

Improving electricity forecasting processes on platinum mines

D.S. Phieros

 orcid.org/0000-0001-8843-1750

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

Supervisor: Dr. J.F. van Rensburg

Co-supervisor: Dr. C. Pelser

Graduation: May 2022

Student number: 37202618

ABSTRACT

Title: Improving electricity forecasting processes on platinum mines

Author: DS Phieros

Supervisor: Dr. Johann van Rensburg

Keywords: Electricity forecasting, Intensity model, Regression modelling

Electricity is a fundamental resource used in mining, and as a result, is a large contributor to mine working cost. It is therefore important for mines to accurately budget their electricity expenditure in order to make educated financial decisions. Inaccurate electricity budgeting can have a negative effect on capital and operational planning.

An important step in the electricity budgeting process is the electricity usage forecast. Electricity usage forecasting is used to predict the mine's electricity usage based on specified system inputs, typically production. This is difficult to perform accurately due to the complexity of mining systems. Furthermore, the existing forecasting process uses overly simplified models – intensity models. These models are relatively inaccurate when compared with forecasting processes used in other industries.

The negligence of detailed operational changes in the forecasting process is the main contributor to the inaccuracies observed. Therefore, a need exists for the application of technical knowledge and more advanced modelling processes to mine data. This will be done in order to develop more accurate and robust electricity forecasts with the aim of improving the mine's financial planning and energy management.

To address the problem, a methodology was developed by evaluating forecasting processes followed in other industries. This method was designed to consider: 1) model definition, 2) data collection and preparation, 3) model development and testing and 4) model interpretation. This methodology was applied to three case study mines, comprised of five deep-level, conventional mine shafts in total, with the exclusion of downstream concentrators, smelters and processing plants.

Using pre-selected input variables, a variety of black-box models of different complexities were fit to the data using ordinary least squares regression for the prediction of electricity usage. The performance of these models was then compared with the use of a range of performance metrics. This was done in order to select the most accurate model for each case study. The final forecasting models were tested by using the models to forecast the electricity usage of each case study. These results were compared with the forecast produced by using the existing method in order to validate the proposed methodology.

The application of the proposed methodology resulted in a reduction in mean absolute percentage error from 10.35% to 7.69%. This is complementary to a 37.5% reduction in root mean squared error from 0.62 GWh, using the intensity method, to 0.38 GWh using the proposed methodology. This reduction illustrates the clear improvement in forecasting accuracy when using the proposed method.

Finally, an improved forecasting methodology was developed and applied to three platinum mines, comprised of five shafts in total. The new methodology involves the application and comparison of different models and offers improvements in forecasting accuracy.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr Chantelle Pelser for her detailed guidance, expert advice and time that she dedicated to assist me in completing this dissertation. Furthermore, I would like to thank my family for providing me with the means to complete this degree. To my fiancé, Anna, thank you very much for your consistent support and understanding throughout the writing process.

Finally, I would like to thank Mr. George Mathews and Prof. Edward Mathews from ETA Operations (Pty) Ltd for providing me with the funding and resources to be able to complete this dissertation.

TABLE OF CONTENTS

ABSTRACT	I
ACKNOWLEDGEMENTS	III
TABLE OF CONTENTS.....	IV
LIST OF TABLES	VIII
LIST OF FIGURES.....	X
CHAPTER 1 LITERATURE REVIEW.....	1
1.1 Platinum mine complexity and variability.....	1
1.1.1 Background on platinum mine operations and electricity forecasting	1
1.1.2 Sub-system complexities and electricity usage	2
1.2 Electricity forecasting methods.....	10
1.2.1 Electricity forecasting processes in the mining industry	10
1.2.2 Electricity forecasting processes in other industries	15
1.2.3 Critical comparison	53
1.3 Problem statement	55
1.4 Need for the study	56
1.5 Study objective	56
1.6 Dissertation overview.....	57
CHAPTER 2 METHODOLOGY	59
2.1 Introduction	59
2.2 Identification of model purpose.....	59

2.3	Data collection and preparation	60
2.3.1	Selection of system boundary	61
2.3.2	Data acquisition	62
2.3.3	Data cleaning	63
2.3.4	Data transformation	64
2.3.5	Data splitting.....	64
2.3.6	Summary	66
2.4	Model development	66
2.4.1	Model type selection	67
2.4.2	Variable correlation assessment.....	67
2.4.3	Model fitting and testing.....	68
2.4.4	Model verification.....	68
2.4.5	Summary	69
2.5	Model interpretation	69
2.6	Methodology validation.....	70
2.7	Conclusion.....	71
CHAPTER 3 RESULTS	73	
3.1	Introduction	73
3.2	Identification of model purpose.....	73
3.2.1	Case study background	74
3.2.2	Model purpose.....	76
3.3	Data collection and preparation	77
3.3.1	Selection of system boundary.....	77

3.3.2	Data acquisition	79
3.3.3	Data cleaning	81
3.3.4	Data transformation	83
3.3.5	Data splitting.....	85
3.3.6	Summary	89
3.4	Model development	89
3.4.1	Model type selection	90
3.4.2	Variable correlation assessment.....	91
3.4.3	Model fitting and testing.....	93
3.4.4	Model verification.....	97
3.4.5	Summary	101
3.5	Model interpretation	105
3.6	Methodology validation.....	107
3.6.1	Electricity intensity forecasting	107
3.6.2	Monthly forecast simulation and error comparison.....	109
3.7	Conclusion.....	114
CHAPTER 4 CONCLUSIONS AND RECOMMENDATIONS		116
4.1	Introduction	116
4.2	Overview of study.....	116
4.3	Satisfying objectives	118
4.3.1	Extraction of modelling process flow from literature	118
4.3.2	Research of techniques applied in general modelling process.....	118

4.3.3	Development of a strategy for deep-level platinum mine electricity usage forecasting	118
4.3.4	Verification of proposed strategy with case studies	119
4.3.5	Objective comparison of proposed strategy to the existing forecasting method	119
4.4	Recommendations for further research	119
4.4.1	Inclusion of mine shift data as an input variable	119
4.4.2	Inclusion of time-series analysis to electricity forecasting methods	120
4.4.3	The utilisation of more granular data	120
4.4.4	Application of the forecasting process to other mining groups	120
4.4.5	Investigation into the independent variables effecting mine electricity usage ..	121
4.4.6	Addressing how models can be kept accurate and representative of the dynamic mining systems	121
4.4.7	Improvement and evaluation of mine production forecasting	121
4.5	Closure of study	121
REFERENCES		123
APPENDICES		134
Appendix A – Literature review detailed findings		134
Appendix B – Shaft electricity consumption models		150
	Data collection and preparation	150
	Model development	156
Appendix C – Compressed air electricity consumption models		171
	Data collection and preparation	171

Model development	177
Appendix D – Ventilation electricity consumption models	192
Data collection and preparation	192
Model development	195

LIST OF TABLES

Table 1.1: Statistical summary of platinum shaft production and compressed air generation electricity usage	5
Table 1.2: Statistical summary of platinum shaft production and pumping electricity usage	7
Table 1.3: Statistical summary of platinum shaft production and hoisting electricity usage	8
Table 1.4: Statistical summary of platinum shaft production and ventilation electricity usage	10
Table 1.5: Evaluating electricity forecasting models against the general modelling approach [46]	17
Table 1.6: Summary of the types of variables considered by the evaluated studies	22
Table 1.7: Summary of the data resolution considered in each of the evaluated studies	22
Table 1.8: Interpretation of Pearson's correlation coefficient [110],[112]	44
Table 1.9: Abbreviations of performance metrics and their meanings	45
Table 2.1: Examples of the minimum number of data points required as a function of the number of input variables	63
Table 3.1: Summary of system boundaries and associated sub-systems considered for case study mines	78
Table 3.2: Variables selected for the electricity forecasting on the case study mines	79
Table 3.3: Summary of data acquired for forecasting process	80
Table 3.4: Summary of outlier detection for all systems	82

Table 3.5: Shaft A variable data ranges.....	84
Table 3.6: Sizes of train/test and verification datasets for each system	86
Table 3.7: Summary of data acquisition and preparation process for all systems	89
Table 3.8: Summary of Pearson's correlation coefficients for each input variable and each output variable to be forecasted	93
Table 3.9: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft C.....	95
Table 3.10: Training set and verification set performance metrics of the final linear model forecasting the electricity consumption of shaft C	98
Table 3.11: Summary of final models and percentage difference between the performance metrics of the verification and training sets for all systems considered	103
Table 3.12: Summary of the most influential variables for each of the modelled systems	105
Table 3.13: Summary of the operating ranges of all of the modelled systems	106
Table 3.14: Performance metrics for the intensity models of each system calculated on the data used to fit the models	108

LIST OF FIGURES

Figure 1.1: Platinum mine compressed air ring with multiple compressor houses and shafts	3
Figure 1.2: Monthly compressed air generation electricity usage of a platinum mine shaft	4
Figure 1.3: Example of the underground dewatering system of a platinum mine shaft.....	6
Figure 1.4: Monthly pumping electricity usage of a platinum mine shaft	6
Figure 1.5: Monthly hoisting electricity usage of a platinum mine shaft.....	8
Figure 1.6: Monthly ventilation electricity usage of a platinum mine shaft	9
Figure 1.7: Summary of forecasting process used in the majority of deep-level platinum mines	13
Figure 1.8: Comparison between electricity intensity and linear relationships on data	14
Figure 1.9: Generic procedure for modelling a variable as a function of its inputs [46].....	16
Figure 1.10: Data acquisition and cleaning step in overall modelling process.....	20
Figure 1.11: Summary of the number of variables considered by each evaluated study.....	21
Figure 1.12: Summary of the number of data points considered by each evaluated study	23
Figure 1.13: Examples of models that are underfit (left), fit well (centre) and overfit (right)	30
Figure 1.14: Summary of train-test split ratio used in evaluated studies	30
Figure 1.15: Summary of the train-test split methods used by the evaluated studies	31
Figure 1.16: Model development step in overall modelling process	33
Figure 1.17: Summary of the number of models considered by the evaluated references	34
Figure 1.18: Model complexity of the models considered by the evaluated studies	35
Figure 1.19: Model linearity of the models considered by the evaluated studies	36
Figure 1.20: Example of the histogram and boxplot for a normal distribution	39

Figure 1.21: Example of the histogram and boxplot for a skewed distribution.....	39
Figure 1.22: Example residual plot against predicted values to test heteroscedasticity assumption.....	41
Figure 1.23: Example distribution plot of residuals.....	41
Figure 1.24: Examples of scatter plots between variables and related Pearson's correlation coefficients[110].....	44
Figure 1.25: Summary of the performance metrics used in the evaluated studies	45
Figure 1.26: Example and desired relationship between forecasted values and actual values.....	50
Figure 1.27: Generic procedure for modelling a variable as a function of its inputs [46].....	52
Figure 1.28: General modelling/forecasting procedure used in other industries.....	53
Figure 1.29: Critical comparison between the mining industry forecasting process (left) and the forecasting process used in other industries (right).....	54
Figure 2.1: Four main steps for the proposed method of electricity usage forecasting on deep-level platinum mines.....	59
Figure 2.2: Identification of model purpose sub-steps.....	60
Figure 2.3: Data collection and preparation sub-steps.....	60
Figure 2.4: System boundaries that can be selected for forecasting depending on data availability.....	61
Figure 2.5: Illustration of the randomisation and separation of the total data into verification and training/test sets	65
Figure 2.6: Illustration of k-fold cross-validation using the training/testing dataset	65
Figure 2.7: Model development and testing sub-steps.....	66
Figure 2.8: Model interpretation sub-steps.....	69
Figure 3.1: Proposed forecasting methodology steps and sub-steps.....	73
Figure 3.2: Case study mine 1 layout	74

Figure 3.3: Case study mine 2 layout	75
Figure 3.4: Case study mine 3 layout	76
Figure 3.5: Shaft E electricity usage outliers and boxplot method	83
Figure 3.6: Frequency plots showing the distributions of variables considered for forecasting shaft A electricity and compressed air electricity consumption	85
Figure 3.7: Data points contained in the train/test and verification sets for shaft A.....	87
Figure 3.8: K-fold test splits for the scaled data used for the modelling of shaft C electricity consumption.....	88
Figure 3.9: Example scatter plots between input variables and target variables for shaft A and shaft B.....	90
Figure 3.10: Scatter plots and Pearson's correlation coefficients between all input variables and the shaft electricity usage for shaft C.....	92
Figure 3.11: Power, linear and polynomial forecast models fit and tested on four folds for the prediction of the shaft electricity usage of shaft C.....	94
Figure 3.12: Final linear model fit to the total training set to model the electricity consumption of shaft C.....	96
Figure 3.13: Final linear model for the forecasting of the shaft electricity consumption of shaft C plotted against the training and verification sets	98
Figure 3.14: Scatter plot between forecasting and actual electricity consumption of shaft C ($R^2=0.807$).....	99
Figure 3.15: Frequency plot of residuals between actual and forecasted values of shaft electricity consumption of shaft C	100
Figure 3.16: Scatter plot between forecasted electricity consumption values and residuals for the shaft electricity consumption of shaft C	101
Figure 3.17: Comparison of the performance of the proposed models and the intensity models for all forecast systems on the validation set	110

Figure 3.18: Residual distribution plot comparing the proposed model and intensity model residuals.....	112
Figure 3.19: Scatter plot of forecasted and actual electricity usage for the intensity ($R^2=0.968$) and proposed models ($R^2=0.986$) of all systems.....	113

CHAPTER 1 LITERATURE REVIEW

1.1 Platinum mine complexity and variability

1.1.1 Background on platinum mine operations and electricity forecasting

Mining is an electricity-intensive industry in South Africa and is responsible for the consumption of 15% of the total electricity generated by Eskom [1]. Platinum mines consume approximately a third of this total supply [2]. This electricity is used to power a variety of sub-systems that comprise the total mining facility. The main sub-systems include: 1) compressed air generation, 2) pumping, 3) hoisting and 4) ventilation [3].

The continuous maintenance and optimisation of each of these sub-systems is required for safe and effective mining operation. This includes the management of the various pieces of equipment making up each sub-system. The variability in equipment and system interconnection makes management of platinum mines a complex task. Detailed planning and budgeting is thus required to ensure profitability [4],[5]. This requires the planning of operational and capital expenses of the facility [6].

Budgets are financial plans that are used to communicate the goals of a company [7]. The relationship between operational and capital targets are considered by dividing the master budget into financial, static, and operational budgets [7]. These budgets address the capital, fixed and day-to-day expenses of the facility, respectively. The correctness and transparency of each of these sections are important for overall budget accuracy [8],[9].

These budgets are typically generated by financial departments, with limited technical operational understanding. As such, technical operation of the mine is often assumed to remain consistent year-on-year [10]. This assumption leads to significant inaccuracy in the operational budget to the point where the master budget becomes obsolete [11]. Inaccurate budgets result in the allocation of unrealistic goals and expectations to mine personnel, which negatively effects the mine's operational management [12].

Electricity usage is a large contributor to mine operational expense, contributing up to a quarter of the total [13]. Therefore, a large proportion of the mine's operational budget is related to the accurate budgeting of electricity usage. Currently, electricity budgets are calculated by considering production changes as well as planned growth projects [10]. Production changes are considered with forecasting techniques. Electricity usage forecasting is where expected changes in electricity usage, relative to planned production, are calculated and incorporated into the total budget [14].

The existing forecasting process is based on the calculation and assumption of an average electricity intensity (electricity used per tonne of production hoisted) [15]. This is typically used as a rough estimator of facility performance, and due to its simplicity, is also used for forecasting purposes [15]. This method, however, tends to oversimplify the relationship between the variables, electricity usage and production. This over-simplification ultimately leads to reduced budget accuracy [16].

In order to improve forecasting accuracy, the current electricity forecasting processes used in other industries were investigated. These were compared with forecasting methods used on deep-level platinum mines. The forecasting processes used in other industries were finally applied to mine electricity forecasting and compared with the existing forecasting method used by mines. The purpose is to improve the current methods used for forecasting by platinum mines by incorporating relevant techniques used in other industries.

1.1.2 Sub-system complexities and electricity usage

The different sub-systems that comprise platinum mine operations are discussed in this chapter. The purpose is to provide insight into the basic operation of each sub-system as well as the role that each sub-system plays in platinum mining. These sub-systems are: 1) compressed air generation, 2) pumping, 3) hoisting and 4) ventilation. The multifaceted relationships that exist between the systems and how they relate to electricity usage will contextualise the necessity for more accurate forecasting techniques.

Compressed air generation

Compressed air consumes up to 21% of the electricity used by a deep-level platinum mine [17]. It is required for the operation of underground pneumatic equipment, such as pumps, actuators, pneumatic loaders and rock drills [18]. It is also required to maintain positive pressure in underground refuge chambers [19]. This equipment requires a consistent, reliable supply of compressed air to meet safety and production requirements underground [20].

This supply is ensured with the use of centrifugal compressors with installed capacities ranging from 1 MW to 15 MW. In some instances multiple machines are often connected to meet demand [21]. Sets of machines are typically localised in buildings referred to as compressor houses, with some networks having multiple sets of interconnected compressor houses [22],[23].

The compressed air is transported from each compressor house to each shaft, as well as underground with the use of steel piping networks that can reach up to 75 km in length [24]. These are complex networks and can contain a significant number of air leaks resulting in large pressure losses in the system [25],[26]. These leaks are prevalent in South African mines due to their size

and age and their presence adds complexity to the system. Figure 1.1 illustrates the layout of a typical compressed air ring on a platinum mine.

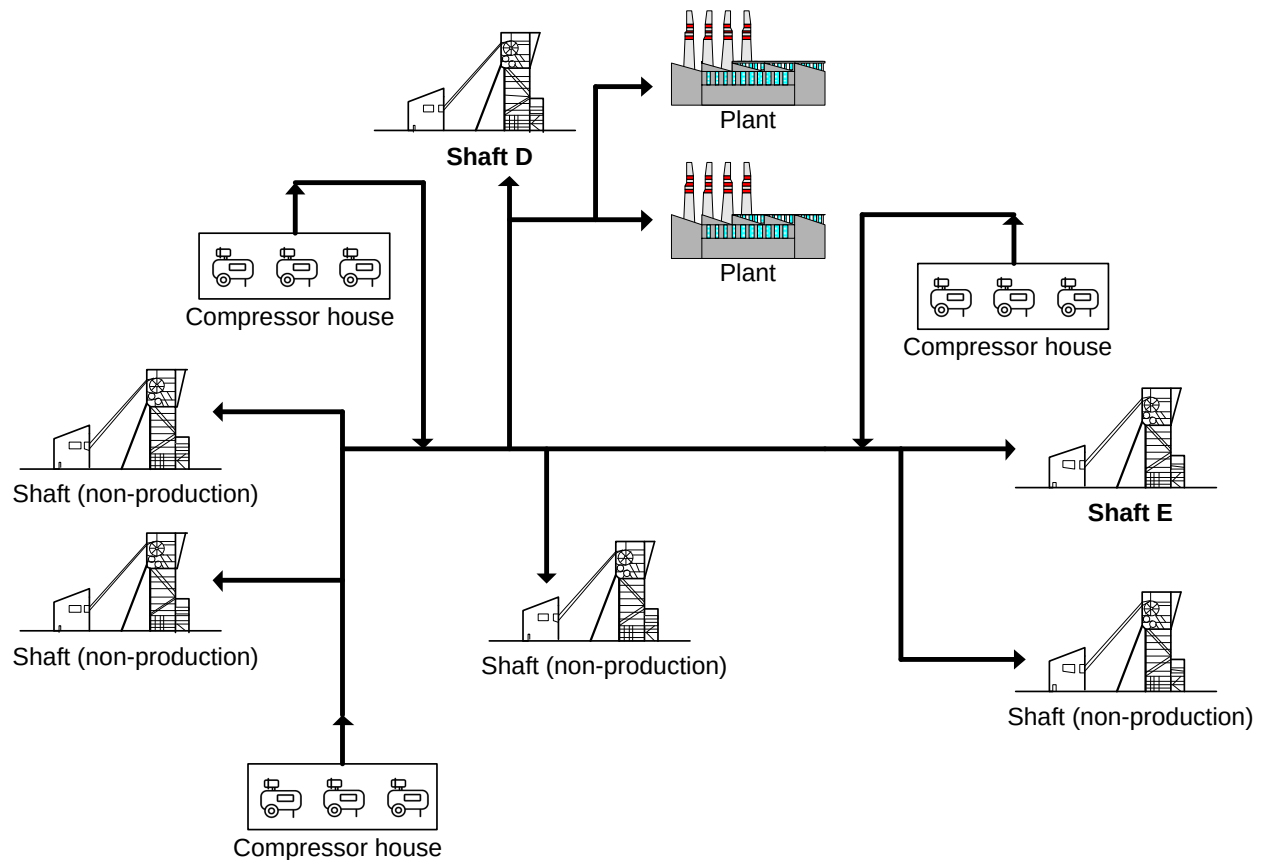


Figure 1.1: Platinum mine compressed air ring with multiple compressor houses and shafts

The compressed air ring depicted in Figure 1.1 shows the complexity of the air flow system underground. Multiple shafts and plants are connected to more than one compressor. As such, the electricity utilisation for compressed air generation in these networks is also divided among the shafts and plants. This is usually done by multiplying the total compressor electricity usage by the fraction of the total generated compressed air used by each shaft or plant.

Figure 1.2 shows the scatter plot between the compressed air generation electricity usage of a platinum mine shaft and the production output of that shaft. The aim of this graph is to illustrate the potential relationship between the two variables as well as outline the ranges in compressed air electricity usage and production. It must be noted that range is calculated as the difference between the maximum and minimum values.

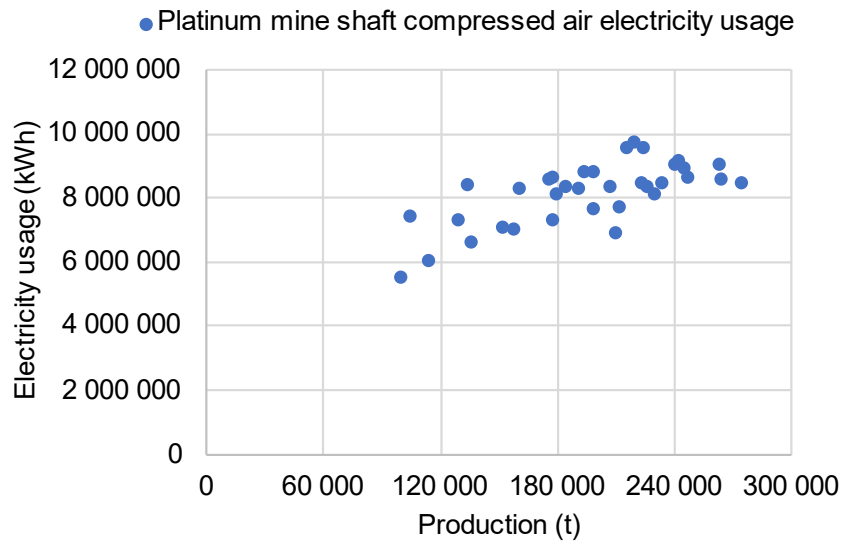


Figure 1.2: Monthly compressed air generation electricity usage of a platinum mine shaft

Figure 1.2 illustrates that the amount of compressed air generation energy increases with production. Furthermore, the relationship between production and compressed air electricity usage is not obviously linear and appears to follow a power fit – plateauing at production values larger than 240 000 t. The data also seems to follow this trend with no immediately apparent outliers. Figure 1.2 shows that the production for this shaft ranges from 99 500 t to 275 000 t and compressed air generation energy for this shaft ranges from approximately 5.52 GWh to 9.73 GWh.

The variability of the data can be quantitatively assessed with the use of the coefficient of variation [27]. The coefficient of variation is a dimensionless measure of the dispersion of data about the mean. The higher the value of the coefficient of variation, the greater the level of dispersion of data about the mean [27]. It is calculated as the quotient of the standard deviation and the mean multiplied by 100 as shown in Equation (1.1).

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 \quad (1.1)$$

This term, therefore, allows the consideration of the standard deviation of the data relative to the mean of the data. Furthermore, since it is dimensionless, it allows for the comparison between distributions of values whose scales of measurements are not comparable [27], specifically, the comparison of the dispersion of different variables to one another. The averages, standard deviations and coefficients of variation are shown Table 1.1.

Table 1.1: Statistical summary of platinum shaft production and compressed air generation electricity usage

Parameter	Mean	Standard deviation	Coefficient of variation
Production	195 180 t	45 721 t	23.4%
Compressed air generation energy	8 157 817 kWh	952 234 kWh	11.6%

Table 1.1 contains the mean and standard deviation of the variables, used for the calculation of the coefficient of variation. Table 1.1 indicates that the monthly coefficient of variation for production and compressed air generation energy is 23% and 11% respectively. This large variation in production and electricity usage indicates that the shaft operates over a large operating range. These large variations in monthly totals and complex relationship between the variables justify the need for system forecasting by means of a more advanced methodology.

Pumping

Water is an essential resource utilised in the platinum mining process. Water sent underground for mining operations, as well as fissure water (water released from fractures in underground rock [28]) must invariably be pumped to surface. The effective removal of water from underground, referred to as dewatering, is necessary for the prevention of mine flooding.

Used process water and fissure water are pumped from working areas into pump dams located on a variety of underground levels [3], [29]. The removal of this water from underground dams is accomplished by pumping the water to sequential pump dams found at increasingly shallower depths. This is repeated until the water is eventually pumped to the surface dam. The depth of platinum mines, which can be over 1 000 m, results in dewatering contributing up to 18% of the electricity consumed by deep-level mines [17].

Dewatering is performed with the use of pump stations located close to underground pump dams. These pump stations can be comprised of numerous centrifugal pumps. Station pumping capacities can be as much as 20 MW depending on the water balance and pumping distance between dams [29]. Figure 1.3 illustrates the presence of multiple pump stations and dams of an underground dewatering system.

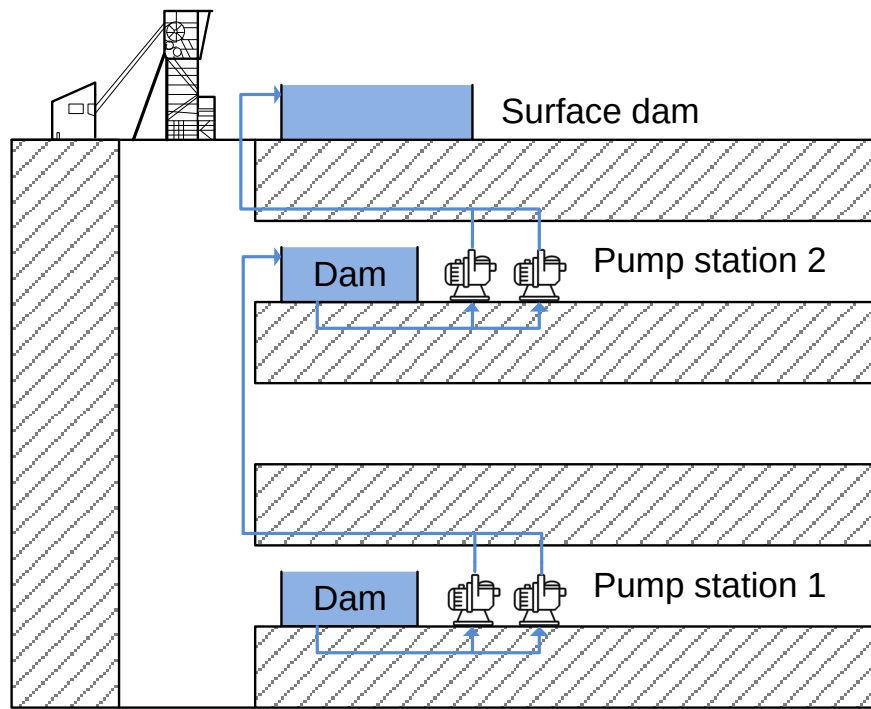


Figure 1.3: Example of the underground dewatering system of a platinum mine shaft

Figure 1.3 illustrates an example of an underground dewatering system on a platinum mining shaft. It is clear from this figure that the dewatering system can be comprised of a number of different pumps and works by pumping the water up sequential levels. It must be noted that it is not infrequent for water to flow between neighbouring shafts underground. This adds an extra layer of complexity when trying to model the electricity usage of these systems. Figure 1.4 illustrates a scatter plot of pumping electricity usage and production for a platinum shaft.

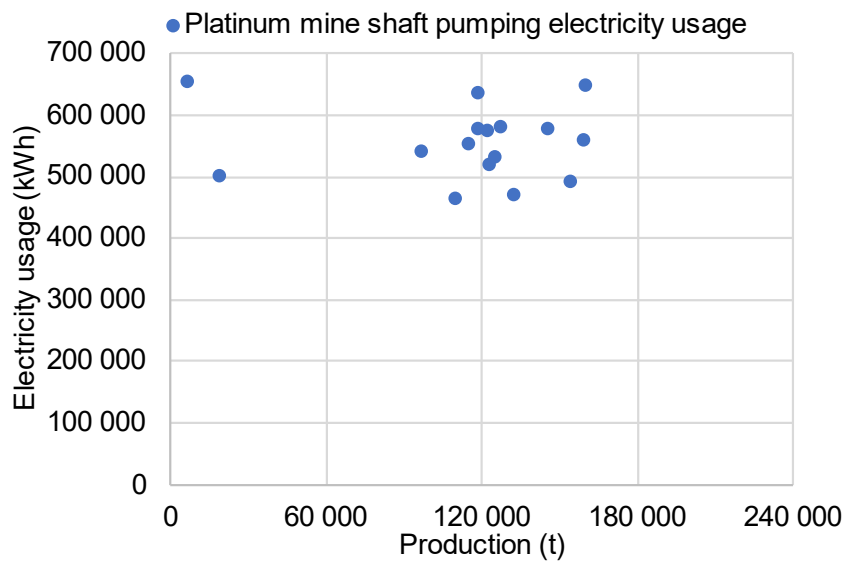


Figure 1.4: Monthly pumping electricity usage of a platinum mine shaft

Figure 1.4 illustrates that there is no correlation between pumping electricity usage and production. This is understandable as there are other drivers contributing to underground water, such as fissure water, as previously mentioned. Figure 1.4 also shows that the production for this shaft ranges from 6 500 t to 160 000 t and the pumping energy for this shaft ranges from approximately 0.46 GWh to 0.65 GWh. The averages, standard deviations and coefficients of variation for the data are shown Table 1.2.

Table 1.2: Statistical summary of platinum shaft production and pumping electricity usage

Parameter	Mean	Standard deviation	Coefficient of variation
Production	114 221	42 247	37.0%
Pumping energy	555 392	57 020	10.3%

Table 1.2 indicates that the variation in the production for this shaft is approximately 3.5 times the variation of the pumping electricity usage. This implies that although this shaft operates over a reasonable production range, there is little variation in the pumping electricity usage. Furthermore, as seen in Figure 1.4, there is no obvious correlation between these two variables.

Hoisting

Hoisting is the process of transporting ore, material, equipment, and employees up and down the mine shaft [30]. It contributes 14% of the electricity consumption on deep-level mines [17]. This is optimally done with the use of a carefully worked out winder schedule [31]. The aim of the winder schedule is to minimise the number of trips required to transport the necessary equipment, material, or people to and from the appropriate levels. This system is therefore used intermittently, depending on the timing of shifts [32].

Metal accounting according to the AMIRA P754 Code is applied in order to ensure a traceable record of the movement of platinum throughout the system [33]. This hoisted ore is used to measure shaft performance, as well as to keep a record of the ore being sent for processing. Once on surface, the ore is transported with the use of train hoppers, trucks, or conveyors from the shafts to the plants. The weighing of this ore, referred to as 'hoisted tonnes', is performed with in-motion rail weighers, truck weighbridges or conveyor belt weighers [34].

The effective weighing of this hoisted ore is fundamental for metal accounting and plays a significant role in production management and forecasting. Since forecasting is dependent on production as an input, the possession of accurate shaft production figures is of the utmost importance. In order to ensure accuracy, weighline systems are calibrated once a week by weighing a hopper of known mass and once every six months by an outside agency [35].

The winder motor typically has a capacity between 1 MW and 4 MW depending on the size of the load [3]. The effective operation of this system is important in order to ensure the safety of mining employees and to maintain consistent production. The electricity usage of the hoisting system of a typical platinum shaft is shown by the scatterplot in Figure 1.5.

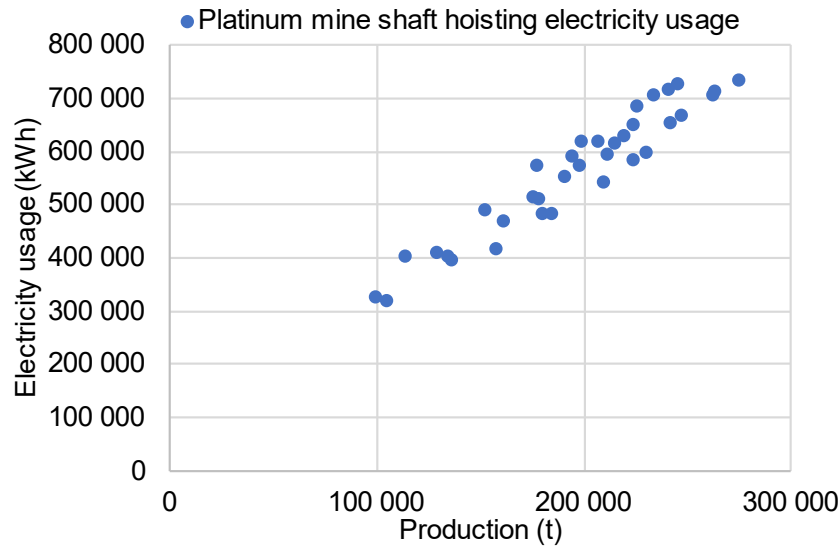


Figure 1.5: Monthly hoisting electricity usage of a platinum mine shaft

Figure 1.5 illustrates that production for this shaft ranges from approximately 99 500 t to 275 000 t and hoisting electricity usage for this shaft ranges from approximately 0.32 GWh to 0.73 GWh. Furthermore, the relationship between production and hoisting electricity usage appears to be linear, unlike that of compressed air electricity usage. This relationship is understandable as the winder system's electricity usage is directly related to the mass of ore hoisted from underground. The averages, standard deviations and coefficients of variation are shown Table 1.3.

Table 1.3: Statistical summary of platinum shaft production and hoisting electricity usage

Parameter	Mean	Standard deviation	Coefficient of variation
Production	195 180 t	45 721 t	23.4%
Hoisting energy	562 870 kWh	116 561 kWh	20.7%

Table 1.3 indicates that the monthly coefficient of variation for production and hoisting electricity usage are 23% and 20.7% respectively. This indicates that that hoisting electricity usage varies by the same magnitude with respect the production. Similar to the previous systems, a forecasting methodology that takes this relationship into account will increase forecasting accuracy.

Ventilation

Ventilation is critically important for maintaining underground safety. It is responsible for the removal of noxious and flammable gases which are liberated from underground mining activity [36]. It is also responsible for maintaining favourable operating conditions underground by removing dust, diesel fumes, and moisture as well as regulating the temperature [37].

Ventilation systems consist of large surface fans, smaller auxiliary or booster fans and a complex duct system used to circulate fresh air underground [37]. The surface fans are typically the largest in size and are seen as the main drivers of ventilation. Auxiliary or booster fans are typically placed in deeper, more intricate areas of the mine to maintain airflow. The rating of the ventilation fans varies significantly depending on the location and function of the fan.

With large variability in fan function, fans can have ratings between 100 kW and 3000 kW [38]. Furthermore, these fans run on a continuous basis at reasonably constant power at operational shafts ensuring a safe working environment at all times. Ventilation contributes 8% of the total electricity consumption on deep-level mines [17]. Figure 1.6 shows the scatter plot between monthly ventilation electricity usage and production of a platinum mine shaft.

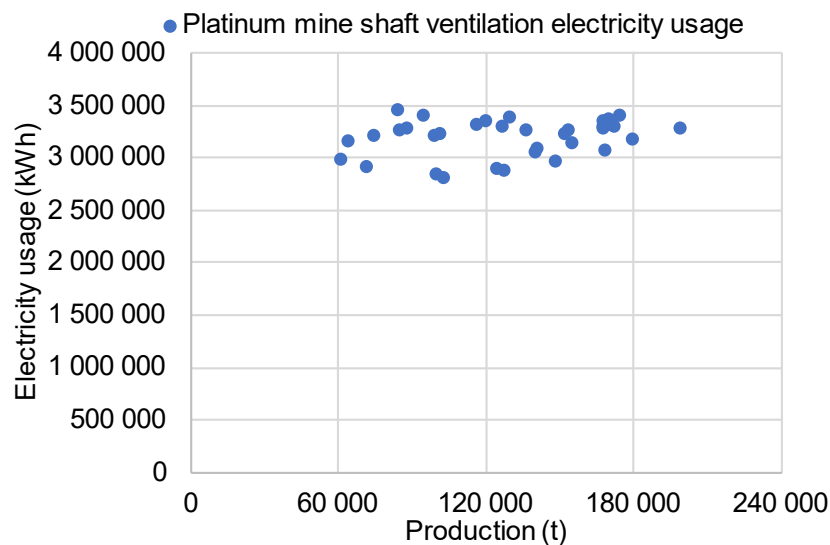


Figure 1.6: Monthly ventilation electricity usage of a platinum mine shaft

Figure 1.6 indicates that although the shaft operates over a large production range, ranging from 60 500 t to 200 000 t, the variation in ventilation electricity usage remains relatively small – with a range between 2.81 GWh and 3.46 GWh. Furthermore, there is no immediately noticeable correlation between these two variables. Since this system runs continuously, it is understandable that there is no correlation between this system's electricity usage and production. The averages, standard deviations and coefficients of variation are shown Table 1.4.

Table 1.4: Statistical summary of platinum shaft production and ventilation electricity usage

Parameter	Mean	Standard deviation	Coefficient of variation
Production	128 974 t	37 340 t	28.9%
Ventilation energy	3 199 123 kWh	174 915 kWh	5.5%

Table 1.4 confirms the insights from Figure 1.6, specifically that the coefficient of variation of ventilation electricity usage is only 5.5%, although that of the shaft production is 28.9%. This indicates that this system does not vary as significantly as production. This insight into system operation must, however, be considered when this system and the total system electricity usage are forecasted.

Summary

The systems described in this section are the main contributors to deep-level mine electricity consumption, contributing up to 80% of the total usage. These systems are complex and have ever-changing layouts due to the dynamic nature of mining. Furthermore, the relationship between each system's electricity consumption and production is different as a result of the nature of each system.

While some systems, such as compressed air generation and hoisting, vary with production, pumping and ventilation do not. Furthermore, systems whose electricity consumption varies with respect to production appear to vary according to different relationships. These relationships between production and electricity usage can also vary between shafts and between mines depending on the system layout and design of the mine.

This variability and the inherent system complexity make the forecasting of the system electricity a complex task. The forecasting process must consider the different mining sub-systems, as well as the relationships between these systems' electricity usage and their associated inputs. The application of a thorough forecasting process that takes this system variability and complexity into account can significantly improve forecasting accuracy.

1.2 Electricity forecasting methods

1.2.1 Electricity forecasting processes in the mining industry

Deep-level platinum mines are made up of a multitude of different electricity consuming systems, as discussed in Section 1.1. Each of these systems is necessary for the recovery of platinum and contributes to the overall electricity usage. Capacity to forecast the electricity usage and cost of

these systems effectively allows for the generation of more accurate budgets. This allows for better resource planning and management which can improve the profitability of the mine [9],[39].

A study was performed in 2020 on the electricity budgeting structures of eight South African mines. The sample group represented 87% of the deep-level mining shares in the country and is, therefore, considered to be representative [10],[40]. The results of this study indicated that production forecasts were included in the budgeting processes of all of the mines surveyed. Furthermore, the change in electricity usage associated with production changes was calculated with the use of electricity intensity models at all of the mines except two [40].

The electricity intensity model is calculated as the amount of electricity used per unit of industrial output, or production in the case of mining [41]. For the mines in question, production is measured in hoisted tonnes. Material extracted from underground is categorised as either reef tonnes or waste tonnes depending whether it contains platinum or not [42]. Hoisted tonnes refer to the number of reef and waste tonnes extracted at each particular mine shaft [40],[43].

Production is therefore paramount for the forecasting process but also introduces variability. Firstly, various shafts mine different platinum reefs, such as the Merensky or UG2 reefs found in South Africa. Secondly, those on the same reefs mine different sections of the reef, typically at varying depths [44]. Production goals and efficiencies, therefore, vary from shaft to shaft and across mines. Mine shafts are also found to practise the stockpiling of material underground [10].

Stockpiling results in a time delay between when material is mined and when it is hoisted and weighed. This implies that the electricity measured on a specific day is not necessarily correlated to the material weighed on that day. This introduces error into the modelling process, as daily production figures might not be a true representation of actual extracted material.

This greatly impacts forecasting in small resolutions, such as hourly or daily intervals. This is because the energy consumed to mine the material will not necessarily correlate with the tonnes hoisted for that same interval. The choice of measurement resolution must, therefore, consider these system nuisances in order to avoid the inclusion of unnecessary error into the process. The selection of measurement resolution is therefore considered to be of high importance.

For mines, electricity intensity is a representation of the amount of electricity utilised for the extraction of a unit of material, platinum bearing or not. It can be calculated for different system boundaries and over different measurement resolutions. The choice of system boundary and measurement resolution is dependent on a range of factors. These include the degree of metering available, the resolution of production planning as well as the mine budget resolution [10].

System boundaries determine which systems' electricity usage will be included in the electricity intensity calculation [7]. For example, an electricity intensity can be calculated for 1) each of the sub-systems of the mine discussed in Section 1.1.2, 2) the different shafts of a mine, or 3) for the mine as a whole. The system boundary selected depends on the desired outcome and conclusions wished to be drawn from the evaluation.

The measurement resolution determines the time intervals between the data points used for the calculation [7]. Depending on the chosen resolution, the electricity intensity can be calculated on a daily, monthly, or quarterly basis, for example. The implication of stockpiling, as discussed above, greatly impacts the effect of production in the energy intensity, and overly granular resolutions are therefore not recommended.

In the forecasting process, once the electricity intensity has been calculated, it is used in conjunction with a production forecast to calculate the expected electricity usage for each system [10]. The production forecast is the production expected for a specific budgeting period. It is typically calculated per shaft and on the same measurement resolution as the electricity intensity. The electricity intensity and electricity usage forecast are calculated with the use of Equation (1.2) and Equation (1.3), respectively [10].

$$\hat{E} = \frac{E_{historical}}{P_{historical}} \quad (1.2)$$

$$E_{forecast,i} = \hat{E} \times P_{forecast,i} \quad (1.3)$$

where:

$\hat{E}$	Electricity-intensity (kWh/tonne)
$E_{historical}$	Historical electricity consumption (kWh)
$P_{historical}$	Historical production (hoisted tonnes)
$E_{forecast}$	Forecasted electricity (kWh)
$P_{forecast}$	Production forecast (hoisted tonnes)
i	Time interval (dependant on measurement resolution)

Equation (1.2) indicates that the electricity intensity is calculated by dividing the historical electricity usage for a specific period by the production for the same period. This will yield the

average electricity intensity over that period. This is then used with the production forecast and Equation (1.3) to calculate the expected electricity required to yield the planned production.

The surveys completed by [40] found that 88% of the mines performed budgeting on a monthly basis, with the remainder budgeting on a quarterly basis [40]. Furthermore, the surveys indicated that the period of time used for the intensity calculation was the previous 12 months. This means that the total electricity usage for the previous 12 months is divided by the total production for the same 12 months. This yields an average electricity intensity using Equation (1.2).

The surveys also indicated that 75% of the mines surveyed developed budgets per mine shaft, with the remainder developing them for the mine as a whole [10],[40]. None of the mines were found to budget on a sub-system level due to the metering limitations, magnitude of data and associated errors [40]. The surveys also indicated that official electricity usage data used for budgeting was obtained from third-party metering companies. These companies typically only consolidate the electricity usage on a shaft or total mine level and on a monthly basis [10].

Therefore, according to the surveys found in the 2020 study, the majority of deep-level mines in South Africa perform budgeting, and therefore forecasting of electricity usage, relative to production changes. This is done with the use of the electricity intensity model described in Equations (1.2) and (1.3). Most of the mines perform the forecast by dividing the mine up into the different shafts and calculating the expected electricity usage on a monthly basis.

Based on the information provided above, the overall forecasting process used in the majority of deep-level platinum mines can be summarised as shown in Figure 1.7 [10],[40].

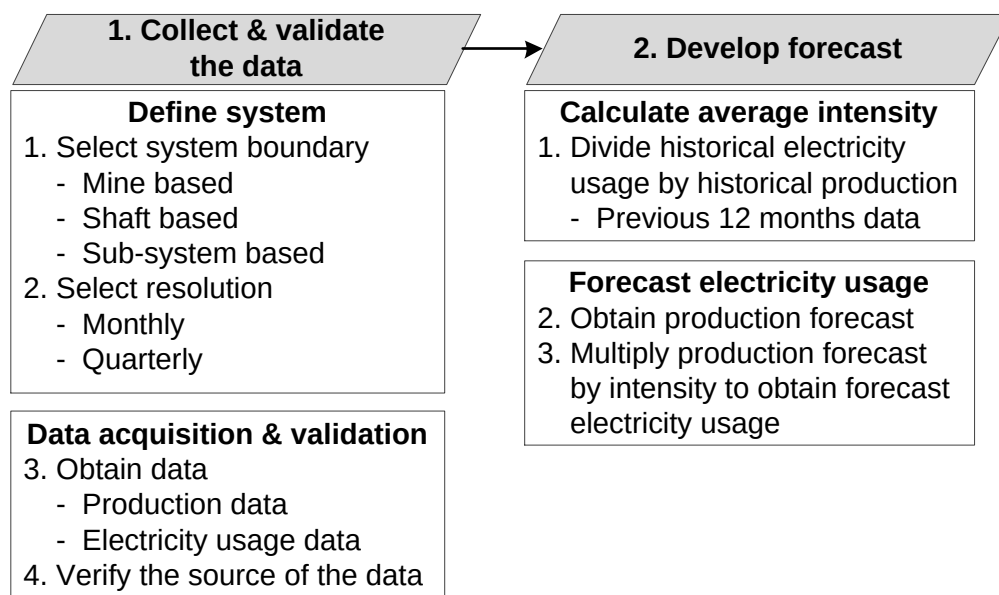


Figure 1.7: Summary of forecasting process used in the majority of deep-level platinum mines

Figure 1.7 illustrates that the main steps of the forecasting process in the mining industry are the acquisition of data (*collect & validate the data*) and the development of the model (*develop forecast*). The first step involves the selection of a system boundary and measurement resolution. Thereafter, the electricity usage and production data are obtained, and the source of the data is validated.

This data is used to calculate an average electricity intensity by considering the previous 12 months' data. The mine's production forecast is then multiplied by the average intensity to obtain the associated electricity usage for the planned production. This is a simple method, using only a single input variable and an average intensity value for the total forecast.

Although this method is easy to implement, it has an associated disadvantage. It does not address the complexity of the correlation between the electricity used and the variable(s) driving it, in this case production [16]. Due to this simplified relationship, electricity intensity does not accurately model the electricity consumption as the function of the input variable. It should rather be used as a rudimentary performance indicator [15].

Electricity intensity is, therefore, simply a metric that should be used for comparative purposes of how electricity usage is changing, but cannot be used for accurate forecasting [45]. This is purely due to the fact that the overly simplified relationship neglects the potential complexities and correlation between the selected variables. Figure 1.8 illustrates the loss of accuracy associated with the electricity intensity method for a hypothetical mining system.

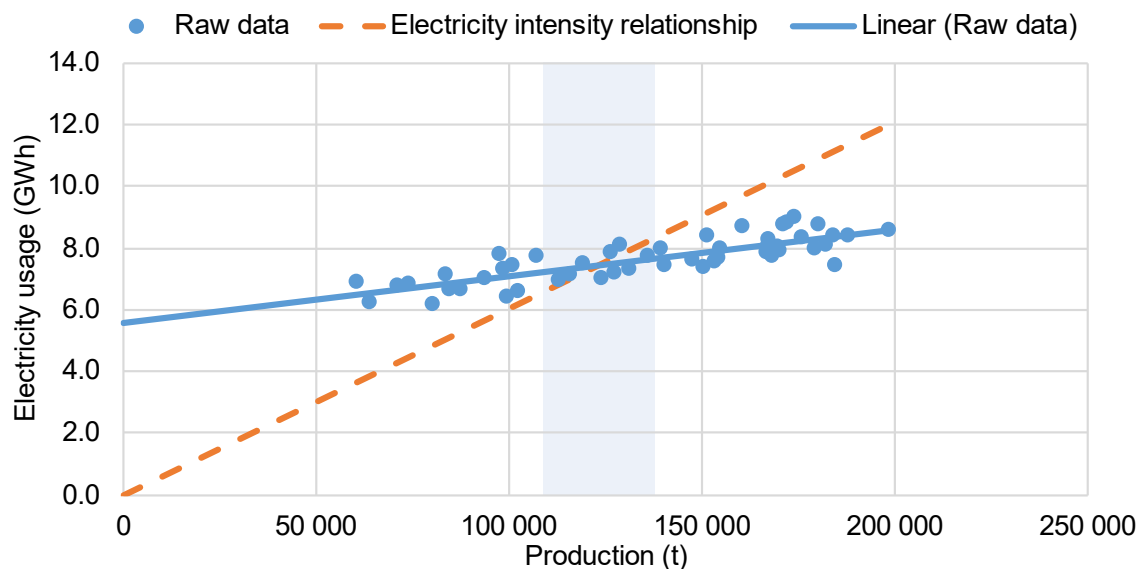


Figure 1.8: Comparison between electricity intensity and linear relationships on data

Figure 1.8 illustrates that the electricity intensity relationship assumes that no electricity is used at zero production. This, however, does not occur in practice, as the mine would still consume electricity in the event of zero production for a variety of reasons. These reasons include the continual dewatering of mine shafts of fissure water to prevent flooding [29] as well as the ventilation of shafts for the performance of maintenance [36].

Figure 1.8 also indicates that the electricity intensity model (indicated by the orange dashed line) can give a rough indication of performance for shafts or mines with limited production ranges. However, for those with wide ranges of production (production outside the blue shaded region) this model fails to accurately describe the relationship between the variables across the entire production range. This is clear due to the deviation from actual data trends outside of a specific operating range indicated by the blue shaded area [15].

For comparative purposes, a wide range of electricity forecasting methods in other industries is considered. This was done to identify whether it would be possible to apply similar methodologies to mine data. The intention of this application is to potentially improve the accuracy of the electricity forecasting methods on platinum mines.

1.2.2 Electricity forecasting processes in other industries

As discussed in section 1.2.1, the current method of electricity forecasting on platinum mines involves the utilisation of an average electricity intensity model. This, coupled with a production forecast, is used to obtain the expected future electricity consumption of the mine. Various sources were investigated in order to identify the electricity forecasting processes used in other industries.

This comparison will allow for the identification of the similarities and differences between the mining industry and others. This will allow for the critical evaluation of the current forecasting process applied on deep-level platinum mines. The aim of this evaluation is to identify whether the techniques used in other industries can be applied to the forecasting of electricity consumption on deep-level mines with the aim of improving accuracy.

A number of sources outline a generic procedure for the modelling of a variable as a function of its input variables [46],[47],[48],[49],[50],[51]. Modelling and forecasting can be seen as interchangeable terms in this instance and for the length of this study. The model should be seen as the mathematical relationship that is used to forecast the output variable given specific input variable values.

This generic procedure summarised in detail in [46] will be used as a template against which all reviewed references will be compared. This will allow for the identification of a general forecasting

process used in other industries. The use of this general process will allow the development of a methodology to be applied to mine electricity forecasting. The generic procedure is shown in Figure 1.9.

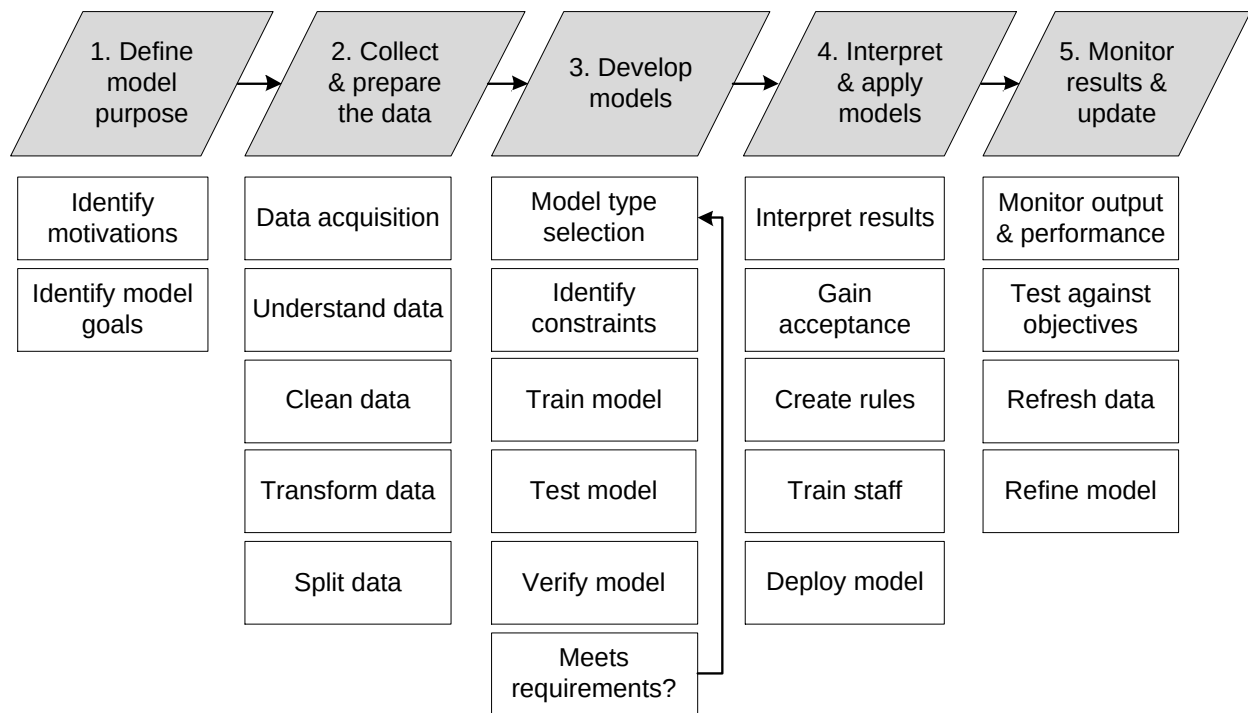


Figure 1.9: Generic procedure for modelling a variable as a function of its inputs [46]

As illustrated in Figure 1.9, the generic modelling procedure is divided into 5 main steps, the first of which is to define the purpose of the model. In the case of electricity usage forecasting, the purpose of the model is to predict the expected electricity usage of a specific system. This forecast can be used for a variety of different outcomes such as budget development, system design, and energy management.

The second step in the process is the acquisition and preparation of data that will be used to develop the model. This step ensures that data is acquired from valid sources. It also involves the removal of outliers or other factors that might introduce error. Furthermore, it ensures that the data is split into training (data used to develop the model) and testing (data used to test the model's accuracy) data sets. This will allow for the training and testing of forecasting models on different sets of data to evaluate overall model accuracy.

The third step involves the model development process. This process involves model selection, model training and model verification. The outcome of this step is to produce a model that performs with an acceptable level of accuracy when predicting on the test set. This step also accounts for the comparison of multiple models. This allows for the selection of the model with the best performance for each application.

The fourth step involves the interpretation and implementation of the developed model. In the case of this study, this involves the application of the forecasting model to actual electricity prediction given a production forecast. Once the model is used for forecasting, the fifth step ensures that the model remains up-to-date and that it continues to meet expected outcomes.

A review of 23 relevant references was conducted to identify the methods and processes used for the forecasting and prediction of electricity consumption in various industries. The modelling processes applied in these references were evaluated based on whether they followed the general process flow presented in Figure 1.9.

In order to evaluate whether the studies followed the general process flow in Figure 1.9, a set of questions was constructed. Each question was formulated to summarise the overall process followed in each of the steps. The questions relating to the steps are as follows:

- Step 1: Are the goals of the model stated?
- Step 2: Is the data collected and manipulated (either cleaned, transformed, or split)?
- Step 3: Are models developed and tested and accuracies compared?
- Step 4: Are the results of the models interpreted; specifically, is insight about the system behaviour drawn from the resulting model?
- Step 5: Are the models deployed and continuously updated?

The results from the evaluation are shown in Table 1.5. Cells marked with an 'X' indicate that the specific process step was followed, while those highlighted in red indicate that there was no evidence of that step being followed. The overall process followed in the reviewed references will allow for the development of a forecasting process that can be applied to deep-level mine electricity forecasting.

Table 1.5: Evaluating electricity forecasting models against the general modelling approach [46]

Reference	Step 1	Step 2	Step 3	Step 4	Step 5
Relevant question:	Are the goals of the model stated?	Is the data collected and manipulated?	Are models developed and tested?	Are the results of the models interpreted?	Are the models continuously updated?
[52]	X	X	X		
[53]	X		X	X	
[54]	X	X	X	X	
[55]	X	X	X	X	
[56]	X	X	X	X	
[57]	X	X	X	X	

Reference	Step 1	Step 2	Step 3	Step 4	Step 5
Relevant question:	Are the goals of the model stated?	Is the data collected and manipulated?	Are models developed and tested?	Are the results of the models interpreted?	Are the models continuously updated?
[58]	X	X	X	X	
[59]	X	X	X	X	
[60]	X	X	X		
[61]	X	X	X	X	
[62]	X	X	X	X	
[63]	X	X	X		
[64]	X	X	X		
[65]	X	X	X	X	
[66]	X	X	X	X	
[67]	X		X		
[68]	X	X	X		
[69]	X		X	X	
[70]	X	X	X	X	
[71]	X	X	X		
[72]	X	X	X	X	
[73]	X	X	X		
[74]	X	X	X	X	
Total (23)	23	20	23	15	0

Table 1.5 illustrates that the majority of the studies that were reviewed followed all of steps in the general modelling process, with the exception of step 5. Step 5, however, is related to the monitoring of models that are operating on real-time data [46] while most of the studies focused on model development with no mention of model implementation.

It must be noted, however, that Step 5 is a crucial step in ensuring that forecast models remain accurate and representative of reality. This is due to the dynamic nature of mining systems. This study, however, will focus on the development of the modelling methodology. As such Step 5 will not be discussed further as it is considered a methodology in itself.

The specific processes followed in each of the steps mentioned in Table 1.5 will be discussed in detail. This will allow for the development of a modelling process that is applied to electricity usage forecasting in general. This process will then be applied to platinum mine data and the results compared with those obtained using the existing method. This will allow for the possible improvement of the accuracy of electricity forecasting on deep-level platinum mines.

Definition of model purpose

The definition of the model purpose is considered an important step in the modelling and forecasting process. It ensures that the model is developed with a clear objective in mind and is designed to meet that objective [46]. For deep-level platinum mines, the goal is to develop a model that can accurately forecast electricity usage based on specific input variables; the current input variable is production. The goal of these forecasts is for monthly budgeting, planning and energy management.

The studies considered were mostly performed for energy management or design purposes. As such, a discussion on the model purpose will be required in order to contextualise the decisions made during the modelling process. From the 23 references considered, the following three purposes were identified, with some references creating the models for more than one of the purposes listed below:

1. Budgeting
2. Energy management
3. Design

17 of the evaluated studies designed the model for *energy management* applications or *budgeting*. Out of the 23 studies, 8 designed the model for *design* purposes, specifically to building design. The key difference between these model functions is that models built for energy management or budgeting purposes typically modelled the electricity demand as a function of time. This is in contrast with models built for design purposes, which typically only took the physical attributes into account for the modelling of electricity usage.

Studies that created models for the purposes mentioned - budgeting, energy management and design - are all considered relevant. This is due to the fact that the modelling process does not differ significantly between the studies. The purpose simply determines the finer details chosen during the modelling process. Thus, a mixture of studies that created models for different functions allows the evaluation of a variety of techniques and methods.

These techniques and methods are determined by the context and outcome of the model. The model context also determines parameters such as the choice of variables, the modelling resolution and the types of models, which will be discussed in detail below. Therefore, the definition of the model purpose is a critical step in contextualising the decisions made during the modelling process and must be kept in mind throughout the entire process.

Data acquisition and cleaning

The importance of accurate and representative data for regression modelling and forecasting cannot be understated [46]. It is vital for the development of accurate forecasts that the data used is representative of the system being forecasted, whether that is a shaft sub-system, a shaft or a total mine. Furthermore, it is important for the data to be validated and cleaned prior to the modelling process [75].

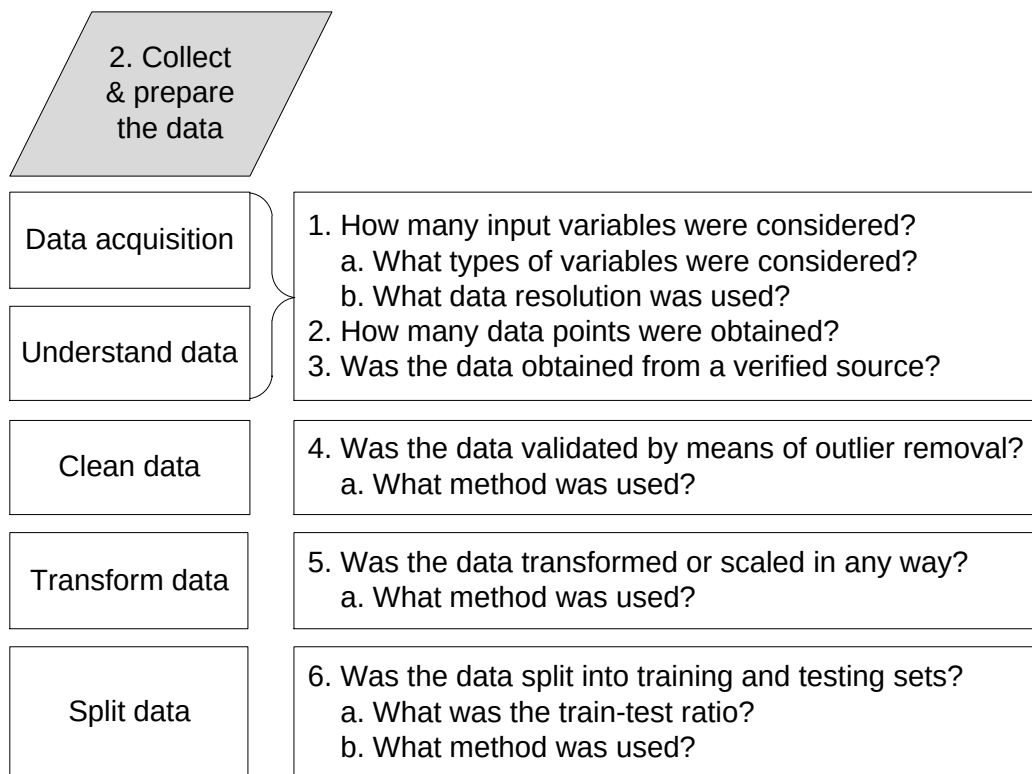


Figure 1.10: Data acquisition and cleaning step in overall modelling process

As illustrated in Figure 1.10, 5 sub-steps are present in this main step of the process. These are: 1) acquire the data, 2) understand the data, 3) clean the data, 4) transform the data and 5) split the data into training and testing sets. These steps will be evaluated by considering the 6 questions pertaining to the acquisition, validation, cleaning, and preparation of data.

Figure 1.10 illustrates that the first 2 sub-steps, *data acquisition* and *understand data*, can be connected as these are related processes. These 2 sub-steps are summarised by evaluating the input variables, the resolution and amount of data, and finally by ensuring that the data is obtained from a verified source. The remaining sub-steps, cleaning, transformation and splitting of the data are each satisfied by evaluating whether that step was performed and identifying what method(s) was/were used.

Data acquisition and understanding data

Questions 1, 2 and 3 in Figure 1.11 relate to the gathering and understanding of data, specifically the number and types of variables as well as the number of data points and the resolution used for each of the studies. Figure 1.11 illustrates the number of variables considered in each of the studies. The red line indicates that only one variable was used by that study. Furthermore, those studies that only used a single variable are highlighted by the orange pattern shading.

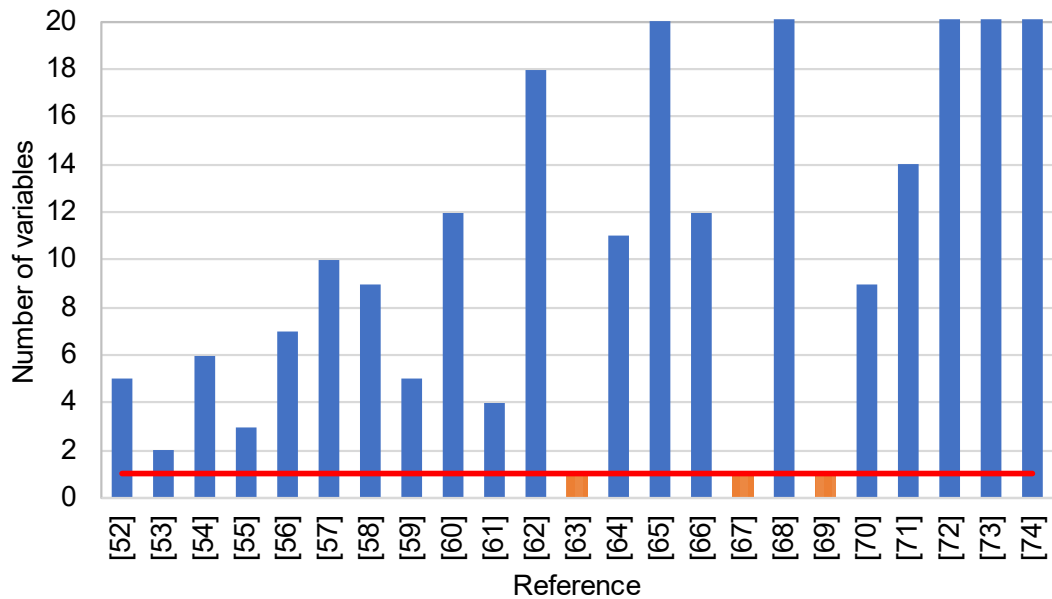


Figure 1.11: Summary of the number of variables considered by each evaluated study

Figure 1.11 indicates that 20 of the studies included more than 1 input variable in model development. The number of input variables used for model development is important as these systems are complex and are influenced by a large number of parameters, much like platinum mines. Including multiple input variables accounts for associations between these variables and the output variable and accounts for more variability in the output variable [76].

Furthermore, the types of input variables considered are also important for model development. This is because different types of variables take different system inputs into account. The input variables have been separated into system/operational variables, time-based variables, and seasonal variables, each of which consider different aspects that could affect the electricity usage of the system.

System variables refer to those inherent in the system, for example, the number of people living in a house or the monthly production of a platinum mine shaft. Time-based variables refer to hours of the day, days of the week or seasons of the year. Finally, seasonal variables

refer to weather-related data such as the ambient temperature or the air humidity. The variables considered by the different studies are shown in Table 1.6.

Table 1.6: Summary of the types of variables considered by the evaluated studies

Variable type	System	Time-based	Seasonal
Number of references	20	9	18

Table 1.6 illustrates that the most commonly used type of variables is system variables. This is understandable as these variables are inherent to the system, and therefore should describe the system electricity usage the best. This was followed by seasonal, and then time-based variables. It must be noted that only 8 of the studies considered all of the types of variables in their modelling processes.

These were also the studies that were capable of modelling electricity usage on an hourly or daily basis within reasonable accuracies. Therefore, the electricity usage forecasting on platinum mines must consider as many classes or number of variables as possible. This is, however, dependent on the degree of metering available at the facility which will determine the amount and type of data available.

The availability of this data and degree of metering must also be evaluated as this will determine the resolution of the data that can be acquired. The resolution of the data will ultimately dictate the resolution of the resulting forecast. The details pertaining to the importance and choice of resolution were discussed in section 1.2.1. The resolution used in each of the sources was evaluated and the results are shown Table 1.7.

Table 1.7: Summary of the data resolution considered in each of the evaluated studies

Resolution				
Hourly	Daily	Weekly	Monthly	Yearly
6	4	1	2	1

Table 1.7 illustrates the variability in data resolution considered by the studies. This variability in resolution is largely related to the outcomes of each of the studies, the model purpose as well as the degree of metering available. The selected resolution will determine the number of data points that will be considered in the modelling process.

Table 1.7 shows that an hourly time resolution was the most favoured by the evaluated studies. This is due to the fact that these studies aimed to predict electricity usage for management purposes or for load forecasting. The daily time resolution was the second most favoured

resolution for the same reason as for the hourly resolution. Predicting the electricity usage on such a granular time scale allows for daily management and reporting.

It is clear from Table 1.7 that the less granular time scales, particularly weekly, monthly and yearly, were not very favourable. Weekly and monthly predictions were also selected by studies that were modelling for energy management; however, it is clear that the more granular the time scale, the better it was for this purpose. Yearly modelling was used for system design and as such a granular scale was not necessary.

The remainder of the studies, 7, did not use any specific time resolution for modelling. These were the studies that focussed on modelling for design purposes. This involves the modelling of electricity usage parameters related to the system instead of the time-based electricity usage itself. An example of this would be the modelling of the heat loss coefficient of a building based on its physical parameters using existing building parameters. Although not directly applicable to forecasting, the methodology can be adapted for forecasting application and as such these studies were still considered.

The chosen resolution will have an effect on the number of datapoints available for the modelling process. The number of data points considered by each of the studies is shown in Figure 1.12. The red circles indicate studies where the number of data points was not explicitly stated.

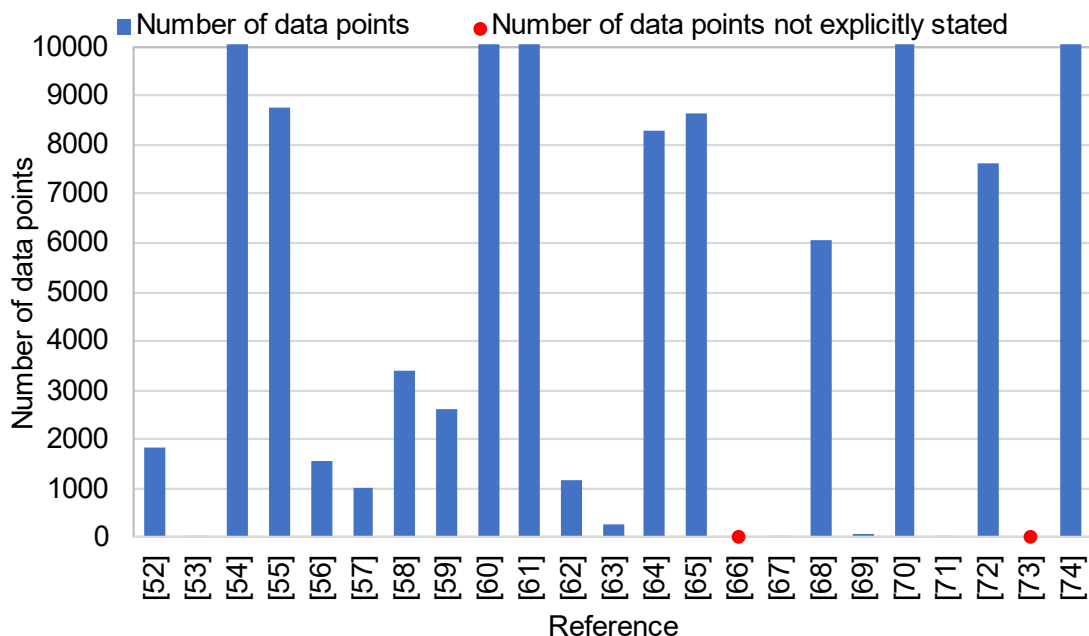


Figure 1.12: Summary of the number of data points considered by each evaluated study

It can be seen in Figure 1.12 that 5 sources were found to use over 10 000 data points and 11 between 1 000 and 10 000. 4 of the sources were found to use less than 100 data points,

specifically [55],[63],[67],[72]. It must be noted that 2 of the sources did not explicitly state the number of data points used, and these are indicated by red circles in Figure 1.12. Furthermore, the studies that considered fewer than 100 data points were also the ones that considered fewer input variables, particularly [63],[67],[72].

The number of data points is an important consideration for regression modelling. This is because the sample size determines whether a model's outputs can be considered to be statistically significant [77]. Several rules of thumb exist that are used to assist with the determination of a statistically significant sample size for a multiple regression analysis [77].

Two of these methods are based on the number of input variables that will be used for the analysis. The first was proposed by Green [77],[78] and the second was proposed by Harris [79]. These are shown in Equation (1.4) and Equation (1.5) respectively.

$$N \geq 50 + 8i \quad (1.4)$$

$$N \geq 50 + i, i \leq 5 \quad (1.5)$$

where:

N	Number of datapoints
i	Number of input variables

Equation (1.4) and Equation (1.5) indicate that the number of data points should be equal to a minimum of 50, plus a specified multiple of the number of input variables, that multiple being equal to 8 for Green's method and 1 for Harris's method. While Green's method is applicable for any number of input variables, Harris's method only applies to analyses having fewer than 6 input variables.

Although other rules of thumb exist, these are related to classification tasks, and as such, will not be discussed further. Therefore, due to their simple and representative application to regression tasks, a combination of Harris's and Green's methods will be used to determine the appropriate sample size. These relationships were applied where applicable to each of the reviewed studies in order to determine if they satisfied these rules of thumb.

It was found that 18 of the studies satisfy the rules of thumb required for the sample to be considered statistically significant. The detailed results can be found in Appendix A. Once the

minimum sample size has been determined, it is important to acquire and clean the data required for the analysis. These are addressed in questions 3 and 4 in Figure 1.10.

Authentication and evaluation are the processes followed to ensure that the data is obtained from a reliable source, cleaned to remove inconsistencies, and checked for quality. Authentication refers to the process of checking the data for accuracy which is done to ensure that the results of the study are reliable and consistent [80],[81].

This includes ensuring that the data is obtained from a reliable source and checking the dataset for overall completeness. Evaluation of the data involves the cleaning of the data to ensure that its quantitative elements are consistent. This includes the datatypes, the range of values as well as outlier removal which will be discussed in the next section.

Question 3, in Figure 1.10, pertains to the authentication of the data, specifically to the verification of the source of the data. 18 of the total reviewed studies ensured that the data was obtained from an official source or consolidated against a verified source to ensure its reliability. This step is important to ensure that the acquired data can be confidently used for the development of models with sufficient accuracy.

Clean data

Once the source of the data has been authenticated and the data has been obtained, the data must be evaluated to ensure that it conforms to a set of prescribed rules. An important process to ensure that data conforms to these rules is the removal of outliers [82]. Outliers are datapoints that deviate significantly from the rest of the data points, falling outside the expected trend of the data [83].

Outliers can cause significant degradation to the performance of a model, and as a result must be removed from the dataset [72]. 10 of the studies explicitly mention the removal of outliers, with the detailed results shown in Appendix A. It must be noted, however, that 3 of the studies used simulation data as their modelling input data, and as such would not require data cleaning as the data was computer generated.

Furthermore, no consistent mathematical method was found amongst the studies that performed outlier removal. All of the studies that removed outliers, with the exception of two, applied intuition, and the removal of values that were not considered relevant was performed. These were typically values that did not follow the overall trend of the data.

Although intuition is an adequate method for outlier removal, it creates the risk of introducing bias into the data [84]. In order to ensure that a more consistent method for outlier removal is

considered for application to the forecasting methodology, statistical approaches of outlier detection will be discussed [83]. The methods that will be considered are the boxplot method for univariate outliers and the Mahalanobis distance statistic for multivariate outliers [83],[85].

The boxplot method involves the utilisation of a 5-number summary. This summary consists of the smallest non-outlier number (the min), the lower quartile (Q1), the median (Q2), the upper quartile (Q3) and the largest non-outlier number (the max). These numbers are used to fully summarise the statistical distribution dataset and outliers are defined by the rules illustrated by Equation (1.6), Equation (1.7) and Equation (1.8).

$$\text{Interquartile Range (IQR)} = Q_3 - Q_1 \quad (1.6)$$

$$\text{Outlier: } x > Q_3 + 1.5 \times \text{IQR} \quad (1.7)$$

$$\text{Outlier: } x < Q_1 - 1.5 \times \text{IQR} \quad (1.8)$$

Equations (1.6), (1.7) and (1.8) illustrate that outliers are defined as values, x that fall above or below the upper quartile or lower quartile by 1.5 times the interquartile range respectively. This is due to the fact that this region should contain 99.3% of objects in a normally distributed dataset [83]. This method will be applied in the event that the forecasting models are built with the use of a single input variable.

The Mahalanobis distance can be used to identify multivariate outliers under the assumption of a normal distribution [83] and is calculated with Equations (1.9) and (1.10) [86],[87].

$$MD(x_i, \bar{x}) = \sqrt{(x_i - \bar{x})C^{-1}(x_i - \bar{x})^T} \quad (1.9)$$

$$C = \frac{1}{n-1} (X - \bar{X})^T (X - \bar{X}) \quad (1.10)$$

where:

$MD(x_i, \bar{x})$	The Mahalanobis distance
x_i	The value of a single object in the dataset
$\bar{x}$	The mean vector of the dataset
C	The covariance matrix of the dataset
X	The total data matrix
n	The number of data points in the data set

The Mahalanobis distance for each datapoint in the set can be calculated with the use of Equation (1.9). Equation (1.10) is used to calculate the covariance matrix of the variables to be used in Equation (1.9). This reduces the dataset to a univariate distance dataset, and as such, the boxplot method can be applied on the Mahalanobis distances of each of the points [83]. If the Mahalanobis distance of a point is considered to be an outlier relative to the distance of the other points, then the point is considered to be an outlier as well.

The application of a traceable method of outlier detection is necessary in order to ensure that outlier removal is consistent and methodical. This is to ensure that a valid statistical method is applied to all data for the prevention of bias introduction and consistent data cleaning.

Transform data

Once the dataset has been authenticated and evaluated it can be transformed or scaled before being used in the modelling process. Data scaling is regarded as an important step in data pre-processing [88]. It is the process of ensuring that all input variables are equally weighted for scenarios where multiple input variables are considered [83]. Scaling does, however, introduce difficulty into the interpretation of the resulting model.

This is because the effect of model coefficients are less intuitive to interpret even if they are rescaled back to the original units [89]. For example, if the regression of electricity consumption and production yields a slope of 5 kWh/tonne, this implies that the electricity usage increases

by 5 kWh for every tonne extracted. With scaling, this intuition falls away as the slope can yield 0.01 scaled kWh/scaled tonnes – which is not interpretable.

Scaling attempts to give all input variables equal weights in the resulting equation by making the variables fall into similar ranges, usually between -1 and 1 or 0 and 1 [83],[90]. Only 6 of the evaluated studies explicitly mention that scaling was performed, with detailed results shown in Appendix A. Scaling was only performed in studies with multiple input variables with two methods being more commonly used.

The two common scaling methods applied in the evaluated studies include min-max scaling and z-score scaling. The aim of min-max scaling is to rescale the data to exist within a new range of values; this is typically chosen to be between 0 and 1 and is achieved with Equation (1.11).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}(b - a) + a \quad (1.11)$$

where:

x'	The scaled value of the datapoint
x	The original value of the datapoint
a	The new minimum value
b	The new maximum value

Equation (1.11) can be used to scale the data points to exist within any range prior to the modelling step in the process. This range is given by values a and b and the new scaled values are calculated by first subtracting the minimum value from the data point and dividing it by the range. The normalised datapoint is then scaled to the new range by multiplying it by the newly chosen range and adding the chosen minimum to that point. Min-max scaling performs a linear transformation on the data and does not change the distribution of the data [91].

Unlike with min-max scaling, the aim of z-score scaling is to scale the data points such that the data has a mean of 0 and a standard deviation of 1. This results in a change to the distribution of the dataset to a Gaussian distribution [91]. This is achieved with the use of Equation (1.12) [83].

$$x' = \frac{x - \bar{x}}{\sigma} \quad (1.12)$$

where:

x'	The scaled value of the datapoint
x	The original value of the datapoint
$\bar{x}$	The mean of the dataset
σ	The standard deviation of the dataset

Equation (1.12) indicates that each of the data points in the dataset is scaled by subtracting the mean of the data from the point and then dividing it by the standard deviation of the dataset. This results in a dataset with a mean of 0 and a unit standard deviation. It must be noted that the same parameters applied to scale the data which is used to create the forecasting model must be employed to scale the input data to the model when the model is used for forecasting [83].

Since z-score scaling alters the distribution of the dataset to be Gaussian, it should be used on datasets that already display Gaussian distributions [92]. Min-max scaling will therefore be used when the dataset does not display a Gaussian distribution or if the user is unsure of the distribution of the data [92].

Split data into training and testing sets

Once the data has been cleaned of outliers and scaled, it must be split into different sets that can be used for fitting (or training) and testing the resulting forecasting models. Splitting the data into different sets ensures the evaluation of the performance of each model on both 'seen' and 'unseen' data [46]. This allows the generalisability of each model to be evaluated and compared when multiple models are considered.

The generalisability of the resulting model refers to the model's capacity to predict on previously unseen or real-world data. A model that shows very little error on the training set but high error on the test set is said to be overfit to the training data [93]. Conversely, a model that shows high error on both the training and test sets is said to be underfit to the data. Specifically, it does not model the general trend of the data [93]. Examples of this are shown in Figure 1.13.

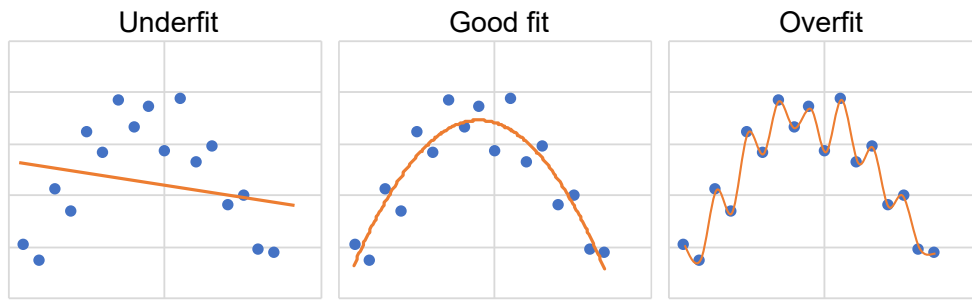


Figure 1.13: Examples of models that are underfit (left), fit well (centre) and overfit (right)

Figure 1.13 illustrates that an underfit model does not adequately follow the general trend of the data. This is usually a result of a model being overly simple [93]. Furthermore, it shows that a model that fits the data well is one that follows the trend without tracking the specific training data points. Finally, Figure 1.13 shows a model that overfits the data is one that follows the specific data points and does not generalise well to the trend of the data [93]. The use of training and testing sets allows the evaluation of the model's generalisability.

Question 6 in Figure 1.10 addresses whether the dataset was split before training the model, what train-test split ratio was used and if a specific method was used to split the data. Each of the studies were evaluated based on these questions and the detailed results are shown in Figure 1.14 and Figure 1.15.

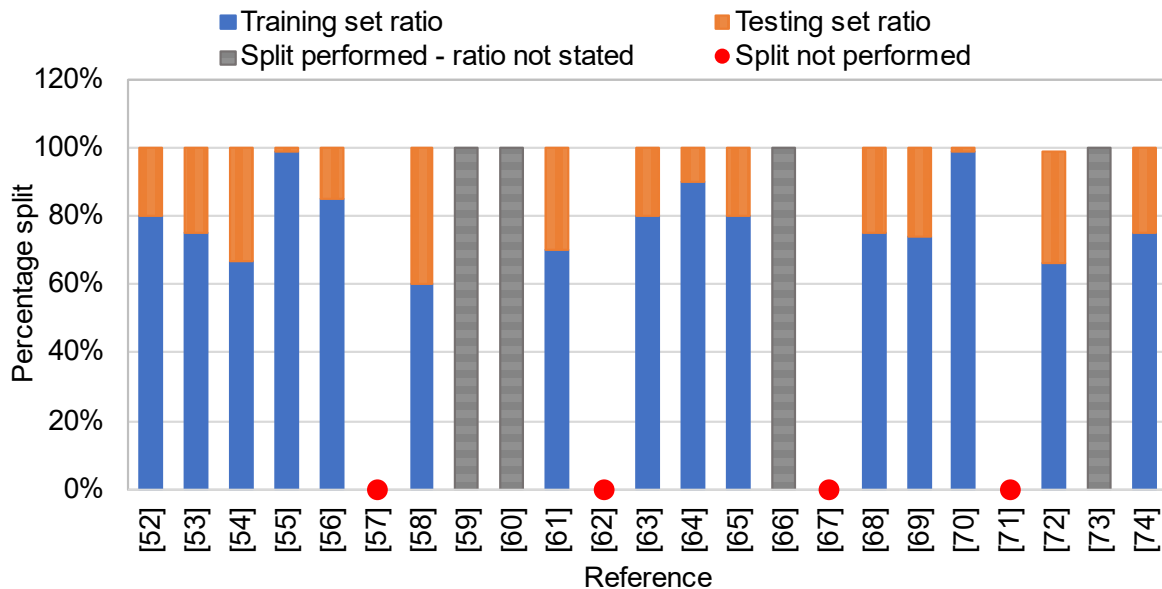


Figure 1.14: Summary of train-test split ratio used in evaluated studies

From Figure 1.14, it is clear that 19 of the 23 studies explicitly mentioned that they split the data before fitting the model. Of those studies, 3 used simulation to generate their training data and tested the models on real buildings [55],[56],[70]. This led to weighted train-test split ratios

between 99:1 and 85:15. The remaining 16 studies had train-test split ratios ranging between 90:10 and 60:40, for those that explicitly stated the ratio used.

Furthermore, as indicated by the red circles in Figure 1.14, 4 of the studies did not perform any train-test split on the dataset. Additionally, 4 of the studies that did perform train-test split did not explicitly mention the ratio used for the split – these are indicated by the bar with the grey horizontal line shading.

Clearly a significant variation in train-test split ratio exists between the various studies. This variation can have a significant impact on the performance of resulting models [94]. The degree of this variation is also a result of the amount of data available, the models used and the context of the model [95]. Specifically, the train-test split ratio is a situational decision with some sources indicating that a good rule of thumb is a 70:30 split [95].

The literature results were compared with the obtained rule of thumb by means of an average. The average of the percentage splits of the 16 references that did not use simulation data as training data was taken. This yielded an average train-test split ratio of 75:25, which is very similar to the acquired rule of thumb.

The various studies also made use of different methods for splitting the data. These include cross-validation, random 2-way holdout split, using the most recent data as the test set and using simulated data as training points and testing on real-world systems. The summary of the number of studies that used the different methods of splitting the data is shown in Figure 1.15.

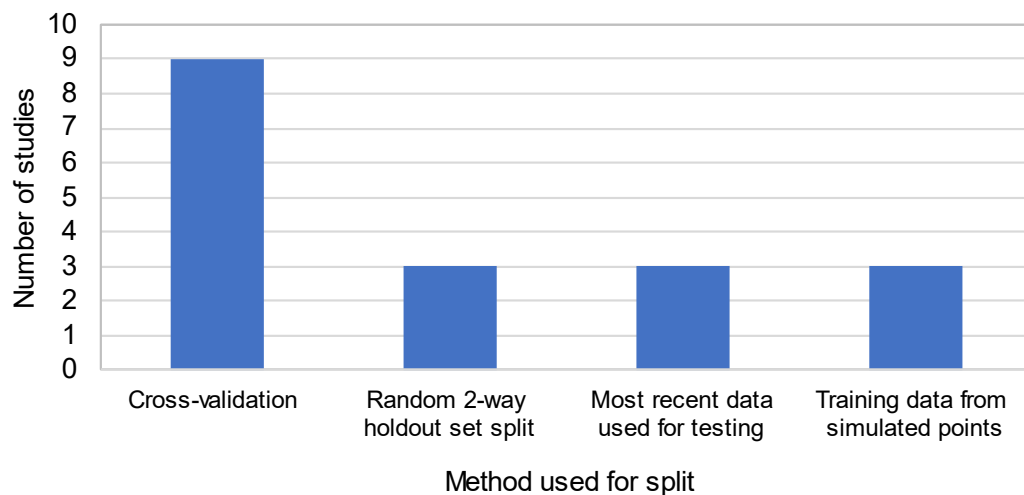


Figure 1.15: Summary of the train-test split methods used by the evaluated studies

Figure 1.15 indicates the 4 different methods used for splitting the data. It must be noted that all the methods except cross-validation are variations of a 2-way holdout split [96]. This is the process where a single training and testing set is created. This is different to cross-validation,

which is a more thorough method, taking bias into account. The most common method identified in the evaluated studies was cross-validation which was used by 9 of the studies.

A 2-way holdout split involves splitting the data into a single set of training and testing data on which to fit and evaluate the model. This is easily achieved by random subsampling [96], as seen in 3 of the studies in Figure 1.15. The risk of using a single set of training and testing data is that either set might not fully represent the whole dataset [96].

This is dependent on the efficacy of the random subsampling or the distribution of the most recent data as used in 3 of the studies. A non-representative distribution in either the training or the testing set will result in bias in model evaluation. This can lead to incorrect conclusions being drawn and sub-optimal models being selected. A method to ensure that misrepresentation does not occur is to use cross-validation.

Cross-validation, specifically k-fold cross-validation, is the process of repeating the 2-way holdout split a known number of times. The average performance for each random assortment of train-test splits can then be calculated to evaluate model performance [96],[97]. This allows a more robust assessment of the model's performance by ensuring that different distributions do not introduce bias into performance evaluation. It also allows the assessment of the model's stability, as the standard deviation of performance across different splits can also be calculated [96].

Therefore, a variety of data-splitting methods exist. These range from using a single holdout set and selecting different data for that set, whether randomised or the most recent data, or using k-fold cross-validation. These can be coupled with a variety of split ratios, which must be decided upon depending on the amount of data available. The splitting of data, regardless of the method, is important to ensure adequate performance evaluation on unbiased/unseen data.

Model development

Once the data has been obtained, cleaned and split, the next step in the overall process is model development [46]. Model development is the process of using the cleaned data to produce models capable of forecasting the output variable, in this case electricity usage, within a reasonable amount of accuracy. The chosen sources were evaluated based on a set of questions related to the sub-steps of step 3 of the overall process shown in Figure 1.9. The detailed questions and how they relate to the sub-steps are shown in Figure 1.16.

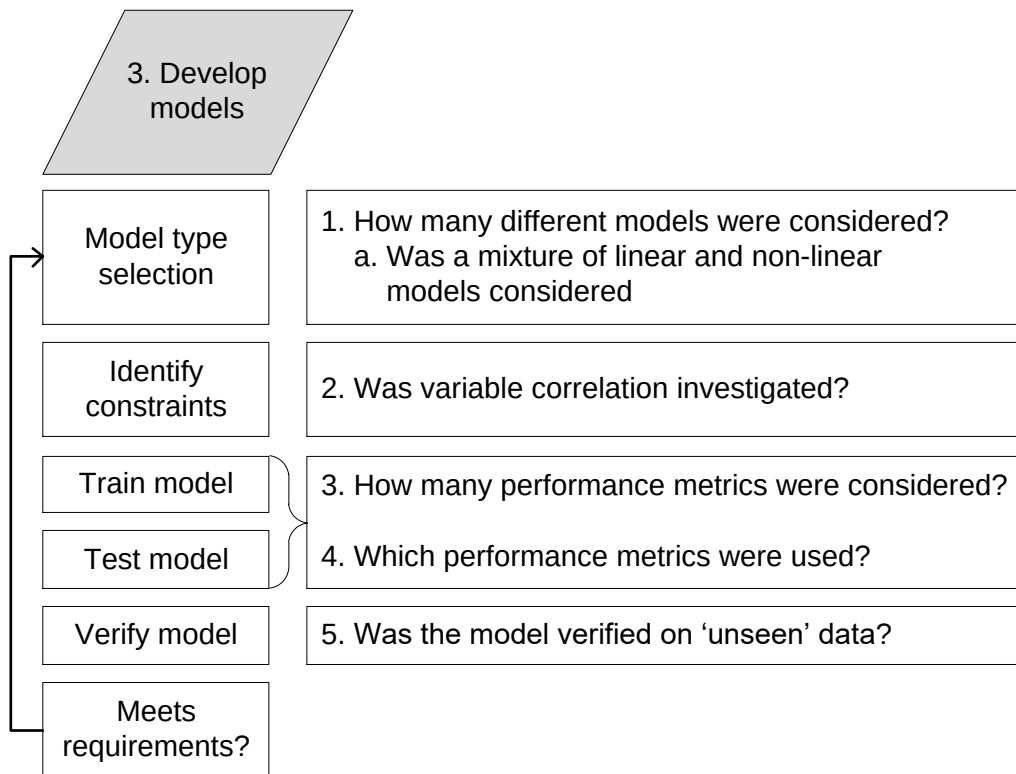


Figure 1.16: Model development step in overall modelling process

As illustrated in Figure 1.16, 4 sub-steps are present in this step of the process. These are: 1) model type selection, 2) identify constraints, 3) train and test the model and 4) validate the model. Model development is an iterative process, as illustrated in Figure 1.16, concluding when the resulting model meets predetermined requirements. These sub-steps will be evaluated by considering the 5 questions shown in Figure 1.16.

Figure 1.16 illustrates that the first sub-step, model type selection, is summarised by evaluating how many different models were considered by each of the studies as well as the complexity of those models. Furthermore, the second sub-step, identify constraints, pertains to the evaluation of correlation between selected input variables and the output variable.

As shown in Figure 1.16, the training and testing sub-steps have been combined, as these are related to the creation of the models. These 2 sub-steps will be evaluated by investigating the number and types of performance metrics used for model testing. Finally, model verification will be investigated to determine whether the trained and tested model was verified on a final unseen dataset to simulate 'real-world' performance and to evaluate whether the model is producing the expected results.

Similar to the data acquisition and cleaning steps of the process, the identification of the steps followed in the studies will assist with the creation of this step in the resulting forecasting strategy.

This will allow the theoretical investigation of the modelling process and the justification of why certain decisions were made in the evaluated studies.

Model type selection

At the end of the data gathering process, a set of authenticated, cleaned data will be available for model development. This data will be used to build and test a regression model that must achieve the model outcome. The first step in the model development process is to select the type of regression model that will be fit to the data.

Model type selection is important as different models will be able to process the data differently, and as such will fit the data differently. The result of these differences is that the performance of various models will differ [98]. This can be seen in the leftmost graph of Figure 1.13, where a linear function is fit to data that follows a quadratic trend. The linear model is not complex enough to fit the trend and will exhibit poor performance compared with the graph in the centre of Figure 1.13.

Since models are expected to have different levels of performance, it is common practice to fit a variety of models to the data and use the train-test split accuracies to select the best model [99], [100]. This allows for model comparison and ensures the selection of the model that produces the best accuracy. The number of models considered by the various studies is shown in Figure 1.17.

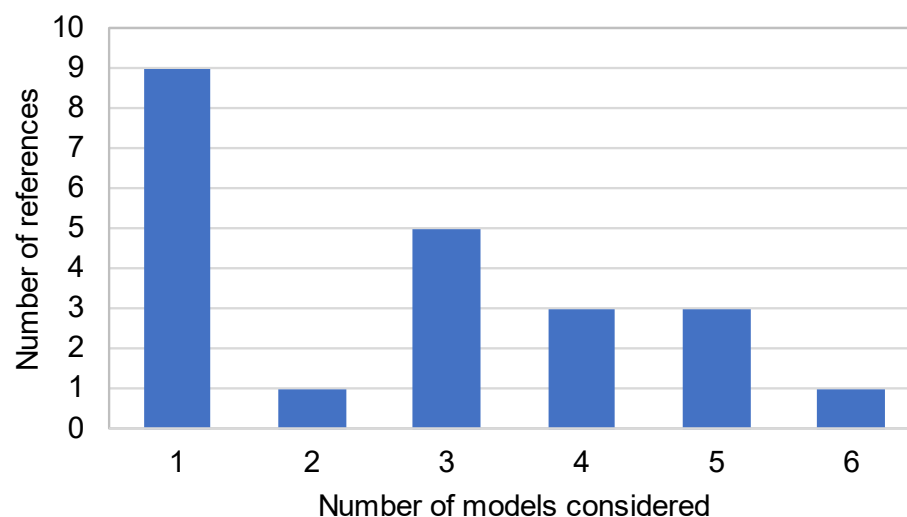


Figure 1.17: Summary of the number of models considered by the evaluated references

Figure 1.17 illustrates that 9 of the 23 evaluated studies considered only a single model. The remaining 14 studies considered at least 2 models, with 3 studies considering 5 models and one of the studies considering 13 models [56]. The detailed data is shown in Appendix A. In

these 14 studies, the performance of the different models was compared, allowing for the selection of the best model for each specific application.

The comparison of multiple models allows for the selection of the model with the best performance on the training and testing data. This is beneficial as it is not possible to know how different models will perform on your specific data before you have fit and tested them. As such, the consideration of multiple models is the most thorough method to ensure that you are modelling the data in the most accurate way.

Due to the consideration of multiple models, it becomes possible to include models of different levels of complexity. Furthermore, it also becomes possible to consider models of a linear and a non-linear nature. This allows the consideration of a large variety of different options, allowing for the selection of a model that generalises well to the data and allows for accurate forecasting [100].

For simplicity, these models have been separated into 2 groups for evaluation of the studies: simple models, which will include physical models and ordinary least squares regression, whether linear or polynomial, and advance models. Advanced models include those that are considered to be machine learning models. These include but are not limited to decision trees, random forests, neural networks and time-series forecasting. A summary of the types of models used by the evaluated studies is shown in Figure 1.18.

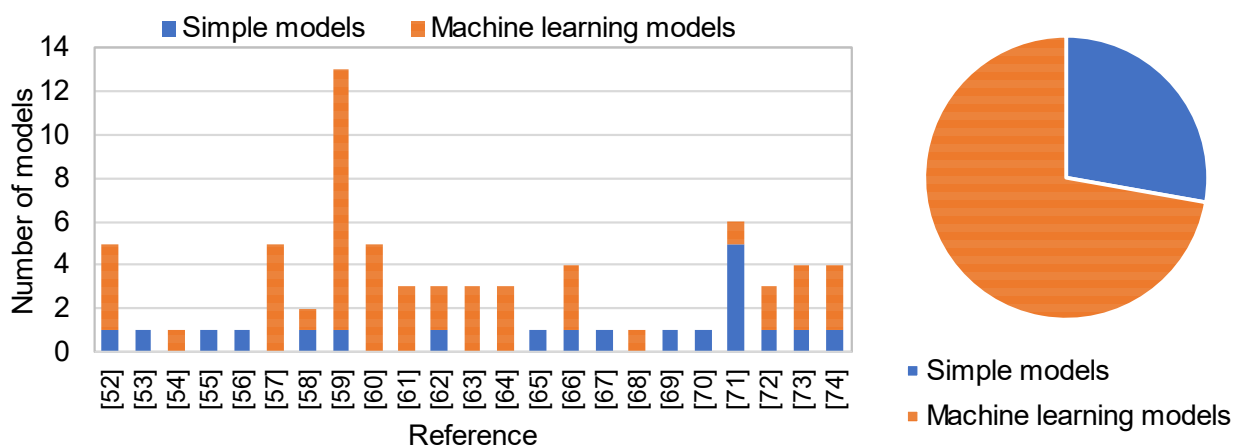


Figure 1.18: Model complexity of the models considered by the evaluated studies

Figure 1.18 shows that 9 of the 14 studies that considered more than 1 model used a mixture of simple models and more complex machine learning models. The simpler models included multiple linear regression, polynomial regression as well as physical system models. The pie chart on the right of Figure 1.18 also indicates that over a quarter of all models used in the evaluated studies were simple models.

In 4 of the 9 studies, ordinary least squares regression was found to produce accuracies similar to those of the more complex models [59],[62],[66],[71], with the simpler models outperforming the complex models in 1 of the studies [66]. Since the application of ordinary least squares regression will already be an increase in model complexity compared with the current intensity model, this method will be discussed below.

It must be noted that more complex machine learning models should be considered if applicable as these can offer significant improvements in performance and accuracy. This is dependent on the specific system and data being forecasted. The model type selection can further be divided by considering the linearity of the models that were considered by the evaluated studies. This is shown in Figure 1.19.

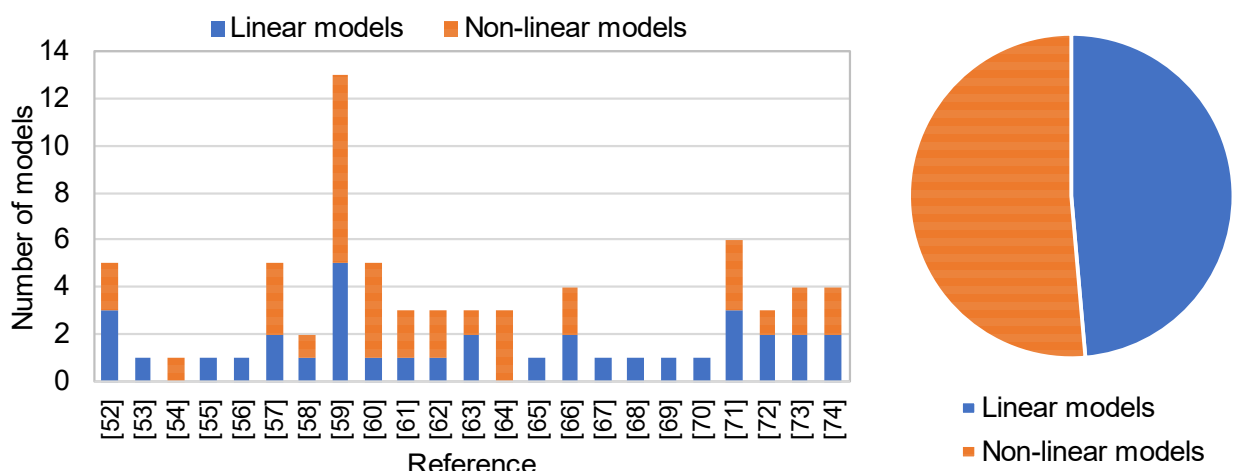


Figure 1.19: Model linearity of the models considered by the evaluated studies

Figure 1.19 illustrates that 13 of the 14 studies that considered multiple models used a mixture of linear and non-linear models. Furthermore, the pie chart on the right of the figure illustrates that almost half of the models considered by all of the studies were non-linear in nature. The exact model types considered by the studies can be seen in Appendix A. These ranged from ordinary least squares regression to random forests and artificial neural networks.

The consideration of models that are non-linear takes into account the possibility of non-linear relationships between the variables. This can be achieved by fitting a variety of different functions using ordinary least squares regression. More complex non-linear functions should, however, also be included as these can offer improvements in model performance.

As discussed in section 1.2.1, the current modelling method in the mining industry involves the multiplication of production by an intensity to forecast electricity usage. The consideration of least squares regression with the application of both linear and non-linear functions will be an improvement on the existing model complexity. This compared with the intensity method will

show the benefit of using more complex functions. Furthermore, it will suffice to illustrate the importance of considering linear and non-linear functions.

Ordinary least squares regression is the process of fitting one or more explanatory variables to a continuous outcome variable by minimizing the sum of square errors [101]. The error is calculated as the difference between the actual and the predicted value of the outcome variable. This relationship can be linear and non-linear depending on the function that is fit to the data [102]. Equation (1.13) and Equation (1.14) illustrate the relationship used for the fitting of a multivariate linear model to the data [102].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (1.13)$$

where:

y	Output/dependent variable
β_i	Regression coefficients
x_i	Input/independent variables
ϵ	Residual error

In Equation (1.13), a linear relationship between the output and input variables is modelled with the use of the regression coefficients. The values of these coefficients are adjusted algorithmically in order to minimise the error between the actual output values and those estimated by the resultant model. The error is calculated with the use of Equation (1.14).

$$E = \sum_{n=1}^N (y_n - \hat{y}_n)^2 \quad (1.14)$$

where:

y	Output/dependent variable
$\hat{y}$	Modelled output/dependent variable
N	Number of data points
n	Data point in dataset

Equation (1.13) illustrates a linear relationship between the input variables and the output variable; this relationship can, however, be substituted with any functional relationship. This relationship can include any linear and non-linear combination of the dependent and independent variables [103]. These can include polynomial or power combinations of input variables.

Several assumptions are made in the process of fitting a linear model to variables. These assumptions must be tested in order to ensure that the resulting model error can be trusted to be representative [104]. These assumptions include 1) normal distribution of input variables, 2) the functional relationship between variables is correct (i.e. linear, power, polynomial relationship), 3) constant variance of the resulting errors (no heteroskedasticity), 4) the residuals are normally distributed [104],[105]. The testing of these assumptions is discussed below.

The normal distribution of the input variables would have been assessed in the data cleaning step of the process. This would specifically be checked and rectified during the '*transform data*' step, as it is important to know the data distribution when deciding which scaling method to use. Assessing the normality of the data can be done with the use of a histogram or box and whisker plot of the data and visually inspecting the distribution [106],[107], [108]. Examples of these plots are shown in Figure 1.20 and Figure 1.21.

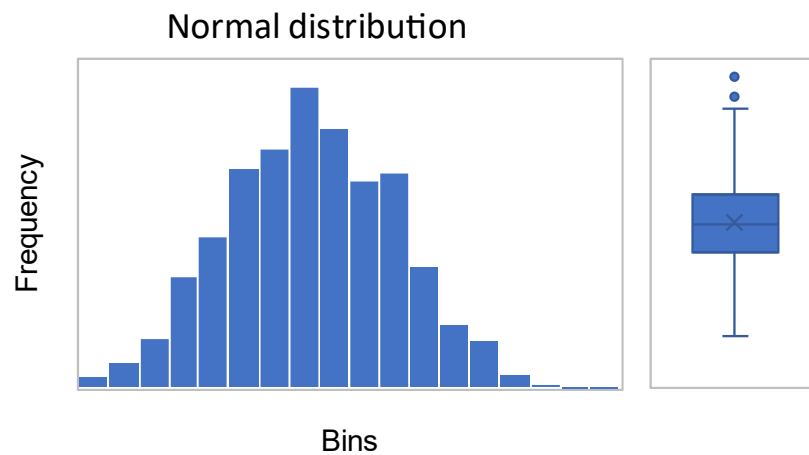


Figure 1.20: Example of the histogram and boxplot for a normal distribution

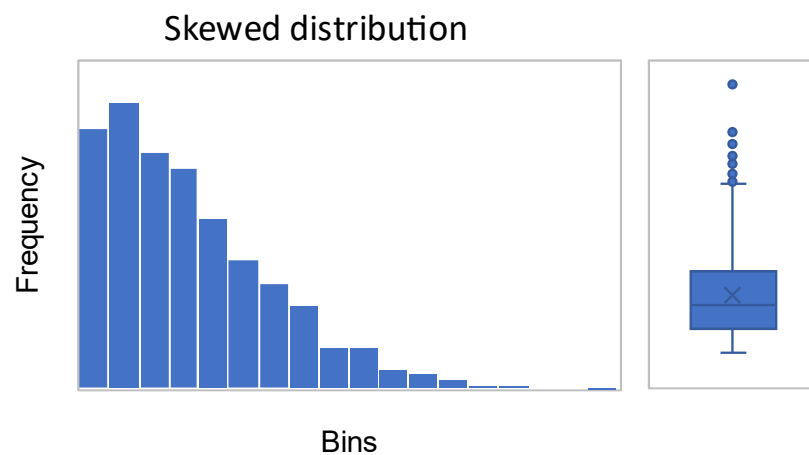


Figure 1.21: Example of the histogram and boxplot for a skewed distribution

Figure 1.20 and Figure 1.21 illustrate the histograms and boxplots for a normal and skewed distribution respectively. Figure 1.20 shows that the histogram of a normal distribution has most of the points in the centre, with fewer points occurring as the data deviates from the mean, which is at the centre of the distribution. Furthermore, this distribution is symmetrical about the central point. This is confirmed by the boxplot in Figure 1.20, which has whiskers of equal lengths which illustrate the symmetry of the distribution.

The boxplot of Figure 1.20 also confirms that the data is symmetrical with a mean, illustrated by the blue cross in the centre of the boxplot, almost equal to the median, illustrated by the blue line in the centre of the box. Any data that does not have similar characteristics to that shown in Figure 1.20 is not considered to be normally distributed. Figure 1.21 is used as an example of data that is not normally distributed; however, this type data can have a variety of distributions.

Figure 1.21, in contrast with Figure 1.20, has a distribution that is skewed, specifically right-skewed. This means that the majority of the datapoints do not occur in the centre, but

towards the left. This type of data is also defined by a long right 'tail' illustrated by the data that extends far to the right of the histogram. Specifically, this data is not symmetrical about the mean, unlike with the normal distribution.

The boxplot in Figure 1.21 confirms these observations, with whiskers that are of different lengths. This illustrates the lack of symmetry and, particularly, the skewness in the data. Furthermore, the boxplot in Figure 1.21 illustrates that the data is skewed with a mean, illustrated by the blue cross, not equal to the median, illustrated by the blue line in the centre of the box.

Therefore, the utilisation of a histogram and boxplot of the data will be used to determine whether the data follows a normal distribution or not. This will allow the selection of the correct scaling technique and will also allow the evaluation of two of the model assumptions, namely, the assumption that the input data is normally distributed and the assumption that the residuals are normally distributed – which will be discussed in this section.

When selecting a relationship between the variables that will be fit using ordinary least squares regression, it is assumed that the selected relationship is correct. Several different functions will be considered in the proposed forecasting strategy in order to take non-linearity of the system into account. This assumption must thus be tested for the selected model to ensure that the selected relationship is true.

The testing of this assumption will be done by plotting a scatter plot of the output variable against each input variable to assess the functional relationship [105]. Furthermore, a scatter plot of the predicted output variable against the actual output variable will be created. This will allow the calculation of a correlation coefficient in order to assess the closeness of fit of the model to the data.

The constant variance of the resulting errors assumption can be tested by plotting a scatter plot between the residuals and the predicted output values. Residuals are simply defined as the difference between the actual and predicted values for each datapoint. An example of this plot is shown in Figure 1.22. For a model where the assumptions are correct this plot will show no systematic pattern between the residuals and the predicted values [109].

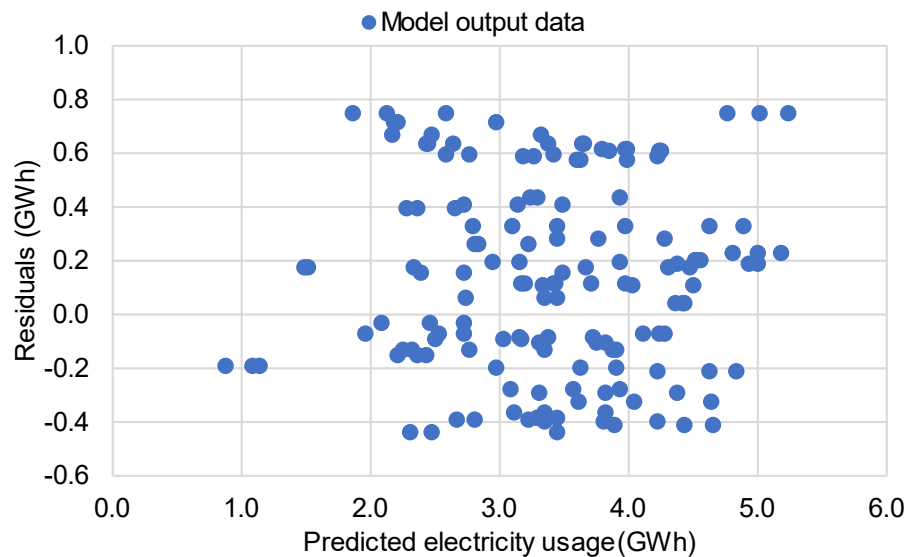


Figure 1.22: Example residual plot against predicted values to test heteroscedasticity assumption

As illustrated in Figure 1.22, the residual plot shows no systematic pattern. This is indicative of data that meets the heteroscedasticity assumption of least squares regression. If the plot showed curvature in the residuals, or inconsistent variance, this assumption would not be considered maintained [104],[109]. This would typically occur in datasets that have a large range between the largest and smallest observed values.

Finally, the final assumption of normally distributed residuals can be tested with a histogram of errors [103],[105]. This provides information that is otherwise lost when an average error is calculated to assess model performance. An example plot is shown in Figure 1.23. If the resulting plot is normally distributed, then the assumption of normality is satisfied.

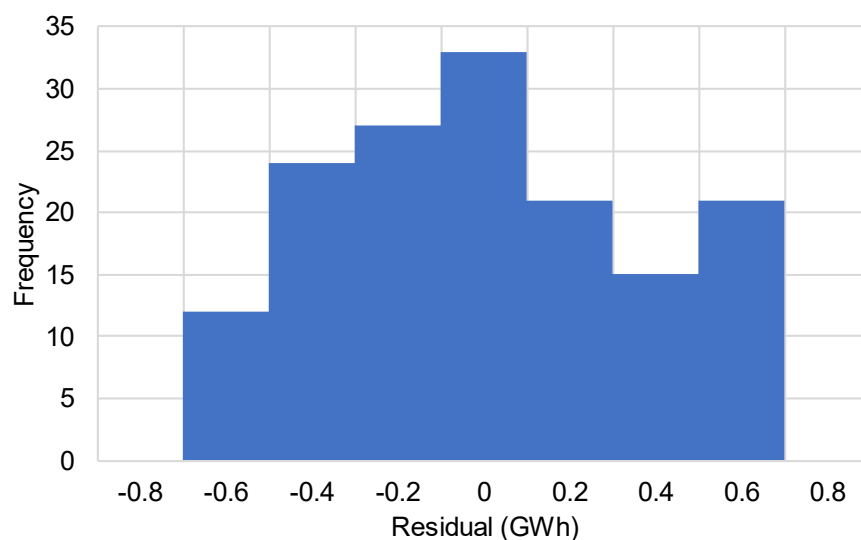


Figure 1.23: Example distribution plot of residuals

The above assumption validation checks will be considered in the model verification step of the process. This is due to the fact that they provide insights into the statistical performance of the resulting model. Furthermore, these assumptions must be validated to ensure that the resulting model is statistically significant and that the resulting model error is representative of performance.

Variable correlation

Since it is possible that a large variety of variables will be considered for the creation of these models, it becomes important to evaluate variable correlation. Correlation is the measure of an association between variables. In correlated data, the change in magnitude of 1 variable, the input variable, is associated with a change in the magnitude of another variable, the output variable [110]. When creating models for forecasting, it is advisable to select variables that have strong correlations to the selected output variable.

The selection of these variables is done by investigating the correlation of each input variable with the output variable. Those that have a weak correlation have no association with the output variable and can be removed. This will result in a simpler, more interpretable model with more predictive value [111]. The considered studies were evaluated to determine whether a correlation analysis was performed on the variables before model fitting. This was to identify whether this is common practice.

It was found that 15 of the 24 studies performed a correlation analysis on the variables. In the majority of the studies this was done with the use of a scatter plot between each input variable and the output variable to visualise the correlation. Pearson's correlation coefficient was also used in the studies to determine the statistical correlation between the variables. Pearson's correlation coefficient, denoted by r is calculated with Equation (1.15) [112],[113].

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1.15)$$

where:

r	Pearson's correlation coefficient
x_i	Input variable x for the i^{th} data point
$\bar{x}$	Mean of variable x
y_i	Output variable for the i^{th} data point
$\bar{y}$	Mean of variable y
n	Number of data points

Equation (1.15) illustrates that Pearson's correlation coefficient is calculated with four terms, the first term on the numerator being the sum of the differences between each data point of the input variable and the mean of the input variable. This term is multiplied by the sum of the differences between each data point of the output variable and the mean of the output variable. This product is divided by the square root of the sum of the squared differences of each variable.

Pearson's correlation coefficient gives an indication of the degree of linear correlation between 2 variables [110]. This parameter should only be used when the 2 variables being compared are normally distributed as it is severely affected by extreme values [112]. It is scaled and exists within a range of -1 and 1. Figure 1.24 shows examples of scatter plots between variables having different Pearson's correlation coefficients [110].

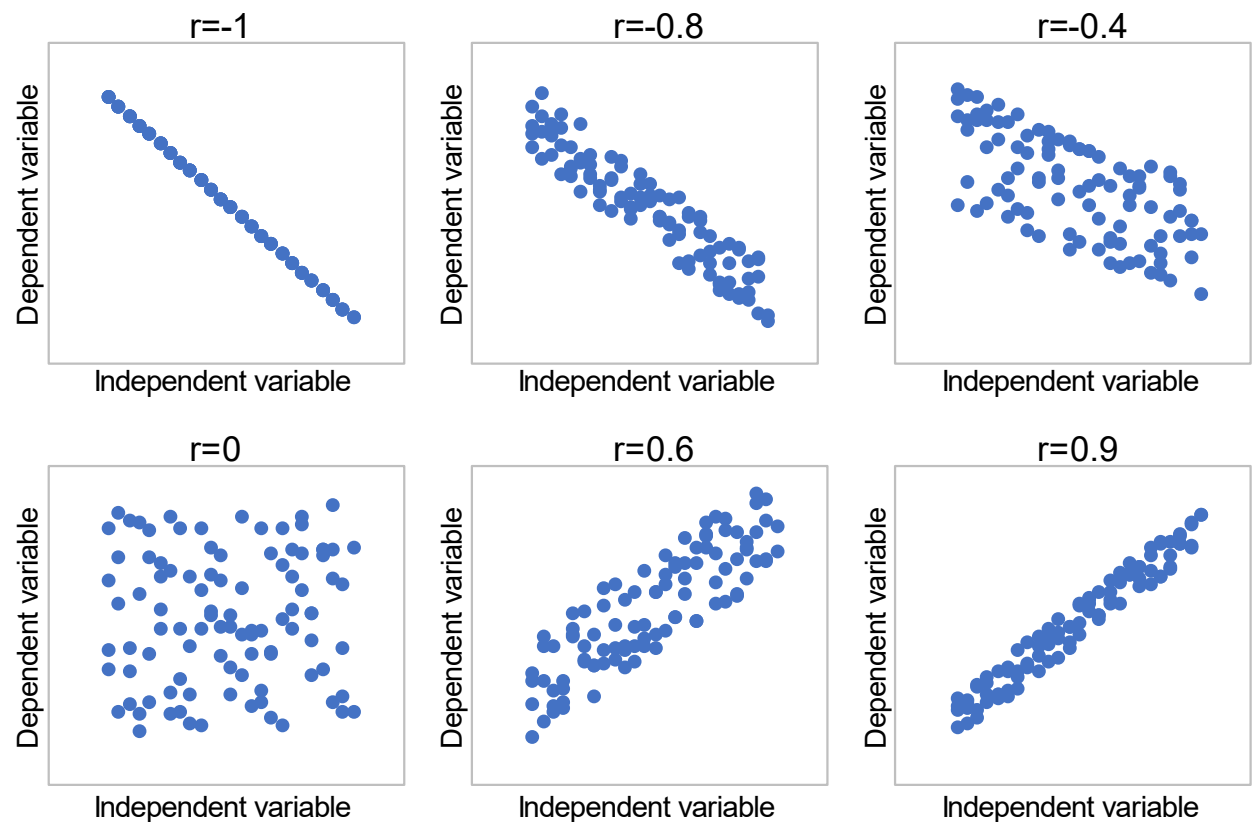


Figure 1.24: Examples of scatter plots between variables and related Pearson's correlation coefficients[110]

Figure 1.24 illustrates the type of scatter plot relationships that exist between 2 variables that exhibit specific Pearson's correlation coefficients, specifically that a strong correlation between variables is seen as the absolute value of the coefficient increases. Furthermore, the sign of Pearson's coefficient will determine whether the variables have a positive and negative correlation. Finally, it shows that as the correlation coefficient gets closer to zero, weaker variable correlation is observed. This is confirmed by Table 1.8.

Table 1.8: Interpretation of Pearson's correlation coefficient [110],[112]

Absolute magnitude of Pearson's correlation coefficient (r)	Interpretation
$0.00 \leq r < 0.10$	Negligible correlation
$0.10 \leq r < 0.40$	Weak correlation
$0.40 \leq r < 0.70$	Moderate correlation
$0.70 \leq r < 0.90$	Strong correlation
$0.90 \leq r \leq 1.00$	Very strong correlation

Table 1.8 indicates that as the absolute value of the correlation coefficient increases, the strength of the linear correlation between the 2 variables increases. This interpretation should be used in order to identify the variables that have the best correlation with the output variable, electricity consumption.

Train model/Test model - Performance evaluation

After the selection of the appropriate model types and the variables that will be included in model development it becomes possible to fit the models. As described in the ‘*Split data into training and testing sets*’ section, the models will be fit and tested on 2 sets of data. The training set will be used for model fitting and the testing set will be used to evaluate the generalisability of the model. The model to be used will be selected based on its performance on the training and test sets.

It is therefore necessary for the performance of the models to be measured and compared [114]. A large variety of performance metrics exist for the evaluation of regression model performance. As such, it is necessary to select an appropriate metric, or metrics, that can be used. As in the previous sections, the selected studies were evaluated to determine the number and type of performance metrics they used for model development. The results are summarised in Figure 1.25 and Table 1.9.

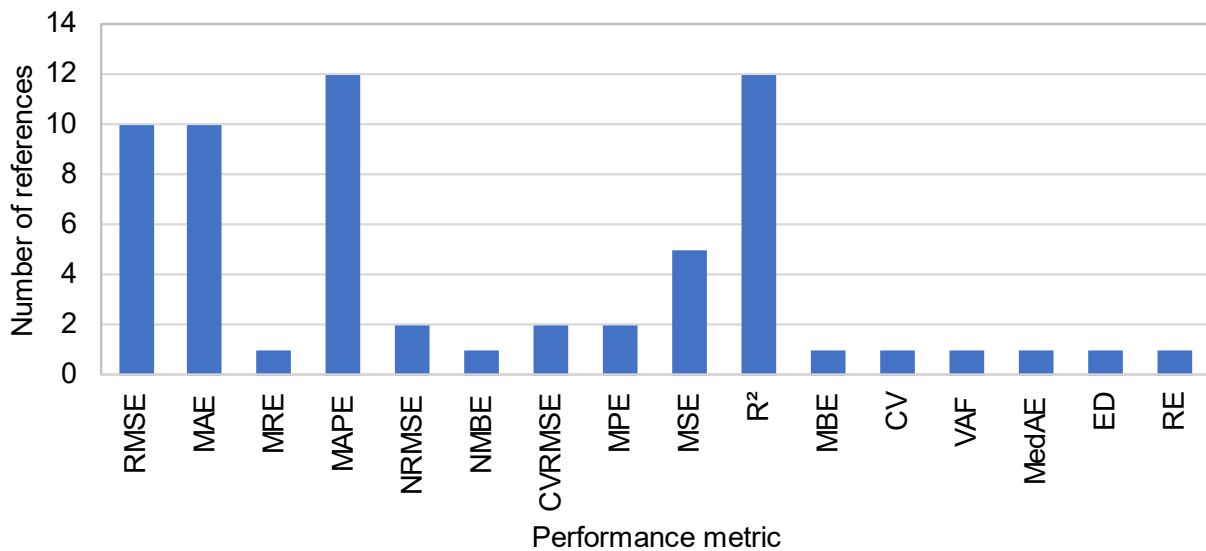


Figure 1.25: Summary of the performance metrics used in the evaluated studies

Table 1.9: Abbreviations of performance metrics and their meanings

Abbreviation	Explanation
RMSE	Root of the mean squared error
MAE	Mean of the absolute error

Abbreviation	Explanation
MRE	Mean of the relative error
MAPE	Mean of the absolute percentage error
NRMSE	Normalised root of the mean squared error
NMBE	Normalised mean of the biased error
CVRMSE	Coefficient of variation of the root of the mean squared error
MPE	Mean of the percentage error
MSE	Mean of the squared error
R ²	Coefficient of determination
MBE	Mean of the biased error
CV	Coefficient of variation
VAF	Variance accounted for
MedAE	Median of the absolute error
ED	Euclidean distance
RE	Relative error

Of the 23 studies evaluated, it was found that 19 used multiple performance metrics to evaluate the models. Figure 1.25 illustrates that 16 different performance metrics were identified from the various studies described in Table 1.9. Furthermore, it can be seen that the most common performance metrics that were selected were MAPE and R² that were used in 12 of the studies. These were closely followed by RMSE and MAE that were used in 10 of the evaluated studies.

A study performed on the popularity of error metrics used [114] also identified RMSE, MAE and MAPE as the most popular. These were identified from independent surveys performed over 25 years [114]. Although these metrics are popular, there has been much contention regarding their accuracy and bias. Various studies have stated that some of these metrics are not representative and can lead to misleading model selection.

A study was performed comparing the value of using RMSE over MAE in 2014 [115]. This was done to address concerns of other researchers who claimed that RMSE was not a good indicator of average model performance. This study found that in certain circumstances RMSE would be more beneficial than MSE, such as when the errors follow a normal distribution.

The study, however, stated that they do not claim that RMSE is superior to MSE [115]. The study concludes by recommending that a combination of metrics, over and above just RMSE and MSE, should be considered when assessing model performance [115]. This is confirmed by the evaluation of the 23 references as 19 were found to consider multiple performance metrics.

A further study, performed on the comparison of R^2 with MAE, MAPE, MSE and RMSE, concluded that R^2 is more informative than the others [116]. This is, however, when the error metrics are considered without context. The study states that an MAE of 0.7, for example, alone does not provide much detail of model performance. This is in contrast with a value of 0.7 for the coefficient of determination (R^2). Therefore, in order to ensure that all error metrics are informative, units and context must be provided [116].

Finally, 19 of the 23 studies that were evaluated considered multiple error metrics, and various other studies state that multiple error metrics should be considered with context [115],[116],[117]. The most common error metrics considered were MAPE, R^2 , RMSE and MAE. These metrics will be discussed in detail and are calculated with the use of Equation (1.16) to Equation (1.20) [116].

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad (1.16)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (1.17)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (\bar{y} - y_i)^2} \quad (1.18)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (1.19)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (1.20)$$

where:

y_i	Actual value of output variable
$\hat{y}_i$	Forecasted value of the output variable
$\bar{y}$	Average of actual output values
N	Number of data points

Equation (1.16) simply defines the average value of the output variable to be used in the other equations. Equation (1.17) defines the MAPE, which should be used when relative variations

should be more strongly considered than absolute variations. The main drawback of this metric is that it is biased towards low forecasts, making it unsuitable for models where large errors are expected [116].

Equation (1.18) describes the calculation of the coefficient of determination. This is defined as the sum of the squared differences between the actual and forecast values divided by the squared difference between the actual values and the mean. This performance metric, although being more informative without context, is highly influenced by measurement error [118].

The coefficient of determination should be considered with caution if data is known to have a high measurement error. It can range from negative infinity to 1 [118], with negative numbers indicating poorer performance than that of a horizontal line equal to the average of the target variable for all training points [116]. Values between 0 and 1 on the other hand indicate increasingly improved performance. The disadvantage of this metric is that negative numbers tend to simply indicate poor performance and are not comparable or quantifiable.

Equation (1.19) defines the RMSE, which, as the name describes, is the square root of the average squared difference between the forecasted output value and the actual output value. Due to the squaring of the error this error metric penalises large error more strictly than MAE, which treats all errors the same [115],[116]. Furthermore, the application of the square root allows the units of the error to be more interpretable than regular MSE [115]. For the purposes of electricity forecasting the units will be in kWh which is easy to interpret.

Finally, Equation (1.20) defines the MAE as the average of the absolute difference between the forecast values and the actual values. As previously mentioned, this metric does not penalise large errors like RMSE or MSE do. As such this metric is not as harshly affected by the presence of outliers in the data [115],[116].

It is clear that much contention exists regarding which error metrics to use for the evaluation of model performance. This is due to the behaviour of different error metrics as a result of differences in the data. As such, a range of error metrics should be applied and compared for model training and testing in order to select the most accurate model [115].

Verify model

Model verification is the process of illustrating that the model can be applied to the forecasting of the actual system [119]. This is accomplished with the use of an unseen dataset that would not have been previously used for model training and testing. The importance of this is to

ensure that the model has been created to perform with sufficient accuracy and expected performance on real-world data.

The separation of this dataset was discussed in the “*Split data into training and testing sets*” section. As with the previous sections, the chosen references were evaluated to assess whether they performed model verification on an unseen dataset.

It was found that only 4 of the 24 references explicitly performed a separate verification step using a completely unseen dataset. However, this is unrepresentative as the majority of the studies concluded by comparing the relative training and testing errors of the different models. This step would have been followed by the selection and deployment of the highest performing model.

The resulting models can be validated by using the same performance metrics detailed in the ‘*Train model/Test model - Performance evaluation*’ section. These metrics are, specifically, MAPE, R^2 , RMSE and MAE. These metrics will simply be calculated on a different set of values. Furthermore, the inclusion of visual methods of comparison will also be included in order to visually illustrate the differences in model performance and to validate model assumptions

These visual techniques include 1) a scatter plot between the actual and predicted values, 2) a frequency plot of the verification error as well as 3) a residual plot of the data against the predicted output values. Each of these plots serves to illustrate a different aspect of model performance which will be explained below.

A scatter plot between the actual and predicted values allows the evaluation of the similarity between the forecast values and the actual expected values. A correlation coefficient can then be calculated between the 2 sets of data points which will give a quantitative representation of model performance. This quantitative value will be between 0 and 1, with 0 representing weak correlation and 1 representing perfect correlation [112]. This can be seen in Figure 1.26.

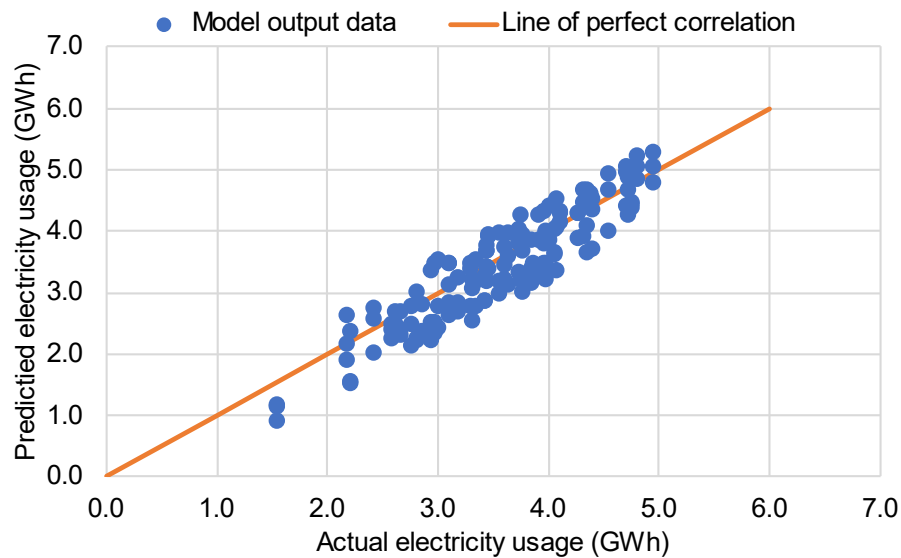


Figure 1.26: Example and desired relationship between forecasted values and actual values

Figure 1.26 illustrates the desired relationship between predicted and actual values, where the actual and forecast points are exactly the same in value. In this scenario the points would fall on the line of perfect correlation shown on the graph. This is in contrast with the example model output points shown in the figure. The correlation coefficient for those points was calculated to be 0.92. This indicates strong correlation, illustrating that the forecast points are close to the actual points.

Secondly, a frequency plot of the verification error distribution allows for the visual representation of error distribution. This provides the deviation of the error and the spread of the error. This will be useful in comparing the proposed methodology with the existing methodology to evaluate performance. Furthermore, it allows for the assessment of one of the assumptions of linear regression, which is that the residuals have a normal distribution as discussed in the '*Model type selection*' section [109].

Finally, a residual plot of the data against the predicted output values will be plotted. The plotting of this graph will allow for the assessment of one of the assumptions of linear regression, that of constant variance or heteroscedasticity [104], [105]. This assumption specifically states that the residuals in ordinary least squares regression have the same but unknown variance and was illustrated in Figure 1.22.

Model interpretation and application

Once the model has been created and validated, it is necessary to interpret the model. This can help to generate insight about the system. The aim of this process is to improve the understanding

of system operation if possible. Furthermore, understanding the model construction and limitations can assist in interpreting model outputs when they are used for forecasting.

It was found that 15 of the 23 studies attempted to extract insight about system behaviour from the resulting models. These insights were drawn by understanding model inputs, evaluating model outputs and comparing a variety of constructed models. The most common insights drawn by the evaluated studies were: 1) the identification of the most influential input variables and 2) the identification of applicable model ranges and conditions. The details of each will be discussed below.

The most common insight drawn from model evaluation is the identification of the most influential variables which was performed by 7 studies. This can be done by altering the input variables included in model development and comparing the resulting model accuracy with and without specific variables [58],[72],[74]. Alternatively, this can also be achieved by evaluating scaled regression coefficients [61],[62],[69] with larger values indicating more influential variables – provided the input variables have been scaled identically.

The second most common insight drawn from 5 of the evaluated studies is the identification of the applicable range and conditions for model application. This insight is drawn from the analysis of the input variables, specifically, the sources and range of the input variables. Understanding the source and ranges of input variables used for model development ensures that the model is not used to forecast with input variables outside of these ranges [59],[65]. This would lead to extrapolation, which would yield results of higher/uncertain accuracy.

Furthermore, this can lead to the identification that the system requires multiple models for different input conditions, such as summer and winter models [53],[57],[66]. The identification of these different operating conditions can lead to the development of more accurate models that are more specifically tailored to system behaviour. This behaviour can be found by identifying consistent errors between model and actual output for a specific operating condition, such as a change in season.

Therefore, it is clear that the interpretation of the final models as well as their expected and actual outputs for specific operating conditions can lead to the development of system insights. These system insights can be used to enrich user understanding of the system and can be recirculated for the development of more accurate models and forecasts in the future.

Summary of overall process

A generic modelling process was presented in section 1.2.2 [46]. This modelling process was used as a template against which sources were compared. This allowed the extraction of a

general modelling process used in electricity forecasting in other industries. The generic modelling process, edited to include only the relevant steps, is shown below as reference followed by the general modelling process extracted by the evaluation of 23 references.

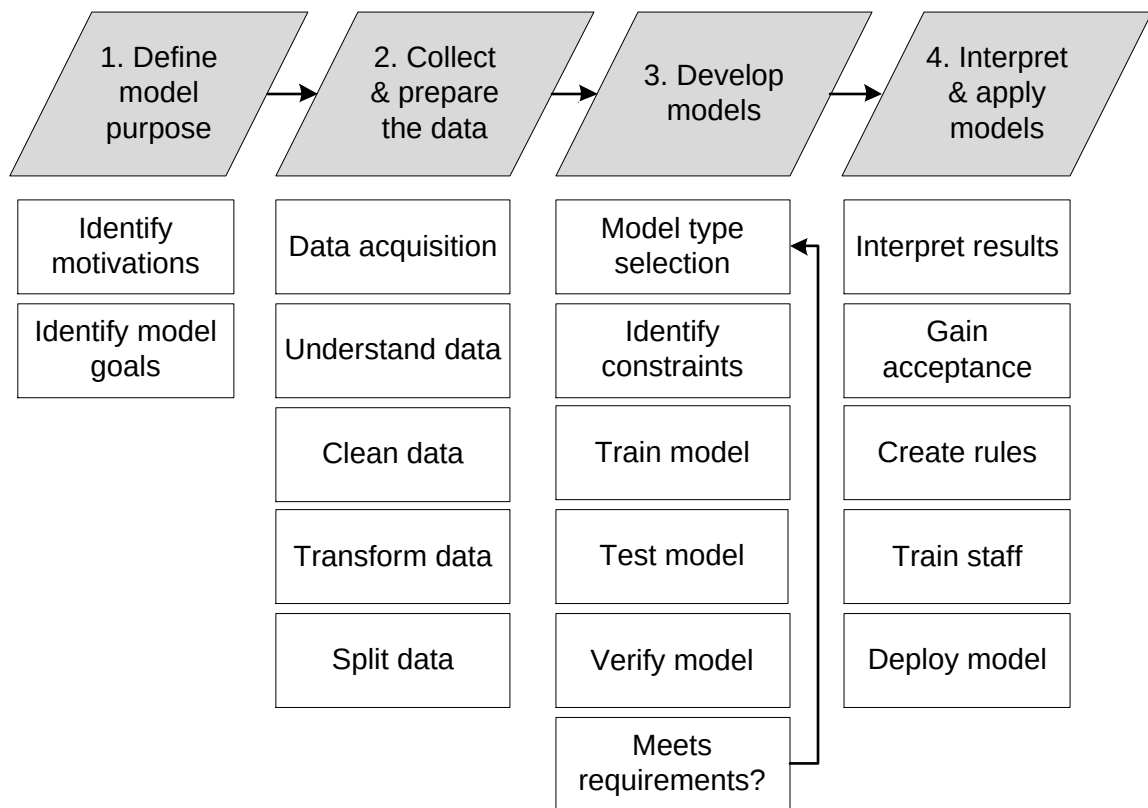


Figure 1.27: Generic procedure for modelling a variable as a function of its inputs [46]

Figure 1.27 illustrates the generic process used to model a variable as a function of its input variables. The modelling process is discussed in steps 2 and 3 of the generic process and as such, only these steps will be explicitly expanded. From the evaluation of the 23 electricity forecasting/modelling sources in the context of the steps in Figure 1.27, it is possible to create the summary diagram shown in Figure 1.28.

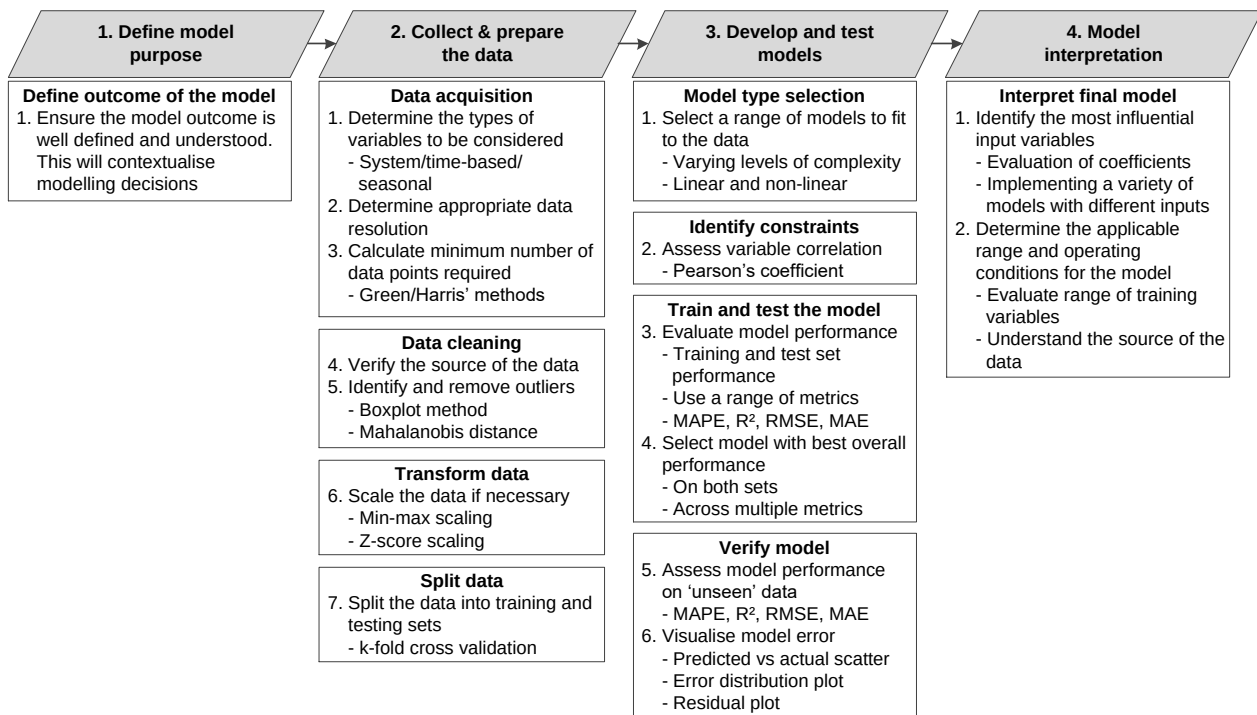


Figure 1.28: General modelling/forecasting procedure used in other industries

Figure 1.28 outlines the general modelling/forecasting procedure extracted from the selected studies and relevant literature when evaluated in the context of the generic process in Figure 1.27. This general modelling process will be used to formulate a proposed forecasting procedure that will be applied to the electricity usage forecasting on platinum mines.

1.2.3 Critical comparison

From the evaluation of the forecasting process used in the mining industry and the processes used in other industries it was possible to summarise the processes shown in Figure 1.7 and Figure 1.28. The comparison of these summaries will allow the critical evaluation of the forecasting process used in the mining industry. This critical comparison will allow the identification of steps that can be added to the mine electricity forecasting process that might improve forecast accuracy.

The comparison was performed by using the general modelling/forecasting process summarised in Figure 1.28 as a template. This process was used as the template as it was formulated from the evaluation of 23 references and relevant literature. The electricity forecasting process used in the mining industry, shown in Figure 1.7, was compared with this template to evaluate how many of the steps are accounted for.

This will allow the identification of pitfalls in the existing mine electricity forecasting process that might be introducing error into their forecasts. The comparison is shown in Figure 1.29. Note that

steps 1 and 4 shown in Figure 1.28 have been excluded so as to only compare the technical modelling steps of each process.

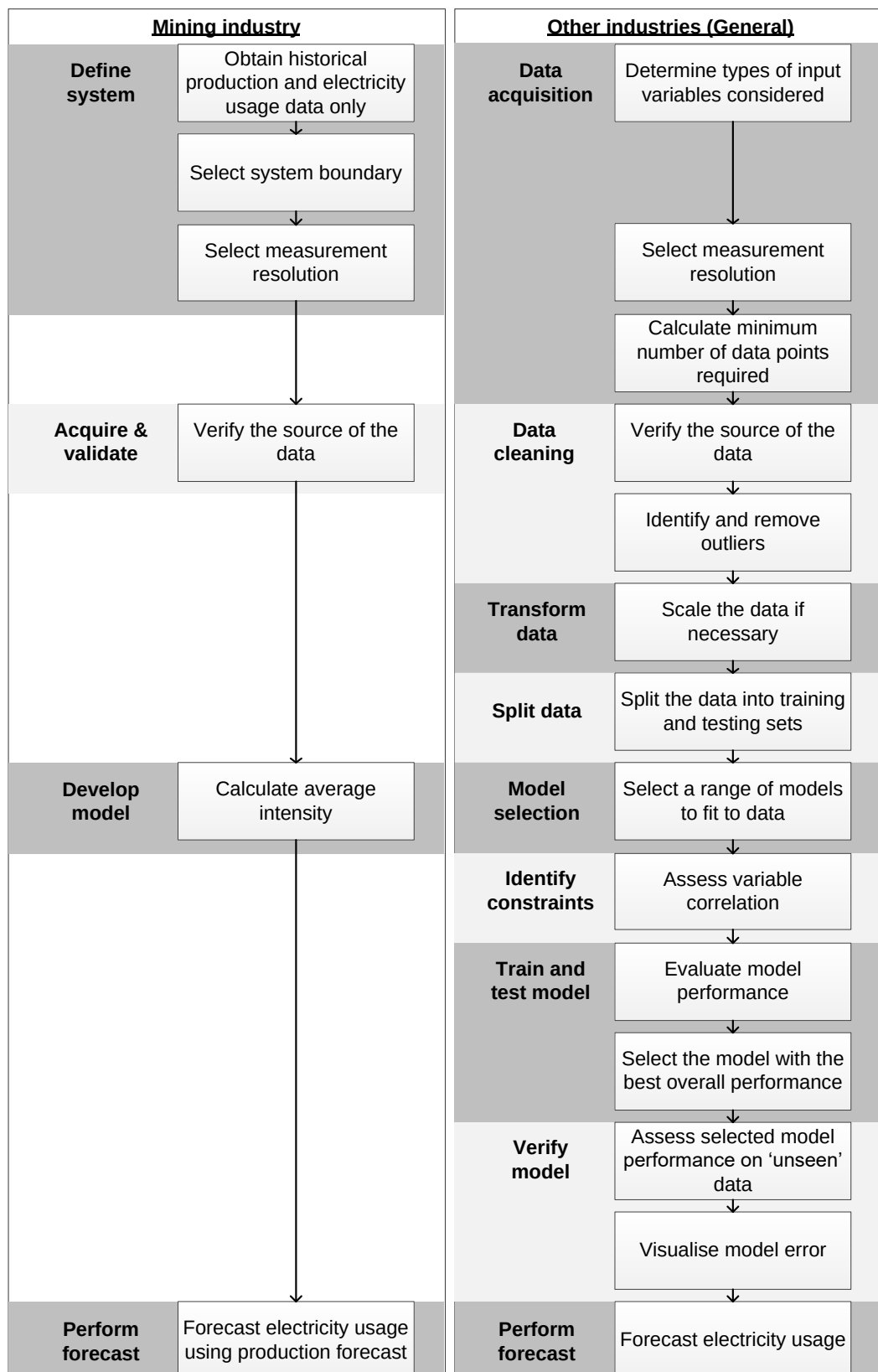


Figure 1.29: Critical comparison between the mining industry forecasting process (left) and the forecasting process used in other industries (right)

Figure 1.29 illustrates the existing forecasting process used in the mining industry on the left and the general forecasting process extracted from other industries on the right. Firstly, it is evident that the mining industry only considers production as an input variable. This is in contrast with other industries that consider system, time-based and seasonal variables. This may lead to more accurate models due to relationships in the system being accounted for.

Secondly, it is evident that the data acquisition and cleaning process used in other industries is more detailed and explicitly follows steps to clean, transform and split the data. Although more detailed, the forecasting process used by other industries does not explicitly define a system boundary, such as in the mining forecasting process.

Figure 1.29 illustrates that like the data-processing step of the process, the modelling process used in other industries is more detailed. The major difference is seen where a single, simple, linear model is considered in the mining industry, while other industries select a range of models. These models also have different levels of complexity and linearity. This allows for the comparison and selection of the model that has the best overall performance on training and testing data.

Model performance evaluation was not explicitly discussed in the mining industry forecasting process. This would be irrelevant, however, as only a single model is used. Therefore, the model's performance would not be comparable with anything. Conversely, due to the consideration of various models by other industries, model performance evaluation and model verification are performed.

Model performance evaluation in other industries is performed by considering a range of performance metrics. The model performance on the training and testing datasets are both evaluated to ensure that the model generalises well to unseen data. Furthermore, the evaluation of multiple metrics allows thorough performance evaluation as some metrics are thought to favour specific data.

In conclusion, the electricity forecasting process used in mining was found to be limited when compared with the general modelling/forecasting process used in other industries. Therefore, there exists scope to include steps from the general process used in other industries to the electricity usage forecasting on deep-level mines. The inclusion of these steps will introduce validated statistical methods for cleaning and evaluating data as well as modelling. This will possibly result in an improvement in model and forecasting accuracy.

1.3 Problem statement

Deep-level platinum mines are incredibly complex and interconnected systems. They utilise a large amount of electricity to maintain a standard of safe, efficient production. Detailed planning

of this electricity usage is required to ensure that the mine can appropriately budget for their monthly expenses. Electricity usage forecasting techniques are used for the planning of this electricity usage.

The forecasting technique currently used on deep-level platinum mines is found to be overly simplistic when compared with that utilised in other industries. The over-simplification of the modelling process might be compromising the accuracy of the resulting forecasts. This is the result of either neglecting relationships between variables or over-simplifying the relationships that are already accounted for.

The following problems are thus highlighted:

- The current forecasting process on deep-level platinum mines is overly simplistic
- Over-simplification neglects the complexity of the system which can lead to inaccuracies

To address these problems, an improved forecasting methodology must be applied to mine electricity usage forecasting. The results of this application can then be compared with the existing method to identify whether improvement is achieved. However, there does not exist a single method for this process. As such, a general method will be extracted from electricity forecasting/modelling performed in other industries.

1.4 Need for the study

A problem with electricity usage forecasting in deep-level mines was identified. The accuracy of the forecasts generated in this industry stand to be improved if a more robust forecasting process is used. However, a detailed, general forecasting process does not exist for this application. Therefore, a need exists to:

- extract a general procedure for electricity usage forecasting from other industries' processes,
- apply this new procedure and the existing forecasting procedure to mine data,
- objectively assess the results of the two applications to quantitatively identify whether changing the methodology offers accuracy improvement and, if so, how much.

1.5 Study objective

The primary objective of this study is to develop an improved modelling process that can be applied to the electricity usage forecasting on deep-level platinum mines. This process should allow for the comparison of different mathematical models and forecast values should be chosen

based on comparative accuracy-based criteria. In order to achieve this, the following objectives will be met:

- a general process flow for model development will be extracted from literature,
- relevant research will be provided on specific techniques applied in the literature,
- these techniques will be used to develop an improved strategy for electricity forecasting on deep-level platinum mines,
- the devised strategy will be verified by applying it to three case study mines,
- the results of the new strategy will be compared with those of the existing method to evaluate accuracy improvement.

By meeting these objectives, the study will provide a general procedure that can be applied for the electricity usage forecasting on deep-level platinum mines. This procedure will include the consideration of different models. Furthermore, model results will be objectively compared in order to allow for the selection of the model that produces the best forecasting accuracy based on preselected performance metrics.

1.6 Dissertation overview

Chapter 1: Literature review

This chapter presents an introduction and background to the study. The problem statement outlines the need for the study while the objectives detail the deliverables that will be met in order to satisfy the need. A thorough literature review is also discussed in Section 1.2 to identify the forecasting processes used in mining and other industries as well as relevant theory related to those processes.

Chapter 2: Methodology

Chapter 2 will present an improved strategy for electricity usage forecasting on deep-level platinum mines. The strategy will consist of the steps identified in Chapter 1: definition of model purpose, data acquisition and cleaning, model development, model interpretation and application and model verification.

Chapter 3: Results and discussion

This chapter will involve the application of the developed forecasting strategy to mine case studies for verification. The strategy will be applied to three case study mines and the results compared with those generated by the existing methodology to validate whether the proposed methodology solves the research problem. The case studies vary according to the size and layout of the shafts.

Electricity usage forecasts will be developed using the devised strategy for each case study. These forecasts will be compared with those developed with the use of the electricity intensity methodology in order to evaluate the change in overall forecast accuracy.

Chapter 4: Conclusion

In Chapter 4 a summary of all conclusions made in the study will be presented. This chapter will refer back to the objectives presented in Chapter 1 in order to prove that all study objectives have been met. Recommendations for further studies will also be proposed in this chapter.

CHAPTER 2 METHODOLOGY

2.1 Introduction

The problem statement in Chapter 1 outlined the need to improve the existing forecasting methods on deep-level platinum mines. A single method for electricity usage forecasting, however, does not exist. As such, a literature study was performed to extract the general method followed in other industries. This extraction was performed by comparing each study with a generic modelling process.

The literature study indicated that electricity forecasting in other industries roughly followed the generic modelling process. The overall modelling process was consequently divided into 4 main steps, 1) identification of model purpose, 2) data collection and preparation, 3) model development and testing and 4) model interpretation. In this chapter these steps will be used to present an improved strategy for forecasting electricity usage on deep-level platinum mines. Figure 2.1 illustrates the flow of the proposed forecasting method.

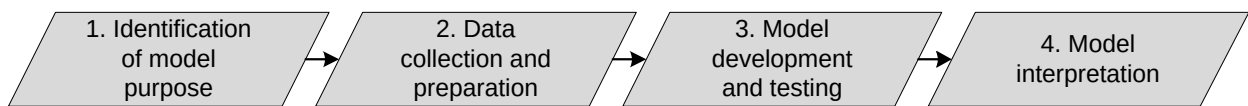


Figure 2.1: Four main steps for the proposed method of electricity usage forecasting on deep-level platinum mines

Each of the main steps shown in Figure 2.1 includes a set of sub-steps outlining the technical details of each process. The discussion of these sub-steps will be performed in this chapter. The discussion of the main steps as well as their respective sub-steps will constitute the first part of the methodology of this study. Additionally, a separate step will be included in order to validate the proposed forecasting method. This step will include the explicit comparison of results using both the proposed method and the existing method.

2.2 Identification of model purpose

To ensure that a model that meets required outcomes is created, it is first necessary to identify those outcomes and the purpose of the model. The intentions of the model must be fully understood, and the goals of the model must be set. The sub-steps of this step in the process are shown in Figure 2.2.

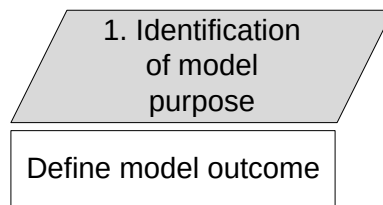


Figure 2.2: Identification of model purpose sub-steps

Figure 2.2 illustrates that the sub-step for this step is to define the outcome of the model. This model outcome will have an impact on decisions made later in the process. Therefore, it is important that the modelling intention is fully understood. The outcome of the electricity forecast is obviously to model the expected electricity consumption of the mine.

The context of this forecast can either be for budgeting purposes, energy management purposes or design purposes, among others. It is important to determine which purpose the model is meant to satisfy at the start of the process as this purpose will contextualise the decisions that will be made. Therefore, this step of the process will simply involve stating what the outcome of the model is intended to be.

2.3 Data collection and preparation

The second main step of the proposed forecasting method is the collection and preparation of the data. Figure 2.3 illustrates the sub-steps that comprise this process. The steps begin with the selection of the system boundary, and thereafter, data within the selected boundary can be obtained. The acquired data must then be cleaned and transformed, if necessary. Finally, the set of validated, cleaned and transformed data must then be split into different sub-sets, one for the training and another one for the testing of the model.

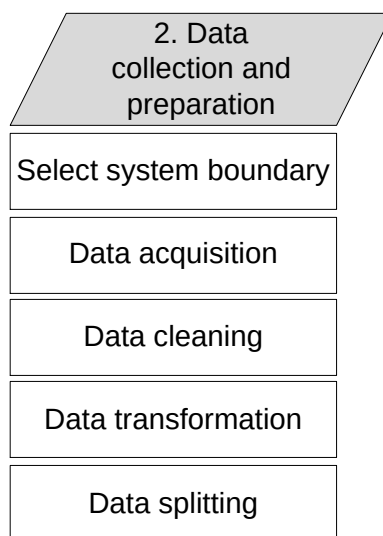


Figure 2.3: Data collection and preparation sub-steps

2.3.1 Selection of system boundary

From section 1.2.1, it was found that the majority of budgets set up on mines were on a shaft level, with few of the studies budgeting for the whole facility as a single unit. This decision will impact the location of the system boundary. The selection of this system boundary is dependent on a wide range of factors. It also has significant implications for the forecasting process.

The level of detail in which the forecast can be performed is primarily dependent on the level of detail of the electrical metering of the facilities [40]. Furthermore, the quality of the data associated with different levels of metering will also determine how detailed the generated forecast can be. The electricity usage data of each system that is forecasted must be verified in order to ensure that the forecast is based on agreed upon values.

This is especially important if the forecast will be used for energy management and budgeting purposes. It ensures that the relevant stakeholders, those responsible for the electricity usage, are held accountable relative to official, verified data. Figure 2.4 illustrates the possible system boundaries that can be selected.

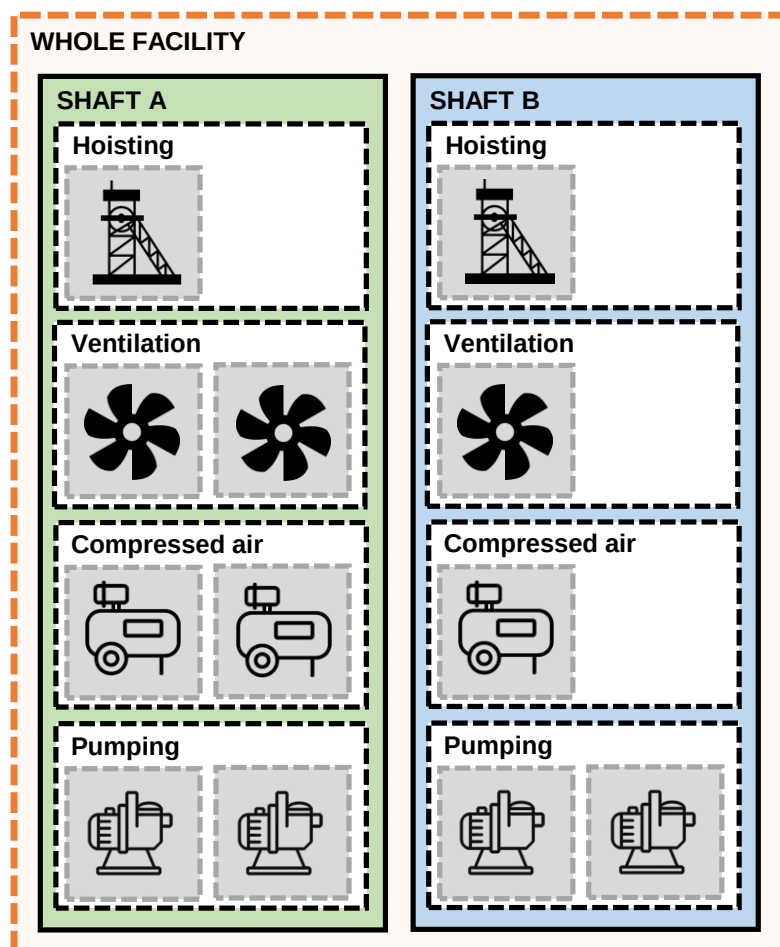


Figure 2.4: System boundaries that can be selected for forecasting depending on data availability

Figure 2.4 illustrates a facility comprised of two shafts, each having their own independent winding, ventilation, compressed air and pumping sub-systems. Each of the sub-systems is also made up of individual components. As such, depending on the degree and accuracy of the metering, system boundaries can be drawn as illustrated by the dashed lines in Figure 2.4. These can range from the overall facility down to the individual components of each sub-system. Once the system boundary has been selected the data that will be used for model development must be acquired.

2.3.2 Data acquisition

As discussed in section 2.3.1, the selection of the system boundary, which is dependent on the level of metering available, will direct this step. The variables, and thus the data, that can be used for modelling will depend on what systems are being modelled exactly. For example, the variables that best describe the electricity usage of the whole mine might be different from those that describe the electricity usage of a compressor on a specific compressed air ring.

Therefore, the selection of the appropriate variables/data is an important consideration. The user must consider a range of different types of variables. These include system, time-based and seasonal variables. This will ensure that a range of inputs affecting the output variable will be considered in the modelling step of the process.

Once the variables have been selected the resolution of the data to be acquired must be selected or determined. As discussed in section 1.2.1, the majority of deep-level platinum mines forecast on either a monthly or quarterly basis. This is the result of their production forecasts and budgets being required and created in these intervals [40]. The required resolution will determine the intervals in which the data must be obtained.

As discussed in section 1.2.1, the electricity usage data on mines is typically obtained from a third-party service provider. These providers are responsible for the auditing and consolidation of the electricity data. As such, the resulting resolution might be determined by the resolution in which the data has been audited and validated.

Once the desired input variables and resolution have been decided upon, or dictated by the system, it is necessary to calculate the minimum data points required. This will be performed by using a combination of Green's or Harris's methods as outlined in section 1.2.2. This will determine the number of data points that must be obtained in order to ensure that the resulting model is statistically significant. The combination of these methods is shown in Equation (2.1).

$$N \geq \begin{cases} 50 + i, & i \leq 5 \\ 50 + 8i, & i > 5 \end{cases} \quad (2.1)$$

Equation (2.1) illustrates the combination of Green and Harris's methods that will be used to determine the minimum sample size required in this study. It is a piecewise combination of the 2 methods, applying Harris's rule of thumb for studies with less than or equal to 5 input variables and Green's for those with more than 5 input variables. Examples of the minimum number of data points required given the number of inputs variables are shown in Table 2.1.

Table 2.1: Examples of the minimum number of data points required as a function of the number of input variables

Example	Number of input variables	Minimum number of data points
1	1	51
2	3	53
3	6	98

Table 2.1 illustrates the application of Equation (2.1) to determine the minimum number of data points required for a model to be statistically significant. Once this has been determined the required data can be obtained from the relevant sources.

2.3.3 Data cleaning

After the data has been obtained, the data quality must be inspected. This is important as poor-quality data can have an adverse effect on the quality and thus the accuracy of the resulting models. The initial step of this inspection is to verify the source of the data. The data must be considered verified and official; this is especially important if the resulting forecast will be used for budgeting or energy management purposes.

In the case of most deep-level platinum mines, the electricity usage data obtained from the third-party metering provider is authenticated by comparing it with the electricity measured by the electricity provider. These values are considered official and are thus used for billing and budgeting. As discussed in section 1.2.2, production data is obtained from mine systems and is verified by the calibration of scales, which is performed on a regular basis.

The verified data must then be inspected for outliers. This is of particular importance as outliers can cause a significant deterioration in model accuracy and performance. Outlier identification will be done with the use of the Mahalanobis distance, and the boxplot method as discussed in section 1.2.2. The Mahalanobis distance will initially be calculated for each data point. Thereafter, the boxplot method will be applied to the Mahalanobis distances to identify outliers.

This is to ensure that outliers are removed by means of a valid statistical method and not just by intuition, which can introduce bias. Any identified outliers must be removed in order to maintain a dataset of high quality.

2.3.4 Data transformation

As discussed in section 1.2.2, the transformation of the data can be performed if the selected input variables exist on significantly different scales. The benefit of data transformation or scaling is to normalise the weighting of different input variables in the resulting model. This, however, reduces the interpretability of the model as the units of input variables are altered by the scaling.

As such, the data will only be transformed when 1) more than a single input variable is considered and 2) the units of the input variables are significantly different. A selection of two scaling methods is available: 1) min-max scaling and 2) z-score normalisation. The selection of which method to use will depend on the distribution of the data.

As discussed in section 1.2.2, min-max scaling scales all values to exist in a decided range and does not change the distribution of the data. Z-score normalisation, however, changes the distribution of the data and results in the data having a mean of 0 and standard deviation of 1. This results in the dataset having a normal or Gaussian distribution.

Z-score normalisation will therefore be applied on datasets that are known to already be Gaussian. Min-max scaling will be applied when datasets are not Gaussian or when the user is unsure of the distribution.

2.3.5 Data splitting

Once the data has been pre-processed by means of outlier removal and scaling, it must be split into different sets for model training, testing and verification. From section 1.2.2, it was found that the rule of thumb of splitting is to use a 70:30 split between training and testing data. This split is typically chosen when a single holdout set for testing is separated from the training data.

This method will be used for the separation of the final verification set. This set will be used for the verification of the final selected model. Therefore, a 70:30 split will be performed to split the data into a training and a verification set respectively. It must be noted that the values will be randomised prior to the split in order to avoid bias. This process is shown in Figure 2.5.

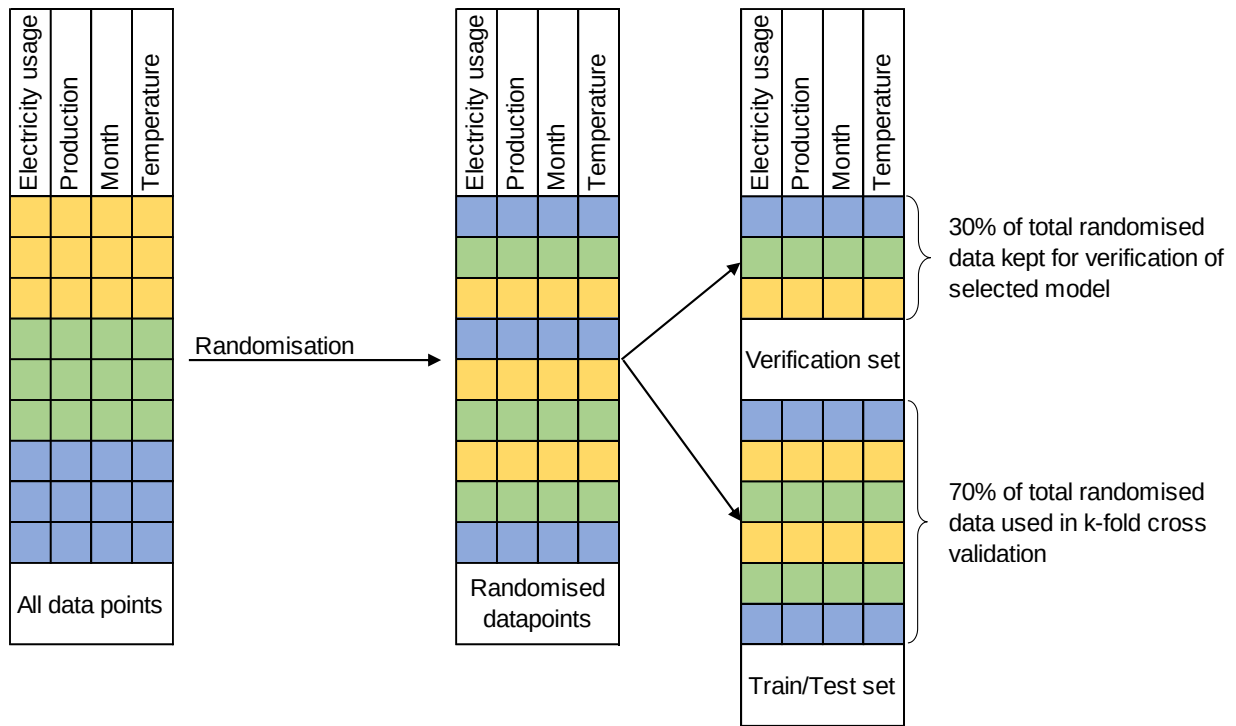


Figure 2.5: Illustration of the randomisation and separation of the total data into verification and training/test sets

Figure 2.5 illustrates the randomisation of the total dataset followed by the splitting of the data into the verification and train/test sets. The verification set will be kept until the end of the process and will be used to evaluate the performance of the final selected model. The train/test set will be used for model fitting using k-fold cross-validation.

The training/test set, which will be comprised of 70% of the total number of data points, will then be used for the training, testing and selection of the final model. This will be performed with the use of k-fold cross-validation. The purpose of using k-fold cross-validation is to avoid bias during model training and selection. The process of k-fold cross-validation is illustrated in Figure 2.6.

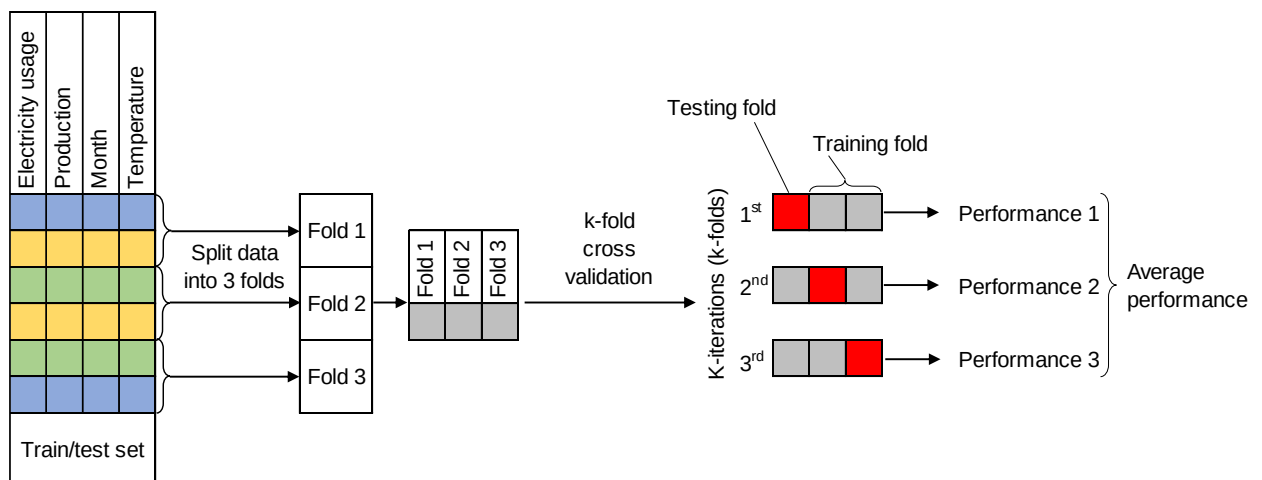


Figure 2.6: Illustration of k-fold cross-validation using the training/testing dataset

Figure 2.6 illustrates the process of k-fold cross-validation. In this process, the randomised training/test dataset is split into a known number of subsets called folds. Each fold will then be used as the testing set sequentially and the model will be trained on the remaining folds. The model will be trained as many times as there are folds and the relevant performance metrics will be calculated. These performance metrics can then be used for the selection of the best model.

2.3.6 Summary

The process of data collection and preparation is crucial to ensure that a verified, high-quality dataset is used for forecasting model development. This is done through five sub-steps. The first step involves the definition of the system by selecting the system boundary. Once the system is defined the appropriate input variable types and resolutions are selected – depending on the degree of metering of the system.

Thereafter, it is possible to calculate and obtain at least the minimum number of data points required for the resulting model to be statistically significant. The raw data is then cleaned through outlier removal and scaled if necessary. The fully transformed dataset is then split into training and verification sets using a 70:30 split. K-fold cross-validation is then used for the training, testing and selection of the final forecasting model.

2.4 Model development

Once the dataset has been processed and verified, it is possible to progress to the model development stage. Model development is performed through four sub-steps, namely model type selection, correlation evaluation, training and testing of the model and finally model verification. These steps are shown in Figure 2.7 and will be discussed in detail in the following sections.

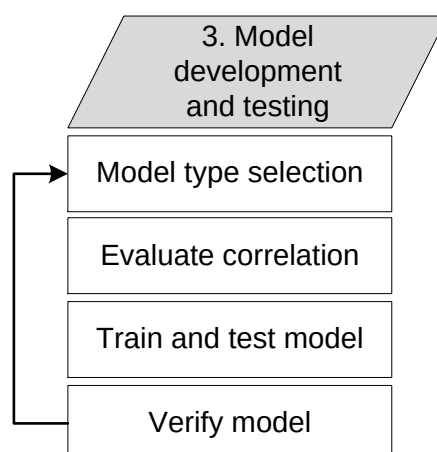


Figure 2.7: Model development and testing sub-steps

2.4.1 Model type selection

The initial step in forecasting model development is to decide on a selection of models to fit to the data and compare. As discussed in section 1.2.2, the majority of the evaluated studies selected a variety of models of different complexity and linearity. Thereafter, through the evaluation of the model performance, the best model for their specific application could be selected.

These studies typically selected models of varying complexity levels as well as a mixture of linear and non-linear models. These models include, in order of decreasing complexity, artificial neural networks, random forests, decision trees and ordinary least squares regression. It was also found that in some instances the simpler models either outperformed or performed just as well as the more complex models [66]. This is dependent on the type of system being modelled and the data available.

For the purposes of this study, ordinary least squares regression will be considered as the forecasting model. This includes the consideration of different functional relationships between the input and output variables so as to consider non-linear models. Since this model offers a complexity improvement over the intensity model, it will be the only other model considered. However, more complex models should be considered as they might result in an improvement in accuracy.

The aim of including models of varying levels of complexity and different linearities is to take the variability of mining systems into account. While some systems might exhibit linear relationships to variables, others might not. This was discussed in section 1.1.2. Thus, the consideration of non-linear relationships using ordinary least squares will be used to illustrate the benefit of this consideration.

2.4.2 Variable correlation assessment

Once the applicable models have been selected it is important to investigate variable correlation. This will allow the selection of the appropriate variables to be included in the model. This is necessary as the inclusion of variables that do not have a strong correlation with the output variable might have an adverse effect on model performance. Furthermore, the removal of non-correlated variables will result in a simpler, more interpretable model.

As discussed in section 1.2.2, this will be done with the use of Pearson's correlation coefficient. A scatter plot will be created between each input variable and the output variable, electricity usage in this case. Pearson's correlation coefficient will then be calculated to supplement the scatter plot. Only the variables with the highest correlation coefficients will then be selected for model

development. The threshold value for selection will be case specific and depend on the discretion of the user.

2.4.3 Model fitting and testing

After the selection of model types and the selection of the most correlated variables, it is possible to fit the models to the training data. As discussed in section 2.3.5, the data will be separated into a train/test set and a verification set by a 70:30 split. Training and testing will then be done on the training set with the use of k-fold cross-validation.

The number of folds to be used will be left to the discretion of the user. However, to keep with the rule of thumb of a 70:30 split between training and testing sets, 4 folds will be used for the purposes of this study. This creates a 75:25 train to test set split for each fold. Therefore, each model will be fit, and its performance evaluated 4 times.

The most important part of model fitting and testing is the selection of the performance metrics that will be used to assess model performance. From section 1.2.2, it was found that the most common metrics considered by the evaluated studies were MAPE, R^2 , RMSE and MAE. Each of these metrics has associated advantages and disadvantages which were discussed in section 1.2.2.

Models will be fit on the training set of each fold and tested on the testing data of that same fold. Thereafter, the training and testing performance will be calculated by using the 4 selected metrics. This will be done for each set of folds, 4 in the case of this study. Finally, the average and standard deviation of the training and testing performance across metrics will be calculated.

The utilisation of multiple metrics allows the mitigation of metric biases from skewing model performance interpretation. Models that perform well across the majority of the selected metrics on both the training and testing folds will thus be selected. Thereafter, the forecasting model will be trained on the total training set to yield the final selected model.

2.4.4 Model verification

After the selection of the model that will be used for the forecasting of the electricity usage it is necessary to evaluate the performance of that model on unseen data. This step is necessary to ensure that the selected model can perform its function with the expected results and with a sufficient amount of accuracy. This will be done by using the verification set, which was separated from the data at the start of the process.

The performance of the model on this set will be evaluated with the use of the same metrics selected for section 2.4.3. The rationale behind this remains the same, in that the consideration of multiple metrics reduces the chance of metric bias skewing the model performance evaluation. The model should perform with similar accuracy on the verification set as it did on the training/test set. If not, the model must be refit and re-evaluated.

Additionally, a set of visual techniques will also be included in order to illustrate the performance of the model and validate model assumptions. The visual techniques include 1) a scatter plot between actual and predicted values, 2) a frequency plot of the verification error distribution and 3) a residual plot. These plots and their insights were discussed in section 1.2.2.

2.4.5 Summary

In this step of the process a forecasting model is developed and validated. The process begins by considering a variety of models exhibiting different levels of complexity and linearity. Variables included in the model are selected with the use of Pearson's correlation coefficient. The model is fit and tested on a training set with the use of k-fold cross-validation.

The most suitable model is then selected through thorough performance evaluation. This model is then validated on unseen data to ensure that it performs with similar accuracy on this set. A combination of MAPE, R^2 , RMSE and MAE is used for the performance evaluation of the model on all sets to ensure that bias does not skew evaluation. Additionally, visual techniques are used to validate model assumptions and assess performance.

2.5 Model interpretation

Once the final model has been selected and its results verified using the unseen verification set, it is necessary to interpret the model. The aim of this is to improve system understanding as well as draw insights from the model that might lead to the development of more accurate models when the process is repeated. The sub-steps of this process are: 1) identify the most influential variables and 2) determine the model operating conditions, and are shown in Figure 2.8.

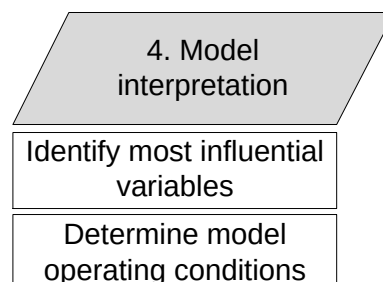


Figure 2.8: Model interpretation sub-steps

Figure 2.8 illustrates the sub-steps of the model interpretation phase of the overall process. This process begins with the identification of the most influential variables. As discussed in section 1.2.2, this can be performed either by comparing the scaled regression coefficients in the resulting models or by building a variety of models and comparing their relative accuracies with the inclusion of different input variables.

For simplicity, and since the model inputs will have been scaled, the importance of variables will be evaluated by comparing the scaled regression coefficients, if multiple variables are considered. Those variables with the largest coefficients will be considered to be the most influential in the system. This will allow the identification of variables that have a large impact on electricity usage and might lead to operational insights that can reduce electricity usage.

The second sub-step in the process involves the definition of the model operating conditions. This will be done by evaluating firstly the source of the input data, to evaluate whether the models can be extended to other systems. Secondly, the operating conditions of the models will be determined by evaluating the ranges of the input variables used for model training. These ranges will be provided with the resulting models to ensure that the models are not used with data outside of these ranges.

As explained in section 1.2.2, the utilisation of models with input data outside of the ranges used for model training causes the extrapolation of outputs. Extrapolation results in outputs with decreased accuracy as the models were not fit to data in the range entered. This insight will allow the operating range of the system to be specified and can lead to energy management capacity if the mine intends on over/under utilising a specific system.

Therefore, the outcome of model interpretation is to define the list of operating conditions of each of the resulting models. These operating conditions will ensure the utilisation of the models within the correct error by avoiding extrapolation. Furthermore, this interpretation will allow the identification of important system variables and operating ranges of systems.

2.6 Methodology validation

In order to validate whether the proposed forecasting strategy offers an improvement over the existing strategy, an additional step will be required. This step will include the comparison of the results obtained from following the proposed forecasting strategy with the results obtained by following the existing forecasting method. Specifically, the results obtained from the steps followed in sections 2.3 and 2.4 will be compared with those obtained by following the method outlined in section 1.2.1.

In order for this validation to be performed, a separate set of data will be required that was not used in the verification of either model. This data will be the most recent data available in order to simulate the electricity usage forecasting of the mine. Each model will perform the electricity usage forecast and the results will be objectively compared.

The results will be compared with the methods outlined in section 2.4.4. The four selected performance metrics will be used to calculate the overall performance of each of the models. Additionally, the visualisation techniques will be used in order to compare the distribution of errors of each model as well as the similarity of the forecast generated to the real values.

The aim of this section is to validate whether the proposed forecasting methodology offers an accuracy improvement in mine electricity usage forecasts. This will satisfy the problem statement of the study claiming that the current forecasting method results in an oversimplification of the relationships between the variables.

2.7 Conclusion

The problem statement in section 1.3 explained that the current forecasting method applied on deep-level platinum mines resulted in inaccurate forecasts. This was due to the oversimplified method resulting in the neglect of the complexity between the variables. This problem statement identified a need to formulate a more advanced forecasting strategy that would result in forecast accuracy improvement.

The improved strategy was presented in this chapter and includes 4 main focus areas: 1) identification of model purpose, 2) data collection and preparation, 3) model development and testing and 4) model interpretation. Each step is comprised of a variety of sub-steps that were explained in this chapter.

Identification of model purpose, discussed in section 2.2, is centred around the definition of the outcomes of the model. This involves setting a goal for the final model to ensure that decisions made during the technical modelling process are justified and contextualised. This step is paramount as it will determine decisions made later in the modelling process.

Data collection and preparation, discussed in section 2.3, involves the definition of the system boundary initially. This allows the selection of relevant variable types as well as resolutions depending on the degree of metering of the systems selected. The minimum number of data points can then be obtained. Once the data is acquired this step involves the cleaning, transformation and splitting of the data into training and verification sets.

The model development stage, discussed in section 2.4, involves the creation of the forecasting model that will finally be used to forecast the electricity usage. This involves the consideration of a range of models and fitting the data to variables that are strongly correlated to the output variable.

This is done with k-fold cross-validation and the models are evaluated on a range of performance metrics. The selected model's performance is evaluated on the verification set to ensure that it performs well on unseen data. This model is then selected as the best forecasting model.

The selection of a final model describing the system implies that this model can be interpreted. Model interpretation, as explained in section 2.5, involves the definition of the operating range of the model as well as the identification of the most influential variables of the system. This step in the process allows insights about the system to be drawn from the model that might assist with further modelling or with system operation.

Finally, in the methodology validation section (section 2.6), the model selected using the proposed methodology is objectively compared with the model created using the existing forecasting method. This is done by performing a simulated forecast on the most recent mine data and objectively comparing the results using the same methods outlined in the model verification step of section 2.4.

The aim of the final section is to address the problem statement of this study: **Will a more advanced forecasting strategy result in accuracy improvement in deep-level platinum mine electricity forecasting?** This method was devised from the evaluation of methods used in literature and allows for the generation of forecasting models using statistically significant methods.

CHAPTER 3 RESULTS

3.1 Introduction

In Chapter 2, a forecasting methodology was developed in order to allow the generation of more accurate electricity forecasts on deep-level platinum mines. This forecasting methodology is illustrated in Figure 3.1 and is comprised of 4 main steps. The details of each of these steps were discussed in Chapters 1 and 2 and were developed by evaluating relevant literature.

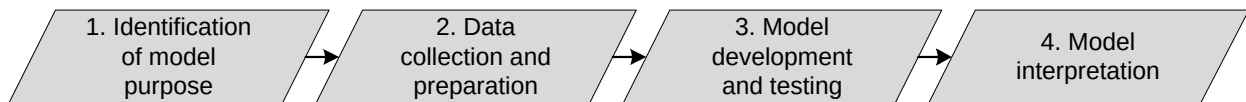


Figure 3.1: Proposed forecasting methodology steps and sub-steps

The *identification of model purpose* step allows the specification of the model outcomes, specifically, what the resulting model will be used for. This could be forecasting for budgeting purposes or for energy management. Secondly, the *data collection and preparation* step of the process focusses on the delivery of an authenticated, cleaned and validated dataset. This set will be used for fitting and validating of the forecasting model.

Model development and testing describes the process of selecting appropriate models to fit to the data. This step includes fitting of the models, comparing their accuracies and ultimately selecting and validating the final model. The process is concluded with the *model interpretation* step. This step involves extracting insights from the resulting model that can lead to increased system understanding and possibly an improvement in model accuracy.

In order to verify the proposed forecasting strategy proposed in Figure 3.1, it was applied to three case studies. These case studies are three deep-level platinum mining areas (mines) located in South Africa, and which fall under a single mining group. The results of this new application were objectively compared to the results obtained by applying the existing forecasting method.

This ensured the assessment of whether the proposed forecasting strategy offered improvements in accuracy when compared with the existing method. This ultimately allowed the assessment of whether the study met the objectives detailed in Chapter 1. While this chapter outlines the main findings from the case studies, the detailed results can be found in Appendix B.

3.2 Identification of model purpose

This section will be used to contextualise the layout and operations of each mine as well as the purpose of the electricity forecast to the reader. This will ensure contextual system understanding

that will justify the decisions made and the nature of the results that will be discussed later in the chapter.

3.2.1 Case study background

Each of the mines that will be forecasted is part of a single facility situated in the North West province in South Africa. The facility is comprised of three mines of different ages. The layouts of the mines in each case study are shown in Figure 3.2, Figure 3.3, and Figure 3.4.

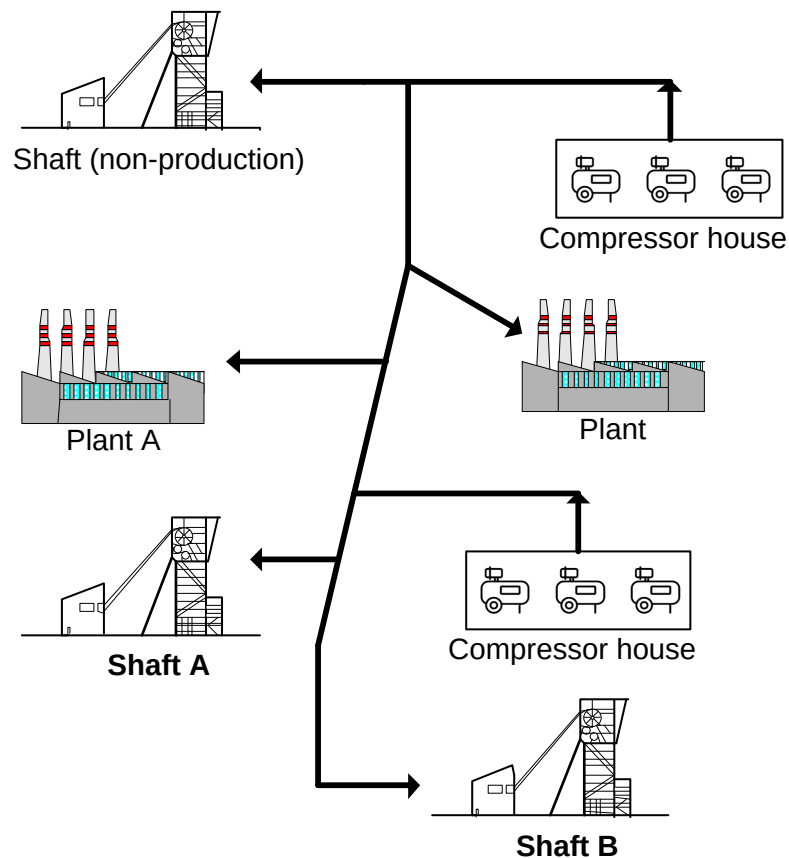


Figure 3.2: Case study mine 1 layout

Figure 3.2 illustrates the layout of case study mining area 1. This mine consists of two production shafts, shaft A and shaft B, and 1 non-production shaft. A single compressed air ring, of two compressor houses (nine compressors) is used to supply the shafts with compressed air. Each of the shafts has its own independent ventilation and dewatering system and no refrigeration is utilised at the mine. The shafts are connected underground, resulting in the flow of water and ventilation between shafts. Shaft A is a vertical shaft of depth of 900 m while shaft B is a decline shaft with a depth of 445 m [120].

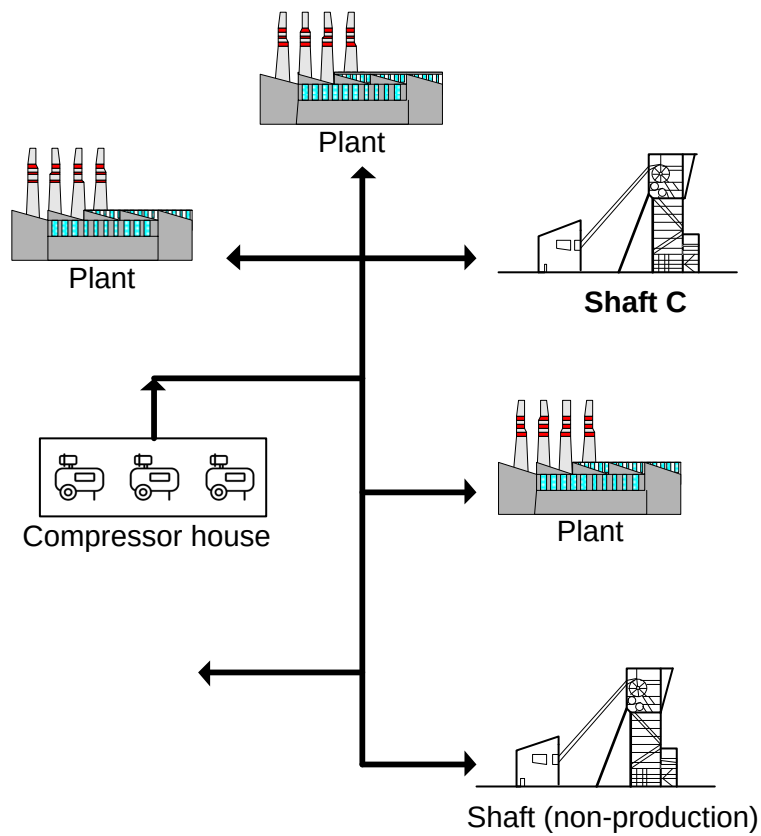


Figure 3.3: Case study mine 2 layout

Figure 3.3 illustrates the layout of case study mining area 2, a mine with a single production shaft, shaft C, and a single compressor house. It must be noted, however, that the second production shaft was decommissioned at the end of 2019 – which will have an impact on compressed air electricity usage of shaft C. Currently, the compressor house is an independent supplier of compressed air to shaft C, which is a vertical shaft with a depth of 1032 m [120]. This shaft has 6 ventilation fans and no refrigeration system.

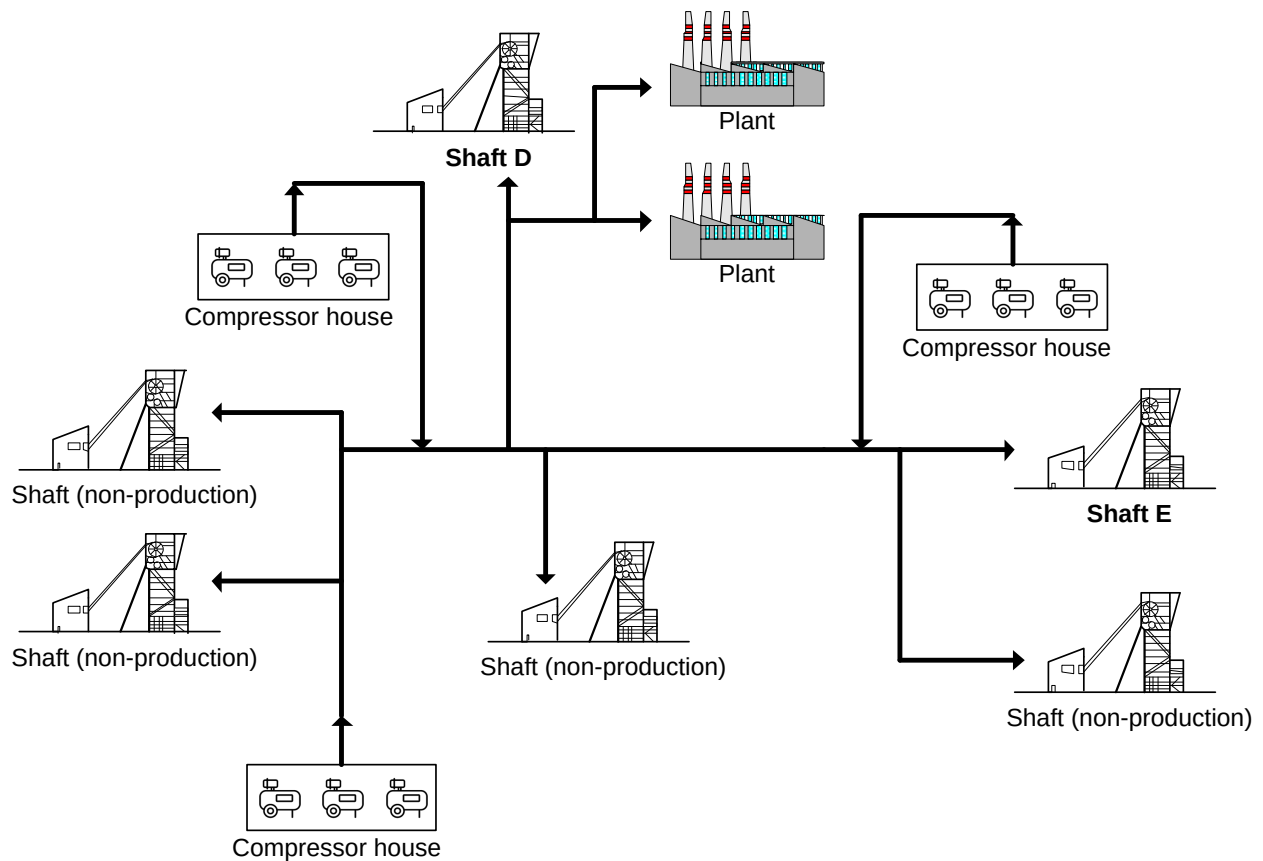


Figure 3.4: Case study mine 3 layout

Finally, Figure 3.4 illustrates the layout of case study mining area 3. This mine is comprised of two production shafts. Similar to mining area 2, 3 of the non-production shafts were decommissioned in 2019. The only remaining production shafts are shaft D, a vertical shaft with a depth of 1000 m, and shaft E, a decline shaft with a depth of 430 m [120]. These shafts are supplied compressed air from a single compressed air ring with 2 compressor houses. Each shaft has its own independent ventilation and dewatering system and no refrigeration.

3.2.2 Model purpose

Each of the case study mining areas outlined above is required to forecast their monthly production and utility consumption for the next year. This is required for capital planning as well as the planning of electricity efficiency projects that are required. Furthermore, it allows the mine to manage the electricity usage based on previously agreed upon values. This ensures that the relevant stakeholders are held accountable and that the mine does not lose profitability as a result of an over-utilisation of electricity.

This study will focus on the utilisation of production as an input variable for the forecasting of electricity usage consumption. It is understood, however, that inaccuracies in the production

forecast can lead to downstream inaccuracies. It is recommended that the investigation of this process be done in future work.

In order to calculate the operational cost, it is first necessary to forecast the electricity required to ensure the planned production. Therefore, a set of models is required that can forecast the electricity consumption of each of the mining areas on a monthly basis. These models can, if possible, forecast on a finer time resolution as these models are also to be used for energy management purposes.

Each of the mining areas requires forecasts for each of the operational shafts (five in total) at least. This is due to the fact that production is planned on a shaft basis, and as a result, the electricity usage of each shaft is monitored and can be managed. The forecast can be set up for smaller units that make up the total electricity usage of each shaft, if possible. This is dependent on the level of metering available, which will be discussed in the next section.

3.3 Data collection and preparation

As discussed in Chapter 2, the *data collection and preparation* step is comprised of 5 sub-steps. These include system boundary selection, data acquisition, data cleaning, data transformation and data splitting. The aim of this section is to produce a clean, verified dataset that can be used confidently for forecasting model development, testing and validation. Each of these steps will be discussed below for each of the case study mining areas.

3.3.1 Selection of system boundary

As stated in Chapter 2, the selection of the system boundary is dependent on the level of detail of the electrical metering at the facility in question. Electrical metering is installed at each of the case study mines measuring the consumption of different systems, sub-systems and pieces of equipment. The overall electricity usage of each shaft and mine is consolidated on a monthly basis by a third-party company that verifies the usage against the electricity provider's invoices.

The purpose of the third-party metering company is to assist the mining group in allocating the electricity usage and cost to different operational mining units. This is necessary as the electricity provider charges the mine for the main electricity supply. This generally includes a variety of shafts, processing plants and auxiliaries, such as offices. Each specific operational unit is held responsible for the electricity usage that is calculated by the third-party provider.

Therefore, the system boundary and its location are dictated by the degree of separation offered by the third-party metering company. This electricity consolidation might only cover the overall

facility, individual shafts, different sub-systems or pieces of equipment. For the mines in question, the electricity consolidation is done on a per shaft and per compressor house basis.

That is, the total electricity usage of each shaft and each compressor house is measured and consolidated against the electricity provider's invoices. For the majority of the shafts, this includes all major sub-systems discussed in Chapter 1, namely ventilation, pumping, and hoisting. Compressed air is independently measured and split between the shafts for those that are on shared compressed air rings.

An exception to this generalisation is the ventilation electricity usage for shaft C, which is independently measured and consolidated. As such, system boundaries were drawn around each of the shafts as well as the compressed air electricity consumption of each shaft for all shafts. Additionally, a separate system boundary was drawn around the ventilation electricity usage for shaft C. Table 3.1 summarises the system boundaries and the systems they encapsulate for each of the case study mines. Green cells indicate the inclusion of the system within the system boundary while red cells indicate the exclusion of that system.

Table 3.1: Summary of system boundaries and associated sub-systems considered for case study mines

Mine	System boundary	Sub-systems				
		Hoisting	Pumping	Underground equipment	Ventilation	Compressed air
1	Shaft A					
1	Shaft A compressors					
1	Shaft B					
1	Shaft B compressors					
2	Shaft C					
2	Shaft C compressors					
2	Shaft C ventilation					
3	Shaft D					
3	Shaft D compressors					
3	Shaft E					
3	Shaft E compressors					

Table 3.1 outlines the system boundaries, and their associated systems, selected for the forecasting of the case study mines. Due to the level of metering available, a large number of the systems are grouped together under the overall shaft electricity usage. The splitting of these sub-systems would allow for more accurate modelling and management of electricity usage. This is, however, not possible with the level of available metering, which is considered to be representative in the industry [10].

3.3.2 Data acquisition

With the system boundaries defined, the data acquisition sub-step of the process was followed. This sub-step begins with the selection of variables, ensuring to include system, time-based and seasonal variables. Thereafter, the selection of the data resolution, and the calculation of the minimum number of data points required was performed. Due to the similarity of the systems included in the chosen system boundaries, the same variables, shown in Table 3.2, were considered universally.

Table 3.2: Variables selected for the electricity forecasting on the case study mines

Variable (Type)	Justification
Production (system)	Main driver of electricity usage. Future forecast available and authenticated. The only variable included in existing method will allow for the direct comparison of the two methodologies.
Temperature (seasonal)	Readily available data.
Average rainfall (seasonal)	Readily available data.
Number of week days (excluding public holidays) (time-based)	Represents the number of production days available to meet production requirements.

With the variables selected, it was possible to select the resolution of the data that could be used for the forecast. With the mines in question, the data resolution was determined by the resolution of the production forecast as well as the historical, authenticated electricity usage data. Both sets of data were only available in monthly intervals. This was the resolution as provided by the third-party metering company.

This is because the mine forecasts production in these intervals for planning purposes and finer resolutions are not manageable due to stockpiling as explained in section 1.2.1. Furthermore, the metering company only segments the electricity usage on a monthly basis as the electricity provider only invoices on a monthly basis. Therefore, the resolution of the forecast was determined by the resolutions of the input variables and was decided to be on a monthly basis.

Green/Harris's methods were then applied to determine the minimum number of data points required. This was necessary in order to ensure that the resulting forecasting model could be statistically significant. Since the minimum number of points is determined by the number of input variables, and the number of input variables for each of the shafts/systems was the same (3 variables), each system required the same minimum number of data points.

Using Equation (2.1), the minimum number of data points required was calculated to be 53 if 3 input variables are used in the forecast. It must be noted that this value will decrease by one for every input variable that is not considered in the modelling process. This will be discussed when variable correlation is calculated, and input variables are chosen to forecast the electricity usage.

With the resolution and minimum number of points known, the data was acquired from the relevant sources. It was necessary to obtain the data from official, authenticated sources. This is important to ensure that the resulting forecast is based on values considered to be official and representative by the mines in question. The details of the acquired data are illustrated in Table 3.3.

Table 3.3: Summary of data acquired for forecasting process

Data obtained	Source	Number of points	Source authenticated
Shaft electricity usage (all shafts)	Third-party metering company	67	Yes – Authenticated against provider invoices
Compressor electricity usage (all shafts)	Third-party metering company	59	Yes – Authenticated against provider invoices
Ventilation electricity usage (shaft C)	Third-party metering company	67	Yes – Authenticated against provider invoices
Production (all shafts)	Mining group reporting system	67	Yes – official reported production
Ambient temperature Average rainfall	Website ¹	67	Yes – Authenticated against other weather systems
Number of weekdays	Calendar	67	Yes – Obtained from official calendar

¹ "Rustenburg monthly climate averages," 2021. <https://www.worldweatheronline.com/rustenburg-weather-averages/north-west/za.aspx> (accessed Sep. 21, 2021)

Table 3.3 illustrates the sources, number and authentication status of the data obtained for the modelling process. All electricity usage data was obtained from the third-party metering company hired by the mining group. Production data was obtained from the mining group's official reporting system. The remaining input variable data was obtained from online weather sources, while the calendar data was obtained through intuition.

Table 3.3 indicates that only 59 points were obtained for the compressed air electricity usage of the shafts while 67 were obtained for the shaft electricity usage. This is because the splitting of the compressed air electricity usage was only started in September 2016, while that for the shafts was started in January 2016. As such, fewer data points were available for the modelling of the shaft compressed air electricity usage.

3.3.3 Data cleaning

The acquired data was then inspected and cleaned in order to ensure that the resulting dataset used in the modelling process was of high quality. During the data acquisition step of the process, the sources of the data were ensured to be valid and authenticated. This was done by ensuring that the data is measured using calibrated equipment and that the measured data is consolidated and agreed upon by stakeholders.

The production data is known to be accurate as the mine must ensure the consistent calibration of weighing systems for product tracking and control. Furthermore, the electricity usage data, obtained from the third-party metering company is consolidated against the electricity provider's invoices. This data is used for billing purposes and is considered official by the mining group.

The weather data, specifically the temperature and rainfall data, was obtained from a source that declares that their data is run alongside other weather recording systems to ensure accuracy². Furthermore, the calendar data, number of weekdays, was obtained from an official calendar. Intuition was used to extract the number of weekdays, excluding public holidays from the calendar – this information was not obtained from the mine explicitly, but is considered to be verified.

With the sources of the data known and verified, it possible to perform outlier detection and removal. As discussed in section 1.2.2, this is necessary as outliers can have a significantly negative impact on model performance. Outlier detection was done with the use of the

² "World weather declaration." <https://www.worldweatheronline.com/aboutus.aspx> (accessed Sep. 21, 2021).

Mahalanobis distance and the boxplot method, as discussed in section 2.3.3. The Mahalanobis distance was chosen as it allows the consideration of multiple variables simultaneously.

With the use of the Mahalanobis distance formula, it was possible to consider the distance of each data point relative to the whole dataset. The distance was calculated by considering all input variables as well as the target variable, electricity usage. The calculation of the distances reduces the data set to a single variable. The boxplot method, as described in section 1.2.2, was then applied to identify the outliers for each system boundary. Table 3.4 summarises the findings for each system.

Table 3.4: Summary of outlier detection for all systems

Shaft electricity	No. Outliers	Remaining data points	Other system electricity	No. Outliers	Remaining data points
Shaft A	3	(57) 64	Shaft A compressors	1	(51) 58
Shaft B	2	(58) 65	Shaft B compressors	2	(50) 57
Shaft C	3	(57) 64	Shaft C compressors	1	(51) 58
Shaft D	1	(59) 66	Shaft D compressors	3	(49) 56
Shaft E	6	(54) 61	Shaft E compressors	2	(50) 57
			Shaft C ventilation	1	(59) 66

Table 3.4 above indicates that outliers were identified for all systems in question. One of the outliers for all the systems fell in the month of April 2020 and was the result of the COVID-19 pandemic. During this month the facility closed and only ran the absolute minimum to prevent flooding and underground equipment damage. The other outliers were the result of distances that fell outside of the considered range. Shaft E's electricity usage distances and datapoints are shown in Figure 3.5 as an example.

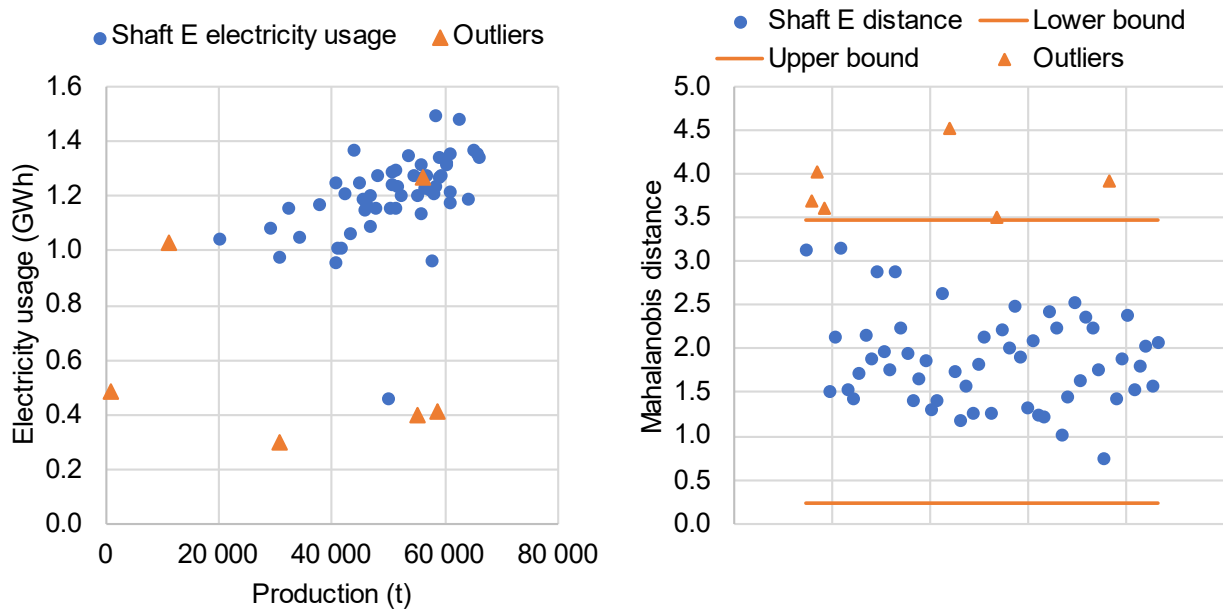


Figure 3.5: Shaft E electricity usage outliers and boxplot method

Figure 3.5 illustrates the electricity usage of shaft E plotted against its production, for visualisation purposes, on the left. On the right of Figure 3.5 are the associated Mahalanobis distances of each data point as well as the interquartile range boundaries as calculated from the boxplot method. Similar figures for the other shafts can be seen in the Appendix B.

Figure 3.5 illustrates that the data points with distances outside of the considered range were labelled as outliers. It must be noted that the other variables, such as rainfall and temperature, also affect these distances. The points were simply plotted against production for visualisation purposes. The outlier points were removed from the data sets in all instances. The datasets, cleaned of outliers using statistical methods, can now be confidently used for modelling.

3.3.4 Data transformation

The next step in data collection and preparation is data transformation. This is necessary due to the large disparity between the ranges of the data obtained. As explained in section 1.2.2, these disparities can lead to different variables having larger weighting in the final model. As a result, the data must be transformed such that multiple input variables can be used to forecast the electricity usage of each of the considered systems. Table 3.5 shows the ranges of the input data obtained for shaft A.

Table 3.5: Shaft A variable data ranges

Variable	Range	Difference between max and min of the range
Temperature	14 – 29°C	15°C
Rainfall	0 – 262 mm	262 mm
Weekdays	18 – 23 days	5 days
Production	99 460 – 292 922 t	193 462 t
Shaft electricity usage	7 406 065 – 11 768 431 kWh	4 362 366 kWh
Compressor electricity usage	5 527 097 – 9 731 011 kWh	4 203 914 kWh

Table 3.5 illustrates the ranges of each of the input variables considered as well as the ranges of the target variables. While temperature and number of weekdays vary by unit values, rainfall has a range of over 200 mm. The difference in range is even more extreme when production is considered, due to its range of over 190 000 t. These large disparities can lead to reduced model performance if all variables are considered for forecasting.

Furthermore, the values of electricity usage possess the largest range of all the variables, in excess of 4 000 000 kWh. The transformation of these variables was performed in order to ensure that the disparities in their numerical ranges did not negatively affect model performance. The transformation of the variables, however, reduces model interpretability, and forecasted values had to be rescaled to their original ranges for interpretation.

As explained in section 1.2.2, two methods of scaling exist, and the appropriate method was chosen based on the distribution of the data points for each variable. Variables with a Gaussian distribution were scaled using z-score normalisation, while those that did not display a Gaussian distribution were scaled using min-max scaling, as justified in section 1.2.2. The variable distributions for the variables of shaft A are shown in Figure 3.6.

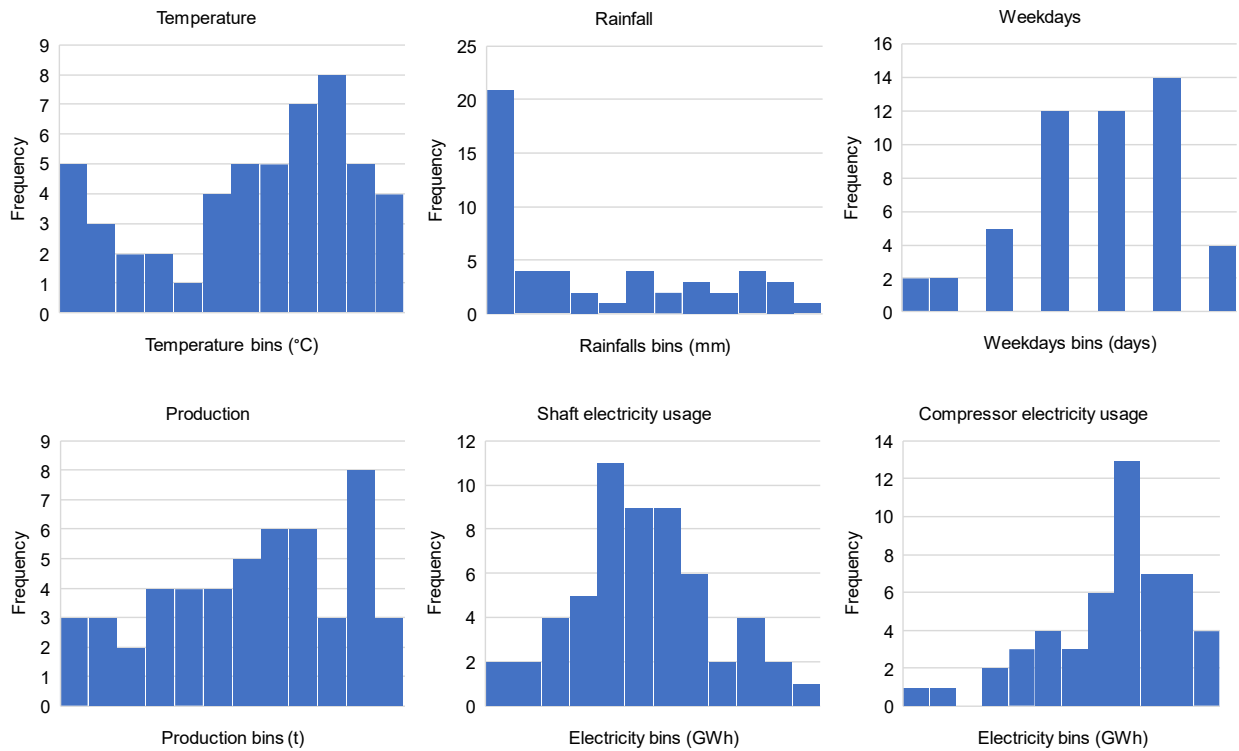


Figure 3.6: Frequency plots showing the distributions of variables considered for forecasting shaft A electricity and compressed air electricity consumption

Figure 3.6 illustrates the distributions for all considered input variables as well as the shaft electricity and compressed air electricity consumption of shaft A. Figure 3.5 illustrates that temperature, rainfall, production and compressor electricity consumption do not show Gaussian distributions. Conversely, the distributions of the weekday data and the shaft electricity consumption appear to be Gaussian.

As such, the temperature, rainfall, production and compressor electricity usage data were scaled using min-max scaling. Furthermore, the weekday and shaft electricity consumption datasets were scaled using z-score normalisation. A similar method was used for all other shafts in order to select the appropriate scaling method. The associated graphs and chosen scaling methods are shown in Appendix B.

3.3.5 Data splitting

The final sub-step of the data collection and preparation phase of the process is the splitting of the data into train/test and verification sets. This is necessary in order to ensure that the forecasting models are fit and tested on representative sets of data. Furthermore, the verification of the forecasting model on previously unseen data is required in order to ensure that the model's accuracy remains consistent with that seen in the training and testing sets.

Firstly, it must be noted that the most recent data, January 2021 – July 2021, was separated from the total dataset before data splitting was performed. As stated in section 2.6, the purpose of this is for the validation of the proposed forecasting methodology relative to the existing method. This validation was done by simulating the forecasting process for the first 7 months of 2021 and assessing the differences in accuracy between the two methods. This serves to further assess the performance of the resulting model on unseen data and to validate whether the proposed solution solves the research problem.

As explained in section 2.3.5, the splitting of the remaining data was first performed by randomising the dataset, which was performed with the *Rand()* function in Microsoft Excel™. Thereafter, 30% of the data was separated and retained as the model verification set. The remaining 70% of the data was used for model training and testing. The sizes of the verification sets were dependant on the number of data points that remained after the removal of outliers and the methodology validation set. The number of data points in each set for each system are shown in Table 3.6.

Table 3.6: Sizes of train/test and verification datasets for each system

Shaft electricity	Train/test set size	Verification set size	Other system electricity	Train/test set size	Verification set size
Shaft A	40	17	Shaft A compressors	36	15
Shaft B	41	17	Shaft B compressors	35	15
Shaft C	40	17	Shaft C compressors	36	15
Shaft D	41	18	Shaft D compressors	34	15
Shaft E	39	15	Shaft E compressors	35	15
			Shaft C ventilation	41	18

Table 3.6 shows the relative sizes of the verification and train/test datasets for each of the modelled systems. The randomisation of the data sets prior to splitting is imperative in order to ensure that the train/test and verification sets contain data that is representative. This will ensure that the accuracies obtained from both sets encapsulate the entire range of the data. The respective data points in each set are shown in Figure 3.7.

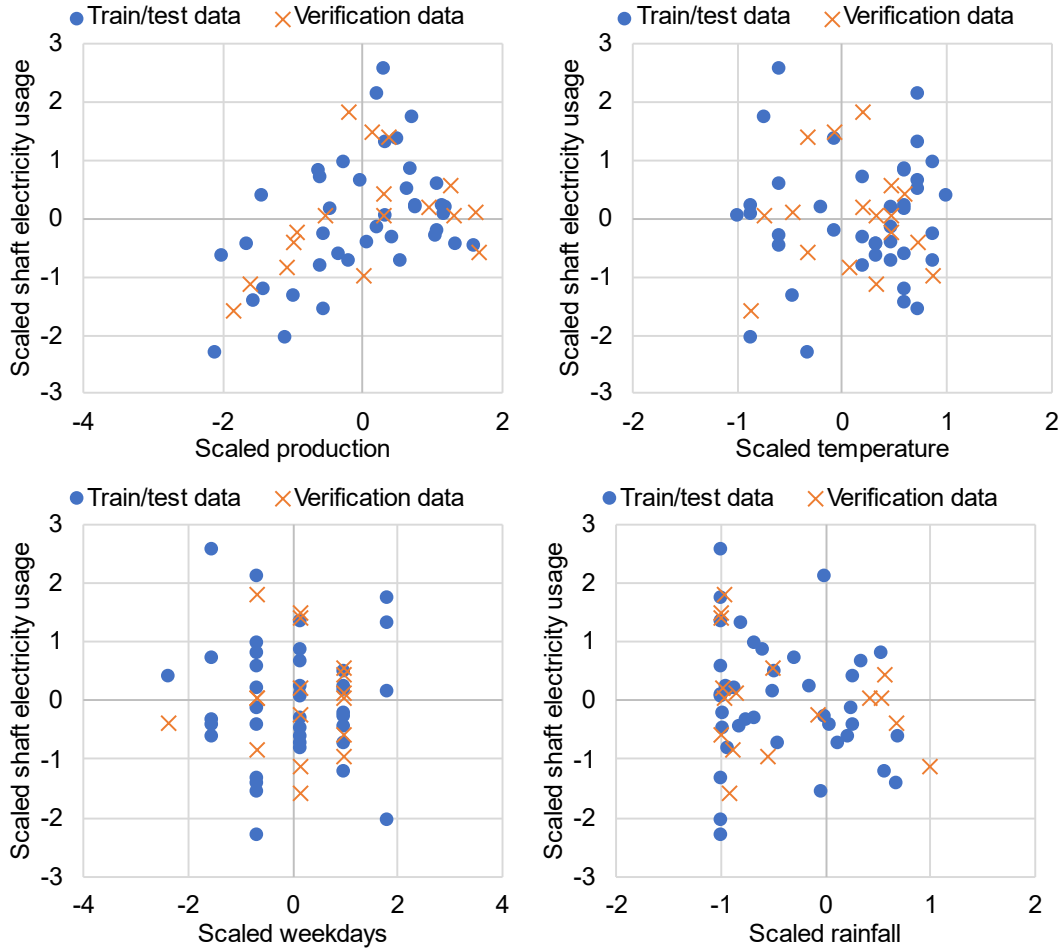


Figure 3.7: Data points contained in the train/test and verification sets for shaft A

Figure 3.7 illustrates the effect of randomisation, ensuring that the train/test and verification sets are comprised of data that is representative. This implies that the data covers the range of input and output variables for both sets. This is necessary in order to ensure that the accuracies obtained in the training and verification steps are representative of overall model performance. Figure 3.7 shows the sets for shaft A electricity usage and similar graphs for the other shafts can be found in Appendix B.

The extracted verification set was retained until the end of the modelling process in order to verify the accuracy of the final chosen model – to evaluate whether the modelling process was correctly implemented. The train/test set was then separated into separate folds that were used for k-fold cross-validation in the modelling process. Four folds were selected in order to ensure a 75:25 train/test split. The use of k-fold cross-validation was necessary in order to avoid bias in the separation of the test set. The data points used in each fold of model training for the shaft electricity usage of shaft C are shown Figure 3.8.

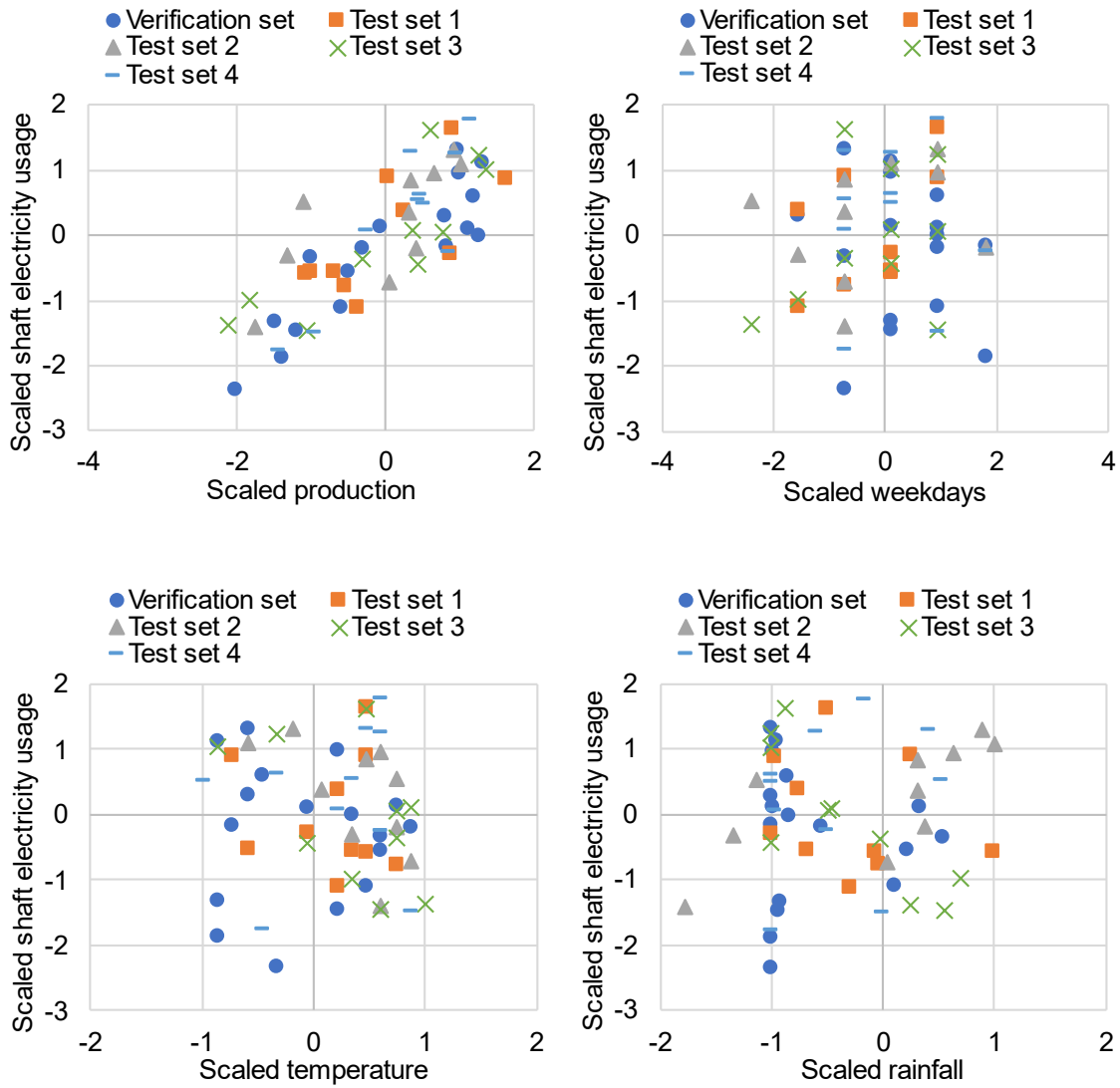


Figure 3.8: K-fold test splits for the scaled data used for the modelling of shaft C electricity consumption

Figure 3.8 illustrates the separation of the respective test sets from each of the folds used for the modelling of the shaft electricity consumption of shaft C. As with the separation of the verification set, the randomisation was necessary in order to ensure that each test fold was comprised of representative data. Figure 3.8 indicates that each of the test folds is comprised of data points that span the ranges of the input and target data points.

The model will be trained on 75% of the train/test dataset and tested on the remaining 25% (test set) of each respective fold. This will mitigate potential randomisation bias that might occur when only a single test set is used. Thereafter, it is possible to calculate summary statistics on model performance across folds. This allows for a more detailed analysis of model performance. This process explained in detail in section 3.4.3.

3.3.6 Summary

The system boundaries for the systems on each of the case study mines were identified. This allowed for the selection of variables and the acquisition of data from authenticated sources. The obtained data was then evaluated for outliers with statistical methods and outliers were removed in order to produce a dataset of high quality for modelling. The data was then split into 3 different datasets, a training/testing set for fitting the models, a set for the verification of the final selected model and finally a set for the validation of the proposed methodology.

Example graphs for each of these process steps were provided for different shafts. The detailed data collection and preparation process for each of the systems can be found in Appendix B. A summary of the resulting datasets for each of the systems is provided in Table 3.7.

Table 3.7: Summary of data acquisition and preparation process for all systems

System	No. data points	Outliers removed	Train/ test size	Model verification size	Method validation size	Complies with Green's minimum sample size
Shaft A	67	3	40	17	7	Yes
Shaft B	67	2	41	17	7	Yes
Shaft C	67	3	40	17	7	Yes
Shaft D	67	1	41	18	7	Yes
Shaft E	67	6	39	15	7	Yes
Shaft A compressors	59	1	36	15	7	Yes
Shaft B compressors	59	2	35	15	7	Yes
Shaft C compressors	59	1	36	15	7	Yes
Shaft D compressors	59	3	34	15	7	Yes
Shaft E compressors	59	2	35	15	7	Yes
Shaft C ventilation	67	1	41	18	7	Yes

Table 3.7 illustrates the summary of the data sets after the data preparation process. The table indicates that outliers were identified and removed from the data sets of all the systems. Furthermore, Table 3.7 shows that the resulting datasets comply with Green's method of minimum sample size. These datasets can now be used for the development of the forecasting models for each of the systems.

3.4 Model development

The following section will discuss the overall model development process of the proposed forecasting methodology. Results from the application of the various sub-steps of this process to

the electricity consumption of shaft C will be illustrated in detail. A summary of the results of the model development process for the other systems will be presented at the end of this section. Detailed results can be found in Appendix B.

3.4.1 Model type selection

As explained in section 1.2.2, the consideration of a variety of models of varying complexity levels and linearities is preferred, since this allows for the consideration of various relationships between system variables. Furthermore, it allows for the comparison and selection of the model that has the best fit to the data. This leads to a model that more accurately represents the data and thus leads to more accurate forecasts.

It was decided to use ordinary least squares regression in the proposed forecasting methodology. This is because it offers an increase in complexity over the current intensity model. Furthermore, in order to account for non-linear relationships between the variables, different functional relationships were also considered together with ordinary least squares regression. These include linear, polynomial and power relationships. These functions were selected by observing potential relationships between the variables as shown in Figure 3.9.

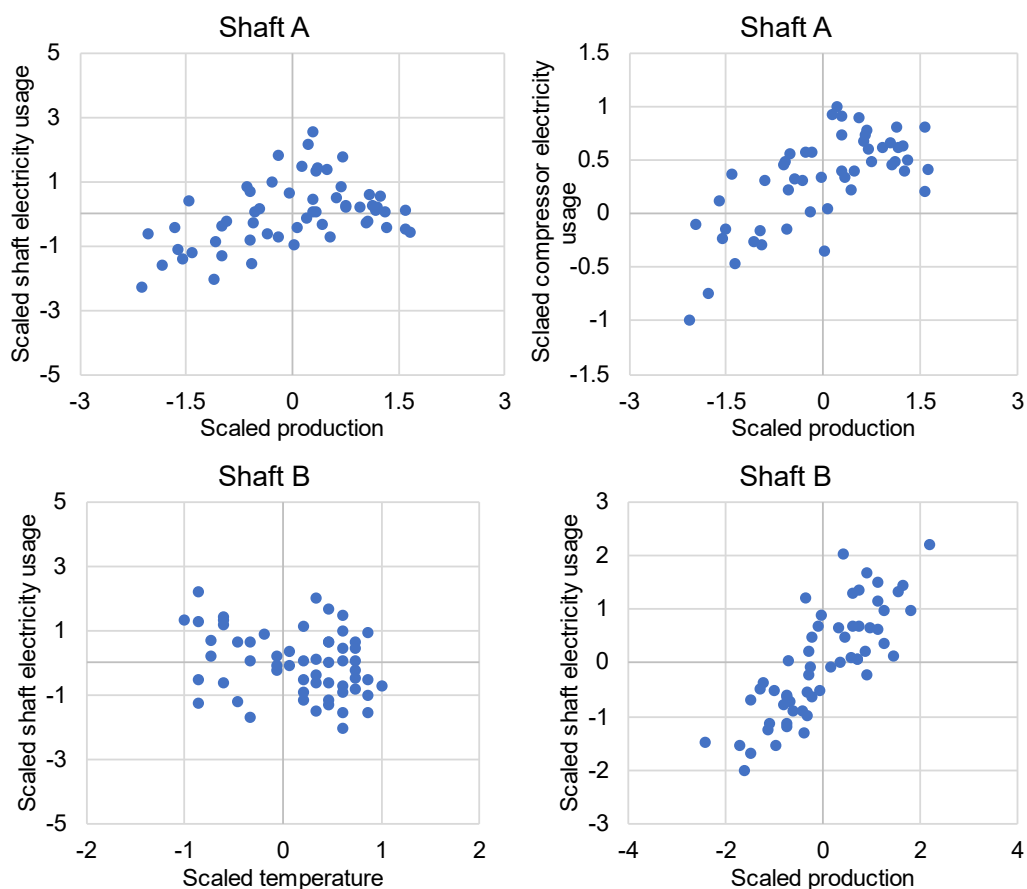


Figure 3.9: Example scatter plots between input variables and target variables for shaft A and shaft B

Figure 3.9 illustrates example scatter plots between input and target variables for shaft A and shaft B. The top two graphs illustrate that the relationships between shaft A electricity usage (left) and shaft A compressor electricity usage (right) against production are not linear. These relationships would be more accurately modelled by power or polynomial fits. Additionally, the bottom left graph, indicating shaft B electricity usage against temperature, does not seem to be correlated to temperature by any functional relationship. Finally, the bottom right graph, shaft B electricity usage and production, is more linear than the other relationships.

These insights allowed the selection of the specific functional relationships, linear, polynomial and power fits, to be specific. It must be stated that more advanced modelling processes might result in more accurate forecasts. The chosen models were limited to multiple functions and ordinary least squares regression to illustrate the benefit of a more complex model, the inclusion of non-linear relationships and the use of multiple models.

3.4.2 Variable correlation assessment

With the types of models selected, it was necessary to evaluate the variable correlation. As explained in section 1.2.2, it is of the utmost importance to only include input variables that have reasonable correlation with the output variable. This is due to the fact that the inclusion of variables with weak correlation might have a negative impact on model performance.

This was done by plotting scatter plots between each input variable and the target variable, as well as calculating Pearson's correlation coefficient. As shown in Table 1.8, variables with Pearson's correlation coefficients of absolute value less than 0.40 are defined as having weak to negligible correlation. It was, therefore, decided that only input variables with absolute correlation coefficients larger than 0.40 would be selected for model development.

If no variables had correlation coefficients larger than 0.40, all variables with correlation coefficients larger than 0.10 were selected. This, however, is not ideal as this indicates that the electricity usage is not even moderately correlated to any of the input variables. The scatter plots between all input variables and the shaft electricity usage of shaft C are shown in Figure 3.10.

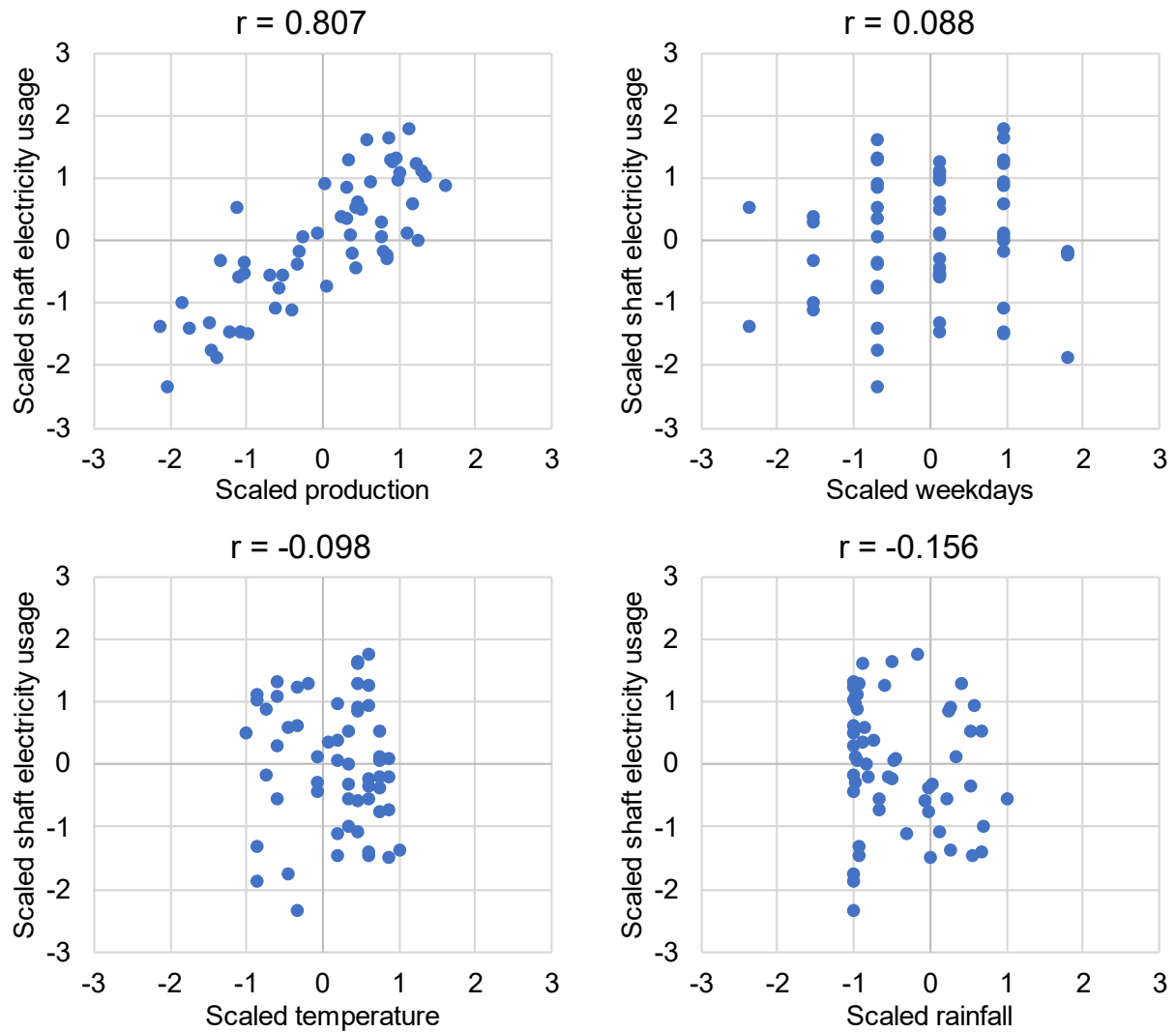


Figure 3.10: Scatter plots and Pearson's correlation coefficients between all input variables and the shaft electricity usage for shaft C

Figure 3.10 illustrates the relationships and correlation coefficients between all input variables and the shaft electricity usage of shaft C. The scatter plots and correlation coefficients indicate that the only variable that has strong correlation to the shaft electricity usage of shaft C is production. The remaining input variables all have absolute Pearson's correlation coefficients less than 0.40 and as such were not considered in the modelling process.

The same analysis was performed for all systems to be modelled and a summary of the results is shown in Table 3.8. In Table 3.8, the correlation coefficients of the variables selected for modelling each specific system are highlighted in green.

Table 3.8: Summary of Pearson's correlation coefficients for each input variable and each output variable to be forecasted

System	Pearson's correlation coefficient (r)			
System/Input variable	Production	Weekdays	Temperature	Rainfall
Shaft A	0.436	-0.037	0.032	-0.180
Shaft B	0.807	0.050	-0.292	-0.291
Shaft C	0.807	0.088	-0.098	-0.156
Shaft D	0.321	0.064	-0.279	-0.082
Shaft E	0.630	0.095	-0.135	-0.237
Shaft A compressors	0.691	-0.059	0.063	-0.208
Shaft B compressors	0.149	0.028	-0.224	0.021
Shaft C compressors	0.648	0.087	0.075	0.009
Shaft D compressors	0.550	0.067	-0.101	-0.173
Shaft E compressors	0.572	0.055	0.036	-0.289
Shaft C ventilation	0.404	0.172	-0.401	-0.432

Table 3.8 illustrates that production is the highest correlated variable to most of the output variables. Furthermore, in most cases it is the only variable that has a correlation coefficient larger than 0.40. As such, most of the forecasts were reduced to being univariate, except for the electricity consumption of shaft D, the shaft B compressors and the ventilation of shaft C.

For the electricity consumption of shaft D and the shaft B compressors, none of the input variables had higher than moderate correlation. As such, the variables with the highest absolute coefficients were selected – production and temperature in both cases. This is disadvantageous for these models but will be completed to illustrate the detrimental effect of poor correlation. Additional, more correlated variables should be investigated based on what is measured/forecasted at the selected mine. This was, however, not the focus of this study.

Finally, production, temperature and rainfall had absolute correlation coefficients above 0.4 for the ventilation electricity usage of shaft C. These variables were all chosen for the modelling and forecasting of this system.

3.4.3 Model fitting and testing

Using ordinary least squares regression in Microsoft Excel™, it was possible to fit the chosen functional relationships to the data. This step in the process was performed for all the systems in question, with some systems having a single input variable and others having multiple input

variables. The results from performing this step on the electricity consumption of shaft C are shown in this section.

As discussed in section 2.3.5, the training and testing of the model was done with the use of k-fold cross-validation. The training set, separated in section 3.3.5, was divided into four folds. The model was subsequently fit on three of the folds at a time and tested on the fourth fold. This was performed four times, changing the test set for a different fold each time. This ensures that randomisation bias is avoided when fitting the model. The power, linear and polynomial models fit and tested on the four folds for shaft C are shown in Figure 3.11.

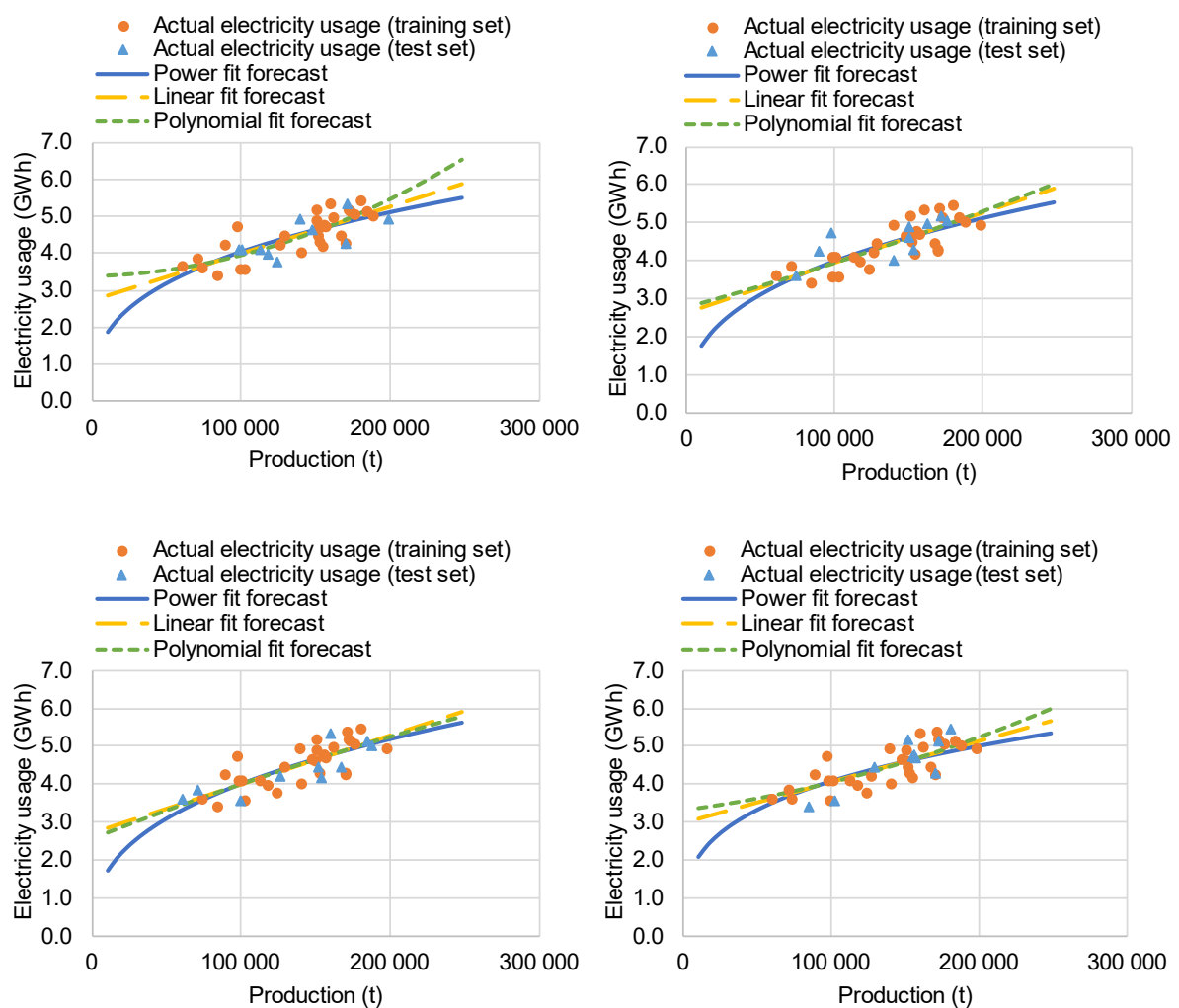


Figure 3.11: Power, linear and polynomial forecast models fit and tested on four folds for the prediction of the shaft electricity usage of shaft C

Figure 3.11 illustrates the resulting models, power, linear and polynomial, that were fit on the respective folds of shaft C. Each graph represents a different fold of cross-validation. It is important to note the variation in the testing set in each graph. Furthermore, the resulting models

are slightly different in each training fold. Particular variation is seen in the polynomial model which is concave in the bottom left graph and convex in the remaining graphs.

After the fitting of the model on each of the folds, four performance metrics were calculated on the resulting outputs of the model. These metrics were MAPE, R^2 , RMSE and MAE and were all used in order to ensure that metric bias was avoided when evaluating the models. Finally, the average and standard deviations of these performance metrics were calculated across the four folds. This allowed the selection of the model with the best overall training and test performance. These results for shaft C are shown in Table 3.9 with the best and worst values highlighted in green and red respectively.

Table 3.9: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft C

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	6.61	6.43	6.37	0.11	0.11	0.13
	Test	7.06	6.80	6.94	0.45	0.36	0.37
R^2	Train	0.597	0.610	0.613	0.031	0.035	0.036
	Test	0.527	0.541	0.526	0.110	0.125	0.130
RMSE (kWh)	Train	353 289	347 104	345 652	3 674	5 733	6 181
	Test	370 402	362 817	368 233	19 812	21 038	19 381
MAE (kWh)	Train	291 404	283 576	281 692	3 197	3 821	5 136
	Test	309 832	299 725	306 738	19 498	16 367	15 400

Table 3.9 illustrates the average and standard deviations for each of the performance metrics across the four folds with the best and worst values highlighted in green and red respectively. Models were selected based on their relative train and test set averages and standard deviations. As explained in section 1.2.2, models with low error on the training set and high error on the test set are said to be overfit to the training data and as such do not generalise well, while those with high train and test set errors are said to be underfit.

Firstly, it is important to note that the power model exhibited the poorest average performance across all metrics on both training and test sets. It did, however, have the lowest standard deviation in R^2 on both sets as well as in RMSE and MAE on the training set. Although this model was only slightly outperformed by the other models it was excluded from consideration in this case due to its poor average performance.

Secondly, when considering the average training set performance, the polynomial model slightly outperformed the linear model across all metrics. However, the linear model outperformed the polynomial model on the test set across all metrics – indicating that it generalises better to unseen data while the polynomial was likely overfitting the data. This is also clear when observing the polynomial models in Figure 3.11, with some folds having convex fits to the data.

The polynomial model exhibited the highest standard deviations for R^2 on both sets and for RMSE and MAE on the training sets. This indicates that although it had the best average performance measured by RMSE and MAE on the training sets it had the highest deviation across folds. This model did, however, have the lowest standard deviation in RMSE and MAE for the test sets accuracies. These are not low enough to justify the use of this model.

Finally, the linear model exhibited lower standard deviations than the polynomial model across most metrics and sets. This is with the exception of the RMSE and MAE on the test set, where the polynomial model slightly outperformed the linear model. It must be noted, however, that these differences in standard deviation are not large enough to justify the use of the polynomial model over the linear model.

The linear model was therefore selected to model the shaft electricity consumption of shaft C. This is due to its reasonable performance on the training set and high performance on the test set across all metrics. These performances indicate that the model generalises well to unseen data and does not overfit the training set. Its comparable standard deviations also indicate that it performs as consistently as the other models do across the training folds. The final linear model was fit to the total training set. The resulting model is shown in Figure 3.12.

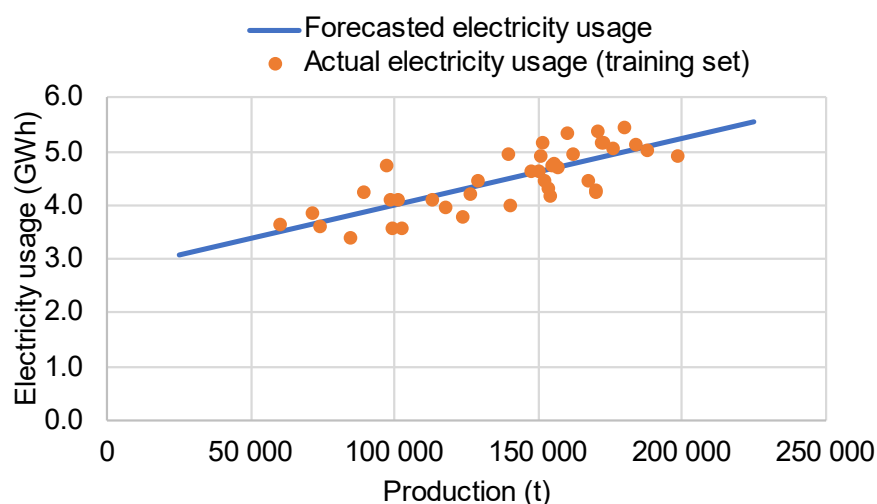


Figure 3.12: Final linear model fit to the total training set to model the electricity consumption of shaft C

Figure 3.12 shows the final linear model fit to the total training set. The model appears to fit the trend of the data well. The model has the following final training performance errors: MAPE of

6.50%, R^2 of 0.609, RMSE of 349 252 kWh and MAE of 286 373 kWh. These are comparable with the performance metrics obtained across the training folds. The final model describes the shaft electricity usage with the use of Equation (3.1).

$$E = 12.36P + 2\,762\,692 \quad (3.1)$$

where:

E Shaft electricity usage (kWh)

P Production (t)

The evaluation of Equation (3.1) leads to the insight that the relationship between production and shaft electricity is positive. Specifically, the electricity usage increases by 12.36 kWh per ton of production. Furthermore, at zero production, the shaft is expected to have an electricity consumption of 2.76 GWh. These relationships make intuitive sense as the shaft would use more electricity to mine more and would need to run basic equipment even if production is not occurring.

It must be noted, however, that extrapolating the modelled relationship is not recommended. That is, attempting to forecast far beyond the range of the training data would yield a higher error rate than forecasting within the training range. This is because the model has not been fit to data beyond this range and as such does not have a confident error rate for values beyond this range. Therefore, this model would have a 6.50% accuracy within a production range of between 50 000 t and 200 000 t approximately.

3.4.4 Model verification

The final step in fitting the model is model verification. In this process, the model, fit on the total training set, is evaluated on unseen data. This unseen data is the verification set that was separated at the start of the process. As explained in section 1.2.2, this step is necessary in order to ensure that the final model generalises well to unseen data and exhibits similar performance to that seen on the training set. The final linear model for shaft C plotted against the training and verification sets is shown in Figure 3.13.

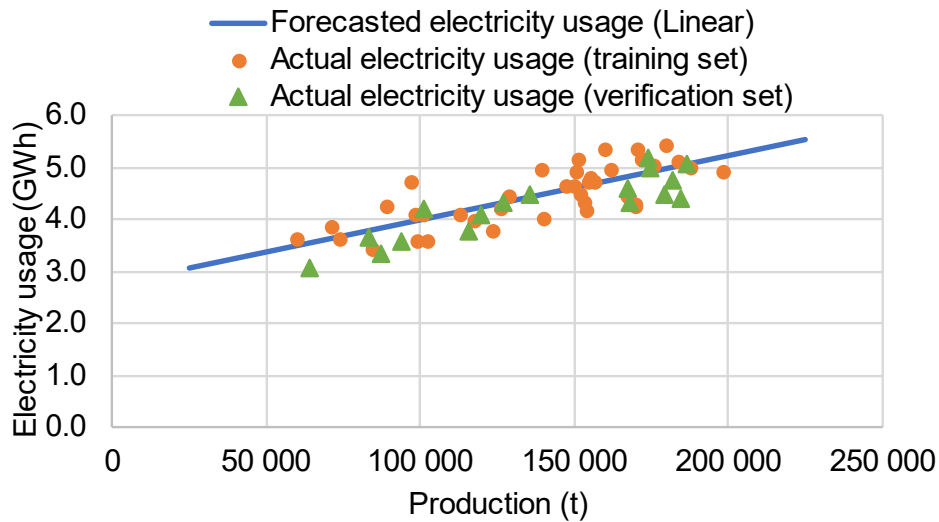


Figure 3.13: Final linear model for the forecasting of the shaft electricity consumption of shaft C plotted against the training and verification sets

Figure 3.13 illustrates the final linear model plotted against the training and verification set data. It is clear from the figure that the verification set spans the production range of shaft C and can therefore be considered to be representative. The performance metrics illustrating the performance of the final linear model on the training and verification sets are shown in Table 3.10.

Table 3.10: Training set and verification set performance metrics of the final linear model forecasting the electricity consumption of shaft C

Performance metric	MAPE (%)	R ²	RMSE (kWh)	MAE (kWh)
Training set	6.50	0.609	349 525	286 373
Verification set	7.02	0.664	340 361	280 239

Table 3.10 illustrates the performance metrics for the final model on the training and verification sets. Both sets are included for comparison purposes in order to indicate that the model performs with similar accuracy on both sets. The model shows slightly better performance on the verification set for R², RMSE and MAE and slightly poorer performance for MAPE. The model is therefore considered to have generalised well and performs with expected accuracy on unseen data.

Additionally to evaluating model performance on unseen data, several visual techniques were also discussed in section 1.2.2. These visual techniques are necessary in further evaluating model performance as well as validating model assumptions. They include 1) a scatter plot between actual and predicted values, 2) a frequency plot of the test and verification error distribution and 3) a residual plot. These plots for the electricity consumption forecast values of shaft C are shown in Figure 3.14, Figure 3.15 and Figure 3.16.

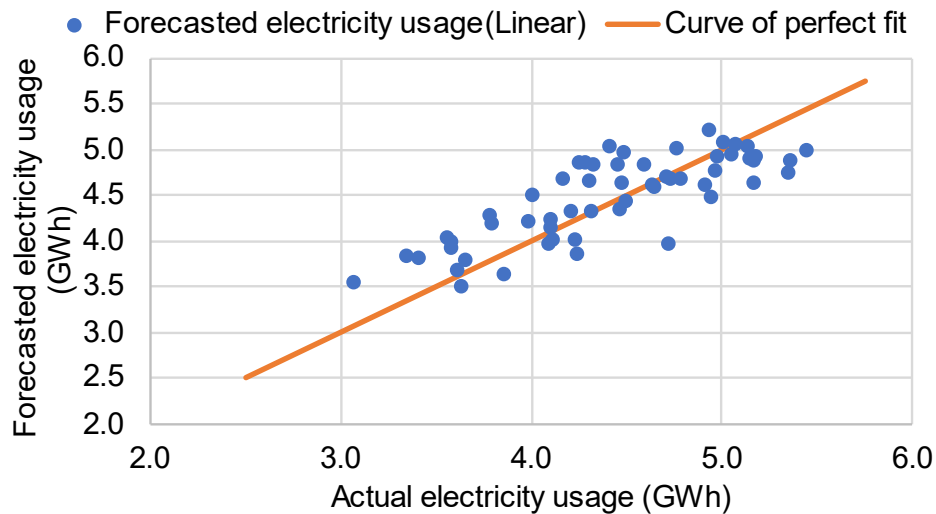


Figure 3.14: Scatter plot between forecasting and actual electricity consumption of shaft C ($R^2=0.807$)

Figure 3.14 is the scatter plot between the actual electricity consumption and forecasted electricity consumption values for shaft C. The orange line indicates the line of perfect fit: forecasted points that fall on this line are the same as the actual points that they were forecasting. Points above the line are higher in value to the actual points while those below are lower in value. The Pearson's correlation coefficient for this data is $R^2=0.807$.

Figure 3.14 firstly illustrates that the forecasted points are within a fixed error to the actual points, illustrated by the points surrounding the line of perfect fit. Secondly, it illustrates that most points fall above the line of perfect fit, meaning that the model tends to overpredict the electricity usage rather than underpredict. Finally, a Pearson's correlation coefficient of $R^2=0.807$ between the actual and forecasted points illustrates that the points have a strong correlation. This implies that the points are strongly related, indicating adequate model performance.

To further evaluate the distribution of the error between the actual and forecasted data, a frequency plot is necessary. This plot illustrates how frequently the model over/underpredicts the electricity usage and by how much. A rough estimation of this can be seen in Figure 3.14. Figure 3.15 illustrates the residual distribution for the electricity consumption of shaft C.

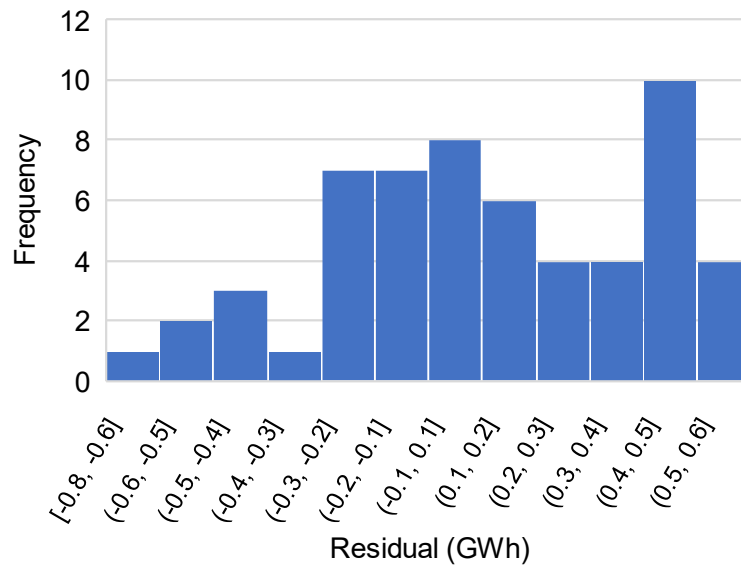


Figure 3.15: Frequency plot of residuals between actual and forecasted values of shaft electricity consumption of shaft C

Figure 3.15 illustrates that the residuals follow a normal distribution. This is with the exception of the high occurrence of error between 0.4 and 0.5 GWh. This satisfies the assumption of linear regression detailed in section 1.2.2, which states that the residuals have a normal error distribution. Figure 3.15 also illustrates that most of the residuals (with the exception of the outlier) are between -0.2 and 0.1 GWh. This indicates that the model has a fair number of points very close to the actual points (zero residual).

Finally, Figure 3.15 illustrates that there are more residuals above 0.3 GWh than residuals less than -0.3 GWh. This indicates that the model overpredicts the electricity usage more frequently than it underpredicts it. This same trend was seen in Figure 3.14. This insight is used to supplement the error metrics shown in Table 3.10. While these metrics indicate a RMSE of 3.4 GWh on the verification set, Figure 3.15 illustrates that this error is usually the result of an overprediction rather than an underprediction.

The final plot is used for the evaluation of another assumption of linear regression, the assumption of constant variance, or heteroskedasticity. This is a plot of the residuals against the forecasted values and is shown in Figure 3.16. For this assumption to be fulfilled there should be no relationship between the datapoints as discussed in section 1.2.2.

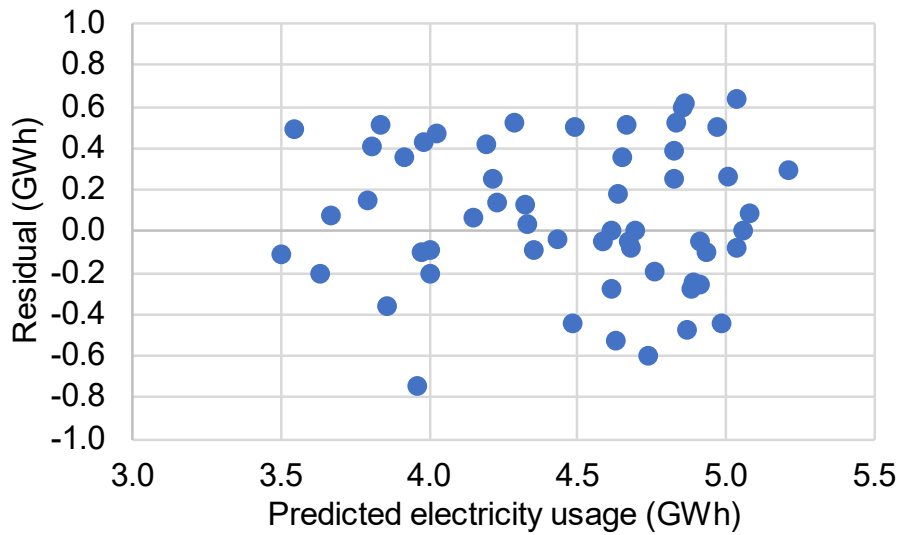


Figure 3.16: Scatter plot between forecasted electricity consumption values and residuals for the shaft electricity consumption of shaft C

Figure 3.16 illustrates the scatter plot between the residuals and the forecasted values. From the figure, no relationship exists between the datapoints. Furthermore, the residuals have a range between -0.8 GWh and 0.6 GWh across the entire range of prediction. This indicates that the residuals do not vary as a function of predicted value. Therefore, the assumption of heteroskedasticity is satisfied.

3.4.5 Summary

Models were fit to the cleaned and prepared data from the previous section. A range of different functional fits were applied together with ordinary least squares regression to fit the models. These models offer an improvement in complexity and a variation in terms of linearity. The range of input variables were filtered using Pearson's correlation coefficient. This was done in order to select only the variables that had the highest correlation with electricity usage.

The data was fit using ordinary least squares regression on linear, power and polynomial functional relationships across four folds of k-fold cross-validation. The average performance of the models was assessed across the four training and test folds. Four performance metrics were used to assess model performance, namely MAPE, R^2 , RMSE and MAE. This was done to avoid metric bias from skewing performance evaluation.

Each model was selected based on its overall performance as well as standard deviation across the four folds and performance metrics. The selected model, either linear, power or polynomial, was trained on the total training set. The resulting model's performance was evaluated on the verification set that was separated at the start of the process. Model verification was performed

with the same performance metrics used in the model training and testing phase as well as with visual techniques.

The utilisation of visual techniques allowed for the evaluation of the distribution of model as well as the assessment of whether the resulting model satisfied the assumptions of linear regression. The detailed process was performed on the shaft electricity usage of shaft C. A summary of the results obtained for all considered systems is shown in Table 3.11 where T is temperature in °C and R is rainfall in mm.

Table 3.11 illustrates a summary of the final models obtained using the proposed forecasting methodology as well as the percentage difference between the performance metrics obtained on the training and verification sets. This is calculated as the verification set metric minus the training set metric divided by the training set metric. Here, a negative value indicates that the metric on the verification set was lower than that on the training set, i.e. the verification set error was lower. The correlation coefficients of high performing models are highlighted in green, while those with very poor correlation coefficients are highlighted in red.

Table 3.11: Summary of final models and percentage difference between the performance metrics of the verification and training sets for all systems considered

System	Final model equation	Percentage difference verification and training performance metrics				Satisfies assumptions of linear regression	Pearson's coefficient between actual and forecasted values
		MAPE (%)	R ² (-)	RMSE (%)	MAE (%)		
Shaft A	$E = -1.2 \times 10^{-4}P^2 + 54P + 3\,612\,237$	-2.7	0.181	-19.25%	-37.67%	Yes	0.589
Shaft B	$E = 8.01P + 1\,927\,605$	0.27	-0.427	17.01%	9.05%	Yes	0.807
Shaft C	$E = 12.37P + 2\,762\,692$	0.52	0.055	-2.55%	-2.14%	Yes	0.807
Shaft D (scaled)	$E = 0.052P + 0.205T + 0.204$	-1.69	-0.343	9.46%	-10.35%	Yes	-0.113
Shaft E	$E = 8.12P + 803\,627$	0.48	-0.103	8.71%	-4.70%	Yes	0.631
Compressors							
Shaft A	$E = -6.5 \times 10^{-5}P^2 + 36.2 + 3\,875\,716$	0.72	0.328	-0.39%	-4.43%	Yes	0.753
Shaft B (scaled)	$E = 2.08P^{-0.139}T^{0.197}$	-0.57	-1.648	-5.25%	-4.64%	Yes	-0.263
Shaft C	$E = -6.18 \times 10^{-5}P^2 + 29.8P + 3\,750\,184$	-0.72	0.135	-3.18%	-8.03%	Yes	0.665
Shaft D	$E = -1.4 \times 10^{-4}P^2 + 49.04 + 1\,254\,854$	0.08	0.125	-6.58%	8.77%	Yes	0.657
Shaft E	$E = 1\,158P^{0.638}$	0.88	0.086	-9.67%	-8.05%	Yes	0.568
Ventilation							
Shaft C (scaled)	$E = 0.334P^{-0.020}T^{-0.256}R^{0.285}$	1.51	-0.334	17.09%	26.32%	Yes	-0.285

The first result indicated in Table 3.11 is the fact that a variety of different functions are applicable to the different systems. Specifically, a mixture of linear, power and polynomial fits performed the best on different systems. This proves that these systems are complex and can be described by more complex relationships between input variables in some instances.

It can be seen from Table 3.11 that most of the models performed with similar performance on the training and verification set. Specifically, most of the models had a deviation of less than 10% in RMSE and MAE between the training and verification sets. This is an acceptable deviation for the data that was available. There are instances where the models performed with lower error on the verification set than on the test set, which is most likely due to the distribution of the data in these sets. These models were not refit on account of their reasonable performance on both sets.

The largest deviation in performance is seen between the RMSE and MAE of the ventilation electricity consumption of shaft C as well as the RMSE of shaft B's electricity consumption. It must be noted that RMSE penalises large errors, unlike MAE. As result, the large deviation in verification set RMSE of the electricity consumption of shaft B is most likely the result of the distribution of this set and not the result of the model overfitting the training data.

Furthermore, the largest deviation, 26.32%, was seen between the MAEs of the ventilation electricity consumption of shaft C. This is an indication that the forecasted values on the verification set are far less accurate than those on the test set. This could be an indication that the model is overfitting to the training set. However, due to the overall poor correlation of the forecasted and actual values illustrated by the Pearson's correlation coefficient of -0.285, this cannot be the case.

The utilisation of a larger number of points for training and verification could result in a reduction in this deviation. Furthermore, it must be noted that the input variables for this model all had poor correlation to the output variables, which resulted in the expected poor predictability. The inclusion of a wider range of input variables might result in better overall model performance for the forecasting of the ventilation electricity of shaft C. This model was therefore not refit as a result of this large deviation and will be used as the final model.

Table 3.11 indicates that the correlation coefficients for most of the systems are larger than 0.55. This implies that the values forecasted by the final models are relatively close to the actual values. The correlation coefficients for some systems, those highlighted in green, have correlation coefficients larger than 0.75. This indicates that these models forecast the electricity consumption with a high degree of accuracy.

Finally, the systems that had input variables with low correlation to the target variables are highlighted in red. Due to this low correlation, the values forecasted by these models have negative correlation coefficients to the actual values. This indicates that the final models do not forecast the electricity consumption with a high degree of accuracy. This is, however, a result of having low correlation to input variables. This can be improved by the inclusion of more input variables in order to find a variable that is more highly correlated to the electricity consumption of these systems.

These final models can now be interpreted in order to define the operating range and conditions in which the abovementioned accuracies are applicable.

3.5 Model interpretation

With the final models selected and validated, it was possible to define the operating range and conditions in which these models are considered to be accurate. Furthermore, the evaluation of the model inputs allowed for the identification of the most influential variable affecting the electricity usage of each system. The results of this process for all the modelled systems are presented below.

The evaluation of the most influential variables, as discussed in section 2.5, was done by means of the comparison of the scaled regression coefficients. This was only applicable for systems in which multiple variables were considered. Systems in which the other variables were removed via correlation analysis imply that only the remaining variable was correlated enough to be significant. Therefore, only three of the modelled systems could be included in the analysis of the most influential variables. These systems are shown in Table 3.12 with the most influential variable's coefficient highlighted in green.

Table 3.12: Summary of the most influential variables for each of the modelled systems

System	Scaled regression coefficients		
	Production	Temperature	Rainfall
Shaft D electricity usage	0.052	0.205	N/A
Shaft B compressed air electricity usage	-0.139	0.197	N/A
Shaft C ventilation electricity usage	-0.020	-0.256	0.285

Table 3.12 contains the scaled regression coefficients for the three multivariate forecasting models. The table indicates that temperature is the most influential variable for the electricity consumption of shaft D and for the compressors of shaft B. Therefore, large changes in temperature have a more significant impact on the electricity consumption of these systems than the other variables considered.

Furthermore, Table 3.12 indicates that rainfall is the most influential variable impacting the electricity usage of the shaft C ventilation fans. This could be as a result of the humidity causing an increase in the density of the air. This can cause the fans to use more electricity to propel the denser air. Additionally, this table illustrates that production has a very low impact on the ventilation electricity usage of shaft C, which is understandable since these systems run continuously irrespective of production.

The second sub-step of the model interpretation process involves the definition of the operation range and conditions of each of the models. As discussed, this is important to ensure that the models are used within the appropriate limits. This ensures that the reported accuracies are correct and that the models are not extrapolating the ranges in which the operation of the systems might change. The operating ranges of each of the systems are shown in Table 3.13.

Table 3.13: Summary of the operating ranges of all of the modelled systems

	Input variable ranges (training data)		
System	Production (t)	Temperature (°C)	Rainfall (mm)
Shaft A	99 460 – 290 079		
Shaft B	45 392 – 154 094		
Shaft C	60 364 – 198 273		
Shaft D	46 250 – 212 800	15 – 28	
Shaft E	32 298 – 66 009		
Compressors			
Shaft A	104 407 – 290 370		
Shaft B	45 392 – 136 795	15 – 29	
Shaft C	60 364 – 198 273		
Shaft D	77 753 – 212 800		
Shaft E	28 916 – 66 009		
Ventilation			
Shaft C	60 364 – 198 273	15 – 29	0 – 220.2

Table 3.13 contains the input variable ranges for the data used for the fitting of the models of each system. These ranges of the input variables are considered to be the ‘valid’ operating ranges of the model. Any deviation from these ranges would imply that the model is extrapolating to some degree which would lead to forecasts of poorer accuracy. With the operating ranges defined, it is possible for users to apply the models with the understanding of which operating ranges will generate forecasts of the levels of accuracy reported above.

These final models will be used in the next section to validate the proposed forecasting methodology. A simulation will be performed whereby the electricity consumption for the first 7 months of 2021 will be forecasted using the proposed and existing methodologies. The results will then be compared objectively in order to assess whether the proposed method offers a significant accuracy improvement.

3.6 Methodology validation

The objective of this study is to objectively compare the results of the proposed methodology with the results of the existing forecasting method. This comparison will allow the determination of whether a more complex and statistical methodology results in an improvement in forecast accuracy. Results were generated by performing the forecast for the first 7 months of 2021 using both methods.

It must be noted that the input variables to these models must also be forecasted and can contribute to overall error. For this reason, the electricity forecast was performed using actual production values for the validation period in order to objectively evaluate the performance of the electricity usage forecasting methods and validate the resulting methodology. This was done as it was found that the MAPE for the production forecasts of shafts for the 7 months ranged from 7% for shaft A to 14% for shaft C.

The error between the forecasted values and the actual values was calculated using the four error metrics from the previous sections. These errors were compared to identify which method produced the most accurate forecasts. To further evaluate the error, a distribution plot of all residuals was generated. Finally, scatter plots between actual and forecasted values were generated to evaluate overall model performance.

The utilisation of these three methods will allow the results of the methods to be compared and for objective differences to be identified. Finally, they will allow the objective determination of the more accurate modelling process and the reasons for the differences in accuracy between the two methods to be identified. This will satisfy the final objective of the study and will validate the generation of a forecasting methodology that allows for statistical model selection and accuracy improvement.

3.6.1 Electricity intensity forecasting

For comparison purposes, it was necessary to generate electricity intensity-based forecasts using the method explained in section 1.2.1. These forecasts were generated with the use of the electricity and production data from 2020. However, the values for the month of

April 2020 - June 2020 were substituted with those from 2019 due to the COVID-19 pandemic which resulted in the closing of the facility.

The average 2020 electricity intensity was calculated for each of the systems identified in section 3.3.1. This average electricity intensity was then used with the actual production for 2021 to calculate the forecasted electricity for the relevant months. The average electricity intensities and resulting training errors calculated on the 2020 data used to fit the model for each of the considered systems are shown in Table 3.14.

Table 3.14: Performance metrics for the intensity models of each system calculated on the data used to fit the models

System	Final model equation	Performance metrics				Pearson's coefficient between actual and forecasted values
		MAPE (%)	R ² (-)	RMSE (%)	MAE (%)	
Shaft A	$E = 53.66P$	9.23	0.352	943 022	826 685	0.782
Shaft B	$E = 31.08P$	8.73	-1.05	254 060	218 117	0.684
Shaft C	$E = 36.27P$	15.41	-3.98	826 577	622 117	0.773
Shaft D	$E = 37.15P$	12.42	-0.603	762 456	642 686	0.501
Shaft E	$E = 24.41P$	6.30	0.310	112 370	81 447	0.283
Compressors						
Shaft A	$E = 46.44P$	8.35	0.288	785 175	662 050	0.806
Shaft B	$E = 36.01P$	13.93	-2.15	444 093	399 466	0.107
Shaft C	$E = 58.49P$	17.57	-1.07	1 407 660	1 161 353	0.586
Shaft D	$E = 39.68P$	9.72	-0.971	646 063	551 535	0.707
Shaft E	$E = 21.76P$	21.91	0.215	252 289	229 864	0.470
Ventilation						
Shaft C	$E = 30.81P$	21.60	-28.17	910 665	701 779	0.161

Table 3.14 illustrates the system intensities as well as the performance of the respective intensity models on the training sets. The table indicates that some of the models produce results with Pearson's correlation coefficients larger than 0.5, indicating reasonable forecasting accuracy. The errors illustrated in the table, however, have a significant range, with MAPE ranging from as low as 6.30% up to 21.91% which is inconsistent. This is compared with the MAPEs of the models fit using the proposed methodology which range from 3.70% to 15.07%.

Furthermore, this method does not allow the comparison of different models, which implies that there is no method to select or attempt to improve the accuracy illustrated in Table 3.14. This is a significant disadvantage because while intensity has the capacity to fit the system well, as seen with the electricity usage of shaft E, it is not universally applicable, which therefore results in the significant range in accuracies shown in Table 3.14.

3.6.2 Monthly forecast simulation and error comparison

The models that were fit in sections 3.4.3 and 3.6.1 were used to generate an electricity forecast for all the systems for the months of January 2021 – July 2021; this is an unseen validation set. The performance metrics for the respective methods were calculated and are shown in Figure 3.17.

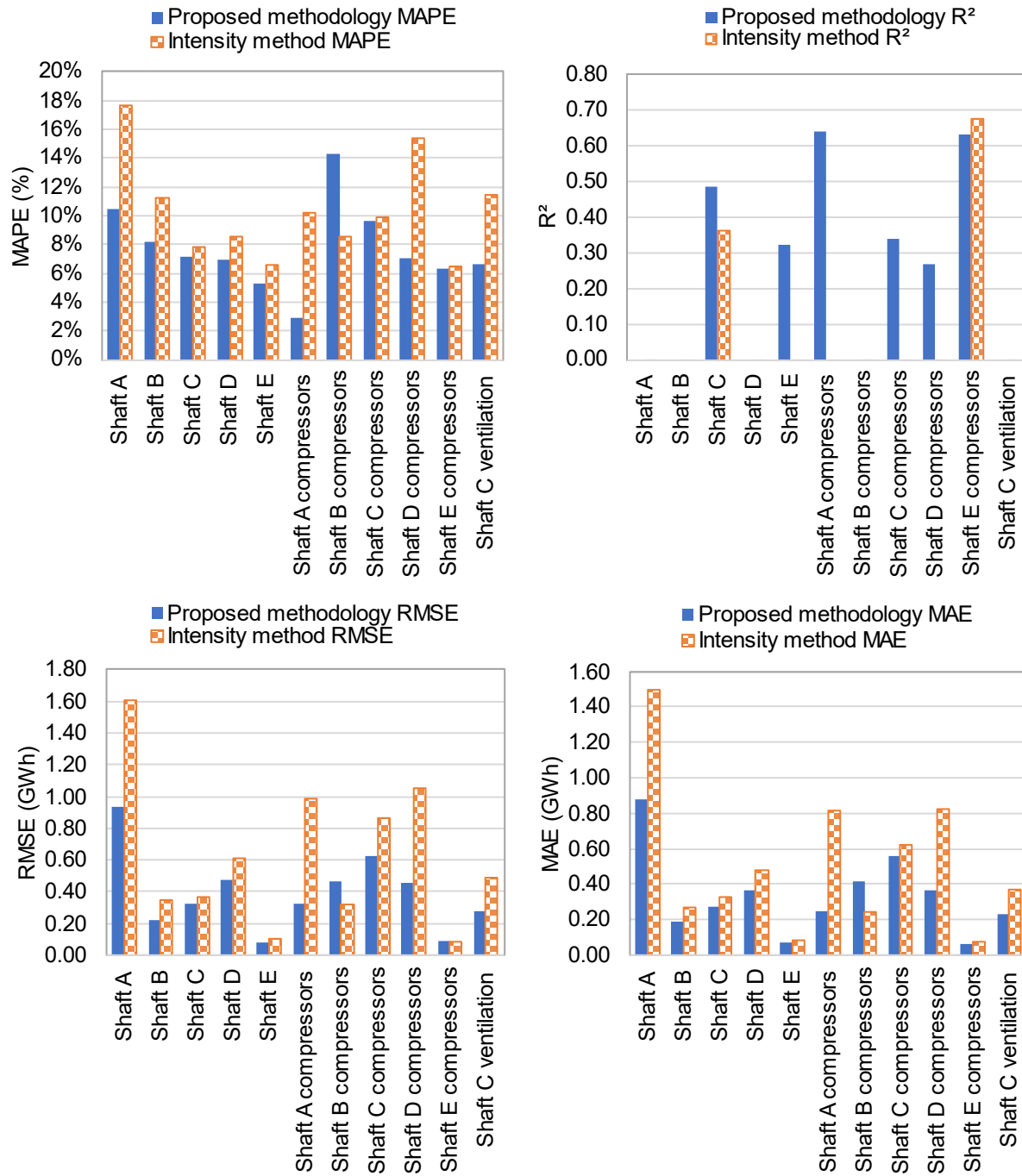


Figure 3.17: Comparison of the performance of the proposed models and the intensity models for all forecast systems on the validation set

Figure 3.17 illustrates the comparison of all the error metrics, MAPE, R², RMSE and MAE, for the intensity and proposed model forecasts of all systems. It must be noted that only positive values of R² were included in the figure as negative values are not comparable. Figure 3.17 illustrates the MAPE on the top left, R² on the top right, RMSE on the bottom left and MAE on the bottom right.

Firstly, it is clear from Figure 3.17 that the proposed models outperform the intensity models in terms of MAPE for all systems except for the shaft B compressor electricity usage. The proposed models also perform with a similar MAPE to the intensity models for the shaft C and shaft E compressors. Furthermore, the average MAPE for the proposed models is 7.69% compared with that of the intensity models which is 10.35%.

Secondly, Figure 3.17 illustrates that the proposed models produce a positive coefficient of determination more often than the intensity models. Additionally, the intensity models only produce a higher coefficient of determination than the proposed models for the shaft E compressors, a value of 0.67 compared to 0.63.

When considering the RMSE, a metric that penalises large errors, Figure 3.17 illustrates that the proposed models again outperform the intensity models on all of the systems except the shaft B compressors. These same insights are shown in the MAE graph at the bottom left of Figure 3.17 although the absolute values are slightly different.

Finally, Figure 3.17 illustrates that the proposed models outperform or at least perform just as well as the intensity models across all metrics for most of the systems. The only system in which the intensity models outperform the proposed models is the shaft B compressors. This is because the system's output variable had very poor correlation with the selected input variables. This clearly illustrates the need to identify and select appropriate, correlated input variables for this application.

Figure 3.17 indicated that the proposed models outperform the intensity models when considering average performance metrics. Further insight into the model performance can be drawn from a distribution of the residuals. This plot allows the evaluation of the range of error of each of the models. The residuals were calculated for all systems and are shown in Figure 3.18.

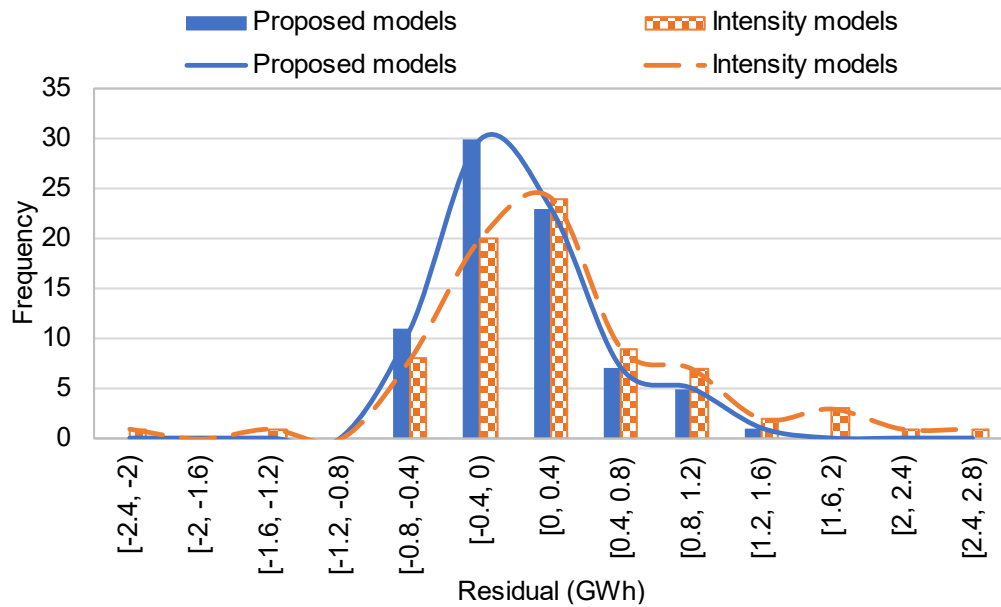


Figure 3.18: Residual distribution plot comparing the proposed model and intensity model residuals

Figure 3.18 illustrates two important insights into the errors of the two methods. Firstly, more of the forecasts, 53, made using the proposed models are between -0.4 and 0.4 GWh. This is compared with 44 forecasts when using the intensity method. This means that the errors for the proposed models are more often closer to zero. This implies that in general the error using the proposed method is smaller than when using the intensity method.

Secondly, Figure 3.18 illustrates that the range of the residuals when using the proposed models is much smaller than when using the intensity models. Specifically, the proposed models have a tighter residual distribution than the intensity models, and the range of residuals of the proposed models is [-0.8,1.6) GWh while that of the intensity models is [-2.4, 2.8) GWh.

This larger residual range implies that the intensity models are less consistent than the proposed models. Specifically, these models can produce significantly large error which is not preferred. This is a result of the ‘one-model-fits-all’ approach taken by the intensity method, leading to either models fitting very well or fitting very poorly to the data. This is dependent on the relationship between the input and output variables. With shafts that have poor correlation between input and output variables typically being forecasted with similar accuracy using the intensity method. This indicates that more focus needs to be applied on selecting appropriate input variables.

A final visual comparison of the forecasting methods is required in order to directly compare the forecasted values with the actual validation values. This is specifically a scatter plot between the forecasted and actual electricity consumption values for all systems. This was done and is shown in Figure 3.19. Note that the orange lines indicate the range of error for the intensity models, while the blue line indicates the range of error for the proposed methodology models.

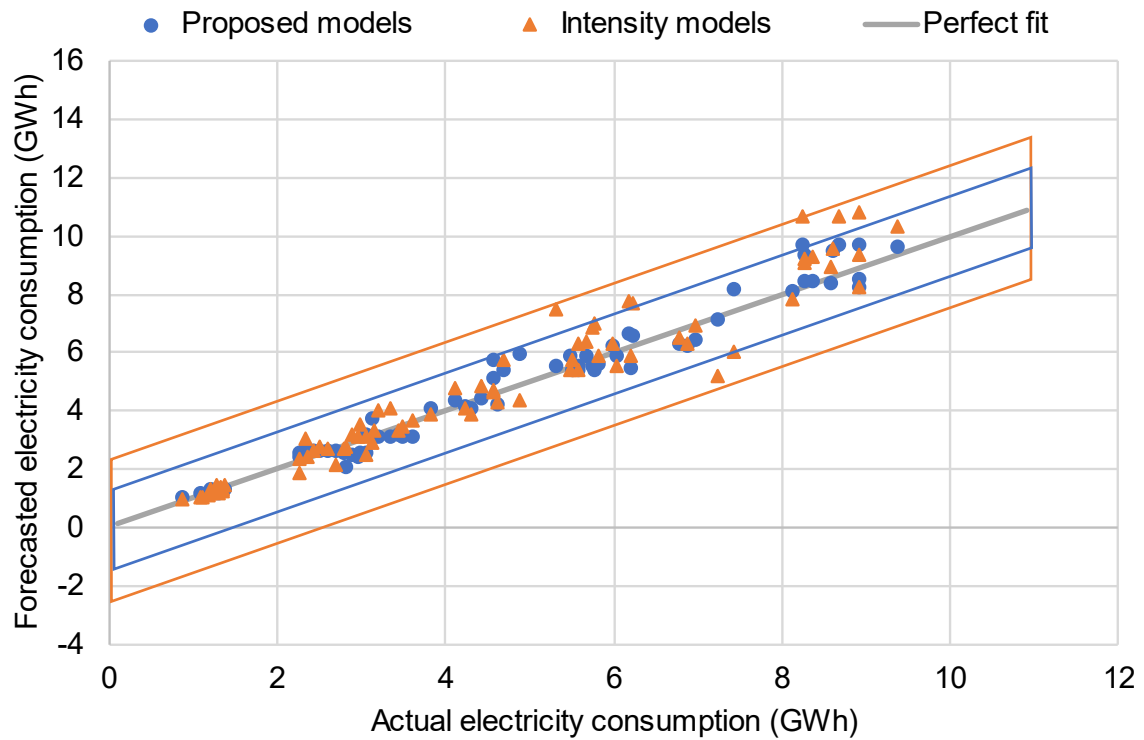


Figure 3.19: Scatter plot of forecasted and actual electricity usage for the intensity ($R^2=0.968$) and proposed models ($R^2=0.986$) of all systems

Figure 3.19 illustrates that both sets of models produce forecasts that closely follow the line of perfect fit. Specifically, the proposed models have an R^2 value of 0.986 compared to that of the intensity models which is 0.968. This indicates that these models generally produce more accurate forecasts for all values forecasted in the validation set.

Furthermore, Figure 3.19 illustrates that the spread of data for the different sets of models is different. Specifically, the set of intensity models have a much larger spread, 2.44 GWh, indicated by the orange lines. This is compared with that of the proposed models which is almost half, 1.43 GWh, indicated by the blue lines. This implies, similarly to Figure 3.18, that although the intensity models have the capacity to be accurate, they also have the tendency to more severely over/underpredict the electricity consumption.

This difference in the spread of data shown in Figure 3.19 can be used to illustrate the economic impact of an increase in forecasting accuracy. Using a model with a spread of 2.44 GWh (1.01 GWh larger than the proposed models) implies that any particular monthly electricity forecast per shaft can be an additional 1.01 GWh too high or too low. Using a tariff of R 0.82 per kWh in summer and R 1.21 per kWh in winter, this correlates to a potential misallocation of R820 000 in summer or R1 220 000 in winter per month for any particular shaft.

It is clear from the overall performance evaluation and from the two additional visualisation techniques that the proposed methodology produces models that can forecast the system electricity usage with greater accuracy. Furthermore, the resulting models have a smaller error range with the average MAPE and RMSE of 7.69% and 0.38 GWh with standard deviations of 2.84% and 0.24 GWh respectively. This is compared with 10.35% and 0.62 GWh for the intensity models with standard deviations of 3.36% and 0.44 GWh.

3.7 Conclusion

This chapter described how the methodology outlined in Chapter 2 was practically applied to obtain results that would satisfy the objectives defined in Chapter 1. The proposed methodology was verified in this chapter by applying it to three case study mines. These mines, although from the same mining group, were of different ages and depths and operated on different reefs and compressed air rings.

Detailed results for all systems contained on the case study mines are presented in Appendix B. This chapter only serves to illustrate the application of the proposed methodology, highlight specific insights and to validate that the proposed method satisfies the outlined objectives.

The *identification of model purpose* stage of the process involved the contextualisation of each of the selected mines. This was followed by outlining the outcomes of the forecasting models. In the case of the mines in question this outcome was the monthly forecasting of electricity usage for the purpose of budget development.

For the *data collection and preparation* phase of the process, system boundaries were chosen for all the mines in question. This was followed by detail on data acquisition and data cleaning for all the chosen systems. Parts of the process for specific systems were shown in detail in order to highlight important insights, particularly with outlier identification and removal, data transformation and data splitting.

The development of high-quality datasets in the previous section allowed for the fitting of forecasting models to all systems. Ordinary least squares regression was used with the application of different functions, specifically linear, power and polynomial functions. The results of the modelling of the electricity consumption of shaft C were presented in detail. A summary of the results for all systems was also presented indicating that all models performed with reasonable accuracy and satisfied initial assumptions.

These models were fit using k-fold cross-validation on a training set and verified on an unseen verification set. This was done to confirm that the chosen models performed with similar accuracy on both sets to prevent over/underfitting to the training data. All models performed consistently,

with verification error within 10% of the training error, except for a single model whose poor performance was a result of poor variable correlation.

The final models were then interpreted in the *model interpretation* stage of the process. This involved the identification of the most influential variables in systems that were multivariate. Furthermore, the operating conditions and ranges of each of the models were identified and defined from the training data. This was done to ensure that the models are not used outside of their valid ranges which would result in extrapolation and a reduction in performance.

The validation of the proposed methodology was then performed by means of a simulation. This simulation involved the forecasting of the electricity consumption of the mines systems for January 2021 – July 2021 using the proposed and existing method. This data was previously unseen to either set of models. The results of these forecasts were then objectively compared to investigate which set of models produced forecasts with the highest accuracy.

These models were objectively compared by means of four error metrics as well as two visualisation techniques. The comparison of error metrics provided a general indication of relative performance while the visualisation techniques provided insight into the distribution of the error of each set of models.

It was found that the proposed models generally produced forecasts with lower error than the intensity models across all performance metrics. Specifically, the models across 11 systems produced an average (standard deviation) MAPE and RMSE of 7.69% (2.84%) and 0.38 GWh (0.24 GWh) for the proposed models. This is compared with 10.35% (3.36%) and 0.62 GWh (0.44 GWh) for the intensity models. This is a 37.5% reduction in average RMSE.

The visualisation techniques illustrated the same insight, specifically that the proposed model forecasts were generally closer to the actual values and that they had a smaller error range than the intensity models. Therefore, validating that the proposed methodology produces models that are capable of forecasting mine electricity usage to a greater degree of accuracy than the current method.

CHAPTER 4 CONCLUSIONS AND RECOMMENDATIONS

4.1 Introduction

This study was conducted to address the oversimplification of deep-level platinum mine electricity forecasting. The need to develop a forecasting methodology capable of offering accuracy improvement to deep-level platinum mine forecasting was addressed. The focus of this study was to address this defined need by satisfying various objectives that were outlined in Chapter 1.

A review of literature was performed in Chapter 1 in order to identify the processes used for model development that are currently used in industry. This knowledge allowed for the development of a general modelling procedure in Chapter 2 that can be applied to deep-level mine electricity forecasting.

Chapter 3 allowed for the application of the proposed modelling procedure to three case study mines and allowed for the generation of eleven validated models. The accuracies of these models were then compared with those produced using the existing method in order to validate the proposed methodology.

The conclusion to the study is provided in this chapter. Section 4.2 and Section 4.3 will provide an overview of the study and detail how the study objectives were met. Section 4.4 will outline proposed recommendations for further studies. Finally, the study will be closed in Section 4.5.

4.2 Overview of study

This study found that the existing method used for forecasting the electricity usage on deep-level platinum mines is overly simplistic. Due to the complex nature of these systems, this over-simplification leads to a reduced forecasting accuracy that can adversely impact planning and energy management. The need to create a more robust modelling process that incorporates statistical rules was therefore highlighted.

A general procedure for modelling a system as a function of its input variables, however, does not explicitly exist. As such, it was necessary to extract a general procedure from those applied to electricity forecasting and prediction in other industries. This proposed general procedure was ultimately applied to the electricity forecasting of a platinum mining group. The results were compared with those generated by the utilisation of the existing method on the same mining group.

A literature review was performed in Chapter 1 in order to extract a general process for the modelling of a system as a function of its inputs. This allowed for the generation of a proposed

forecasting method in Chapter 2. This method was comprised of four main steps: *identification of model purpose, data collection and preparation, model development and testing and model interpretation*. Each of these main steps was comprised of various sub-steps used for the generation of electricity forecasts.

This proposed forecasting method was applied to three case study mines in Chapter 3. For each case study the data was acquired and processed to yield a high-quality, transformed dataset. Each case study mine was comprised of a variety of sub-systems that were individually modelled as the data was available. This led to the generation of eleven electricity forecasting models for the different sub-systems.

Each system was modelled using ordinary least squares regression with the application of linear, power and polynomial functions to account for both linear and non-linear behaviour. The most accurate model for each system was selected with the use of a training/test set and k-fold cross-validation. The model's performance was then verified on an unseen verification set that was separated at the start of the process.

The models created using the proposed methodology were used to generate a simulated electricity forecast for the three case study mines for January 2021 – July 2021. A similar forecast was performed using the existing forecasting method. The accuracies of these forecasts were then compared in order to validate the proposed methodology. This allowed for the quantitative evaluation of the study objective stating that the new methodology should offer accuracy improvement.

It was concluded that the proposed forecasting methodology outperformed the existing method across all considered performance metrics for all systems but one. The outlier system was one with very poor correlation between the input and output variables, which explains the poor performance. The application of the proposed method reduced the average MAPE from 10.35% to 7.69% and the average RMSE across systems from 0.62 GWh to 0.38 GWh which is a 25.7% reduction in MAPE and a 37.5% reduction in RMSE.

Two visualisation techniques were applied in order to evaluate the error distributions of the respective methods. The same conclusion as above was drawn from these techniques, specifically that the proposed methodology models produced forecasts of lower error as well as lower error range. This implies that the existing method, although sometimes accurate, is not as consistent as the proposed methodology and can produce significantly large errors.

The results from this study indicated that the proposed forecasting methodology could generate forecasts of higher accuracy than the existing method.

4.3 Satisfying objectives

The primary objective of this study was to develop an improved modelling process that could be applied to the electricity usage forecasting of deep-level platinum mines. This process had to allow for the comparison of different mathematical models with accuracy-based criteria to ensure the selection of the most applicable model. For this to be achieved, the following objectives had to be met:

- extract a general process flow for model development from relevant literature,
- provide relevant research on the specific techniques applied in general process flow,
- develop a proposed strategy for the electricity consumption forecasting of deep-level platinum mines,
- verify the proposed strategy by applying it to three case study mines,
- objectively compare the results of the proposed strategy with those of the existing method to evaluate accuracy improvement.

Each of these sub-objectives and how they were achieved will be discussed below.

4.3.1 Extraction of modelling process flow from literature

The current industry processes for forecasting or predicting electricity usage as a function of input variables were investigated in Chapter 1. This allowed for the generation of the general modelling process flow presented in section 1.2.2. This general modelling process was compared with the existing forecasting process applied on deep-level platinum mines in section 1.2.3.

4.3.2 Research of techniques applied in general modelling process

The literature review performed in Chapter 1 led to the identification of a wide variety of statistical data preparation and modelling methods used in various industries. The theory and application of these methods were discussed in Chapter 1. This was done in order to compare different methods for the identification of relevant advantages and disadvantages. This allowed for the selection of the most applicable methods during the development of the proposed methodology.

4.3.3 Development of a strategy for deep-level platinum mine electricity usage forecasting

The proposed strategy for electricity usage forecasting of deep-level platinum mines was developed in Chapter 2. This strategy was developed based on the literature review performed in Chapter 1. The proposed strategy was comprised of four main steps, namely *identification of model purpose, data collection and preparation, model development and testing* and

model interpretation. Each of these steps consisted of a variety of sub-steps that were followed to allow for the generation of verified forecasting models.

4.3.4 Verification of proposed strategy with case studies

The proposed forecasting strategy developed in Chapter 2 was applied to three case study mines in Chapter 3. The relevant results of the forecasting process and each case study were discussed in Chapter 3. The detailed results of each specific system are presented in Appendix B. The application of the proposed methodology led to the generation of eleven forecasting models which were verified for consistent accuracy across training and unseen verification datasets.

4.3.5 Objective comparison of proposed strategy to the existing forecasting method

Section 3.6 of Chapter 3 addressed the validation of proposed forecasting methodology. This validation was performed by the objective comparison of the proposed forecasting models with those created using the existing method. This was done by means of forecast accuracy comparison as well as with the utilisation of visual techniques for comparative error evaluation. It was found that the proposed methodology offered significant improvements in accuracy over the existing method.

4.4 Recommendations for further research

This section details recommendations for possible further studies. Seven recommendations are listed and discussed in detail below. The investigation of these recommendations could result in further improvement to the results of this study.

- Inclusion of mine shift data as an input variable
- Inclusion of time-series analysis to electricity forecasting methods
- The utilisation of more granular data
- Application of the forecasting process to other mining groups such as gold mines
- Investigation into the independent variables effecting mine electricity usage
- Addressing how models can be kept accurate and representative of the dynamic mining systems
- Improvement and evaluation of mine production forecasting

4.4.1 Inclusion of mine shift data as an input variable

It was found in the study that most of the systems electricity usage had the highest correlation with production. Although advantageous for monthly modelling of the system, the disadvantage

of this is the potential stockpiling of ore, which would prevent modelling on a more granular time scale, weekly or daily.

Since people directly consume the electricity on the mine and are responsible for production it is expected that their presence would be correlated to the electricity usage. The utilisation of shifts schedules would allow the potential to forecast on a more granular time scale. Furthermore, the inclusion of this input variable might lead to an overall improvement in model accuracy.

4.4.2 Inclusion of time-series analysis to electricity forecasting methods

This study focused on the utilisation of regression models for the forecasting of electricity usage. These models focus on providing the model with known input and output values. The application of time-series analysis as a forecasting method would not require the explicit inclusion of input variables as these models identify trends in time-series data. Since mine electricity consumption is provided as a time-series, this method could produce results of reasonable accuracy, provided the data resolution is granular enough.

4.4.3 The utilisation of more granular data

A limitation of forecasting on a monthly basis, as seen in this study, is the low granularity of the data. This low granularity limits the complexity of the models as well as the number of input variables that can be used. The application of the proposed forecasting methodology to systems that have more granular data, weekly, daily or even hourly, could result in accuracy improvements.

Furthermore, the capacity to forecast electricity usage on small time-scales allows for detailed resource planning and energy management. This application might lead to a system that allows mines to manage their electricity by performing daily forecasts and comparing their actuals against the values predicted by the system.

4.4.4 Application of the forecasting process to other mining groups

This study focussed on the development of a forecasting methodology for the electricity forecasting of deep-level platinum mines. The application/adaptation of this methodology to other types of mines, such as gold or coal, might lead to forecast accuracy improvement for those systems. Additionally, insights drawn from those systems could offer meaningful improvements to the proposed forecasting methodology.

4.4.5 Investigation into the independent variables effecting mine electricity usage

This study aimed at producing a forecasting methodology to improve the accuracy of electricity usage forecasting on deep level platinum mines. This method used independent variables found in literature with a deeper focus on the modelling component of the forecast. It is recommended that a more detailed study is performed on the identification of the variables most highly correlated to electricity usage which are easily measured and forecasted.

4.4.6 Addressing how models can be kept accurate and representative of the dynamic mining systems

In this study it was decided not to address the continuous monitoring and maintenance of models as focus was applied on the development of the forecasting methodology. The investigation of methods to monitor model accuracy, assess model performance and update models would significantly improve the practical application of these models. Furthermore, this would potentially allow these models to be used for more granular forecasting and planning.

4.4.7 Improvement and evaluation of mine production forecasting

It was found that production was the main input variable to the resulting models. Although these models focussed on improving the electricity forecasting process, error in the production forecast can invariably reduce electricity forecasting accuracy. As a result, improving the upstream process will allow for the reduction of overall error.

4.5 Closure of study

In this study an improved modelling process was proposed for the electricity usage forecasting of deep-level platinum mines. This was achieved by reviewing relevant literature and extracting the general process for modelling a system as a function of its inputs in industry. Thereafter, the theory for the relevant techniques applied in industry was investigated and compared. A well-defined problem statement and clear objectives were identified and provided to contextualise the deliverables of the study.

With the objectives defined and the relevant theory investigated, it was then possible to propose a methodology for the electricity usage forecasting of deep-level platinum mines. This methodology was verified by applying it to three case study mines. The methodology was then validated by objectively comparing its results with the results produced by using the existing forecasting method. It was finally found that the proposed method offers significant accuracy improvement compared with the existing method.

This chapter concluded the document and the study. This was done by providing an overview of the study, outlining how the study objectives were achieved and by providing a list of recommendations for further studies.

REFERENCES

- [1] Eskom Holdings Limited, "The energy efficiency series: Towards an energy efficient mining sector," 2010. [Online]. Available: http://www.eskom.co.za/sites/idm/Documents/121040ESKD_Mining_Brochure_paths.pdf %5Cnwww.eskom.co.za/dsm.
- [2] G. Ryan, "Energy management by platinum companies in South Africa: exploring mitigation strategies for reducing the impact of electricity price increases," in *The 6th International Platinum Conference, "Platinum-Metal for the Future,"* 2014, pp. 299–308.
- [3] C. Cilliers, "Benchmarking electricity use of deep-level mines," Ph.D. dissertation, Dept. Mech. Eng., North-West University, Potchefstroom, 2016.
- [4] V. F. N. Torres, G. R. Mateus, A. G. Martins, W. Carneiro, and L. S. Chaves, "Integrated optimization and simulation models for short-term open-pit mine planning," *J. South. African Inst. Min. Metall.*, vol. 120, no. 11, pp. 617–626, 2020, doi: 10.17159/2411-9717/1266/2020.
- [5] M. S. Q. Domingues, A. L. F. Baptista, and M. T. Diogo, "Engineering complex systems applied to risk management in the mining industry," *Int. J. Min. Sci. Technol.*, vol. 27, no. 4, pp. 611–616, 2017, doi: 10.1016/j.ijmst.2017.05.007.
- [6] A. Hassan, K. Ouahada, T. Marwala, and B. Twala, "Optimization of the compressed air-usage in South African mines," in *IEEE Africon*, 2011, pp. 13–15.
- [7] I.-S. Rakos and M. Man, "Cost budgeting – A relevant instrument for improving the performances of a coal mine case study: Jiului Valley, Romania," *J. Account. Audit. Res. Pract.*, vol. 2016, no. 2016, pp. 1–22, 2016, doi: 10.5171/2016.920413.
- [8] S. C. Hansen and W. A. Van der Stede, "Multiple facets of budgeting: An exploratory analysis," *Manag. Account. Res.*, vol. 15, no. 4, pp. 415–439, 2004, doi: 10.1016/j.mar.2004.08.001.
- [9] A.-M. Ríos, M.-D. Guillamón, B. Benito, and F. Bastida, "The influence of transparency on budget forecast deviations in municipal governments," *J. Forecast.*, vol. 37, pp. 457–475, 2018, doi: 10.1002/for.2513.
- [10] C. Pelser, "Improved energy budgeting process using measurement and verification principles," Ph.D. dissertation, Dept. Dev. Eng., North-West University, Potchefstroom,

2020.

- [11] J. Hager, V. S. S. Yadavalli, and R. Webber-Youngman, "Stochastic simulation for budget prediction for large surface mines in the South African mining industry," *J. South. African Inst. Min. Metall.*, vol. 115, no. 6, pp. 531–539, 2015.
- [12] H. Fourie, "Improvement in the overall efficiency of mining equipment: A case study," *J. South. African Inst. Min. Metall.*, vol. 116, no. 3, pp. 275–281, 2016, doi: 10.17159/2411-9717/2016/v116n3a9.
- [13] L. F. van der Zee, "Modelling of electricity cost risks and opportunities in the gold mining industry," Ph.D. dissertation, Dept. Elec. Eng., North-West University, Potchefstroom, 2013.
- [14] L. Webster, J. Granderson, S. Fernandes, E. Crow, and S. Earni, "IPMVP's snapshot on advanced measurement & verification," Berkeley, 2020.
- [15] Department of Energy: Republic of South Africa, "Module 7: Energy monitoring, targeting & reporting," 2015.
- [16] P. Therkelsen, P. Rao, D. Sholes, A. Mckane, B. Meffert, and R. Green, "The value of regression models in determining industrial energy savings," in *ECEEE Industrial Summer Study Proceedings*, 2016, pp. 389–399.
- [17] A. J. Schutte, "An integrated energy efficiency strategy for deep mine ventilation and refrigeration," Ph.D. dissertation, Dept. Mech. Eng., North-West University, Potchefstroom, 2014.
- [18] C. Beals *et al.*, *Improving Compressed Air System Performance*. Washington, DC., 2003.
- [19] F. W. Schroeder, "Energy efficiency opportunities in mine compressed air systems," M.S. dissertation, Dept. Mech. Eng., North-West University, Potchefstroom, 2009.
- [20] P. de Villiers, "Reducing deep-level mine refuge bay compressed air consumption," M.S. dissertation, Dept. Dev. Eng., North-West University, Potchefstroom, 2019.
- [21] B. M. Friedenstien, C. Cilliers, and J. van Rensburg, "Simulating operational improvements on mine compressed air systems," *South African J. Ind. Eng.*, vol. 29, no. 3, pp. 69–81, 2018, doi: 10.7166/29-3-2049.
- [22] H. Naser, "Energy savings through the automatic control of underground air demand," M.S.

- dissertation, Dept. Mech. Eng., North-West University, Potchefstroom, 2008.
- [23] J. A. Jacobs, M. J. Mathews, and M. Kleingeld, "Failure prediction of mine compressors," *J. Fail. Anal. Prev.*, vol. 19, no. 4, pp. 976–985, 2019, doi: 10.1007/s11668-019-00684-0.
 - [24] A. J. M. van Tonder, "Automation of compressor networks through a dynamic control system," Ph.D. dissertation, Dept. Elec. Eng., North-West University, Potchefstroom, 2014.
 - [25] R. Dindorf, "Estimating potential energy savings in compressed air systems," *Procedia Eng.*, vol. 39, pp. 204–211, 2012, doi: 10.1016/j.proeng.2012.07.026.
 - [26] H. P. Reddy, S. Narasimhan, S. M. Bhallamudi, and S. Bairagi, "Leak detection in gas pipeline networks using an efficient state estimator. Part II. Experimental and field evaluation," *Comput. Chem. Eng.*, vol. 35, no. 4, pp. 662–670, 2011, doi: 10.1016/j.compchemeng.2010.10.011.
 - [27] C. E. Brown, "Coefficient of Variation," in *Applied Multivariate Statistics in Geohydrology and Related Sciences*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 155–157.
 - [28] H. Zhang, W. Zhang, L. Lv, and Y. Feng, "Effect of fissure water on mechanical characteristics of rock mass," *Min. Sci. Technol.*, vol. 20, no. 6, pp. 846–849, 2010, doi: 10.1016/S1674-5264(09)60293-3.
 - [29] C. Cilliers, L. Van Der Zee, and M. Kleingeld, "Cost savings on mine dewatering pumps by reducing preparation-and comeback loads," *Proc. Conf. Ind. Commer. Use Energy, ICUE*, 2014, doi: 10.1109/ICUE.2014.6904166.
 - [30] M. M. Khan and G. J. Krige, "Evaluation of the structural integrity of aging mine shafts," *Eng. Struct.*, vol. 24, no. 7, pp. 901–907, 2002, doi: 10.1016/S0141-0296(02)00028-7.
 - [31] W. Badenhorst, J. Zhang, and X. Xia, "A near optimal hoist scheduling for deep level mine rock winders," *Electr. Power Syst. Res.*, vol. 42, no. 9, pp. 209–214, 2010, doi: 10.3182/20090705-4-SF-2005.00038.
 - [32] L. Giraud and B. Galy, "Fault tree analysis and risk mitigation strategies for mine hoists," *Saf. Sci.*, vol. 110, pp. 222–234, 2018, doi: 10.1016/j.ssci.2018.08.010.
 - [33] D. Seke, "From metal to money: The importance of reliable metallurgical accounting," *J. South. African Inst. Min. Metall.*, vol. 114, no. 1, pp. 1–6, 2014.
 - [34] P. G. Gaylard, N. G. Randolph, and C. M. G. Wortley, "Metal accounting in the platinum

- industry: How effective is it?," *South. African Inst. Min. Metall.*, vol. 113, no. 3, pp. 203–212, 2013.
- [35] H. E. Bartlett and M. J. Liebenberg, "Mine to metal: A practical balance for a large platinum producer," *J. South. African Inst. Min. Metall.*, vol. 114, no. 1, pp. 103–108, 2014.
- [36] A. P. Cook, C. F. Meyer, and K. Balt, "Flammable gas emissions in South African gold and platinum mines. Review of incidents and accidents," in *Proceedings of the 8th US Mine Ventilation Symposium*, 2000, vol. 53, no. 1, pp. 105–109.
- [37] M. Tang and Y. Ding, "The reliability of ergonomics in the ventilation system of an underground metal mine," *Procedia Eng.*, vol. 26, pp. 1705–1711, 2011, doi: 10.1016/j.proeng.2011.11.2357.
- [38] A. Chatterjee, L. Zhang, and X. Xia, "Optimization of mine ventilation fan speeds according to ventilation on demand and time of use tariff," *Appl. Energy*, vol. 146, pp. 65–73, 2015, doi: 10.1016/j.apenergy.2015.01.134.
- [39] T. Wnuk-Pel and C. Christauskas, "Analysis of Operational Budgeting Practices in Polish and Lithuanian Companies," *Transform. Bus. Econ.*, vol. 17, no. 3, pp. 102–125, 2018.
- [40] C. van Aarde, A. G. S. Gous, and M. Kleingeld, "Improved energy budgeting process using measurement and verification principles," *South African J. Ind. Eng.*, vol. 31, no. 3, pp. 58–68, 2020.
- [41] O. Davidson and H. Winkler, "South Africa's energy future: Visions, driving factors and sustainable development indicators," Cape Town, 2003.
- [42] J. C. Vosloo, "Control of and underground rock winder system to reduce electricity costs on RSA gold mines," M.S. dissertation, Dept. Elec. Eng., North-West University, Potchefstroom, 2006.
- [43] Gold Fields, "South Deep Gold Mine: Technical Short Form Report," 2011.
- [44] D. I. Bleiwas, "Estimates of electricity requirements for the recovery of mineral commodities, with examples applied to sub-saharan Africa," California, 2011. [Online]. Available: <http://pubs.usgs.gov/of/2011/1253/report/OF11-1253.pdf>.
- [45] B. S. Reddy and B. K. Ray, "Decomposition of energy consumption and energy intensity in Indian manufacturing industries," *Energy Sustain. Dev.*, vol. 14, no. 1, pp. 35–47, 2010, doi: <https://doi.org/10.1016/j.esd.2009.12.001>.

- [46] R. Xu, "Predictive modeling in insurance: Modeling process," *Forecast. Futur.*, no. 8, pp. 27–32, 2013, [Online]. Available: <https://www.soa.org/globalassets/assets/Library/Newsletters/Forecasting-Futurism/2013/december/ffn-2013-iss8-xu.pdf>.
- [47] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, 2021, doi: <https://doi.org/10.1007/s12525-021-00475-2>.
- [48] B. Lantz, *Machine Learning with R*. Birmingham: Packt Publishing, 2013.
- [49] J. Brownlee, *Machine Learning Mastery with Python*. 2020.
- [50] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps*. Bangalore: Springer-Science, 2019.
- [51] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Comput. Mater.*, vol. 5, no. 1, 2019, doi: 10.1038/s41524-019-0221-0.
- [52] K. P. Amber, R. Ahmad, M. W. Aslam, A. Kousar, M. Usman, and M. S. Khan, "Intelligent techniques for forecasting electricity consumption of buildings," *Energy*, vol. 157, pp. 886–893, 2018, doi: 10.1016/j.energy.2018.05.155.
- [53] M. R. Braun, H. Altan, and S. B. M. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," *Appl. Energy*, vol. 130, pp. 305–313, 2014, doi: 10.1016/j.apenergy.2014.05.062.
- [54] A. Bagnasco, F. Fresi, M. Saviozzi, F. Silvestro, and A. Vinci, "Electrical consumption forecasting in hospital facilities: An application case," *Energy Build.*, vol. 103, pp. 261–270, 2015, doi: 10.1016/j.enbuild.2015.05.056.
- [55] T. Catalina, V. Iordache, and B. Caracaleanu, "Multiple regression model for fast prediction of the heating energy demand," *Energy Build.*, vol. 57, pp. 302–312, 2013, doi: 10.1016/j.enbuild.2012.11.010.
- [56] G. Ciulla and A. D'Amico, "Building energy performance forecasting: A multiple linear regression approach," *Appl. Energy*, vol. 253, 2019, doi: 10.1016/j.apenergy.2019.113500.
- [57] J. Hwang, D. Suh, and M.-O. Otto, "Forecasting electricity consumption in commercial buildings using a machine learning approach," *Energies*, vol. 13, no. 22, 2020, doi: 10.3390/en13225885.

- [58] H. Pombeiro, R. Santos, P. Carreira, C. Silva, and J. M. C. Sousa, "Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. fuzzy modeling vs. neural networks," *Energy Build.*, vol. 146, pp. 141–151, 2017, doi: 10.1016/j.enbuild.2017.04.032.
- [59] C. Robinson *et al.*, "Machine learning approaches for estimating commercial building energy consumption," *Appl. Energy*, vol. 208, pp. 889–904, 2017, doi: 10.1016/j.apenergy.2017.09.060.
- [60] S. Seyedzadeh, F. Pour Rahimian, P. Rastogi, and I. Glesk, "Tuning machine learning models for prediction of building energy loads," *Sustain. Cities Soc.*, vol. 47, 2019, doi: 10.1016/j.scs.2019.101484.
- [61] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Dev. Built Environ.*, vol. 5, 2021, doi: 10.1016/j.dibe.2020.100037.
- [62] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007, doi: 10.1016/j.energy.2006.11.010.
- [63] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, "Time series forecasting for building energy consumption using weighted support vector regression with differential evolution optimization technique," *Energy Build.*, vol. 126, pp. 94–103, 2016, doi: 10.1016/j.enbuild.2016.05.028.
- [64] Z. Wang, Y. Wang, and R. S. Srinivasan, "A novel ensemble learning approach to support building energy use prediction," *Energy Build.*, vol. 159, pp. 109–122, 2018, doi: 10.1016/j.enbuild.2017.10.085.
- [65] F. Boulaire, A. Higgins, G. Foliente, and C. McNamara, "Statistical modelling of district-level residential electricity use in NSW, Australia," *Sustain. Sci.*, vol. 9, no. 1, pp. 77–88, 2014, doi: 10.1007/s11625-013-0206-8.
- [66] R. Granell, C. J. Axon, M. Kolokotroni, and D. C. H. Wallom, "Predicting electricity demand profiles of new supermarkets using machine learning," *Energy Build.*, vol. 234, p. 110635, 2021, doi: 10.1016/j.enbuild.2020.110635.
- [67] R. Koen and J. Holloway, "Application of multiple regression analysis to forecasting South Africa's electricity demand," *J. Energy South. Africa*, vol. 25, no. 4, pp. 48–58, 2014, doi:

10.17159/2413-3051/2014/v25i4a2238.

- [68] H. X. Zhao and F. Magoulès, "Feature selection for predicting building energy consumption based on statistical learning method," *J. Algorithms Comput. Technol.*, vol. 6, no. 1, pp. 59–77, 2012, doi: 10.1260/1748-3018.6.1.59.
- [69] S. H. Cho, W. T. Kim, C. S. Tae, and M. Zaheeruddin, "Effect of length of measurement period on accuracy of predicted annual heating energy consumption of buildings," *Energy Convers. Manag.*, vol. 45, no. 18–19, pp. 2867–2878, 2004, doi: 10.1016/j.enconman.2003.12.017.
- [70] T. Catalina, J. Virgone, and E. Blanco, "Development and validation of regression models to predict monthly heating demand for residential buildings," *Energy Build.*, vol. 40, no. 10, pp. 1825–1832, 2008, doi: 10.1016/j.enbuild.2008.04.001.
- [71] S. Farzana, M. Liu, A. Baldwin, and M. U. Hossain, "Multi-model prediction and simulation of residential building energy in urban areas of Chongqing, South West China," *Energy Build.*, vol. 81, pp. 161–169, 2014, doi: 10.1016/j.enbuild.2014.06.007.
- [72] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, "Short-term load forecasting in a non-residential building contrasting models and attributes," *Energy Build.*, vol. 92, pp. 322–330, 2015, doi: 10.1016/j.enbuild.2015.02.007.
- [73] A. Rahman and A. D. Smith, "Predicting fuel consumption for commercial buildings with machine learning algorithms," *Energy Build.*, vol. 152, pp. 341–358, 2017, doi: 10.1016/j.enbuild.2017.07.017.
- [74] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy Build.*, vol. 140, pp. 81–97, 2017, doi: 10.1016/j.enbuild.2017.01.083.
- [75] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, 2015, doi: 10.1186/s40537-014-0007-7.
- [76] M. Allen, *The SAGE Encyclopedia of Communication Research Methods*. Thousand Oaks, California: SAGE Publications, Inc, 2017.
- [77] E. Burmeister and L. M. Aitken, "Sample size: How many is enough?," *Aust. Crit. Care*, vol. 25, no. 4, pp. 271–274, 2012, doi: 10.1016/j.aucc.2012.07.002.

- [78] P. C. Austin and E. W. Steyerberg, "The number of subjects per variable required in linear regression analyses," *J. Clin. Epidemiol.*, vol. 68, no. 6, pp. 627–636, 2015, doi: 10.1016/j.jclinepi.2014.12.014.
- [79] C. R. W. van Voorhis and B. L. Morgan, "Understanding power and rules of thumb for determining sample sizes," *Tutor. Quant. Methods Psychol.*, vol. 3, no. 2, pp. 43–50, 2007, doi: 10.20982/tqmp.03.2.p043.
- [80] J. M. Morse, M. Barrett, M. Mayan, K. Olson, and J. Spiers, "Verification strategies for establishing reliability and validity in qualitative research," *Int. J. Qual. Methods*, vol. 1, no. 2, pp. 13–22, 2002, doi: 10.1177/160940690200100202.
- [81] L. Houston, Y. Probst, and A. Martin, "Assessing data quality and the variability of source data verification auditing methods in clinical research settings," *J. Biomed. Inform.*, vol. 83, pp. 25–32, 2018, doi: 10.1016/j.jbi.2018.05.010.
- [82] R. Rizk, S. McKeever, J. Petrini, and E. Zeitler, "Diftong: a tool for validating big data workflows," *J. Big Data*, vol. 6, no. 1, pp. 1–27, 2019, doi: 10.1186/s40537-019-0204-5.
- [83] S. Agarwal, *Data Mining: Data Mining Concepts and Techniques*, 3rd ed. Massachusetts: Elsevier, 2014.
- [84] W. Hamer, W. Booyesen, 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, doi: 10.7166/28-3-1850.
- [85] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Univ.*, vol. 34, no. 3, pp. 583–595, 2019.
- [86] P. E. Gill, W. Murray, and M. H. Wright, "Linear Algebra Background," in *Numerical Linear Algebra and Optimization*, 2019, pp. 27–42.
- [87] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemom. Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, 2000, doi: 10.1016/S0169-7439(99)00047-7.
- [88] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–10, 2016, doi: 10.1186/s12859-016-1236-x.
- [89] A. Gelman, "Scaling regression inputs by dividing by two standard deviations," *J. Stat.*

Med., vol. 27, pp. 2865–2873, 2008, doi: 10.1002/sim.

- [90] A. Gelman and J. Hill, “Linear regression: before and after fitting the model,” in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2010, pp. 53–78.
- [91] G. Aksu, C. O. Güzeller, and M. T. Eser, “The effect of the normalization method used in different sample sizes on the success of artificial neural network model,” *Int. J. Assess. Tools Educ.*, vol. 6, no. 2, pp. 170–192, 2019, doi: 10.21449/ijate.479404.
- [92] M. M. Suarez-Alvarez, D. T. Pham, M. Y. Prostov, and Y. I. Prostov, “Statistical approach to normalization of feature vectors and clustering of mixed datasets,” *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 468, no. 2145, pp. 2630–2651, 2012, doi: 10.1098/rspa.2011.0704.
- [93] S. Mutasa, S. Sun, and R. Ha, “Understanding artificial intelligence based radiology studies: What is overfitting?,” *Clin. Imaging*, vol. 65, pp. 96–99, 2020, doi: 10.1016/j.clinimag.2020.04.025.
- [94] A. Rácz, D. Bajusz, and K. Héberger, “Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification,” *Molecules*, vol. 26, no. 1111, pp. 1–16, 2021, doi: <https://doi.org/10.3390/molecules26041111>.
- [95] K. K. Dobbin and R. M. Simon, “Optimally splitting cases for training and testing high dimensional classifiers,” *BMC Med. Genomics*, vol. 4, no. 31, pp. 1–8, 2011, doi: 10.1186/1755-8794-4-31.
- [96] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” 2018. [Online]. Available: <http://arxiv.org/abs/1811.12808>.
- [97] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, “Cross-validation pitfalls when selecting and assessing regression and classification models,” *J. Cheminform.*, vol. 6, no. 10, pp. 1–15, 2014, doi: 10.1186/1758-2946-6-10.
- [98] W. Wang *et al.*, “A systematic review of machine learning models for predicting outcomes of stroke with structured data,” *PLoS One*, vol. 15, no. 6, pp. 1–16, 2020, doi: 10.1371/journal.pone.0234722.
- [99] M. Westphal and W. Brannath, “Evaluation of multiple prediction models: A novel view on model selection and performance assessment,” *Stat. Methods Med. Res.*, vol. 29, no. 6, pp. 1728–1745, 2020, doi: 10.1177/0962280219854487.
- [100] S. S. Henley, R. M. Golden, and T. M. Kashner, “Statistical modeling methods: challenges

- and strategies,” *Biostat. Epidemiol.*, vol. 4, no. 1, pp. 105–139, 2020, doi: 10.1080/24709360.2019.1618653.
- [101] B. Zdaniuk, “Ordinary least-squares (OLS) model,” in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed. Dordrecht: Springer Netherlands, 2014, pp. 4515–4517.
- [102] J. T. Pohlmann and D. W. Leitner, “A comparison of ordinary least squares and logistic regression,” *Ohio J. Sci.*, vol. 103, no. 5, pp. 118–125, 2003.
- [103] E. Ostertagová, “Modelling using polynomial regression,” *Procedia Eng.*, vol. 48, pp. 500–506, 2012, doi: 10.1016/j.proeng.2012.09.545.
- [104] J. W. Osborne and E. Waters, “Four assumptions of multiple regression that researchers should always test,” *Pract. Assessment, Res. Eval.*, vol. 8, no. 2, pp. 1–5, 2003.
- [105] R. J. Casson and L. D. M. Farmer, “Understanding and checking the assumptions of linear regression: A primer for medical researchers,” *Clin. Exp. Ophthalmol.*, vol. 42, no. 6, pp. 590–596, 2014, doi: 10.1111/ceo.12358.
- [106] P. Ranganathan and N. J. Gogtay, “An introduction to statistics – data types, distributions and summarizing data,” *Indian J. Crit. Care Med.*, vol. 23, no. 2, pp. 169–170, 2019, doi: 10.5005/jp-journals-10071-23198.
- [107] R. D. Sheats and V. Shane Pankratz, “Understanding distributions and data types,” *Semin. Orthod.*, vol. 8, no. 2, pp. 62–66, 2002, doi: 10.1053/sodo.2002.32075.
- [108] M. Krzywinski and N. Altman, “Visualizing samples with box plots,” *Nat. Methods*, vol. 11, no. 2, pp. 119–120, 2014, doi: 10.1038/nmeth.2813.
- [109] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*, 2nd ed. Hoboken: John Wiley & Sons, Inc., 2007.
- [110] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.
- [111] A. M. Variyath and A. Brobbey, “Variable selection in multivariate multiple regression,” *PLoS One*, vol. 15, no. 7, pp. 1–15, 2020, doi: 10.1371/journal.pone.0236067.
- [112] M. M. Mukaka, “Statistics corner: A guide to appropriate use of correlation coefficient in

- medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.
- [113] I. Jebli, F. Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," *Energy*, vol. 224, p. 120109, 2021, doi: 10.1016/j.energy.2021.120109.
- [114] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," 2018. [Online]. Available: <http://arxiv.org/abs/1809.03006>.
- [115] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
- [116] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE , MAE , MAPE , MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, no. 623, pp. 1–24, 2021, doi: 10.7717/peerj-cs.623.
- [117] A. Botchkarev, "Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio," *SSRN Electron. J.*, pp. 1–16, 2018, doi: 10.2139/ssrn.3177507.
- [118] C. L. Cheng, Shalabh, and G. Garg, "Coefficient of determination for multiple measurement error models," *J. Multivar. Anal.*, vol. 126, pp. 137–152, 2014, doi: 10.1016/j.jmva.2014.01.006.
- [119] M. Analla, "Model validation through the linear regression fit to actual versus predicted values," *Agric. Syst.*, vol. 57, no. 1, pp. 115–119, 1998, doi: 10.1016/S0308-521X(97)00073-5.
- [120] Lonmin, "Mineral resources and mineral reserve statement," 2018. [Online]. Available: <http://reports.sibanyestillwater.com/2018/download/lonmin-RR2018.pdf>.

APPENDICES

Appendix A – Literature review detailed findings

Table A-1: Summary of the purposes of the models created in the reviewed studies

Ref	Purpose	Model purpose
[52]	Budgeting	Building electricity consumption forecasting for budget planning, energy efficiency initiatives, reduction in carbon emissions and controlling energy usage.
[53]	Design	Supermarket electricity consumption forecasting for planning with the inclusion of climate change effects
[54]	Energy management	Hospital electricity consumption forecasting for dynamic management of thermal behaviour for energy saving policy
[55]	Design	Prediction of building heating demand for the design of new buildings
[56]	Design & energy management	Prediction of building electricity consumption for building design and energy management
[57]	Energy management	Prediction of building electricity consumption for efficient power grid operation. Specifically for the prevention of power shortages, and the reduction of social costs as result of unnecessary supply
[58]	Energy management	Prediction of intelligent building electricity consumption for the implementation of energy efficiency initiatives and energy management
[59]	Design	Prediction of commercial building electricity consumption to map the relationship between design and consumption for the improvement of city planning
[60]	Design	Utilisation of machine learning models for building energy load prediction for design and retrofitting
[61]	Energy management	Building energy consumption forecasting for application into a building energy management system for energy wastage reduction
[62]	Energy management	Prediction of Hong Kong electricity usage for future demand forecasting by utility companies
[63]	Energy management	Building electricity load forecasting for operational and management improvement
[64]	Energy management	Building electricity usage prediction for energy management and implementation of electricity efficiency initiatives

Ref	Purpose	Model purpose
[65]	Design & Energy management	Prediction of electricity consumption of residential sector in New South Wales, Australia, for electricity network investment and asset management. For the evaluation of future scenarios on energy consumption
[66]	Design & Energy management	Predicting the electricity daily load profile of proposed new supermarkets for future planning with the consideration of energy management principles
[67]	Budgeting & Energy management	Forecasting long term electricity usage patterns to determine the effect of different scenarios on these patterns. Specific application to strategic planning. Done for multiple sectors – platinum mining included.
[68]	Energy management	Investigation of the effect of feature selection on the performance of machine learning models in the prediction of building energy consumption
[69]	Energy management	Predicting the annual heating energy consumption in buildings. Multiple models to investigate the effect of measurement period on model accuracy
[70]	Energy management	Predicting monthly heating demand for single-family residential sector for utilisation in project design for the implementation of energy efficiency initiatives
[71]	Design	Building energy usage prediction for future energy demand in urban residential areas resulting from rapid urbanisation
[72]	Energy management	Short-term load forecasting for utility provider resource planning for the balance of electricity supply and demand
[73]	Energy management	Building energy prediction for the development of building emission models which can be coupled with climate models
[74]	Energy management	Prediction of energy usage of appliances in buildings for demand side management, detection of abnormal energy usage patterns and for energy management

Table A-2: Reviewed references summary of data acquisition, validation, cleaning and preparation steps

Reference	Variables	Datapoints	Verified	Validated	Transform	Split
Relevant question:	Input variables considered?	Number of datapoints obtained?	Data obtained from verified source?	Outlier and missing values fixed?	Data transformed or scaled?	Data split into different sets?
[52]	5	≈1 825	✓	✗	✗	✓
[53]	2	44	✗	✓	✓	✗
[54]	6	≈35 000	✗	✗	✗	✓
[55]	3	8 765	✓	✗	✓	✓
[56]	7	1 560	✓	✗	✗	✓
[57]	10	1 000	✗	✗	✓	✓
[58]	9	3 381	✗	✗	✗	✓
[59]	5	2 612	✓	✓	✓	✓
[60]	12	25 000	✓	✗	✓	✓
[61]	4	18 628	✓	✗	✓	✓
[62]	18	1166	✓	✗	✓	✗
[63]	1	261	✓	✓	✗	✓
[64]	11	8 297	✓	✓	✗	✓
[65]	20	8 632	✓	✓	✓	✓
[66]	12	-	✗	✓	✗	✓
[67]	1	38	✓	✗	✗	✓
[68]	23	6 072	✓	✓	✓	✓
[69]	1	≈90	✗	✗	✗	✗
[70]	9	18 144	✓	✓	✗	✓
[71]	14	12	✓	✗	✗	✗
[72]	28	7616	✓	✓	✓	✓
[73]	26	-	✓	✗	✗	✓
[74]	32	26 300	✓	✓	✗	✓
Total (23)	-	-	17	10	10	19

Table A-3: Minimum sample size required for each of the considered studies

Reference	Variables	Minimum datapoints	Datapoints	Requirement satisfied
	Input variables considered?	Minimum datapoints required	Number of datapoints obtained?	Was at least the minimum number of points considered?
[52]	5	55	≈1 825	✓
[53]	2	52	44	×
[54]	6	98	≈35 000	✓
[55]	3	53	8 765	✓
[56]	7	106	1 560	✓
[57]	10	130	1 000	✓
[58]	9	122	3 381	✓
[59]	5	55	2 612	✓
[60]	12	146	25 000	✓
[61]	4	54	18 628	✓
[62]	18	194	1166	✓
[63]	1	51	261	✓
[64]	11	138	8 297	✓
[65]	20	210	8 632	✓
[66]	12	146	-	-
[67]	1	51	38	×
[68]	23	234	6 072	✓
[69]	1	51	≈90	✓
[70]	9	122	18 144	✓
[71]	14	162	12	
[72]	28	274	7616	✓
[73]	26	258	-	
[74]	32	306	26 300	✓

Table A-4: Types of variables and data resolution considered in evaluated studies

Reference	System variables	Time based variables	Seasonal variables	Resolution
Relevant question:	Operational/system variables considered?	Time based variables considered?	Seasonal variables considered?	Resolution of data points?
[52]	✓	✓	✓	Daily
[53]	×	×	✓	Weekly
[54]	✓	✓	✓	Daily
[55]	✓	×	✓	Design*
[56]	✓	×	✓	Design*
[57]	✓	×	✓	Monthly
[58]	✓	✓	✓	15 minute
[59]	✓	×	✓	Design*
[60]	✓	×	✓	Design*
[61]	✓	×	×	Monthly
[62]	✓	×	✓	Design*
[63]	×	✓	×	Daily
[64]	✓	✓	✓	Hourly
[65]	✓	×	×	Design*
[66]	✓	×	✓	Hourly
[67]	✓	×	×	Yearly
[68]	✓	×	✓	Hourly
[69]	×	×	✓	Daily
[70]	✓	✓	✓	Design*
[71]	✓	×	×	Yearly
[72]	✓	✓	✓	Hourly
[73]	✓	✓	✓	Hourly
[74]	✓	✓	✓	Hourly
Total	20	9	18	

Table A-5: Summary of outlier removal and methods used in reviewed studies

Reference	Outliers removed	Number removed	Method used
Relevant question:	Were outliers removed?	How many outliers were removed?	What method was used for outlier removal?
[52]	×	-	-
[53]	✓	~ 6	Outliers of residual terms
[54]	×	-	-
[55]	×	-	-
[56]	×	-	-
[57]	×	-	-
[58]	×	-	-
[59]	✓	4 108	Data filtered according to specific data ranges
[60]	×	-	-
[61]	×	-	-
[62]	×	-	-
[63]	✓	-	Removal of zero energy values when airconditioning system was not operational
[64]	✓	271	No mentioned how data was cleaned
[65]	✓	8	Observation of points that didn't follow the trend of the data
[66]	✓	-	-
[67]	×	-	-
[68]	✓	-	Removal of lower electricity consumption data points (Saturday and Sundays have lower consumption)
[69]	×	-	-
[70]	✓	-	Removal of datapoints with variables outside of considered ranges
[71]	×	-	-
[72]	✓	113	Points outside a specific empirically determined percentage of the IQR
[73]	×	-	-
[74]	✓	-	-
Total	10	-	-

Table A-6: Summary of data transformation and methods used in reviewed studies

Reference	Transformation performed	Method used	Reason
Relevant question:	Was the data transformed in any way?	What method was used for data transformation?	Why was the data transformed?
[52]	x	-	-
[53]	x	-	-
[54]	x	-	-
[55]	x		
[56]	x	-	-
[57]	✓	Min-max normalisation	Normalisation to adjust for numerical data differences – prevention of model deterioration
[58]	x	-	-
[59]	✓	Z-score scaling	Energy consumption is observed to follow an approximate log-normal distribution
[60]	✓	Min-max scaling	Removal of numerical differences in data – prevention of model performance degradation
[61]	✓	Z-score scaling	Removal of numerical differences in data – prevention of model performance degradation
[62]	x	-	-
[63]	x	-	-
[64]	x	-	-
[65]	x	-	-
[66]	x	-	-
[67]	x	-	-
[68]	✓	Min-max scaling	Removal of numerical differences in data – prevention of model performance degradation
[69]	x	-	-
[70]	x	-	-
[71]	x	-	-

Reference	Transformation performed	Method used	Reason
Relevant question:	Was the data transformed in any way?	What method was used for data transformation?	Why was the data transformed?
[72]	✓	Unit length scaling	Removal of numerical differences in data – prevention of model performance degradation
[73]	x	-	-
[74]	x	-	-
Total	6		

Table A-7: Summary of train/test split ratios and methods applied in reviewed studies

Reference	Split performed	Ratio		Method
Relevant question:	Was the data split into training and test sets?	Training set ratio	Testing set ratio	What method was used for the split?
[52]	✓	80%	20%	Most recent year chosen for testing
[53]	✓	75%	25%	New data obtained for testing
[54]	✓	67%	33%	Most recent days chosen for testing
[55]	✓	99%	1%	Training data was obtained from simulated points. Testing data obtained from real buildings
[56]	✓	85%	15%	Training data was obtained from simulated points. Testing data obtained from real buildings
[57]	x	-	-	-
[58]	✓	60%	40%	Random split – cross-validation
[59]	✓	-	-	Random split – cross-validation
[60]	✓	-	-	Random split – cross-validation
[61]	✓	70%	30%	Random split – Single set
[62]	x	-	-	-
[63]	✓	80%	20%	Random split - cross-validation
[64]	✓	90%	10%	Random split - cross-validation
[65]	✓	80%	20%	Random split – Single set
[66]	✓	-	-	-
[67]	x	-	-	-
[68]	✓	75%	25%	Most recent years used as test data – cross-validation
[69]	✓	74%	26%	Random split – Single set
[70]	✓	99%	<1%	Real buildings used for testing. Simulation data used for training
[71]	x	-	-	-
[72]	✓	66%	33%	Random split – cross-validation
[73]	✓	?	?	Random split – cross-validation
[74]	✓	75%	25%	Random split – cross-validation
Total (23)	19			

Table A-8: Summary of model number, model types and number of performance metrics used for modelling in reviewed studies

Reference	Number of models	Least squares	Machine learning	Performance metrics	Model verification
Relevant question:	How many different models were considered?	Least squares regression model considered?	Advanced machine learning models considered?	Number of performance metrics considered?	Were models verified on an 'unseen' dataset?
[52]	5	✓	✓	5	×
[53]	1	✓	×	2	×
[54]	1	×	✓	5	×
[55]	1	✓	×	4	×
[56]	1	✓	✓	5	×
[57]	5	×	✓	4	×
[58]	2	✓	✓	3	×
[59]	13	✓	✓	3	✓
[60]	5	×	✓	3	×
[61]	3	×	✓	5	×
[62]	3	✓	✓	1	×
[63]	3	×	✓	5	×
[64]	3	×	✓	3	×
[65]	1	✓	×	1	×
[66]	4	✓	✓	3	×
[67]	1	✓	×	1	×
[68]	1	×	✓	2	×
[69]	1	✓	×	1	×
[70]	1	✓	×	2	×
[71]	6	✓	✓	2	×
[72]	3	✓	✓	2	✓
[73]	4	✓	✓	2	✓
[74]	4	✓	✓	2	✓
Total		16	16		4

Table A-9: Summary of types of models considered by evaluated studies

Reference	Number of models	‘Simple’ models considered	Machine learning models considered	Best performing model
Relevant question:	How many different models were considered?	Models included that are considered to be simple (non-machine learning)?	Models included that are considered to be machine learning models?	Which model provided the best performance
[52]	5	MR	GP, ANN, DNN, SVM	ANN (6%) MR(8.5%) MAPE
[53]	1	MR	-	-
[54]	1	-	ANN	-
[55]	1	MR	-	-
[56]	1	MR	-	-
[57]	5	-	ANN, DNN, LSTM, SVR and SARIMAX	SARIMAX
[58]	2	MR	Fuzzy model, NN	79.3% Fuzzy model values variance accounted for
[59]	13	MR	XGBoost, Bagging, MLP, RF, ET, SVR, KNN, ADaB, linear SVR, ridge, elasticnet lasso	Gradient boosting $r^2=0.82$
[60]	5	-	ANN, SVM, GP, RF, GBRT	Gradient boosting
[61]	3	-	KNN, SVM, ANN	SVM MAPE = 8.2%
[62]	3	MR	ANN, Decision tree	Decision tree

Reference	Number of models	'Simple' models considered	Machine learning models considered	Best performing model
Relevant question:	How many different models were considered?	Models included that are considered to be simple (non-machine learning)?	Models included that are considered to be machine learning models?	Which model provided the best performance
[63]	3	-	Eps-SVM, nu-SVM and DE incl.	-
[64]	3	-	CART, EBT, compact EBT	Bagging trees (2.97 - 4.63% MAPE)
[65]	1	MR	-	-
[66]	4	MR	KNN, SVR, ANN	Similar. MR/KNN outperforming SVR and ANN
[67]	1	MR	-	-
[68]	1	-	SVM (2 kernel functions)	-
[69]	1	LR	-	-
[70]	1	MLR	-	-
[71]	6	2 Grey models, MLR, Polynomial model, polynomial regression	ANN	ANN
[72]	3	MLR	ANN, SVM	MLR
[73]	4	MLR	ANN, GP, Ridge regression	
[74]	4	MLR	SVM, GBM, RF	

Table A-10: Summary of variable correlation assessments performed by reviewed studies

Reference	Correlation	Removed
Relevant question:	Was variable correlation investigated?	Were poorly correlated variables removed?
[52]	x	x
[53]	✓	x
[54]	✓	x
[55]	✓	x
[56]	✓	x
[57]	✓	x
[58]	✓	✓
[59]	x	
[60]	✓	✓
[61]	x	
[62]	✓	x
[63]	x	
[64]	✓	✓
[65]	✓	✓
[66]	✓	x
[67]	x	
[68]	✓	✓
[69]	x	
[70]	x	
[71]	x	
[72]	✓	✓
[73]	✓	✓
[74]	✓	✓
Total	15	8

Table A-11: Summary of number and type of performance metrics used by reviewed studies

Reference	Multiple metrics	Validation	Specific metrics used
Relevant question:	Were multiple performance metrics considered?	Was a model chosen and validated on 'unseen' data?	What specific performance metrics were used for model evaluation?
[52]	✓	×	RMSE, MAE, MRE, MAPE, NRMSE
[53]	✓	×	NMBE, CVRMSE
[54]	✓	×	MAPE, Daily Peak MAPE, CVRMSE, MPE, Percentage of Test set with a MAPE under 5%
[55]	✓	×	MAE, MSE, RMSE, MAPE
[56]	✓	×	R2, MAE, MSE, RMSE, MAPE
[57]	✓	×	MAPE, RMSE, MBE, CV
[58]	✓	×	VAF, MAE, MSE
[59]	✓	✓	MAE (10 [^]), medAE (10 [^]), r2
[60]	✓	×	R2, RMSE, MAE
[61]	✓	×	R2, RMSE, MAE, NRMSE, MAPE
[62]	×	×	RMSE
[63]	✓	×	ME, RMSE, MAE, MPE, MAPE
[64]	✓	×	R2, RMSE, MAPE
[65]	×	×	R2
[66]	✓	×	Euclidean distance, MSE, NP
[67]	×	×	R2
[68]	✓	×	MSE, R2
[69]	×	×	MAPE
[70]	✓	×	R2 and standard error of the estimate
[71]	✓	×	MAPE, R2
[72]	✓	✓	MAPE, R2
[73]	✓	✓	MAE, Relative error
[74]	✓	✓	RMSE, R2, MAE, MAPE
Total	19	4	

Table A-12: Summary of insights drawn from reviewed studies

Reference	Interpretation	Insights drawn
Relevant question:	Was the model interpreted?	Was specific insights were drawn from model interpretation?
[52]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[53]	✓	The models dependence on different variables changes over the operating range as expected becoming less linearly dependant on some variables. Relating black-box model to expected operation of known system
[54]	✓	Evaluation of actual model behaviour compared to expected model behaviour. For example the model predicts lower electricity usage in winter due to absence of airconditioning.
[55]	✓	Squaring specific input variables to account for non-linearity of the system and its variables shows insight
[56]	✓	Multiple different MR models created with different variables. Shows that they attempted to show that more variables are more descriptive of system.
[57]	✓	Application of different models to different operating conditions.
[58]	✓	Models indicated that occupancy is a strong predictor variable. Consideration to include more occupancy related variables in the next iteration
[59]	✓	Limitations of models discussed. Realisation that the resulting models require a large amount of data and that they only work in the USA.
[60]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[61]	✓	Analysis performed to identify which input variables have the largest effect on the output variable
[62]	✓	Analysis performed to identify which input variables have the largest effect on the output variable
[63]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.

Reference	Interpretation	Insights drawn
Relevant question:	Was the model interpreted?	Was specific insights were drawn from model interpretation?
[64]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[65]	✓	Evaluation of the predictive capacity of the model based on location. Draws insights into the validity of the model for different areas and hence its ability to predict
[66]	✓	Models are built by splitting the supermarket data initially by temperature and season. This indicates that there is understanding that the system is affected by these variables enough to model them as independent models. Shows that the model was built with system understanding in mind
[67]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[68]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[69]	✓	Analysis performed to identify which input variables have the largest effect on the output variable
[70]	✓	Analysis performed to identify which input variables have the largest effect on the output variable
[71]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[72]	✓	Modelling helped identify that ambient data collected inside the building does not improve model performance – HVAC operation is fixed.
[73]	x	Only tuning and selection of models discussed. No insights about system behaviour interpreted.
[74]	✓	Models helped to identify which variables have the largest impact on electricity consumption. Insights into system behaviour.

Appendix B – Shaft electricity consumption models

Data collection and preparation

Outlier detection

The Mahalanobis distance was calculated for all data points for the electricity consumption forecasting of all the shafts considered. This was used together with the boxplot method to identify outlier data points from the multivariate datasets. The graphs illustrating the distance outliers are shown in Figure B-1 below.

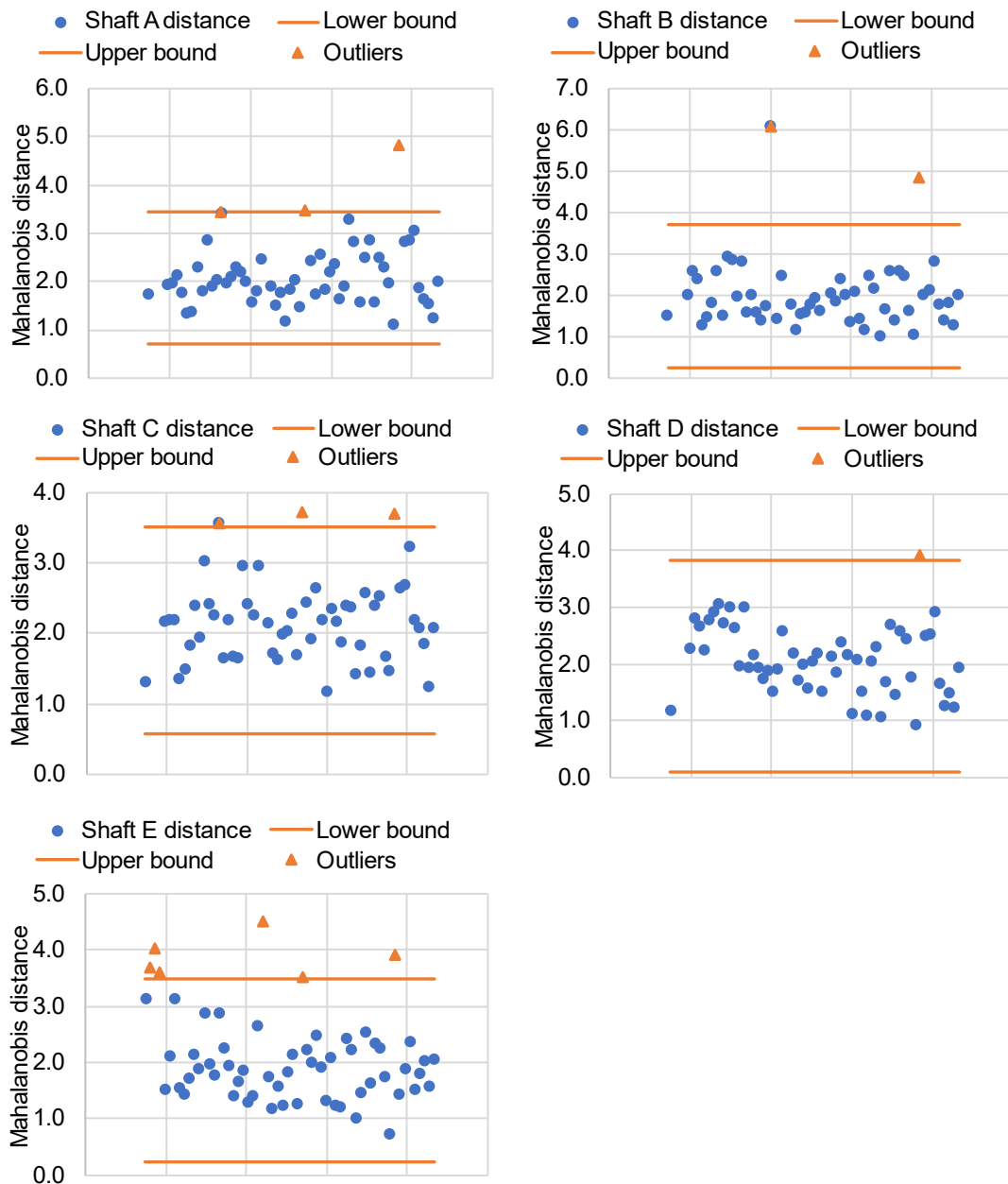


Figure B-1: Outlier detection using the Mahalanobis distance and boxplot method on shaft electricity usage for shafts A-E

Figure B-1 illustrates the retained data points, the outlier data points as well as the quartile 1 and quartile 3 boundaries used to identify the outliers. The graphs indicate that outliers were identified for all shafts considered. Furthermore, the graphs indicate that the distances of the retained data points are clustered closely together. The removal of the outliers serves to ensure that the final data set is clean, reducing the amount of unnecessary error that would result from the presence of non-representative data points.

Data transformation

Once the outliers were removed, it was necessary to scale the data. This is done in order to prevent large differences in the units of input variables from causing weighting differences between the variables. Two scaling methods were considered, min-max scaling and z-score scaling. The former is used for datasets that are not normally distributed, while the latter is used on normally distributed data sets. Therefore, prior to scaling, it was necessary to evaluate the data distributions of all variables that were scaled – this is shown in Figure B-2 and Figure B-3.

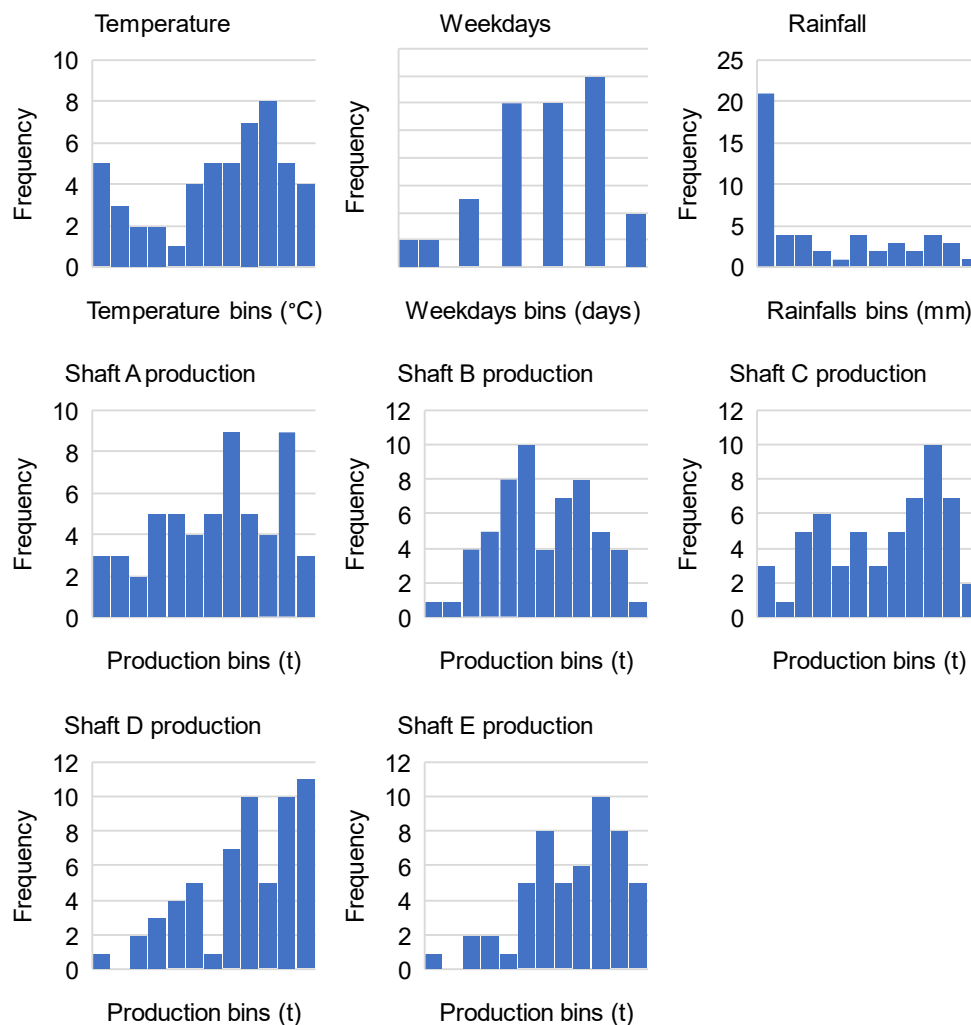


Figure B-2: Input variable distributions for the determination of the most applicable scaling method for shaft electricity usage

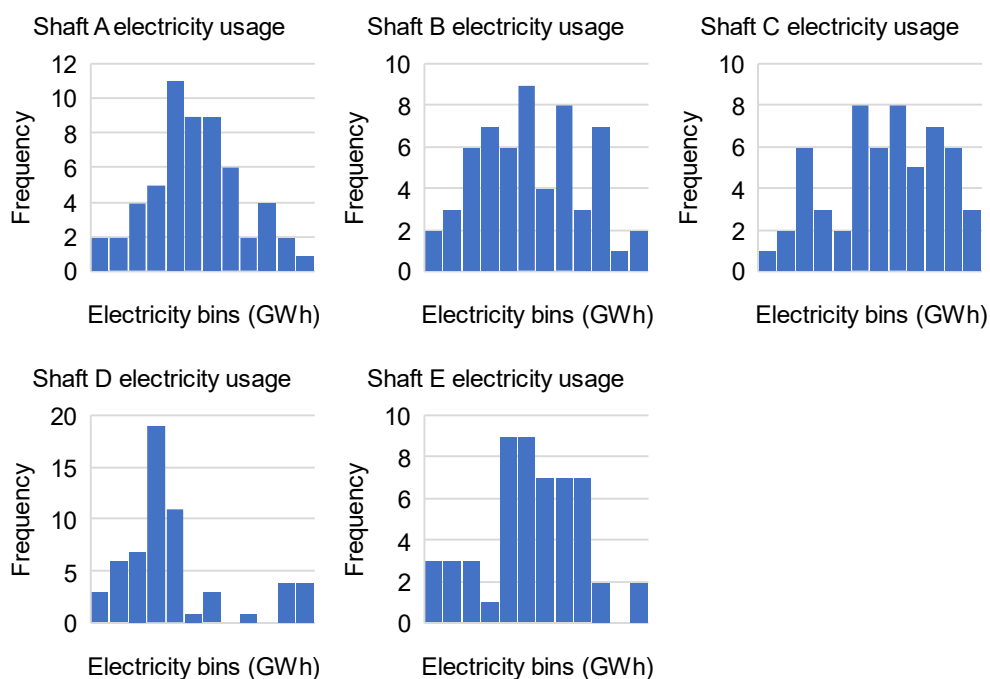


Figure B-3: Output variable distributions for the determination of the most applicable scaling method for shaft electricity usage

Figure B-2 and Figure B-3 show the distribution of the input and output variables that were used to generate models for forecasting the shaft electricity usage. The scaling method used for each variable is dictated by the distribution of the data and is shown in Table B-1.

Table B-1: Summary of scaling methods applied to the variables used for the modelling of the shaft electricity consumption

Variable	Min-max scaling	Z-score scaling
Temperature	X	
Weekdays		X
Rainfall	X	
Shaft A production		X
Shaft B production		X
Shaft C production		X
Shaft D production	X	
Shaft E production	X	
Shaft A electricity consumption		X
Shaft B electricity consumption		X
Shaft C electricity consumption		X
Shaft D electricity consumption	X	
Shaft E electricity consumption		X

Data splitting

Once the data was cleaned of outliers and transformed, it was necessary to split it into train/test and verification sets. The datasets were first randomised in order to ensure that split datasets were representative of the total dataset. The resulting train/test set was used with k-fold cross-verification to train the resulting model.

Furthermore, the separation of a verification set was necessary to ensure that the resulting model was tested on unseen data. The data points that comprised the train/test and verification sets for each shaft are shown in Figure B-4 to Figure B-8 below.

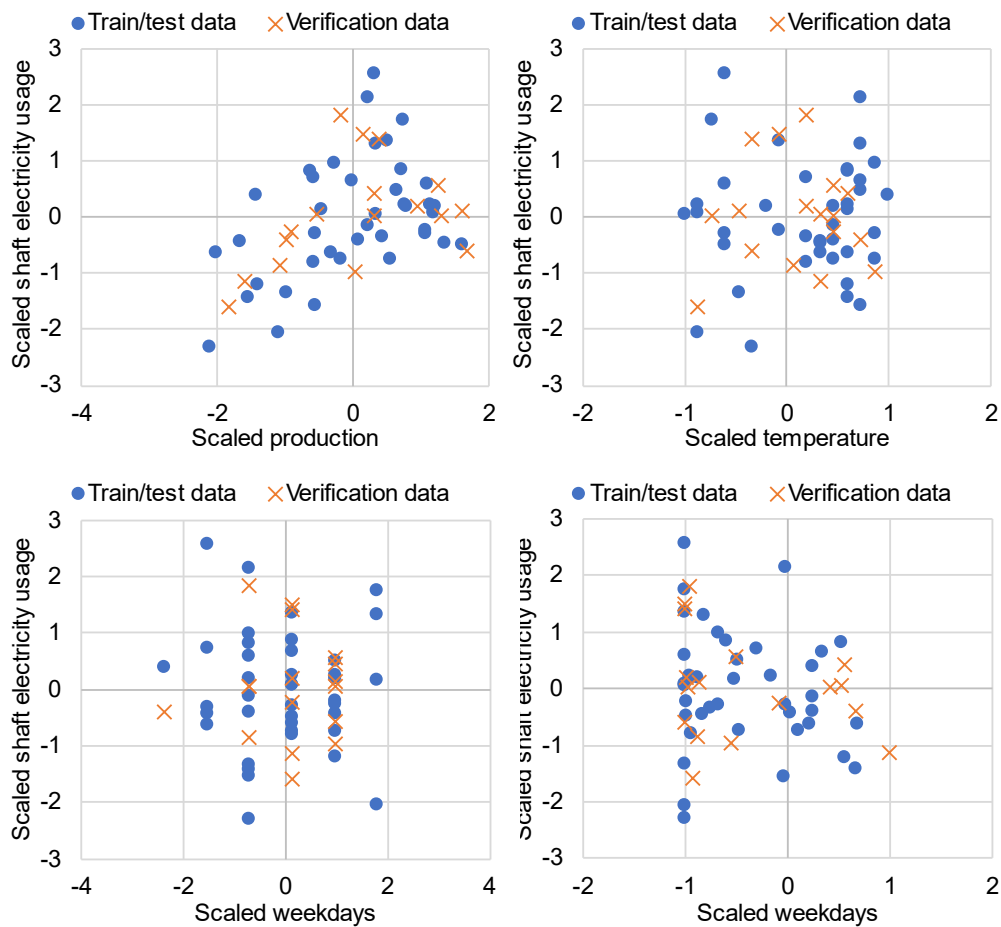


Figure B-4: Data splits showing the train/test and verification sets for shaft A

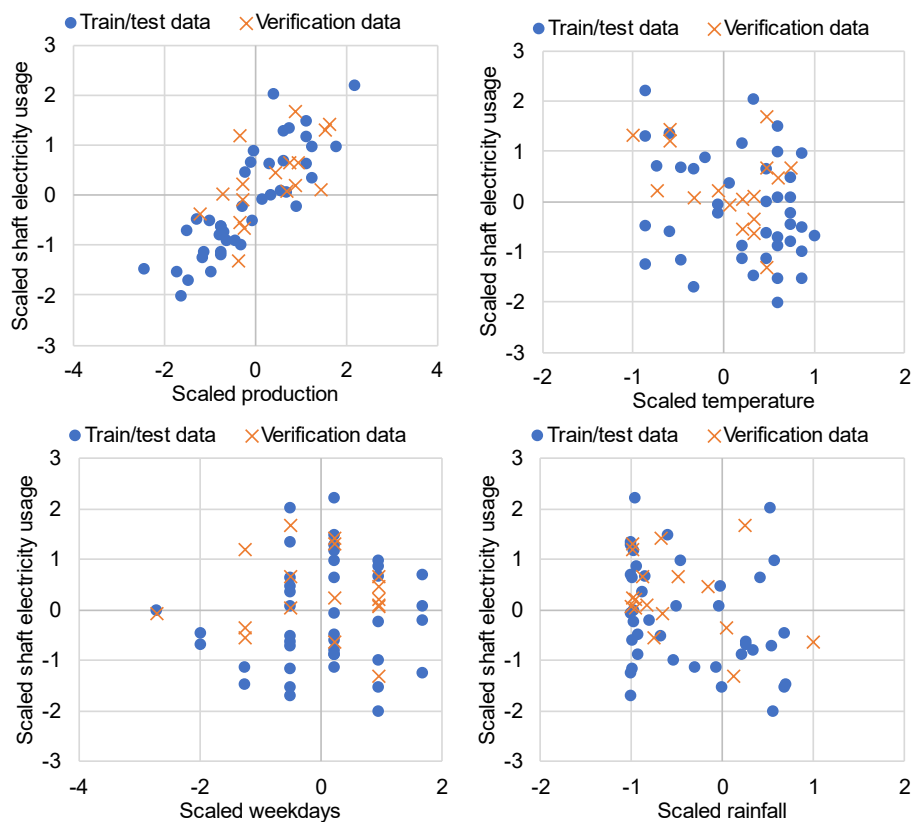


Figure B-5: Data splits showing the train/test and verification sets for shaft B

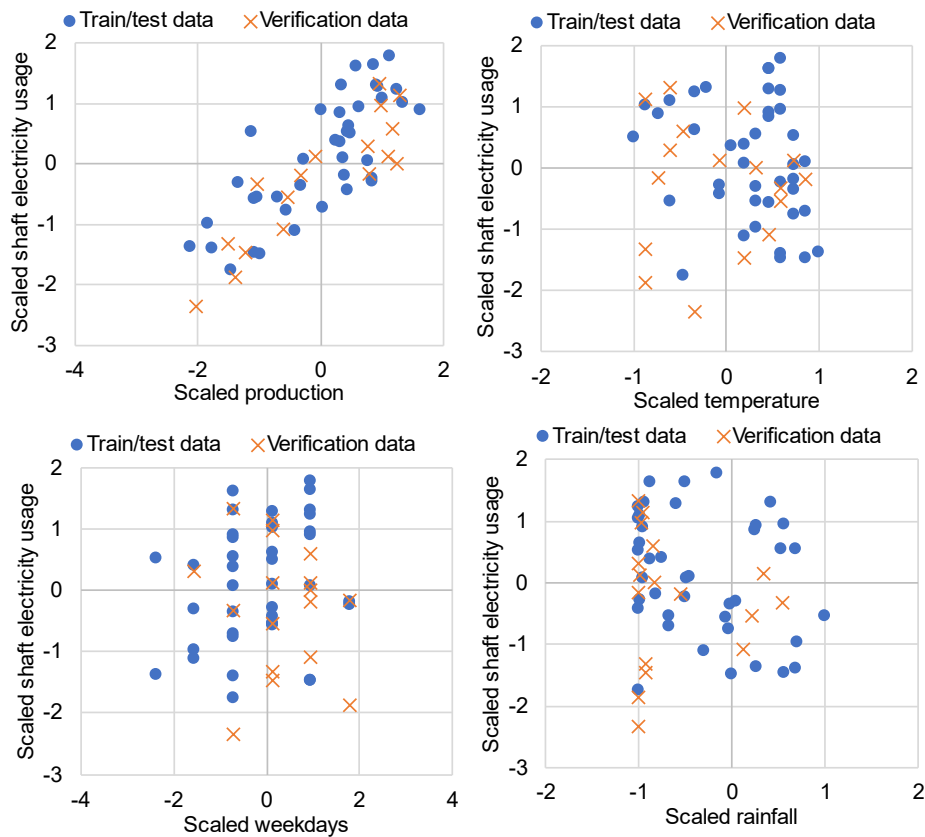


Figure B-6: Data splits showing the train/test and verification sets for shaft C

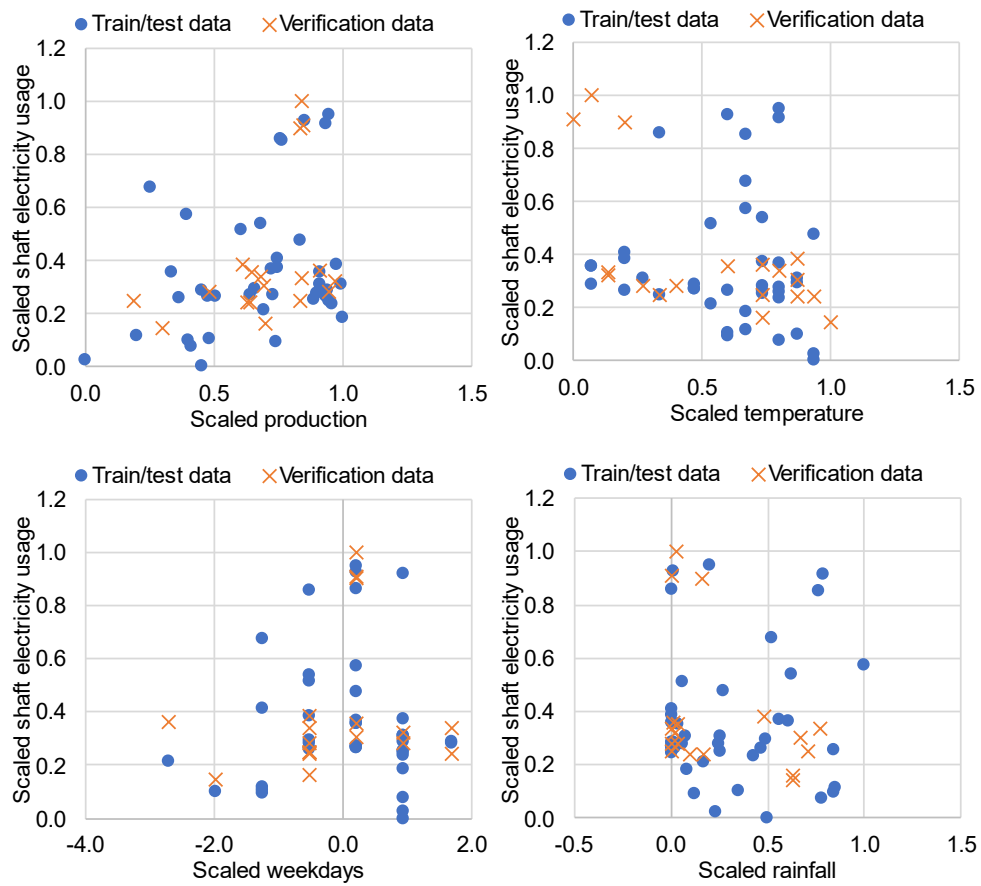


Figure B-7: Data splits showing the train/test and verification sets for shaft D

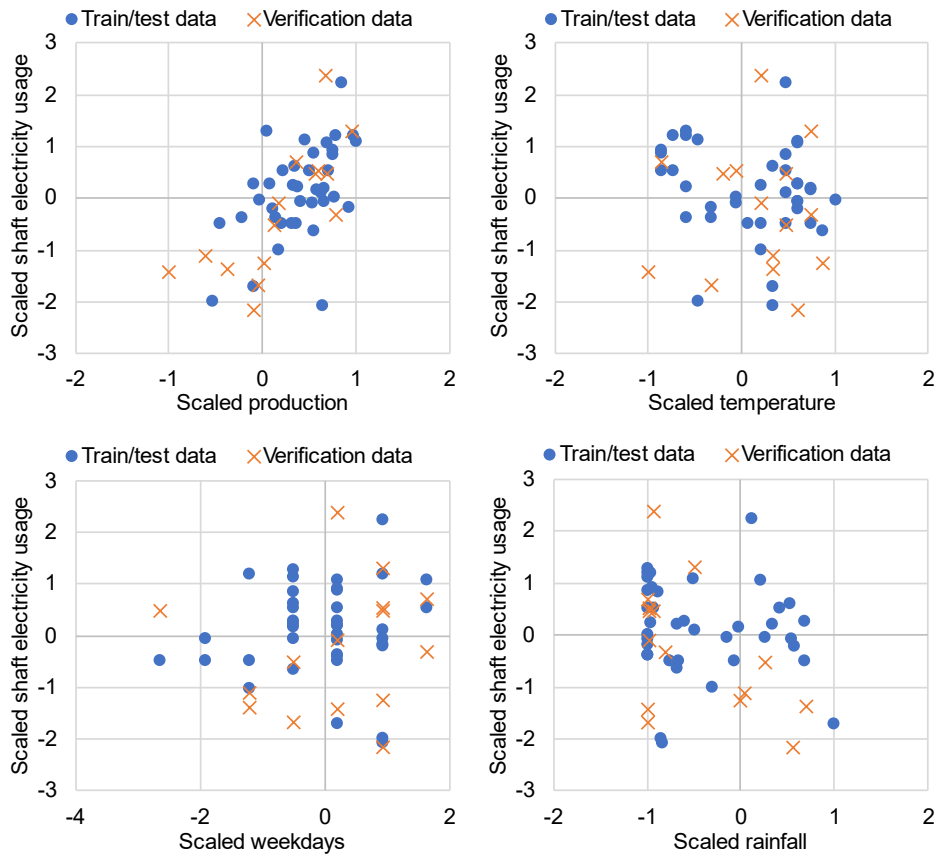


Figure B-8: Data splits showing the train/test and verification sets for shaft E

Figure B-4 to Figure B-8 illustrate that the values of the variables in the train/test and verification sets for each of the shafts span the range of each variable. This shows that the values used in each set are representative of the dataset and will result in accuracies that can be considered valid. This check is necessary in order to ensure that the verification and train/test data sets are similar in distribution.

Model development

Variable correlation assessment

After cleaning and splitting of the datasets it was possible to begin with the model development step of the process. This step began with the determination of input variables that were most strongly correlated to each output variable to be forecasted. This determination was performed with the use of Pearson's correlation coefficient. According to Table 1.8, variables with absolute correlation coefficients less than 0.4 are considered to have weak to negligible correlation.

As such, only input variables with absolute correlation coefficients larger than or equal to 0.4 to the output variables were considered. It is necessary to remove variables with low correlation as the inclusion of these can have an adverse effect on model performance. The scatter plots between variables and associated Pearson's correlation coefficients are shown in Figure B-9, Figure B-10, Figure B-11, Figure B-12 and Figure B-13 below.

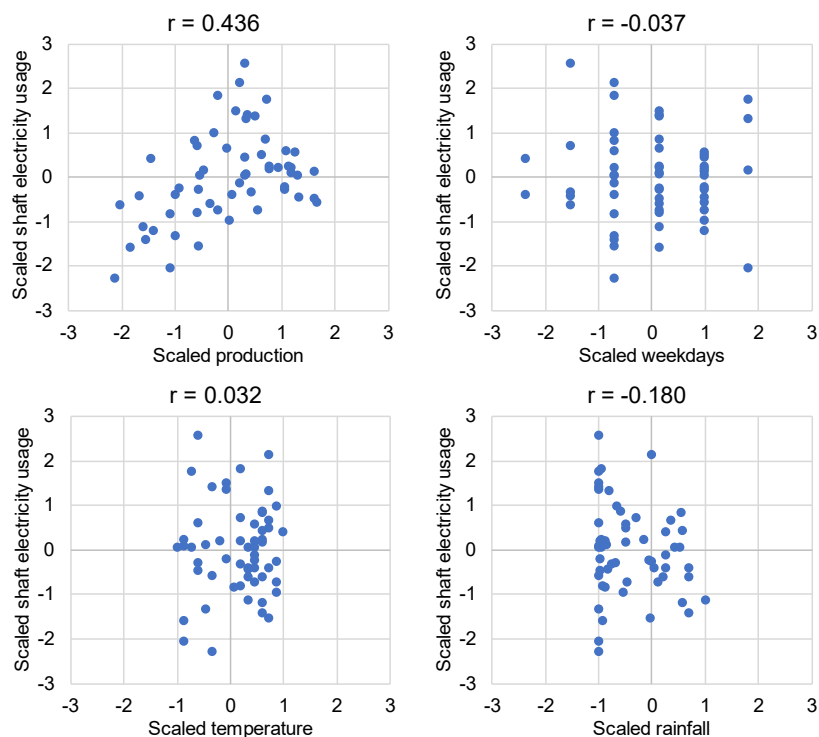


Figure B-9: Variable correlation graphs and Pearson's coefficients for the shaft electricity usage of shaft A

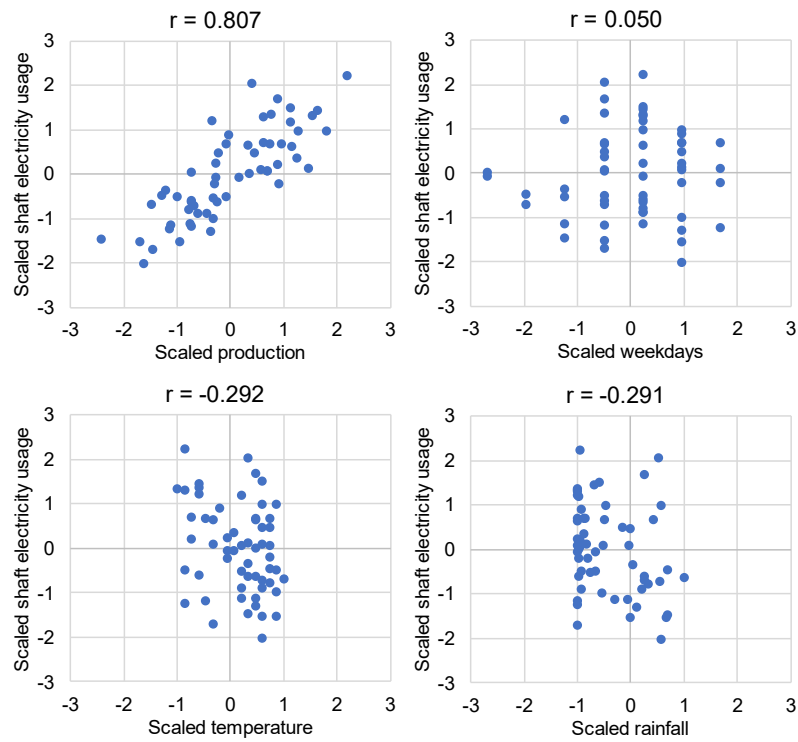


Figure B-10: Variable correlation graphs and Pearson's coefficients for the shaft electricity usage of shaft B

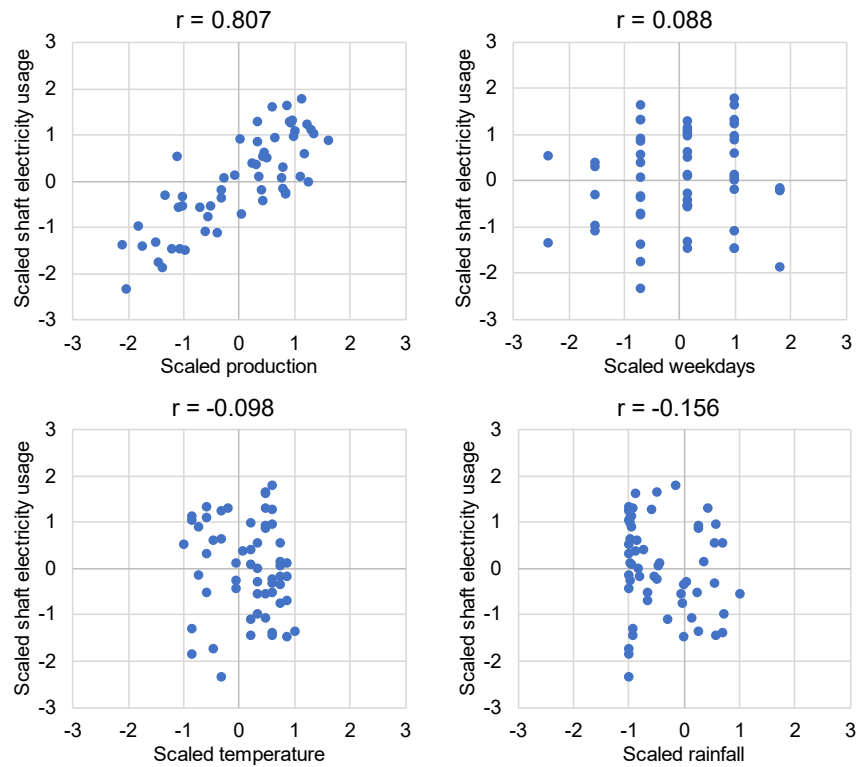


Figure B-11: Variable correlation graphs and Pearson's coefficients for the shaft electricity usage of shaft C

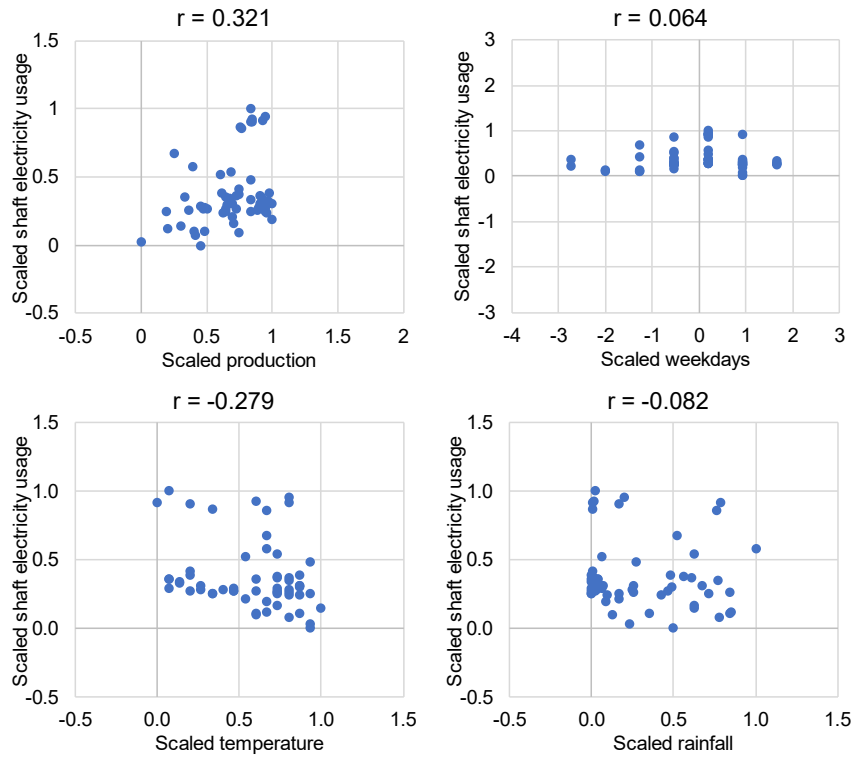


Figure B-12: Variable correlation graphs and Pearson's coefficients for the shaft electricity usage of shaft D

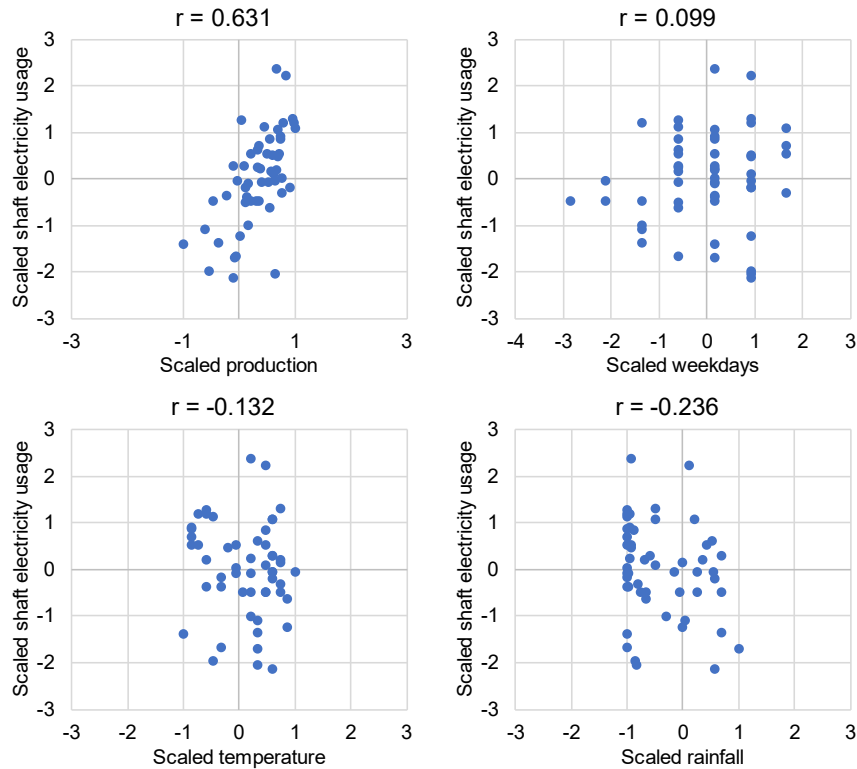


Figure B-13: Variable correlation graphs and Pearson's coefficients for the shaft electricity usage of shaft E

Figure B-9 to Figure B-13 illustrate that the most strongly correlated variable in all systems is production. This is, in most cases except for in the case of shaft D, the only variable with a

correlation coefficient larger than 0.4. Therefore, this was the only input variable considered for the modelling of shaft electricity usage for shafts A, B, C and E. Figure B-12 illustrates that none of the considered variables had absolute correlation coefficients larger than 0.4. As a result, production and temperature were both considered for the forecasting of the electricity usage of shaft D.

Model fitting and testing

With the splitting of the datasets and the selection of the relevant variables completed, it was possible to fit the models to the data. This was done with the use of k-fold cross-validation. As explained in section 2.4.3, four folds were selected with the model being trained on three of the folds and tested on the remaining fold. This was performed four times in order to remove bias that would have caused accuracy misrepresentation.

Furthermore, a variety of models were considered and compared in order to allow the selection of the most accurate model for each system. These models included power, linear and polynomial fits. The average and standard deviation of the accuracies across the four folds with the use of four performance metrics were used for model selection.

The results for the modelling of the shaft electricity usage for the five shafts are shown in Table B-2. Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3, The average columns in Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3 illustrates that for the modelling of the shaft electricity consumption of shaft B, the polynomial model performed the best on the training set and the worst on the test set. This indicates that this model is most likely overfitting to the training set. The linear model, however, performed comparatively well on the training set and the best on the test set. This model also

produced the lowest standard deviation across most metrics. The linear model was therefore selected for this system.

Table B-4, Table B-5 and Table B-6 with the best and worst performing models highlighted in green and red respectively for each metric .

Table B-2: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft A electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	6.95%	7.08%	6.51%	0.59%	0.64%	0.67%
	Test	7.77%	7.84%	7.70%	2.05%	1.86%	2.47%
R ²	Train	0.225	0.200	0.308	0.047	0.042	0.083
	Test	-0.636	-0.670	-0.629	0.945	0.927	1.266
RMSE (kWh)	Train	805 088	818 057	759 175	53 512	55 734	58 441
	Test	880 929	893 364	856 957	167 909	164 419	209 566
MAE (kWh)	Train	661 881	671 392	614 992	59 100	63 980	58 395
	Test	738 110	742 040	723 620	172 537	161 064	199 468

Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft B electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	3.88%	3.71%	3.65%	0.30%	0.22%	0.21%
	Test	3.89%	3.74%	4.07%	0.65%	0.57%	0.71%
R ²	Train	0.694	0.710	0.714	0.023	0.022	0.020
	Test	0.675	0.685	0.648	0.076	0.068	0.069
RMSE (kWh)	Train	126 542	123 157	122 174	8 853	6 401	7 163
	Test	126 099	123 180	131 623	27 498	20 654	29 347
MAE (kWh)	Train	106 347	101 849	100 170	8 645	6 372	6 287
	Test	106 109	102 711	111 572	20 646	17 298	21 793

The average columns in Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3 illustrate that for the modelling of the shaft electricity consumption of shaft B, the polynomial model performed the best on the training set and the worst on the test set. This indicates that this model is most likely overfitting to the training set. The linear model, however, performed comparatively well on the training set and the best on the test set. This model also produced the lowest standard deviation across most metrics. The linear model was therefore selected for this system.

Table B-4: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft C electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	6.61%	6.43%	6.37%	0.11%	0.11%	0.13%
	Test	7.06%	6.80%	6.94%	0.45%	0.36%	0.37%
R ²	Train	0.597	0.610	0.613	0.031	0.035	0.036
	Test	0.527	0.541	0.526	0.110	0.125	0.130
RMSE (kWh)	Train	353 289	347 104	345 652	3 674	5 733	6 181
	Test	370 402	362 817	368 233	19 812	21 038	19 381
MAE (kWh)	Train	291 404	283 576	281 692	3 197	3 821	5 136
	Test	309 832	299 725	306 738	19 498	16 367	15 440

The average columns in Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3 illustrate that for the modelling of the shaft electricity consumption of shaft B, the polynomial model performed the best on the training set and the worst on the test set. This indicates that this model is most likely overfitting to the training set. The linear model, however,

performed comparatively well on the training set and the best on the test set. This model also produced the lowest standard deviation across most metrics. The linear model was therefore selected for this system.

Table B-4 illustrates similar insights to Table B-2 indicates that the polynomial model performed the best across all folds when the average of the performance metrics was considered. This model, however, also produced the highest standard deviation across all performance metrics on training and testing sets. This high standard deviation, however, is not significantly worse than that of the other models. The polynomial model was therefore selected for the forecasting of the shaft electricity usage of shaft A.

Table B-3. The average columns indicate that the polynomial model performed the best on the training sets, while the linear model performed the best on the test sets and comparatively well on the training sets. The polynomial model also produced the highest standard deviation across folds for the majority of the metrics. Therefore, the linear model was chosen for the forecasting of the shaft electricity usage of shaft C.

Table B-5: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft D electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	12.47%	12.72%	17.30%	0.56%	0.62%	4.33%
	Test	12.30%	12.03%	18.20%	1.93%	1.30%	7.57%
R ²	Train	-0.244	-0.043	-1.335	0.073	0.053	1.202
	Test	-0.241	-0.048	-1.295	0.111	0.047	0.992
RMSE (kWh)	Train	1 088 889	997 677	1 428 216	37 421	43 369	296 305
	Test	1 059 547	972 733	1 434 689	191 163	161 778	525 196
MAE (kWh)	Train	794 794	780 631	1 105 894	35 310	44 005	256 061
	Test	781 284	741 582	1 154 078	152 927	115 285	504 814

Table B-5 indicates that the linear model produced the best average accuracies across all folds for majority of the metrics considered. Furthermore, the polynomial model produced the worst average performance across all folds as well as the highest standard deviation. Due to its overall high average performance across metrics and lowest standard deviation, the linear model was selected for forecasting the electricity consumption of shaft D.

Table B-6: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft E electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	5.81%	5.76%	5.64%	0.63%	0.69%	0.79%
	Test	6.64%	6.60%	6.85%	1.74%	1.83%	1.81%
R ²	Train	0.384	0.392	0.401	0.089	0.090	0.092
	Test	-0.115	-0.093	-0.141	0.793	0.746	0.744
RMSE (kWh)	Train	88 459	87 962	87 356	9 033	9 381	9 906
	Test	99 241	98 904	101 755	25 236	24 929	24 981
MAE (kWh)	Train	70 894	70 188	68 794	8 500	9 111	10 410
	Test	80 961	80 342	83 354	25 767	26 641	26 592

Finally, Table B-6 illustrates that the polynomial model produced the best average performance across all metrics on the training sets, but the worst average performance on the test sets. This indicates that the model is most likely overfitting to the training data. Conversely, the linear model produced the best average performance on the tests sets across all metrics with comparable performance on the training sets and comparable standard deviations. Therefore, this model was selected for the forecasting of the shaft electricity usage of shaft E.

The summary of results showing the final models selected for all systems is shown in Table 3.11.

Model verification

With the final models selected by comparing average k-fold cross-validation accuracies, it was possible to fit the models to the total training set. These final models were validated using the verification set that was split from the total dataset at the start of the process. The purpose of this is to verify whether the final models perform with similar accuracy on the training and verification sets. This would imply that the models are capable of performing well on unseen data. The final models, only for univariate models, are shown in Figure B-14.

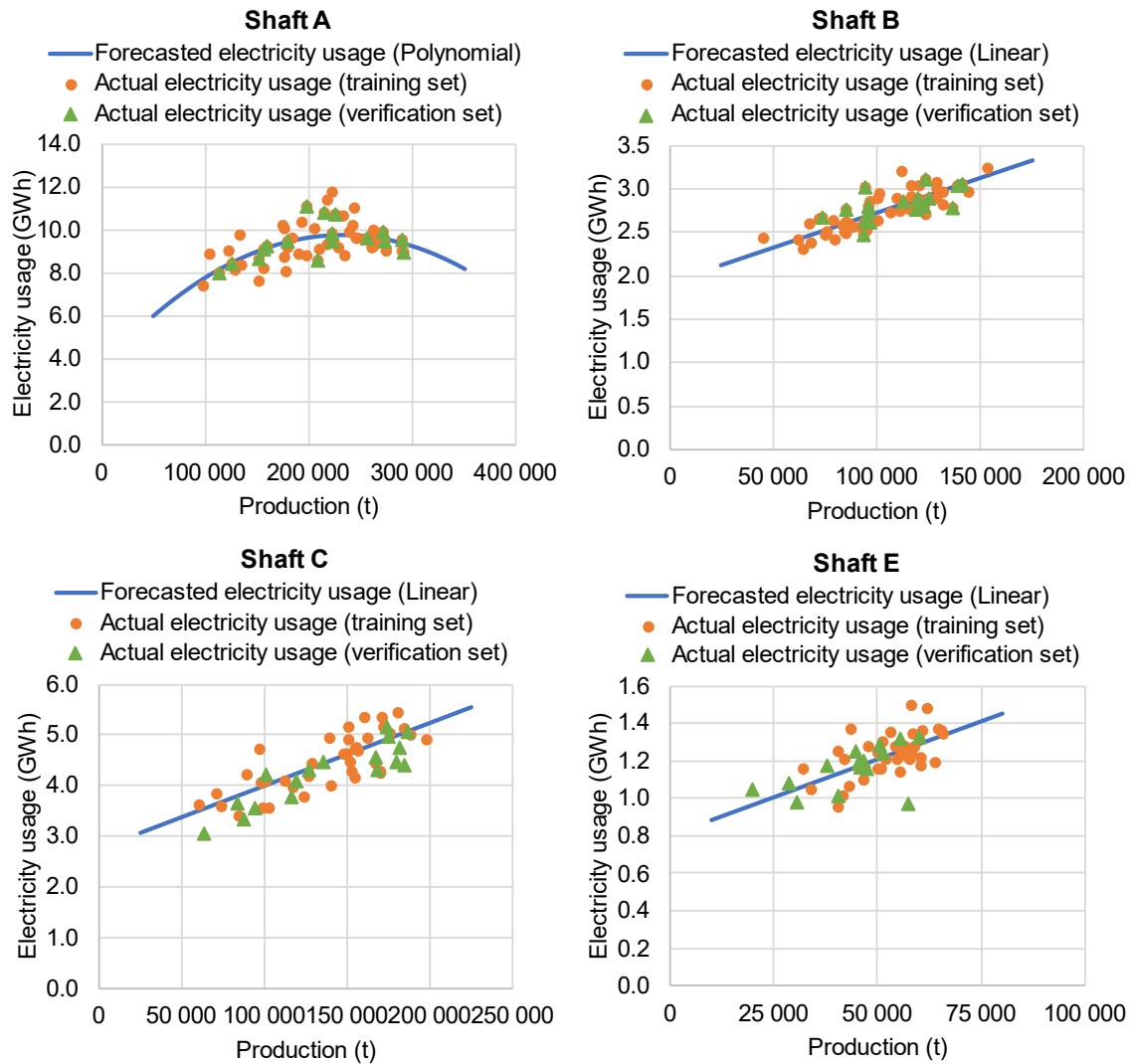


Figure B-14: Final models for the forecasting of the shaft electricity consumption of shafts A, B, C and E plotted against the training and verification sets

Figure B-14 illustrates the final forecasting models for univariate systems plotted against the training sets and verification sets for each system. Figure B-14 illustrates that the verification and training sets both span the total operating range of each of the systems. Therefore, the verification set can be considered to be representative. Table B-7 contains the training set and verification set accuracies for the models shown in Figure B-14. These accuracies were compared to ensure that the model has comparative performance on both sets.

Table B-7: Training set and verification set performance metrics of the final models forecasting the electricity consumption of shafts A to E

Performance metric	MAPE (%)	R ²	RMSE (kWh)	MAE (kWh)
Shaft A				
Training set	6.71	0.300	730 250	632 924
Verification set	4.01	0.481	589 662	394 487

Performance metric	MAPE (%)	R ²	RMSE (kWh)	MAE (kWh)
Shaft B				
Training set	3.70	0.713	123 308	101 718
Verification set	3.97	0.686	144 284	110 925
Shaft C				
Training set	6.50	0.609	349 525	286 373
Verification set	7.02	0.664	340 361	280 239
Shaft D				
Training set	12.35	-0.016	985 308	760 301
Verification set	10.66	-0.359	1 078 517	681 640
Shaft E				
Training set	5.89	0.366	90 521	71 795
Verification set	6.37	0.263	98 401	68 419

Table B-7 indicates that all models produced similar performance across all metrics on both the training and verification sets. With some models producing better performance on the verification set than on the training set, these are specifically the models for the electricity usage of shaft A and shaft D. This indicates that all of the models are capable of producing similar performance on unseen data and can therefore be used for forecasting the system electricity usage.

A final step in the verification of the final models is the verification of the initial model assumptions. This is performed by evaluating the similarity of the predicted data points to the actual data points used for training and verification with the use of a scatter plot. Furthermore, a residual plot and distribution plot of residuals is used to validate the assumption of heteroskedasticity and normally distributed residuals. These plots are shown in Figure B-15, Figure B-16, Figure B-17, Figure B-18 and Figure B-19

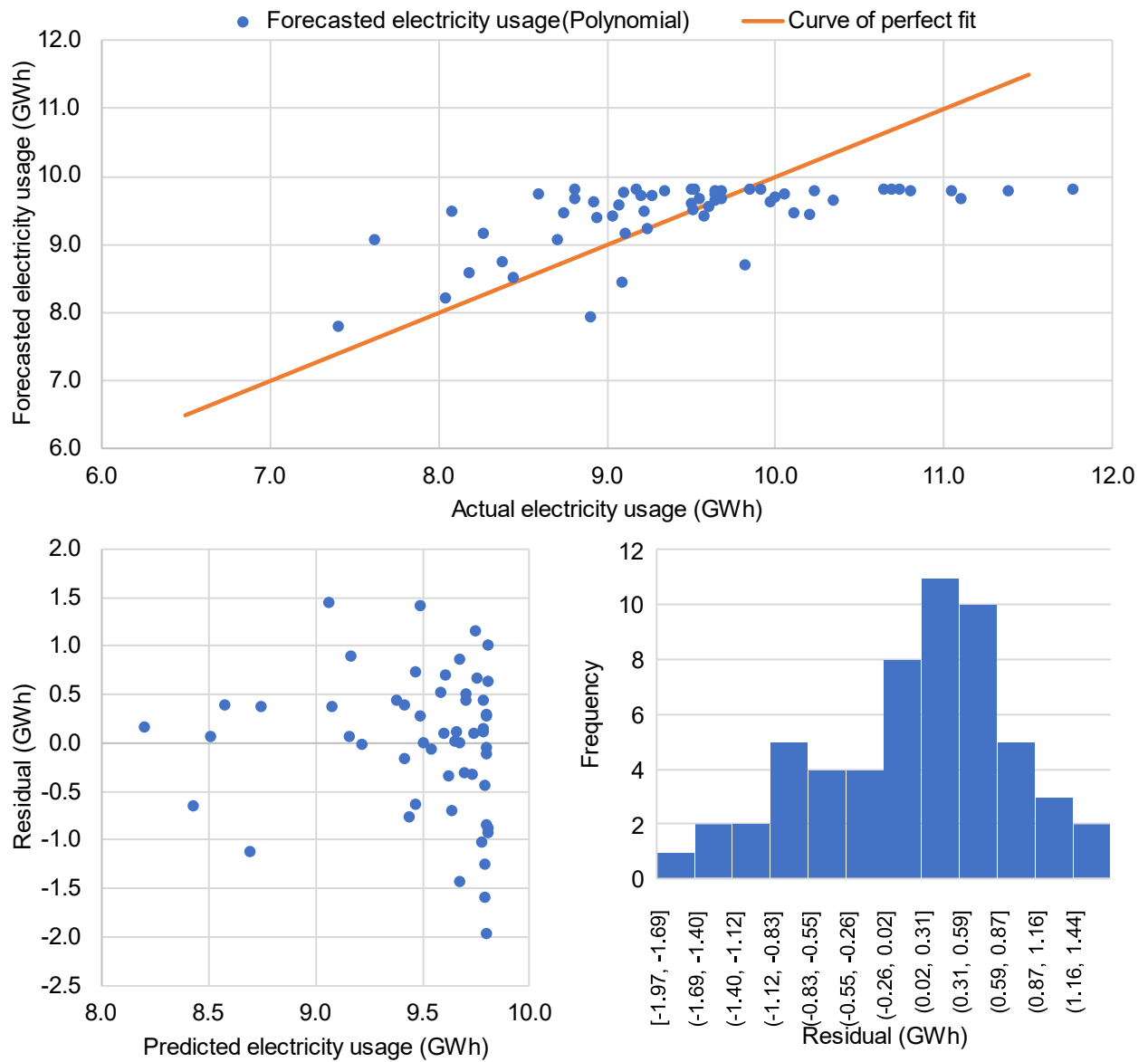


Figure B-15: Scatter plot ($R^2=0.589$), residual plot and residual distribution plot of shaft A electricity consumption to evaluate model assumptions

Figure B-15 illustrates that the relationship between the actual and forecasted electricity usage for shaft A yields a correlation coefficient of 0.589. This implies that the forecasted values are close to the actual values but not incredibly accurate as this value indicates moderate correlation according to Table 1.8. This is due to the large variability in the actual electricity usage of the shaft at approximately 9.9 GWh forecasted electricity usage.

The residual plot on the bottom left of Figure B-15 illustrates that there is not relationship between the variance and the predicted electricity usage. This satisfies the assumption of heteroskedasticity. The plot on the bottom right of Figure B-15 indicates that the residuals follow a normal distribution with the range of the residuals existing from -1.98 GWh to 1.44 GWh.

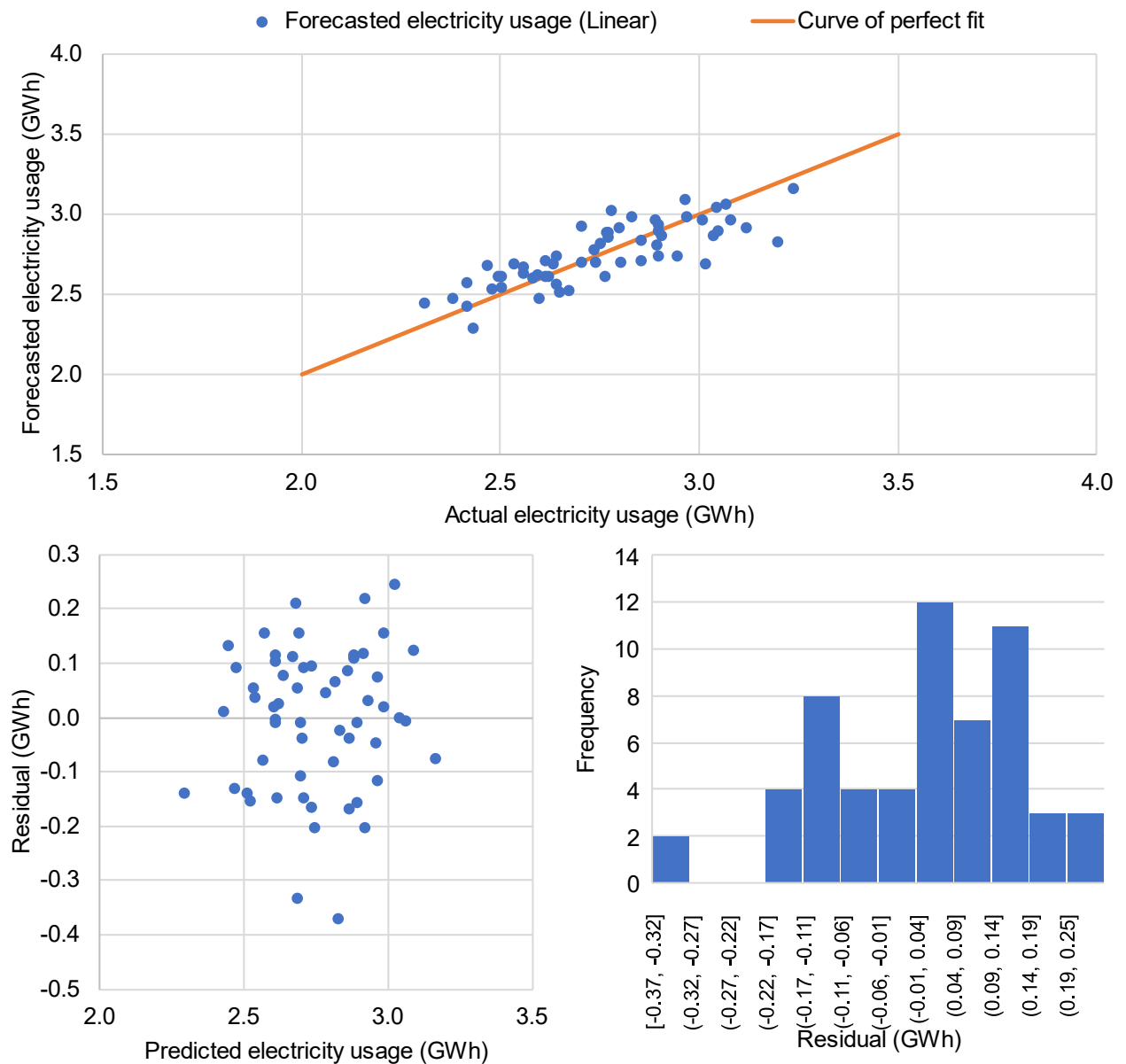


Figure B-16: Scatter plot ($R^2=0.807$), residual plot and residual distribution plot of shaft B electricity consumption to evaluate model assumptions

Figure B-16 illustrates that the relationship between the actual and forecasted electricity usage for shaft B yields a correlation coefficient of 0.807. This implies this model performs well since the forecasted values have a strong correlation to the actual values. This graph also visually illustrates that these values follow the line of perfect fit with a constant error, either above or below the line.

The distribution plot on the bottom right of Figure B-16 confirms the constant error observation shown in the scatter plot. This is shown as the residuals range from -0.22 GWh to 0.25 GWh (not considering the outlier points between -0.37 GWh and -0.32 GWh) with a normal distribution and majority of the points centred around zero. This also satisfies the assumption of normal residual distribution. Finally, the residual plot at the bottom left of Figure B-16 illustrates that there is no

relationship between the residuals and the predicted values. This satisfies the assumption of constant variance of least squares regression.

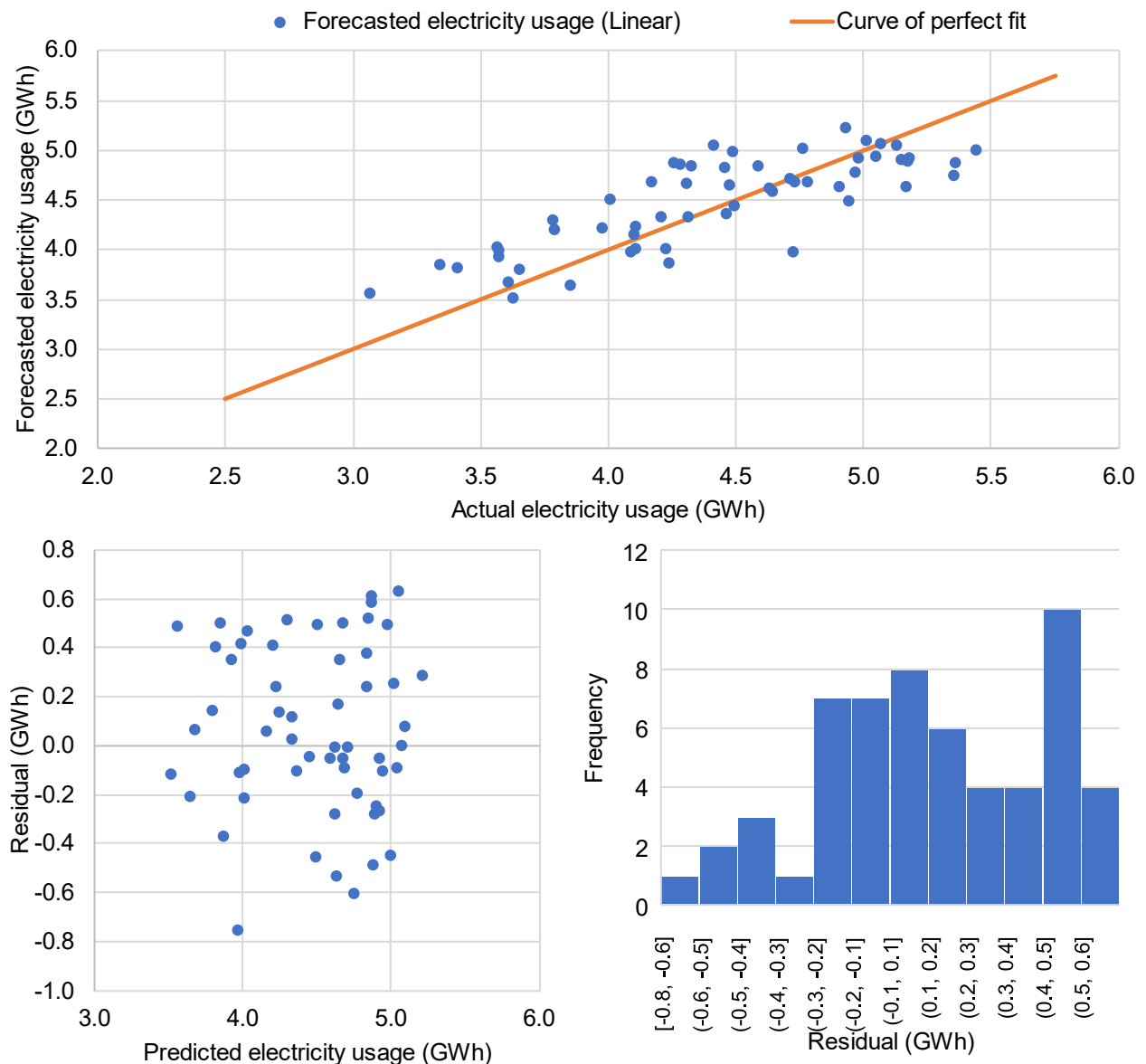


Figure B-17: Scatter plot ($R^2=0.807$), residual plot and residual distribution plot of shaft C electricity consumption to evaluate model assumptions

Figure B-17 illustrates that the relationship between the actual and forecasted electricity usage for shaft C yields a correlation coefficient of 0.807. This implies this model performs well since the forecasted values have a strong correlation to the actual values. This graph also visually illustrates that these values follow the line of perfect fit with a constant error, either above or below the line.

The distribution plot on the bottom right of Figure B-17 confirms the constant error observation shown in the scatter plot. This is shown as the residuals range from -0.6 GWh to 0.6 GWh (not considering the outlier points between -0.8 GWh and -0.6 GWh) with a normal distribution and

majority of the points centred around zero. This also satisfies the assumption of normal residual distribution. Finally, the residual plot at the bottom left of Figure B-17 illustrates that there is no relationship between the residuals and the predicted values. This satisfies the assumption of constant variance of least squares regression.

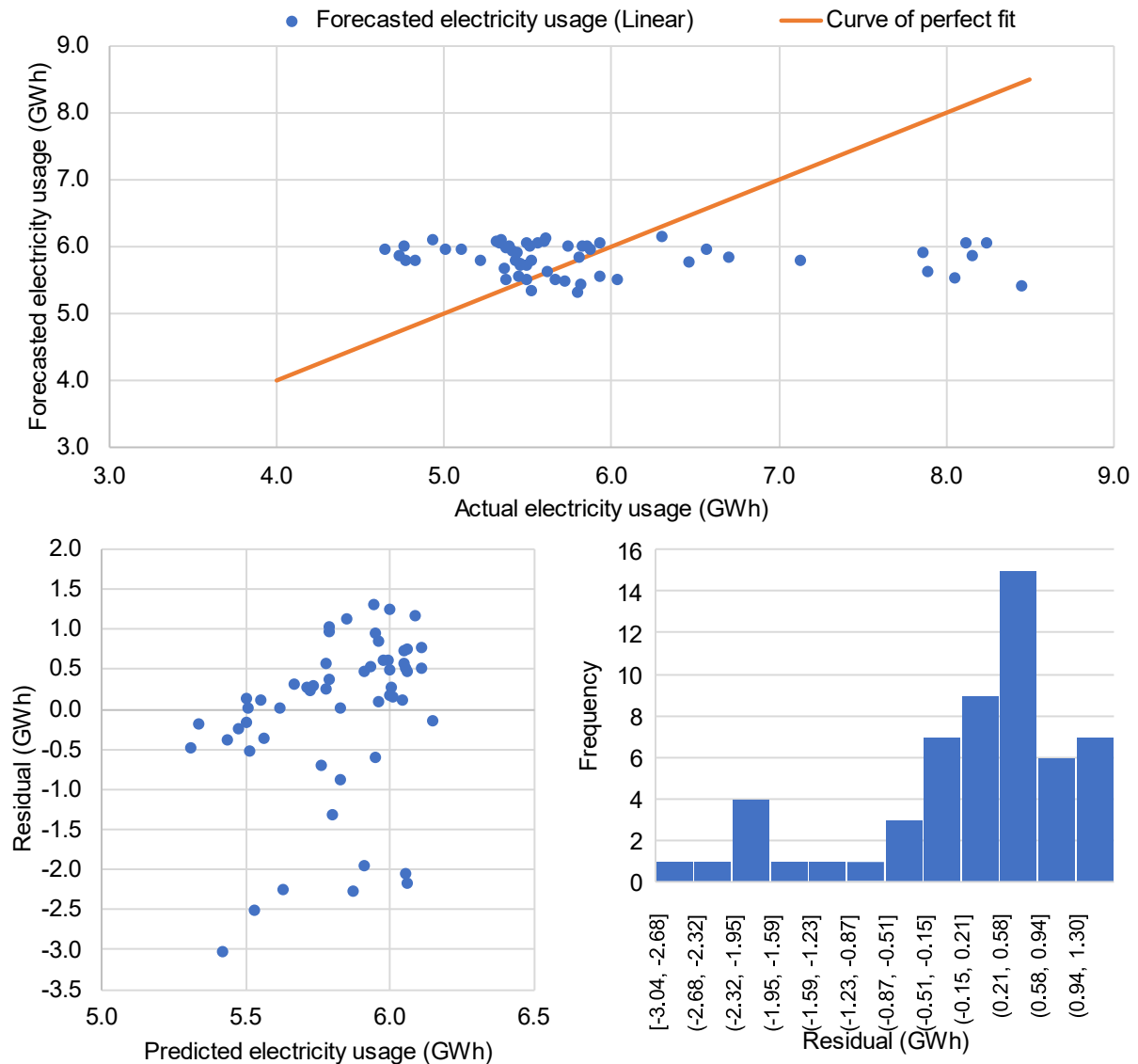


Figure B-18: Scatter plot ($R^2=-0.113$), residual plot and residual distribution plot of shaft D electricity consumption to evaluate model assumptions

The scatter plot in Figure B-18 illustrates that there is no relationship between the actual values and predicted values for the forecasting of the electricity consumption of shaft D. This model was, however, fit on variables that were not correlated to the output variable – so this poor performance is expected.

The distribution of residuals at the bottom of Figure B-18 also illustrates that the residuals are skewed to the left and do not follow a normal distribution. Finally, the residual plot also shows that

the residuals do not follow constant variance. This model should be revisited with new input variables in order to produce a model with better performance.

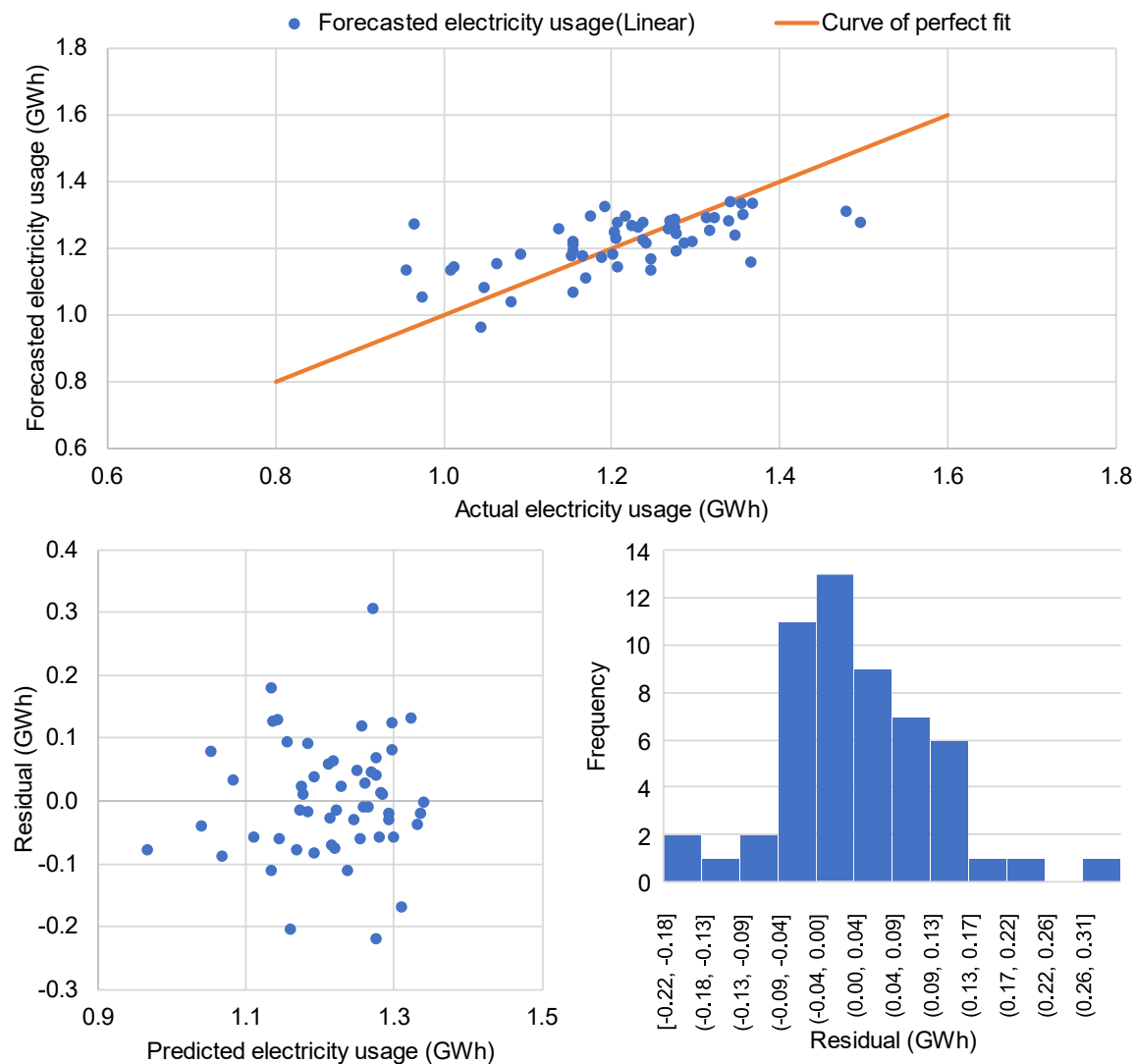


Figure B-19: Scatter plot ($R^2=0.631$), residual plot and residual distribution plot of shaft E electricity consumption to evaluate model assumptions

The scatter plot between the actual and predicted values shown in Figure B-19 illustrates that this model performs adequately at predicting the electricity usage, with a correlation coefficient of 0.631. This model also produces a distribution plot of residuals with a normal distribution shown at the bottom right of Figure B-19 – satisfying the model assumption. This model, much like those of shaft B and C produces residuals that are centred around zero with an error range from - 0.22 GWh to 0.31 GWh.

Finally, the residual plot shown at the bottom left of Figure B-19 illustrates that the model residuals are constant and do not vary with predicted output. This satisfies the assumption of constant variance of least squares regression.

Appendix C – Compressed air electricity consumption models

Data collection and preparation

Outlier detection

The Mahalanobis distance was calculated for all data points for the compressed air electricity consumption forecasting of all the shafts considered. This was used together with the boxplot method to identify outlier data points from the multivariate datasets. The graphs illustrating the distance outliers are shown in Figure C-1 below.

Compressed air electricity usage

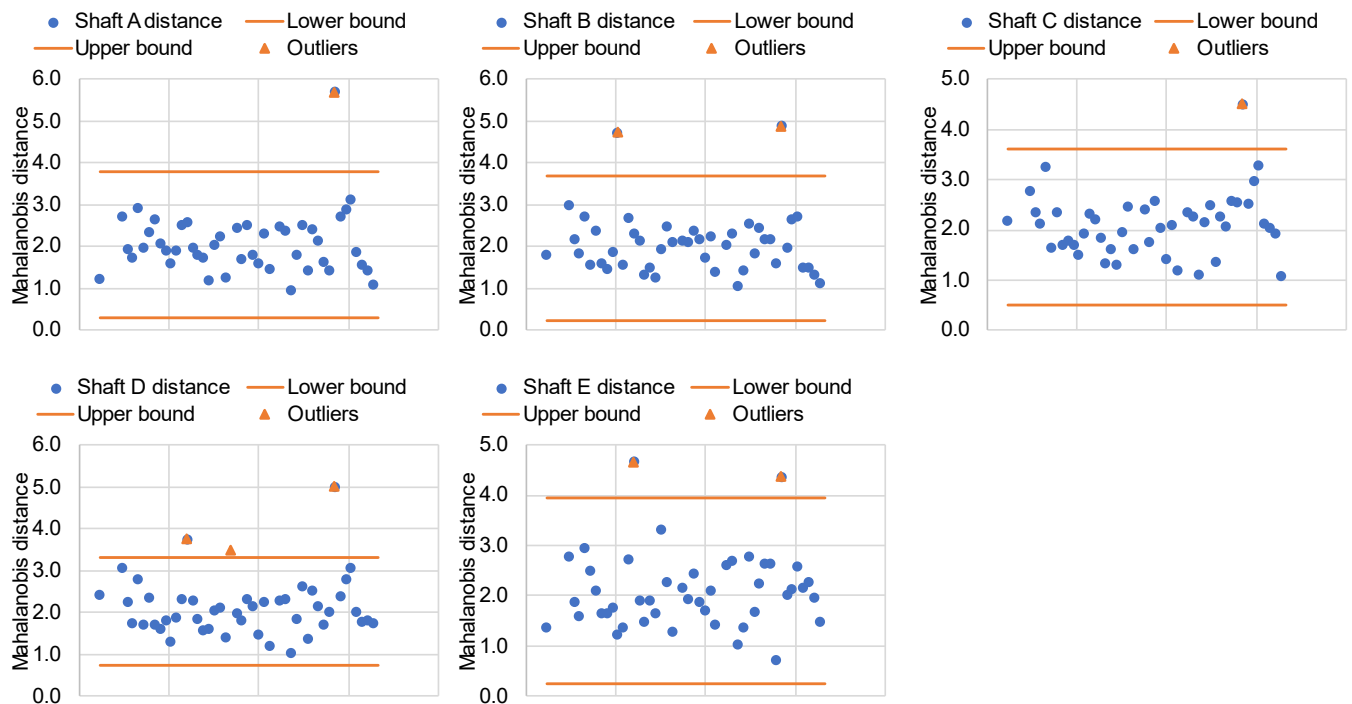


Figure C-1: Outlier detection using the Mahalanobis distance and boxplot method on compressed air electricity usage for shafts A-E

Figure C-1 illustrates the retained data points, the outlier data points as well as the quartile 1 and quartile 3 boundaries used to identify the outliers. The graphs indicate that outliers were identified for all shafts considered. Furthermore, the graphs indicate that the distances of the retained data points are clustered closely together. The removal of the outliers serves to ensure that the final data set is clean, reducing the amount of unnecessary error that would result from the presence of non-representative data points.

Data transformation

Once the outliers were removed, it was necessary to scale the data. This is done in order to prevent large differences in the units of input variables from causing weighting differences between the variables. Two scaling methods were considered, min-max scaling and z-score scaling. The former is used for datasets that are not normally distributed, while the latter is used on normally distributed data sets. Therefore, prior to scaling, it was necessary to evaluate the data distributions of all variables that were scaled – this is shown in Figure C-2 and Figure C-3.

Input variables

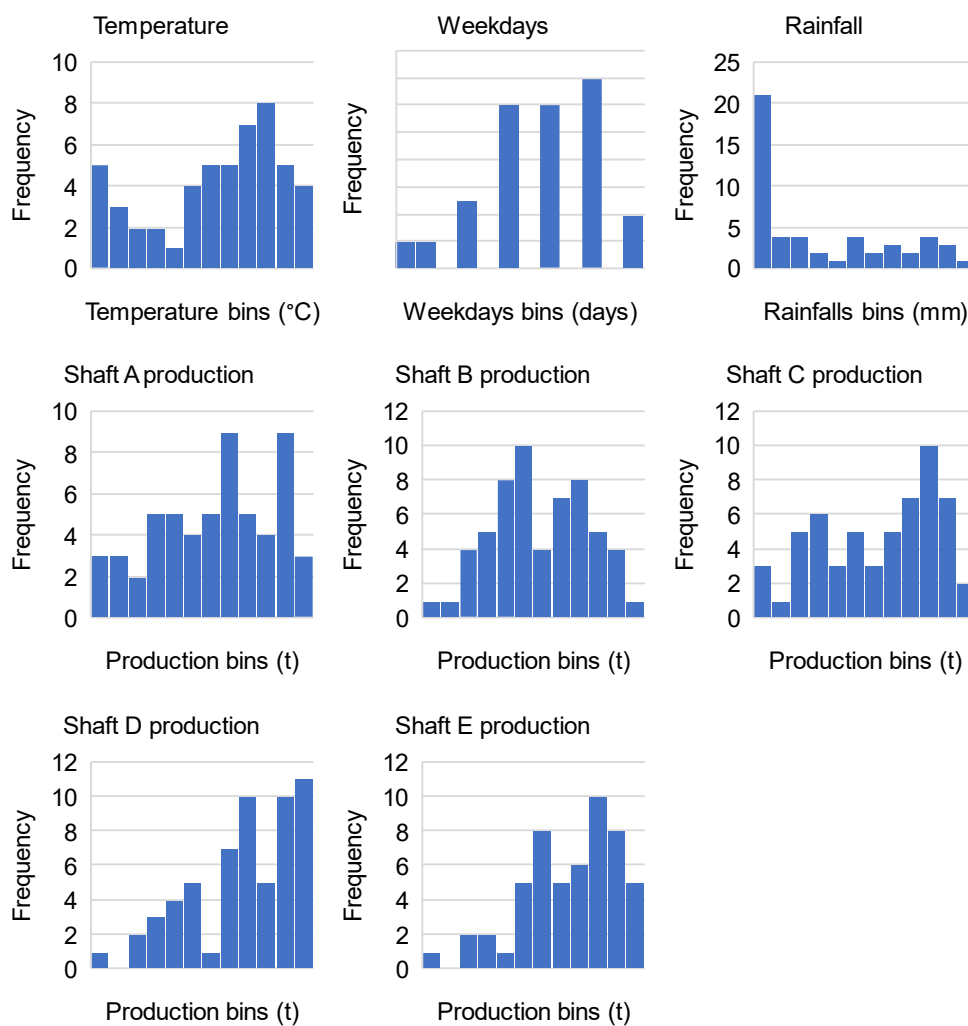


Figure C-2: Input variable distributions for the determination of the most applicable scaling method for compressed air electricity usage

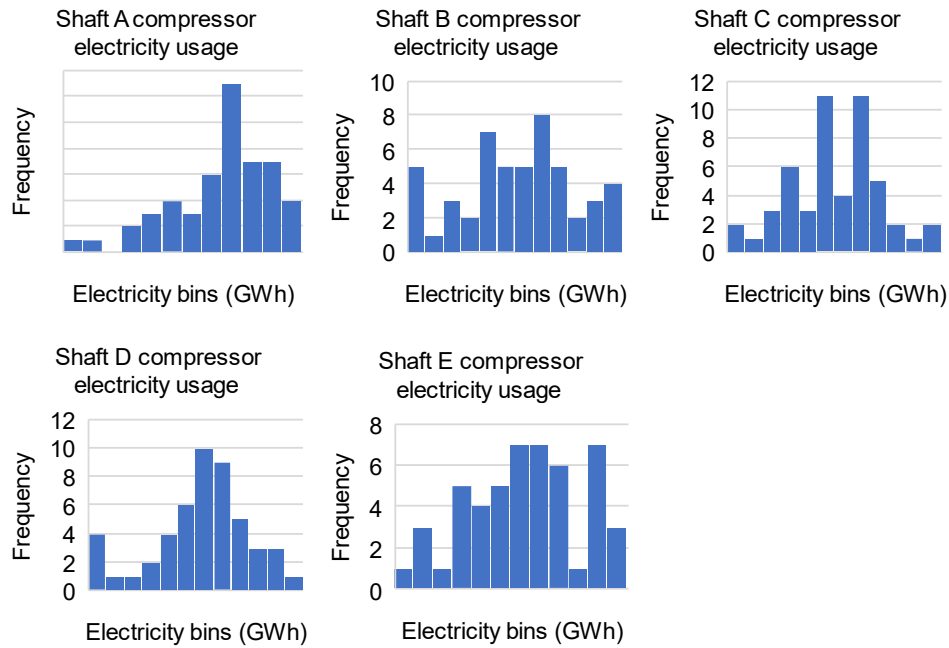


Figure C-3: Output variable distributions for the determination of the most applicable scaling method for compressed air electricity usage

Figure C-2 and Figure C-3 show the distribution of the input and output variables that were used to generate models for forecasting the shaft electricity usage. The scaling method used for each variable is dictated by the distribution of the data and is shown in Table B-1.

Table C-1: Summary of scaling methods applied to the variables used for the modelling of the compressed air electricity consumption

Variable	Min-max scaling	Z-score scaling
Temperature	X	
Weekdays		X
Rainfall	X	
Shaft A production		X
Shaft B production		X
Shaft C production		X
Shaft D production	X	
Shaft E production	X	
Shaft A compressed air electricity consumption	X	
Shaft B compressed air electricity consumption		X
Shaft C compressed air electricity consumption		X
Shaft D compressed air electricity consumption		X
Shaft E compressed air electricity consumption		X

Data splitting

Once the data was cleaned of outliers and transformed, it was necessary to split it into train/test and verification sets. The datasets were first randomised in order to ensure that split datasets were representative of the total dataset. The resulting train/test set was used with k-fold cross-verification to train the resulting model.

Furthermore, the separation of a verification set was necessary to ensure that the resulting model was tested on unseen data. The data points that comprised the train/test and verification sets for each shaft are shown in Figure C-4 to Figure C-8 below.

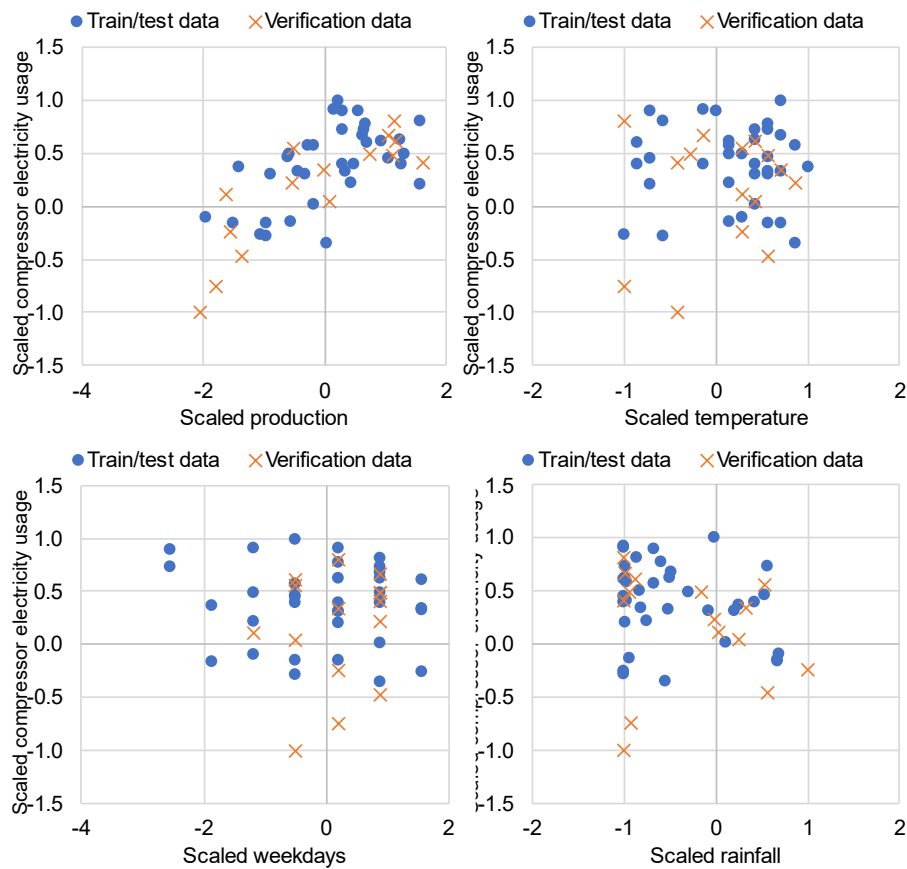


Figure C-4: Data splits showing the train/test and verification sets for shaft A

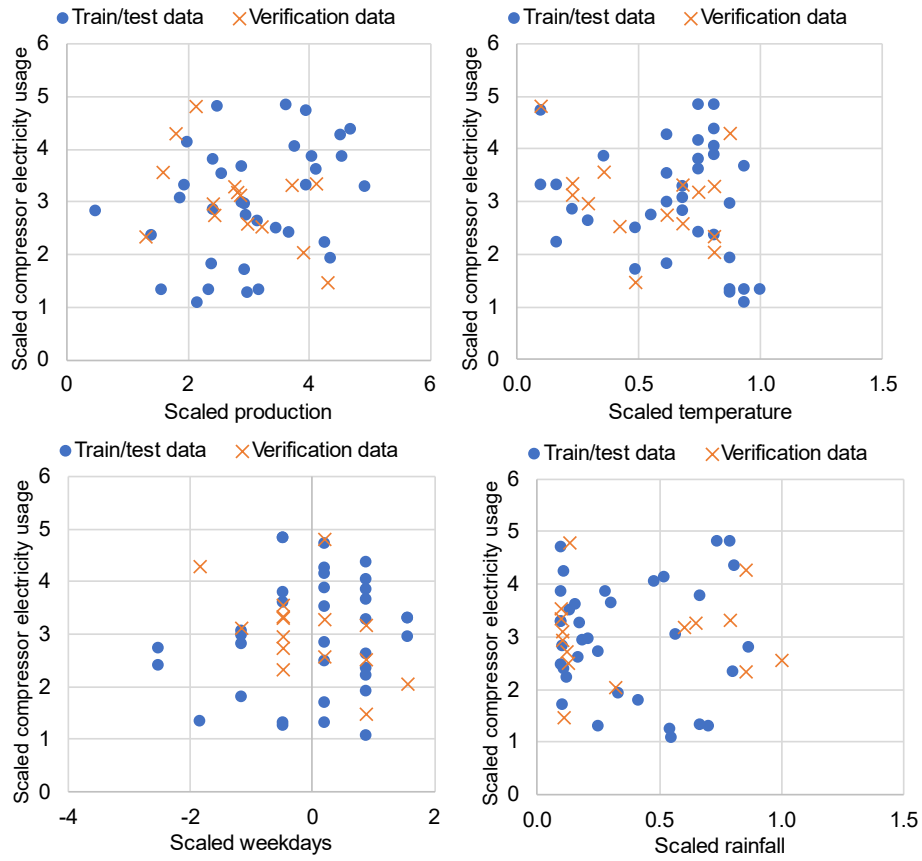


Figure C-5: Data splits showing the train/test and verification sets for shaft B

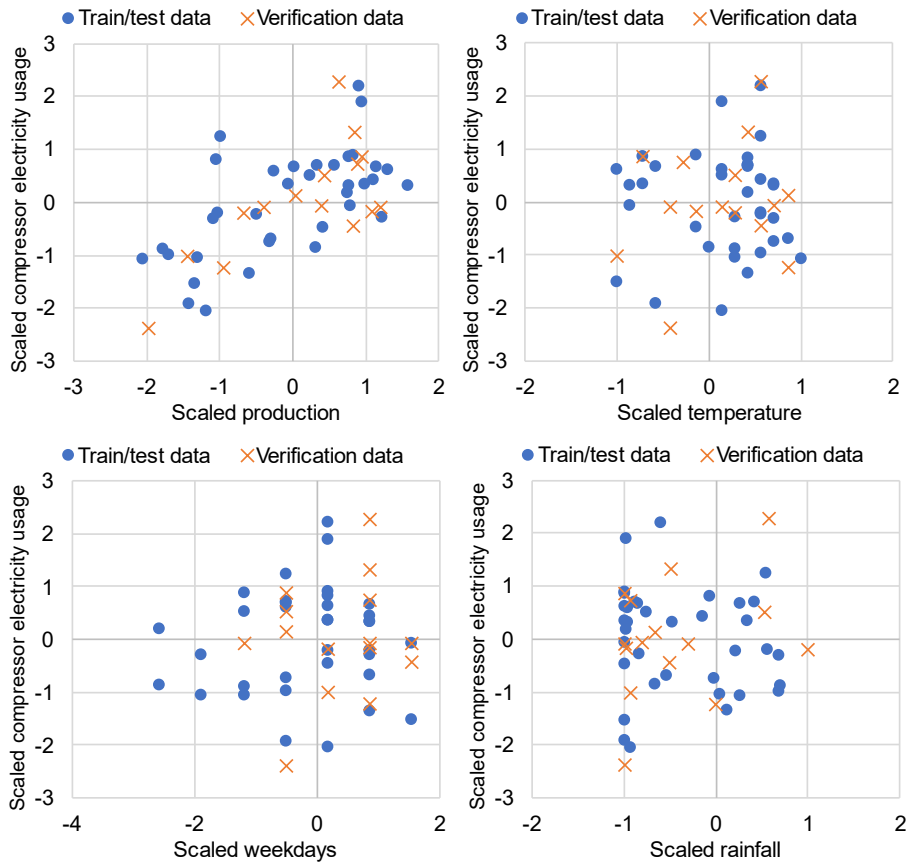


Figure C-6: Data splits showing the train/test and verification sets for shaft C

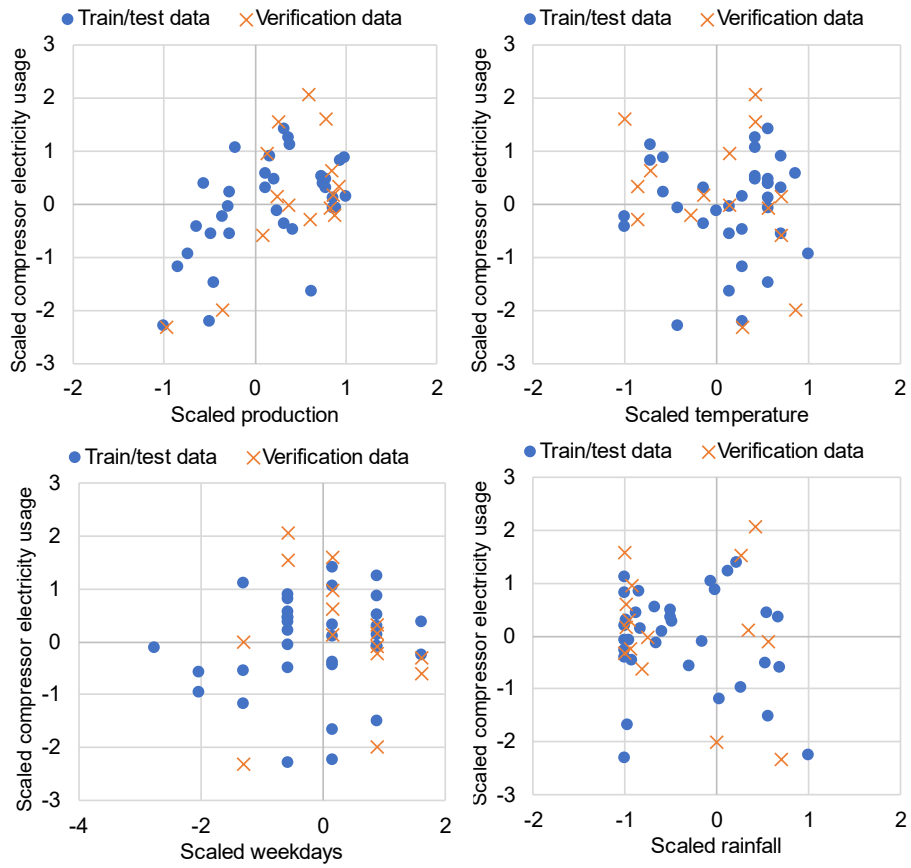


Figure C-7: Data splits showing the train/test and verification sets for shafts D

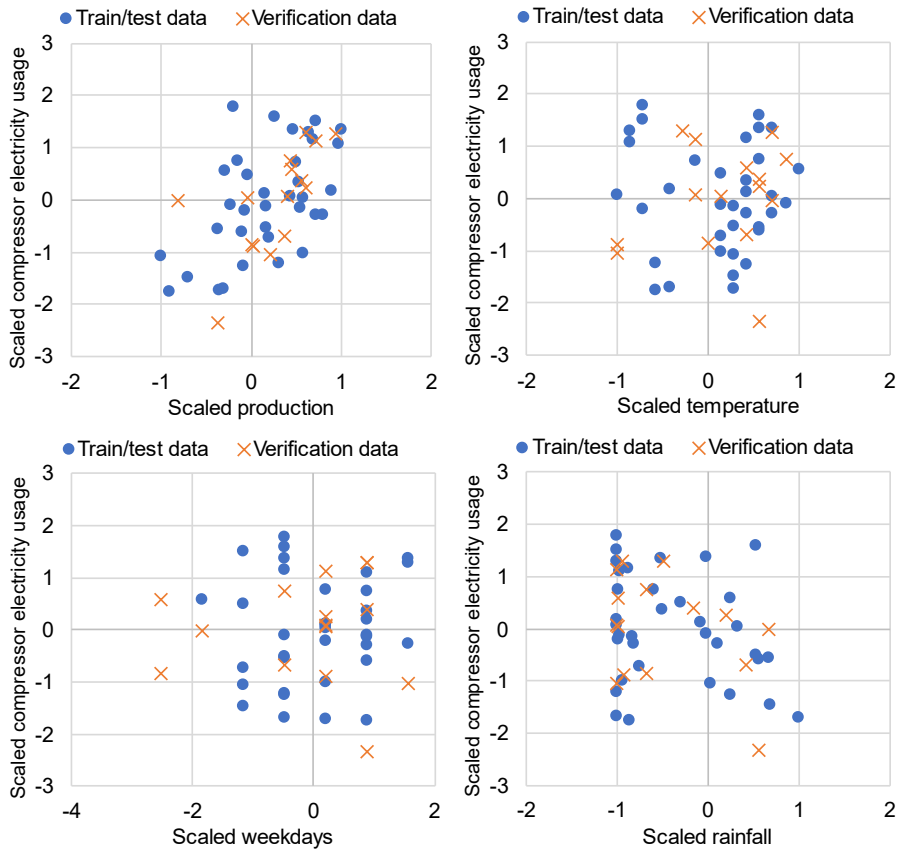


Figure C-8: Data splits showing the train/test and verification sets for shaft E

Figure C-4 to Figure C-8 illustrate that the values of the variables in the train/test and verification sets for each of the shafts span the range of each variable. This shows that the values used in each set are representative of the dataset and will result in accuracies that can be considered valid. This check is necessary in order to ensure that the verification and train/test data sets are similar in distribution.

Model development

Variable correlation assessment

After cleaning and splitting of the datasets it was possible to begin with the model development step of the process. This step began with the determination of input variables that were most strongly correlated to each output variable to be forecasted. This determination was performed with the use of Pearson's correlation coefficient. According to Table 1.8, variables with absolute correlation coefficients less than 0.4 are considered to have weak to negligible correlation.

As such, only input variables with absolute correlation coefficients larger than or equal to 0.4 to the output variables were considered. It is necessary to remove variables with low correlation as the inclusion of these can have an adverse effect on model performance. The scatter plots between variables and associated Pearson's correlation coefficients are shown in Figure C-9, Figure C-10, Figure C-11, Figure C-12 and Figure C-13 below.

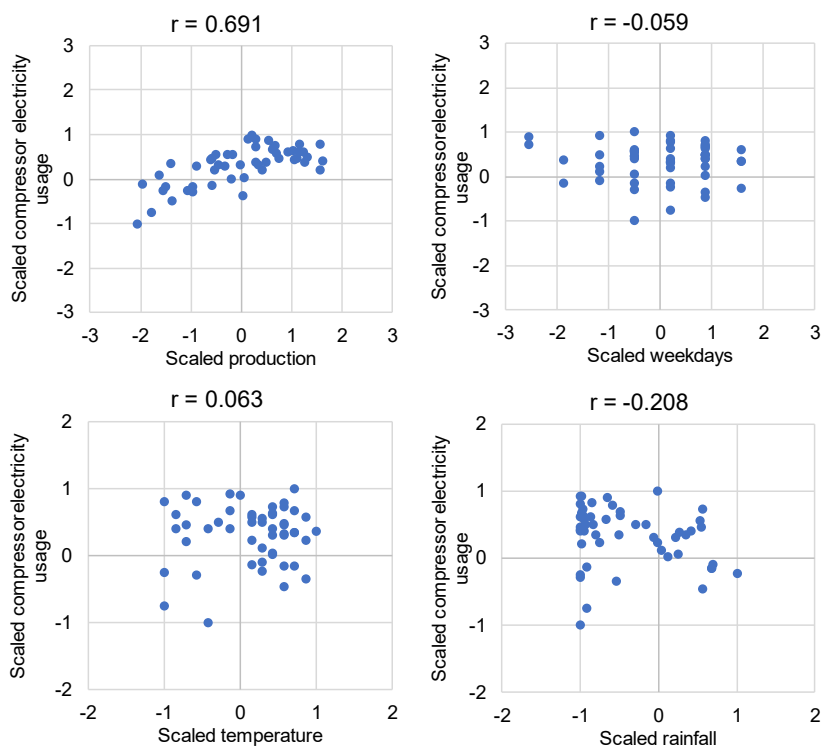


Figure C-9: Variable correlation graphs and Pearson's correlation coefficients for the compressed air electricity usage of shaft A

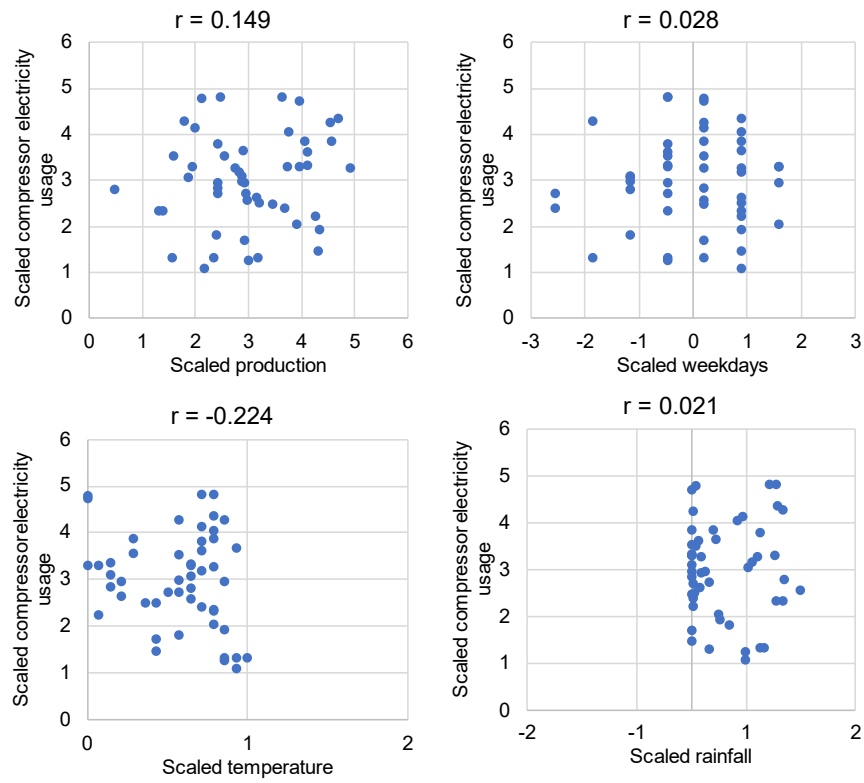


Figure C-10: Variable correlation graphs and Pearson's correlation coefficients for the compressed air electricity usage of shaft B

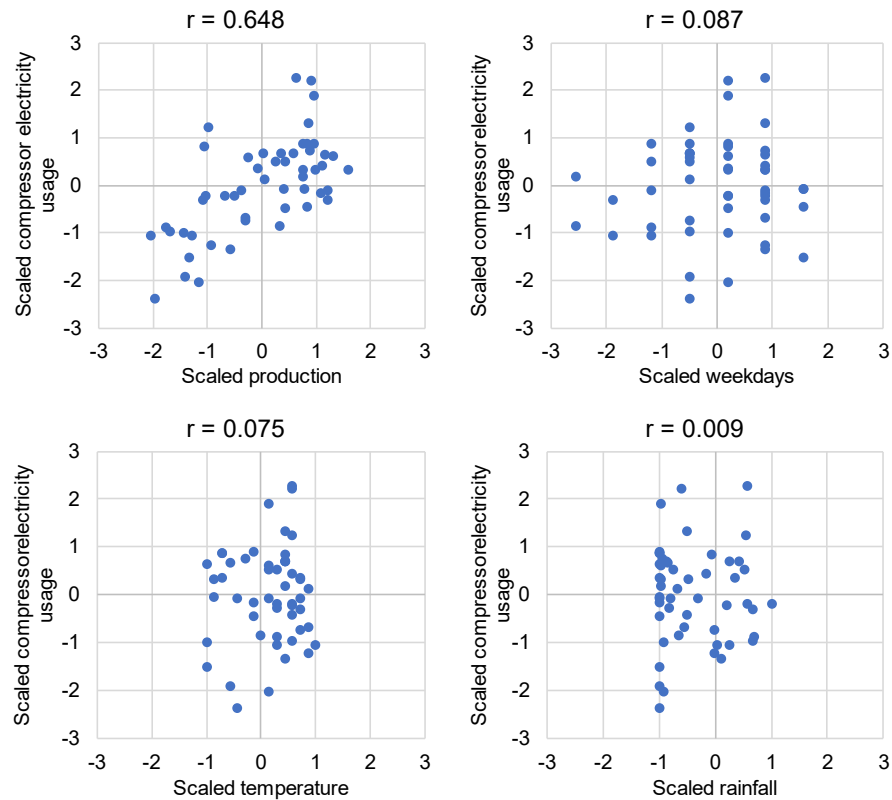


Figure C-11: Variable correlation graphs and Pearson's correlation coefficients for the compressed air electricity usage of shaft C

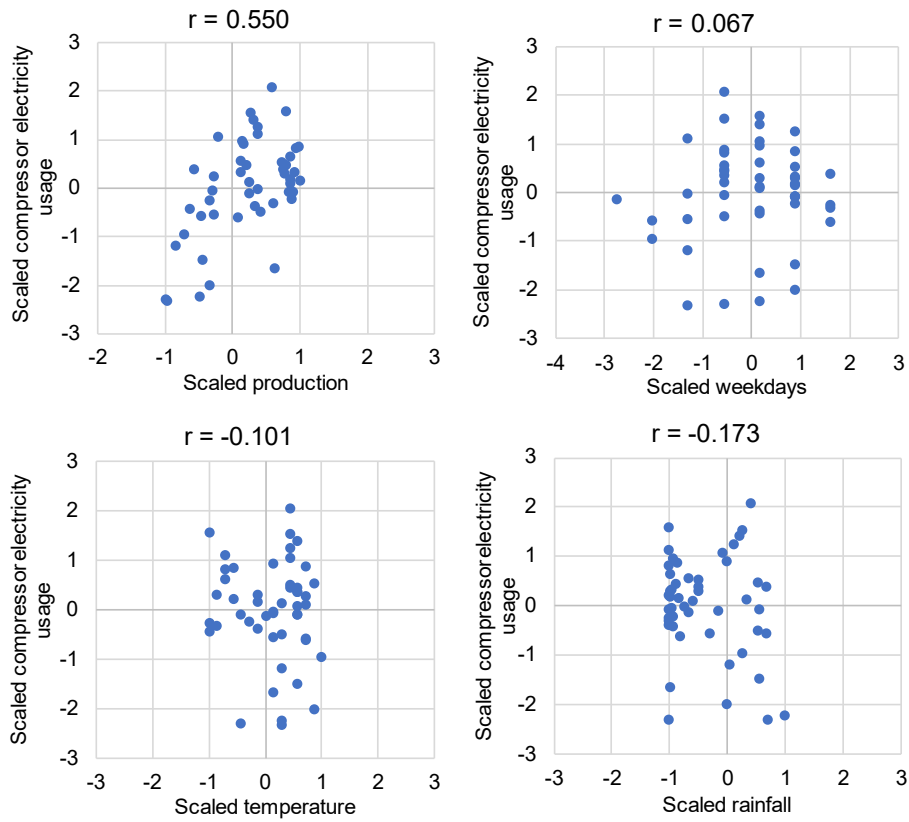


Figure C-12: Variable correlation graphs and Pearson's correlation coefficients for the compressed air electricity usage of shaft D

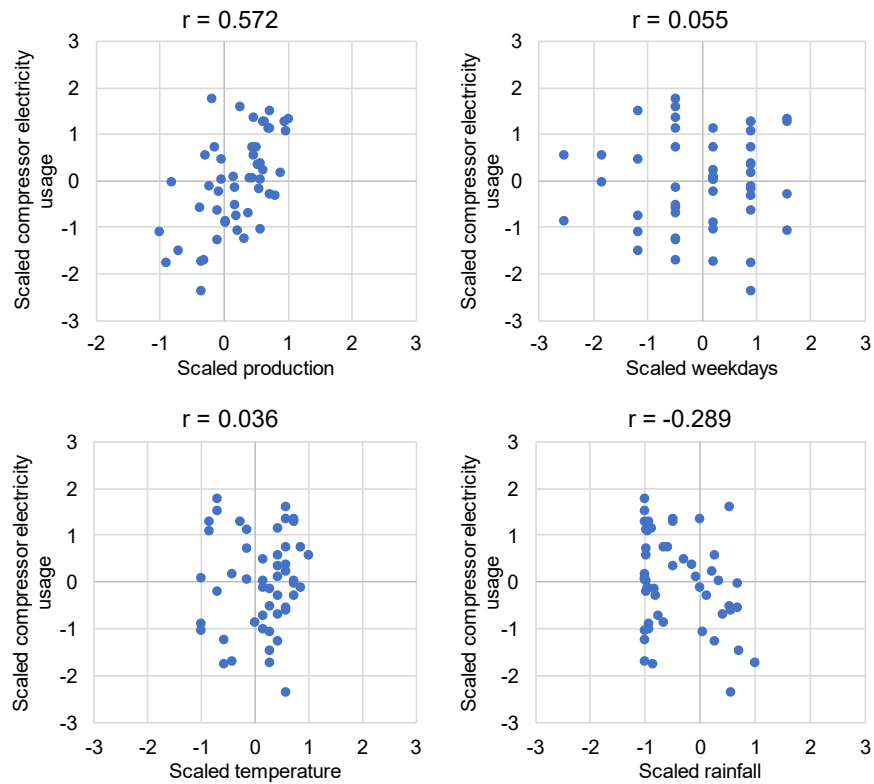


Figure C-13: Variable correlation graphs and Pearson's correlation coefficients for the compressed air electricity usage of shaft E

Figure C-9 to Figure C-13 illustrate that the most strongly correlated variable in all the systems, except shaft B, is production. Production produces an absolute Pearson's correlation coefficient larger than 0.4 for all shafts except for shaft B where the most strongly correlated variable was found to be rainfall ($r=-0.224$). Therefore, production was selected as the only input variable for modelling the compressed air electricity consumption of shafts A, C, D and E. While rainfall and production were both selected for the modelling of shaft B's compressed air electricity usage.

Model fitting and testing

With the splitting of the datasets and the selection of the relevant variables completed, it was possible to fit the models to the data. This was done with the use of k-fold cross-validation. As explained in section 2.4.3, four folds were selected with the model being trained on three of the folds and tested on the remaining fold. This was performed four times in order to remove bias that would have caused accuracy misrepresentation.

Furthermore, a variety of models were considered and compared in order to allow the selection of the most accurate model for each system. These models included power, linear and polynomial fits. The average and standard deviation of the accuracies across the four folds with the use of four performance metrics were used for model selection.

The results for the modelling of the compressed air electricity usage for the five shafts are shown in Table C-2, Table C-3, Table C-4, Table C-5 and Table C-6 with the best and worst performing models highlighted in green and red respectively for each metric.

Table C-2: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft A compressed air electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	6.27%	6.32%	6.10%	0.48%	0.60%	0.42%
	Test	6.54%	6.51%	6.75%	1.62%	2.05%	1.42%
R ²	Train	0.332	0.315	0.371	0.052	0.073	0.049
	Test	0.122	0.110	0.085	0.331	0.439	0.327
RMSE (kWh)	Train	623 301	631 013	604 752	41 787	52 234	39 570
	Test	649 515	654 083	658 359	128 356	163 681	125 066
MAE (kWh)	Train	519 988	523 186	501 684	34 409	43 353	31 614
	Test	541 025	538 503	552 628	111 051	134 274	101 074

Table C-2 illustrates the average and standard deviation of the performance of the four models across the four performance metrics for the compressed air electricity consumption modelling of shaft A. Table C-2 indicates that the polynomial model produced the best average performance

on the training sets and the worst performance on the test sets. The performance of this model on the test sets was not significantly worse than the other models considered, however. Furthermore, the linear model produced the best performance on the test sets and the worst performance on the training sets.

Additionally, the polynomial model produced the lowest standard deviation across training folds, while the linear model produced the highest standard deviation. The polynomial model performed the best on the training sets, competitively (although worst) on the testing sets and produced the lowest standard deviation across folds. As such this model was selected for this forecasting the electricity consumption of this system.

Table C-3: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft B compressed air electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	14.26%	22.45%	254.69%	1.82%	8.73%	114.97%
	Test	14.91%	23.38%	268.96%	2.63%	10.11%	112.97%
R ²	Train	-1.806	-6.288	-1116.725	0.596	5.048	1053.084
	Test	-2.487	-8.332	-1189.781	1.404	6.830	827.387
RMSE (kWh)	Train	516 230	788 261	9 285 534	66 429	278 699	4 185 177
	Test	525 047	811 004	9 698 108	82 283	337 126	4 169 155
MAE (kWh)	Train	440 121	684 856	7 584 859	57 381	256 119	3 434 434
	Test	457 987	717 530	7 978 300	90 752	322 615	3 310 652

Table C-3 illustrates the performance of the models across all performance metrics for the compressed air electricity consumption forecasting of shaft B. This table clearly indicates that the power model produced the best average performance across all performance metrics on both the training and test sets. Additionally, this model produced the lowest standard deviation across training folds on both training and testing sets for all metrics. Therefore, the power model was chosen for the compressed air electricity forecasting of shaft B.

Table C-4: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft C compressed air electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	8.04%	8.05%	7.99%	0.86%	0.82%	0.83%
	Test	8.64%	8.62%	8.75%	2.29%	2.36%	2.33%
R ²	Train	0.392	0.390	0.400	0.071	0.074	0.070
	Test	0.207	0.186	0.195	0.200	0.209	0.188
RMSE (kWh)	Train	655 267	656 683	651 388	56 420	56 949	56 782
	Test	686 360	693 202	692 566	134 475	129 859	129 583
MAE (kWh)	Train	519 794	520 684	513 340	50 050	46 994	47 462
	Test	558 626	558 859	563 697	131 047	135 697	130 800

Table C-4 illustrates the performance of the models across all performance metrics for the compressed air electricity consumption forecasting of shaft C. Table C-4 initially indicates that the average performance of all models on training and test sets were similar across all metrics. This similarity in value indicates that the models performed with similar accuracy. The polynomial model performed slightly better than the others on the training sets and the power model on the test sets.

A similar behaviour can be seen when comparing the standard deviations shown in Table C-4. It can be seen that the standard deviation of the polynomial model across most metrics is the lowest on the test sets, while that for the power model is the lowest on the training sets. Due to its comparative average performance on the test sets, best performance on the training sets and lowest standard deviation on the test sets, the polynomial model was selected for this system.

Table C-5: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft D compressed air electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	6.55%	6.72%	5.81%	0.67%	0.74%	0.51%
	Test	7.30%	7.37%	6.48%	2.18%	2.26%	2.18%
R ²	Train	0.301	0.267	0.399	0.100	0.092	0.128
	Test	-0.583	-0.571	-0.350	1.140	1.103	0.907
RMSE (kWh)	Train	425 921	436 743	392 367	34 366	37 276	28 600
	Test	465 804	470 612	431 589	113 096	114 656	106 122
MAE (kWh)	Train	335 201	341 510	294 801	33 339	36 050	23 345
	Test	372 560	374 450	326 412	104 335	107 761	104 888

Table C-5 illustrates the performance of the models across all performance metrics for the compressed air electricity consumption forecasting of shaft D. Table C-5 indicates that the polynomial model produced the best average performance across all metrics on both the training and test sets. Furthermore, this model also produced the lowest standard deviation for the majority of the metrics on the training and test sets. As such, the polynomial model was selected for the compressed air electricity forecasting of shaft D.

Table C-6: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft D compressed air electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	14.98%	15.22%	15.18%	0.61%	0.68%	0.69%
	Test	15.52%	15.88%	18.36%	1.87%	2.07%	4.25%
R ²	Train	0.283	0.286	0.298	0.102	0.101	0.095
	Test	0.126	0.118	-0.011	0.253	0.264	0.143
RMSE (kWh)	Train	210 222	209 791	208 033	11 049	11 314	10 268
	Test	210 995	211 818	231 984	31 125	32 588	42 171
MAE (kWh)	Train	171 255	171 111	170 265	7 255	7 615	7 206
	Test	176 565	177 428	199 907	20 376	21 819	33 324

Table C-6 illustrates the performance of the models across all performance metrics for the compressed air electricity consumption forecasting of shaft E. Table C-6 illustrates that the polynomial model generally produced the best average performance on the training sets. This is compared to the power model that produce the best average performance across the test sets, with competitive average performance on the training sets.

Additionally, the polynomial model also produced the worst performance on the test sets. This paired with its performance on the training sets is indicative of possible overfitting to the training data. Furthermore, Table C-6 illustrates that the power model produced the lowest standard deviation across all folds on the majority of the metrics on the test set, with competitively low standard deviation on the training set. As such, the power model was selected for the compressed air electricity usage forecasting of shaft E.

The summary of results showing the final models selected for all systems is shown in Table 3.11.

Model verification

With the final models selected by comparing average k-fold cross-validation accuracies, it was possible to fit the models to the total training set. These final models were validated using the

verification set that was split from the total dataset at the start of the process. The purpose of this is to verify whether the final models perform with similar accuracy on the training and verification sets. This would imply that the models are capable of performing well on unseen data. The final models, only for univariate models, are shown in Figure C-14.

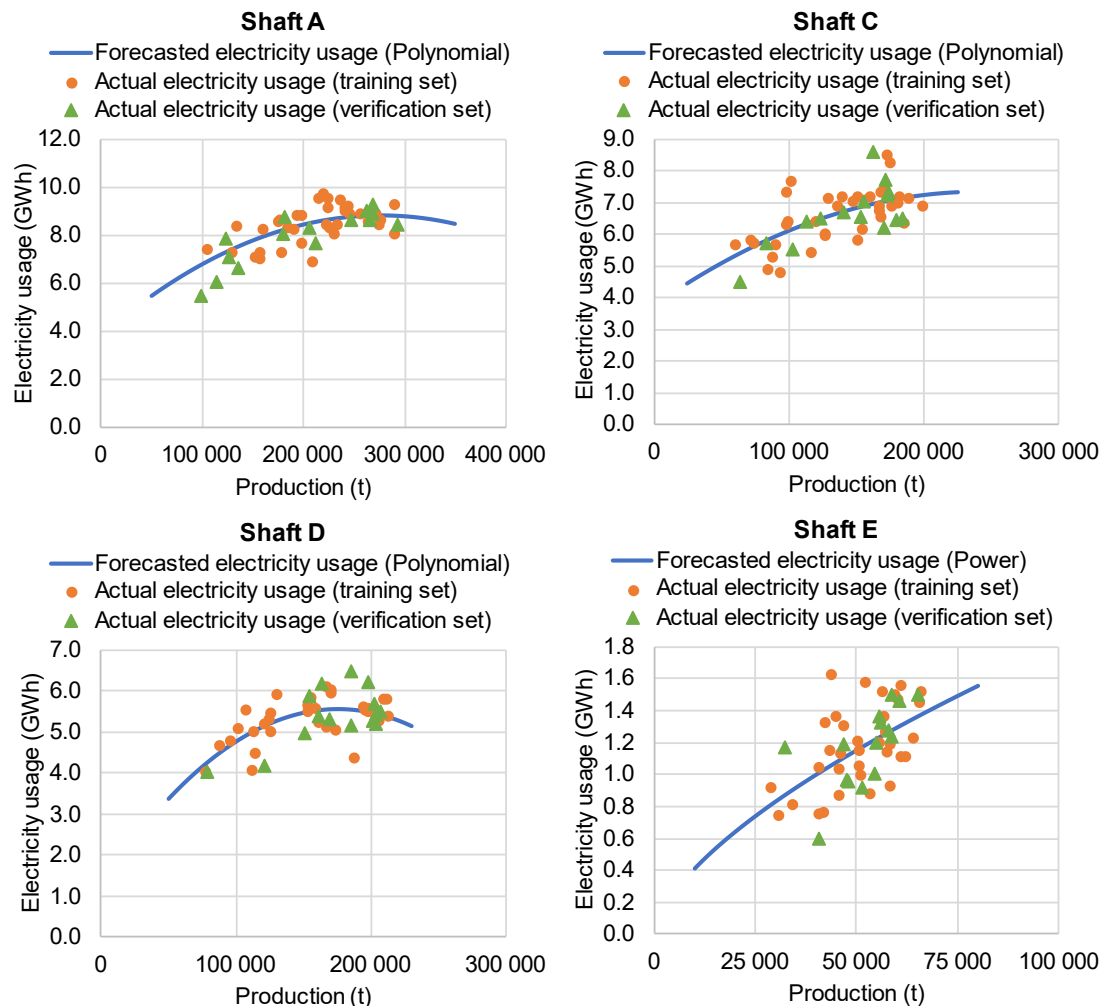


Figure C-14: Final models for the forecasting of the compressed air electricity consumption of shafts A, C, D and E plotted against the training and verification sets

Figure C-14 illustrates the final forecasting models for univariate systems plotted against the training sets and verification sets for each system. Figure C-14 illustrates that the verification and training sets both span the total operating range of each of the systems. Therefore, the verification set can be considered to be representative. Table C-7 contains the training set and verification set accuracies for the models shown in Figure C-14. These accuracies were compared to ensure that the model has comparative performance on both sets.

Table C-7: Training set and verification set performance metrics of the final models forecasting the compressed air electricity consumption of shafts A to E

Performance metric	MAPE (%)	R ²	RMSE (kWh)	MAE (kWh)
Shaft A				
Training set	6.19	0.360	615 785	508 858
Verification set	6.91	0.688	613 362	486 337
Shaft B				
Training set	14.30	-1.792	517 229	441 508
Verification set	13.73	-3.444	490 063	421 026
Shaft C				
Training set	8.10	0.393	660 289	520 029
Verification set	7.38	0.528	639 261	478 263
Shaft D				
Training set	5.73	0.407	398 137	290 047
Verification set	5.81	0.532	371 958	315 490
Shaft E				
Training set	15.07	0.296	210 665	172 431
Verification set	15.95	0.382	190 297	158 533

Table C-7 indicates that all models produced similar performance across all metrics on both the training and verification sets. With some models producing better performance on the verification set than on the training set, these are specifically the models for the compressed air electricity usage of shaft A, shaft B, shaft C and shaft D. This indicates that all of the models are capable of producing similar performance on unseen data and can therefore be used for forecasting the system compressed air electricity usage.

A final step in the verification of the final models is the verification of the initial model assumptions. This is performed by evaluating the similarity of the predicted data points to the actual data points used for training and verification with the use of a scatter plot. Furthermore, a residual plot and distribution plot of residuals is used to validate the assumption of heteroskedasticity and normally distributed residuals. These plots are shown Figure C-15, Figure C-16, Figure C-17, Figure C-18 and Figure C-19.

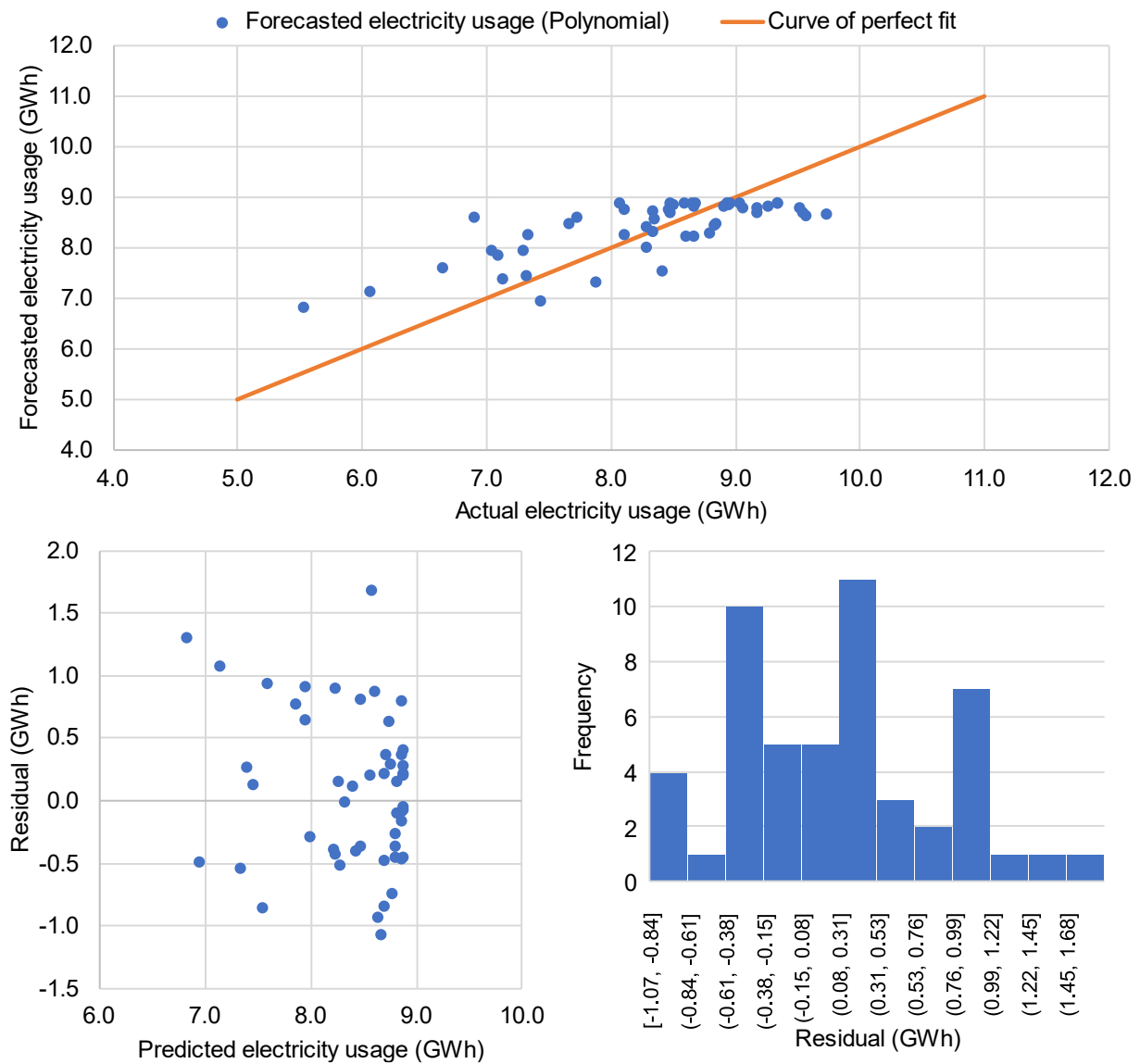


Figure C-15: Scatter plot ($R^2=0.753$), residual plot and residual distribution plot of shaft A compressed air electricity consumption to evaluate model assumptions

The scatter plot in Figure C-15 illustrates that the data points lie close to the curve of perfect fit and produce a correlation coefficient of 0.753. This is indicative of a strong correlation between values and illustrates that the forecasting model produces values close to the actual values. The residual distribution plot at the bottom right of Figure C-15 illustrates that the residuals have a normal distribution centred around zero and that the error ranges from -1.07 GWh to 1.68 GWh. This satisfies the assumption of normally distributed residuals.

Finally, the residual plot on the bottom left of Figure C-15 illustrates no relationship between residuals and predicted values. Furthermore, these residuals have the same range across the range of predicted values. Therefore, the assumption of constant variance is also satisfied for this model.

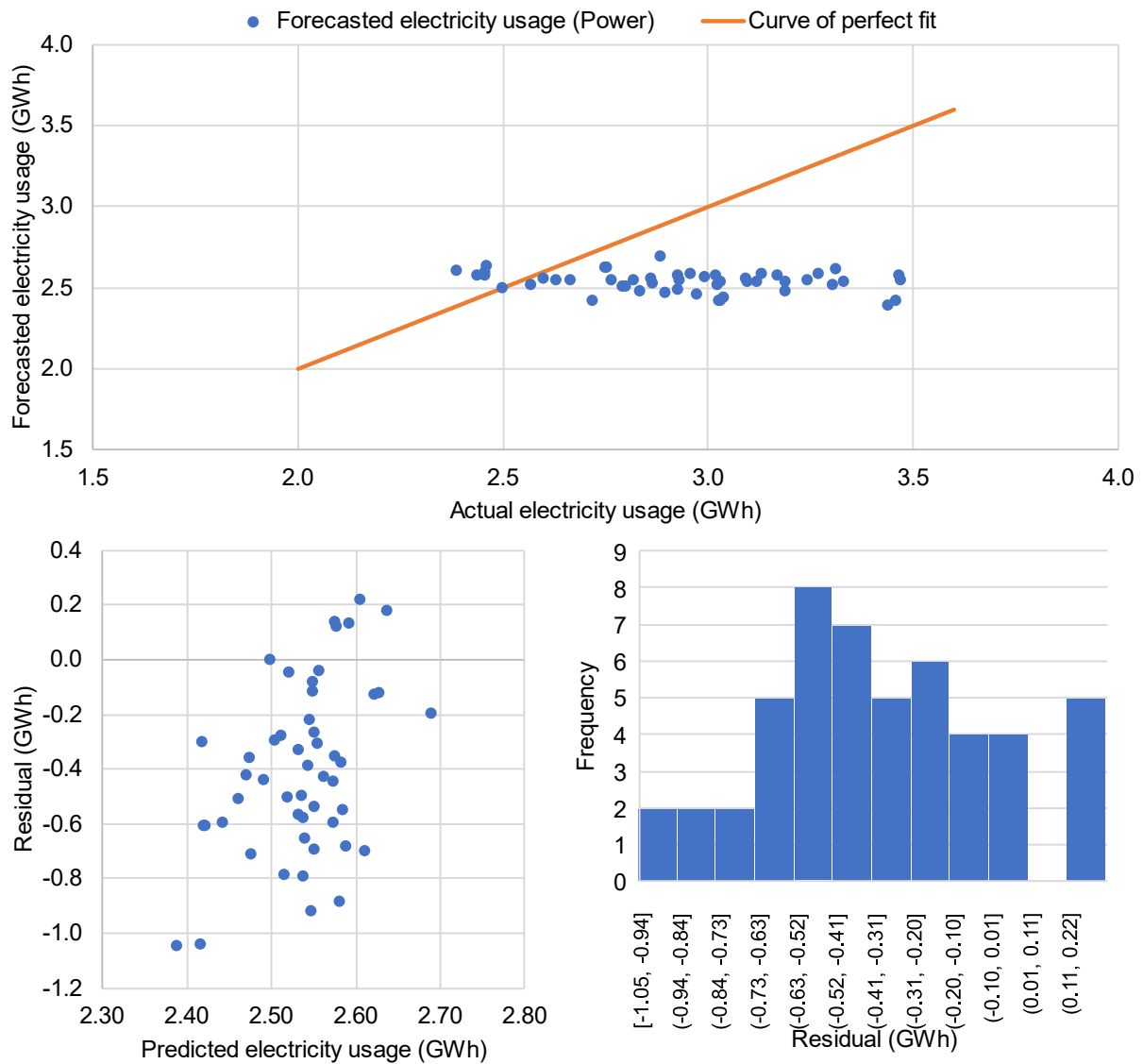


Figure C-16: Scatter plot ($R^2=-0.263$), residual plot and residual distribution plot of shaft B compressed air electricity consumption to evaluate model assumptions

Figure C-16 illustrates that the forecasted data points do not accurately track the actual values for the compressed air electricity forecasting of shaft B. This is confirmed by the correlation coefficient of -0.263. This indicates poor correlation between the actual and forecasting points. This is, however, not unexpected as this model was fit with poorly correlated input data. The model inputs should be revisited in order to produce a more competitive model.

Additionally, the residual distribution plot at the bottom right of Figure C-16 shows a normal distribution, therefore satisfying the assumption of normally distributed residuals for this model. These residuals, however, are not centred around zero and have a range from -1.05 GWh to 0.22 GWh. This implies that this model underpredicts the electricity consumption significantly.

Finally, the residual plot on the bottom left of Figure C-16 illustrates no relationship between the residuals and the predicted values. Furthermore, these residuals do not vary across the predicted value range. Therefore the assumption of constant variance is satisfied for this model.

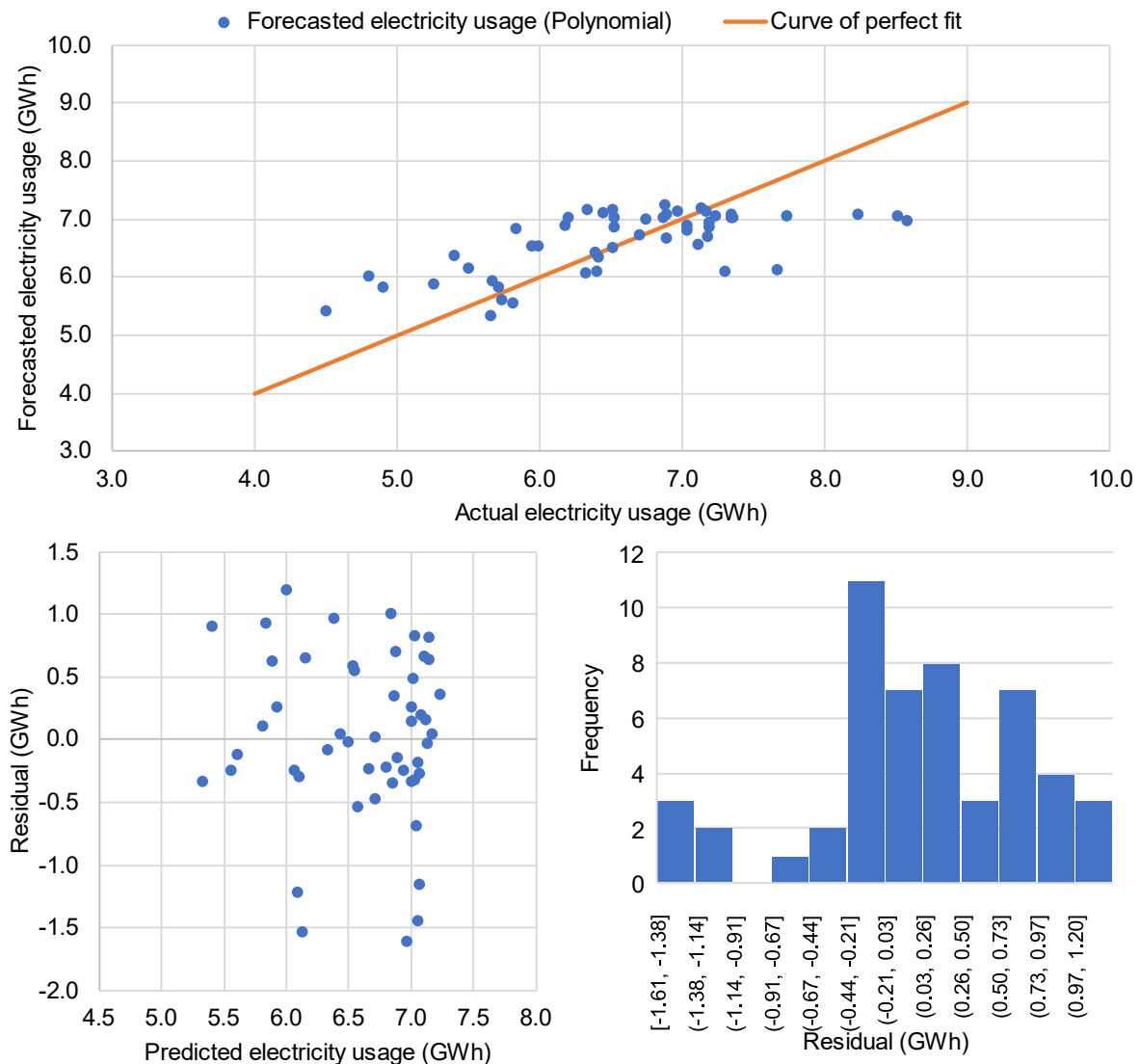


Figure C-17: Scatter plot ($R^2=0.665$), residual plot and residual distribution plot of shaft C compressed air electricity consumption to evaluate model assumptions

The scatter plot in Figure C-17 illustrates that majority of the data points follow the curve of perfect fit, with majority of the points lying above the curve. Furthermore, at high values of actual electricity usage the forecasted values plateau and severely underpredict the electricity usage. The correlation coefficient for this data is 0.665 which describes moderate correlation, indicating that this model performs adequately at forecasting the actual electricity usage.

The residual distribution plot illustrates a normal distribution with majority of the datapoints centred around zero therefore satisfying the assumption of normally distributed residuals. The range of

error for this data is -1.61 GWh to 1.20 GWh illustrating that the model over and underpredicts with similar error.

Finally, the residual plot shown at the bottom left of Figure C-17 illustrates no relationship between the residuals and the predicted values. Additionally, the points with large negative values are the same plateaued points shown in the scatter plot. The assumption of constant variance is therefore satisfied for this model.

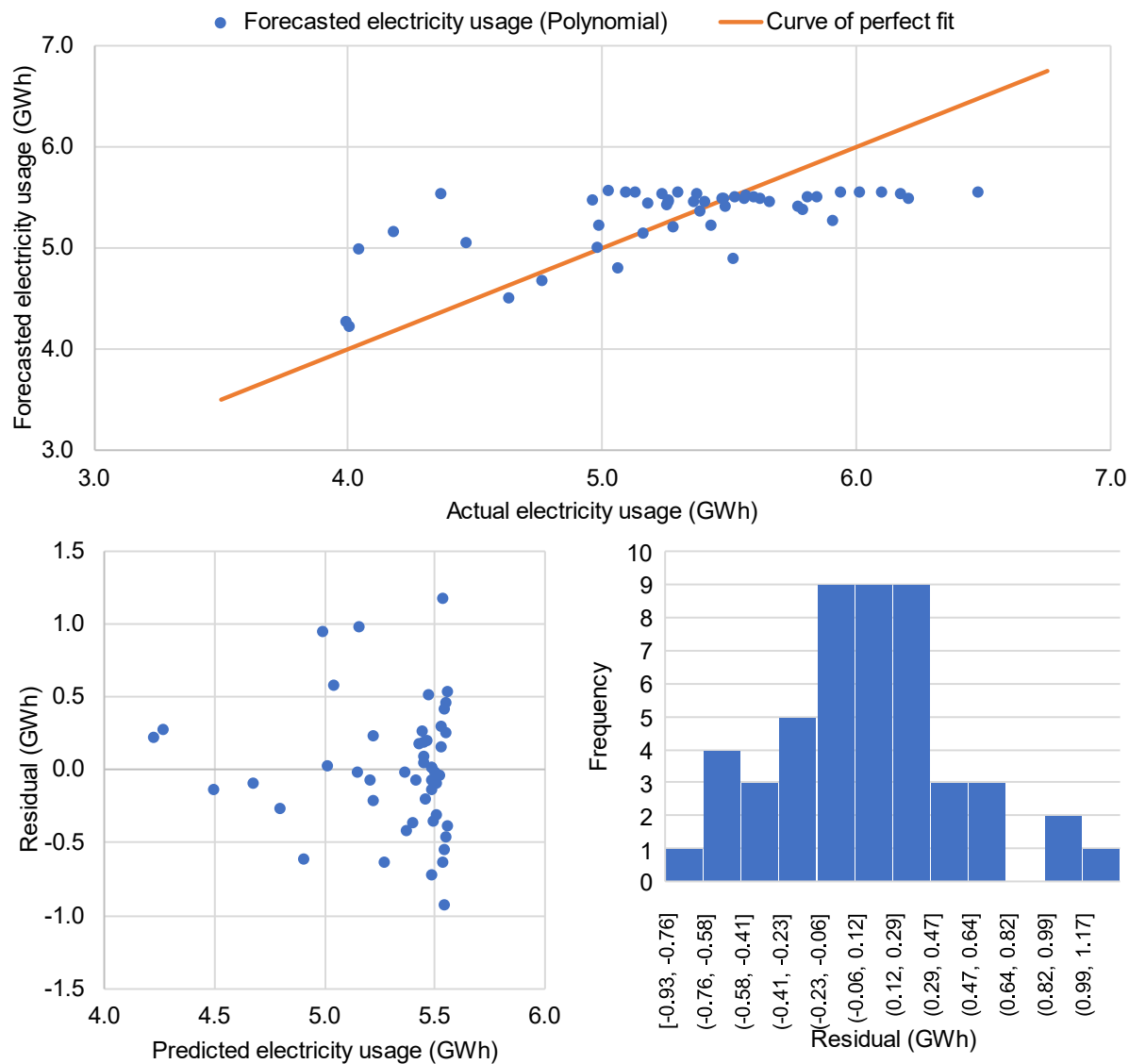


Figure C-18: Scatter plot ($R^2=0.657$), residual plot and residual distribution plot of shaft D compressed air electricity consumption to evaluate model assumptions

The scatter plot in Figure C-18 illustrates that the model overpredicts the electricity usage at low actual values and underpredicts at high actual values. Furthermore, the forecasted values plateau and are constant at around 5.5 GWh while the actual values range from 5 to 6.5 GWh. This is indicative of high variability of actual values. This model, however, still produces a correlation

coefficient of 0.657 indicated moderate correlation and decent performance at forecasting the compressed air electricity usage of shaft D.

The residual distribution plot illustrates a normal distribution centred around zero with a residual range from -0.93 GWh to 1.17 GWh. This illustrates that the model over and underpredicts by similar amounts. The assumption of normal residual distribution is therefore satisfied for this model.

Finally, the residual plot at the bottom left of Figure C-18 illustrates no relationship between the residuals and the forecasted values. The large number of residuals accruing at a predicted value of 5.5 GWh is indicative of the high variability of actual electricity values at similar values of production. This model, however, produces constant variance and as such this assumption is satisfied.

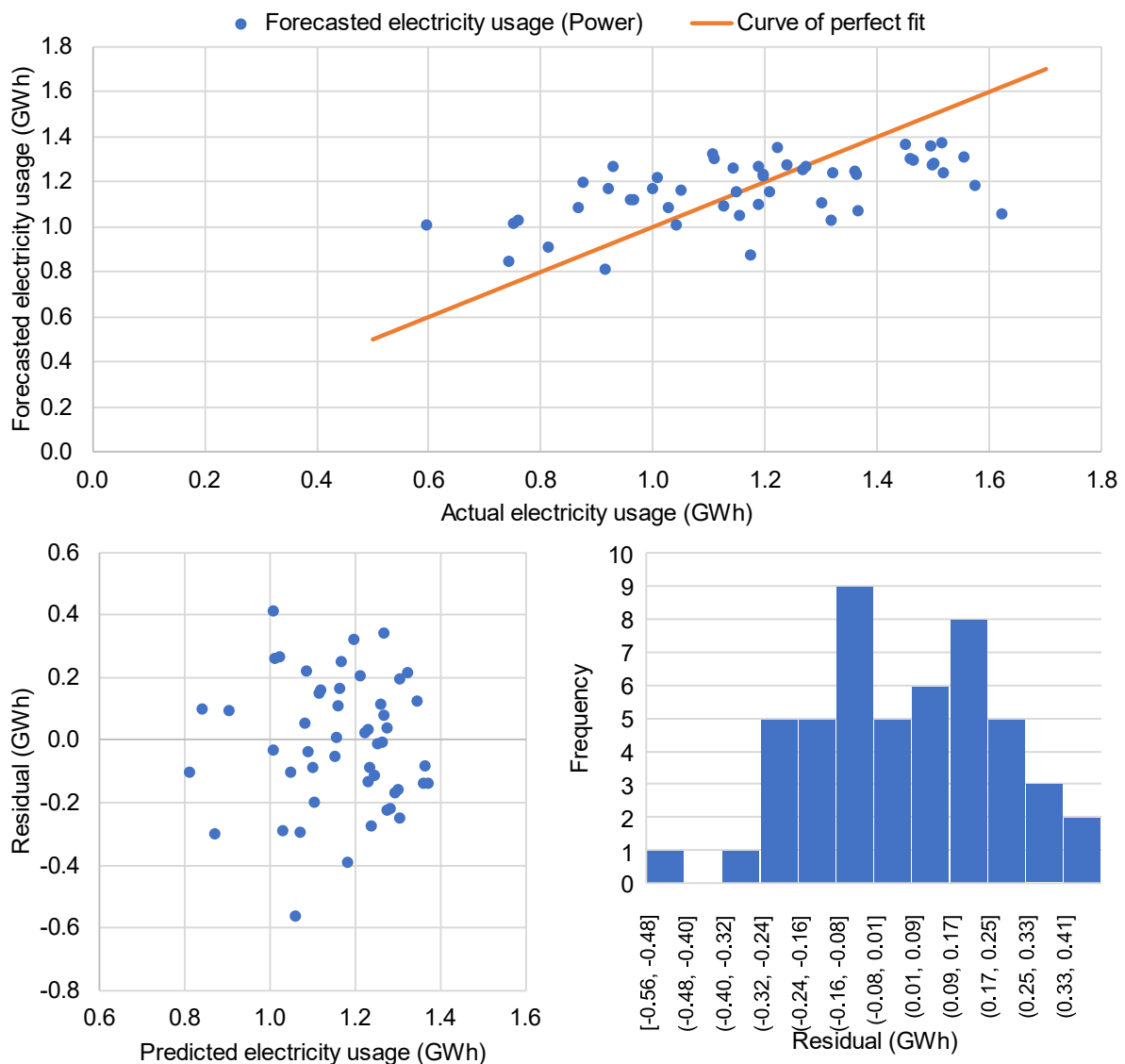


Figure C-19: Scatter plot ($R^2=0.568$), residual plot and residual distribution plot of shaft D compressed air electricity consumption to evaluate model assumptions

Figure C-19 illustrates that the points on the scatter plot fall close to the curve of perfect fit. Furthermore, the model underpredicts at low actual values and overpredicts at high actual values. This model produces a correlation coefficient of 0.568 with the actual values, indicating moderate correlation. Therefore, this model performs adequately at forecasting the compressed air electricity usage of shaft E.

The residual distribution plot illustrates that the assumption of normally distributed residuals is satisfied for this model. This graph illustrates a normal distribution centred around zero with residuals ranging from -0.56 GWh to 0.41 GWh. This indicates that the model over and underpredicts by similar amounts.

Finally, the residual plot on the bottom left of Figure C-19 illustrates that the constant variance assumption is satisfied for this model. This is because the residuals have no relationship to the predicted values, as well as the fact that the residuals are constant across the range of predicted values.

Appendix D – Ventilation electricity consumption models

Data collection and preparation

Outlier detection

The Mahalanobis distance was calculated for all data points for the ventilation electricity consumption forecasting of shaft C. This was used together with the boxplot method to identify outlier data points from the multivariate dataset. The graph illustrating the distance outliers are shown in Figure D-1 below.

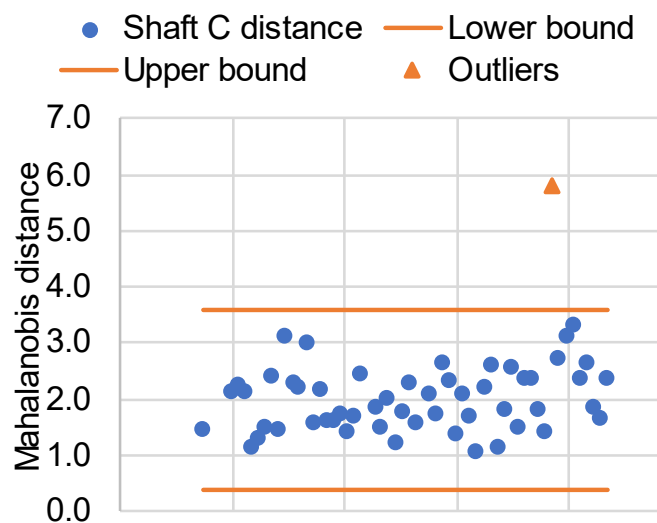


Figure D-1: Outlier detection using the Mahalanobis distance and boxplot method on ventilation electricity usage for shaft C

Figure D-1 illustrates the retained data points, the outlier data points as well as the quartile 1 and quartile 3 boundaries used to identify the outliers. The graph indicates that a single outlier was identified for the ventilation usage of shaft C. Furthermore, the graph indicates that the distances of the retained data points are clustered closely together. The removal of the outliers serves to ensure that the final data set is clean, reducing the amount of unnecessary error that would result from the presence of non-representative data points.

Data transformation

Once the outliers were removed, it was necessary to scale the data. This is done in order to prevent large differences in the units of input variables from causing weighting differences between the variables. Two scaling methods were considered, min-max scaling and z-score scaling. The former is used for datasets that are not normally distributed, while the latter is used on normally distributed data sets. Therefore, prior to scaling, it was necessary to evaluate the data distributions of all variables that were scaled – this is shown in Figure D-2.

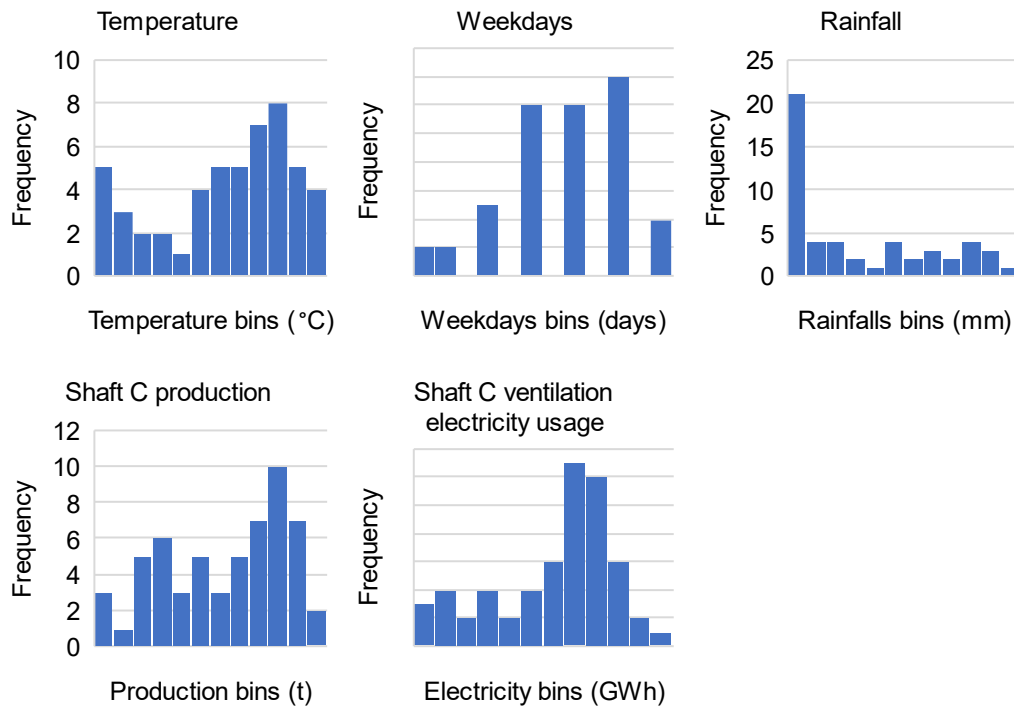


Figure D-2: Variable distributions for the determination of the most applicable scaling method for ventilation electricity usage variables of shaft C

Figure D-2 shows the distribution of the input and output variables that were used to generate models for forecasting the ventilation electricity usage. The scaling method used for each variable is dictated by the distribution of the data and is shown in Table B-1.

Variable	Min-max scaling	Z-score scaling
Temperature	X	
Weekdays		X
Rainfall	X	
Shaft C production		X
Shaft C ventilation electricity consumption	X	

Data splitting

Once the data was cleaned of outliers and transformed, it was necessary to split it into train/test and verification sets. The datasets were first randomised in order to ensure that split datasets were representative of the total dataset. The resulting train/test set was used with k-fold cross-verification to train the resulting model.

Furthermore, the separation of a verification set was necessary to ensure that the resulting model was tested on unseen data. The data points that comprised the train/test and verification sets for the forecasting of the ventilation electricity usage of shaft C are shown in Figure D-3.

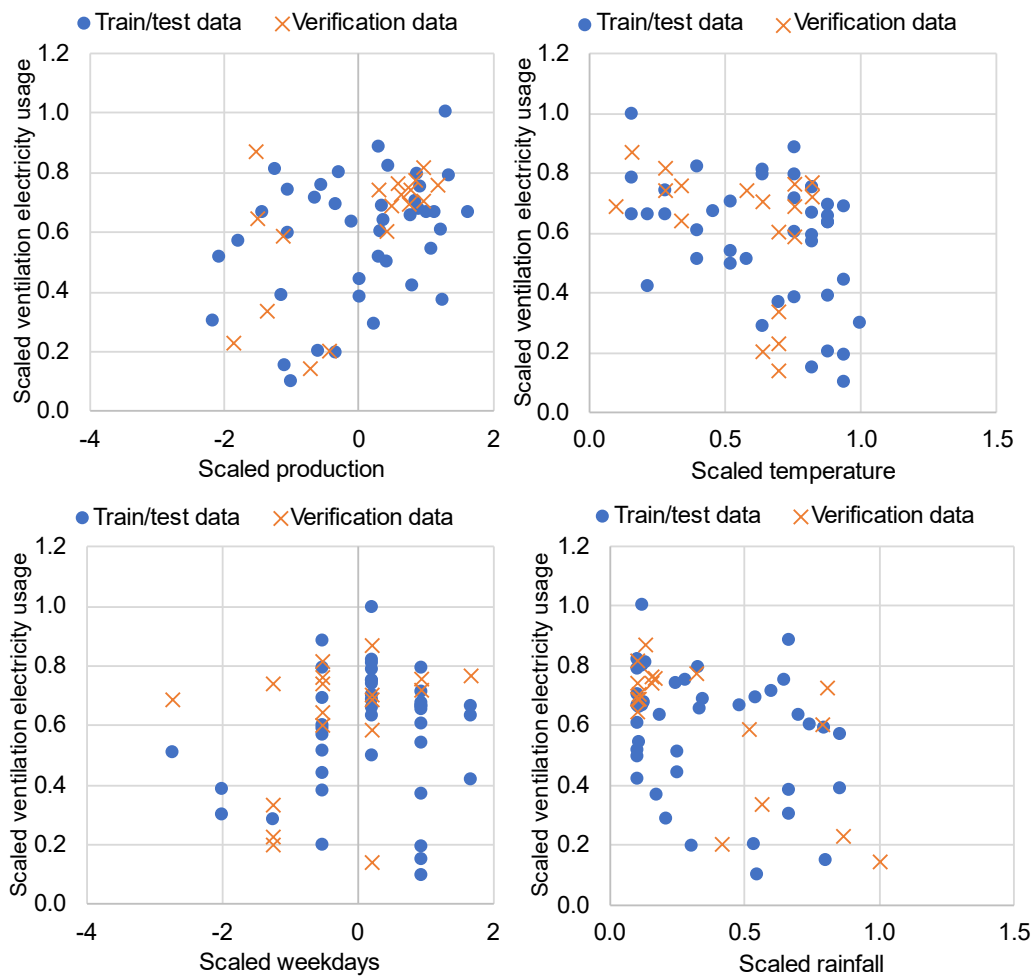


Figure D-3: Data splits showing the train/test and verification sets for shaft C

Figure D-3 illustrates that the values of the variables in the train/test and verification sets for shaft C span the range of each variable. This shows that the values used in each set are representative of the dataset and will result in accuracies that can be considered valid. This check is necessary in order to ensure that the verification and train/test data sets are similar in distribution.

Model development

Variable correlation assessment

After cleaning and splitting of the datasets it was possible to begin with the model development step of the process. This step began with the determination of input variables that were most strongly correlated to the ventilation electricity usage of shaft C. This determination was performed with the use of Pearson's correlation coefficient. According to Table 1.8, variables with absolute correlation coefficients less than 0.4 are considered to have weak to negligible correlation.

As such, only input variables with absolute correlation coefficients larger than or equal to 0.4, to the ventilation electricity usage of shaft C were considered. It is necessary to remove variables with low correlation as the inclusion of these can have an adverse effect on model performance. The scatter plots between variables and associated Pearson's correlation coefficients are shown in Figure D-4 below.

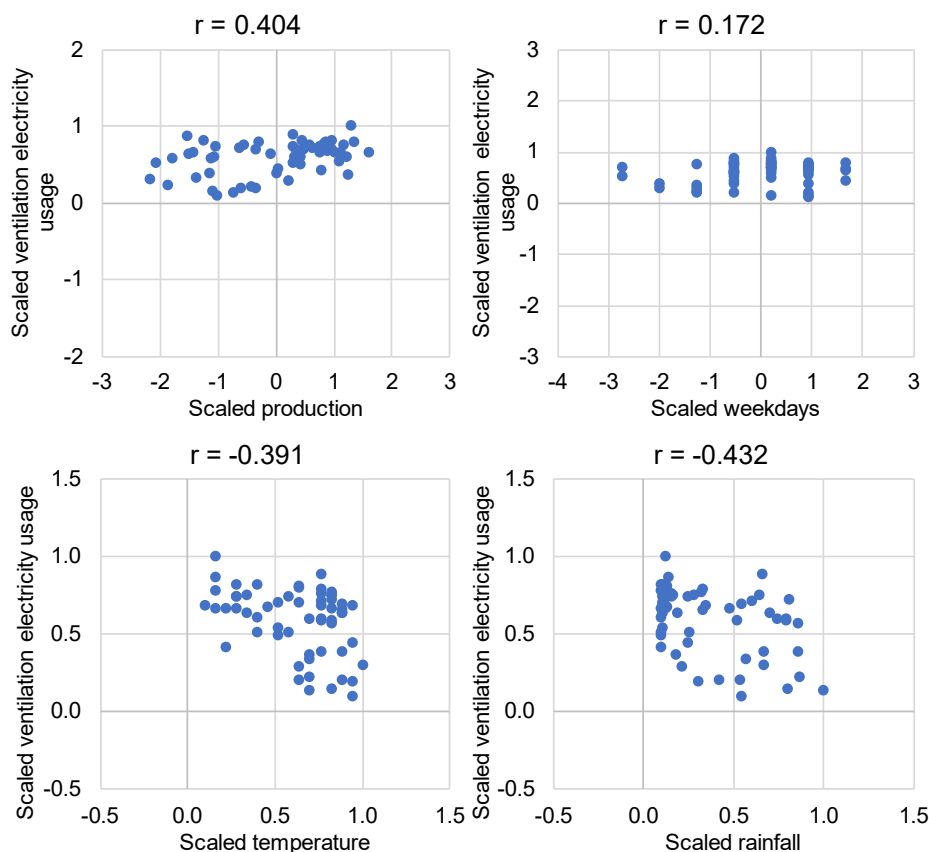


Figure D-4: Variable correlation graphs and Pearson's correlation coefficients for the ventilation electricity usage of shaft C

Figure D-4 illustrates that production, temperature and rainfall have absolute Pearson's correlation coefficients close to 0.4 when correlated to ventilation electricity usage of shaft C. Since none of the variables have a particularly strong correlation to the output variable and all variables produce similar correlation coefficients, all variables were considered for the modelling of the ventilation electricity usage of shaft C.

Model fitting and testing

With the splitting of the datasets and the selection of the relevant variables completed, it was possible to fit the models to the data. This was done with the use of k-fold cross-validation. As explained in section 2.4.3, four folds were selected with the model being trained on three of the folds and tested on the remaining fold. This was performed four times in order to remove bias that would have caused accuracy misrepresentation.

Furthermore, a variety of models were considered and compared in order to allow the selection of the most accurate model for each system. These models included power, linear and polynomial fits. The average and standard deviation of the accuracies across the four folds with the use of four performance metrics were used for model selection. The results for the modelling of the ventilation electricity usage for the shaft C is shown in Table D-1 with the best and worst performing models highlighted in green and red respectively for each metric.

Table D-1: Average and standard deviation of performance metrics calculated for the models fit to the train/test folds of shaft C ventilation electricity consumption

Metric	Set	Average			Standard deviation		
		Power	Linear	Poly	Power	Linear	Poly
MAPE (%)	Train	5.11%	8.26%	119.03%	1.46%	3.00%	64.18%
	Test	5.50%	7.66%	134.35%	0.75%	3.38%	85.86%
R ²	Train	-0.375	-2.861	-933.616	0.505	2.660	623.847
	Test	-0.934	-2.014	-2156.691	1.404	2.126	2614.963
RMSE (kWh)	Train	202 483	315 396	4 740 413	49 714	113 812	2 607 903
	Test	204 956	285 397	5 223 572	26 530	127 869	3 308 274
MAE (kWh)	Train	169 314	271 773	3 977 517	52 001	103 580	2 160 214
	Test	175 640	243 906	4 360 805	27 475	111 690	2 804 124

Table D-1 illustrates that the power model produced the best average performance across all performance metrics on the training and test sets. Furthermore, the polynomial model produced the worst performance across all metrics for the training and test sets. Additionally, the power model produced the lowest standard deviation on both the training and test sets across all metrics,

with the polynomial model producing the highest standard deviation. Therefore, the power model was selected for the forecasting of this system's electricity usage.

Model verification

With the final models selected by comparing average k-fold cross-validation accuracies, it was possible to fit the models to the total training set. These final models were validated using the verification set that was split from the total dataset at the start of the process. The purpose of this is to verify whether the final models perform with similar accuracy on the training and verification sets. This would imply that the models are capable of performing well on unseen data.

Table D-2 contains the training set and verification set accuracies for the model used for the prediction of the ventilation electricity usage of shaft C. These accuracies were compared to ensure that the model has comparative performance on both sets.

Table D-2: Training set and verification set performance metrics of the final model forecasting the ventilation electricity consumption of shaft C

Performance metric	MAPE (%)	R ²	RMSE (kWh)	MAE (kWh)
Shaft C				
Training set	5.65	-0.430	209 189	183 207
Verification set	7.16	-0.764	244 937	231 418

Table D-2 illustrates the power model used for the forecasting the ventilation electricity usage of shaft C performs with similar accuracy across all metrics on both the training and verification sets. This illustrates that the model is capable of producing reasonable forecasts on unseen data and can therefore be used for forecasting the system electricity usage.

A final step in the verification of the final model is the verification of the initial model assumptions. This is performed by evaluating the similarity of the predicted data points to the actual data points used for training and verification with the use of a scatter plot. Furthermore, a residual plot and distribution plot of residuals is used to validate the assumption of heteroskedasticity and normally distributed residuals. These plots are shown in Figure D-5.

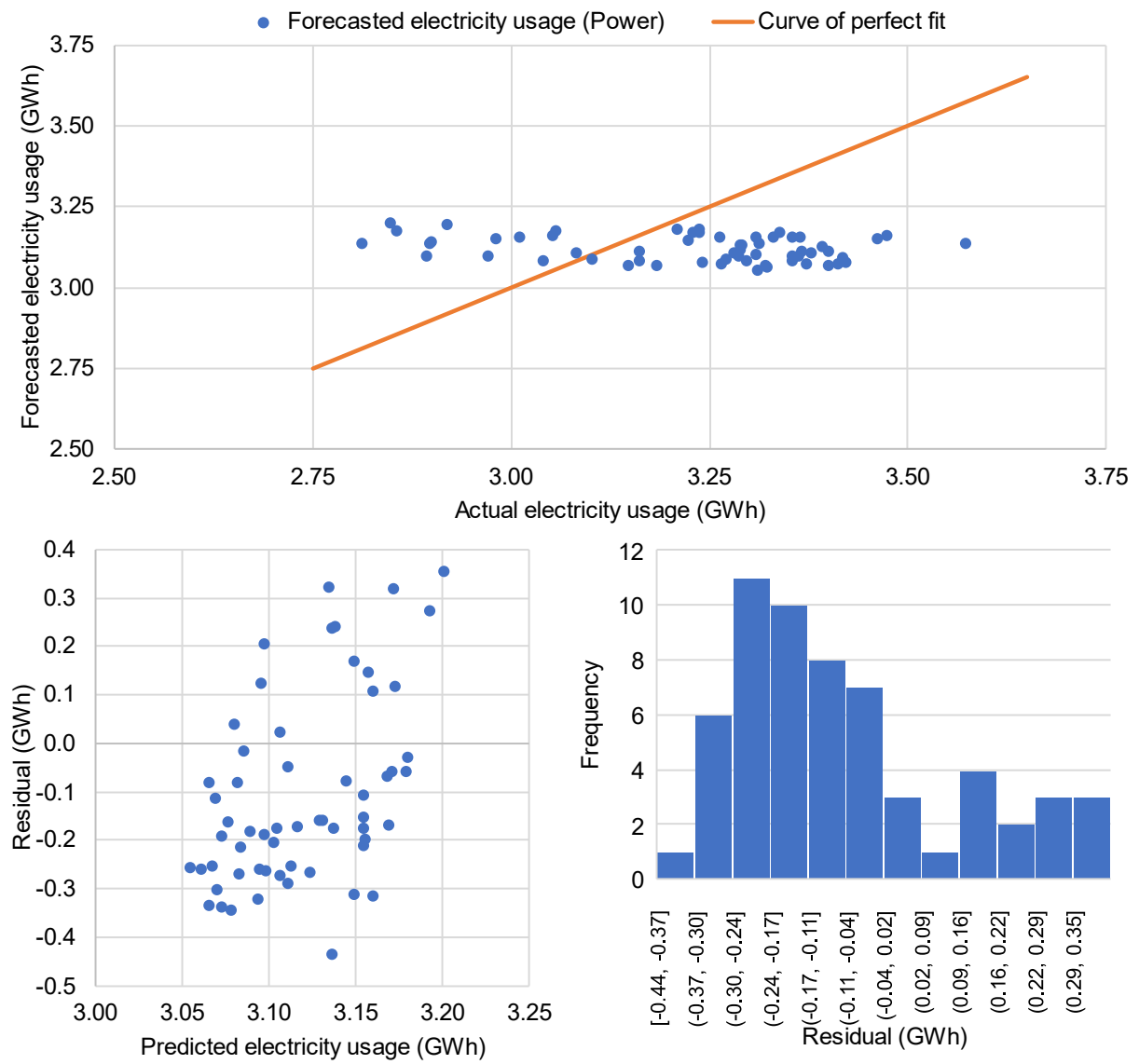


Figure D-5: Scatter plot, residual plot and residual distribution plot of shaft C ventilation electricity consumption to evaluate model assumptions