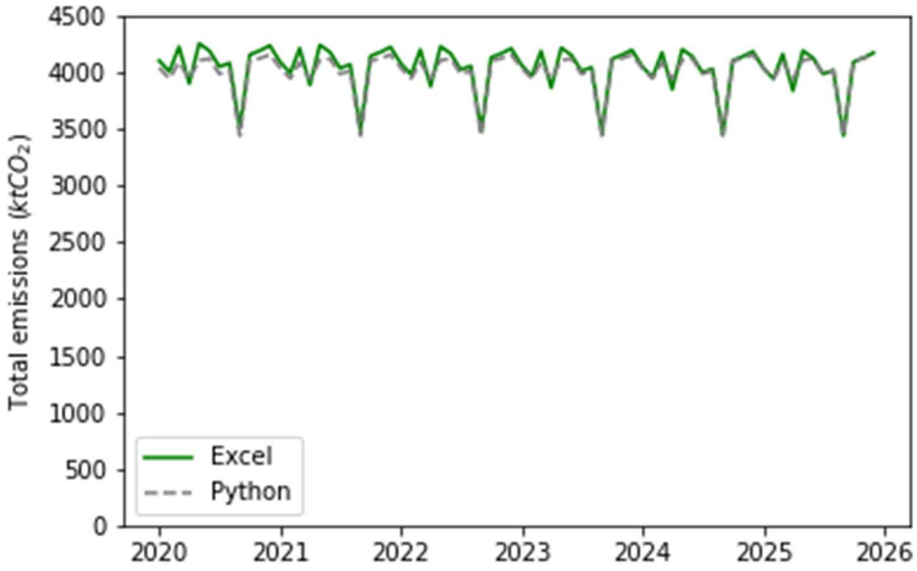


a) Comparison of forecasted fuel combustion emissions using Excel and Python

b) Comparison of forecasted process emissions using Excel and Python



c) Comparison of forecasted fugitive emissions using Excel and Python



d) Comparison of the total forecasted emissions using Excel and Python

Figure B-14: Comparison of the forecasted emissions using Excel and Python.

In Figure B-14, a similar pattern can be observed between the Excel and Python models for all four graphs. For the fuel combustion emissions (Figure B-14 (a)) the Excel model produced much higher values than the Python model. For the process and fugitive emissions (Figure **B-14** (b) and (c)), the two models initially have similar results, whereafter a deviation between the models can be observed. The total emissions (Figure B-14 (d)), shows a slight deviation between the Excel and Python models initially, whereafter similar results are observed by 2025.

The magnitude if the deviation between the Excel and Python models were determined by comparing the total value for 2025 for the respective emissions as well as the total emissions. The results are summarised in Table B-9.

Table B-9: Comparison of the total 2025 emissions using Excel and Python.

Emissions	2025 Excel total	2025 Python total	Difference
Fuel combustion emissions ($ktCO_2$)	23 942	23 108	3.61%
Process emissions ($ktCO_2$)	22 200	22 943	3.24%
Fugitive emissions ($ktCO_2$)	1 917	1 928	0.56%
Total emissions ($ktCO_2$)	48 059	47 980	0.16%

From Table B-9, it is evident that the fuel combustion emissions have the largest difference (3.61%) when comparing the Excel model to the Python model, while a difference if 3.24%

and 1.75% was observed for the process and fugitive emissions, respectively. When combining all the emissions, a difference of 0.69% can be observed between the Excel model and the Python model. Thus, the Excel model for the respective emissions can be validated since the calculated differences are less than 5%.