



A general approach to develop and assess models estimating coal energy content

C van Aarde

 **orcid.org 0000-0003-1382-4562**

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

Supervisor:

Prof Marius Kleingeld

Graduation May 2019

Student number: 23380950

ABSTRACT

Title: A general approach to develop and assess models estimating coal energy content

Author: C van Aarde

Supervisor: Prof Marius Kleingeld

Keywords: Calorific value, proximate analysis, model development, objective comparison, visualisation

The gross calorific value (GCV) is an important property defining the energy content of fuels such as coal. Many industries require the GCVs to accurately quantify and report the energy efficiency of the operations. Unfortunately, measuring the GCVs of coal can be a time-consuming and expensive process. Multiple researchers have therefore developed models to predict the GCVs based on more accessible variables such as proximate and ultimate compositions.

This study presents a review of models available from literature as background to the problem statement. These models were developed for specific coal types from various geographical locations. Since the properties of coal differs between locations, it is questionable whether these models are applicable for coal from a new location. This limiting factor, along with the various options of existing models identifies the problem of this study: Will the literature models be applicable to a new coal dataset and which one would be best?

A preliminary evaluation testing the application of existing literature models on new coal data showed significant discrepancies in the results. This evaluation demonstrated that the literature models perform differently on new data with errors ranging from 3.7% to 72.1%. The evaluation also shows that various approaches are used in the development of these models. The significant variance in results together with the different model attributes makes it challenging to objectively assess and compare suitable models. The evaluation findings necessitate the need to devise a general industry “best practice” approach to develop new models in a consistent way. Furthermore, it also requires a methodology to objectively compare the different model characteristics and subsequent results.

A detailed literature study was conducted to determine the general steps required for model development. The literature study identified three common focus areas namely data preparation, model development and model validation. An additional focus area was also

added to assess techniques to visually evaluate, assess and compare the various models. This fourth focus area introduces a new and practical method to ultimately identify the most suitable model to use on new coal data.

The four focus areas identified in the literature study were combined to devise a methodology to use for the development and comparison of GCV models. The methodology layout follows the four focus areas from literature: *data preparation*, *model development*, *model validation* and *comparison of results*. Each focus area consists of several sub-steps based on the industry and academic best practices obtained from literature.

The methodology was tested and verified by applying it to three different case studies. The case studies consist of coal data from South Africa, India and Alaska. A new GCV model is developed for each dataset and presented errors in the range of 0.26% to 2.63%. These results verifies the new methodology's ability to consistently deliver high quality models.

The visualisation technique is further used to investigate and validate results. The original assessment combining available models and new unrelated data was repeated. The results show that there is a significant difference between model results obtained using model focus area / region specific data vs. unrelated data (0.26% – 2.63% vs. 8.26% – 72.3%). However, the new technique now allows the user to quickly assess model quality and accuracy. This ultimately enables the user to select the most appropriate model for a specific dataset.

The study identified the need to objectively develop and compare the performance and applicability of GCV models. A wide literature survey was conducted to find academic and industry best practice techniques required to create a structured approach to develop new GCV models. The method to compare the models was applied to case studies and enabled the user to identify which model would be best in a practical and objective manner.

ACKNOWLEDGEMENT

I would firstly like to express my sincere gratitude to Dr Andries Gous and Miss Lee-Ann Botes for their expert guidance, involvement and time to help compile a quality dissertation. Your assistance and mentoring made this study possible.

I would further like to give special thanks to Dr Walter Booysen for his outstanding leadership and contributions to this study.

Finally, I would like to thank Prof Edward Mathews and Prof Marius Kleingeld for providing the opportunity and resources to conduct this research. Thank you Enermanage (Pty) Ltd and its sister companies for financial support to complete this study.

CONTENTS

Abstract	ii
Acknowledgement	iv
Contents	v
List of Figures	vii
List of Tables.....	ix
List of acronyms and Abbreviations	xi
Nomenclature.....	xiii
Chapter 1 - Introduction	1
1.1 Preamble	2
1.2 Background on the quantification of energy	2
1.3 Problem statement.....	8
1.4 Objectives and scope of investigation	9
1.5 Conclusion.....	10
Chapter 2–Literature study.....	11
2.1 Preamble	12
2.2 Best practice approach to model development.....	12
2.3 Dataset preparation	16
2.4 Model development.....	28
2.5 Model validation.....	41
2.6 Model comparison	44
2.7 Conclusion.....	53
Chapter 3 – Methodology	54
3.1 Preamble	55
3.2 Dataset preparation	55
3.3 Model development.....	60
3.4 Model validation.....	63
3.5 Model comparison	67
3.6 Conclusion.....	73
Chapter 4 – Verification and validation	75
4.1 Preamble	76
4.2 Dataset preparation	76
4.3 Model development.....	84
4.4 Model validation.....	88
4.5 Model comparison	90
4.6 Conclusion.....	97
Chapter 5 – Conclusion and recommendations	99
5.1 Preamble	100
5.2 Overview of study	100
5.3 Meeting the objectives	101
5.4 Recommendation for further research.....	102

5.5	Closure of study.....	104
Reference List		105
Appendix A – Case study results.....		113
A.1	Case study 1: South African coal	114
A.2	Case study 2: Indian coal.....	126
A.3	Case study 3: Alaskan coal.....	136
Appendix B – Visualisation discussion.....		146
B.1	New model: New Zealand and Australian coal	147
B.1.1	Dataset preparation	147
B.1.2	Model development.....	150
B.1.3	Model validation.....	152

LIST OF FIGURES

Figure 1-1: Total primary energy supply (TPES) in South Africa 2016.....	2
Figure 2-1: Best practice approach for model development [28].....	13
Figure 2-2: Outlier identification method.....	24
Figure 2-3: Effect of the proximate variables on the GCV of coal[18]	29
Figure 2-4: Effect of the proximate and ultimate variables on the GCV of coal [21]	29
Figure 2-5: Distribution of data with a linear and nonlinear fit	34
Figure 2-6: MATLAB results for error percentages	38
Figure 2-7: MATLAB results for model fit.....	39
Figure 2-8: Desired relationship between the predicted and experimental GCVs	42
Figure 2-9: DIKW hierarchy.....	45
Figure 2-10: Visualisation suggestion chart [77]	47
Figure 2-11: Relationship between the FC and GCV (a) and the sample size (b).....	48
Figure 2-12: Comparison of the proximate variables of 3 groups of data.....	48
Figure 2-13: Distribution of the GCV of the five groups	49
Figure 2-14: Composition of group 1's coal	50
Figure 2-15: Visual comparison of models	51
Figure 2-16: Combined visualisation of models	52
Figure 3-1: Four focus areas for GCV modelling of coal.....	55
Figure 3-2: Sub-steps for the data preparation focus area	56
Figure 3-3: Data split into modelling and validation sets.....	59
Figure 3-4: Sub-steps for the model development focus area	60
Figure 3-5: Sub-step for the model validation focus area	64
Figure 3-6: MAE of the developed model on validation data	65
Figure 3-7: Experimented vs predicted GCVs	66
Figure 3-8: Sub-steps for the model comparison focus area	67
Figure 3-9: Data used for visualisation parameters	69
Figure 3-10: Scatterplot area with the identified limits	70
Figure 3-11: Scatterplot with the limits and the models	71
Figure 3-12: Visualisation of models performance on a) own data and b) new data	72
Figure 3-13: Devised approach for GCV modelling of coal.....	74
Figure 4-1: devised approach for GCV modelling of coal	76
Figure 4-2: Proximate analysis versus GCV relationship on all three sources' data	79
Figure 4-3: Proximate analysis versus GCV relationship on source 1 and 3 data.....	80

Figure 4-4: Proximate analysis versus GCV relationship on the clean dataset	82
Figure 4-5: Proximate analysis versus GCV relationship on the data split	83
Figure 4-6: Relationship between A (a) and IM (b) vs the GCV	86
Figure 4-7: Relationship between FC (a) and VM (b) vs the GCV	87
Figure 4-8: Validation of the developed model	89
Figure 4-9: Visualisation 1(a) – comparison on published modelling data	93
Figure 4-10: Visualisation 1(b) – comparison on published validation results	94
Figure 4-11: Visualisation 2(a) - comparison on new data	96
Figure 4-12: Visualisation 2(b) – Comparison on new data with new model	97
Figure A-1: Proximate analysis versus GCV relationship on all three source's data	116
Figure A-2: Proximate analysis versus GCV relationship on source 1 and 3 data	117
Figure A-3: Proximate analysis versus GCV relationship on the clean dataset	121
Figure A-4: Proximate analysis versus GCV relationship on the data split	122
Figure A-5: Relationship between A (a) and IM (b) vs the GCV	124
Figure A-6: Relationship between FC (a) and VM (b) vs the GCV	125
Figure A-7: Validation of the developed model	126
Figure A-8: Case study 2 raw data distribution	128
Figure A-9: Clean case study 2 data	130
Figure A-10: Data split of case study 2 data	131
Figure A-11: Relationship between A (a) and IM (b) vs the GCV	134
Figure A-12: Relationship between FC (a) and VM (b) vs the GCV	134
Figure A-13: Validation of the developed model	136
Figure A-14: Case study 2 raw data distribution	138
Figure A-15: Clean case study 3 data	140
Figure A-16: Data split of case study 3 data	141
Figure A-17: Relationship between A (a) and IM (b) vs the GCV	143
Figure A-18: Relationship between FC (a) and VM (b) vs the GCV	144
Figure A-19: Validation of the developed model	145
Figure B-1: Proximate analysis versus GCV relationship (new model)	148
Figure B-2: Proximate analysis versus GCV relationship on clean data	149
Figure B-3: Proximate analysis versus GCV relationship data split	150
Figure B-4: Validation of the new model	153

LIST OF TABLES

Table 1-1: Characteristics of typical coal types[10].....	4
Table 1-2: Literature GCV models tested with random coal samples	5
Table 1-3: Main methods followed by literature models.....	6
Table 2-1: Evaluating literature models with Xu's best practise approach [28].....	14
Table 2-2: Conversions between different coal bases [39]	17
Table 2-3: Gather and understand data of GCV models.....	18
Table 2-4: Outlier removal of literature models.....	22
Table 2-5: Transformed data of literature GCV models	25
Table 2-6: Data splitting of GCV models	26
Table 2-7: Types of models developed by the literature GCV models	30
Table 2-8: Variables included in the literature GCV models.....	31
Table 2-9: Investigated literature GCV models.....	33
Table 2-10: Nonlinear regression types.....	35
Table 2-11: LINEST function output (matrix)	36
Table 2-12: LINEST matrix statistics explanation [60]	37
Table 2-13: Validation procedures of literature GCV models.....	43
Table 2-14: Data / information table	45
Table 2-15: Boundary ranges for R^2 and MAE	46
Table 3-1: Example of coal samples including missing data.....	57
Table 3-2: Coal samples containing outliers.....	58
Table 3-3: LINEST matrix.....	61
Table 3-4: Testing criteria for model checks [27], [37], [88]	63
Table 3-5: Model validation parameters and criteria.....	66
Table 3-6: Visualisation parameters and limits	69
Table 4-1: Summary of obtained data for case study 1	77
Table 4-2: Sample list of the potential outliers.....	81
Table 4-3: Case studies data summary	84
Table 4-4: LINEST matrix results	85
Table 4-5: Developed models with criteria for each case study	88
Table 4-6: Case studies validation results	90
Table 4-7: Reported R^2 and MAE parameters of the literature models.....	91
Table 4-8: Calculated R^2 and MAE parameters of the developed models	92
Table 4-9: Calculated CC and MAE on the new dataset.....	95

Table A-1: Summary of obtained data.....	114
Table A-2: Full list of the potential outliers.....	118
Table A-3: LINEST matrix for South African coal regression model.....	123
Table A-4: Summary of case study 2 data.....	127
Table A-5: Sample list of the potential outliers	128
Table A-6: LINEST matrix results	132
Table A-7: Summary of case study 3 data.....	137
Table A-8: Sample list of the potential outliers	138
Table A-9: LINEST matrix results	142
Table B-1: LINEST matrix for new coal regression model	151

LIST OF ACRONYMS AND ABBREVIATIONS

Abbreviation	Description
A	<i>Ash</i>
ad	<i>Air-dried</i>
ADR	<i>Average relative deviation</i>
ar	<i>As-received</i>
ANFIS	<i>Adaptive neuro-fuzzy system</i>
ANN	<i>Artificial neural network</i>
ANOVA	<i>Analysis of variance</i>
ASTM	<i>American society for testing and materials</i>
C	<i>Carbon (element)</i>
CC	<i>Comparison coefficient</i>
daf	<i>Dry, ash-free</i>
DIKW	<i>Data-information-knowledge-wisdom</i>
VM	<i>Volatile matter</i>
M	<i>Moisture</i>
IM	<i>Inherent moisture</i>
FC	<i>Fixed carbon</i>
GCV	<i>Gross calorific value</i>
H	<i>Hydrogen (element)</i>
LINEST	<i>Line statistics</i>
MAE	<i>Mean absolute error</i>
MATLAB	<i>Matrix laboratory</i>
N	<i>Nitrogen (element)</i>

O	<i>Oxygen (element)</i>
P	<i>Sulphur (element)</i>
R^2	<i>Coefficient of determination</i>
RMSE	<i>Root mean square error</i>
SE	<i>Standard error</i>
TPES	<i>Total primary energy supply</i>
VBA	<i>Visual basic for application</i>

NOMENCLATURE

Term	Description
Basis of analysis	The type of base
Coalification stage	<i>See basis of analysis</i>
Developed model	<i>The models developed in this study</i>
Gross calorific value	<i>Heat produced by combustion of unit quantity of a coal sample when burned at constant volume in an oxygen comb calorimeter under specified conditions</i>
Literature model	<i>The GCV models found in literature used in this study</i>
Outlier	<i>An observation that lies an abnormal distance from the other values in a random sample from a population</i>
Parameter	<i>The parameters associated with validation criteria (e.g. R^2, MAE and RMSE)</i>
Proximate analysis	<i>A quantitative analysis of the moisture, volatile matter, fixed carbon (by difference) and ash content present in a coal sample</i>
Ultimate analysis	<i>A quantitative analysis of various elements present in a coal sample, such as carbon, hydrogen, sulphur, oxygen and nitrogen</i>
Regression	<i>A technique for determining the statistical relationship between two or more variables where one or more variables are dependent on one or more other variables</i>
Regressor	<i>The independent variables in a regression equation</i>
Variables	<i>The variables associated with the proximate and ultimate analysis (e.g. ash, moisture, hydrogen, oxygen, etc.)</i>

CHAPTER 1 - INTRODUCTION

1 Introduction

1.1 Preamble

Coal and the energy it provides are of fundamental importance for the world[1]. Many industries rely on coal energy content to satisfy their industrial need [2]. This chapter aims to provide background and context into the importance of understanding coal as an energy source.

This chapter further provides general background on the inherent energy in coal and the problems that are faced when quantifying it. These associated problems are highlighted (Section 1.3) and identified the need and objectives for the study (Section 1.4).

1.2 Background on the quantification of energy

1.2.1 Coal as an energy source

Global energy demands are predominately satisfied with fossil fuels such as fuel-oil, natural gas and coal [2]. Among the fossil fuels, coal is the most abundant energy resource contributing approximately 40% to the world's energy[1]. Figure 1-1 illustrates that energy derived from coal provides approximately 70% of South Africa's total energy supply [3]. The remaining supplies are produced from crude oil (14.8%), biofuels and waste (10.7%), nuclear (2.4%), gas (2.9%), and other sources (0.1%). Understanding coal as an energy source is therefore vital for the energy landscape.

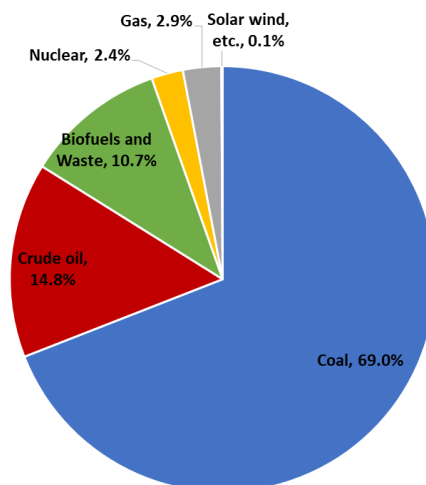


Figure 1-1: Total primary energy supply (TPES) in South Africa 2016

Coal is primarily used as a solid fuel to produce electricity and heat through combustion [2]. The electricity industry, along with several other industries such as the ferrochrome, iron and steel, and petrochemical facilities, use coal as a primary energy source [3]. It is therefore expected that coal usage would receive high priority within the respective energy management strategies.

Energy management forms a fundamental part in any energy intensive industry [4]. To determine and improve their energy efficiency, industries need to accurately know the energy content of their coal [5]. Knowing the accurate energy content of coal is also important for energy reporting purposes and other tax incentives, such as section 12L of the income tax bill [6]. It is, therefore, crucial to know the exact energy content of the coal for industrial applications.

1.2.2 Coal energy quantification

Coal defines several materials that have a wide range of properties [7]. Coal can simplistically be described as the altered remains of prehistoric vegetation that originally accumulated in swamps and peat bogs. This material has solidified over the years and captured the carbon properties of the material [1]. Differences in the material and the solidification process produced different types of coal with inherently different properties.

It is a common practice within the coal industry to assess the quality of coals and to express it using calorific values (CV), proximate and ultimate analyses. The CV is usually expressed as the gross calorific value (the higher heating value) or the net calorific value (the lower heating value). The difference between the gross calorific value (GCV) and the net calorific value (NCV) is the latent heat of condensation of the water produced during the combustion process [8]. NCV varies since it does not include the condensation heat of water [9].

The proximate analysis includes the weight percentage of the ash (A), moisture (M), fixed carbon (FC) and volatile matter (VM) present in the coal. The ultimate analysis is an elemental analysis done on the coal and includes the carbon (C), hydrogen (H), nitrogen (N), sulphur (S) and oxygen (O) content in weight percentage. Both these analyses can be measured in several different bases. The different bases, also known as the coalification stages, mainly depend on the moisture taken into consideration when analysing the sample.

As-received (ar) basis takes all the variables into consideration and uses the total weight as the basis of the measurement, therefore including the total moisture. Air-dried (ad) basis neglects the presence of moisture other than inherent moisture while dry-base (db) leaves out

all moisture. Dry, ash-free (daf) and dry, mineral-free (dmf) leave out all moisture and the ash and the mineral matter respectively. Conversions from one base to another can be performed using mass balance equations [8].

Table 1-1 presents typical coal samples with ranges for the analyses discussed on an as-received basis. It should be noted that these properties can / will differ for other bases (refer to Chapter 2).

Table 1-1: Characteristics of typical coal types[10]

Variables	Coalification stage	Lignite	Sub - bituminous	Bituminous	Anthracite
Proximate analysis					
Ash (%)	Air-dried	>40	25-40	15-25	5-15
Moisture (%)	As-received	30-55	10-25	10-15	<10
Inherent moisture (%)	Air-dried	8-12	6-10	5-8	1-4
Fixed carbon (%)	Dry ash-free	55-65	60-70	70-85	>80
Volatile matter (%)	Dry ash-free	>50	40-50	30-40	5-25
Ultimate analysis					
Carbon (%)	Air-dried	65-72	72-76	70-75	>75
Hydrogen (%)	Air-dried	4.5-5.5	4-4.5	3-4.5	2-3
Nitrogen (%)	Air-dried	~1-2	~1-2	~1-2	~1-2
Sulphur (%)	Air-dried	0-1.5	1.5-3.5	1.5-3.5	0-1.5
Oxygen (%)	Air-dried	20-30	15-20	10-12	3-8
Energy content					
GCV (MJ/kg)	Ash-free	12.5-15.5	17.6 - 24.2	26.4 – 31.9	30.0-38.0

It can be seen that there is a difference between the proximate and the ultimate variables of the different types of coal. The respective variables do not only differ in type but are also not limited to a single value. Variables from different types of coal can potentially be the same, thus indicating that a single value does not define specific coal. Differences in the proximate and ultimate analyses are not limited to the coal type but would further vary based on the geographical location of the coal [11].

The energy content of the coal is dependent on the different properties as expressed by the proximate and ultimate variables [12]. Any variation within the analysis variables will therefore cause a variation in the energy content. This energy content needs to be quantified for the different types to ensure that it remains suitable for the application.

The energy content of the coal (GCV) is usually measured in MJ/kg[13]. The most suitable and accurate apparatus for determining the CVs of coal is the bomb calorimeter [8]. This apparatus is adopted in the ASTM D5865-12 (Standard test method for gross calorific value of coal and coke) and the ISO 1928:2009 (Solid mineral fuel – Determination of gross calorific value) [8]. However, this is an expensive process and requires trained personnel to perform [14]. Therefore, many companies do not measure the GCV of coals if it is not required.

Contradictory to the GCV, the proximate and ultimate analyses are usually readily available due to the ease of determination. The analysis process to obtain these variables is much simpler and cheaper than the process to measure the GCV [15]. These variables are furthermore measured instead, since some of this information on the properties of the coal is required for downstream processes [16].

Research is continuously performed to identify methods on how to use these proximate and ultimate analyses to determine the GCV. The dependence of the GCV on the proximate and ultimate analyses enables the development of models using these variables [17]. One of the limitations of these existing models is that they were developed on different types of coal from different locations, which creates the potential for differences since the properties of coal can vary between various locations and types. The question therefore arises whether these models could be used for different coal.

A small control test was performed to evaluate this question. Table 1-2 shows five existing models from literature (from various locations) with their reported error values on their validation sample sets. The validation datasets are obtained from a similar sampling area as the data used to develop the model. Ten random South African (SA) coal samples (available properties: GCV, proximate and ultimate variables) were used to calculate the standard error when the literature models were applied respectively.

Table 1-2: Literature GCV models tested with random coal samples

Location based literature GCV models		Reported error %	Tested error %
Model A: India [18]	$GCV[MJ/kg] = -0.03(A) - 0.11(M) + 0.33(VM) + 0.35(FC)$	1.49%	3.68%
Model B: USA [15]	$GCV[MJ/kg] = 91.4621 - 0.0556(M) + 0.025(V) - 0.9039(A) - 0.5687(C) - 0.6972(N) - 1.1252(O) - 0.8775(S)$	5.31%	46.32%
Model C: China [19]	$GCV[J/g] = 18960.16 - 225.27(A)$	3.20%	72.05%

Location based literature GCV models		Reported error %	Tested error %
Model D:SA [16]	$GCV [MJ/kg] = 37.927 - 0.076(VM) + 1.830(IM) - 0.338(A) + 0.0014(VM^2) + 0.203(IM) - 0.027(IM)(A)$	2.11%	8.22%
Model E: Turkey [20]	$GCV [MJ/kg] = 33.078 - 0.72(M) + 0.012(M^2) - 1.163(M^3) - 0.324(A^4)$	3.94%	2.30%

The literature models in Table 1-2 predict the gross calorific value (GCV) with the ash (A), moisture (M), inherent moisture (IM), volatile matter (VM), fixed carbon (FC), carbon (C), nitrogen (N), oxygen (O), sulphur (S) and hydrogen (H) content of coal.

The test results showed that a conclusive deduction could not be made. The calculated errors varied from errors smaller than the reported errors to values exceeding the reported error. However, it is evident that the models act differently on a single sample set. Even Model D, which was originally developed on SA coal, did not give the expected results. It would therefore be difficult to choose a suitable option from existing literature models and be confident that it is accurate.

Therefore, it is suggested to develop a new model on the specific set of data for which one wishes to predict the GCV by capturing the specific properties of that coal and using them for estimations. However, the question arises “*how do you develop a GCV model?*” The default answer will be to use the same method as in literature to develop such new model. This will allow for the comparison between the newly developed model and those from literature.

The methods used within 12 literature models were surveyed for commonalities and differences. The main steps that these models followed were highlighted in order to replicate the development process. Table 1-3 indicates the different aspects of the methods that were followed in the literature.

Table 1-3: Main methods followed by literature models

Referencing GCV model	Data preparation		Model development		Model validation	
	Obtain a representative sample	Analyse the data for outliers	Develop model for specific application	Evaluate the performance of the model	Validate the model on new data	Give verdict on the application of the model
Literature model 1: Majumder et al. (2008) [18]	X	X	X	X	X	X

Referencing GCV model	Data preparation		Model development		Model validation	
	Obtain a representative sample	Analyse the data for outliers	Develop model for specific application	Evaluate the performance of the model	Validate the model on new data	Give verdict on the application of the model
Literature model 2: <i>Patel et al. (2007)</i> [21]	X		X	X	X	X
Literature model 3: <i>Gulec & Gulbandilar</i> (2018) [22]	X		X	X	X	
Literature model 4: <i>Parikh et al. (2005)</i> [23]	X		X	X	X	
Literature model 5: <i>Behnamfard & Alaei</i> .(2016) [12]	X		X	X	X	X
Literature model 6: <i>Nhuchen & Afzal</i> (2008) [24]	X		X	X	X	X
Literature model 7: <i>Verma et al. (2010)</i> [15]	X	X	X	X	X	
Literature model 8: <i>Mesroghli et al. (2009)</i> [25]	X	X	X	X	X	X
Literature model 9: <i>Huang et al. (2008)</i> [19]	X		X	X	X	X
Literature model 10: <i>Akkaya et al. (2009)</i> [20]	X		X	X	X	X
Literature model 11: <i>Mason & Gandi</i> (1994) [26]	X		X	X	X	X
Literature model 12: <i>Kock & Franzidis</i> (1973) [16]	X		X	X		

Table 1-3: Main methods followed by literature modelsThe different elements within Table 1-3 were grouped under three main areas. These areas were common within the different models even though the underlying detail differed. The three areas are data preparation, model development and model validation. The sub-steps that were followed by the various models were grouped under the relevant focus areas, aiding in identifying the general approach that was followed by all of the models.

However, from Table 1-3 it can be noted that the models did not follow the same approach to develop the respective models. The survey highlights the steps that were not included or not explicitly mentioned for each model, increasing the difficulty of identifying a common generic approach. Without such approach it remains difficult to objectively compare the models and the results thereof.

The different approaches increase the uncertainty related to the energy content of coal. The need therefore exists to devise a general approach that can be used to develop GCV models on new datasets, while incorporating the different properties that are quantifiable through specific analyses.

1.3 Problem statement

Industries worldwide use coal as a primary energy source for multiple processes. These industries rely on the energy content of the coal for energy management and reporting. The type of coal and other chemical properties influence the energy content of the coal that is used. Measuring the energy content of coal is a time-consuming and expensive process and therefore not routinely done. Alternative energy quantification methods are therefore a necessity.

Alternative methods include models that predict the energy content of coal based on the coal properties as determined with proximate and ultimate analyses. These analyses are obtained easily and are usually readily available. However, such models were developed for coal from different locations and with varying properties. The assumption is that these models might not work on coal from various locations. The following problems are thus highlighted:

- Will these models be applicable to any coal dataset?
- If applicable, which one of these models would be preferred for a specific coal dataset?

To address these problems a new developed model for a specific coal dataset is required that can be used as reference for the comparison. However, there is no clear guideline to the method that needs to be followed to develop a GCV model. Different methods are currently used by the literature models and hinder the ability to objectively compare the models with each other.

In short, there is a need to:

- devise a general approach to develop a new model for a specific coal dataset,
- objectively assess and compare the new and existing models to identify a preferred model.

1.4 Objectives and scope of investigation

1.4.1 Objectives

The primary objective of this study is to devise a general approach that can be used to **develop GCV models** on new datasets and objectively **assess and compare the results** from various models. In order to achieve this, the following objectives will be met:

- providing relevant guidelines regarding the model development process,
- providing relevant research regarding techniques for comparison,
- devising a general approach for GCV model development,
- verifying the approach by applying it to three case studies, and
- objectively assessing and comparing the performance of the various models on a new dataset.

By meeting these objectives the study will identify whether it is necessary to develop new models for specific datasets. Furthermore, the study will provide a general approach to model development for comparison.

1.4.2 Scope of investigation

Chapter 1: Introduction and background

This chapter presents the introduction and background to this study. The problem statement section emphasises the need for the study. The objectives that must be met throughout the course of this study are also detailed in Chapter 1.

Chapter 2: Literature study

This chapter presents a study of the relevant literature regarding the main ideas to model development. Firstly, a general approach to model development is discussed. This is then followed by the three main component discussion: data preparation, model development and model validation. This chapter ends with the visualisation and comparison research done on how to present data visually and how to compare different results with each other.

Chapter 3: Methodology

In this chapter the approach to develop a new GCV model is established. The approach consists of three main steps identified from Chapter 2; data preparation, model development and model validation. The approach also includes a step to allow for the visualisation and comparison of the results.

Chapter 4: Results and discussion

In Chapter 4 the methodology is verified by applying it to actual case studies. Three case studies are discussed in detail. These case studies vary according to the types of coal and the geographical locations. New models were developed using the approach for each case study. These models were compared with each other and with the literature models to ultimately highlight a potential best model.

Chapter 5: Conclusion and recommendations

Chapter 5 provides a summary of the conclusions made from this study. This chapter refers back to the objectives stated in Chapter 1 to prove that all the objectives were met. Furthermore, recommendations for further studies are proposed in this chapter.

1.5 Conclusion

In this chapter, it was identified that coal is a main energy source used by many industries. These industries require the energy content of the coal, which is not always available, for efficiency purposes. Several models exist in literature that can predict this energy content, but the applicability of these models is questionable on new datasets.

It was suggested to evaluate the literature models performance on new coal data. The performance of these models on new data can be compared with a model developed specifically for that new coal dataset. An overview on the literature models approach to model development is given. This overview established the need to devise a general approach to develop and assess GCV models.

The evaluation and comparison will ultimately address the problem statement highlighted in this chapter: **Will the literature models be applicable to a new coal dataset and which one would be best?** The provided objectives devise a best practice methodology that can be used to predict the GCV and to compare the various developed models with each other.

CHAPTER 2–LITERATURE STUDY

2 Literature study

2.1 Preamble

It was suggested in Chapter 1 to develop a new model to predict the GCV of coal, since the applicability of the existing literature models on new data is questionable. This chapter will focus on available literature regarding general model development and the visualisation and objective comparison thereof.

A best practice approach for modelling (Section 2.2) will be evaluated. This evaluation will show the main steps to use when developing a new model. Each of the main steps during the model development process will then be discussed in detail. These include: dataset preparation (Section 2.3), model development (Section 2.4) and model validation (Section 2.5). Finally, visualisation techniques for model comparison (Section 2.6) will be discussed.

The research in this chapter will be used to devise a general approach to develop and assess GCV models (Chapter 3).

2.2 Best practice approach to model development

2.2.1 Overview

Statistical modelling is a mathematical technique to approximate reality and to make predictions from this approximation. Building a statistical model involves constructing a mathematical description of some real-world phenomena that accounts for the uncertainty and randomness involved in that system [27]. Such models are used for various applications, e.g. improving risk management in the banking sector, predicting consumer purchase habits and detecting fraud [28]. Similar to these predictions, statistical models can be used to predict the GCV of coal based on its chemical properties [29].

There are various existing methods to develop new models with unique characteristics [30]. However, certain key characteristics within these methods are commonly included when developing new models. It is thus important to understand what these common characteristics are and the relevance of each.

In this section, an existing best practice approach to develop models will be investigated. This approach is continually used as a reference point for the model development process, aiding in the identification of the common steps that are included within GCV models found in

literature. This will be done to ultimately identify the steps and characteristics that should be included to develop GCV models.

2.2.2 General process steps

It is necessary to identify the general steps within the modelling process [31], ensuring that these steps are universally included and necessary. Research on the general modelling process was found within the actuarial field of study [28]. This research looked at numerous techniques and presented a comprehensive methodology.

Basic techniques for statistical modelling that have a general application, was published in the Forecasting and Futurism journal of July 2013 [28]. An approach was compiled with the purpose of better understanding predictive modelling and of preparing for a modelling approach. This approach is given in the flow diagram in Figure 2-1.

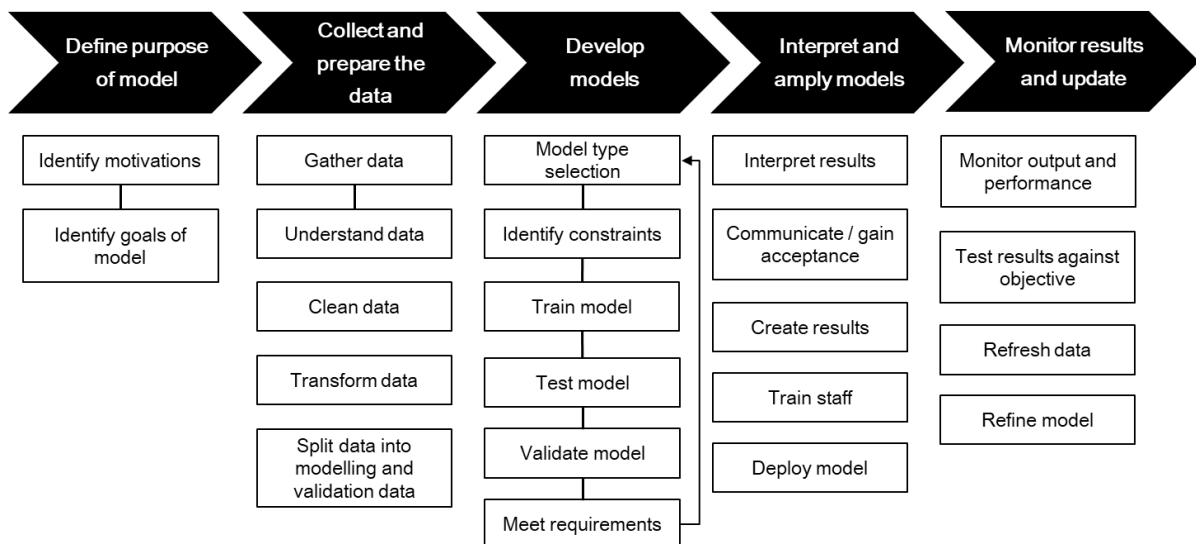


Figure 2-1: Best practice approach for model development [28]

Richard Xu [28], the author of this approach, believed that if a modeller follows procedures that had been proved to be effective, mistakes can be avoided and the final model will have a better likelihood of succeeding [32]. These procedures were captured with main focus areas that consisted of sub-components.

The approach is based on five main focus areas. The first focus area in his diagram is to define the purpose of the model. This is the basic background step where one decides why a model is required and what the model should achieve. For GCV models, this step is similar for all of the models, since the aim is to use a set of input variables to predict the GCV of a type of coal [18].

The second focus area in the approach is to collect and prepare the data that will be used to develop the model. This focus area ensures that the information that is used as input is free of errors that could influence the model [6]. This ensures that the models developed in the next focus area are fundamentally built on high quality data [33].

The third focus area gives attention to the development and validation of the model. The aim is to ultimately present a model that will be statistically correct and performs accurately [34].

The fourth and fifth focus areas in the approach are dedicated to the actual implementation of a model on real data. The fourth step focusses on the interpretation of the results to ensure that the model will be feasible to use. Continuous improvement of the model is emphasised and captured within the last step, presenting a developed model that is specific and of high quality for future use.

To test the approach, the various GCV models found in literature were evaluated to see if the same main steps were followed. Table 2-1 shows the main focus areas of Xu's approach [28], and indicates whether or not the GCV models followed these steps.

Table 2-1: Evaluating literature models with Xu's best practise approach [28]

Referencing GCV model	Focus area 1	Focus area 2	Focus area 3	Focus area 4	Focus area 5
Relative question asked?	Are motives given on the objectives and desired outcomes?	Data obtained in an orderly manner?	Develop a model with reasonable accuracy?	Were models validated against a set criteria?	Were models continuously updated with latest data?
Literature model 1: <i>Majumder et al. (2008)</i> [18]	X	X	X	X	
Literature model 2: <i>Patel et al. (2007)</i> [21]	X	X	X	X	
Literature model 3: <i>Gulec & Gulbandilar (2018)</i> [22]	X	X	X	X	
Literature model 4: <i>Parikh et al. (2005)</i> [23]	X	X	X	X	
Literature model 5: <i>Behnamfard & Alaei. (2016)</i> [12]	X	X	X	X	
Literature model 6: <i>Nhuchen & Afzal (2008)</i> [24]	X	X	X	X	
Literature model 7: <i>Verma et al. (2010)</i> [15]	X	X	X	X	
Literature model 8:	X	X	X	X	

Referencing GCV model	Focus area 1	Focus area 2	Focus area 3	Focus area 4	Focus area 5
Relative question asked?	Are motives given on the objectives and desired outcomes?	Data obtained in an orderly manner?	Develop a model with reasonable accuracy?	Were models validated against a set criteria?	Were models continuously updated with latest data?
<i>Mesroghli et al. (2009)</i> [25]					
Literature model 9: <i>Huang et al. (2008)</i> [19]	X	X	X	X	
Literature model 10: <i>Akkaya et al. (2009)</i> [20]	X	X	X	X	
Literature model 11: <i>Mason & Gandi (1994)</i> [26]	X	X	X	X	
Literature model 12: <i>Kock & Franzidis (1973)</i> [16]	X	X	X	X	

A first glance of Table 2-1 shows that all of the models followed the same approach. The only difference is observed with regards to the last focus area. The last focus area in Xu's approach [28] is specifically included for simulations and models that use real time data, such as within the actuarial science industry [28]. Since the GCV models are developed and tested on a set of existing data, this focus area is probably insignificant for this study.

It was previously stated that the first focus area of the approach is similar for the models, since the motivation is to calculate the GCV [35]. This indicates that the first and last focus areas of this approach are respectively generic and specific in the context of this study. These focus areas could thus be excluded from a specific approach for GCV models, highlighting the three focus areas in the middle as potential points of interest that should be investigated further.

2.2.3 Summary

Xu [28] compiled a best practice approach for model development, providing a reference point to compare and identify the steps taken by GCV models found in literature. The steps taken by the literature models are similar to the focus areas provided by Xu's approach [28]. However, a difference was observed for the last focus area. This is because the last focus area is specifically applicable for real time development of models and since the GCV models are based on a static set of data, this is of less importance.

Three main focus areas were highlighted from the comparison. These are the various different methods used (1) to prepare the data, (2) to develop the models and (3) to validate the models. Each of these focus areas will therefore be discussed in more detail within the next sections.

2.3 Dataset preparation

2.3.1 Overview

Data is not without practical implementation problems and therefore complicates the process of developing a statistical model [36]. It is vital to deal with these implementation problems in the data to develop successful models. Before the modelling process can commence, it is important to gather a reliable and qualitative dataset [35]. Not all data can be classified as being of high quality. It is therefore necessary to prepare a consistent dataset [37].

In order to obtain a consistent dataset, Xu [28] gave four sub-steps in his approach. These sub-steps are:

1. Gather and understand data
2. Clean data
3. Transform data
4. Split data into modelling and validation data

Each of these sub-steps will be investigated individually, firstly to see what the literature GCV models did and secondly to choose a suitable option for the overall approach.

2.3.2 Gather and understand data

Xu [28] explains that in this sub-step, data needs to be obtained from reliable sources, ensuring the quality of the data [28]. The quality of the data depends greatly on the type of data obtained, where the data is from and what the data includes or excludes [38]. The sampling procedure, chemical analyses and testing of the coal samples should thus be according to the ASTM standards [8].

To comply with ASTM standards the data used should be representable [5]. A representable dataset gives assurances that all of the possible properties are included. For a representable coal dataset it would thus be vital to ensure that the sampling data used covers a wide range of terrestrial area. This usually requires the integration of various samples and datasets [8].

A general rule for combining data is to ensure similarity in the multi-source datasets [8]. For coal data obtained from multiple sources the base of analyses should be similar. This means

that a source containing proximate analysis data on a specific base cannot be directly integrated with a source containing data measured on another base. Conversion from one basis to another should instead be performed using mass balance equations [8]. This will allow for similarity between the source's data.

Conversions between the various bases are presented in Table 2-2, where IM and TM present the type of moisture (in weight percentage) present in the coal. There are different types of moisture in coal [39]. Coal is mined wet due to the interaction with the groundwater and therefore would consist of total moisture (TM). This water can be readily evaporated at room temperature, leaving the coal with only the moisture within the coal itself. This moisture is known as the inherent moisture (IM).

Table 2-2: Conversions between different coal bases [39]

Current base:	Multiply current base by:	To obtain:		
		Air-dry	Dry-base	As-received
As-received		$\frac{100 - \text{IM}\%}{100 - \text{TM}\%}$	$\frac{100}{100 - \text{TM}\%}$	-
Air-dry		-	$\frac{100}{100 - \text{IM}\%}$	$\frac{100 - \text{TM}\%}{100 - \text{IM}\%}$
Dry-base		$\frac{100 - \text{IM}\%}{100}$	-	$\frac{100 - \text{TM}\%}{100}$

Table 2-2 allows for the conversion of any of the four proximate variables on a specific base to another. A proximate variable is converted by multiplying the value at the current base (rows) with the formula given to obtain the value at the desired base (columns). For example, a variable converted from an as-received base (row 1) to an air-dry base (column 1) would be multiplied by the formula in the upper left corner (Table 2-2) to get the new value.

Once a complete dataset is compiled, the data can be further inspected to ensure that it does not include any incorrect data or missing values [6]. The data should further be representable [5]. A representable sample is a sample that is representative of the environment it is taken from. The samples should therefore cover a wide range of terrestrial area. Finally, the number of samples obtained should also be sufficient for modelling purposes [40].

Table 2-3 shows the steps the literature GCV models followed to *gather and understand* the data.

Table 2-3: Gather and understand data of GCV models

Referencing GCV model	Type of data included	Basis of Analysis	Number of samples	Source of data	Sample location	Methods used to gather data
Literature model 1: <i>Majumder et al. (2008)</i> [18]	- Proximate analysis - GCVs	As-received	250	Manual sampling using ASTM procedures	Central India	- Mentioned that errors can occur due to difficulty in producing a representative sample.
Literature model 2: <i>Patel et al. (2007)</i> [21]	- Proximate analysis - Ultimate analysis - GCVs	As-received & dry-base	79	CFRI, Dhanbad database	India	- Samples taken from six different coal-producing regions.
Literature model 3: <i>Gulec & Gulbandilar (2018)</i> [22]	- Proximate analysis - Ultimate analysis - GCVs	As-received	2500	Manual sampling using ASTM procedures	Turkey	- Samples taken of raw coal and drilled coal; - Samples taken from underground mines and open pit mines, periodically.
Literature model 4: <i>Parikh et al. (2005)</i> [23]	- Proximate analysis - GCVs	Dry-base	577	Manual sampling using ASTM procedures	International	- Different coal and lignite samples have been used to cover a wide variety of proximate analysis values.
Literature model 5: <i>Behnamfard & Alaei.(2016)</i> [12]	- Proximate analysis - GCVs	As-received & dry-base	270	U.S. Bureau of mines	United States	- Samples are obtained from different coalfields.

Referencing GCV model	Type of data included	Basis of Analysis	Number of samples	Source of data	Sample location	Methods used to gather data
Literature model 6: <i>Nhuchen& Afzal</i> (2008) [24]	- Proximate analysis - Ultimate analysis - GCVs	Dry-base	246	Manual sampling using ASTM procedures	Not mentioned	- Samples were taken from two different seams; - Ensured that samples are from the same geological age and history.
Literature model 7: <i>Verma et al. (2010)</i> [15]	- Proximate analysis - Ultimate analysis - GCVs	As-received	4540	U.S. Geological Survey Coal Quality database	United States	- Excluded all samples with A content higher than 50%; - Excluded all samples with a total proximate / ultimate variables higher than 100.
Literature model 8: <i>Mesroghli et al. (2009)</i> [25]	- Proximate analysis - Ultimate analysis - GCVs	As-received	4540	USA open-file report 97-134 [Bragg et al. 2009]	United States	- Excluded all samples with A content higher than 50%; - Excluded all samples with a total proximate / ultimate variables higher than 100.
Literature model 9: <i>Huang et al. (2008)</i> [19]	- Proximate analysis - GCVs	Dry-base	222	Manual sampling using ASTM procedures	China	- Samples are obtained from different coalfields.
Literature model 10: <i>Akkaya et al.</i> (2009)[20]	- Proximate analysis - GCVs	As-received	270	As published by Palmer et al.[41]	Turkey	- Nationwide coal sampling – covering a wide range of terrestrial area, approximately 110'000km².

Referencing GCV model	Type of data included	Basis of Analysis	Number of samples	Source of data	Sample location	Methods used to gather data
Literature model 11: <i>Mason &Gandi (1994)</i> [26]	- Proximate analysis - Ultimate analysis - GCVs	Dry-base	775	Coal Research Section of Pennsylvania state University	United States	Only mentions that the data represents a large deposit of coal.
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	- Proximate analysis - GCVs	Dry-base & ash-free	250	41 Annual Report of Fuel Research Board (1971)	South Africa	No mention of data handling other than the sample size and source.

The 12 literature models discussed in Table 2-3 obtained data either by taking manual samples or by gathering samples from a nationally recognised database. The models that used manual sampling (literature model 1, 3, 4, 6 and 9) followed ASTM sampling procedures and ensured that the samples are representative of the coalfields they were taken from.

In the most cases, the literature CV models were developed on a set of data obtained from a nationally recognised database, instead of through manual sampling itself. In these cases, it is very important that the sources are reliable and that the data is of good quality since errors can be easily found in data [42]. This is especially true if the data is obtained from an unknown source or multiple different sources. In these cases, data can be mismatched, invalid or include missing and incorrect data [43]. The data quality can thus be checked and controlled by data validation and data verification [43].

Data validation is a process that follows prescribed rules about the value of the elements, such as the type of data, the range of values, missing values and consistency [44]. Data verification is a process in which different types of data are checked for accuracy and consistency after the data entry is completed (total checks for completeness, reconciliation of data sources, source comparison and consistency with different datasets) [45].

Most of the literature models used a dataset containing at least 200 samples. The number of samples used for modelling is important since the result of the developed model will be more statistically significant if a big enough sample size is used [43]. Several tests exist that give the minimum required number of samples for a desired model [46].

The simplest test that can quickly indicate whether a sample size is representable for modelling purposes is Green's method [47]. This method is specifically applicable for multiple regression analyses and is based on the number of predictable variables. For example, if all four proximate variables will be used in a model to predict the GCV, the number of predictors (p) will be four, and the required sample size (n) will be calculated as 84, as in Equation 2-1.

$$n = 50 + 8(p) \quad (2-1)$$

- n - Sample size
- p - Number of predictors

Cornish gave two other approaches that can also be used to determine the minimum sample size [48]. The first approach is precision-based, where the precision is known and the formula

is used to solve for the standard error (SE) instead. The formula for any standard error always contains the sample size, and thus, the sample size can be solved. This method includes various assumptions that need to be made, and since it is not that easy to use, it is not the preferred method.

Cornish's second approach is the power-based sample size calculation. This method is very useful since it allows the user to calculate the precision and degree of certainty as well. The limit of this method is, however, that it only relates to hypothesis testing, which can be difficult to use when a prediction model is required [43].

To ensure a high-quality dataset, the dataset will thus need to be checked for completeness by investigating the dataset for abnormalities such as missing data or erroneous data and to ensure that the sample is of a representative size [43]. For simplicity, Green's method will be used to test the required sample size.

2.3.3 Clean data

A representable dataset that is of high quality can still contain abnormalities or outliers [49]. Outliers represent abnormal events such as measurement errors or unplanned occurrences [37]. This data point deviates so much from the other observations that it arouses suspicions that it was generated by a different mechanism. For this reason, Xu [28] included the clean data sub-step into the approach. It is stated that this sub-step is one of the most important sub-steps in data preparation, since outliers can influence the final model tremendously [50].

Some of the literature models also emphasised the importance of outlier removal. Table 2-4 gives the literature models' approaches to remove outliers.

Table 2-4: Outlier removal of literature models

Referencing GCV model	Remove outliers?	Number of outliers removed:	Method used?
Literature model 1: <i>Majumder et al. (2008)</i> [18]	-		-
Literature model 2: <i>Patel et al. (2007)</i> [21]	-		-
Literature model 3: <i>Gulec & Gulbandilar (2018)</i> [22]	-		-
Literature model 4: <i>Parikh et al. (2005)</i> [23]	Yes	2	Intuition

Referencing GCV model	Remove outliers?	Number of outliers removed:	Method used?
Literature model 5: <i>Behnamfard & Alaei. (2016) [12]</i>	-		-
Literature model 6: <i>Nhuchen & Afzal (2008) [24]</i>	Yes	2	Intuition
Literature model 7: <i>Verma et al. (2010) [15]</i>	-		-
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	-		-
Literature model 9: <i>Huang et al. (2008) [19]</i>	Yes	27	Mahalanobis distance
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	Yes	6	-
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	-		-
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	-		-

Five literature models removed outliers during the development process and this was subsequently stated. Literature model 4 and 6 removed outliers through an intuition approach. This was appropriate to their studies, since both removed the data with an abnormally high ash content [9, 14]. This method, however, is not always preferable since it introduces the risk of “cherry picking” data [6].

According to Booyesen [33], there are other, more reliable existing statistical methods that can be used to identify potential outliers in a set of data. Booyesen introduces a new method that can be used to remove outliers. This method includes the process of investigating the leverages (horizontal differences between the scatter plot and the estimated least squares regression line) and residuals (vertical differences between the scatter plot and the estimated least squares regression line). The methodology combines simple statistical formulas with a visual presentation of data to deliver an intuitive evaluation process.

Booyesen’s method is similar to the Mahalanobis distance method used by literature model 9. The Mahalanobis distance is a statistical method that can be used for multivariable outlier detection [49]. This method works especially well with multivariable outlier detection, such as when several proximate analysis variables are being taken into consideration. This method, though, has a masking and swamping effect which can make it difficult to detect potential outliers as real outliers [51].

Several other, less complex methods exist that can also identify potential outliers [42]. These methods usually include the calculation of some referencing point, like the average of a set of values. Data points that are extremely larger than the average are then seen as outliers. Several of these functions already exist within programming platforms, such as Microsoft Excel™ and MATLAB, which makes it more user-friendly [31].

When a program such as MATLAB is used, caution must be taken when the outliers are automatically removed. It can happen that MATLAB identifies a point as an outlier, where in fact, it was a real data point. It would be recommended to only use the programming to identify the outliers [43]. The outliers can then be evaluated manually, and it can be decided whether the data point is an outlier or not.

One such a function that only identifies the potential outliers is the *is-outlier* function [32]. Figure 2-2 shows an example of a complete dataset that is evaluated in MATLAB for possible outliers. The green dotted line indicates the mean of the total sample set. MATLAB identifies all data points with 2 standard deviations from the mean as outliers, ensuring a 95% confidence interval. The green band shows all the data within three standard deviations from the mean.

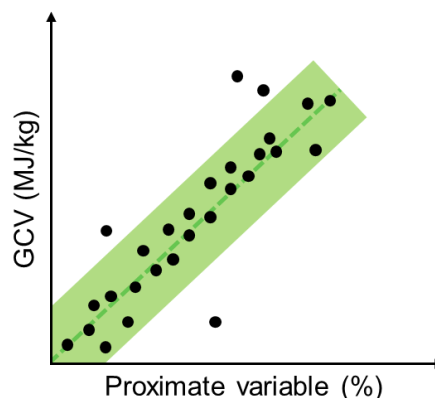


Figure 2-2: Outlier identification method

The potential outliers (the samples outside the green band) are investigated to ensure whether or not they are true outliers. The manual evaluation performed on the data points outside the green band (potential outliers) should be justifiable to prevent unnecessary removal.

Since programming platforms are easier to use when outliers need to be identified in a large set of data, it is recommended to use a platform, such as MATLAB, for the identification of potential outliers in a set of proximate variables.

2.3.4 Transform data

Data transformation is the application of a deterministic mathematical function to each point in a dataset [32]. Thus, each point, x_i , is replaced with the transformed value $y_i = f(x_i)$, where f is a function. Transformations are usually applied so that the data meets the assumptions of a statistical inference procedure, or to improve the interpretability or appearance of graphs.

Xu [28] included this step in his approach for when a non-normal distribution of data is used for modelling purposes such as linear regression. The concept is to firstly evaluate the normality of the gathered data. This can be done by calculating the mean and standard deviation and then by adding 1, 2 & 3 standard deviations from the mean to create a range. This can then be compared with the normal standard distribution that can be calculated in Microsoft Excel™ with the *norm.dist function*, to test each standard deviation value [24]. If the data is not normally distributed, it can then be re-expressed to make it closer to the normal [52].

Though this step is important in some scenarios, it is not always required. If the obtained data seems linear of nature and measured on the same scale, for instance, transformation is not required [53]. Since the models are all developed using the proximate variables or ultimate analysis, which is data that directly correlates to the desired dependent variable, GCV, this step would be unnecessary.

The literature models were investigated to see if any of them transformed their datasets to fit a normal distribution. Table 2-5 shows the literature models' approach to transformation.

Table 2-5: Transformed data of literature GCV models

Referencing GCV model	Transformed data?	Approach used:
Literature model 1: <i>Majumder et al. (2008)</i> [18]	No	-
Literature model 2: <i>Patel et al. (2007)</i> [21]	No	-
Literature model 3: <i>Gulec & Gulbandilar (2018)</i> [22]	No	-
Literature model 4: <i>Parikh et al. (2005)</i> [23]	No	-
Literature model 5: <i>Behnamfard & Alaei (2016)</i> [12]	No	-

Referencing GCV model	Transformed data?	Approach used:
Literature model 6: <i>Nhuchen& Afzal (2008) [24]</i>	No	-
Literature model 7: <i>Verma et al. (2010) [15]</i>	Yes	Axon transfer function
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	No	-
Literature model 9: <i>Huang et al. (2008) [19]</i>	No	-
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	No	-
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	No	-
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	No	-

Only one model was developed with transformed data. This was for an artificial neural network (ANN), which sometimes requires transformed data, especially if multiple variables would be included [11]. Most of the literature models did not transform their data. For this reason, this step seems unnecessary and could therefore be excluded from a modelling approach for a GCV model.

2.3.5 Split data into modelling and validation datasets

Xu's final sub-step in the data preparation focus area of the approach is to split the data into modelling and validation data [28]. This sub-step is important to ensure that the model can be validated. The data used for the modelling cannot also be used for validation, thus, a different set of data must be used to validate the model [38]. This validation data can be obtained in various ways. Table 2-6 shows how the validation dataset is obtained from the total set of data from the various literature models.

Table 2-6: Data splitting of GCV models

Referencing GCV model	Data split ratio		Method used
	Modelling	Validation	
Literature model 1: <i>Majumder et al. (2008) [18]</i>	164	86	-
Literature model 2: <i>Patel et al. (2007) [21]</i>	63	16	Randomly selected
Literature model 3: <i>Gulec & Gulbandilar (2018) [22]</i>	2100	400	-

Referencing GCV model	Data split ratio		Method used
	Modelling	Validation	
Literature model 4: <i>Parikh et al. (2005) [23]</i>	450	100	-
Literature model 5: <i>Behnamfard & Alaei.(2016) [12]</i>	270	27	New samples taken for validation
Literature model 6: <i>Nhuchen& Afzal (2008) [24]</i>	220	26	-
Literature model 7: <i>Verma et al. (2010) [15]</i>	4371	69	Randomly selected
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	2000	800	Randomly selected
Literature model 9: <i>Huang et al. (2008) [19]</i>	178	44	New samples taken for validation
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	143	33	-
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	700	75	Randomly selected
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	250	-	-

From Table 2-6 it can be seen that all of the literature models performed some split to obtain the validation dataset. However, none of the models mentioned a formal method of how the split was obtained.

Several structured methods exist that can be used to obtain this split in the data. A general rule of thumb is the 70-30 split, where 70% of the data is chosen at random and the remaining 30% is used for validation [46]. Although none of the literature models mentioned this method used, it can be seen that many of them have about 70% of their total dataset as the modelling data and the remaining 30% as the testing data.

This 70-30 rule was altered by Gupta [46], to have an even bigger split. He introduced the 30-40-30 split, where the 40% is used for modelling and the two sets of 30% are used to test the model twice. The benefit of this method over the normal 70-30 split is that the model can be validated twice, each time with a new validation set.

There are other methods that can be used as well, like cross-validation or the k-fold method [54]. The disadvantage of these methods is that they are somewhat time-consuming since it is required to do several repetitions to find the optimal split.

Since most of the literature models used a general rule of thumb method to split their data, it is safe to say that the 70-30 split or the improved split on that, the 30-40-30 split, will be sufficient to use for the modelling and validation of GCV models.

2.3.6 Summary

The quality of the data used to develop models is important. For this reason, Xu [28] introduced a data preparation step in his approach that consists of four important sub-steps. Each one of these sub-steps was discussed and the optimal methods to obtain this were chosen.

2.4 Model development

2.4.1 Overview

Once a quality dataset is compiled, the modelling data can be used to develop a model for the prediction of the GCV of coal [28]. To develop an effective and robust model, a solid foundation in statistics and practical experience in statistical applications are required. For this reason, several sub-steps were introduced in the model development step of Xu's approach [28]. These sub-steps are:

1. Model type selection
2. Identify variables
3. Develop and test models
4. Check requirements

Each one of these steps will be investigated and applied to the literature GCV models to compare their approach with Xu's approach [28].

2.4.2 Model type

The first step to model development is to decide what type of model should be developed [55]. There are different types of models, and usually more than one type can work on a given set of data [28]. It is important to choose the model that will fit the data the best. To ensure that the right model is chosen, it would be best to start by looking at the directionality and relationship of the data. To do this, individual scatterplots for every independent variable in the analysis can be drawn. It is important to evaluate the scatterplots and to understand the significance thereof.

Majumder *et al.* [18] investigated the impact of a combination of the proximate analysis variables on the GCV of coal. Figure 2-3 shows his findings.

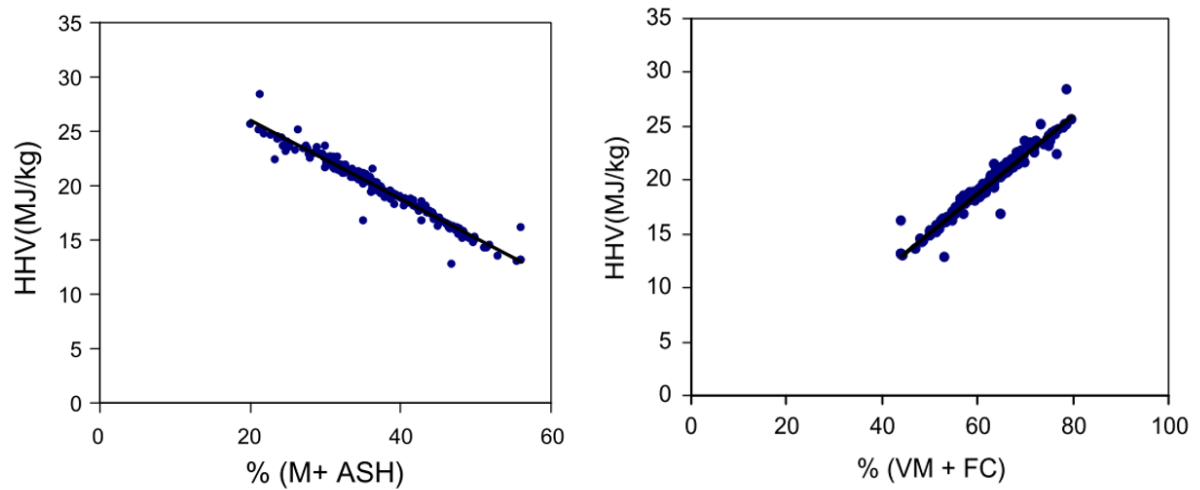


Figure 2-3: Effect of the proximate variables on the GCV of coal[18]

Majumder *et al.* [18] concluded that a linear relationship is obvious and therefore developed a linear regression model to predict a GCV model for coal based on the proximate variables.

However, Patel *et al.* [21] introduced the possibility of the existence of nonlinear relationships between some of the proximate and ultimate variables. Figure 2-4 shows his findings.

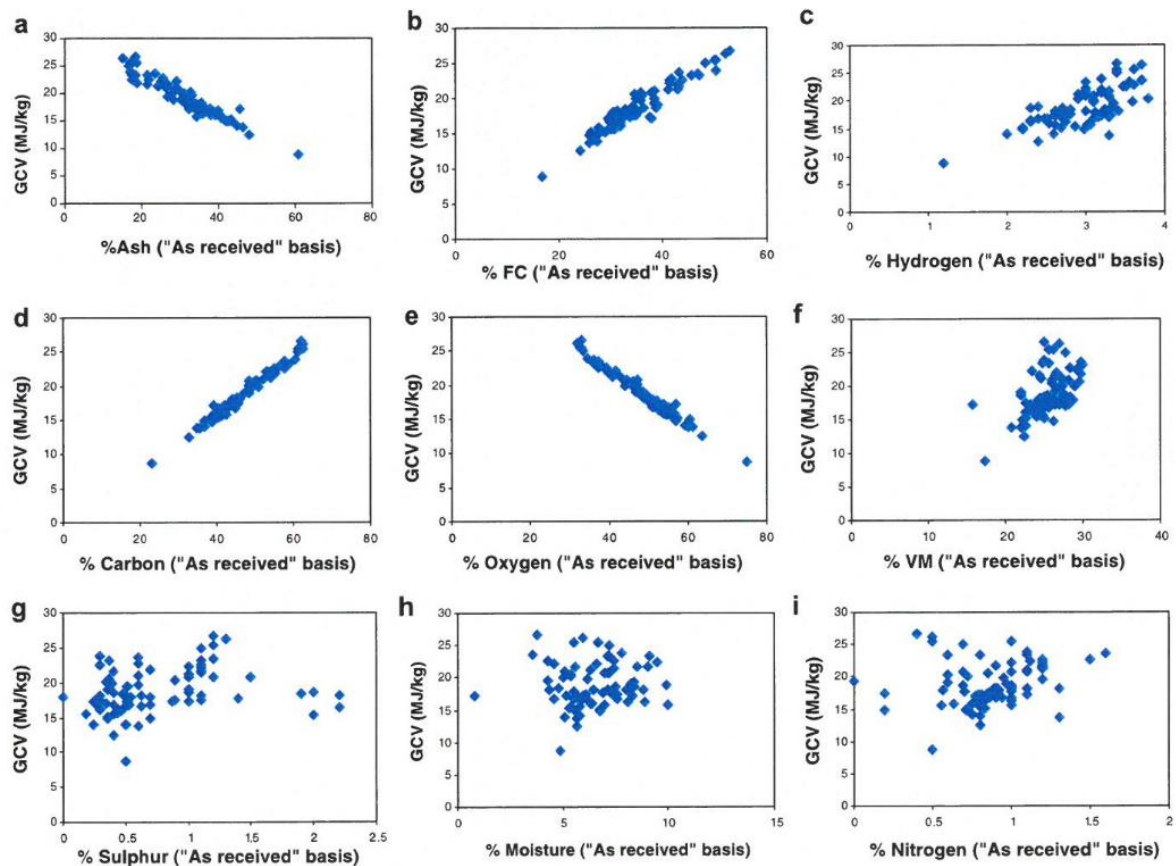


Figure 2-4: Effect of the proximate and ultimate variables on the GCV of coal [21]

He therefore proposed the use of an artificial neural network (ANN) as opposed to a regression model, to overcome the difficulty of choosing the correct form of the linear / nonlinear data-fitting function. Patel's [21] ANN model gave a slightly more accurate prediction (AAE =1.23 %) on Indian coal than the one Majumder *et al.* [18] (AAE = 1.49 %) obtained on Indian coal.

The performance of an ANN and a regression model for GCV predictions of coal were compared on multiple case studies by Mesroghli *et al.* [25]. The results from both models were very similar (>0.5% difference). It was concluded that for the problems with straightforward relationships between inputs and outputs, it is better to use a common and understood technique as regression instead of a more complicated method as ANN.

Most of the literature models are regression models (both linear and nonlinear). In some instances, ANNs and inference systems (IS) such as ANFIS and CANFIS were developed. Table 2-7 indicates the types of models that were developed for each of the 12 investigated literature models.

Table 2-7: Types of models developed by the literature GCV models

Referencing GCV model	Type of model(s) developed
Literature model 1: <i>Majumder et al. (2008) [18]</i>	- Multiple linear regression
Literature model 2: <i>Patel et al. (2007) [21]</i>	- ANN (<i>artificial neural network</i>)
Literature model 3: <i>Gulec & Gulbandilar (2018) [22]</i>	- Multiple linear regression - ANN
Literature model 4: <i>Parikh et al. (2005) [23]</i>	- Multiple linear regression
Literature model 5: <i>Behnamfard & Alaei.(2016) [12]</i>	- Multiple linear regression - ANFIS (<i>adaptive neuro-fuzzy inference system</i>)
Literature model 6: <i>Nhuchen & Afzal (2008) [24]</i>	- Multiple linear regression
Literature model 7: <i>Verma et al. (2010) [15]</i>	- Multiple linear regression - CANFIS (<i>co-active neuro-fuzzy inference system</i>)
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	- Multiple linear regression
Literature model 9: <i>Huang et al. (2008)[19]</i>	- Multiple linear regression - Multiple nonlinear regression - ANN
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	- Multiple nonlinear regression

Referencing GCV model	Type of model(s) developed
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	- Multiple linear regression
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	- Multiple nonlinear regression

For each literature model, the relationship of the data was always inspected before deciding what model type to develop. Most of the literature models developed only on proximate analysis variables identified a linear relationship and therefore developed a linear regression model. The nonlinear regression models and ANNs were developed when ultimate analysis variables were also taken into consideration [12, 22, 25].

Due to this observation it is possible that linear regressions are better if only proximate analysis variables will be taken into consideration, where nonlinear models should be considered if other variables will be included.

2.4.3 Identify variables

Once it is known what type of model will be developed, the variables that should be included in the model can now be selected. According to Xu [28], the core of the model is based on the selection of the right group of variables. It is important to consider all variables initially and to also look at the variables' interactions with each other [28]. This will allow the right group of variables that has the biggest effect on the model to be picked.

In literature, models were developed to predict the GCV of coals based on the proximate, ultimate and or a combination of other properties of coals. Table 2-8 shows the variables included in each of the developed GCV models.

Table 2-8: Variables included in the literature GCV models

Referencing GCV model	Group of variables	Specific variables in %
Literature model 1: <i>Majumder et al. (2008) [18]</i>	- Proximate analysis	- A, M, FC, VM
Literature model 2: <i>Patel et al. (2007) [21]</i>	- Proximate analysis	- A, M, FC, VM
	- Ultimate analysis	- H, N, O, S
Literature model 3: <i>Gulec & Gulbandilar (2018) [22]</i>	- Proximate analysis	- A, M, VM
	- Ultimate analysis	- S
Literature model 4:	- Proximate analysis	- A, FC, VM

Referencing GCV model	Group of variables	Specific variables in %
<i>Parikh et al. (2005) [23]</i>		
Literature model 5: <i>Behnamfard & Alaei.(2016) [12]</i>	- Proximate analysis	- A, M, FC
Literature model 6: <i>Nhuchen & Afzal (2008) [24]</i>	- Proximate analysis	- FC, VM
	- Ultimate analysis	- H, N, C
Literature model 7: <i>Verma et al. (2010) [15]</i>	- Proximate analysis	- A, FC, VM
	- Ultimate analysis	- C, N, O, S
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	- Proximate analysis	- A, M, FC, VM
	- Ultimate analysis	- C, N, O, S
Literature model 9: <i>Huang et al. (2008) [19]</i>	- Proximate analysis	- A
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	- Proximate analysis	- A, M, FC, VM
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	- Proximate analysis	- A
	- Ultimate analysis	- C, H, S
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	- Proximate analysis	- A, M, FC, VM

It is noted that the proximate variables are included in most cases, with lesser occurrences of the ultimate analysis. Majumder *et al.*[18] investigated several models in order to determine the optimal proximate and ultimate variables to include in the prediction of the GCV of coals.

The models that were included in the study consisted of only the proximate and ultimate analysis variables. Each property was discussed and it was concluded that the effect of all the proximate analysis variables should be included when developing a model to predict the GCV of coal.

2.4.4 Develop model

Now that it is certain what type of model is desired and what variables to include in the model, the model can be developed. The data selected as the modelling data in the data split sub-step can be used to develop the model. The investigated literature models are presented in Table 2.9.

Table 2-9: Investigated literature GCV models

Referencing GCV model	Model
Literature model 1: <i>Majumder et al. (2008) [18]</i>	$GCV [MJ/kg] = -0.03(Ash) - 0.11(IM) + 0.33(VM) + 0.35(FC)$
Literature model 2: <i>Patel et al. (2007) [21]</i>	ANN
Literature model 3: <i>Gulec & Gulbndilar (2018) [22]</i>	$GCV [kJ] = 8637.113 - 109.277(M) - 90.556(A) + 10.885(S) - 11.245(VM)$
	ANN
Literature model 4: <i>Parikh et al. (2005) [23]</i>	$GCV [kJ/kg] = 0.3536(FC) + 0.1559(VM) - 0.0078(A)$
Literature model 5: <i>Behnmfard & Alai.(2016) [12]</i>	$GCV [btu/lb] = 12204 - 170(M) + 46.8(FC) - 127(A)$
	ANFIS
Literature model 6: <i>Nhuchen & Afzal (2008) [24]</i>	$GCV[MJ/kg] = 0.1846(VM) + 0.3525(FC)$
	$GCV[MJ/kg] = 32.7934 + 0.0653(C^2) - 0.5321(C) - 2.8769(H) + 0.0608(CH) - 0.2401(N)$
Literature model 7: <i>Verma et al. (2010) [15]</i>	$GCV[MJ/kg] = 91.4621 - 0.0556(M) + 0.025(V) - 0.9039(A) - 0.5687(C) - 0.6972(N) - 1.1252(O) - 0.8775(S)$
	CANFIS
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	$GCV[MJ/kg] = 37.777 - 0.647(M) - 0.387(A) - 0.089(VM)$
	$GCV[MJ/kg] = -26.29 + 0.275(A) + 0.605(C) + 1.352(H) + 0.840(N) + 0.321(S)$
	$GCV[MJ/kg] = 6.971 + 0.269(C) + 0.195(N) - 0.061(A) - 0.251(O) + 1.08(H)$
Literature model 9: <i>Huang et al. (2008) [19]</i>	$GCV[J/g] = 18960.16 - 225.27(A)$
	$GCV [J/g] = 8765423.92 - 1029388.80(A) - \frac{43824257.34}{A} + 74155.88(A^2) + \frac{113019575.18}{A^2} - 3216.28(A^3) - \frac{82925424.38}{A^3} + 77.08(A^4) + \frac{108635322.66}{A^4} - 0.78(A^5) - \frac{58745733.52}{A^5}$
	ANN
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	$GCV [MJ/kg] = 33.078 - 0.72(M) + 0.012(M^2) - 1.163(M^3) - 0.324(A^4)$

Referencing GCV model	Model
	$GCV[MJ/kg] = 0.561(M)^{-6.137}(VM)^{0.381}(FC)^{0.626}$
	$GCV[MJ/kg] = 0.836(M)^{-8.155}(A)^{-3.559}(VM)^{0.35}(FC)^{0.626}$
Literature model 11: <i>Mason & Gandhi (1994) [26]</i>	$GCV [MJ/kg] = 0.472(C) + 1.48(H) + 0.1935(S) + 0.107(A) - 12.29$
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	$GCV [MJ/kg] = 37.927 - 0.076(VM) + 1.830(M) - 0.338(A) + 0.0014(VM^2) + 0.203(M) - 0.027(M)(A)$
	$GCV [MJ/kg] = -3.793 + 0.812(FC) - 6.212\left(\frac{FC}{VM}\right) - 0.00002(FC)^3 + 0.710\left(\frac{FC}{VM}\right)^2 - 0.028\left(\frac{FC}{VM}\right)^3$
	$GCV [MJ/kg] = 37.639 - 0.077(VM) - 1.254(M) - 0.052(A) + 0.0012(VM^2) + 0.145(M^2) - 0.00027(VM)(A) - 0.025(A)(M)$

Table 2-9 indicates that some of the literature models developed more than one model. These studies were explicitly done to compare the various modelling options. Consequently, a variety of models is presented. This section focusses on the two methods used to develop each of these types of models: regression models and soft computing models.

Regression models

Regression analysis is a statistical technique for demonstrating the relationship between independent and dependent variables [56]. This relationship can be linear or nonlinear [57] as illustrated by Figure 2-5.

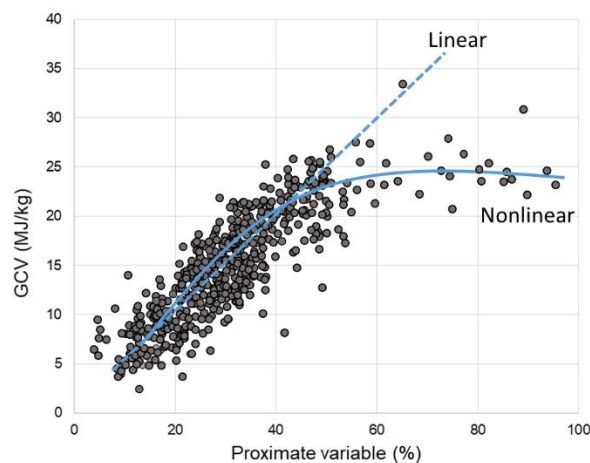


Figure 2-5: Distribution of data with a linear and nonlinear fit

For a multiple linear relationship, the data is expected to fit a straight line, and typically takes the form represented in Equation 2-2.

$$y = x_1X_1 + x_2X_2 + \dots + x_kX_k + C \quad (2-2)$$

- y - Output (dependent) variable
- X_i - Input (independent) variables
- x_i - Regressors (constants associated with each of the input variables)
- C - Residual error (constant value)

where x_i represents different allocated weights (regressor values) to each one of the input variables and C the residual error terms of the model [58].

In multiple nonlinear regression the data is modelled by a function which is a nonlinear combination of the dependent and independent variables [27]. A nonlinear regression model can be presented by Equation 2-3.

$$y = f(x_i, \beta) + C \quad (2-3)$$

- y - Output (dependent) variable
- x_i - Input (independent) variables
- $f()$ - Function describing the nonlinearity
- β - Vector with n parameters
- C - Residual error (constant value)

The function describing the nonlinearity given in Equation 2-3 can take many forms. The most commonly used nonlinear regressions are given in Table 2-10.

Table 2-10: Nonlinear regression types

Type of nonlinear regression	Typical form for n parameters
Polynomial regression	$y = \beta_0 + \beta_1x_i + \beta_2x_i^2 + \dots + \beta_nx_i^n + C$
Power regression	$y = ax_i^\beta$
Logistic regression	$y = \frac{1}{1 - e_i^{x\beta}} + C_i$

Where β is a vector with n parameters, x_i the independent variables and y the dependent variables.

Both the linear and nonlinear regression models can be developed in Microsoft Excel™, using the line statistics (LINEST) function [59]. This function is a complete linear least squares curve fitting routine that produces uncertainty estimates for the fit values.

There are two ways to access the LINEST functionality in Excel™: through the function directly and through the "analysis tools" set of macros. Both of these are built-in Excel™ functions and can be easily used [59].

The output of the LINEST function is a matrix that includes the regressors and residual error term, as well as some other statistics. An explanation of the LINEST function output (matrix) is presented in Table 2-11.

Table 2-11: LINEST function output (matrix)

LINEST matrix explanation				
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	v_2	#N/A	#N/A	#N/A
SS_{reg}	SS_{res}	#N/A	#N/A	#N/A

The first row of the matrix gives the regressors as well as the residual error term (C) that is required for the developed model represented in Equation 2-2 and Equation 2-3. It is important to evaluate the regressors of the model to see if the model is as expected. If all the regressors are zero, it means that the output will be completely dependent on the residual error value (C) irrespective of what the proximate variables are.

The evaluation of the regressors can also indicate the type of influence each of the regressors has on the output of the model (positive or negative). If the relationship between an input variable and the output variable is negative, it is expected that the regressor associated with this input value will also be negative, and vice versa.

The second row in the matrix gives the standard error of each regressor (SE_{x_i}). Each SE_{x_i} indicates the error amount that the associated regressors add to the standard error of the model [56].

The other statistics in the LINEST matrix give information on the predictability and accuracy of the model. These statistics are based on the same data used to develop the model with (the 40% modelling data). Table 2-12 gives a summary of these statistics.

Table 2-12: LINEST matrix statistics explanation [60]

Statistical item (Parameters)	Formula	Application
Coefficient of determination (R^2)	$\frac{\sum_{k=1}^p (\hat{y}_k - \bar{y})^2}{\sum_{k=1}^p (y_k - \bar{y})^2}$	- Range: $0 \leq R^2 \leq 1$ - If close to 1, model is good
Root mean square error (RMSE)	$\sqrt{\frac{SS_{reg}}{v_2}}$	- Smaller value means model predicted points are close to data points
F-value (F)	$\frac{SS_{reg}/v_1}{SS_{res}/v_2}$	- If greater than $F_{critical}$, then regression is justified with an α – confidence level
Degree of freedom for residuals (v_2)	$P - n$	- If zero then solution demands an equality instead of approximation
Sum of squares of regressor errors (SS_{reg})	$\sum_{K=1}^P (\hat{y}_k - \bar{y})^2$	- Indicates variability of regressed data
Sum of squares of residual errors (SS_{res})	$\sum_{K=1}^P (\hat{y}_k - y_k)^2$	- Indicates variability of the data

These statistics can be used to test the developed model (Equation 2-2). The parameters and variables in the formulas are listed in the Table of Abbreviations, and the full application can be found in the referencing source [60]. These statistics can typically be used to test if the model meets the desired requirements discussed in Section 2.4.5.

Soft computing models: Neural Networks and Fuzzy Logic Systems

Artificial neural networks (ANNs) are computer programs designed to simulate the human brain's processing ability. An adaptive neuro-fuzzy inference system (ANFIS) is a type of ANN based on a fuzzy inference system. A coactive neuro-fuzzy inference system (CANFIS) is an extended version of the ANFIS.

These soft computing models are powerful tools that have the ability to identify the underlying highly complex relationships between input-output data [18]. For this reason, several of the literature models investigated developed such models to improve the accuracy of the GCV prediction [15, 16, 18, 19, 22, 24, 25].

The neural network code of MATLAB (R201a) mathematical software was used by most of the literature models to devise soft computing models. This code uses the best possible combination of input variables (with specific weights allocated to each variable) to predict a GCV using training data. Validation data is then ultimately used to test the model and the results are visually portrayed. Figure 2-6 and Figure 2-7 show the error percentage output and the R^2 graphs given by MATLAB.

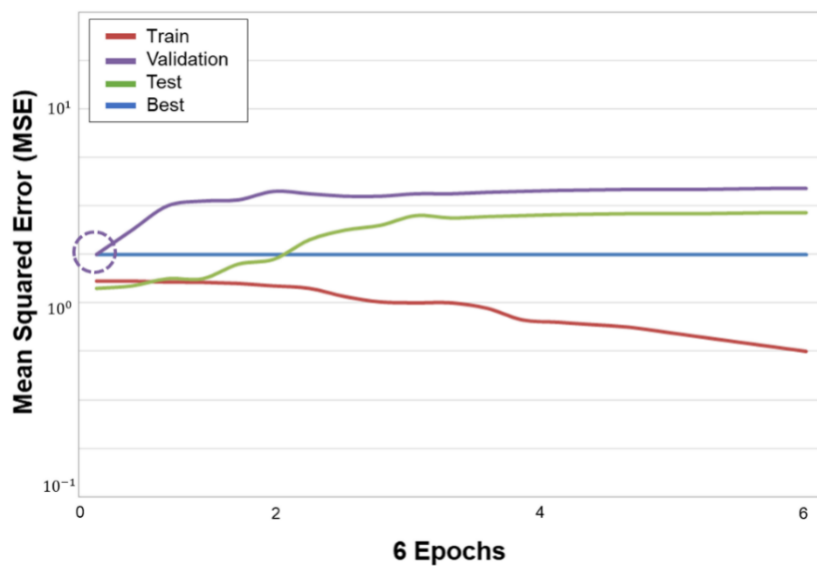


Figure 2-6: MATLAB results for error percentages

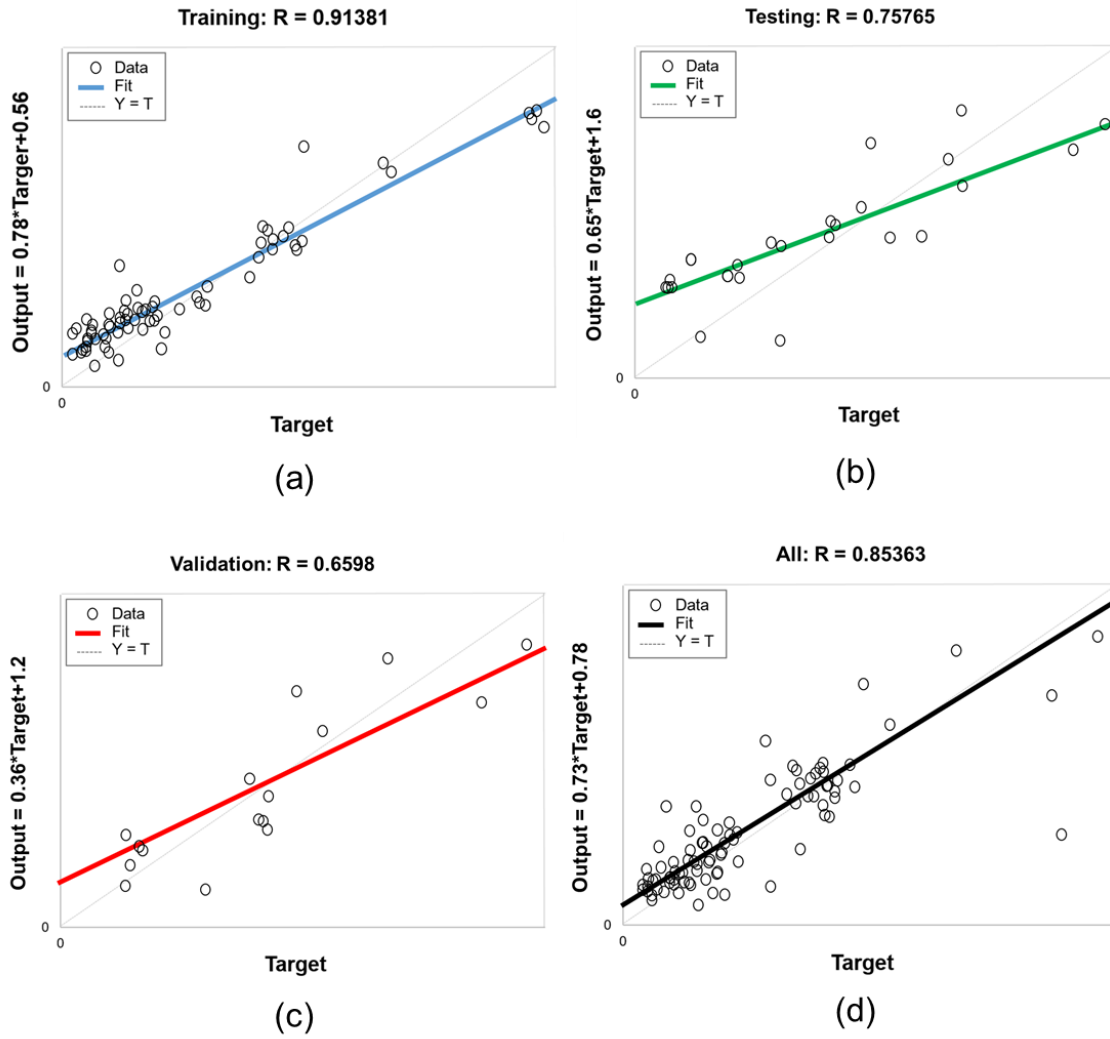


Figure 2-7: MATLAB results for model fit

The results from these graphs are interpreted to test and validate the models in the following sections (Section 2.4.5 and Section 2.5).

2.4.5 Check requirements

Once the model is developed, it should be tested to check that it meets the basic requirements. The aim of this step is to quickly see if the developed model will be adequate. Xu [28] included this sub-step in the model development step to keep the modeller from spending time on the model validation step if the model is not good enough to begin with [28]. It is important to note that during this sub-step the modelling dataset is still used (and not the validation dataset).

Xu [28] claims that there are various statistics that can give information on the credibility of the developed model [61]. Akkaya *et al.* [20] stated that the R^2 , RMSE and MAE values on the

modelling data are good indicators of the GCV prediction accuracy of the model. These three parameters are discussed in detail.

Coefficient of correlation (R^2)

The R^2 is the statistical measure of how close the data is to the fitted line and is almost always included in the testing of a regression model [58]. The R^2 value is always between 0% and 100%, where 0% indicates that the model explains none of the variability of the response data around the mean. On the other hand, 100% indicates that the model explains all the variability of the response around its mean [42]. In general, the higher the R^2 value, the better the model fits the data. The R^2 values are given by the LINEST matrix and are explained in Table 2-11

Root mean square error (RMSE)

Xu [28] suggested to use the RMSE value to test the models, although literature models mainly used the MAE. Both the MAE and the RMSE express average model prediction error and are indifferent to the direction of the errors. Xu [28] explains that the RMSE squares the errors before they are averaged, which gives a relatively higher weight to larger errors [48]. This is thus more useful than the MAE when large errors are particularly undesirable. The RMSE value is given by the LINEST matrix output and is explained in Table 2-11.

Mean absolute error (MAE)

The MAE measures how far the predicted values are from the observed values. The MAE is not given by the LINEST matrix and can be calculated using Equation 2-2.

$$MAE = Average(|Y_{m,model} - Y_{p,model}|) \quad (2-2)$$

MAE	-	Mean absolute error of the model on the modelling data
$Y_{m,model}$	-	Measured values of modelling data
$Y_{p,model}$	-	Predicted values on modelling data

For this study the R^2 , RMSE and the MAE, calculated on the modelling data, will thus be included in the *check requirement* sub-step.

If these statistics fall within the desired ranges, the model is deemed fit for the prediction. If the statistics deteriorate much from the desired ranges, fitting is most likely the problem. It is then recommended to restart the model development step from the beginning and rather choose a different model type and / or variables to develop a new model [28].

2.4.6 Summary

Before a model can be developed, it is important to analyse the data that will be used for the model to understand the type of model that will be required. Once the type of model is selected, the variables can be chosen that will be used for the model.

The investigated literature models are mainly developed as a linear regression model; this is due to the linear relationship between the proximate analysis variables and the GCV. It was also investigated by a literature model study, that the use of all four proximate variables should be included to develop a GCV model.

2.5 Model validation

2.5.1 Overview

Validation is the task of demonstrating that the model is a reasonable representation of the actual system, i.e. that it is a *good* model [44]. What constitutes a *good* model is subjective, but a performance modelling point of view presents a way to make it more objective. Performance modelling judges the goodness of models based on how accurately measures extracted from the model correspond to the measures which would be obtained from the represented system [44], which is the difference between the predicted and measured values.

To validate the predictive ability of a model, data other than that used to develop the model is required [62]. For this sub-step, a set of validation data is required.

A model is inherently abstract when compared with the system it represents [44], hindering the potential to obtain an error free comparison. Xu [28] therefore discussed various steps to calculate the accuracy between the predicted and experimented results. These steps, as well as the steps followed by the literature models, will be discussed in this section.

2.5.2 Validation assessment

A model's performance is overestimated when it is tested on the same data used to construct the model [61]. Several steps have therefore been proposed to estimate the performance of the model on independent data (validation data) more accurately [62]–[67].

Xu [28] discussed the importance of the validation step at the hand of the MAE between the experimental and the predicted data and the comparison coefficient (CC). These statistical parameters are calculated on the *validation data* placed aside in Section 2.1 (data split) and it presents a quantitative measure of validation. It should be noted that the MAE calculated in

this sub-step is not the same MAE calculated in the check requirements sub-step. The data used in these two sub-steps differs.

Mean absolute error of the model on the validation data ($MAE_{(val)}$)

This MAE between the experimental and the predicted data is as the name suggests – the mean of the absolute errors. The absolute error is the absolute value of the difference between the predicted value and the experimental value, indicating what a probable error could be when the specific model is used. This can be calculated using Equation 2-4.

$$MAE_{(val)} = \text{Average}(|Y_{m,val} - Y_{p,val}|) \quad (2-4)$$

- $MAE_{(val)}$ - Mean absolute error of the model on the validation data
- $Y_{m,val}$ - Measured values of validation data
- $Y_{p,val}$ - Predicted values on validation data

Comparison coefficient (CC)

The comparison coefficient (CC) measures the strengths of a linear comparison between the experimented and the predicted values [68]. It ranges from 0 to 1, with a value close to 1, the closer the absolute value to 1, the stronger the relationship. A relationship of zero indicates that there is no relationship [37]. Figure 2-8 indicates the desired relationship between the predicted and experimental values, with a CC of 1.

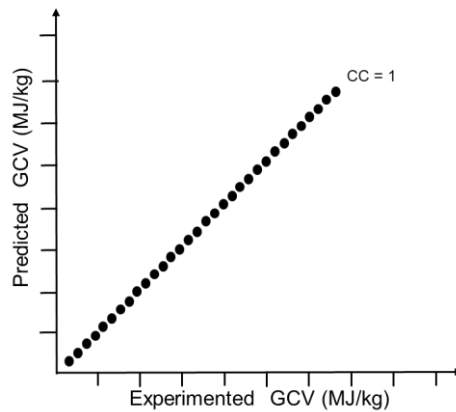


Figure 2-8: Desired relationship between the predicted and experimental GCVs

These two validation criteria (MAE and CC) are also used by the literature models to validate their models. Some of the literature models included other validations, as can be seen in Table 2-13.

Table 2-13: Validation procedures of literature GCV models

Referencing GCV model	Validation procedure
Literature model 1: <i>Majumder et al. (2008)</i> [18]	- Coefficient of multiple determinations (R^2) for model - Mean absolute error (MAE)
Literature model 2: <i>Patel et al. (2007)</i> [21]	- Comparison coefficient (CC) - Root mean square error (RMSE) - Mean absolute error (MAE)
Literature model 3: <i>Gulec & Gulbandilar (2018)</i> [22]	- Coefficient of multiple determinations (R^2) for model - Comparison coefficient (CC) on validation - Root mean square error (RMSE) - Mean absolute error (MAE)
Literature model 4: <i>Parikh et al. (2005)</i> [23]	- Coefficient of multiple determinations (R^2) for model - Average absolute error (AAE) - Average bias error (ABE)
Literature model 5: <i>Behnamfard & Alaei. (2016)</i> [12]	- Coefficient of multiple determinations (R^2) for model - Mean absolute error (MAE) - P-value (at $\alpha = 0.95$) - Mallows' Cp
Literature model 6: <i>Nhuchen & Afzal (2008)</i> [24]	- Coefficient of multiple determinations (R^2) for model - Mean absolute error (MAE) - Average bias error (ABE)
Literature model 7: <i>Verma et al. (2010)</i> [15]	- Coefficient of multiple determinations (R^2) for model - Comparison coefficient (CC) on validation - Mean standard error (MSE)
Literature model 8: <i>Mesroghli et al. (2009)</i> [25]	- Coefficient of multiple determinations (R^2) for model - Standard error (SD) - Mean absolute error (MAE)
Literature model 9: <i>Huang et al. (2008)</i> [19]	- Coefficient of multiple determinations (R^2) for model - Comparison coefficient (CC) on validation - Root mean square error (RMSE)
Literature model 10: <i>Akkaya et al. (2009)</i> [20]	- Coefficient of multiple determinations (R^2) for model - Standard error of estimation (SEE) - Standard squared error (SSE) - Mean absolute error (MAE)
Literature model 11: <i>Mason & Gandi (1994)</i> [26]	- Standard deviation of error (SDE) - Mean absolute error (MAE)
Literature model 12: <i>Kock & Franzidis (1973)</i> [16]	- Coefficient of multiple determinations (R^2) for model - Standard error (SE)

The validations that the literature models included in their study are all very common in many practices. A study done on baseline development for the Section 12L Income Tax Act by Hamer [69] discussed all of these validation methods. Here follows the list he gives of common indicators as well as referencing literature relating to its application.

- Coefficient of determination [6,71,72]
- Standard error [50]
- Average error [37]
- Root-mean squared error (RMSE) [51,71]
- Coefficient of variance of RMSE [71]
- Mean bias error [37]
- Absolute and relative precision [50]
- F-statistic and degree of freedom [71]
- T-statistic and relevance of independent variables [71]
- Error frequency distributions [37]
- Cumulative sum of deviations (CUSUM) [37]

These common indicators present additional criteria to validate a model. However, the two parameters (MAE, CC) are used for validation in most of the literature models, indicating that these parameters are most likely practical and easy to determine, which allows them to be used for general purposes and are not limited to specific situations.

2.5.3 Summary

In this section, Xu [28] introduced the validation of the developed model in his approach.

Several criteria are investigated that can be used to validate a model. The mean absolute error (MAE) and the comparison coefficient (CC) can be identified as the most popular methods to use and give valuable information on the validity of the model.

2.6 Model comparison

2.6.1 Overview

Visualisation techniques are the use of statistical graphs and plots to represent data and information. The strength of visualisation is that it can communicate large amounts of information that will help the user to analyse and understand the data better [72]. It makes complex data and information more accessible, comparable, understandable and useable.

A Data-Information-Knowledge-Wisdom (DIKW) hierarchy [73] was constructed to show how wisdom can be obtained from a set of raw data. Visualisation is the golden thread that transfers data into information, information to knowledge, and knowledge to wisdom [73]. This is achieved by presenting the data with its underlying context. Figure 2-9 shows this hierarchy.

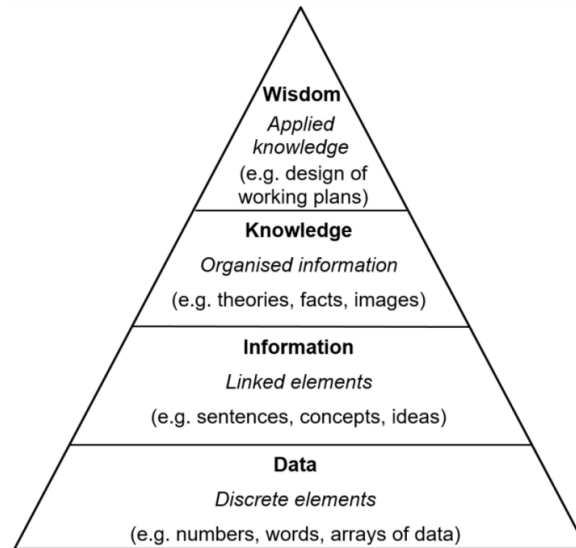


Figure 2-9: DIKW hierarchy

In this section, the criteria (parameters and limits) for desired model performances will be discussed. Various techniques will then be considered that can be used to visualise the various coal datasets and literature models. These visualisation techniques and desired criteria ranges can then be used to objectively compare the literature models with each other.

2.6.2 Parameters and limits

There are several ways in which one can visualise data. The first step is to identify and understand the story that one wants to portray and the relationship one wants to show [74]. Table 2-14 shows random data for five different groups of coal used to develop five random models. Table 2-14 is for illustration purposes only.

Table 2-14: Data / information table

Group	Sample size	Average A (%)	Average FC (%)	Average GCV(MJ/kg)	R^2	MAE (%)
1	135	20	40	22	0.85	3.5
2	200	30	30	25	0.84	2.9
3	1'800	10	15	28	0.90	5.9
4	50	15	50	30	0.75	1.8
5	500	30	40	25	0.60	2.5

Table 2-14 contains data which can give a variety of information. It can be noted that each of the different data groups contain a different number of samples. These samples also differ with regards to the ash content, fixed carbon content and the GCV value. By using this data to construct a model, it is noted that the goodness of fit for these data differs, as well as the standard errors.

These differences complicate the process to make conclusions on which developed model is best. To simplify this, it would be beneficial to only identify a few key factors to focus on [75], for example, to only consider the average ash content, the goodness of fit and the standard error of the model. More information can be gained from Table 2-14 by limiting the factors that are taken into consideration.

These limited factors are now the point of focus for visualisation. It is expected that each key factor would have criteria that would indicate whether they are acceptable or not. Such criteria can include a minimum or maximum value, or a range in which this factor should be [76]. These criteria define a boundary or limit that can visually be presented on a figure, giving an easy method to analyse data.

Table 2-15 presents a summary of the criteria for the R^2 and MAE values as discussed in Section 2.3. If R^2 and MAE values fall within the desired ranges, the model is deemed fit for the prediction.

Table 2-15: Boundary ranges for R^2 and MAE

Parameter	Range	Criteria
R^2	$0 \leq R^2 \leq 1$	As close to 1 as possible
MAE	0% to a max %	As close to 0% as possible

Taking these boundaries into consideration, it is easier to see which of the models comply or not. Visual illustration of these boundaries would make the process of analysing data quicker and therefore less time-consuming. In a similar manner, for examples, the ash content in Table 2-14 can be analysed by visualising a maximum value of 25%.

Defining boundaries and the subsequent visualisation thereof provides a simplified manner to analyse data. The simplification improves the transition from data to wisdom. Specific knowledge is captured, and interpretations can be made based on the single evaluation.

2.6.3 Visualisation techniques

Visualisation is when the identified boundaries and limits are used to visually represent the data. The goal of visualisation is to simplify the data to better understand it and to gain knowledge from it.

Data can be visualised in several different ways. It is important to choose the right visualisation type that will highlight the informative regions identified by the boundaries. Figure 2-10 gives several suggestions as to what type of visualisation technique to use [77].

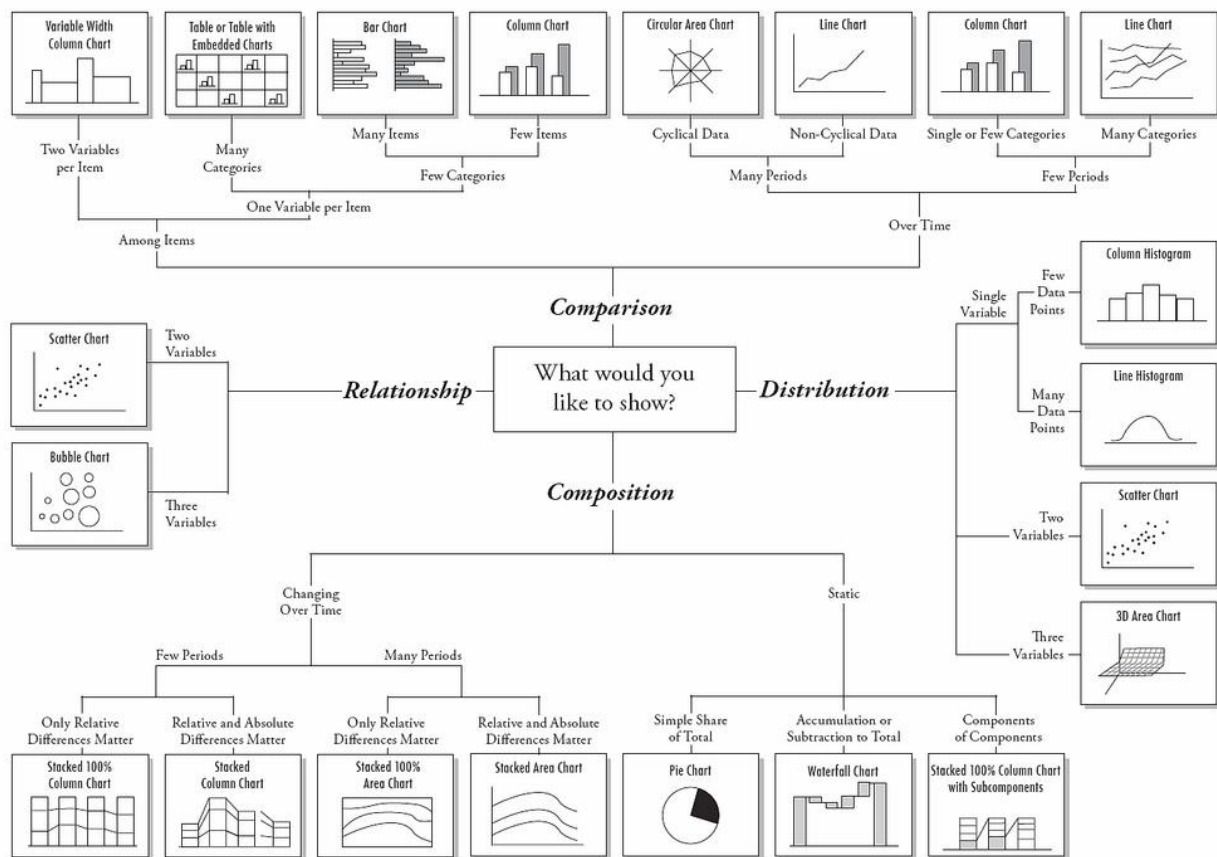


Figure 2-10: Visualisation suggestion chart [77]

The chart indicates four main ideas that are typically visualised [77]. Each of these ideas will be discussed in more detail by using the data in Table 2-14 to give examples where necessary.

Main idea 1: Relationship

If the relationship between several variables needs to be identified, a scatterplot or a bubble chart can be used. Scatter plots use horizontal (X) and vertical (Y) axes to plot quantitative, independent, or dependent variables in order to visualize the relationship between two variables [50]. Bubble charts have an extra component which is the size of the bubble. This allows the user to take a third aspect into consideration [78].

The advantages of scatterplots and bubble graphs are that they can show large quantities of data and make it easy to see relationships between parameters / variables and clustering effects [79]. These types of charts also work with almost any continuous scale data.

Figure 2-11 shows a scatterplot (left) and a bubble chart (right). The scatterplot shows the relationship between the average fixed carbon content of the coal groups and the average GCV. A positive linear relationship is observed between these two variables. This conclusion would not have been as easy to make simply by looking at Table 2-14. The same can be

observed from the bubble graph. This graph includes the sample size which is shown by the size of the bubble.

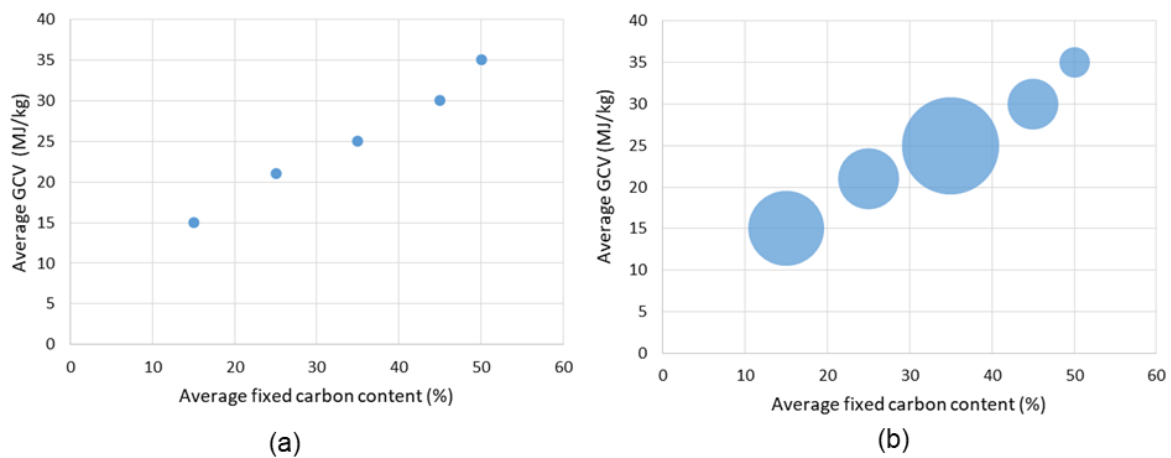


Figure 2-11: Relationship between the FC and GCV (a) and the sample size (b)

Main idea 2: Comparison

To compare variables with one another, bar charts are typically used. Bar charts use a horizontal (X) axis and a vertical (Y) axis to plot categorical data or longitudinal data [80]. Bar charts are very versatile. They are best used to show change over time, to compare different categories, or to compare parts of a whole [74].

Figure 2-12 shows the comparison of the proximate variables of the first three groups' data. From the bar chart it can easily be seen that Group 1 has a very high content of fixed carbon when compared with the other two groups. Such observations should be analysed further to understand why this might have occurred. However, this section only shows the ease of identifying differences between data.

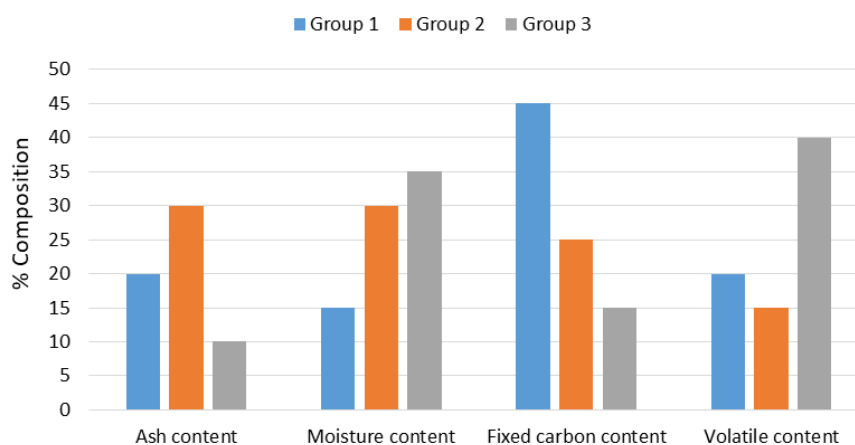


Figure 2-12: Comparison of the proximate variables of 3 groups of data

Main idea 3: Distribution

Both scatter and bar charts can be used to show the distribution of the data or parameters associated with the data over a period. Bar graphs will be used if you want to show the distribution of one variable, where scatterplots will rather be used when you want to show the distribution of two or more variables [75]. Line graphs are also commonly used to show the data's distribution over time.

Figure 2-13 shows the distribution of the GCV of all five groups of data, illustrating that it is possible to see which group gave the highest or lowest GCV. Assuming that groups 1 to 5 were taken over a period of time (e.g. 5 consecutive months), this would indicate how the GCV has changed over the time period.

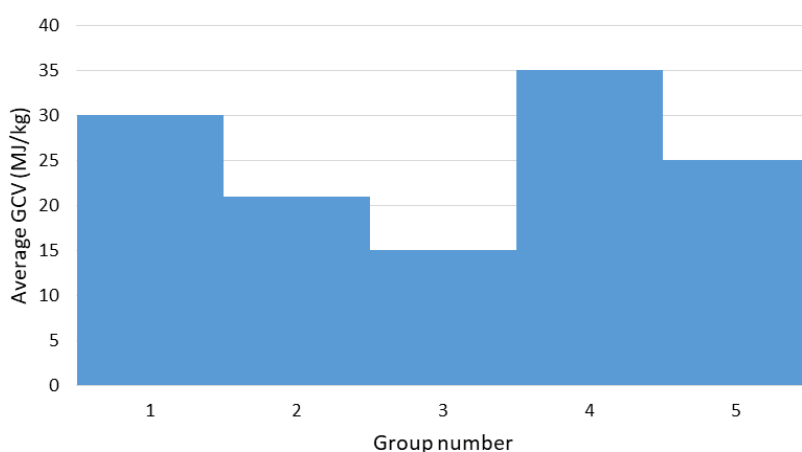


Figure 2-13: Distribution of the GCV of the five groups

Main idea 4: Composition

Pie charts are the most popular types of graphs to use when trying to show the composition [31]. It is advisable to be cautious when using pie charts with too many slices. The pie slices are easily gauged when the sizes are familiar percentages (25%, 50%, 75%, 100%), because they are common angles [81]. Other angles can be interpreted inconsistently, making it difficult to compare relative sizes and therefore less effective.

Since the four proximate variables for group 1's data is in familiar percentages, a pie chart is constructed to show the composition, highlighting how the composition can visually be illustrated. From the example, fixed carbon clearly contributes the biggest percentage towards the total analysis. However, without the percentage labels, room for misinterpretation is left for the remaining variables, bringing uncertainty into the observation, especially if a quantitative result is required. Pie charts should therefore be used with caution, when the quantitative

value is less important. Figure 2-14 gives an example of a pie chart with a set percentage A, M, FC and VM composition.

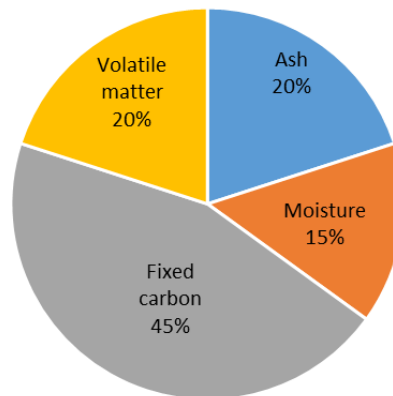


Figure 2-14: Composition of group 1's coal

2.6.4 Objective comparison

An objective of this study is to compare the various GCV models in order to identify a desired model. The dissimilarities in the data and methodologies used to develop the various literature models proves difficulty for **objective** comparison. This problem can be addressed by assessing and comparing the validity and reliability of each model [82].

Validity (accuracy) describes how well the model represents the true state of affairs. Reliability (precision) describes how closely multiple models with a specific test agree with each other [83]. The validity and reliability can be visually compared using the visualisation techniques discussed in Section 2.6.3. To allow for a comprehensive but still simplistic form of comparison the various visualisation techniques can be combined [72].

The objective comparison of the literature models will be addressed by visually comparing the validity and reliability of the models.

Objective comparison of the validity of models

The validity of the models can be visually compared with a number of visualisation techniques [84]. The aim of this comparison is to give a verdict on each model's reported performance. The literature models mostly reported on sample size, accuracy of the model, error percentages and the model's fit (R^2), all of which can give information on the model's validity [66].

Figure 2-15 presents three visualisations, each giving information with regards to one of four GCV model's validity. This is a simple method that can be used to easily identify the ranges of GCV predictions, sample size and accuracy for each model.

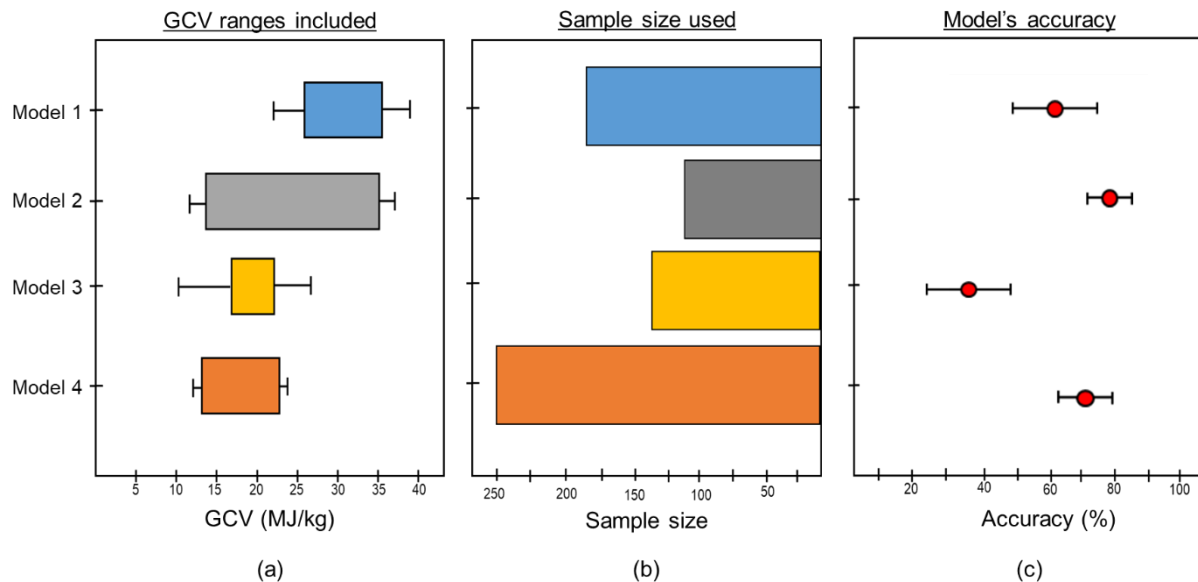


Figure 2-15: Visual comparison of models

Figure 2-15 gives valuable information and several conclusions can be made from the three figures. One conclusion is that model 2 gives the widest range of consistent GCV predictions as well as the highest accuracy whilst having the smallest sample size. The visualisation thus enabled the simple demonstration of data per model.

When trying to decide which model is ultimately the best when taking all these validity parameters into consideration, it gets more difficult [84]. To simplify this process, all of the parameters are combined on the same graph. Figure 2-16 illustrates the combination of these parameters on one graph. From Figure 2-16 it is easier to see which model generally performs better, since all of the validity parameters can be evaluated simultaneously [31].

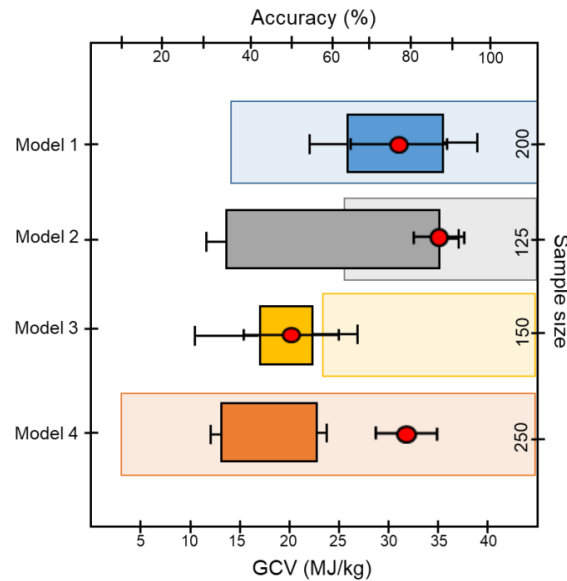


Figure 2-16: Combined visualisation of models

The visualisation allows the user to easily identify which model's accuracy is the best whilst simultaneously considering the sample sizes and ranges of prediction. Conclusions, such as the sample size does not seem to effect accuracy of prediction in Figure 2-16, can be made easily.

Objective comparison of the reliability of models

The first thing to consider when comparing the reliability of models (test models on new data), is to ensure that the criteria for comparison is similar [85]. This implies that it is desirable for certain factors used in the modelling process to be similar for all models. These factors can include, but are not limited, to the [47]:

- sample size,
- type of coal (lignite, bituminous, anthracite, etc),
- basis of analysis (air-dry, as-received, dry-base, etc),
- type of independent variables (proximate and / or ultimate), and
- model type (regression, ANN, etc.).

Comparing models that were developed within similar criteria ranges on a new dataset will allow unbiased results of the performance on the new dataset [36]. This is impractical for the objective comparison of the reliability of the literature models since most of these factors differ.

Due to the differences in the factors listed above it is crucial to ensure that the data then used for comparison complies with all of the model's specifications [74]. For coal GCV model

comparisons this will mean that caution should be taken not to test literature models with its designed limitations.

The same visualisation techniques demonstrated in Figure 2-15 and Figure 2-16 can be applied for the comparison of the reliability of the models as well.

2.6.5 Summary

The aim of this section is to visualise the developed model and to ultimately compare it with the literature models. Different techniques were discussed that can be used to visualise various aspects of the new model. These techniques focus on comparison, distribution, relationship and composition. A combination of the visualisation techniques can then be used to compare the various models with each other. Using this, one or more validation criteria for the models can be illustrated on the same graph to easily see which model outperforms the other.

2.7 Conclusion

This chapter gave comprehensive research on the best practice methodology to use when developing a new GCV model. The general best practice approach for model development, presented by Xu [28], was investigated alongside the methodologies of 12 literature models.

This investigation identified three focus areas that are required to develop a new GCV model. These three focus areas are data preparation, model development and model validation.

Each of these focus areas is followed by various sub-steps. It was found that the literature models followed different techniques to comply with the sub-steps. The various techniques were discussed with respect to its advantages, disadvantages and application. Ultimately the most suitable techniques for GCV modelling were identified.

In addition to the three focus areas identified in Xu's approach [28], a model validation focus area is included and relevant literature provided. This focus area is newly introduced with the aim of helping the user to visualise the final outcome of the first three focus areas.

Various visualisation techniques were discussed alongside their applicability on GCV models. The objective of this last focus area is to objectively compare the various models with each other and to evaluate the methodology.

CHAPTER 3 – METHODOLOGY

3 Methodology

3.1 Preamble

The problem statement in Chapter 1 emphasised the need to identify a general approach to develop new CV models for coal and ultimately compare them with existing models. Different aspects applicable to the development of new GCV models were identified in the literature study (Chapter 2). Additionally, literature regarding visualisation techniques illustrated how different models may be compared.

The literature study showed that existing GCV models focus on three focus areas when building a new model. The three focus areas are: (1) data preparation, (2) model development, and the consequential (3) model validation. In this chapter these three focus areas as well as a newly introduced fourth focus area, model comparison (4), will be used to present a general approach for GCV modelling. The diagram in Figure 3-1 shows the chronology of the four focus areas of the devised approach.



Figure 3-1: Four focus areas for GCV modelling of coal

Each one of the four focus areas, as in Figure 3-1, includes several sub-steps. These sub-steps will be discussed thoroughly in this chapter. The combination of the focus areas and the respective sub-steps presents the necessary methodology for this study.

3.2 Dataset preparation

3.2.1 Overview

The first focus area of the methodology is the data preparation. Figure 3-2 shows the sub-steps included under the data preparation focus area. Firstly, data is gathered, where after the quality of the dataset is evaluated (inspect data). If necessary, the dataset is cleaned to improve its quality (clean data). The final sub-step divides the dataset so that it can be used for modelling and validation (split data). Each of these sub-steps will be discussed in more detail in the following sections.

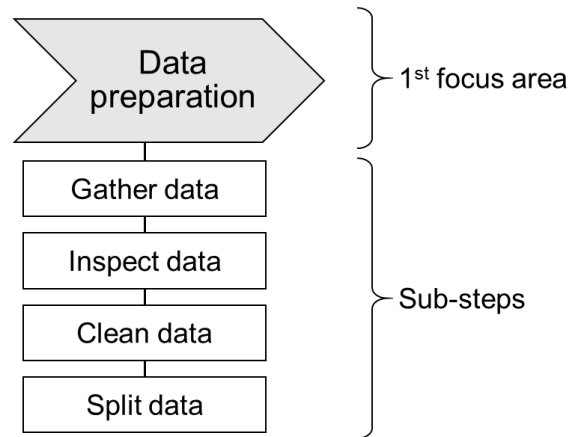


Figure 3-2: Sub-steps for the data preparation focus area

3.2.2 Gather data

It is evident from Chapter 1 that coal consists of a diverse set of properties. When obtaining coal datasets it is therefore important to ensure samples are consistent (of good quality [58] and from a reliable source [47]). A consistent dataset will enable building representative models.

The required datasets are most commonly obtained from either the coal mines or from the industrial processes that utilise the coal [3]. These industries can analyse the coal samples at different conditions [5]. This can be as-received (original coal sample), air-dried (coal sample taken after it has air dried), dry-base (total moisture has been completely removed) or dry-ash-free (total moisture and ash has been removed) [9]. It is firstly important to ensure that the various coal data are on the same base.

These datasets may consist of the proximate (A, FC, M and VM), the ultimate analysis (C, H, N, O and S) with the corresponding GCVs, along with the coal type and geographic location.

The different coal types include bituminous, sub-bituminous, lignite or anthracite [86]. It is important to have information regarding the geographical origin of the coal, since the location can influence the properties even for a similar type of coal [87]. Not distinguishing between the geographical origins of the coal can lead to inherent variability and an inconsistent dataset [5].

The data should therefore be inspected to ensure that it is correctly interpreted and used for specific applications.

3.2.3 Inspect data

After understanding the data, the quality of the dataset needs to be evaluated. It is important to ensure a high quality dataset since it will have a direct impact on the quality and results of the developed models. In this sub-step the data quality is checked for completeness by investigating the dataset for abnormalities and ensuring a representative sample size.

Before the sample size can be established, the data must be inspected for any abnormalities. These abnormalities in the data can include samples containing missing, incorrect or irregular data.

Table 3-1 gives an example of five samples of coal data and the respective proximate analysis variables. This example focusses on the proximate variables only but the same concept can be applied to the ultimate analysis. These properties are the proximate analysis variables which consist of the ash (A), inherent moisture (IM), fixed carbon (FC) and volatile matter (VM). Sample 1 contains coal with a very high IM content. This seems irregular since the IM content usually ranges between 2% and 15% [5]. It can further be seen that samples 3 and 5 include missing data. In such a case, it is suggested that the entire sample is removed from the dataset.

Table 3-1: Example of coal samples including missing data

Sample	A (%)	IM (%)	FC (%)	VM (%)	GCV (MJ/kg)
Sample no 1	12.4	60.1	19.9	7.5	34.24
Sample no 2	19.6	4.1	52.5	23.8	23.49
Sample no 3	28.6	2.6		22.2	
Sample no 4	34.3	3.6	41	21.1	19.43
Sample no 5	32.3	2.6	42.3		19.53

It is possible that a significant number of samples are removed during the evaluation. To ensure that the dataset remains representative, the number of samples used should be above a minimum required sample size. To assess the sample size, Green's method is used due to its simplicity [47]. The formula used to determine the sample size (n) is given in Equation 3-1.

$$n = 50 + 8p \quad (3-1)$$

n - minimum sample size

p - variables per sample

where n denotes the minimum required sample size and p the number of variables associated with each sample. In this study four proximate variables are considered, i.e. ash, inherent moisture, fixed carbon and volatile matter. The corresponding minimum sample size (n) if four variables are considered ($p=4$) will be 82. Therefore, it should be ensured that all datasets have a minimum of 82 samples.

It should be noted that a complete dataset on its own does not prove high quality. The quality of the data is similarly dependent on the metering and analysis equipment. To assure a high data quality it is recommended to verify whether the bomb calorimeter, used to obtain the gross calorific value (GCV) of the samples, is calibrated. Data obtained from a calibrated calorimeter will be perceived as high quality due to external assurance that the information is accurate [69].

3.2.4 Clean data

As discussed in Chapter 2, outliers can significantly influence the quality of a dataset which may have a direct influence on the developed models. In this sub-step the dataset is evaluated to identify and remove possible outliers. This is done to ensure a high quality dataset that can be used to develop accurate coal models.

Potential outliers are easily identified in MATLAB using the *is-outlier* function (as discussed in Chapter 2). MATLAB flags all data points larger than two standard deviations from the mean. Table 3-2 shows a set of five more exemplary coal samples containing potential outliers. Each flagged sample should be evaluated as to whether it should be removed or not.

Table 3-2: Coal samples containing outliers

Sample	Flagged?	A (%)	IM (%)	FC (%)	VM (%)	GCV (MJ/kg)
Sample no 6	YES	26.9	56.3	12.4	15.3	12.4
Sample no 7	NO	31.9	8.3	32.9	26.9	12.6
Sample no 8	NO	15.1	3.1	58.2	23.6	26.4
Sample no 9	NO	22.5	3.3	52.5	21.7	23.9
Sample no 10	YES	0	15.8	56.3	27.9	12.3

Evaluating the different properties highlights the potential cause of the outlier. Outcomes from the evaluation will be based on logic and specific knowledge on the samples. The evaluation of sample 1 shows that the sum of the sample's proximate analysis values does not add up to 100%. Specific knowledge that a proximate analysis should total 100% allows the logical deduction that the sample is not correct. The specific error cannot be identified from the data

alone, although it is likely to be the high IM content. This sample can therefore be removed as an outlier due to the underlying errors.

Evaluation of sample 5 shows a 0% for A content. A zero value should have been picked up during the initial *check data* sub-step but could have been overlooked. A 0% ash value is unlikely since coal is usually composed of ash [15]. This shows that the *clean data* sub-step would identify bad quality data points that may have been missed during previous sub-steps.

3.2.5 Split data into modelling and validation datasets

Once a clean dataset is obtained, the dataset is split into data used for modelling and validation purposes, respectively. It is necessary to randomise the data first. The purpose of the randomisation is to ensure that the validation and modelling datasets include a representative and fair distribution of data. The data is randomised using the *Rand* function in Microsoft Excel™.

Figure 3-3 illustrates a total dataset, where the different colours distinguish between samples from different sources (each row). After randomising the samples, the 30/40/30 split method is used to split the samples into modelling and validation data [45].

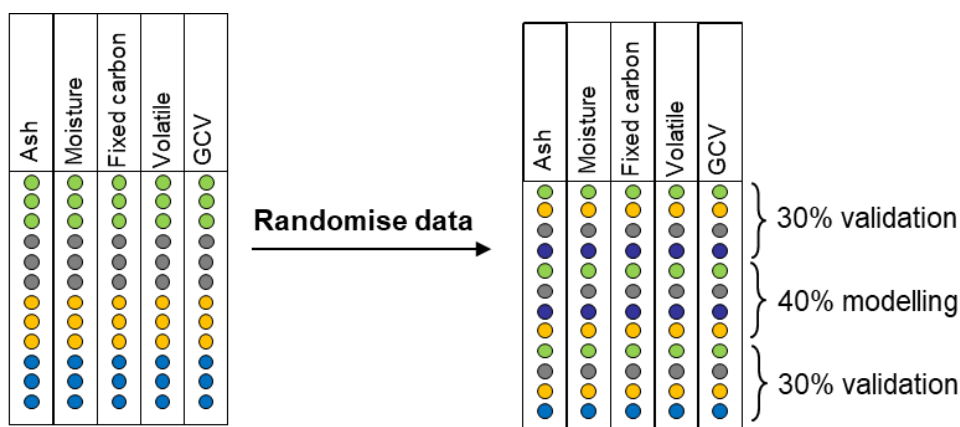


Figure 3-3: Data split into modelling and validation sets

The 40% modelling data will be used to develop the model in the next step. The two groups of 30% validation data will be used to validate the model twice. The average of the two validations will then be calculated and used to validate the final model. Splitting the data thus presents three different but comparable datasets as the last step for the data preparation focus area. These datasets can now be used to develop and validate GCV models.

3.2.6 Summary

The sub-steps in the data preparation step focus on understanding the data and ensuring that the dataset is of high quality. This is done through four sub-steps. Data is gathered from different sources and evaluated for missing data or substantial errors. If necessary, the data quality is further improved by removing outliers to clean the data. The cleaned dataset is then split into three smaller datasets that can be used to develop and validate GCV models.

3.3 Model development

3.3.1 Overview

In the second focus area the high quality dataset from focus area one is used to develop GCV models. This is done in four sub-steps, namely model type selection, identify modelling variables, develop model and test model. Figure 3-4 shows the sub-steps of the model development focus area. Each of these sub-steps will be discussed in more detail in the following sections.

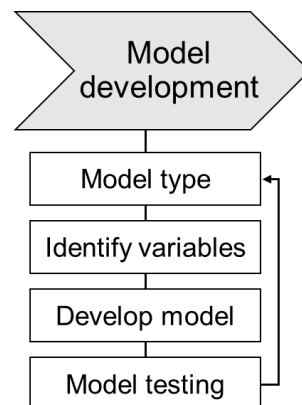


Figure 3-4: Sub-steps for the model development focus area

3.3.2 Model type

The first step to building a new model is to decide what type of model to build. The most prevalent options for the modelling of coal GCV are Regression and Artificial Neural Networks (ANNs). These methods were discussed in detail in Chapter 2 and it was concluded that both the regression model and the ANN give similar results. Regression models are, however, more straightforward and user friendly [11, 12, 25].

It is recommended to evaluate and confirm the relationship between the GCV and the individual proximate variables on a new set of data. For the instances where a clear linear relationship is observed, a linear regression model can be devised [18]. For nonlinear relationships either nonlinear regression models or an ANN will suffice.

3.3.3 Identify variables

Once the desired model type has been selected, the next sub-step is to identify the variables to include in the model. The variables should present a reasonable representation of the reality. Majumder *et al.* [18] explicitly investigated the type of variables that should be included to develop a GCV model. He concluded that the effect of all four of the proximate analysis outweighs the effect of the ultimate analysis when developing GCV models. The proximate variables are also more commonly available than the ultimate analysis [23].

It is therefore recommended to include the proximate variables for GCV models. However, the incorporation of the ultimate variables will also be sufficient.

3.3.4 Develop model

After determining the model type and variables to be used, the model can be build. This is done by using the 40% modelling data as prepared in the data preparation section. In this study the LINEST function in Excel™ is used to build linear regression models. As discussed in Chapter 2, the LINEST function gives a matrix as output which gives the constants for each variable (x_1 to x_4) as well as the intercept. The influence of each of these variables is also given within this matrix. Table 3-3 gives the LINEST matrix.

Table 3-3: LINEST matrix

LINEST matrix				
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	v_2	#N/A	#N/A	#N/A
SS_{reg}	SS_{res}	#N/A	#N/A	#N/A

The first row of the matrix gives the coefficients for each variable, known as the regressors, as well as the residual error, which is a constant value. The second row contains the standard error values associated with each regressor. The standard error values are an estimate of how

much the errors in individual variable measurements contribute to the overall error and ultimately indicate the influence the regressor has on the output.

The regressors (first row) are used in the model and presented in Equation 3-2.

$$GCV = x_1(Ash) + x_2(IM) + x_3(FC) + x_4(VM) + C \quad (3-2)$$

GCV	-	Gross calorific value in MJ/kg
x_i	-	Constant associated with each of the four proximate variables
C	-	Residual error (constant value)
A	-	Ash content of the coal in %
IM	-	Moisture content of the coal in %
FC	-	Fixed carbon content of the coal in %
VM	-	Volatile matter content of the coal in %

Equation 3-2 is the developed model. It is important to evaluate the regressors of the model to see if the model is as expected. If all the regressors are zero, it means that the GCV will be completely dependent on the residual error value (C) irrespective of what the proximate variables are.

The evaluation of the constants can also indicate the type of influence each of the constants has on the GCV (positive or negative). If the relationship between a proximate variable and the GCV is negative, it is expected that the constant associated with this proximate value will also be negative, and vice versa.

If the model is not as expected, it is advisable to inspect the SE_{x_i} values alongside the predictors. This will allow the user to identify the influence that the regressor with the unexpected constant has on the overall error of the model. If the influence is significantly small, it is still practical to use the model. If the influence is big, it is recommended to re-evaluate the data used for modelling and the type of model selection [66].

A developed model that looks correct also does not necessarily constitute a viable outcome. It is therefore recommended to test the model to ensure that the model is representative of the dataset. Some of the other statistics in the LINEST matrix are used to test the model and will be discussed in the next section (check requirements).

3.3.5 Check requirements

In this sub-step the developed regression model is evaluated against a set criteria. This is done to ensure that the model is representative of the dataset. For this evaluation, the modelling data is used to test the model.

The R^2 , the RMSE and the MAE (on the modelling data) are evaluated to give information on the credibility of the developed model. These statistics and their applicability are discussed in Chapter 2. The R^2 and the RMSE value are obtained from the LINEST matrix while the MAE are calculated using Equation 3-2.

Table 3-4 indicates the desired criteria ranges for these statistics to adhere to. If the models comply with these criteria, they can be validated and used. However, if the model does not adhere to the criteria in this sub-step, it is advised to develop a different model that is acceptable.

Table 3-4: Testing criteria for model checks [27], [37], [88]

Statistic	Criteria
R^2	Close to 1
RMSE	Small value
MAE	Small value

3.3.6 Summary

In this focus area, a new GCV model is developed and statistically tested. A suitable model type and the necessary variables are confirmed from relational comparisons. The model is then built and tested to see whether it meets the requirements stipulated in Chapter 2. If the model meets the requirements it can be deemed statistically sound and be validated in the next focus area.

3.4 Model validation

3.4.1 Overview

In this focus area the developed model is validated. This is done to demonstrate that the model is representative of the actual data characteristics. This ensures that the model reproduces authentic behaviour with enough fidelity to satisfy the analysis objective. Therefore, to validate the newly developed model, the two 30% datasets are used. Figure 3-5 shows the validation focus area of the methodology.

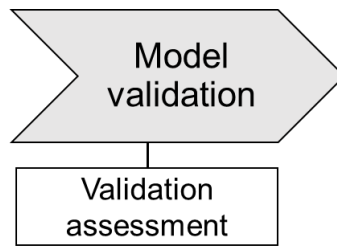


Figure 3-5: Sub-step for the model validation focus area

3.4.2 Validation assessment

There are two steps to judging how well the model predicts the desired outcome (experimented GCVs). The first is to assess whether the predictions which have been made are reasonable with respect to the real behaviour (model validation), which will be discussed in this section of the study. The second is to ascertain whether the developed model implements the predictions correctly (model verification), which will be done by applying the methodology on several case studies in Chapter 4. For both these evaluations it is important to note that the two sets of 30% validation data placed aside during the data preparation step (Section 3.2) are used.

For the validation of the developed model, which is the multivariable function presented by Equation 3-2, two statistical parameters will be tested. These parameters are the MEA and the CC, both of which present a quantitative measure of validation.

Mean absolute error

The MAE is the absolute average errors calculated between the experimented GCVs(from the analyses) and the predicted GCVs. Figure 3-6 shows an x-y axis along with the relationship between the experimental GCV and the predicted GCV. The closer the black dots are to the blue line, the more accurate the model. The difference between the black dot and the blue line thus indicates the error value associated with that prediction.

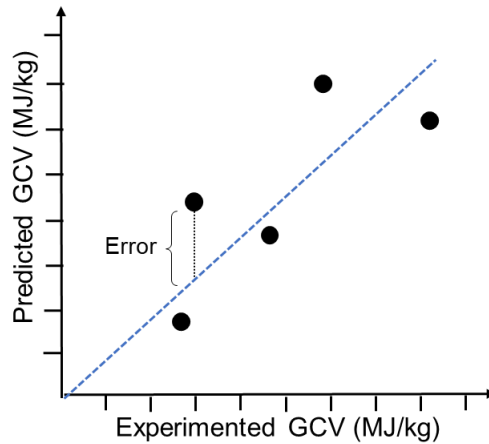


Figure 3-6: MAE of the developed model on validation data

The MAE for the model is calculated by determining the absolute value of all of these errors between all the dots and the blue line.

Comparison coefficient (CC)

In addition to the MAE parameter, the model is validated by comparing the experimented and predicted GCVs. This is done by calculating the CC on the relationship which will further evaluate the accuracy of the prediction.

The GCVs of the two sets of 30% validation datasets are calculated using the developed model and illustrated in Figure 3-7. The black dots illustrate the desired outcome ($CC = 1$). The two striped lines in Figure 3-7(a) indicate the outcome when each of the validation sets is used in the developed model to predict the outcome. Figure 3-7(b) illustrates the average CC value between the two validations.

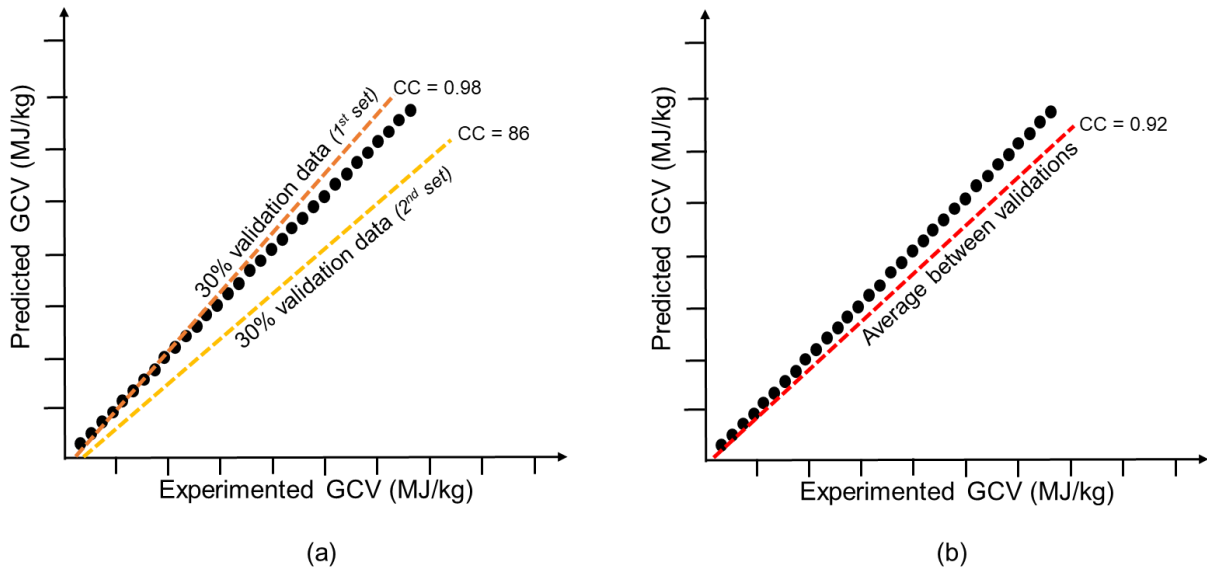


Figure 3-7: Experimented vs predicted GCVs

The average CC in Figure 3-7 (b) along with the MAE in Figure 3-6 gives judgment on the validity of the model. Table 3-5 gives a summary of these two statistical parameters and criteria used to validate the model.

Table 3-5: Model validation parameters and criteria

Model validation parameter		Criteria
MAE	Absolute average error	Small value
CC	Comparison coefficients	As close to 1 as possible

3.4.3 Summary

In this focus area the approach to validate the model is described. The relationship evaluation is performed on the two 30% validation datasets. The model is firstly validated by calculating the MAE between the experimented and the predicted GCVs. The model is then further validated by using the relationship between the experimented and predicted GCVs, known as the CC. This focus area ultimately ensures that the model represents a reasonable representation of the experimental GCVs.

3.5 Model comparison

3.5.1 Overview

The last focus area in the model development approach is model comparison. The intention of this focus area is to objectively compare the performance of the developed and literature GCV models. The purpose of the comparison is to ultimately identify a preferred model on new coal data. Figure 3-8 shows the sub-steps for the model comparison focus area.

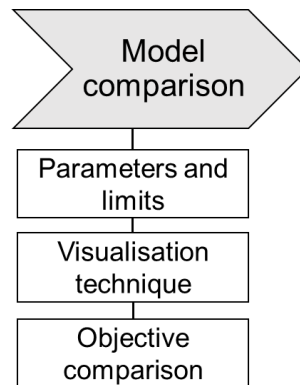


Figure 3-8: Sub-steps for the model comparison focus area

The sub-steps of this focus area aim to identify the parameters (Section 3.5.2) and techniques (Section 3.5.3) that will be used for the comparison of the model's validity and reliability (Section 3.5.3). Each of these sub-steps is discussed and gives the final steps of this study's methodology.

3.5.2 Parameters and limits

In this sub-step the limits of the desired statistical parameters that the user wants to compare are established. The limits refer to what is acceptable for the identified statistical parameters. This is done to compare the statistical relevance of the different models using visual representation. These limits will enable users to easily identify the accuracy and representativeness of each of the different models.

One or more statistical parameters can be used during the evaluation. These parameters include, but are not limited, to the R^2 , F -value, SSE , sample size and minimum and maximum boundaries of a variable.

To meet the main objective of this study, two individual visualisations will be composed for the performance of the GCV models. The first visualisation will focus on the model's validity. The validity of the models evaluates the performance of each model (developed and literature) on their own datasets. The second visualisation will focus on the model's reliability. The reliability

of the models evaluates the performance of each model (developed and literature) on new coal datasets. These two visualisations will allow for the comparison of the model's intended performance (*Visualisation 1*) as well as for the model's real performance on new data (*Visualisation 2*).

Figure 3-9 demonstrates the data and parameters that will be used for each of the two visualisations.

For *Visualisation 1*, the R^2 and MAE parameters of the developed model and the literature models will be used. The R^2 value for the developed model was calculated in the *check requirements* sub-step of the methodology (Section 2.4.5). The MAE were calculated in the *data validation* sub-step with the validation dataset (Section 2.5). The R^2 of each literature model, along with the MAE on their respective validation dataset, are published and available [12, 15-19, 21–25]. This allows for the objective comparison of the models based on these parameters. The objective of *Visualisation 1* is to compare the performance of each model on their original datasets.

For *Visualisation 2*, a new coal dataset is used to validate each of the models. This new dataset contains coal samples different from any of the models that were originally developed. This allows for the calculations of the CC and the MAE of each model on the new dataset. The objective of *Visualisation 2* is to compare the performance of each model on a new coal dataset.

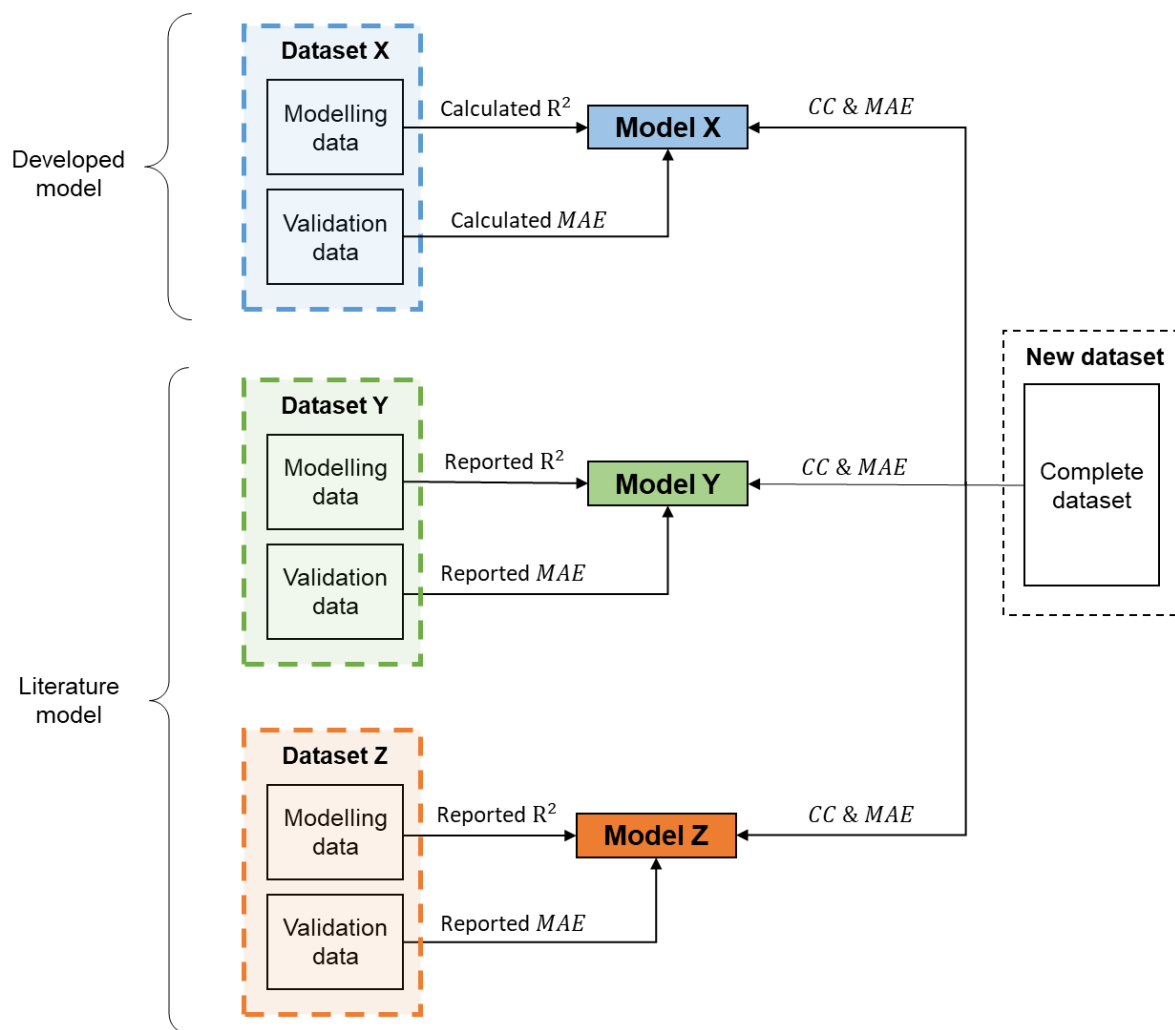


Figure 3-9: Data used for visualisation parameters

Now that the specific parameters are identified, these parameter's criteria will determine the limits chosen. The limits are based on the desirable ranges for these parameters as discussed in Chapter 2. Table 3-6 gives a summary of the criteria of each parameter.

Table 3-6: Visualisation parameters and limits

Parameter		Ranges	Desired limits
Visualisation1 (a) (Modelling data used)	R^2	$0 < R^2 < 1$	Close to 1
	MAE	0 to max	Close to 0%
Visualisation1 (b) (Validation data used)	R^2	$0 < R^2 < 1$	Close to 1
	MAE	0 to max	Close to 0%
Visualisation2 (New dataset used)	CC	$0 < CC < 1$	Close to 1
	MAE	0 to max	Close to 0%

The desired limits are very broad and do not indicate specific boundaries. Definite boundaries can therefore be identified by using the average values reported by the literature models as the limits. The literature models are all published and acceptable models; it is therefore assumed that parameters in the same ranges as these models will be accepted.

3.5.3 Visualisation techniques

After the parameters are identified the type of visualisation technique can be selected. The visualisation chart in Chapter 2 (Figure 2-10) is used to identify the visualisation technique. For *Visualisation 1* and *Visualisation 2* the relationship between the two limiting parameters will be illustrated.

Since the relationship between two parameters will be shown, a scatterplot will be used. The advantage of the scatterplot is that it easily illustrates the relationships between the parameters. Figure 3-10 illustrates a scatterplot axis with only the limits identified.

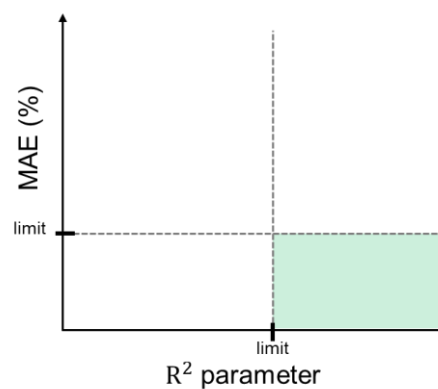


Figure 3-10: Scatterplot area with the identified limits

The ideal relationship point should be where the two parameters satisfy the specific criteria. The optimal performance area is indicated with the green box in Figure 3-10. An undesired relationship would therefore be a point with a high MAE and a low R^2 .

The respective MAE and R^2 parameters of each model can then be added to the plot area containing the limits, as shown in Figure 3-11.

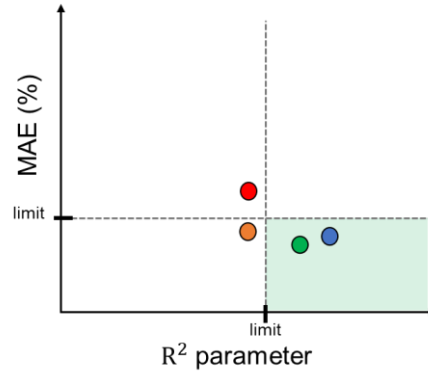


Figure 3-11: Scatterplot with the limits and the models

This visualisation technique enables the user to easily identify which models shows better relationships. In Figure 3-11 it can also be easily identified that the red dot is in an undesired range, making this model less preferable to use, compared to the green and blue dots.

This concept is applied to *Visualisation 1* and *Visualisation 2*. For *Visualisation 2* the CC parameter instead of the R^2 parameter is used. These visual representation of the statistical parameters, as shown in Figure 3-11, ensures that the results are easily interpreted [6].

3.5.4 Objective comparison

For the objective comparison of the models, two visualisations are considered, *Visualisation 1* and *Visualisation 2*.

The first visualisation focusses on the validity (accuracy) of the models which describes how well the models perform on their own data. This visualisation enables the comparison of various GCV models to ultimately choose a preferred model. The published MAE and R^2 parameter values on a) the modelling data and b) the validation data are used.

The second visualisation focusses on the reliability (precision) of the models which describes how well the models perform on new coal data. A new MAE and CC will thus be calculated using the new coal dataset for each model used in *Visualisation 1*. The new dataset includes coal sample analyses from a location different from that on which the models are developed. This allows for the comparison of the model's performance on new data.

The validity and reliability of the models are visually compared with the use of scatterplots. The limits for the set out parameters are the same for both visualisations. Figure 3-12 shows the expected outcomes of a) *Visualisation 1* and b) *Visualisation 2*.

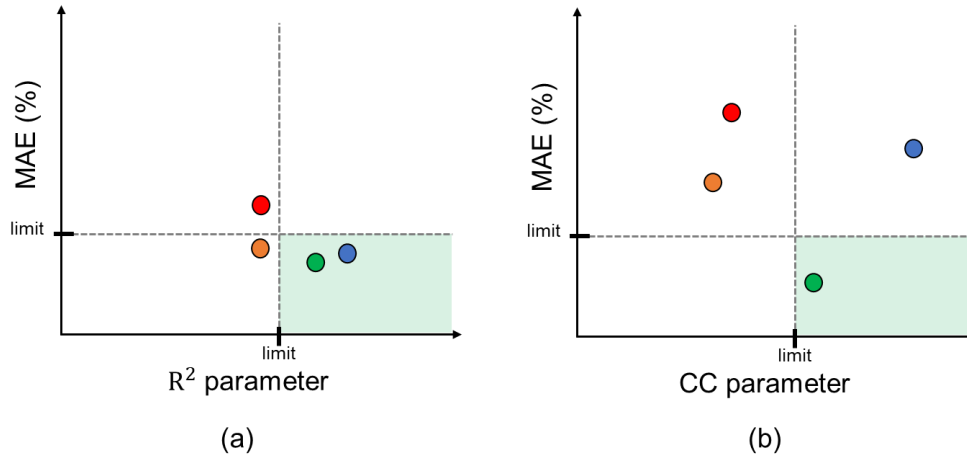


Figure 3-12: Visualisation of models performance on a) own data and b) new data

Figure 3-12 (a) represents the expected performance of the models (validity). This will typically be the model's optimal performance. Figure 3-12 (b) illustrates how the models perform on new data (reliability). From this illustration it can be noted that only the green model (green dot) performs within the optimal range when new data is used.

This application is practically applicable for identifying which model will be sufficient to be used for coal at a specific location. Such visual indication demonstrates how the models perform on new data, allowing the user to easily identify a desired model to use. This sub-step does not eliminate the other models but provides an indication on how each will perform when used.

3.5.5 Summary

The first sub-step of this focus area addressed the various parameters that can be included for the objective comparison of the various models. These parameters were identified as the R^2 of the model, the MAE calculated on the modelling and validation data, and lastly the CC between the predicted and measured GCVs.

The limits for each of these parameters were discussed and identified as the average between the literatures models' results for these parameters. The limits are used as guidelines that will ultimately allow the user to identify the optimal ranges for the models' performances.

The visualisation technique decided on to best describe the relationship between the parameters was the scatterplot visuals. The scatterplots allowed for the objective comparison of the various models' performance based on their validity and reliability.

3.6 Conclusion

The problem statement in Chapter 1 identified the need to devise a general approach to develop new GCV models for coal and to ultimately compare them to the literature models. In this chapter a general approach was devised with a specific focus on modelling the coal energy content and comparing the new model to the existing models.

The general approach devised in this chapter includes four focus areas: (1) data preparation, (2) model development, (3) model validation and (4) model comparison. Each one of these focus areas consists of various sub-steps. The application of each focus area with the respective sub-steps was discussed in the various sections of this chapter.

In the first focus area (Section 3.2), data is gathered and evaluated to ensure a consistent and representable dataset. This allows for clean and representable data to be used for modelling purposes.

In the second focus area (Section 3.3) the relationship between the desired variables is assessed and a model based on the observed relationships is developed. A general statement to develop a linear regression model containing all four proximate variables was made based on literature models' outcome.

The third focus area (Section 3.4) allows the user to validate the model and to ensure that a statistically sound model is used.

Finally, the various models are visualised and compared in the last focus area (Section 3.5) of the devised approach. This focus area is important to objectively compare the models in order to decide which model will be preferred for the specific dataset. The aim of this last focus area is to address the problem statement of this study: **Will the literature models be applicable to a new coal dataset and which one would be best?**

The devised approach is shown in Figure 3-13. This approach enables the development of coal GCV models and comparison thereof to existing models.

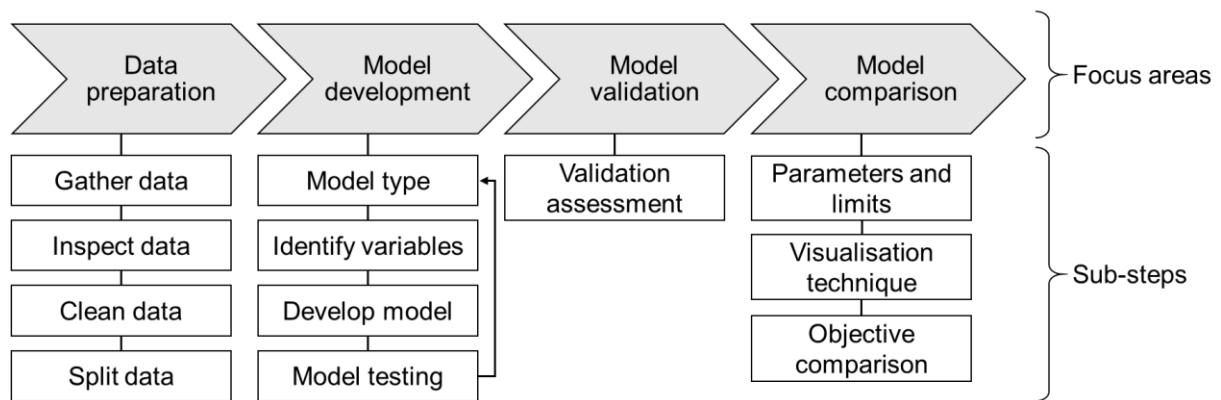


Figure 3-13: Devised approach for GCV modelling of coal

CHAPTER 4 – VERIFICATION AND VALIDATION

4 Verification and Validation

4.1 Preamble

In Chapter 3 a general approach (Figure 4-1) was devised to develop and assess models estimating coal energy content. The approach focusses on four focus areas and incorporates several sub-steps to achieve the desired outcome.

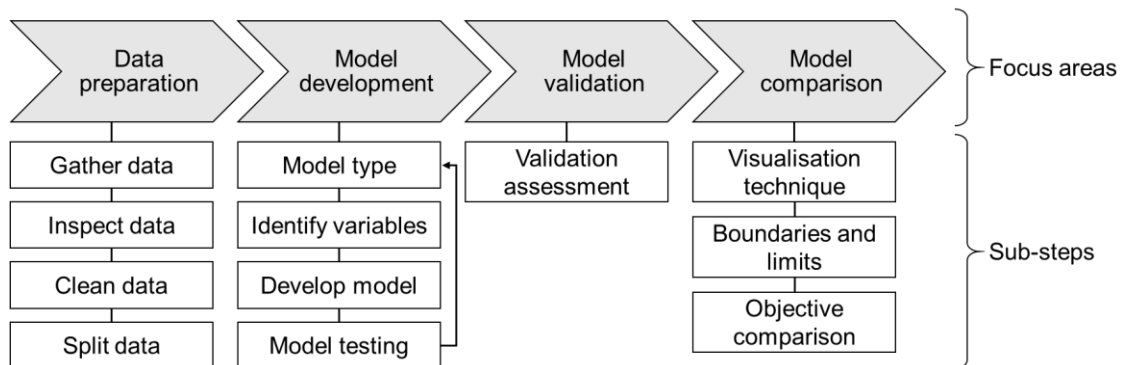


Figure 4-1: devised approach for GCV modelling of coal

The *data preparation* focus area delivers a high-quality dataset which is divided into modelling and validation data. The modelling data can be used to develop a representable model in the *model development* focus area. The developed model is then validated using the validation data set aside in the first step. The validation of the model ensures that the model would be able to accurately predict the GCV of coal. If the model met the validation criteria it can be compared to the literature models in the *model comparison* focus area. Visual comparison presents an objective manner to assess the performance of the various models.

To verify the general approach, presented in Figure 4-1, it will be applied to three case studies. Each case study is directed on coal from a different geographical location (South Africa, India and Alaska). Detailed results for each of these case studies can be found in Appendix A. This chapter only highlights significant observations and findings of each case study.

4.2 Dataset preparation

4.2.1 Overview

In this section, relevant results from case study 1 will be presented and discussed. Data from South African coal were prepared to deliver a high-quality dataset. Additional information can be found in Appendix A.1.

4.2.2 Gather data

For case study 1, a total of 2'850 samples were obtained from three different reliable sources. The data is summarised as follows:

- 1'650 samples from an industrial plant in Mpumalanga from 2013 to 2016. This plant gets its coal from mines across South Africa.
- 720 samples from a third party¹, covering data from 2009 up until 2016. These samples are taken from various mines across South Africa.
- 480 samples from multiple South African coal bulletins. These bulletins are freely available on the Department of Energy of the Republic of South Africa's website [89].

The coal data from all three of these sources is on an *air-dried* basis (contains IM only) and includes lignite, bituminous and anthracite types of coal. The data further includes all four of the proximate variables, the ultimate analysis as well as the corresponding GCV values in MJ/kg. In total, the data represents the coal fields located in KwaZulu-Natal, Mpumalanga, Limpopo and the Free State, with lesser regard to Gauteng, the North West Province and the Eastern Cape.

Table 4-1 gives a summary of the data obtained for case study 1. All the samples include the proximate analysis (%), the ultimate analysis (%) and the corresponding GCV (MJ/kg).

Table 4-1: Summary of obtained data for case study 1

Source	Sample size	Coal types included	Geographical origin (RSA provinces)	Sample information included?		
				Proximate analysis	Ultimate analysis	GCV
Industrial plant (S1)	1'650	- Lignites - Bituminous - Anthracite	- KwaZulu-Natal - Mpumalanga - Limpopo - Free State	Yes	Yes	Yes
Third party (S2)	750	- Bituminous	- Mpumalanga - North West	Yes	Yes	Yes
Coal bulletins (S3)	480	- Lignites - Bituminous - Anthracite	- Gauteng - North West - Eastern Cape	Yes	Yes	Yes

¹Details on the third party cannot be made public due to confidentiality agreements.

4.2.3 Inspect data

The gathered data was inspected to ensure that the dataset is representable and of good quality, ultimately promoting consistent results from reliable data obtained through verified methods. The bomb calorimeter from source 1 and 2 (S1 and S2) is calibrated which gives external assurance that the GCV are accurately measured. The sampling and analysis procedures followed to obtain the proximate and ultimate variables also follow the ASTM standards[5], [9], [14]. The data from source 3 (S3) is published data and believed to be reliable and an accurate representation of the real data.

The data quality is further inspected for completeness by investigating the dataset for any abnormalities (missing or incorrect data). Only the proximate variables and corresponding GCVs are inspected since it was already established in Chapter 3 that a model will be developed on these variables.

Two samples of data were observed to have missing information within S1. These data samples had the proximate variables available, but not the corresponding GCV. Using these samples would not provide the necessary information needed to develop a model. These two data samples were therefore removed from the dataset.

Data obtained from the respective sources was further evaluated for incorrect data. The three sources' data was combined into a single large dataset. Each of the proximate variables were respectively plotted against the corresponding GCV, as can be seen in Figure 4-2 (a) to (d).

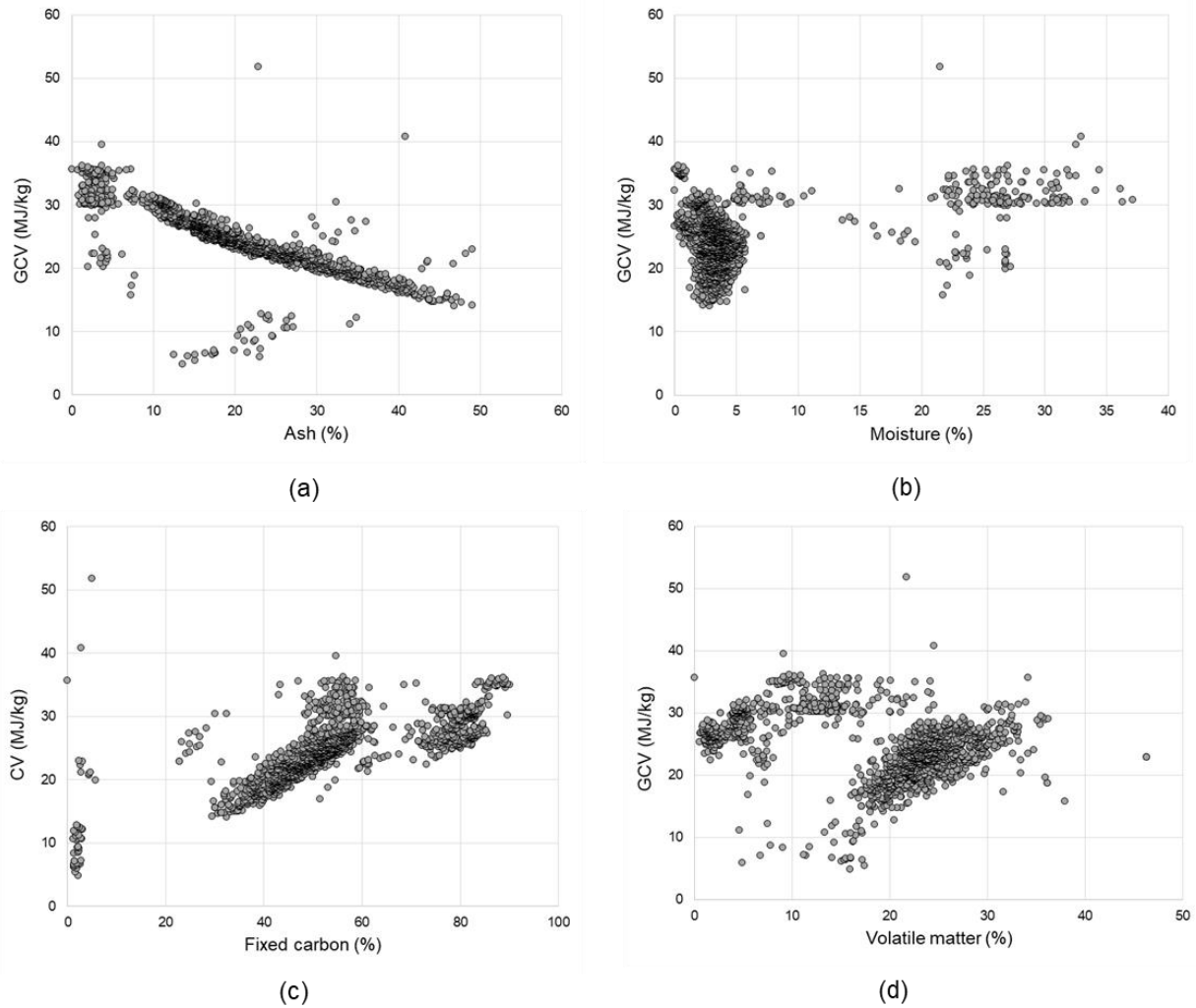


Figure 4-2: Proximate analysis versus GCV relationship on all three sources' data

From Figure 4-2 it is observed that there are two distinct groups of data. This phenomenon was investigated, and a difference in the reported data was found. Data from S1 contained total moisture (TM), whereas the first and third sources' data contained inherent moisture (IM). The reason for this is possibly that the second source's samples were analysed on an *as-received* basis and incorrectly labelled as *air-dried* basis.

Analyses done on an *as-received* basis and on an *air-dried* basis give different percentage compositions of each of the proximate and ultimate variables [5]. The variables cannot be converted to air-dried basis either, since the IM moisture for these samples is unavailable.

These datasets can therefore not be combined and still be classified as a complete dataset. It was decided to include only the IM and not the TM in the final dataset since the IM presents a better representation of the coal [5], [9], [16]. The second source's data was thus removed from the dataset.

Figure 4-3 (a) to (d) repeats the plots of the proximate variable and the corresponding GCV, with the second source's data removed.

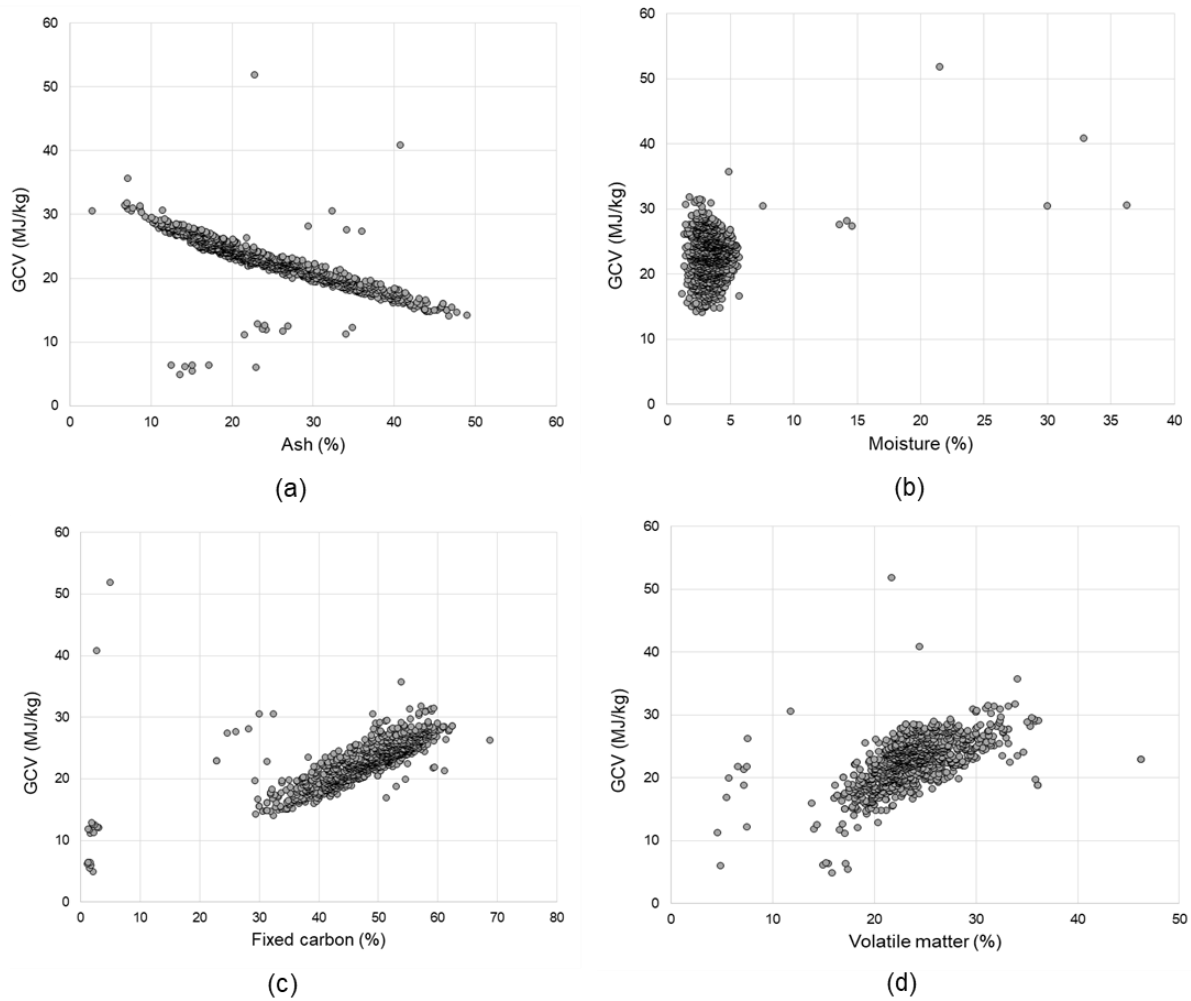


Figure 4-3: Proximate analysis versus GCV relationship on source 1 and 3 data

Removing a complete data source from the dataset could have a significant impact on the sample size. The sample size was therefore tested using Green's method, as discussed in Chapter 3. This new dataset contained 2'098 samples and still complied with Green's minimal sample size of 84. The dataset was therefore deemed a representable and complete sample set.

4.2.4 Clean data

Before the modelling procedure could be started, MATLAB was used to identify possible outliers in the new dataset. The *is-outlier* build-in function in MATLAB was used and 45 potential outliers were identified. The reason behind the identification as an outlier was investigated before data was removed. After evaluation, 43 of the identified outliers were

removed. Two flagged outliers could not be identified as true outliers and therefore kept in the final sample.

Table 4-2 gives a sample of the list of identified outliers. The full list can be found in Appendix A.1. Table 4-2 also gives the reason for the outlier, as well as whether or not the outlier was removed.

Table 4-2: Sample list of the potential outliers

Proximate variable				GCV (MJ/kg)	Reason for outlier	Removed
A (%)	IM (%)	FC (%)	VM (%)			
13.6	68.4	2.1	15.9	24.82	Abnormal high IM	Yes
15.0	3.5	41.5	11.4	29.6	Sample seems normal	No
33.4	8.6	31.2	26.8	71.8	Abnormal high GCV	Yes
31.1	2.2	59.2	7.5	6.8	Low GCV, high FC	Yes
31.6	2.3	59.5	6.6	21.81	Sample seems normal	No

The main reasons for the outliers were found to be a high moisture content in 34 of the potential outliers. Upon further evaluation of the data, it was found that these 34 samples were part of S1's data and were obtained from a different mine from the rest of this source's data. A possible explanation could be that these samples were incorrectly labelled with inherent moisture, but contained total moisture. These points were removed from the dataset due to their high moisture values that were not comparable with the other samples.

Additional reasons were identified with the evaluation of the proximate values. Some of the proximate variables exceeded 100% which could logically not be the case. This is a common mistake that can happen when recording the results for an analysis [10]. These points were deemed outliers and were removed from the dataset.

A third reason was identified based on the normal expected values that the samples should be. Several samples included a different GCV than what was expected. In most of these cases the GCV were low compared to other samples that had similar high values for fixed carbon. These were therefore deemed as outliers and removed from the dataset.

The final dataset contained 2'055 datasets, which complied with Green's minimum sample size. Figure 4-4 (a) to (d) illustrates each of the proximate variables against the respective GCV of the clean dataset.

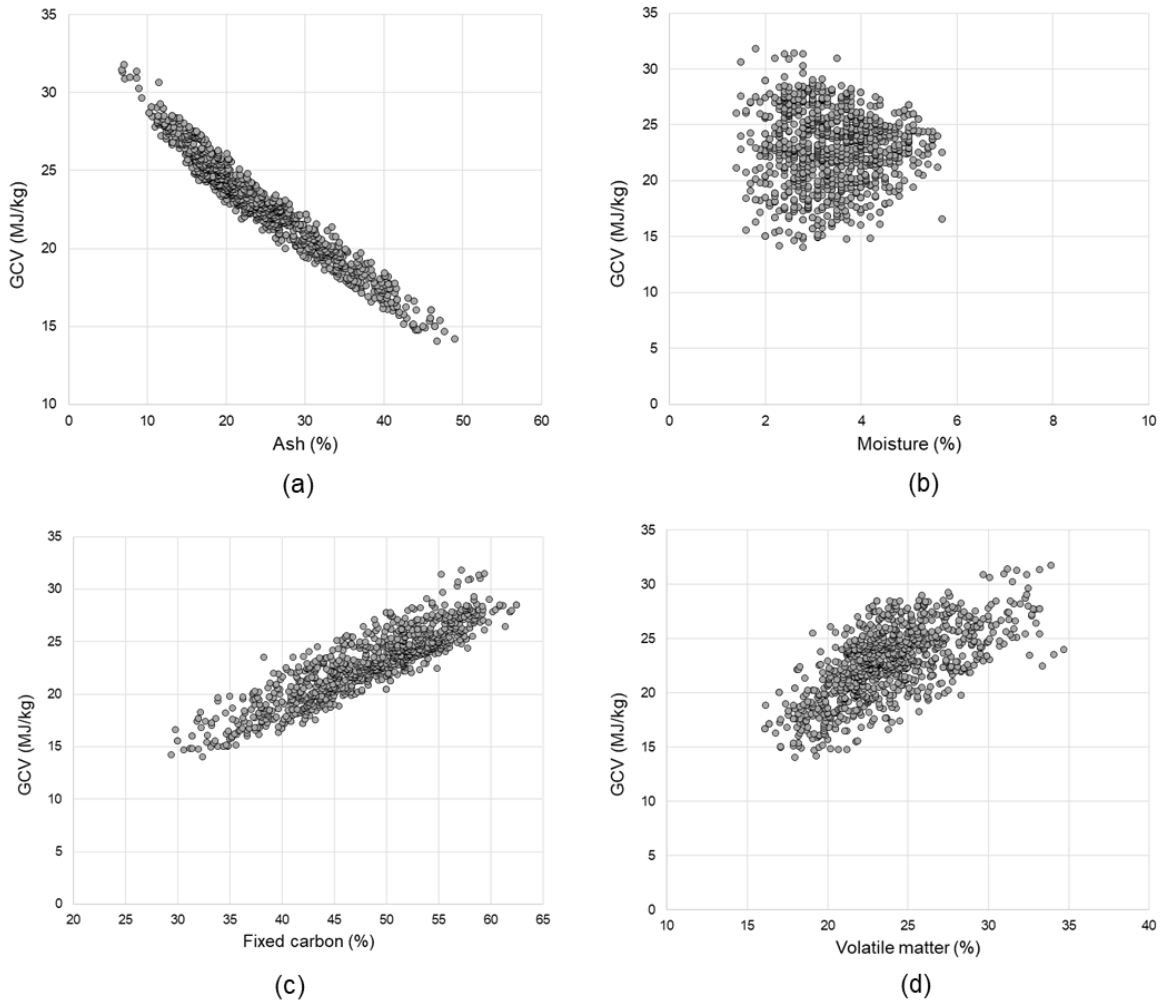


Figure 4-4: Proximate analysis versus GCV relationship on the clean dataset

The clean dataset was considered a better representation of the real relationship and can therefore be used for modelling.

4.2.5 Split data into modelling and validation datasets

The final sub-step for data preparation was to split the data into modelling and validation datasets. This was done to ensure that the different sets of data all represent the total dataset, allowing development and validation of the models using different datasets.

To split the data, the complete clean set was firstly randomised with the *Rand* function in Microsoft Excel™, after which, the first 30% and the last 30% of the data, consisting of 617

samples each, were kept aside for the respective first and second validation sets. The remaining 40% of the data (821 samples) was used to develop the models.

Figure 4-5 (a) to (d) again show the relationship between the proximate variables and the corresponding GCV with the clean data. However, Figure 4-5 distinguishes between the different validation sets along with the modelling set, illustrating that the three individual sets were representative of the total set.

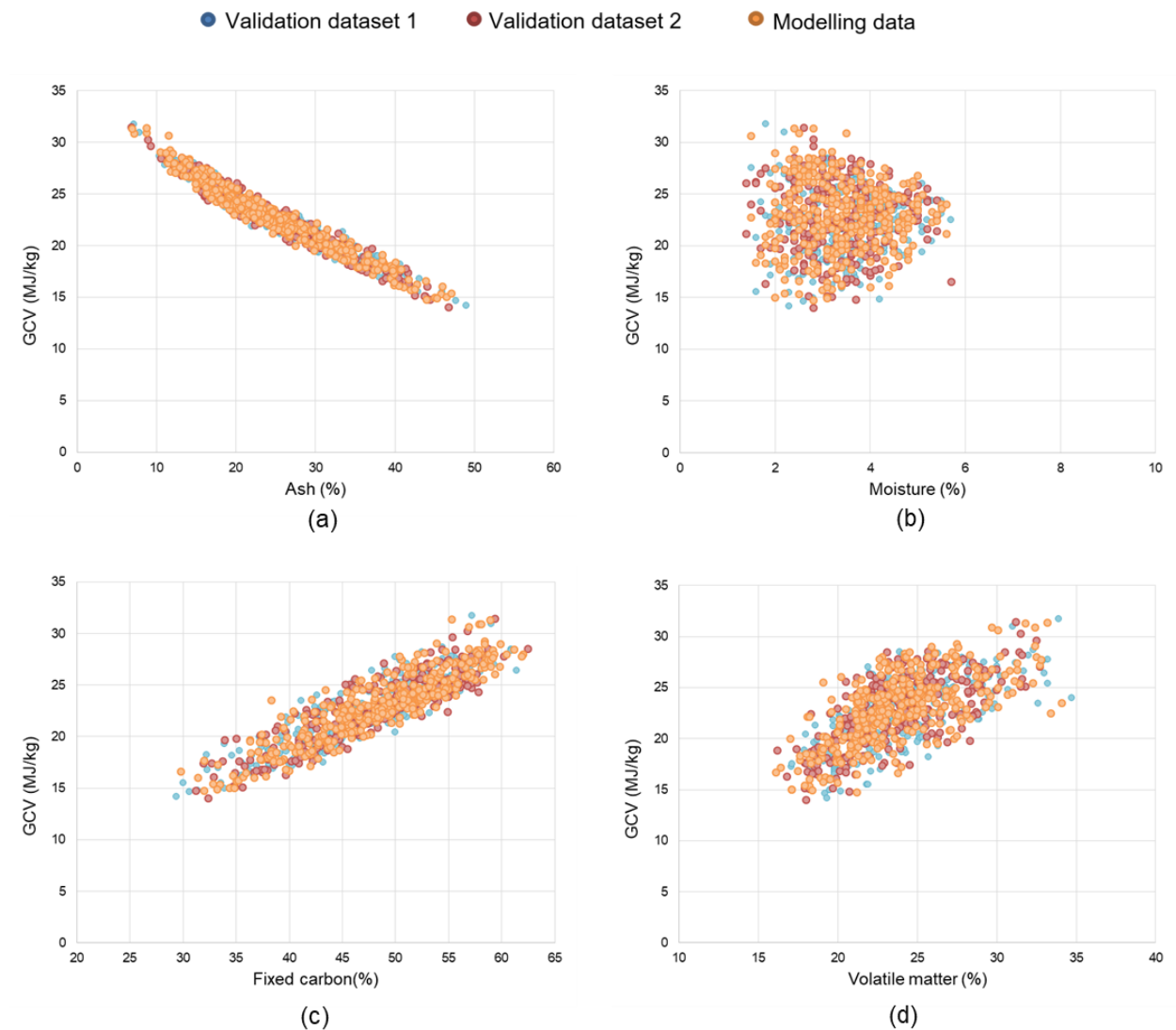


Figure 4-5: Proximate analysis versus GCV relationship on the data split

The blue and red dots illustrate Validation dataset 1 and 2 respectively, while the orange dots represent the modelling data.

The modelling set was used to develop a new model for GCVs in Section 4.3, whereas the validation sets were used to validate the model in the Model validation focus area (Section 4.4).

4.2.6 Summary of case studies

Data was obtained from various sources for each of the three case studies. These datasets were analysed to evaluate the quality of the data and to ultimately use the data for modelling purposes. This section focussed specifically on the data for case study 1. The detailed data preparation process for case study 2 and case study 3 data can be found in Appendix A.1. A summary is presented in Table 4-3 to indicate the final outcome of this focus area.

Table 4-3: Case studies data summary

Case study (CS)	Gather and understand data		Clean data		Usable dataset	
	Number of data sources	Samples obtained	Abnormalities identified	Outliers removed	Final number of samples	Comply with Green's minimal sample
CS1	3	2850	752	43	2055	Yes
CS2	3	254	0	6	248	Yes
CS3	5	243	0	2	241	Yes

Significant high quality datasets were obtained for each of the case studies and new models can now be developed during the next focus area.

4.3 Model development

In this section, relevant results from case study 2 will be presented and discussed. Data from Indian coal were used in this case study to develop GCV models. Additional information can be found in Appendix A.2.

4.3.1 Model type

In the previous section, proximate analysis data was plotted on several instances against the corresponding GCV values. In most instances a form of linear relationship was noted between the variables and the GCV. Due to this linearity the method of multiple linear regression was chosen to develop each of the case studies' models. Multiple linear regression will capture the

linear relationship while ensuring that the effect of each respective variable influences the model accordingly, as Majumder *et al.* already confirmed [18].

4.3.2 Identify variables

A developed model needed to include the variables that would present a reasonable representation of the reality. Literature indicated that the variables from the proximate analysis are commonly used, due to their availability and ease of access. Models developed with the approach presented in this study include the four proximate variables. This ensured that the effect of the different variables was captured and that the models were similar to other literature models. All models developed with this approach would therefore be multiple linear regression models which use the four proximate variables as inputs.

4.3.3 Develop and test models

A multiple regression model was developed for each case study. This section showcases this sub-step with the results from case study 2.

The modelling dataset for case study 2 consisted of 100 samples which represented 40% of the data. These 100 samples were used to develop the regression model using the LINEST function in Excel™. Table 4-4 gives the LINEST matrix obtained for case study 2.

Table 4-4: LINEST matrix results

LINEST matrix				
$x_{VM} = 0.151$	$x_{FC} = 0.615$	$x_{IM} = 0.111$	$x_A = -1.564$	$C = 12.51$
$SE_{x_{VM}} = 0.021$	$SE_{x_{FC}} = 0.094$	$SE_{x_{IM}} = 0.003$	$SE_{x_A} = 0.103$	$SE_C = 0.743$
$R^2 = 0.945$	$RMSE = 0.411$	#N/A	#N/A	#N/A
$F = 2541.51$	$v_2 = 243$	#N/A	#N/A	#N/A
$SS_{reg} = 0.031$	$SS_{res} = 0.0741$	#N/A	#N/A	#N/A

The developed model for case study 2 is presented in Equation 4-1.

$$GCV = 12.51 - 1.564(A) + 0.111(IM) + 0.615(FC) + 0.151(VM) \quad (4-1)$$

- GCV - Calorific value in MJ/kg
- A - Ash content of the coal in %
- IM - Moisture content of the coal in %
- FC - Fixed carbon content of the coal in %
- VM - Volatile matter content of the coal in %

Where GCV is expressed in MJ/kg and A, IM, FC and V are the percentage of ash, inherent moisture, fixed carbon and volatile matter, respectively.

Evaluating the A regressors of the model, they were, as foreseen, negative. This is expected since the A negatively affects the GCV [15]. It can be seen in the data that the relationships between the A are negative with regards to the GCV, which validates the negative regressor associated with the A.

The IM are also expected to negatively influence the GCV [15, 17, 18]. However, the regressors associated with the IM are positive. The data was evaluated and the positive relationship was observed in the relationship between the IM and GCV as well. Figure 4-6 illustrates the A and IM relationships with the respective GCV.

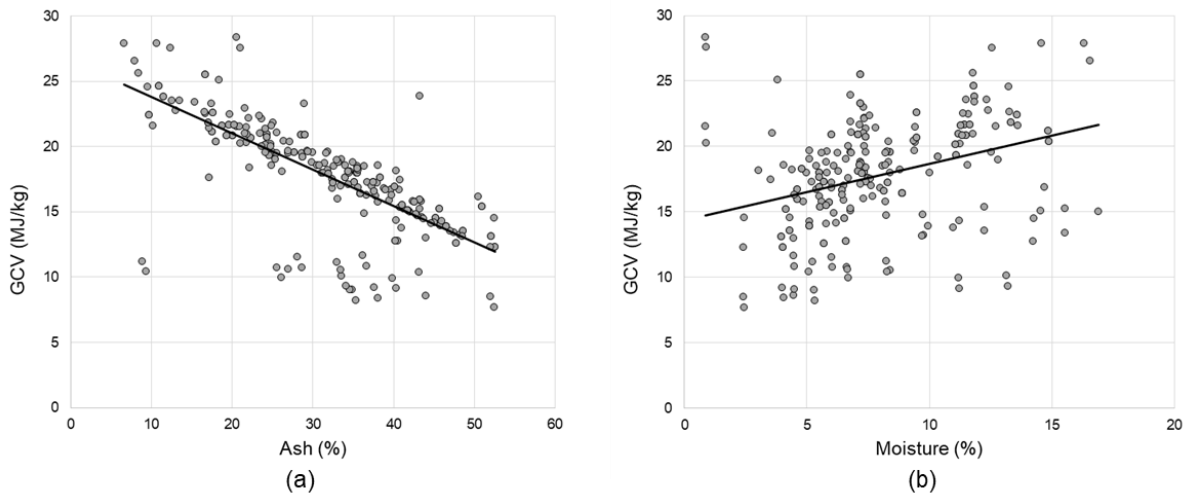


Figure 4-6: Relationship between A (a) and IM (b) vs the GCV

This phenomena was investigated at the hand of the $SE_{x_{IM}}$ value given in the LINEST matrix. The $SE_{x_{IM}}$ indicated that the influence of the IM on the overall error was small compared with

the variables influences. This implies that the impact of the IM regressor in the GCV was minimum and that it will not affect the viability of the model.

Further evaluation of Equation 4-1 shows the expected positive FC and VM regressors. The relationship between the FC and GCV and the VM and GCV was plotted (Figure 4-7) and the expected positive relationships (linear model) were observed.

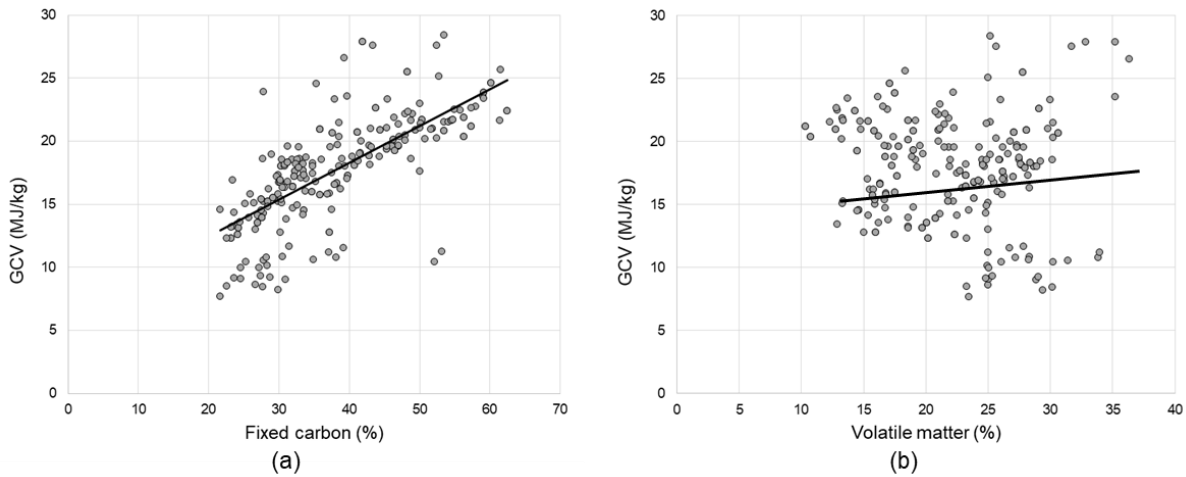


Figure 4-7: Relationship between FC (a) and VM (b) vs the GCV

From Figure 4-7, it can be seen that the comparison of the VM with the GCV is scattered. A clear positive relationship is not observed. The $SE_{x_{VM}}$ for Equation 4-1 is also very small which indicates that these values do not influence the GCV prediction very much.

The developed model was therefore deemed feasible because it met the other prescribed statistics reviewed in the next section.

4.3.4 Check requirements

A developed model does not necessarily constitute a viable outcome. The R^2 , RMSE and the MAE statistics were evaluated to ensure that the model was usable and presented accurate results.

The LINEST matrix given in Table 4-4 calculated the $R^2 = 0.945$ and the RMSE = 0.411%. The MAE of the developed model on the modelling data is calculated as 0.421%. These values were evaluated against the criteria presented in Chapter 2 and 3.

The evaluation showed that the model satisfied each of the criteria without areas for concern, indicating that using model would probably produce reasonable and accurate results.

4.3.5 Summary of case studies

This section presented details on the model developed for case study 2. The modelling process for the other two case studies can be found in Appendix A. Table 4-5 gives the results for each respective case study. The developed model is given, as well as the model's R^2 , RMSE and MAE. The final column indicates whether the model complied with the required criteria or not.

Table 4-5: Developed models with criteria for each case study

Case study model	Developed model	Results on modelling data			
		R^2	RMSE	MAE	Check
CS1 (Equation 5-1)	$GCV = 21.5 - 0.743(A) - 0.365(IM) + 0.512(FC) + 0.743(VM)$	0.994	0.47%	0.86%	Yes
CS2 (Equation 4-1)	$GCV = 12.51 - 1.564(A) + 0.111(IM) + 0.615(FC) + 0.151(VM)$	0.945	0.41%	0.42%	Yes
CS3 (Equation 5-3)	$GCV = 24.25 - 2.31(A) + 0.121(IM) + 0.356(FC) + 0.123(VM)$	0.953	0.55%	1.12%	Yes

Models for the three case studies produced high R^2 values and acceptable RMSE and MAEs. The models were tested against set criteria and passed without concerns. Further validation is required to ensure that the model produces results that are trustworthy and have a high likelihood of being close to reality.

4.4 Model validation

In this section, relevant results from case study 3 will be presented and discussed. The model developed from Alaskan coal was validated to indicate its representation of reality. Additional information can be found in Appendix A.3.

4.4.1 Model validation

The model developed according to the previous section needed to be validated. Validation was performed to assess whether the predicted GCVs were reasonable with respect to the experimented GCVs.

The two sets of 30% validation data were used to calculate two MAE and the CC parameters (one for each validation set). The average between the two parameters was then used as the reporting answer.

The MAE between the experimented and the predicted GCVs was calculated as 2.63%, indicating that the predicted GCV were relatively similar to the respective experimental counterpart. This then indicated that the model was likely to estimate trustworthy and reliable GCVs.

The relationship between the experimented and predicted GCVs was plotted and the CC was determined as 0.942. Figure 4-8 shows this relationship.

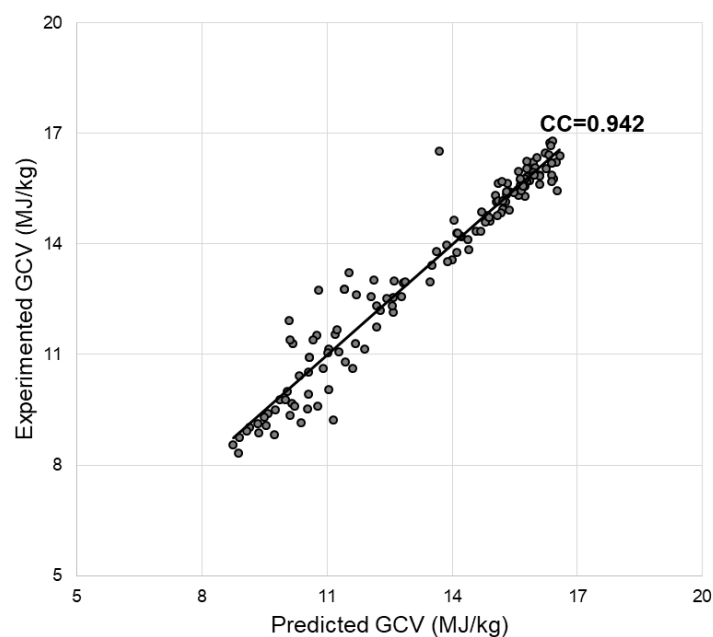


Figure 4-8: Validation of the developed model

Both the aforementioned results complied with the desired criteria which established the validity of the model.

4.4.2 Summary of case studies

Validating of the models formed a critical component of the approach. The validation was performed as part of the approach for case studies 1, 2 and 3. Each of the developed models was validated using its respective two validation datasets. Table 4-6 gives a summary of the validation results for each of the case studies.

Table 4-6: Case studies validation results

Model validation parameter	Case study 1		Case study 2		Case study 3	
	Average between validation sets	Applicable?	Average between validation sets	Applicable?	Average between validation sets	Applicable?
MAE	0.26%	Yes	1.36%	Yes	2.63%	Yes
CC	0.989	Yes	0.891	Yes	0.942	Yes

Models were developed and validated using the best practice modelling approach. It was found that the models represented the respective datasets and outcomes sufficiently. This approach generated several additional models that can be used but does not give an easy indication towards a preferred choice. The following section aims to provide a mechanism that could assist with such decisions.

4.5 Model comparison

The objective of this section of the study is to compare the developed and literature models' performance. The comparison will assist in selecting the most suitable model for a given dataset. This section therefore aims to provide a visual manner to assist with such decision making. The results from all three case studies were combined to illustrate the benefit of visualisation.

Two visualisations are presented in this chapter, *Visualisation 1* and *Visualisation 2*. The first visualisation addresses the validity of the various models. This allows for the comparison of the designed performance of each model. The second visualisation focusses on the model's reliability. This indicates the model's ability to predict the GCV of a new coal dataset and ultimately answers the main problem identified in this study: **Will the literature models be applicable to a new coal dataset and which one would be best?**

4.5.1 Visualisation 1

The aim of *Visualisation 1* is to see if the developed models perform better, worse or similar to the literature models. For this visualisation the limits used will be the R^2 of the models and the MAE of the models on their a) modelling data and b) validation data.

Table 4-7 presents these parameters as reported by the literature models. The rows highlighted in orange imply that these models did not use MAE as a validation parameter, but

other statistical parameters such as EES or ABE. Several of the models also did not give any results, other than the model's R^2 , on the modelling data. These cells are also coloured in orange within Table 4-7. For this reason, only the models that reported on MAE specifically will be used for the comparison.

Table 4-7: Reported R^2 and MAE parameters of the literature models

Referencing GCV model	Model type	Reported results		
		R^2 Modelling (a) and Validation (b)	MAE Modelling (a)	MAE Validation (b)
Literature model 1: <i>Majumder et al. (2008)</i> [18]	Regression	0.980	-	1.49%
Literature model 2: <i>Patel et al. (2007)</i> [21]	ANN	0.987	1.23%	2.07%
Literature model 3: <i>Gulec & Gulbandilar</i> (2018) [22]	Regression	0.999	1.27%	1.29%
	ANN	0.946	1.12%	1.18%
Literature model 4: <i>Parikh et al. (2005)</i> [23]	Regression	0.995	3.74%	5.51%
Literature model 5: <i>Behnamfard & Alaei</i> (2016) [12]	Regression	0.965	-	-
	ANFIS	0.999	-	-
Literature model 6: <i>Nhuchen & Afzal</i> (2008) [24]	Regression 1	0.896	-	1.38%
	Regression 2	0.921	-	1.13%
Literature model 7: <i>Verma et al. (2010)</i> [15]	Regression	0.995	2.85%	5.31%
	CANFIS	0.987	0.73%	0.55%
Literature model 8: <i>Mesroghli et al. (2009)</i> [25]	Regression 1	0.970	0.28%	1.14%
	Regression 2	0.990	0.17%	0.50%
	Regression 3	0.995	0.06%	0.46%
Literature model 9: <i>Huang et al. (2008)</i> [19]	Regression 1	0.870	1.90%	3.2%
	Regression 2	0.880	1.85%	2.8%
	ANN	0.896	1.10%	1.21%
Literature model 10: <i>Akkaya et al. (2009)</i> [20]	Regression 1	0.967	-	3.94%
	Regression 2	0.967	-	3.79%
	Regression 3	0.969	-	3.79%
Literature model 11: <i>Mason & Gandi</i> (1994) [26]	Regression	0.923	-	-
Literature model 12: <i>Kock & Franzidis</i> (1973) [16]	Regression 1	0.880	-	2.56%
	Regression 2	0.840	-	2.11%

Referencing GCV model	Model type	Reported results		
		R^2 Modelling (a) and Validation (b)	MAE Modelling (a)	MAE Validation (b)
	Regression 3	0.910	-	2.06%

The same parameters were also calculated for each of the case studies in Section 4.4 and are presented in Table 4-8.

Table 4-8: Calculated R^2 and MAE parameters of the developed models

Developed models for each case study	Model type	Calculated results		
		R^2 (of model)	MAE(modelling)	MAE (validation)
CS1 developed model: South African coal (Equation 5-1)	Regression	0.994	0.86%	0.26%
CS2 developed model: Indian coal (Equation 4-1)	Regression	0.945	0.42%	1.36%
CS3 developed model: Alaskan coal (Equation 5-3)	Regression	0.953	1.12%	2.63%

These parameters are visually presented on a scatterplot with the R^2 and the respective MAE's chosen as the limiting parameters. In Chapter 3, the limits for these parameters were identified as the average values obtained from all of the literature models. These parameters and their respective limits are calculated and listed:

R^2	-	0.90
MAE (modelling)	-	1.5%
MAE (validation)	-	3.0%

Figure 4-9 shows the respective R^2 and MAE (modelling) results of each model. The green box represents the area which will satisfy both the identified limits of the R^2 and MAE (modelling) parameters. Figure 4-9 also indicates the optimal performance of these models, as it represents the data used to develop the model.

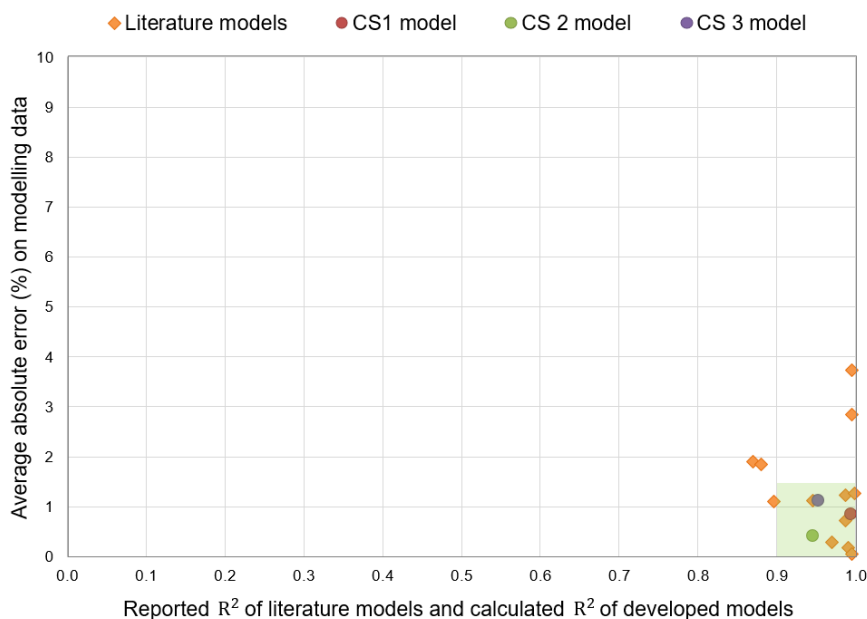


Figure 4-9: Visualisation 1(a) – comparison on published modelling data

From Figure 4-9 it can be seen that the developed models (coloured dots) perform similarly to the literature models (orange diamonds) which validates the developed models. The devised approach given in Chapter 3 is therefore verified.

The models' performance on their respective validation data are presented in Figure 4-10. This figure indicates the expected performance for these models on a new dataset. The green box again represents the area which will satisfy both the identified limits of the R² and MAE as calculated on each model's validation datasets.

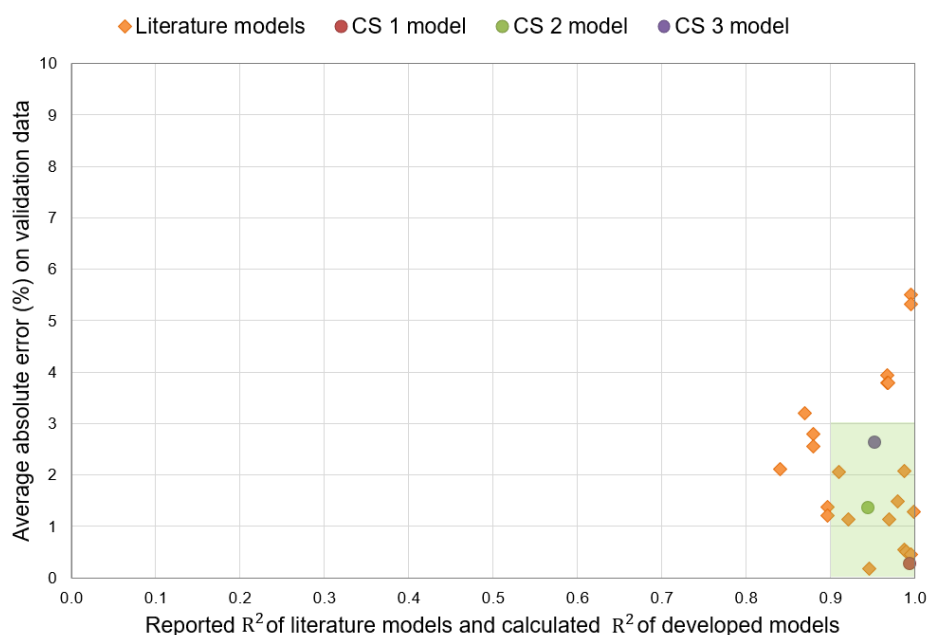


Figure 4-10: Visualisation 1(b) – comparison on published validation results

From Figure 4-10 it is expected that the use of the models within the green box will give a MAE of less than 3% when used on a new dataset.

4.5.2 Visualisation 2

A significant difference between each of the models was the data it was based on. This could create the scenario where each model could be suitable for its specific data but would not produce similar results on another dataset. For this reason, each of the models (literature and developed) was applied to a new dataset, thus investigating how each model would be able to predict the GCV on a completely different dataset.

The new dataset was obtained from a published article by Beamish [90], which contains 92 New Zealand and Australian coal analyses. This dataset contains the proximate and ultimate analysis (on air-dried and as-received basis), as well as the corresponding GCVs in MJ/kg. The samples covered a wide range of coal fields and coal types.

The aim of *Visualisation 2* is to evaluate the performance of the GCV models on this new dataset. As mentioned in Section 4.5.1, an MAE of less than 3% is expected. For this validation, the CC of the predicted GCVs against the experimented GCVs is calculated and used as a limiting parameter.

Table 4-9 gives the CC and the MAE of each model (developed and literature) on the new dataset. The validation data of this new dataset was used. The CC and MAE parameters could not be calculated for the ANNs, ANFIS and CANFIS models and are therefore not included.

Table 4-9: Calculated CC and MAE on the new dataset

Referencing GCV model	Model type	New results	
		CC	MAE
CS1 developed model: <i>South African coal</i> (Equation 5-1)	Regression	0.989	2.89%
CS2 developed model: <i>Indian coal</i> (Equation 4-1)	Regression	0.806	4.32%
CS3 developed model: <i>Alaskan coal</i> (Equation 5-3)	Regression	0.723	4.01%
Literature model 1: <i>Majumder et al. (2008) [18]</i>	Regression	0.895	3.12%
Literature model 2: <i>Patel et al. (2007)[21]</i>	ANN		
Literature model 3: <i>Gulec & Gulbandilar (2018) [22]</i>	Regression	0.448	9.32%
	ANN		
Literature model 4: <i>Parikh et al. (2005) [23]</i>	Regression	0.249	24.32%
Literature model 5: <i>Behnamfard & Alaei.(2016) [12]</i>	Regression	0.861	8.95%
	ANFIS		
Literature model 6: <i>Nhuchen & Afzal (2008) [24]</i>	Regression 1	0.546	6.20%
	Regression 2	0.411	8.64%
Literature model 7: <i>Verma et al. (2010) [15]</i>	Regression	0.856	36.99%
	CANFIS		
Literature model 8: <i>Mesroghli et al. (2009) [25]</i>	Regression 1	0.765	7.63%
	Regression 2	0.339	6.32%
	Regression 3	0.249	15.35%
Literature model 9: <i>Huang et al. (2008) [19]</i>	Regression 1	0.961	2.80%
	Regression 2	0.469	11.35%
	ANN		
Literature model 10: <i>Akkaya et al. (2009) [20]</i>	Regression 1	0.652	35.60%
	Regression 2	0.689	29.99%
	Regression 3	0.905	63.89%
Literature model 11: <i>Mason & Gandi (1994) [26]</i>	Regression	0.564	9.50%
Literature model 12: <i>Kock & Franzidis (1973) [16]</i>	Regression 1	0.652	9.10%
	Regression 2	0.721	8.26%
	Regression 3	0.778	6.98%

The data in Table 4-9 can be visually represented on a scatterplot with the CC and MAE limits of 0.90 and 3% respectively. Figure 4-11 illustrates *Visualisation 2*, where the green box indicates the CC and MAE limits.

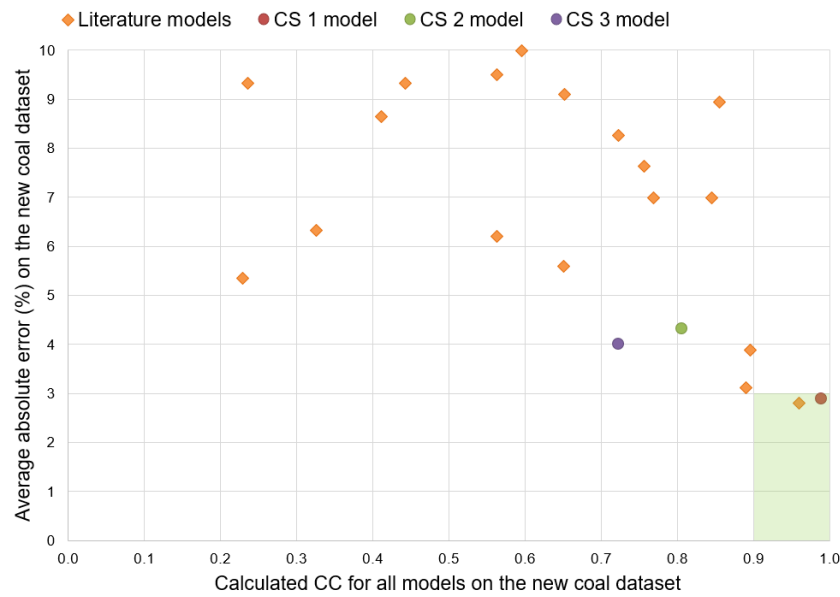


Figure 4-11: Visualisation 2(a) - comparison on new data

Figure 4-11 shows how the models reacted on the new dataset. Figure 4-11 shows the CC and MAE results of the various models (developed models represented by the dots, literature models represented by the diamonds). It was expected that the respective models would give a CC and MAE that fell within the desired ranges (green box), irrespective of the dataset.

It was found that almost all of the models gave results outside these expected limits. Figure 4-11 illustrates this visually with numerous results all over plot. Literature model 9 (Huang *et al.* [19]) and the developed model for case study 1 (South African coal) are within the green box. These two models could therefore be identified as more suitable for implementation on the specific dataset. It would still be recommended to test these models before using them.

The scatter observed in the model's performance indicates that it would not be feasible to blindly use any existing GCV model on a new dataset. For this reason, the approach devised in Chapter 3 was used to develop a fourth model.

The fourth model was developed specifically on this New Zealand and Australian coal dataset. The model development and validation results are in Appendix B. The R^2 and MAE of this new model on its validation data are 0.943 and 1.63% respectively. Figure 4-12 illustrates the performance of the various other models on the new dataset along with this fourth model (*new model* illustrated with the black triangle).

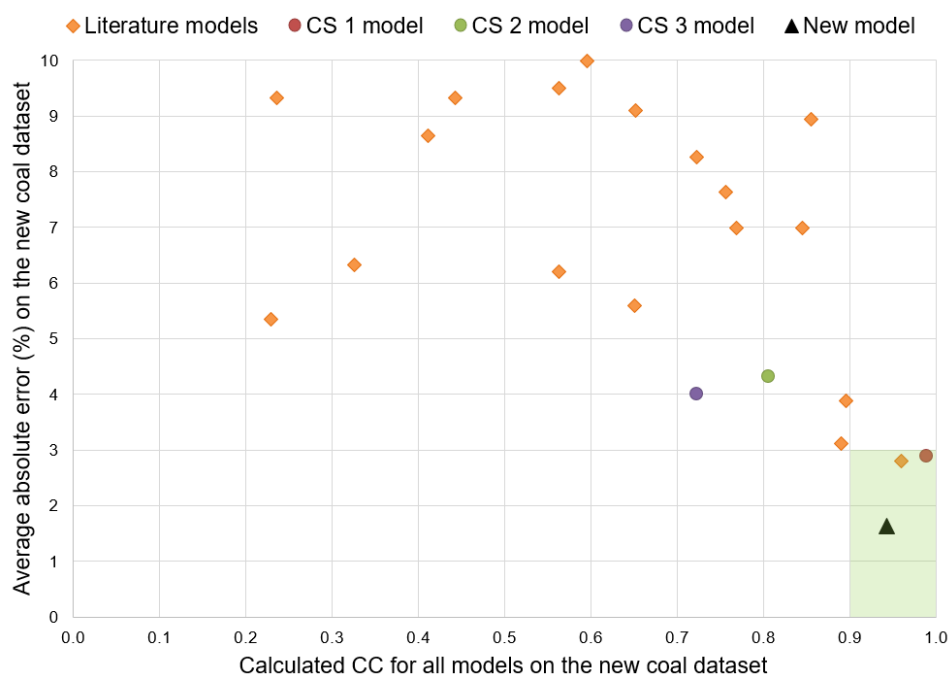


Figure 4-12: Visualisation 2(b) – Comparison on new data with new model

This visualisation and comparison showed that a GCV model would not be suitable to be used blindly on any dataset and that it is advisable to rather develop a new model for a new dataset.

4.6 Conclusion

This chapter described how the devised approach in Chapter 3 was practically applied to obtain results that would satisfy the objectives of the study (presented in Chapter 1). The approach was verified in this chapter by applying it to three case studies. The data for each case study covered a wide range of coal properties for South African, Indian and Alaskan coal.

Detailed results on the data preparation, model development and model validation focus area for each case study are presented in Appendix A. However, in this chapter only significant observations were highlighted with respect to each focus area of the methodology.

For the *data preparation* focus area, relevant results from case study 1 were presented. The importance of understanding a dataset was accentuated and a high quality dataset was prepared by removing errors and outliers.

High quality datasets enabled the accurate development of new models. Case study 2's results were presented for the *model development* focus area. A multiple linear regression model was used to estimate GCV using the four proximate variables. The model was tested against statistical criteria (R^2 , RMSE and MAE) and passed without concerns.

The developed model for case study 3 was then additionally validated using a specific dataset (validation data) to ensure that it is an accurate representation of reality. The main component of the validation was the comparison between the experimental and predicted GCVs. The validations indicated that the new model was trustworthy and would probably be suitable to estimate GCVs.

The developed models for all three case studies were then objectively compared with the literature models. Two visualisations were illustrated. The first visualisation compared the performance of the developed models with the predicted performance of the literature models. From this visualisation the conclusion could be made that the developed models operate in similar ranges to the literature models. The development of accurate and comparable models as a result verified the devised approach and methodology of this study.

The second visualisation focussed on addressing the problem statement given in Chapter 1: **Will the literature models be applicable to a new coal dataset and which one would be best?** Through this visualisation it was observed that the different models could not achieve the same performance as predicted on a new dataset. It was ultimately detected that a model developed for a specific dataset would be most suitable for that dataset.

The visualisation technique used, however, did provide a simplistic manner for illustrating the performance of the models. The results from this chapter showed that the best practice approach could effectively be applied to develop GCV models and enable objective comparison thereof.

CHAPTER 5 – CONCLUSION AND RECOMMENDATIONS

5 Conclusion and recommendations

5.1 Preamble

This study was conducted to address the uncertainty regarding the applicability of existing GCV models on new coal datasets. The need to devise a general approach to model and assess coal energy content was therefore addressed. This study focussed on meeting this need by addressing various objectives set out in Chapter 1.

Chapter 2 gave context on a best practice approach to model development. The literature models were reviewed with the aim of identifying the main steps required to develop a representable GCV model. The knowledge obtained in the literature study was applied in Chapter 3 to deliver a general approach that could be used to develop and compare GCV models. In Chapter 4 this devised approach was verified by applying it to three case studies which produced three validated models.

A conclusion on the study is given in this chapter. Section 5.2 and 5.3 will provide an overview of the study and demonstrate how the required objectives were met. Recommendations for further studies will be proposed in Section 5.4. Finally, the study will be closed in Section 5.5.

5.2 Overview of study

This study showed that multiple literature models exist that can predict the GCVs of coal based on its proximate and ultimate variables. A problem was introduced when the applicability of these models was questioned on a new dataset. The need to objectively compare the performance of the literature models on a new dataset was therefore highlighted. This raised a second problem, “*how does one develop a GCV model?*” It was hence suggested to devise a general approach that could be used to develop GCV models and to ultimately compare the performance of the various models on a new dataset.

A literature study was compiled on the best practice approach for modelling and, together with the methods used by each of the literature models, a new approach was devised in Chapter 3. The new approach consisted of four focus areas: *data preparation*, *model development*, *model validation* and *model comparison*. Each of these focus areas consists of various sub-steps applicable specifically to GCV modelling.

The approach was applied to three case studies. For each case study a qualitative dataset was compiled and used to develop a multiple linear regression model, dependent on the four proximate variables. A validation dataset was used to validate each of the developed models.

It was concluded in *Visualisation 1* that the developed models performed similarly to the literature models, with R^2 and errors in the same ranges. The developed models are therefore deemed fit for GCV prediction through validation. The development of successful GCV models as a result verified the general approach devised in Chapter 3.

Whether these models would work on a new dataset was still a problem that had to be addressed. For *Visualisation 2*, a new dataset was introduced. All the GCV models (developed and literature) were applied to this new dataset. It was ultimately found that most of the models gave errors much larger than expected. It is therefore advisable to develop a new GCV model for the new coal dataset as opposed to using existing GCV models.

The results from this study showed that the devised approach could effectively be applied to develop GCV models and enable objective comparison thereof.

5.3 Meeting the objectives

The primary objective of this study was to devise a general approach that could be used to develop GCV models on new datasets and objectively compare the results from various models. In order to achieve this, the following objectives had to be met:

- provide relevant guidelines regarding the model development process,
- provide relevant research regarding techniques for comparison,
- devise a general approach for GCV model development,
- verify the approach by applying it to case studies, and
- objectively assess and compare the performance of the various models on a new dataset.

The remaining part of this section will demonstrate how all of the objectives, as stated in Section 1.4, were achieved.

5.3.1 Guidelines regarding model development

The basic requirements for model development were provided in the literature study (Chapter 2) for which a general approach for model development could be evaluated (Section 2.2). This was supplemented by investigating methodologies that were followed by

each of the literature models. Subsequently, the most applicable focus areas and sub-steps were consolidated as the relevant guidelines for model development.

5.3.2 Research regarding techniques for comparison

Section 2.5 provided the basic research regarding methods to objectively compare multiple models. Several visualisation techniques were introduced and discussed with regard to its application on comparison methods, highlighting the usefulness of combining individual techniques to present a holistic overview.

5.3.3 Devise a general approach for GCV model development

In Chapter 3, an approach to GCV modelling was devised. The approach was devised based on the relevant literature provided in Chapter 2. The devised approach consisted of four focus areas, namely data preparation, model development, model validation and model comparison. Each of these focus areas comprised various sub-steps consolidated to specifically address the modelling aspects of coal's energy content.

5.3.4 Verify the approach with case studies

The devised approach was applied to three case studies in Chapter 4. Each case study was discussed in detail in Appendix A, while the relevant results of each case study were presented in Chapter 4. In the end three new models were developed and validated with a given set of criteria.

5.3.5 Compare and assess the performance of the models on a new dataset

The developed models were assessed on a new dataset to ultimately compare their performance with the literature models (also on the new dataset). The results were visually presented on scatter plots in Section 4.3 to assist in the objective comparison, enabling several conclusions that could be made based on the results obtained in Chapter 4.

5.4 Recommendation for further research

This section provides recommendations for further studies. Five recommendations are listed below and discussed in more detail subsequently. The execution of these recommendations could further improve the results of this study.

- Develop and compare models specific to coal types and operational sites
- Include the ultimate analysis in the models for comparison
- Investigate the effect of the number of samples used for modelling
- Effect of integrating different data sources

- Signs associated with each regressor and relationship between the variables

5.4.1 Develop and compare models specific to coal types

This study focused on coal from different regions (e.g. South Africa, India and Alaska). It was found that it is advisable to develop new models specific to a region's coal. It is recommended to limit the evaluation further by only focussing on specific coal types for each region. Comparisons can be made on the applicability of the various type-specific models within a specific region. The various region's type-specific models can also be evaluated and compared.

5.4.2 Include the ultimate analysis in the models for comparison

The GCV models developed for each case study included the effect of the four proximate variables only. The decision was based on Majumder *et al.*'s [18] conclusion on the variables affecting GCV, as discussed in Chapter 3. However, it is recommended to include the effect of the ultimate analysis when comparison will be made with models containing these variables. This will allow for a better and more objective comparison between the various models.

5.4.3 Investigate the effect of the number of samples used for modelling

In Chapter 4 it was seen that case study 1's developed models gave predictions on the new dataset that were well within the desired ranges. This can possibly be due to the large dataset available for modelling. It is suggested to evaluate the effect of the number of samples used to develop a model.

5.4.4 Effect of integrating different data sources

Combining data from various sources to compile an integrated dataset can be challenging. Several alternative methods exist that can be used to compile a multiple dataset. It is recommended to identify and apply the best method that can be used to compile a representative dataset. It is also recommended to further evaluate the reliability and validity of the datasets used within these models.

5.4.5 Regressor signs and variable relationships

Within the developed models, the regressor signs did not correlate with the relationship between the individual parameters. The standard error of the regressors was found to be small and it was therefore recommended to accept the sign none the less. This scenario can be further evaluated by evaluating the reliability and validity of the data to identify the effect of parameters within multiple linear regression models.

Additionally, this study can be applied to different datasets to re-evaluate the findings in Chapter 4.

5.5 Closure of study

In this study a general approach was devised to develop and assess models estimating coal energy content. This was achieved by providing a well-defined problem statement and clear objectives for the study. Thereafter, relevant literature was reviewed and the best practices for model development and comparison were identified. The knowledge obtained from the literature study was then used to devise a general approach for GCV model development. This approach was verified by applying it to three different case studies. The results from the case studies were visually presented to allow for the objective comparison of the results of the various models.

This chapter concluded the document by providing an overview of this study. Furthermore, it was stated how the objectives of the study were met and recommendations were also provided for further studies.

REFERENCE LIST

- [1] E. Mathews, Burt, R. W. Stanton, C. S. Singh, and R. van Liere, "A comprehensive overview of coal," *World Coal Inst.*, vol. 6, no. 1, pp. 1–44, 2005.
- [2] H. Hagen and P. Robinson, "Derivation of methodology capable of identifying suitable collectors for coal flotation by using surface dependant techniques," Cape Peninsula University of Technology, 2002.
- [3] T. Yoshida, M. Lino, T. Takanohashi, and T. Katoh, "Study on thermoplasticity of coals by dynamic viscoelastic measurement: effect of coal rank and comparison with Gieseler fluidity.," *Energy Policy*, vol. 79, no. 29, pp. 339–405, 2000.
- [4] South African Bureau of Standards (SABS), "SANS 17246:2011 - Coal and Proximate Analysis," 2011.
- [5] South African Bureau of Standards (SABS), "SANS 10320:2004 - Evaluation of coal resources and coal reserves," 2004.
- [6] W. Hamer, W. Booysen, and E. H. Mathews, "A practical approach to managing uncertainty in the measurement and verification of energy efficiency savings," *South African J. Ind. Eng.*, vol. 28, no. 3, pp. 128–146, 2017.
- [7] South African Bureau of Standards (SABS), "SANS 13909-6: 2002 - Mechanical sampling : Coke Preparation of test samples." 2002.
- [8] S. Mauia, S. Kajitani, N. Suzuki, N. Ashizawa, and S. Hara, "CO₂ gasification rate analysis of coal char in entrained flow coal gasifier," *Fuel*, no. 85, pp. 163–169, 2006.
- [9] E. Furimsky, J. Ripmeester, and S. Gregg, "The classification of coal macerals and their relation to certain chemical and physical parameters of coal," *Geol. Soc. South Africa*, vol. 81, no. 2, pp. 95–102, 2005.
- [10] Q. P. Campbell, M. D. Barnardo, and J. R. Bunt, "Moisture adsorption and desorption characteristics of some South African coals," *J. South. African Inst. Min. Metall.*, vol. 113, no. 11, pp. 803–807, 2013.
- [11] Q. Feng, J. Zhang, X. Zhang, and S. Wen, "Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating

- conditional expectation and artificial neural network,” *Fuel Process. Technol.*, vol. 129, pp. 120–129, 2015.
- [12] A. Behnamfard and R. Alaei, “Estimation of coal proximate analysis factors and calorific value by multivariable regression method and adaptive neuro-fuzzy inference system (ANFIS),” *Int. J. Min. Geo-Engineering*, vol. 1, pp. 51–1, 2017.
- [13] V. Trent, J. H. Medlin, S. L. Coleman, and R. W. Stanton, “Chemical analyses and physical properties of 12 coal samples from the Pocahontas Field, Tazewell County, Virginia, and McDowell County, West Virginia,” *Geol. Surv. Bull.* 1528, pp. 1–37, 1982.
- [14] S. Falcon, R. M and P. Snyman, C, “An introduction to coal petrography: Atlas of petrographic constituents in the bituminous coals of Southern Africa.,” *Geol. Soc. South Africa.*, no. 483, pp. 1–11, 2007.
- [15] K. Verma, A, N. Singh, T, and M. Monjezi, “Intelligent prediction of heating value of coal,” *Iran. J. Earth Sci.*, vol. 2, no. 10, pp. 32–38, 2010.
- [16] J. Kock and J. Franzidis, “The estimation of some properties of South African coals from proximate analyses,” *J. South African Inst. Min. Metall.*, no. 43, pp. 420–429, 1973.
- [17] A. Sharma, B. K. Saikia, and B. P. Baruah, “Maceral Contents of Tertiary Indian Coals and their Relationship with Calorific Values,” *Int. J. Innov. Res. Dev.*, vol. 1, no. 7, pp. 196–208, 2012.
- [18] A. K. Majumder, R. Jain, P. Banerjee, and J. P. Barnwal, “Development of a new proximate analysis based correlation to predict calorific value of coal,” *Fuel*, vol. 87, no. 13–14, pp. 3077–3081, 2008.
- [19] C. Huang, L. Han, X. Liu, and Z. Yang, “Models predicting calorific value of straw from the ash content,” *Int. J. Green Energy*, vol. 5, no. 6, pp. 533–539, 2008.
- [20] A. Akkaya, “Proximate analysis based multiple regression models for higher heating value estimation of low rank coals,” *Fuel Process. Technol.*, no. 90, pp. 165–170, 2009.
- [21] S. Patel, B. Kumar, and B. Yogesh, “Estimation of gross calorific value of coals using artificial neural networks,” *Fuel*, vol. 86, pp. 334–344, 2007.
- [22] M. Gulec and E. Gulbandilar, “Determination of the lower calorific and ash values of the

- lignite coal by using artificial neural networks and multiple regression analysis,” *Physicochem. Probl. Miner. Process.*, vol. 31, no. 6, pp. 12–23, 2018.
- [23] J. Parikh, S. A. Channiwala, and G. K. Ghosal, “A correlation for calculating HHV from proximate analysis of solid fuels,” *Fuel*, vol. 84, no. 10, pp. 487–494, 2005.
 - [24] D. Nhuchhen and M. Afzal, “HHV Predicting Correlations for Torrefied Biomass Using Proximate and Ultimate Analyses,” *Bioengineering*, vol. 4, no. 1, p. 7, 2017.
 - [25] S. Mesroghli, E. Jorjani, and S. Chehreh Chelgani, “Estimation of gross calorific value based on coal analysis using regression and artificial neural networks,” *Int. J. Coal Geol.*, vol. 79, no. 1–2, pp. 49–54, 2009.
 - [26] D. Gandhi and K. Mason, “Formulas for calculating the heating value of coal and coal char: development, tests and uses,” *Fuel*, vol. 34, no. 2, pp. 235–245, 1994.
 - [27] T. Allan and J. Peters, “Introduction to Regression Techniques,” *J. Mach. Learn. Res.*, vol. 8, no. 12, p. 144, 2011.
 - [28] R. Xu, “Modeling Process: Predictive model in insurance,” *Forecast. Futur.*, vol. 11, no. 8, 2013.
 - [29] A. V. Akkaya, “Proximate analysis based multiple regression models for higher heating value estimation of low rank coals,” *Fuel Process. Technol.*, vol. 90, no. 2, pp. 165–170, 2009.
 - [30] H. Schultheis, A. Singhaniya, and S. Chaplot, Devendra, “Comparing Model Comparison Methods,” *J. Comput. Sci. Eng.*, vol. 6, no. 8, pp. 1294–1299, 2017.
 - [31] C. Chen, H. Wolfgang, and A. Unwin, *Handbook of Data Visualization*. Springer, 1998.
 - [32] D. Snell, D. Norris, J. Heaton, C. Ramsey, and M. Lindstrom, “Forecasting and Futurism,” *Forecast. Futur.*, no. 7, p. 12, 2013.
 - [33] W. Booysen, L. A. Botes, and W. Hamer, “A practical methodology for the systematic identification of outliers,” in *Industrial and commercial use of energy conference (ICCUE)*, 2017, no. 1, pp. 1–6.
 - [34] G. Keeble and H. Wainer, “Data Quality Control,” *Am. Stat.*, vol. 38, no. 2, pp. 137–147, 2015.

- [35] K. Seervi, "Prediction of Calorific Value of Indian Coals By Artificial Neural Network," National Institute of Technology Rourkela, 2015.
- [36] M. Kuhn and K. Johnson, "Applied predictive modeling," *Appl. Predict. Model.*, pp. 1–600, 2013.
- [37] W. Booysen, "Measurement and verification of industrial DSM projects," Potchefstroom Campus of the North-West University, 2014.
- [38] United Nations Economic Commission, "Making data meaningful Part 2," in *Data and Metadata Reporting and Presentation Handbook*, UN, 2009, pp. 1–50.
- [39] D. van Prevelen, "Coal typology and conversion factors," *Appl. Energy*, no. 8, p. 514, 2007.
- [40] P. R. Holland, *Programming and Data Visualization Techniques*, 1st ed. New York: Apress Media, 2015.
- [41] P. B. Palmer and D. G. O'Connell, "Research Corner Regression Analysis for Prediction: Understanding the Process," *Cardiopulm. Phys. Ther. J.*, vol. 20, no. 3, pp. 23–26, 2009.
- [42] H. Wickham, D. Cook, and H. Hofmann, "Visualizing statistical models," *Ann. Appl. Stat.*, vol. 62, no. 09, pp. 1–21, 2000.
- [43] R. Cornish, K. Cybenko, Anne, F. E. Varesio, and G. V. S. Sharma, "An introduction to sample size calculations," *Math. Appl. Dev.*, vol. 2, pp. 1–5, 2006.
- [44] D. Zimmerman, "Model validation and verification of large and complex space structures," *Inverse Probl. Sci. Eng.*, vol. 8, no. 2, pp. 93–118, 2000.
- [45] D. P. MacKinnon, M. J. Valente, and I. C. Wurpts, "Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis," *Personal. Soc. Psychol. Rev.*, no. 3, pp. 111–132, 2018.
- [46] R. K. Gupta, "Regression models use in data splitting techniques," *J. Math.*, vol. 5, no. 5, pp. 53–56, 2013.
- [47] 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.
- [48] L. A. Botes, “Objective evaluation of industrial energy efficiency models for the RSA Section 12L tax incentive,” Potchefstroom Campus of the North-West University, 2017.
 - [49] I. Ben-gal, “Outlier Detection,” *Data Min. Knowl. Discov. Handb.*, pp. 131–146, 2005.
 - [50] U.S. Department of Energy, “International Performance Measurement and Verification Protocol,” vol. 1, no. September, 2002.
 - [51] R. Warren, R. Smith, and A. Cybenko, “Use of Mahalanobis distance for detecting outliers and outlier clusters in markedly non-normal data,” *J. Appl. Stat.*, vol. 81, no. 14, pp. 35–49, 2011.
 - [52] L. Bavoil and S. P. Callahan, “Interactive Multiple-View Visualizations,” *J. Sci. Comput. Imaging*, vol. 63, no. 3, pp. 162–169, 2005.
 - [53] T. McCune, “Data transformations,” in *Analysis of Ecological Communities*, Wiley, 2003, pp. 303–305.
 - [54] I. Murray, “Training, Testing, and Evaluating Different Models,” *J. Appl. Stat.*, pp. 1–4, 2016.
 - [55] A. Arseven, “Mathematical Modelling Approach in Mathematics Education,” *Univers. J. Educ. Res.*, vol. 3, no. 12, pp. 973–980, 2015.
 - [56] C. D. Thomas, “Successful application and evaluation of a Linear Regression model,” *J. Casualty Actuar. Sci.*, vol. 3, no. 8, pp. 1–13, 2008.
 - [57] R. Williams, “Review of Multiple Regression,” University of Notre Dame, 2015.
 - [58] M. Bremer, “Multiple Linear Regression,” in *Handbook of Regression Analysis*, 2013, pp. 223–231.
 - [59] M. Bremer, “Regression model development with Software packages,” in *Handbook of Regression Analysis*, 2013, pp. 632–639.
 - [60] M. Marioara and A. Kamer-Ainur, “Errors and Limitations associated with Regression and Correlation Analysis,” *J. Am. Stat. Assoc.*, vol. 82, no. 41, pp. 710–712, 2008.

- [61] E. Steyerberg, F. Harrell, and G. Borsboom, "Internal validation of predictive models," *J. Clin. Epidemiol.*, vol. 54, no. 8, pp. 774–781, 2001.
- [62] E. W. Steyerberg and F. E. Harrell, "Prediction models need appropriate internal, internal-external, and external validation," *J. Clin. Epidemiol.*, vol. 69, pp. 245–247, 2016.
- [63] E. W. Steyerberg *et al.*, "Poor performance of clinical prediction models: the harm of commonly applied methods," *J. Clin. Epidemiol.*, vol. 98, pp. 133–143, 2018.
- [64] M. Bonato, F. Sella, I. Berteletti, and C. Umiltà, "Neuropsychology is nothing without control: A potential fallacy hidden in clinical studies," *Cortex*, vol. 48, no. 3, pp. 353–355, 2012.
- [65] E. W. Steyerberg, M. Schemper, and F. E. Harrell, "Logistic regression modeling and the number of events per variable: Selection bias dominates," *J. Clin. Epidemiol.*, vol. 64, no. 12, pp. 1464–1465, 2011.
- [66] T. Debray and Y. Vergouwe, "A new framework to enhance the interpretation of external validation studies of clinical prediction models," *J. Clin. Epidemiol.*, vol. 68, no. 3, pp. 279–289, 2015.
- [67] E. O. Ogundimu, D. G. Altman, and G. S. Collins, "Adequate sample size for developing prediction models is not simply related to events per variable," *J. Clin. Epidemiol.*, vol. 76, pp. 175–182, 2016.
- [68] K. Callaghan, "The Correlation Coefficient," *Explor. Coll. Algebr. Discov. Databased Appl.*, no. c, pp. 553–557, 1996.
- [69] W. Hamer, "A practical approach to quantify RSA Section 12L EE tax incentives for large industry," no. November, pp. 1–268, 2016.
- [70] C. Martani, D. Lee, P. Robinson, R. Britter, and C. Ratti, "Studying the dynamic relationship between building occupancy and energy consumption," *Energy Build.*, vol. 47, pp. 584–591, 2012.
- [71] I. Birnbaum, "How to Determine and Verify Operating and Maintenance Savings in Energy Savings Performance Contracts," *Can. J. Stat.*, no. 62, pp. 185–188, 2017.

- [72] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann, "Multifield visualization using local statistical complexity," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1384–1391, 2007.
- [73] S. Baškarada and A. Koronios, "Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension," *Australas. J. Inf. Syst.*, vol. 18, no. 1, pp. 5–24, 2013.
- [74] R. Crooks, J. Lankow, and J. Ritchie, *Data Visualization: The Power of Visual Storytelling*, Wiley. Germany: Wiley, 2015.
- [75] M. Gleicher, D. Albers, and R. Walker, "Visual comparison for information visualization," *Appl. Predict. Model.*, vol. 10, no. 4, pp. 289–309, 2011.
- [76] A. Skhosana, "Atheletics South Africa (ASA) - Domestic competition rules and regulations," no. 1. pp. 1–198, 2018.
- [77] A. V Abela, *Communicating Successfully With Visual Presentations*, 1st ed. Extreme Presentations, 2010.
- [78] E. De Jonge, "Data visualisation," in *Statistics Methods*, 3rd ed., no. 201209, Netherlands: The Hague/Heerlen, 2012, pp. 1–46.
- [79] J. L. Vosloo and P. Snyman, C, "Data Presentation Types and Application," *J. Am. Stat. Assoc.*, no. 35, pp. 1–4, 2014.
- [80] C. Talmadge and J. Gale, "Introduction to Data Visualization Techniques Using Web-based Tools," *Natl. Inst. Stat.*, no. 23, pp. 1–15, 2013.
- [81] T. L. Schroeder and J. Lankow, "Visual Analysis Best Practices: Simple Techniques for Making Every Data Visualization Useful," *J. Stat. Appl.*, vol. 6, p. 41, 2014.
- [82] C. Blessing, H. Bradsher-fredrick, H. Miller, R. Miller, and R. Rutchik, "Cognitive Testing of Statistical Graphs," *J. energy*, no. 12, pp. 15–20, 2006.
- [83] T. L. Schroeder, B. B. Rainey, D. A. Goss, and T. P. Grosvenor, "Reliability of and comparisons among methods of measuring dissociated phoria," *Optom. Vis. Sci.*, vol. 73, no. 6, pp. 389–397, 1996.

- South Africa's electricity demand," *J. Energy South. Africa*, vol. 25, no. 4, pp. 49–54, 2014.
- [85] A. Kuznetsova, M. Monjezi, and R. S. Laramée, "Evaluating possible models for Multiple Regression," *J. Stat. Appl.*, vol. 6, pp. 12–19, 2016.
- [86] M. S. Krishnan, "Classification of coal," *J. Geol.*, vol. 6, no. 3, pp. 543–559, 1987.
- [87] L. B. van Heerden, "Critical analyses of coals," University of Pretoria, 2016.
- [88] T. P. Speed and B. Yu, "Model selection and prediction: Normal regression," *Stat. J. IAOS*, vol. 45, no. 1, pp. 308–327, 2012.
- [89] T. Maimane, "Department: Energy Republic of South Africa," *Coal resources*, 2018. [Online]. Available: http://www.energy.gov.za/files/coal_frame.html. [Accessed: 10-Aug-2018].
- [90] B. Beamish, "Proximate analysis of New Zealand and Australian coals by thermogravimetry," *J. Geol. Geophys.*, vol. 37, no. 4, pp. 387–392, 1994.
- [91] P. Antonia, N. R. Cheepurupalli, G. V. S. Sharma, C. S. Singh, and E. N. D. Rao, "Petrographic Analysis and its Relationship with Calorific Values of Non- Coking Coals," *Int. J. Eng. Trends Technol.*, vol. 30, no. 1, pp. 13–20, 2015.
- [92] G. Lopez-Campos, "Presenting data for Biomedical Informatics," in *Centre of Research Excellence in Patient Safety*, 2015, pp. 16–29.
- [93] F. Barnes, "Coal Resources of Alaska," *Contrib. to Econ. Geol.*, pp. 1–63, 1967.

APPENDIX A – CASE STUDY RESULTS

A. Case study results

A.1 Case study 1: South African coal

A.1.1 Dataset preparation

Gather data

For case study 1, a total of 2'850 samples was obtained from three different reliable sources. The data is summarised as follows:

- 1'650 samples from an industrial plant in Mpumalanga from 2013 to 2016. This plant gets its coal from mines across South Africa.
- 720 samples from a third party², covering data from 2009 up until 2016. These samples are taken from various mines across South Africa.
- 480 samples from multiple South African coal bulletins. These bulletins are freely available on the Department of Energy of the Republic of South Africa's website [89].

The coal data from all three of these sources is on an *air-dried* basis (contains IM only) and includes lignite, bituminous and anthracite types of coal. The data further includes all four of the proximate variables, the ultimate analysis as well as the corresponding GCV values in MJ/kg. In total, the data represents the coal fields located in KwaZulu-Natal, Mpumalanga, Limpopo and the Free State, with lesser regard to Gauteng, the North West Province and the Eastern Cape.

Table A-1 gives a summary of the data obtained. All the samples include the proximate analysis (%), the ultimate analysis (%) and the corresponding GCV (MJ/kg).

Table A-1: Summary of obtained data

Source	Sample size	Coal types included	Geographical origin (RSA provinces)	Sample information included?		
				Proximate analysis	Ultimate analysis	GCV
Industrial plant (S1)	1'650	- Lignites - Bituminous - Anthracite	- KwaZulu-Natal - Mpumalanga - Limpopo - Free State	Yes	Yes	Yes

²Details on the third party cannot be made public due to confidentiality agreements.

Source	Sample size	Coal types included	Geographical origin (RSA provinces)	Sample information included?		
				Proximate analysis	Ultimate analysis	GCV
Third party (S2)	750	- Bituminous	- Mpumalanga - North West	Yes	Yes	Yes
Coal bulletins (S3)	480	- Lignites - Bituminous - Anthracite	- Gauteng - North West - Eastern Cape	Yes	Yes	Yes

Inspect data

The gathered data was inspected to ensure that the dataset is representable and of good quality, ultimately promoting consistent results from reliable data obtained through verified methods. The bomb calorimeter from source 1 and 2 (S1 and S2) is calibrated which gives external assurance that the GCV is accurately measured. The sampling and analysis procedures followed to obtain the proximate and ultimate variables also follow the ASTM standards [5], [9], [14]. The data from source 3 (S3) is published data and believed to be reliable and an accurate representation of the real data.

The data quality is further inspected for completeness by investigating the dataset for any abnormalities (missing or incorrect data). Only the proximate variables and corresponding GCVs are inspected since it was already established in Chapter 3 that a model will be developed on these variables.

Two samples of data were observed to have missing information within S1. These data samples had the proximate variables available, but not the corresponding GCV. Using these samples would not provide the necessary information needed to develop a model. These two data samples were therefore removed from the dataset.

Data obtained from the respective sources was further evaluated for incorrect data. The three sources' data was combined into a single large dataset. Each of the proximate variables were respectively plotted against the corresponding GCV, as can be seen in Figure A-1 (a) to (d).

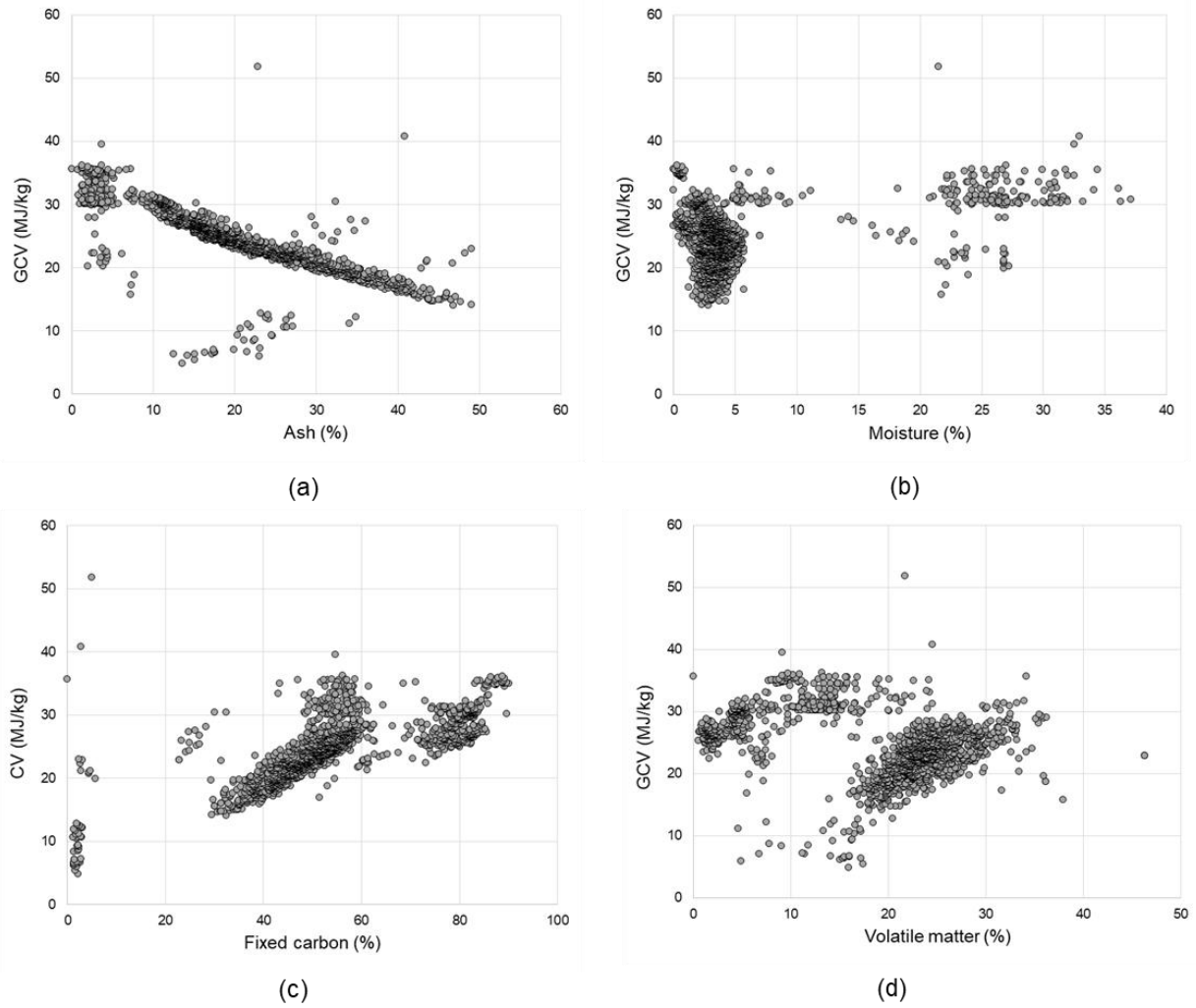


Figure A-1: Proximate analysis versus GCV relationship on all three source's data

From Figure A-1 it is observed that there are two distinct groups of data. This phenomenon was investigated, and a difference in the reported data was found. Data from S1 contained total moisture (TM), whereas the first and third sources' data contained inherent moisture (IM). The reason for this is possibly that the second source's samples were analysed on an *as-received* basis and incorrectly labelled as *air-dried* basis.

Analyses done on an *as-received* basis and on an *air-dried* basis give different percentage compositions of each of the proximate and ultimate variables [5]. These datasets can therefore not be combined and still be classified as a complete dataset. It was decided to include only the IM and not the TM in the final dataset since the IM presents a better representation of the coal [5], [9], [16]. The second source's data was thus removed from the dataset.

Figure A-2 (a) to (d) repeats the plots of the proximate variable and the corresponding GCV, with the second source's data removed.

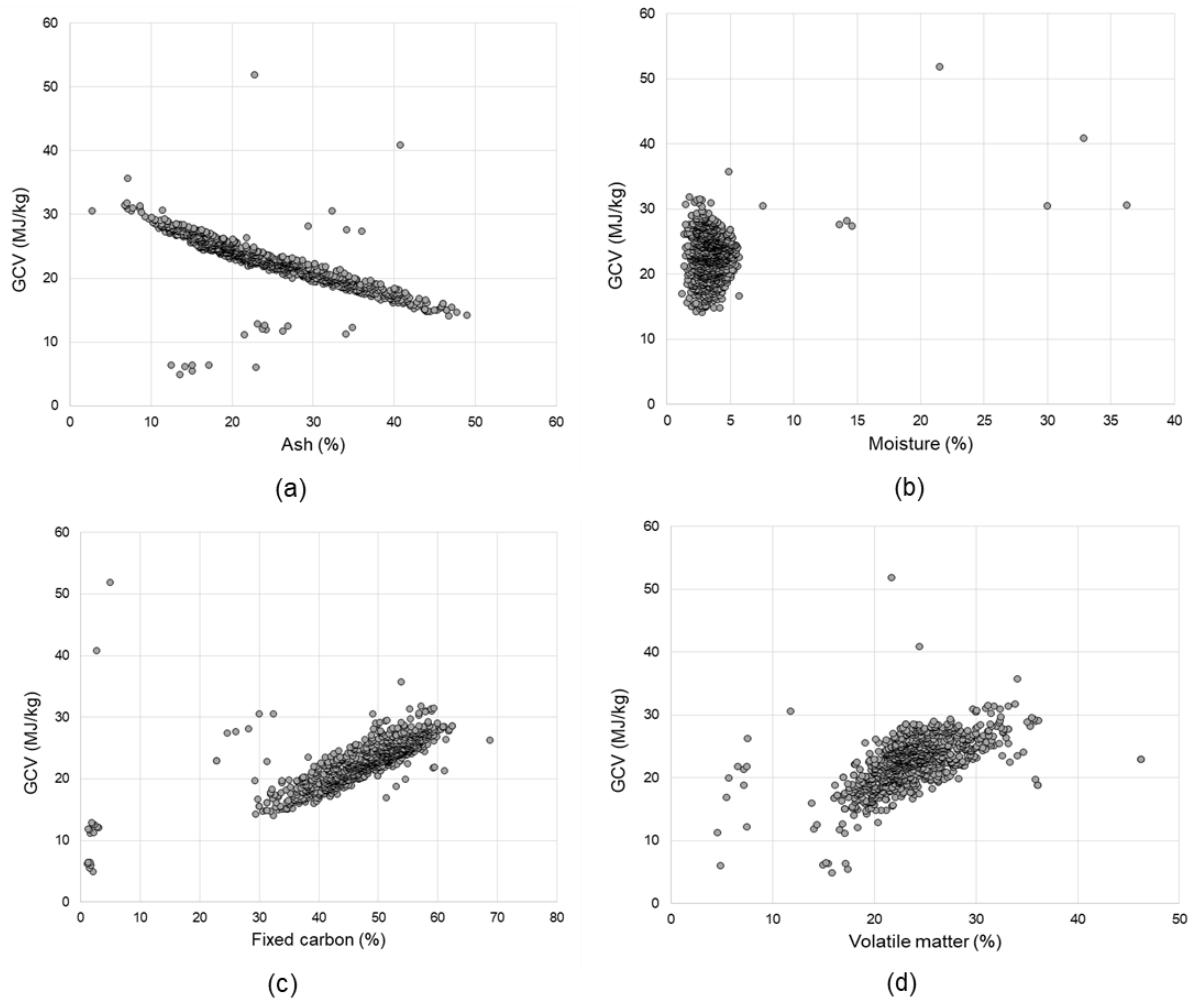


Figure A-2: Proximate analysis versus GCV relationship on source 1 and 3 data

Removing a complete data source from the dataset could have a significant impact on the sample size. The sample size was therefore tested using Green's method, as discussed in Chapter 3. This new dataset contained 2'095 samples and still complied with Green's minimum sample size of 84. The dataset was therefore deemed a representable and complete sample set.

Clean data

Before the modelling procedure could be started, MATLAB was used to identify possible outliers in the new dataset. The *is-outlier* build-in function in MATLAB was used and 45 potential outliers were identified. The reason behind the identification as an outlier was investigated before data was removed. After evaluation, 43 of the identified outliers were

removed. Two flagged outliers could not be identified as true outliers and therefore kept in the final sample.

Table A-2 gives a list of identified outliers. Table A-2 also gives the reason for the outlier, as well as whether or not the outlier was removed.

Table A-2: Full list of the potential outliers

Proximate variables				GCV(MJ/kg)	Reason for outlier	Removed?
A (%)	IM (%)	FC (%)	VM (%)			
68.4	15.9	2.1	13.6	4.82	High moisture	Yes
65.9	17.4	1.6	15.1	5.4	High moisture	Yes
70.3	4.9	1.8	23	5.96	High moisture	Yes
69.6	15	1.2	14.2	6.14	High moisture	Yes
66.5	17.2	1.3	15.1	6.36	High moisture	Yes
65.5	15.5	1.7	17.2	6.36	High moisture	Yes
70.9	15.3	1.3	12.5	6.39	High moisture	Yes
67	15.5	1.2	16.3	6.61	High moisture	Yes
63.8	16	2.6	17.6	6.62	High moisture	Yes
62.1	14.1	2.3	21.5	6.7	High moisture	Yes
64.6	16	1.9	17.5	6.73	High moisture	Yes
32.1	11.4	41.5	15.0	18.6	Sample seems normal	No
64.6	16	1.9	17.5	6.73	High moisture	Yes
62.9	11.2	2.8	23.1	7.22	High moisture	Yes
67.4	9	1.3	22.3	8.38	High moisture	Yes
64.9	11.8	2.2	21.1	8.5	High moisture	Yes
67.4	7.8	2.3	22.5	8.67	High moisture	Yes

Proximate variables				GCV(MJ/kg)	Reason for outlier	Removed?
A (%)	IM (%)	FC (%)	VM (%)			
58.8	14.3	2.3	24.6	9.16	High moisture	Yes
61.4	16.2	2	20.3	9.32	High moisture	Yes
57	16.3	2.2	24.5	9.38	High moisture	Yes
60.8	16.5	1.9	20.7	10.36	High moisture	Yes
56.3	15.5	2.2	26	10.6	High moisture	Yes
59.8	17.2	1.1	21.9	10.62	High moisture	Yes
54.7	16	2.9	26.4	10.62	High moisture	Yes
56.8	13.3	2.8	27.1	10.75	High moisture	Yes
59.5	17.1	1.7	21.6	11.08	High moisture	Yes
59	4.6	2.3	34.1	11.19	High moisture	Yes
55.5	16.6	1.6	26.3	11.71	High moisture	Yes
60.2	14.1	1.3	24.2	11.85	High moisture	Yes
54.7	18.4	3.1	23.8	12.07	High moisture	Yes
54.7	7.5	2.9	34.9	12.19	High moisture	Yes
56.3	14.4	2.4	26.9	12.47	High moisture	Yes
57	16.9	2	24.1	12.6	High moisture	Yes
54.6	20.4	1.9	23.2	12.81	High moisture	Yes
32.9	24.5	2.7	40.8	40.8	High moisture	Yes
21.5	21.7	5	22.8	61.8	Abnormal high GCV	Yes
2.2	7.5	59.2	31.1	5.8	Low GCV, high FC	Yes
2.3	6.6	59.5	31.6	21.81	Sample seems normal	No

Proximate variables				GCV(MJ/kg)	Reason for outlier	Removed?
A (%)	IM (%)	FC (%)	VM (%)			
2.1	7.2	61.2	29.5	21.28	Low GCV, high FC	Yes
5.1	5.7	54.6	34.6	19.89	Low GCV, high FC	Yes
2.1	7.2	53.1	37.6	18.77	Low GCV, high FC	Yes
1.2	5.5	51.4	41.9	16.88	Low GCV, high FC	Yes
59	4.6	84.3	0	4.1	Calculation error	Yes
62.1	14.1	63.1	21.5	6.7	Calculation error	Yes
64.6	16	45.7	17.5	6.73	Calculation error	Yes

The main reasons for the outliers were found to be a high moisture content in 34 of the potential outliers. Upon further evaluation of the data, it was found that these 34 samples were part of S1's data and were obtained from a different mine than the rest of this source's data. A possible explanation could be that these samples were incorrectly labelled with inherent moisture, but contained total moisture. These points were removed from the dataset due to this high moisture value that was not comparable with the other samples.

Additional reasons were identified with the evaluation of the proximate values. Some of the proximate variables exceeded 100% which could logically not be the case. This is a common mistake that can happen when recording the results for analysis. These points were deemed outliers and were removed from the dataset.

A third reason was identified based on the normal expected values that the samples should be. Several samples included a different GCV than what was expected. In most of these cases the GCV was low compared with other samples that had similar high values for fixed carbon. These were therefore deemed as outliers and removed from the dataset.

The final dataset contained 2'089 datasets, which complied with Green's minimum sample size. Figure A-3 (a) to (d) illustrates each of the proximate variables against the respective GCV of the clean dataset.

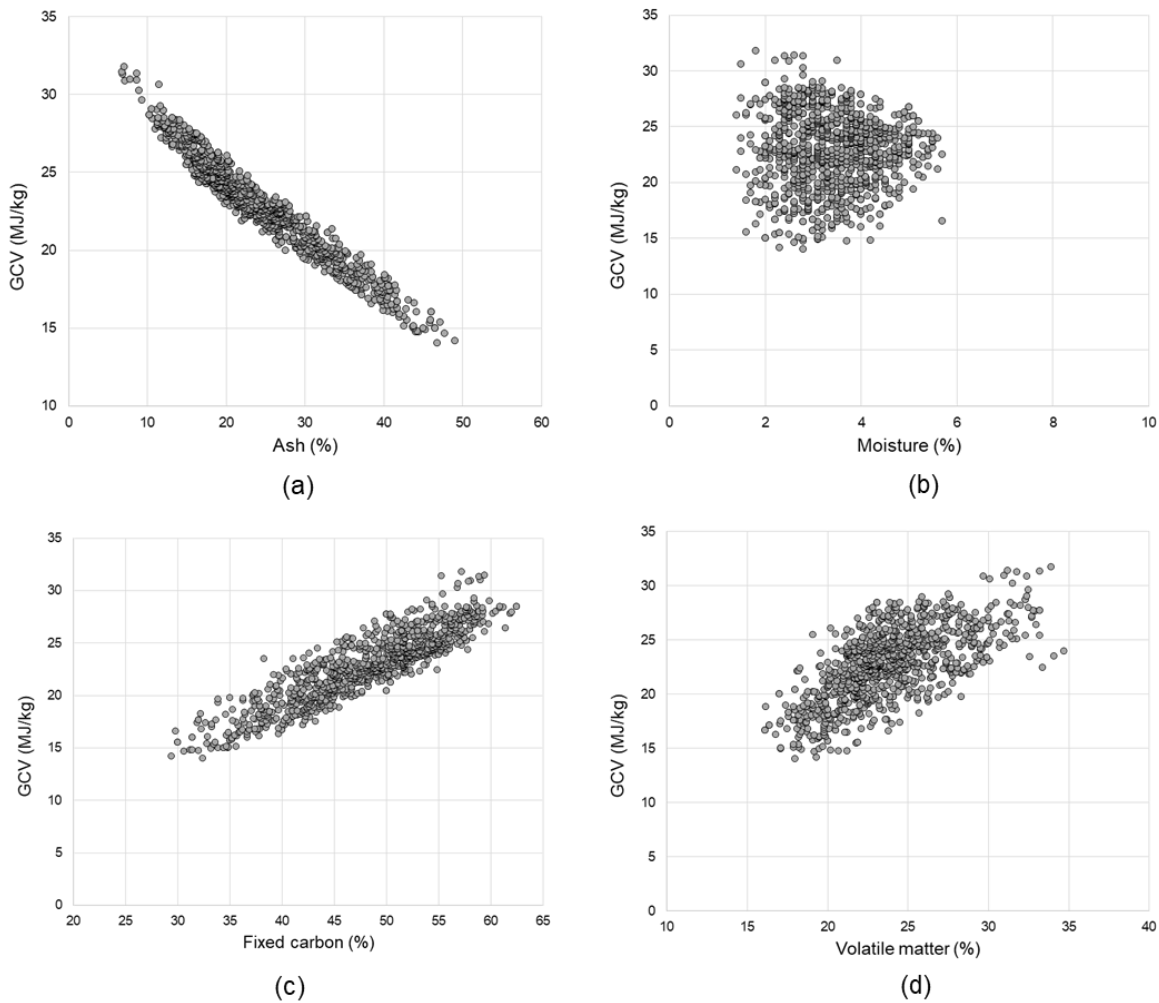


Figure A-3: Proximate analysis versus GCV relationship on the clean dataset

The clean dataset was considered a better representation of the real relationship and can therefore be used for modelling.

Split data into modelling and validation data

The final sub-step for data preparation was to split the data into modelling and validation datasets. This was done to ensure that the different sets of data all represent the total dataset, allowing development and validation of the models using different datasets.

To split the data, the complete clean set was firstly randomised with the *Rand* function in Microsoft Excel™, after which the first 30% and the last 30% of the data, consisting of 617 samples each, were kept aside for the respective first and second validation sets. The remaining 40% of the data (821 samples) was used to develop the models.

Figure A-4 (a) to (d) again show the relationship between the proximate variables and the corresponding GCV with the clean data. However, these figures distinguish between the

different validation sets along with the modelling set, illustrating that the three individual sets were representative of the total set.

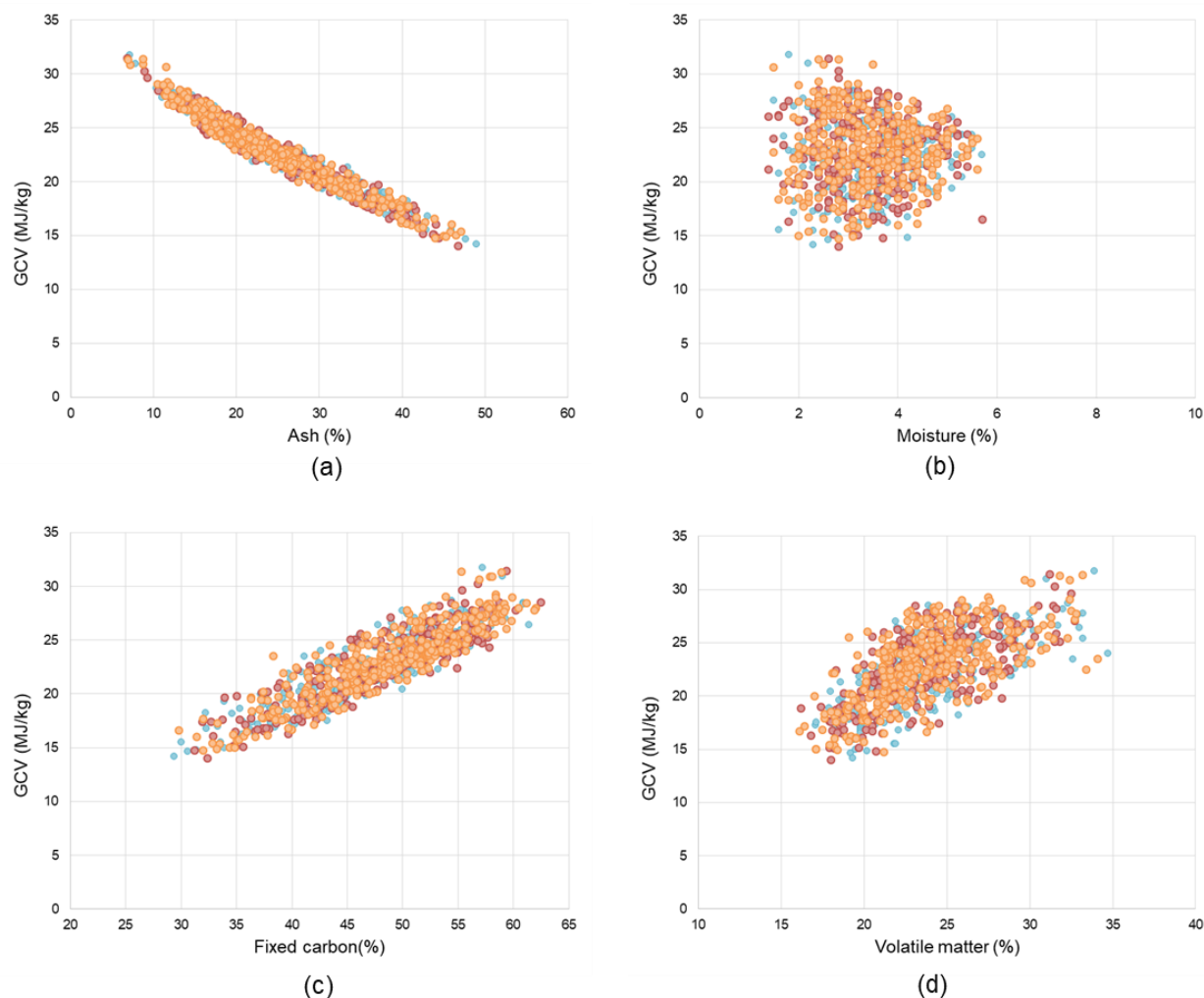


Figure A-4: Proximate analysis versus GCV relationship on the data split

The blue and red dots illustrate Validation dataset 1 and 2 respectively, while the orange dots represent the modelling data.

The modelling sets were used to develop a new model for GCVs in Section 4.3, whereas the validation sets were used to validate the model in the Model validation focus area (Section 4.4).

A.1.2 Model development

Type selection

The relationship between the proximate variables and the GCVs seen in Figure A-3 shows a linear relationship. For this reason, it was decided to develop a linear regression model for this dataset.

Identify constraints

Literature indicated that the variables from the proximate analysis are commonly used, due to their availability and ease of access. Models developed with the approach presented in this study include the four proximate variables. This ensured that the effect of the different variables was captured and that the models were similar to other literature models. All models developed with this approach would therefore be multiple linear regression models which use the four proximate variables as inputs.

Develop and test models

During the data split step, 821 samples were grouped into the modelling dataset group. These samples were used to develop a linear regression model using the LINEST function in Excel. Table A-3 gives the LINEST matrix for this case study.

Table A-3: LINEST matrix for South African coal regression model

LINEST matrix for South African coal model				
0.743	0.512	-0.365	-0.743	21.51
0.153	0.234	0.0136	0.214	1.201
0.994	0.469	#N/A	#N/A	#N/A
4354.8	816	#N/A	#N/A	#N/A
0.203	0.0069	#N/A	#N/A	#N/A

The developed model was:

$$GCV = 21.51 - 0.743(A) - 0.365(IM) + 0.512(FC) + 0.743(VM) \quad (7-1)$$

- GCV - Calorific value in MJ/kg
A - Ash content of the coal in %

- IM* - Moisture content of the coal in %
- FC* - Fixed carbon content of the coal in %
- VM* - Volatile matter content of the coal in %

where GCV is expressed in MJ/kg and A, IM, FC and V are the percentage of ash, inherent moisture, fixed carbon and volatile matter, respectively.

Evaluating the A and IM regressors of the model, it was as foreseen, negative. This is expected since the A and the IM negatively effects the GCV [15]. It can also be seen in the data that the relationship between the A is negative with regards to the GCV. The relationship between the IM and the GCV seems neutral when the data is inspected. Figure A-5 illustrates the A and IM vs GCV relationship respectively.

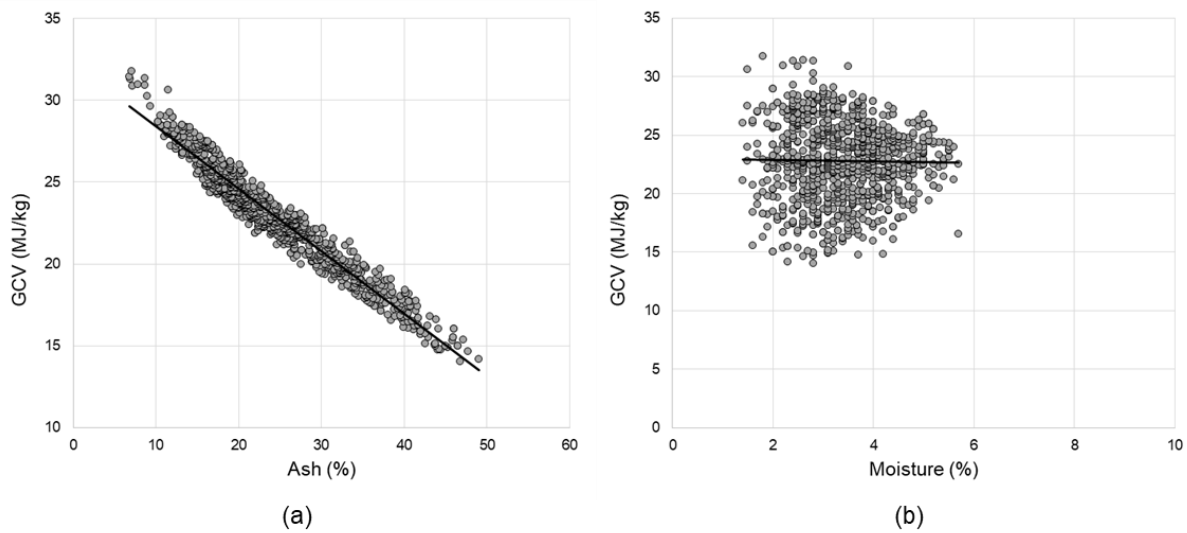


Figure A-5: Relationship between A (a) and IM (b) vs the GCV

In general, the GCV is positively affected by the FC and VM. The relationship between the FC and GCV and the VM and GCV were plotted (Figure A-6) and the expected positive relationships were observed.

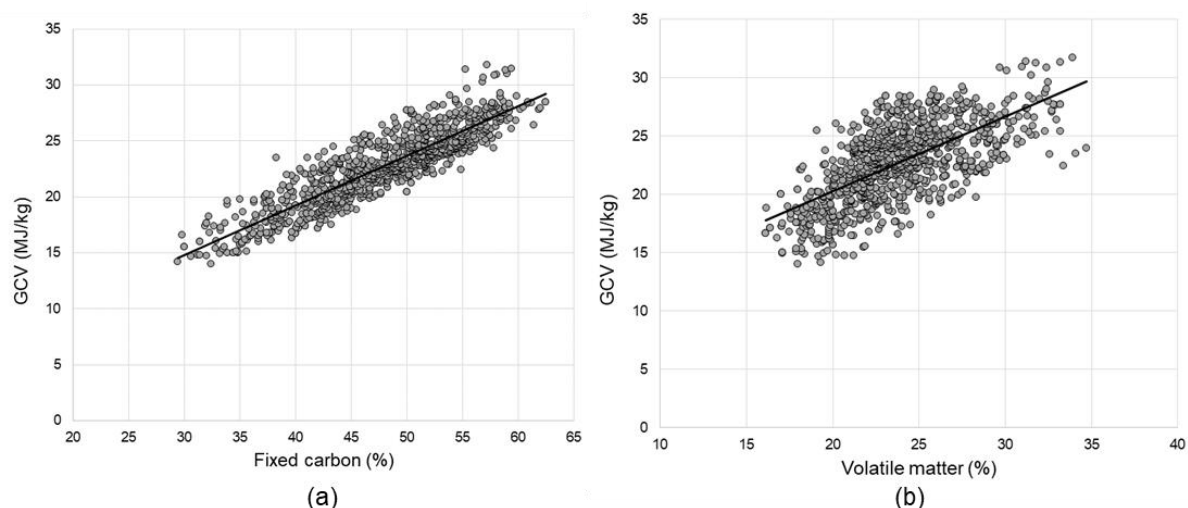


Figure A-6: Relationship between FC (a) and VM (b) vs the GCV

All the regressors in the developed model (Equation 7-1) correlate with the observed relationships. The developed model was therefore deemed feasible since it met the other prescribed statistics reviewed in the next section.

Check requirements

A developed model does not necessarily constitute a viable outcome. The R^2 , RMSE and the MAE statistics were evaluated to ensure that the model was usable and presented accurate results.

The LINEST matrix given in Table A-3 calculated the $R^2 = 0.994$ and the $RMSE = 0.469\%$. The MAE of the developed model on the modelling data is calculated as 0.86% . These values were evaluated against the criteria presented in Chapter 2 and 3.

The evaluation showed that the model satisfied each of the criteria without areas for concerns, indicating that using the model would probably produce reasonable and accurate results.

A.1.3 Model validation

The model developed according to the previous section needed to be validated. Validation was performed to assess whether the predicted GCVs were reasonable with respect to the experimented GCVs. The two sets of 30% validation data were used to calculate two MAE and the CC parameters (one for each validation set). The average between the two parameters were then used as the reporting answer.

The MAE between the experimented and the predicted GCVs were calculated as 2.89% , indicating that the predicted GCV was relatively similar to the respective experimental

counterpart, thus indicating that the model was likely to estimate trustworthy and reliable GCVs. The relationship between the experimented and predicted GCVs was plotted and the CC was determined as 0.989. Figure A-7 shows this relationship.

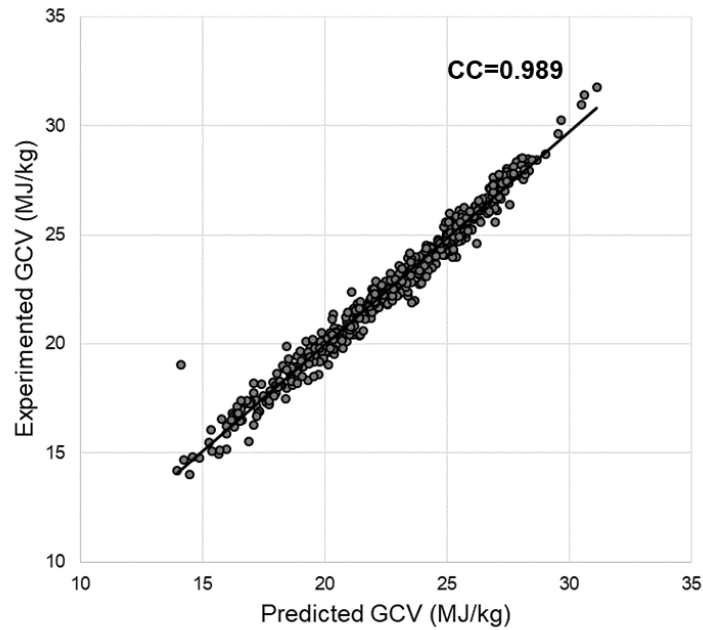


Figure A-7: Validation of the developed model

Both the aforementioned results complied with the desired criteria which established the validity of the model.

A.2 Case study 2: Indian coal

A.2.1 Dataset preparation

Gather data

For case study 2, 254 Indian coal samples were obtained from three different sources. A summary of the samples follows:

- 50 samples from *Majumder's* published article[18]. These samples cover a wide range of proximate variables and GCVs and were obtained according to ASTM standards.
- 75 samples from *Patel's* published article[21]. These samples contain proximate and ultimate analysis, as well as the corresponding GCVs.
- 129 samples from the *Indian coal bulletin* of 2010 to 2012[13].

The data from all three of these sources is readily available online. The coal samples are all based on an air-dried basis and cover different coal regions within India.

A summary of the obtained data is given in Table A-4.

Table A-4: Summary of case study 2 data

Source	Sample size	Coal types included	Geographical origin	Sample information included?		
				Proximate analysis	Ultimate analysis	GCV
<i>Majumder</i> (S1)[18]	50	- Lignites	- Northern India	Yes	No	Yes
<i>Patel</i> (S2)[21]	75	- Lignites	- Did not specify the exact regions of India	Yes	Yes	Yes
<i>Coal bulletins</i> (S3)[89]	129	- Lignites - Bituminous - Anthracite	- All coal regions of India	Yes	Yes	Yes

Inspect data

The data from all three sources was combined to form a total dataset of 254 samples. The dataset was inspected to ensure that it is representable and of good quality, ultimately promoting consistent results from reliable data obtained through verified methods.

Since all the data was obtained from published papers the data is deemed reliable and an accurate representation of the coal. Figure A-8 shows the distribution of the complete dataset.

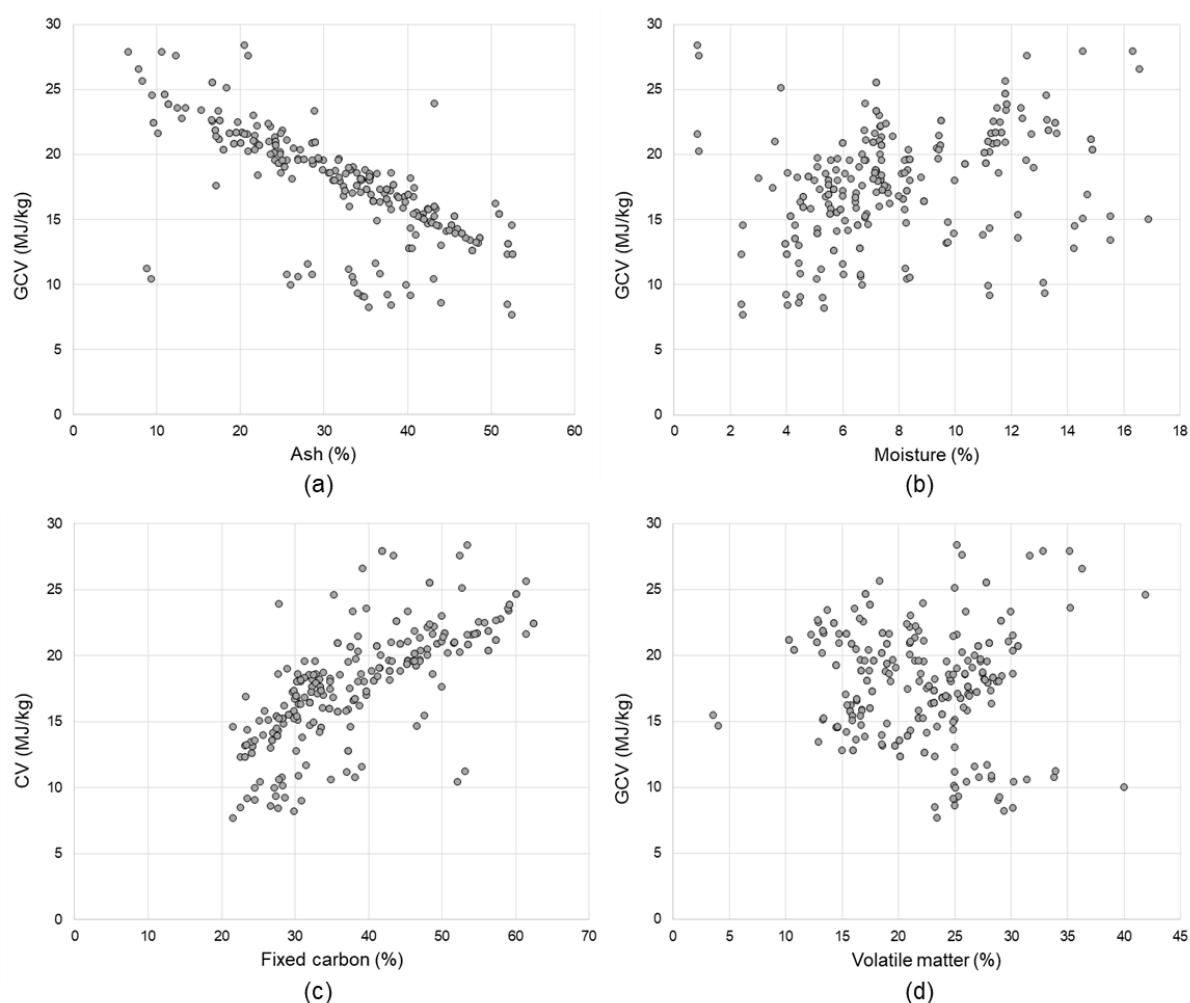


Figure A-8: Case study 2 raw data distribution

The complete dataset contained 254 samples and still complied with Green's minimal sample size of 84. The dataset was therefore deemed a representative and complete sample set.

Clean data

Before the modelling procedure could be started, MATLAB was used to identify possible outliers in the new dataset. The *is-outlier* build-in function in MATLAB was used and 6 potential outliers were identified. The reason behind the identification as an outlier was investigated before data was removed. After evaluation, all 6 of the identified outliers were removed.

Table A-5 gives the list of identified outliers. Table A-5 also gives the reason for the outlier, as well as whether or not the outlier was removed.

Table A-5: Sample list of the potential outliers

Proximate variable				GCV (MJ/kg)	Reason for outlier	Removed
IM (%)	VM (%)	FC (%)	A (%)			
11.5	35.24	39.75	13.51	8.59	High FC, low GCV	Yes
13.25	41.9	35.37	9.48	9.62	High FC, low GCV	Yes
6.32	35.21	41.89	16.58	8.62	High FC, low GCV	Yes
14.51	32.85	41.92	10.66	6.66	High FC, low GCV	Yes
12.56	31.69	43.39	12.36	11.65	High FC, low GCV	Yes
4.5	15.04	54.56	14.6	8.99	High FC, low GCV	Yes

The main reasons for the outliers were found to be a high FC and a corresponding low GCV, which is not the expected ranges. By removing these 6 samples, the final dataset now contained 248 samples, which still complies with Green's minimum sample size.

The final dataset is illustrated in Figure A-9, where the relationship of each of the proximate variables against the GCVs is given.

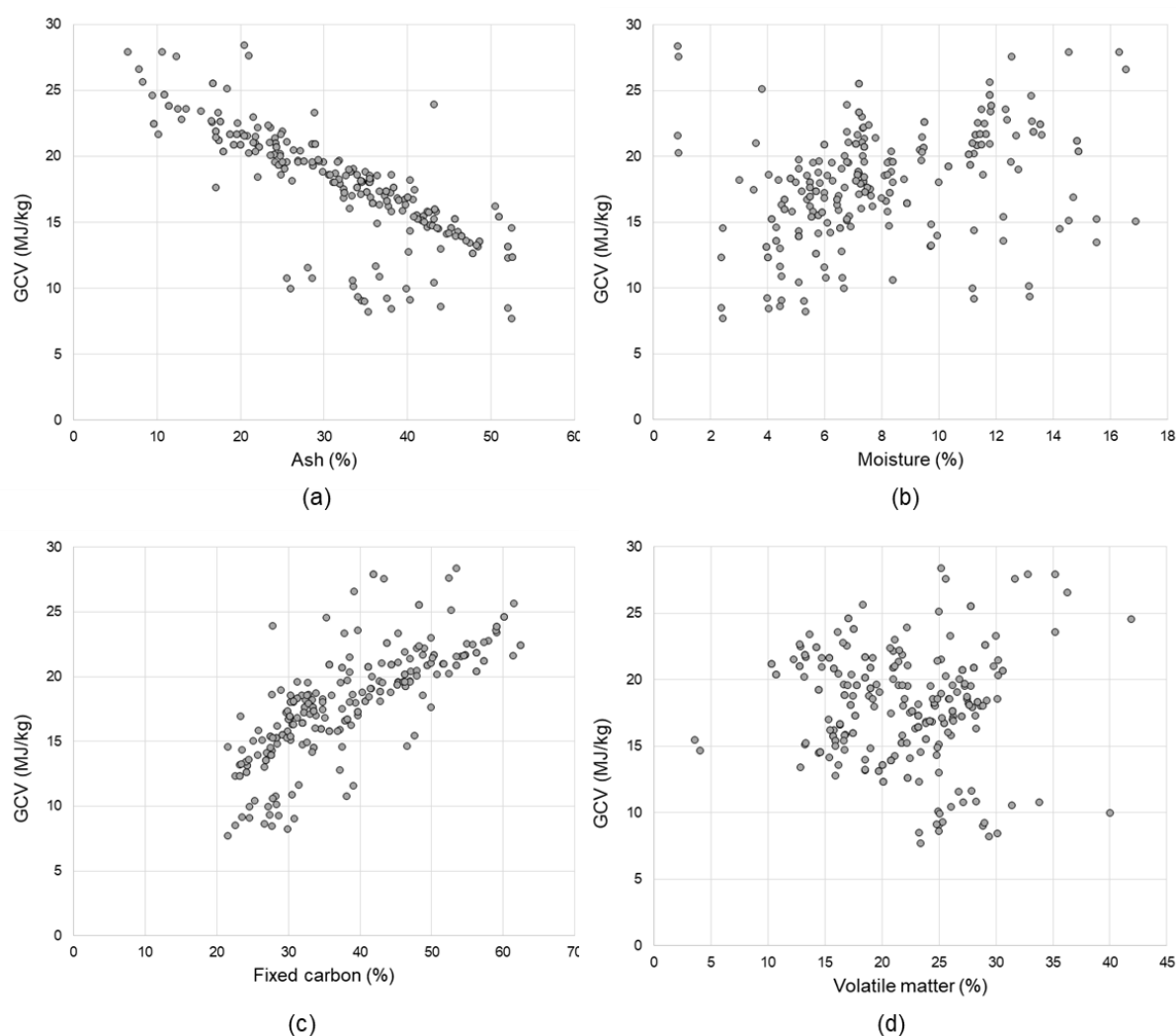


Figure A-9: Clean case study 2 data

The clean data is now considered a better representation of the real data and can therefore be used for modelling.

Split data into modelling and validation data

The next step was to split the clean data into modelling data and validation data. In order to split the data, the clean dataset was first randomised to ensure that the different groups of data will still represent the original sample. The *Rand* function was used in Microsoft excel to randomise the data. The 30/40/30 split was then used to split the data into three groups: the first 30% as the first validation set of data, the next 40% as the modelling dataset and the final 30% as the second set of validation data. The 248 data points were thus divided as follows:

- Group 1: 74 data points (first 30%) for Validation set 1
- Group 2: 100 data points (next 40%) for Modelling

- Group 3: 74 data points (last 30%) for Validation set 2

Figure A-10 shows the distribution of the three groups of data for each of the proximate variables. These figures indicate that all three groups of data are representative of the original clean dataset.

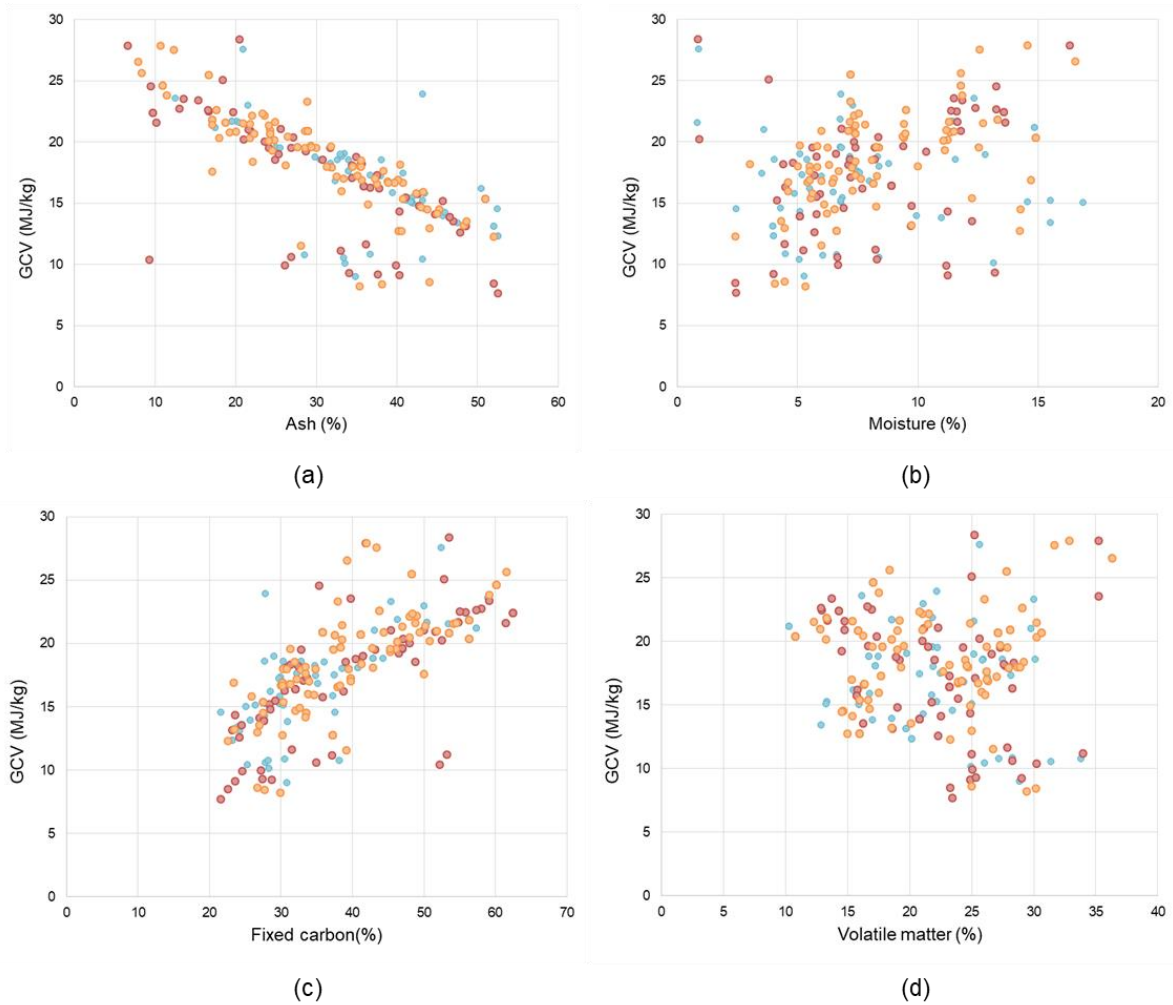


Figure A-10: Data split of case study 2 data

The blue and red dots illustrate Validation dataset 1 and 2 respectively, while the orange dots represent the modelling data.

The modelling set is used to develop a new model. The validations sets are used to validate the model in the Model validation section.

A.2.2 Model development

Type selection

In the previous section, proximate analysis data was plotted on several instances against the corresponding GCV values. In most instances, a form of linear relationship was noted between the variables and the GCV. Due to this linearity the method of multiple linear regression was chosen to develop each of the case studies' models. Multiple linear regression will capture the linear relationship while ensuring that the effect of each respective variable influences the model accordingly, as Majumder *et al.* [18] already confirmed.

Identify constraints

A developed model needed to include the variables that would present a reasonable representation of the reality. Literature indicated that the variables from the proximate analysis are commonly used, due to their availability and ease of access. Models developed with the approach presented in this study include the four proximate variables. This ensured that the effect of the different variables was captured and that the models were similar to other literature models. All models developed with this approach would therefore be multiple linear regression models which use the four proximate variables as inputs.

Develop and test models

A multiple regression model was respectively developed for each case study. This section showcases this sub-step with the results from case study 2.

The modelling dataset for case study 2 consisted of 100 samples which represented 40% of the data. These 100 samples were used to develop the regression model using the LINEST function in ExcelTM. Table A-6 gives the LINEST matrix obtained for case study 2.

Table A-6: LINEST matrix results

LINEST matrix				
0.151	0.615	0.111	-1.564	12.51
0.021	0.094	0.003	0.103	0.743
0.945	0.411	#N/A	#N/A	#N/A
2541.51	243	#N/A	#N/A	#N/A
0.031	0.0741	#N/A	#N/A	#N/A

The developed model for case study 2 is presented in Equation 7-2:

$$GCV = 12.51 - 1.564(A) + 0.111(IM) + 0.615(FC) + 0.151(VM) \quad (7-2)$$

- GCV* - Calorific value in MJ/kg
- A* - Ash content of the coal in %
- IM* - Moisture content of the coal in %
- FC* - Fixed carbon content of the coal in %
- VM* - Volatile matter content of the coal in %

where GCV is expressed in MJ/kg and A, IM, FC and V are the percentage of ash, inherent moisture, fixed carbon and volatile matter, respectively.

Evaluating the A regressors of the model, it was as foreseen, negative. This is expected since the A negatively affects the GCV [15]. It can be seen in the data that the relationships between the A are negative with regards to the GCV, which validates the negative regressor associated with the A.

The IM are also expected to negatively influence the GCV. However, the regressors associated with the IM are positive. The data was evaluated and the positive relationship was observed in the relationship between the IM and GCV as well. Figure A-11 illustrates the A and IM relationship with the respective GCV.

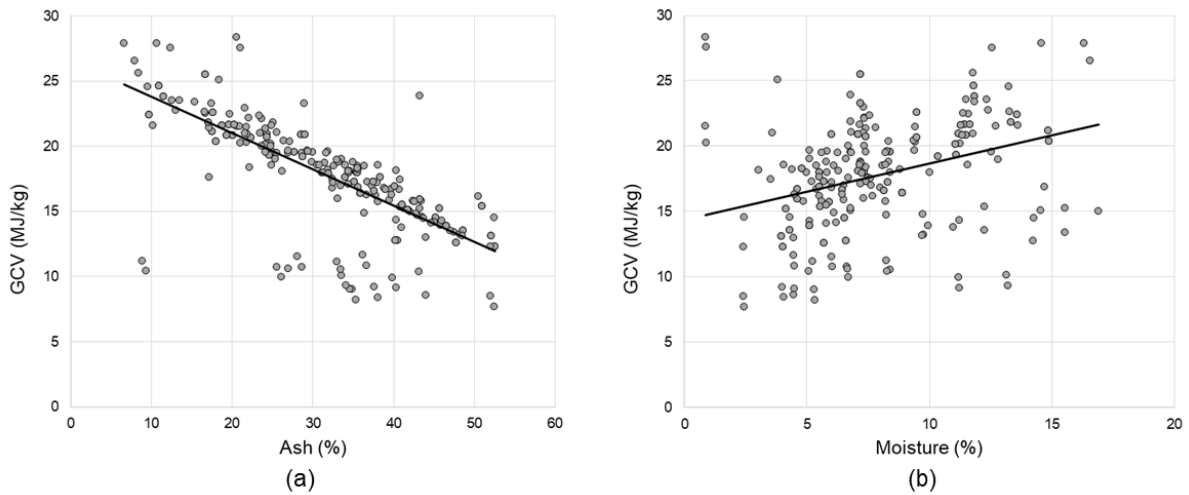


Figure A-11: Relationship between A (a) and IM (b) vs the GCV

This phenomena was investigated at the hand of the $SE_{x_{IM}}$ value given in the LINEST matrix. The $SE_{x_{IM}}$ indicated that the influence of the IM on the overall error was small compared to the variables influences. This implies that the impact of the IM regressor in the GCV was minimum and that it will not affect the viability of the model.

Further evaluation of Equation 7-2 shows the expected positive FC and VM regressors. The relationship between the FC and GCV and the VM and GCV was plotted (Figure A-12) and the expected positive relationships were observed.

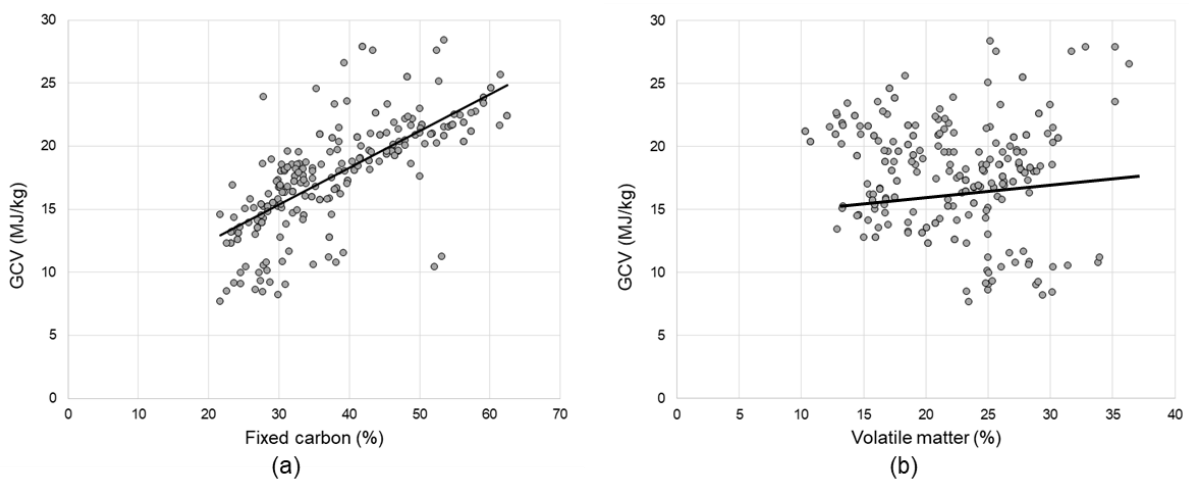


Figure A-12: Relationship between FC (a) and VM (b) vs the GCV

From Figure A-12, it can be seen that the comparison of the VM with the GCV is scattered. A clear positive relationship is not observed. The $SE_{x_{VM}}$ for Equation 7-2 is also very small which indicates that these values do not influence the GCV prediction very much.

The developed model was therefore deemed feasible since it met the other prescribed statistics reviewed in the next section.

Check requirements

A developed model does not necessarily constitute a viable outcome. The R^2 , RMSE and the MAE statistics were evaluated to ensure that the model was usable and presented accurate results.

The LINEST matrix given in Table A-6 calculated the $R^2 = 0.945$ and the $RMSE = 0.411\%$. The MAE of the developed model on the modelling data is calculated as 0.42%. These values were evaluated against the criteria presented in Chapter 2 and 3.

The evaluation showed that the model satisfied each of the criteria without areas for concerns, indicating that using model would probably produce reasonable and accurate results.

A.2.3 Model validation

The model developed according to the previous section needed to be validated. Validation was performed to assess whether the predicted GCVs were reasonable with respect to the experimented GCVs.

The two sets of 30% validation data were used to calculate two MAE and the CC parameters (one for each validation set). The average between the two parameters was then used as the reporting answer.

The MAE between the experimented and the predicted GCVs were calculated as 4.32%, indicating that the predicted GCV was relatively similar to the respective experimental counterpart, thus indicating that the model was likely to estimate trustworthy and reliable GCVs.

The relationship between the experimented and predicted GCVs was plotted and the CC was determined as 0.891. Figure A-13 shows this relationship.

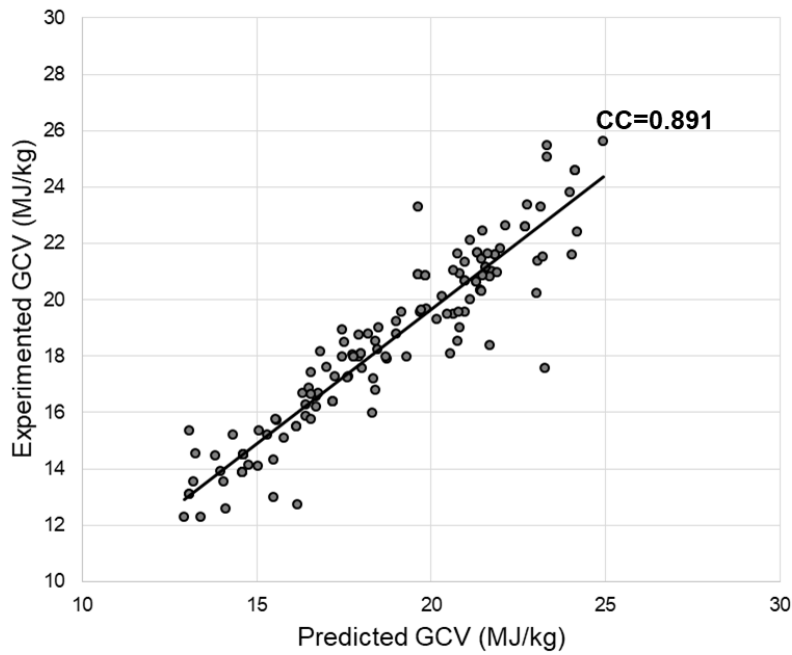


Figure A-13: Validation of the developed model

Both the aforementioned results complied with the desired criteria which established the validity of the model.

A.3 Case study 3: Alaskan coal

A.3.1 Dataset preparation

Gather data

For this case study, 243 Alaskan coal samples were obtained from five different sources. A summary of the samples follows:

- 15 samples from a published paper by Peters, 2015[21],
- 183 samples from coal bulletin of 2014[50]
- 20 samples from a published article by Antonio, 2009[91]
- 15 samples from a published article by Guillian, 2009[92]
- 10 samples from a published article by Maue, 2009 – 2011[93]

The data from all these sources is readily available online. The coal samples are all based on an as-received basis and cover different coal regions within Alaska.

A summary of the obtained data is given in Table A-7.

Table A-7: Summary of case study 3 data

Source	Sample size	Coal types included	Geographical origin	Sample information included?		
				Proximate analysis	Ultimate analysis	GCV
<i>Peters (S1)[21]</i>	15	- Bituminous	- Arctic coastal plain	Yes	No	Yes
<i>Coal bulletin (S2)[50]</i>	183	- Bituminous	- Northern Cordillera - Southern Cordillera	Yes	Yes	Yes
<i>Antonio (S3)[91]</i>	20	- Lignites - Bituminous	- Intermountain plateau	Yes	Yes	Yes
<i>Guillian(S4)[92]</i>	15	- Lignites - Bituminous	- Arctic coastal plain - Intermountain plateau	Yes	No	Yes
<i>Mauie(S5)[93]</i>	10	- Bituminous	- Northern Cordillera - Southern Cordillera	Yes	Yes	Yes

Inspect data

The data from all five sources was combined to form a total dataset of 243 samples. The dataset was inspected to ensure that it is representative and of good quality, ultimately promoting consistent results from reliable data obtained through verified methods.

Since all the data was obtained from published papers, the data is deemed reliable and an accurate representation of the coal. Figure A-14 shows the distribution of the complete dataset.

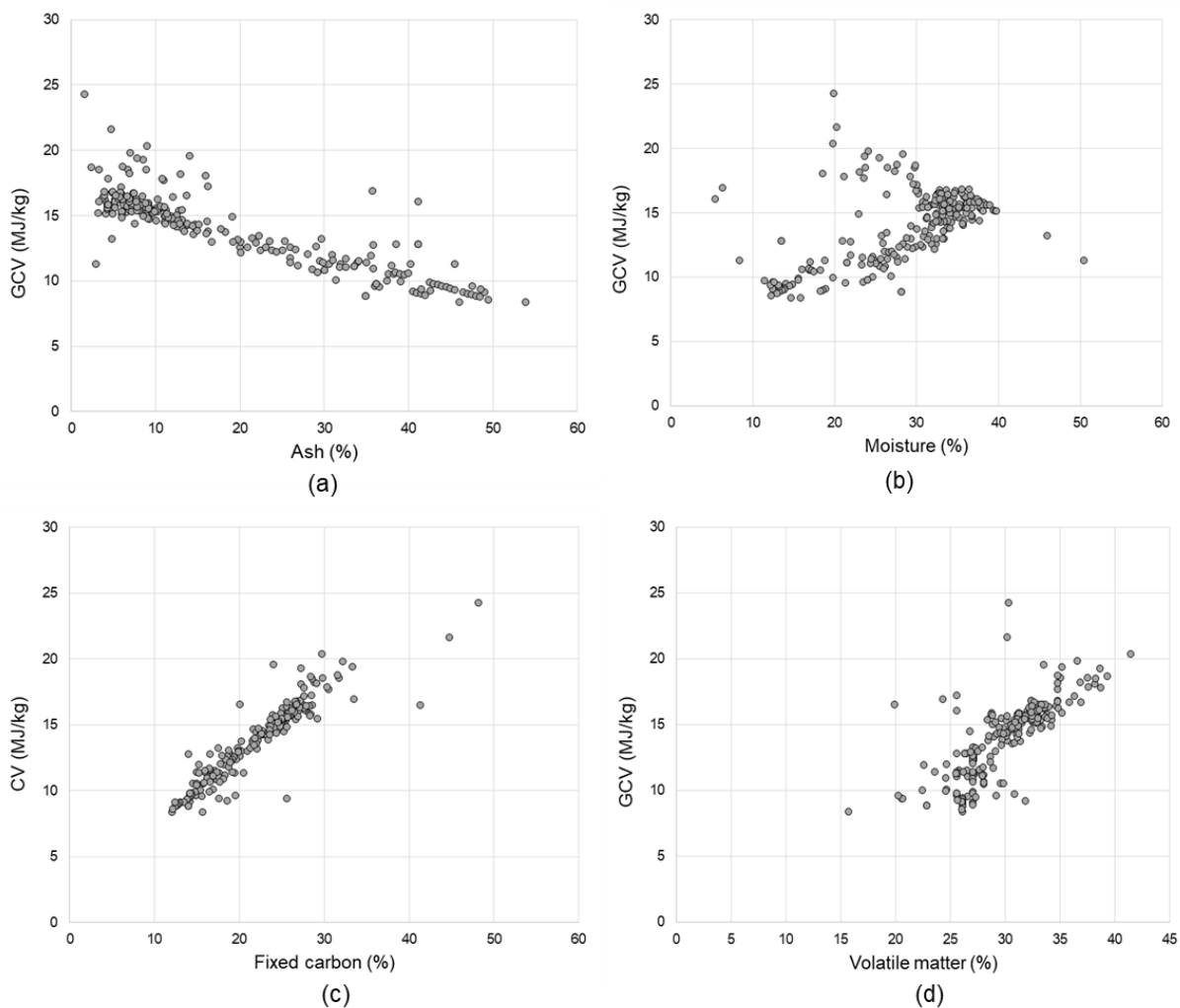


Figure A-14: Case study 2 raw data distribution

The complete dataset contained 243 samples and still complied with Green's minimal sample size of 84. The dataset was therefore deemed a representative and complete sample set.

Clean data

Before the modelling procedure could be started, MATLAB was used to identify possible outliers in the new dataset. The *is-outlier* build-in function in MATLAB was used and 2 potential outliers were identified. The reason behind the identification as an outlier was investigated before data was removed. After evaluation, both of the identified outliers were removed.

Table 4-2 gives the list of identified outliers. The table also gives the reason for the outlier, as well as whether or not the outlier was removed.

Table A-8: Sample list of the potential outliers

Proximate variable				GCV (MJ/kg)	Reason for outlier	Removed
M (%)	VM (%)	FC (%)	A (%)			
20.28	29.20	45.75	4.77	21.61	High FC compared with Alaskan averages	Yes
19.88	30.31	48.21	1.60	24.25	High FC compared with Alaskan averages	Yes

The average FC range determined from five different sources is between 10% and 35%. After inspecting the types of coal within these sources, it was identified that Alaska typically produces lignite and bituminous coal types. Both these coal types have a literature reported maximum FC value of less than 35% as well.

The two outliers identified in MATLAB had an FC value higher than 45%. For this reason, these two samples were removed from the dataset. The final dataset now contained 241 samples, which still complies with Green's minimum sample size.

The final dataset is illustrated in Figure A-15, where the relationship of each of the proximate variables against the GCVs is given.

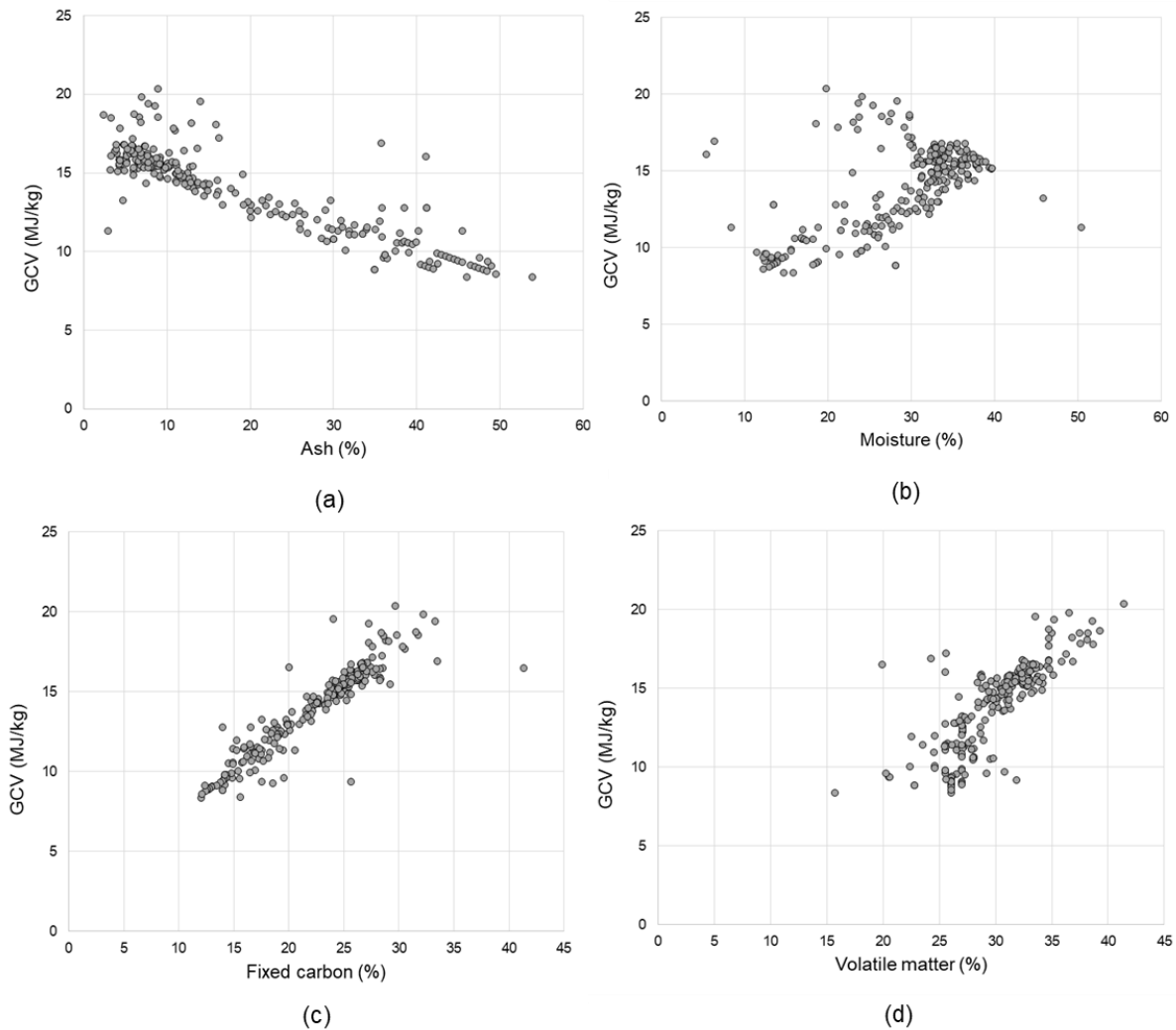


Figure A-15: Clean case study 3 data

The clean data is now considered a better representation of the real data and can therefore be used for modelling.

An unusual relationship is observed between the moisture content of the Alaskan coal compared with the GCV. From Figure A-15 it seems like the GCV is increasing with an increase in the moisture content of the coal. It was discussed in Chapter 2 that the moisture in the coal decreases the GCV.

Split data into modelling and validation data

The next step was to split the clean data into modelling data and validation data. In order to split the data, the clean dataset was first randomised to ensure that the different groups of data will still represent the original sample. The *Rand* function was used in Microsoft excel to randomise the data. The 30/40/30 split was then used to split the data into three groups: the

first 30% as the first validation set of data, the next 40% as the modelling dataset and the final 30% as the second set of validation data. The 241 data points were thus divided as follows:

- Group 1: 72 data points (first 30%) for Validation set 1
- Group 2: 97 data points (next 40%) for Modelling
- Group 3: 72 data points (last 30%) for Validation set 2

Figure A-16 shows the distribution of the three groups of data for each of the proximate variables. These figures indicate that all three groups of data are representative of the original clean dataset.

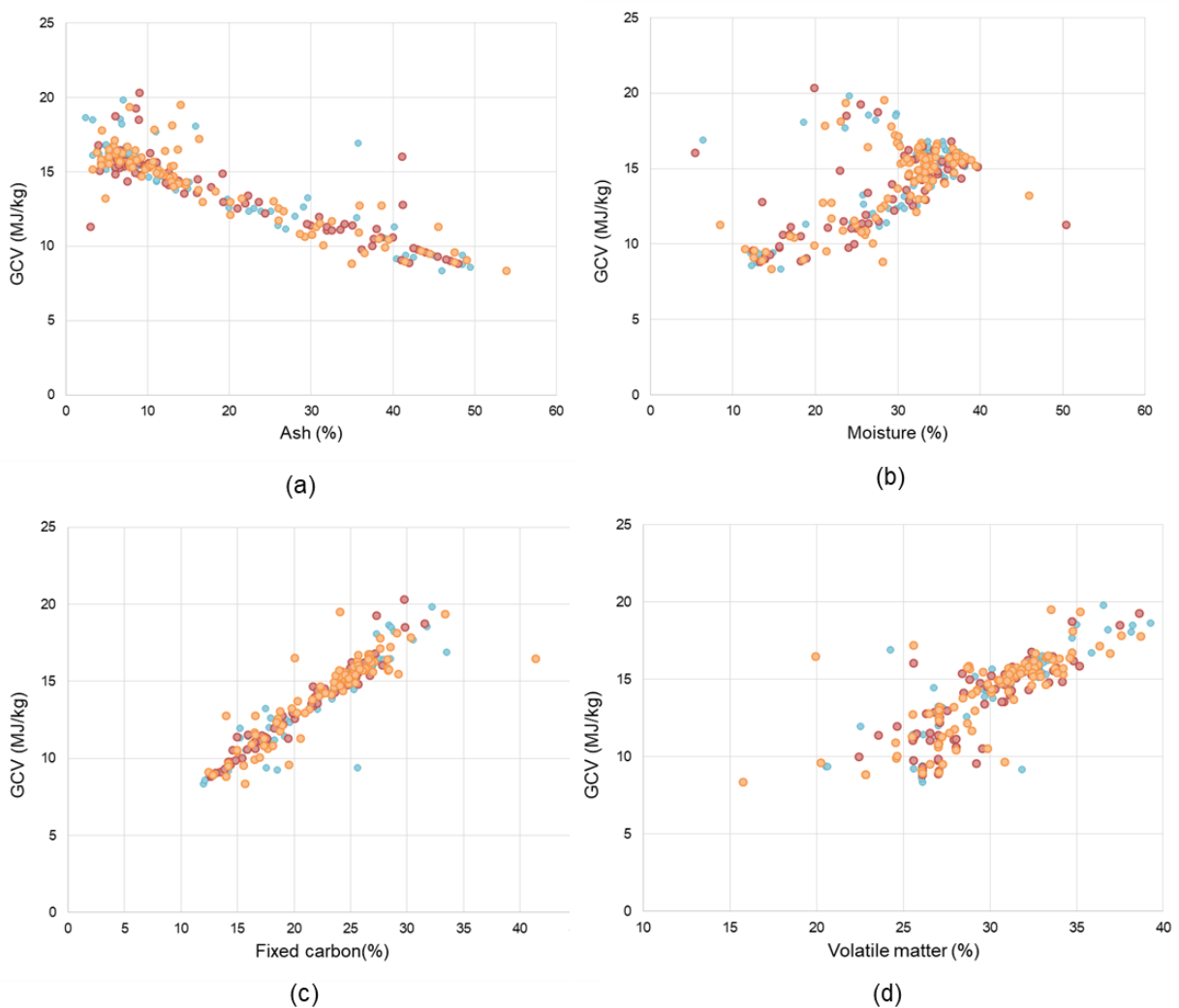


Figure A-16: Data split of case study 3 data

The blue and red dots illustrate Validation dataset 1 and 2 respectively, while the orange dots represent the modelling data. The modelling set is used to develop a new model. The validations sets are used to validate the model in the Model validation section.

A.3.2 Model development

Type selection

In the previous section, proximate analysis data was plotted on several instances against the corresponding GCV values. In most instances a form of linear relationship was noted between the variables and the GCV. Due to this linearity the method of multiple linear regression was chosen to develop each of the case studies' models. Multiple linear regression will capture the linear relationship while ensuring that the effect of each respective variable influences the model accordingly, as Majumder *et al.* [18] already confirmed.

Identify constraints

A developed model needed to include the variables that would present a reasonable representation of the reality. Literature indicated that the variables from the proximate analysis are commonly used, due to their availability and ease of access. Models developed with the approach presented in this study include the four proximate variables. This ensured that the effect of the different variables was captured and that the models were similar to other literature models. All models developed with this approach would therefore be multiple linear regression models which use the four proximate variables as inputs.

Develop and test models

The modelling dataset for this case study consisted of 97 samples which represented 40% of the data. These 97 samples were used to develop the regression model using the LINEST function in Excel™. Table A-9 gives the LINEST matrix obtained for case study 3.

Table A-9: LINEST matrix results

LINEST matrix for case study 3				
0.125	0.356	0.123	-2.310	24.25
0.0072	0.045	0.0018	0.065	2.136
0.953	0.548	#N/A	#N/A	#N/A
699.76	236	#N/A	#N/A	#N/A
0.136	1.432	#N/A	#N/A	#N/A

The developed model for case study 3 is presented in Equation 7-3:

$$GCV = 24.25 - 2.31(A) + 0.123(IM) + 0.356(FC) + 0.125(VM) \quad (7-3)$$

- GCV - Calorific value in MJ/kg
- A - Ash content of the coal in %
- M - Moisture content of the coal in %
- FC - Fixed carbon content of the coal in %
- VM - Volatile matter content of the coal in %

where GCV is expressed in MJ/kg and A, M, FC and V are the percentage of ash, moisture, fixed carbon and volatile matter, respectively.

Evaluating the A regressors of the model, it was as foreseen, negative. This is expected since the A negatively affects the GCV [15]. It can be seen in the data that the relationship between the A are negative with regards to the GCV, which validates the negative regressor associated with the A.

The IM are also expected to negatively influence the GCV. However, the regressors associated with the M are positive. The data was evaluated and the positive relationship was observed in the relationship between the IM and GCV as well. Figure A-17 illustrates the A and IM relationship with the respective GCV.

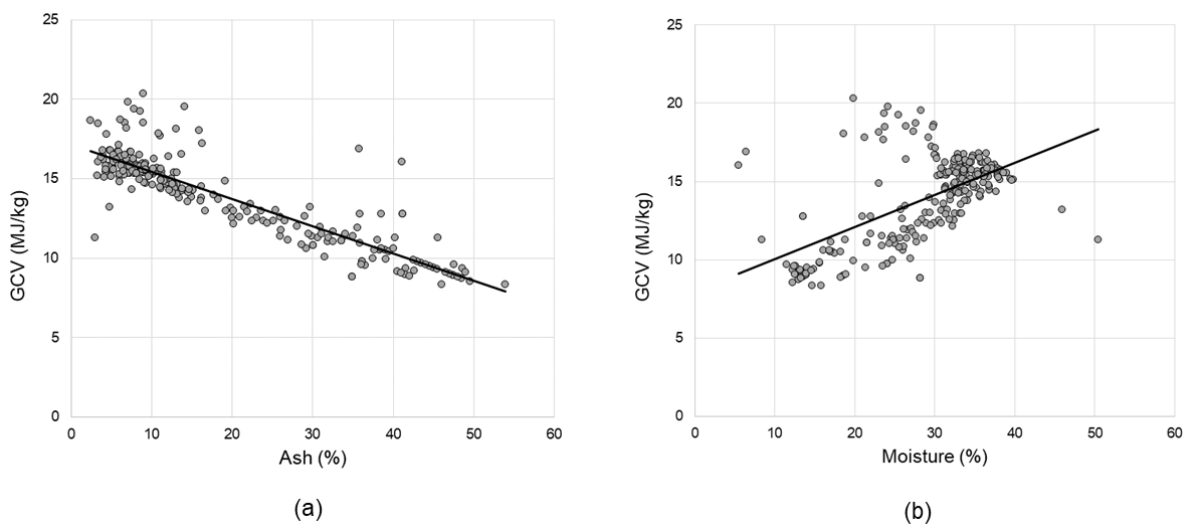


Figure A-17: Relationship between A (a) and IM (b) vs the GCV

This phenomena was investigated at the hand of the SE_{x_M} value given in the LINEST matrix.

The SE_{x_M} indicated that the influence of the M on the overall error was small compared with

the other variables. This implies that the impact of the M regressor in the GCV was minimum and that it will not affect the viability of the model.

Further evaluation of Equation 7-3 shows the expected positive FC and VM regressors. The relationship between the FC and GCV and the VM and GCV were plotted (Figure A-18) and the expected positive relationships were observed.

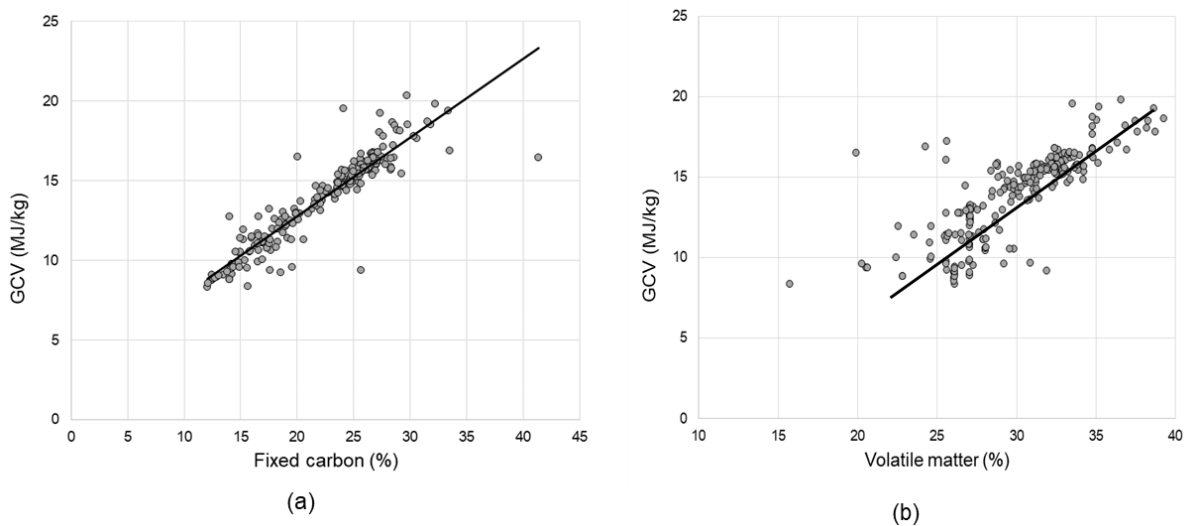


Figure A-18: Relationship between FC (a) and VM (b) vs the GCV

The developed model was therefore deemed feasible since it met the other prescribed statistics reviewed in the next section.

A.3.3 Model validation

The model developed according to the previous section needed to be validated. Validation was performed to assess whether the predicted GCVs were reasonable with respect to the experimented GCVs.

The two sets of 30% validation data were used to calculate two MAE and the CC parameters (one for each validation set). The average between the two parameters was then used as the reporting answer.

The MAE between the experimented and the predicted GCVs was calculated as 4.01%, indicating that the predicted GCV was relatively similar to the respective experimental counterpart, thus indicating that the model was likely to estimate trustworthy and reliable GCVs.

The relationship between the experimented and predicted GCVs was plotted and the CC was determined as 0.942. Figure A-19 shows this relationship.

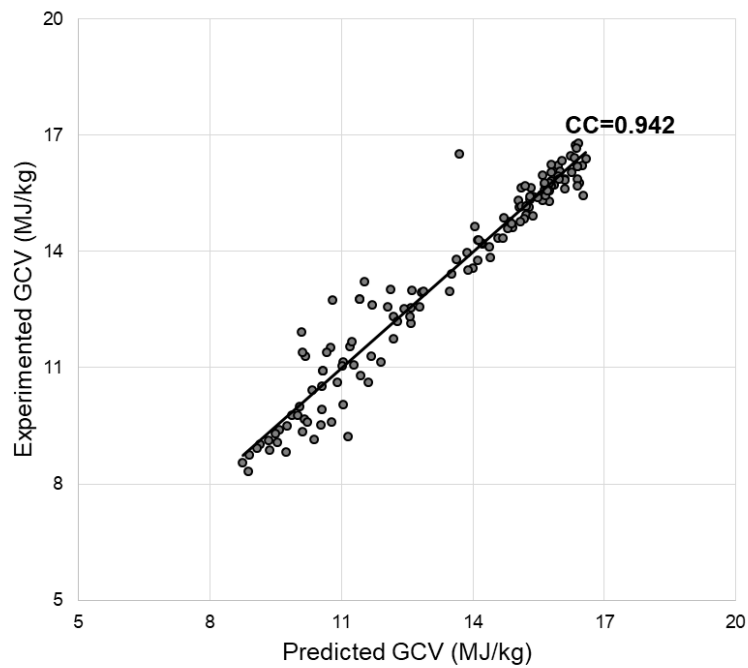


Figure A-19: Validation of the developed model

Both the aforementioned results complied with the desired criteria which established the validity of the model.

APPENDIX B – VISUALISATION DISCUSSION

B. Visualisation results

B.1 New model: New Zealand and Australian coal

B.1.1 Dataset preparation

Gather data

For the new dataset, a total of 92 New Zealand and Australian coal samples are used. The data was obtained from a published article by Beamish [90] in 1994. The dataset is compiled through the combination of the various tables of coal data given within the article. The various samples are on different bases, and therefore need to be converted to the same base before further usage of the dataset can commence.

Inspect data

The gathered data was inspected to ensure that the dataset is representative and of good quality. No incorrect or missing data was obtained. Since the data is obtained from a published paper, the quality of the data is trusted. To illustrate the distribution of the data, each of the proximate variables is respectively plotted against the corresponding GCV, as can be seen in Figure B-1 (a) to (d).

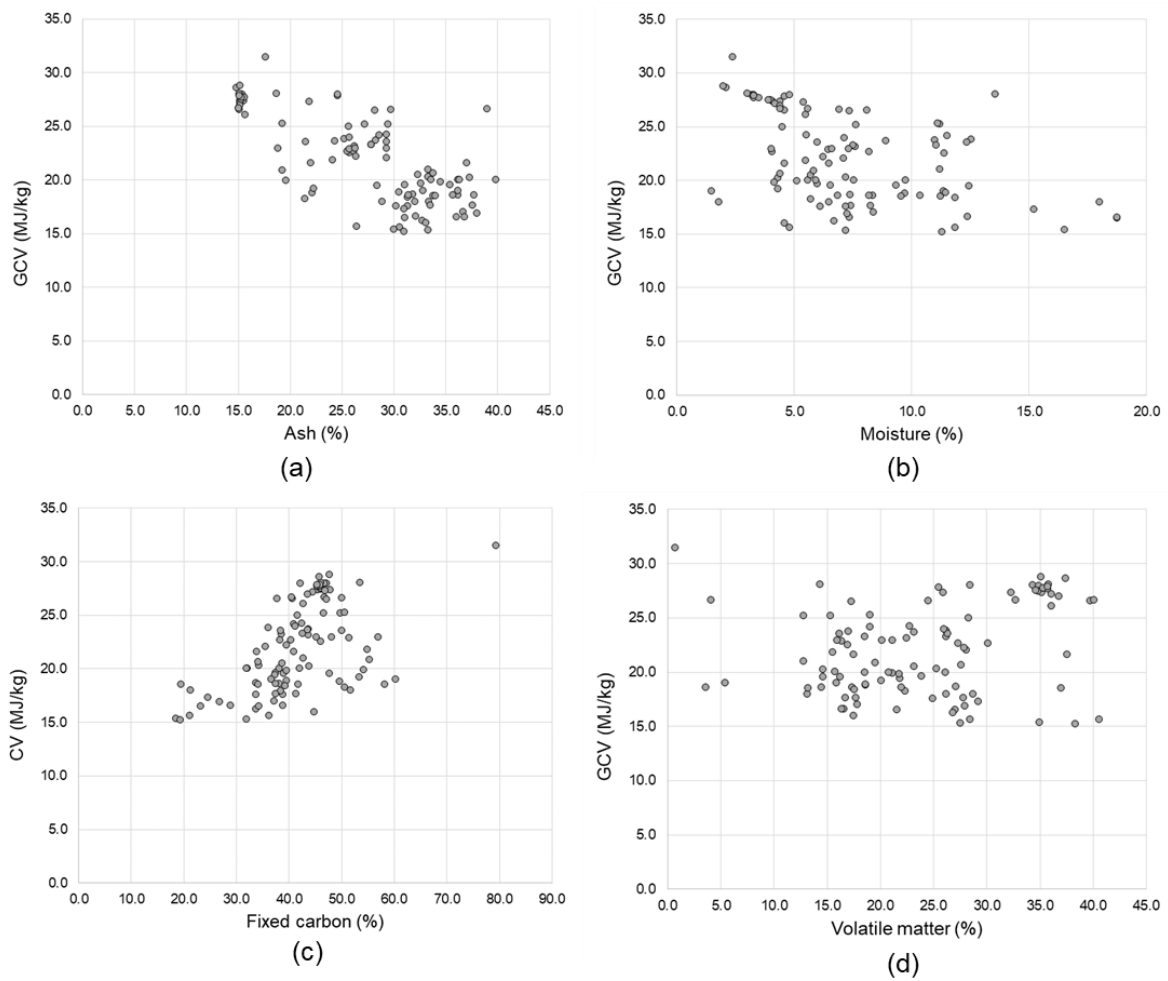


Figure B-1: Proximate analysis versus GCV relationship (new model)

Clean data

Before the modelling procedure could be started, MATLAB was used to identify possible outliers in the new dataset. The *is-outlier* build-in function in MATLAB was used and 2 potential outliers were identified. The reason behind the identification as an outlier was investigated and the two outliers were removed.

The new dataset therefore consists of 90 samples and the distribution is shown in Figure B-2.

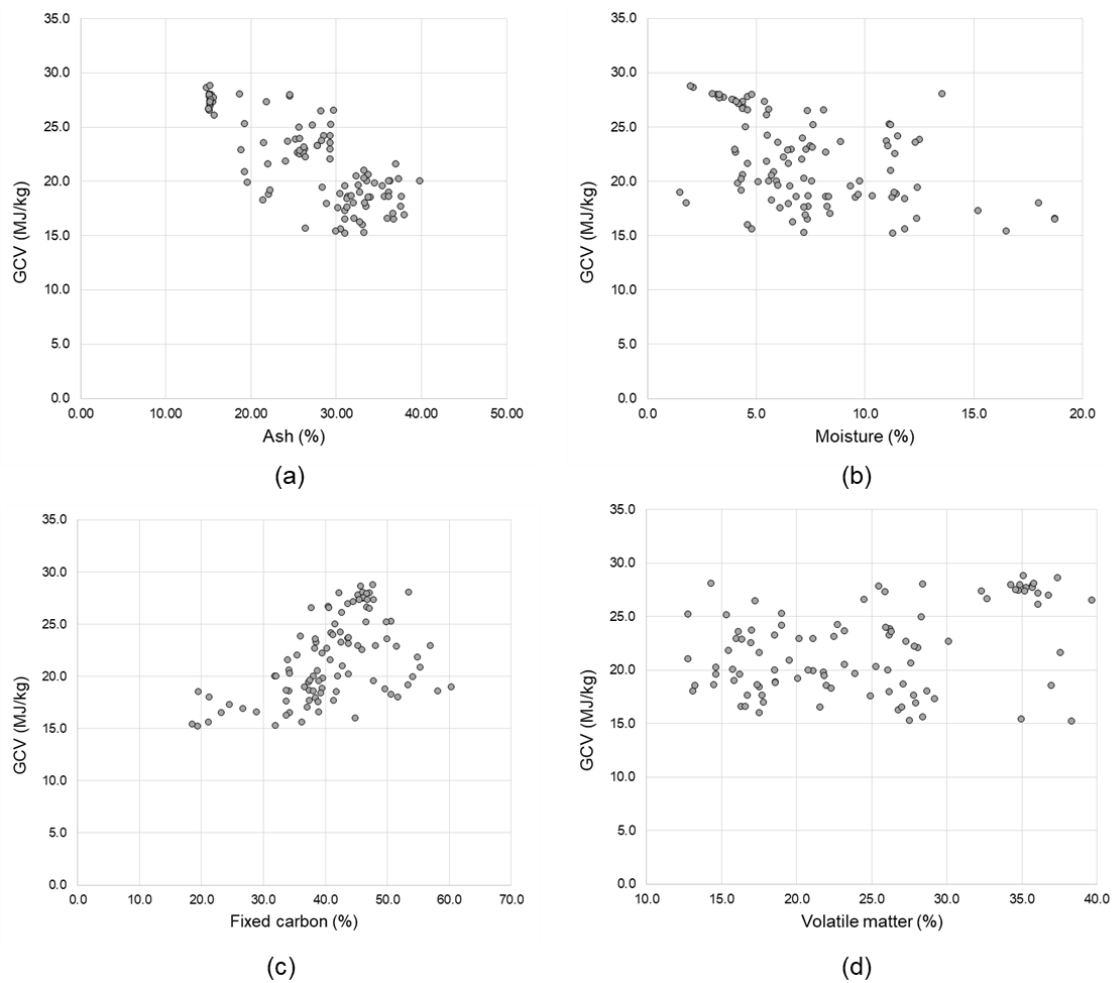


Figure B-2: Proximate analysis versus GCV relationship on clean data

The clean dataset was considered a better representation of the real relationship and can therefore be used for modelling.

Split data into modelling and validation data

The final sub-step for data preparation was to split the data into modelling and validation datasets. This was done to ensure that the different sets of data all represent the total dataset, allowing development and validation of the models using different datasets.

To split the data, the complete clean set was firstly randomised with the *Rand* function in Microsoft Excel™, after which the first 30% and the last 30% of the data, consisting of 27 samples each, were kept aside for the respective first and second validation sets. The remaining 40% of the data (36 samples) was used to develop the models.

Figure B-3 (a) to (d) again shows the relationship between the proximate variables and the corresponding GCV with the clean data. However, these figures distinguish between the

different validation sets along with the modelling set. Illustrating that the three individual sets were representable of the total set.

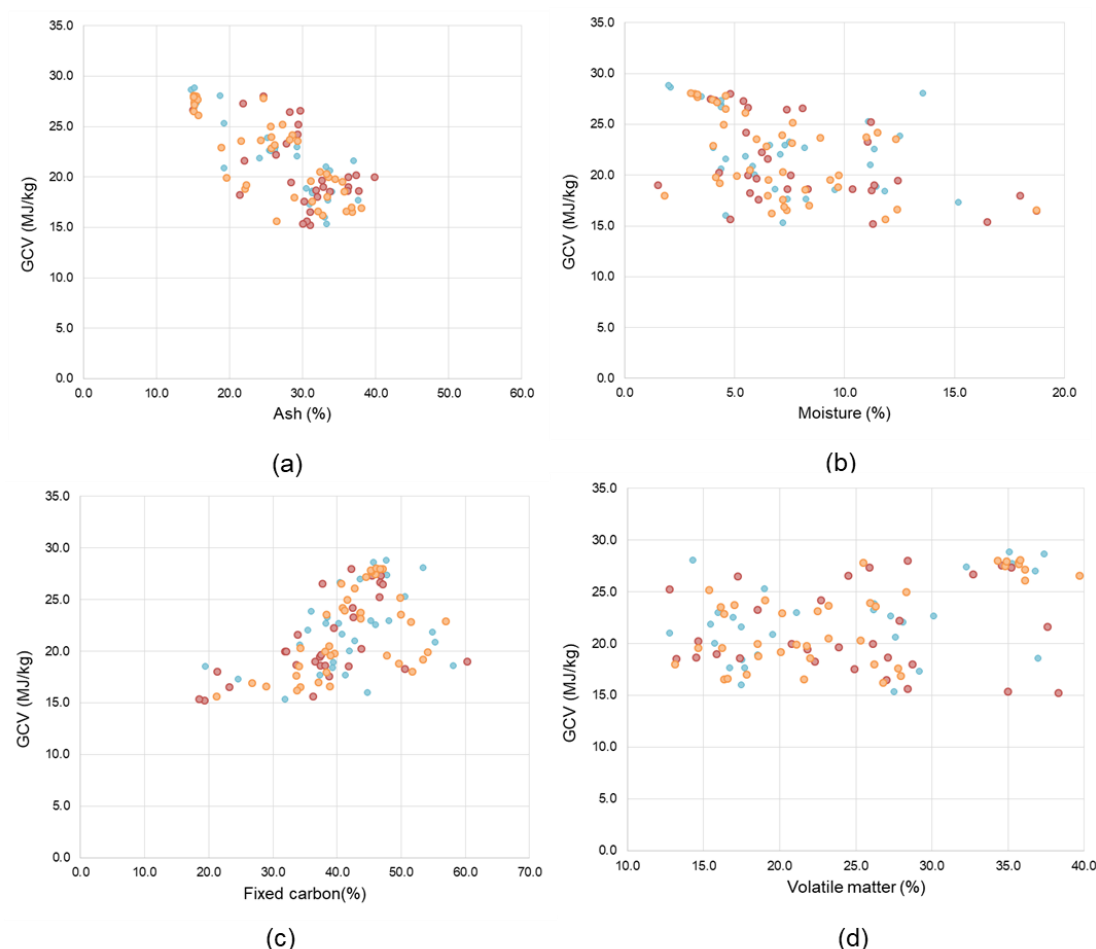


Figure B-3: Proximate analysis versus GCV relationship data split

The blue and red dots illustrate validation dataset 1 and 2 respectively, while the orange dots represent the modelling data.

The modelling data is used to develop the new model. The validation data is used to validate this new model and test the existing models for *Visualisation 2*.

B.1.2 Model development

Type selection

The relationship between the proximate variables and the GCVs seen in Figure B-3 all show a linear relationship. For this reason, it decided to develop a linear regression model for this dataset.

Identify constraints

Literature indicated that the variables from the proximate analysis are commonly used, due to their availability and ease of access. Models developed with the approach presented in this study include the four proximate variables. This ensured that the effect of the different variables was captured and that the models were similar to other literature models. All models developed with this approach would therefore be multiple linear regression models which use the four proximate variables as inputs.

Develop and test model

During the data split step, 36 samples were grouped into the modelling dataset group. These samples were used to develop a linear regression model using the LINEST function in Excel. Table B-1 gives the LINEST matrix for the newly developed model.

Table B-1: LINEST matrix for new coal regression model

LINEST matrix for South African coal model				
0.001	0.126	-0.020	-0.161	23.31
0.003	0.234	0.001	0.193	4.231
0.994	0.685	#N/A	#N/A	#N/A
484.8	31	#N/A	#N/A	#N/A
0.856	0.969	#N/A	#N/A	#N/A

The developed model was:

$$GCV = 23.31 - 0.161(A) - 0.02(IM) + 0.126(FC) + 0.001(VM) \quad (8-1)$$

- GCV - Calorific value in MJ/kg
- A - Ash content of the coal in %
- IM - Moisture content of the coal in %
- FC - Fixed carbon content of the coal in %
- VM - Volatile matter content of the coal in %

where GCV is expressed in MJ/kg and A, IM, FC and V are the percentage of ash, inherent moisture, fixed carbon and volatile matter respectively.

From the developed model (Equation 8-1) it can be seen that both the VM and the M have a very small influence on the output. This could be due to the scatter in the dataset for these two variables.

Check requirements

A developed model does not necessarily constitute a viable outcome. The R^2 , RMSE and the MAE statistics were evaluated to ensure that the model was usable and presented accurate results.

The LINEST matrix given in Table B-1, with the $R^2 = 0.994$ and the $RMSE = 0.685\%$. The MAE of the developed model on the modelling data is calculated as 1.23%. These values were evaluated against the criteria presented in Chapter 2 and 3.

The evaluation showed that the model satisfied each of the criteria without areas for concerns, indicating that using the model would probably produce reasonable and accurate results.

B.1.3 Model validation

The model developed according to the previous section needed to be validated. Validation was performed to assess whether the predicted GCVs were reasonable with respect to the experimented GCVs.

The two sets of 30% validation data were used to calculate two MAE and the CC parameters (one for each validation set). The average between the two parameters was then used as the reporting answer.

The MAE between the experimented and the predicted GCVs was calculated as 11.63%, indicating that the predicted GCV were relatively similar to the respective experimental counterpart, thus indicating that the model was likely to estimate trustworthy and reliable GCVs.

The relationship between the experimented and predicted GCVs was plotted and the CC was determined as 0.943. Figure B-4 shows this relationship.

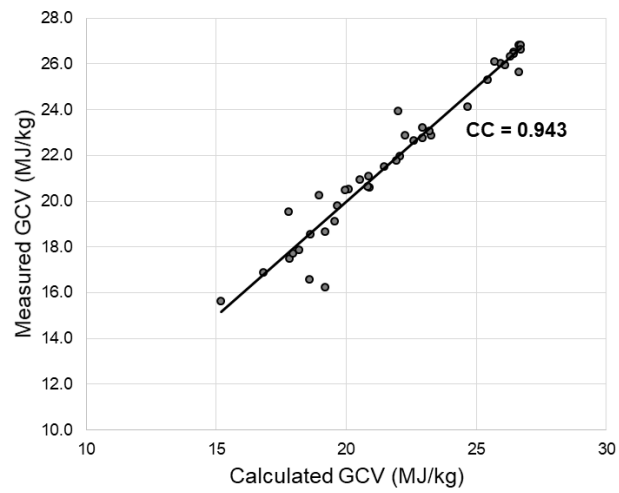


Figure B-4: Validation of the new model

Both the aforementioned results complied with the desired criteria which established the validity of the model.

